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*

Correction

The Annals of Applied Statistics
2016, Vol. 10, No. 2, 549–574
DOI: 10.1214/16-AOAS921
© Institute of Mathematical Statistics, 2016

WHAT WE LOOK AT IN PAINTINGS: A COMPARISON BETWEEN EXPERIENCED AND INEXPERIENCED ART VIEWERS

BY ANNA-KAISA YLITALO[*,1], AILA SÄRKKÄ[†,2] AND PETER GUTTORP[‡,§]

University of Jyvaskyla[*],
Chalmers University of Technology and University of Gothenburg[†],
University of Washington[‡] and Norwegian Computing Center[§]

How do people look at art? Are there any differences between how experienced and inexperienced art viewers look at a painting? We approach these questions by analyzing and modeling eye movement data from a cognitive art research experiment, where the eye movements of twenty test subjects, ten experienced and ten inexperienced art viewers, were recorded while they were looking at paintings.

Eye movements consist of stops of the gaze as well as jumps between the stops. Hence, the observed gaze stop locations can be thought of as a spatial point pattern, which can be modeled by a spatio-temporal point process. We introduce some statistical tools to analyze the spatio-temporal eye movement data, and compare the eye movements of experienced and inexperienced art viewers. In addition, we develop a stochastic model, which is rather simple but fits quite well to the eye movement data, to further investigate the differences between the two groups through functional summary statistics.

REFERENCES

ANONYMOUS (2000). Brochure of Art Centre Salmela. Mäntyharju, Finland.

ANTES, J. R. (1974). The time course of picture viewing. *J. Exp. Psychol.* **103** 62–70.

BADDELEY, A., RUBAK, E. and TURNER, R. (2015). *Spatial Point Patterns: Methodology and Applications with R*. Chapman & Hall/CRC Press, London. Available at http://www.taylorandfrancis.com/books/details/9781482210200/.

BAILEY, T. C. and GATRELL, A. C. (1995). *Interactive Spatial Data Analysis*. Longman, Harlow.

BARLOW, H. B. (1952). Eye movements during fixation. *J. Physiol.* **116** 290–306.

BARNARD, G. A. (1963). Contribution to the discussion of professor Bartlett's paper. *J. Roy. Statist. Soc. Ser. A* **25** 294.

BARTHELMÉ, S., TRUKENBROD, H., ENGBERT, R. and WICHMANN, F. (2013). Modelling fixation locations using spatial point processes. *J. Vis.* **13** 1–34.

BERMAN, M. and DIGGLE, P. (1989). Estimating weighted integrals of the second-order intensity of a spatial point process. *J. Roy. Statist. Soc. Ser. B* **51** 81–92. MR0984995

BUSWELL, G. T. (1935). *How People Look at Pictures—A Study of the Psychology of Perception in Art*. The Univ. Chicago Press, Chicago, IL.

DIGGLE, P. J. (1985). A kernel method for smoothing point process data. *J. R. Stat. Soc. Ser. C.* **34** 138–147.

DOKSUM, K. A. and SIEVERS, G. L. (1976). Plotting with confidence: Graphical comparisons of two populations. *Biometrika* **63** 421–434. MR0443210

Key words and phrases. Coverage, intensity, point process, shift function, transition probability.

DUCHOWSKI, A. T. (2002). A breadth-first survey of eye tracking applications. *Behav. Res. Methods Instrum. Comput.* **34** 455–470.

ENGBERT, R., TRUKENBROD, H. A., BARTHELMÉ, S. and WICHMANN, F. A. (2015). Spatial statistics and attentional dynamics in scene viewing. *J. Vis.* **15** 1–17.

FINDLAY, J. M. (2009). Saccadic eye movement programming: Sensory and attentional factors. *Psychol. Res.* **73** 127–135.

GATRELL, A. C., BAILEY, T. C., DIGGLE, P. J. and ROWLINGSON, B. S. (1996). Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British Geographers, New Series* **21** 256–274.

HAZELTON, M. L. (2007). Kernel estimation of risk surfaces without the need for edge correction. *Stat. Med.* **27** 2269–2272.

HENDERSON, J. M. and HOLLINGWORTH, A. (1999). High-level scene perception. *Annu. Rev. Psychol.* **50** 243–271.

KELSALL, J. E. and DIGGLE, P. J. (1995a). Kernel estimation of relative risk. *Bernoulli* **1** 3–16. MR1354453

KELSALL, J. E. and DIGGLE, P. J. (1995b). Non-parametric estimation of spatial variation in relative risk. *Stat. Med.* **14** 2335–2342.

KINSLER, V. and CARPENTER, R. H. S. (1995). Saccadic eye movements while reading music. *Vis. Res.* **35** 1447–1458.

KOMOGARTSEV, O. V., RYU, Y. S. and KOH, D. H. (2009). Quick models for saccade amplitude prediction. *Journal of Eye Movement Research* **1** 1–13.

KRISTJANSON, A. F. and ANTES, J. R. (1989). Eye movement analysis of artists and nonartists viewing paintings. *Vis. Arts Res.* **15** 21–30.

LOCHER, P. J. (2006). The usefulness of eye movement recordings to subject an aesthetic episode with visual art to empirical scrutiny. *Psychol. Sci.* **48** 106–114.

MACKWORTH, N. H. and MORANDI, A. J. (1967). The gaze selects informative details within pictures. *Atten. Percept. Psychophys.* **2** 547–552.

MANOR, B. R. and GORDON, E. (2003). Defining the temporal threshold for ocular fixation in free-viewing visuocognitive tasks. *J. Neurosci. Methods* **128** 85–93.

MIKKOLA, K., ED. (1997). Risto Suomi. Amos Anderson Art Museum, publications, new series, no 25.

MYLLYMÄKI, M., MRKVICKA, T., GRABARNIK, P., SEIJO, H. and HAHN, U. (2016). Global envelope tests for spatial processes. *J. Roy. Statist. Soc. Ser. B*. DOI:10.1111/rssb.12172.

MYLLYMÄKI, M., GRABARNIK, P., SEIJO, H. and STOYAN, D. (2015). Deviation test construction and power comparison for marked spatial point patterns. *Spat. Stat.* **11** 19–34. MR3311854

NAGASAWA, S., YIM, S. and HONGO, H. (2005). Feasibility study on marketing research using eye movement: An investigation of image presentation using an "eye camera" and data processing. *J. Adv. Comput. Intell. Intell. Informa.* **9** 440–452.

NOTON, D. and STARK, L. (1971). Scanpaths in saccadic eye movements while viewing and recognizing patterns. *Vis. Res.* **11** 929–942.

PENTTINEN, A. and YLITALO, A.-K. (in press). Deducing self-interaction in eye movement data using sequential spatial point processes. *Spatial Statistics*. DOI:10.1016/j.spasta.2016.03.005.

THE YORCK PROJECT (2002). 10.000 Meisterwerke der Malerei, DVD-ROM, 2002. ISBN 3936122202. Distributed by DIRECTMEDIA Publishing GmbH.

RAYNER, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychol. Bull.* **124** 372–422.

RAYNER, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *Q. J. Exp. Psychol., A Hum. Exp. Psychol.* **62** 1457–1506.

RIPLEY, B. D. (1988). *Statistical Inference for Spatial Processes*. Cambridge Univ. Press, Cambridge. MR0971986

SUNDELL, D. (1986). *Eero Järnefelt (1863–1937), Retretti 25.5.–21.9.1986*. Retretti.
VOGT, S. and MAGNUSSEN, S. (2007). Expertise in pictorial perception: Eye-movement patterns and visual memory in artists and laymen. *Perception* **36** 91–100.
WADE, N. J. (2010). Pioneers of eye movement research. *I-Perception* **1** 33–68.
WAKEFIELD, J. C., KELSALL, J. E. and MORRIS, S. E. (2000). Clustering, cluster detection, and spatial variation in risk. In *Spatial Epidemiology*: *Methods and Applications* (P. Elliot, J. C. Wakefield, N. G. Best and D. J. Briggs, eds.) 128–152. Oxford Univ. Press, London.
WILK, M. B. and GNANADESIKAN, R. (1968). Probability plotting methods for the analysis of data. *Biometrika* **55** 1–17.
YARBUS, A. L. (1967). *Eye Movements and Vision*. Plenum Press, New York.
YLITALO, A.-K., SÄRKKÄ, A. and GUTTORP, P. (2016a). Supplement to "What we look at in paintings: A comparison between experienced and inexperienced art viewers." DOI:10.1214/16-AOAS921SUPPA.
YLITALO, A.-K., SÄRKKÄ, A. and GUTTORP, P. (2016b). Supplement to "What we look at in paintings: A comparison between experienced and inexperienced art viewers." DOI:10.1214/16-AOAS921SUPPB.
YLITALO, A.-K., SÄRKKÄ, A. and GUTTORP, P. (2016c). Supplement to "What we look at in paintings: A comparison between experienced and inexperienced art viewers." DOI:10.1214/16-AOAS921SUPPC.
YLITALO, A.-K., SÄRKKÄ, A. and GUTTORP, P. (2016d). Supplement to "What we look at in paintings: A comparison between experienced and inexperienced art viewers." DOI:10.1214/16-AOAS921SUPPD.
YLITALO, A.-K., SÄRKKÄ, A. and GUTTORP, P. (2016e). Supplement to "What we look at in paintings: A comparison between experienced and inexperienced art viewers." DOI:10.1214/16-AOAS921SUPPE.

The Annals of Applied Statistics
2016, Vol. 10, No. 2, 575–595
DOI: 10.1214/16-AOAS908
© Institute of Mathematical Statistics, 2016

PREDICTIVE MODELING OF CHOLERA OUTBREAKS IN BANGLADESH

BY AMANDA A. KOEPKE[*,1], IRA M. LONGINI, JR.[†,1], M. ELIZABETH HALLORAN[*,‡,1], JON WAKEFIELD[‡,2] AND VLADIMIR N. MININ[‡,3]

Fred Hutchinson Cancer Research Center[*], University of Florida[†] and University of Washington[‡]

Despite seasonal cholera outbreaks in Bangladesh, little is known about the relationship between environmental conditions and cholera cases. We seek to develop a predictive model for cholera outbreaks in Bangladesh based on environmental predictors. To do this, we estimate the contribution of environmental variables, such as water depth and water temperature, to cholera outbreaks in the context of a disease transmission model. We implement a method which simultaneously accounts for disease dynamics and environmental variables in a Susceptible-Infected-Recovered-Susceptible (SIRS) model. The entire system is treated as a continuous-time hidden Markov model, where the hidden Markov states are the numbers of people who are susceptible, infected or recovered at each time point, and the observed states are the numbers of cholera cases reported. We use a Bayesian framework to fit this hidden SIRS model, implementing particle Markov chain Monte Carlo methods to sample from the posterior distribution of the environmental and transmission parameters given the observed data. We test this method using both simulation and data from Mathbaria, Bangladesh. Parameter estimates are used to make short-term predictions that capture the formation and decline of epidemic peaks. We demonstrate that our model can successfully predict an increase in the number of infected individuals in the population weeks before the observed number of cholera cases increases, which could allow for early notification of an epidemic and timely allocation of resources.

REFERENCES

ANDRIEU, C., DOUCET, A. and HOLENSTEIN, R. (2010). Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 269–342. MR2758115

ANDRIEU, C. and ROBERTS, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.* **37** 697–725. MR2502648

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. Statist.* **41** 164–171. MR0287613

BEAUMONT, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics* **164** 1139–1160.

BHADRA, A., IONIDES, E. L., LANERI, K., PASCUAL, M., BOUMA, M. and DHIMAN, R. C. (2011). Malaria in Northwest India: Data analysis via partially observed stochastic differential equation models driven by Lévy noise. *J. Amer. Statist. Assoc.* **106** 440–451. MR2866974

Key words and phrases. Hidden Markov model, particle filter, MCMC, Bayesian, SIR.

BRETÓ, C., HE, D., IONIDES, E. L. and KING, A. A. (2009). Time series analysis via mechanistic models. *Ann. Appl. Stat.* **3** 319–348. MR2668710

CAO, Y., GILLESPIE, D. T. and PETZOLD, L. R. (2005). Avoiding negative populations in explicit Poisson tau-leaping. *J. Chem. Phys.* **123** 054104.

CAUCHEMEZ, S. and FERGUSON, N. M. (2008). Likelihood-based estimation of continuous-time epidemic models from time-series data: Application to measles transmission in London. *J. R. Soc. Interface* **5** 885–897.

CODEÇO, C. (2001). Endemic and epidemic dynamics of cholera: The role of the aquatic reservoir. *BMC Infect. Dis.* **1** 1.

COLWELL, R. R. and HUQ, A. (1994). Environmental reservoir of Vibrio cholerae: The causative agent of cholera. *Ann. N.Y. Acad. Sci.* **740** 44–54.

DIEKMANN, O., HEESTERBEEK, J. A. P. and METZ, J. A. J. (1990). On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. *J. Math. Biol.* **28** 365–382. MR1057044

DOUCET, A., DE FREITAS, N. and GORDON, N., eds. (2001). *Sequential Monte Carlo Methods in Practice*. Springer, New York. MR1847783

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

EDDELBUETTEL, D. (2013). *Seamless R and C++ Integration with Rcpp*. Springer, New York.

EDDELBUETTEL, D. and FRANÇOIS, R. (2011). Rcpp: Seamless R and C++ integration. *J. Stat. Softw.* **40** 1–18.

EISENBERG, M. C., ROBERTSON, S. L. and TIEN, J. H. (2013). Identifiability and estimation of multiple transmission pathways in cholera and waterborne disease. *J. Theoret. Biol.* **324** 84–102. MR3041644

FEARNHEAD, P., GIAGOS, V. and SHERLOCK, C. (2014). Inference for reaction networks using the linear noise approximation. *Biometrics* **70** 457–466. MR3258050

FERM, L., LÖTSTEDT, P. and HELLANDER, A. (2008). A hierarchy of approximations of the master equation scaled by a size parameter. *J. Sci. Comput.* **34** 127–151. MR2373035

FINKENSTÄDT, B. F. and GRENFELL, B. T. (2000). Time series modelling of childhood diseases: A dynamical systems approach. *J. Roy. Statist. Soc. Ser. C* **49** 187–205. MR1821321

GILLESPIE, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.* **22** 403–434. MR0503370

GILLESPIE, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81** 2340–2361.

GILLESPIE, D. T. (2001). Approximate accelerated stochastic simulation of chemically reacting systems. *J. Chem. Phys.* **115** 1716–1733.

GINSBERG, J., MOHEBBI, M. H., PATEL, R. S., BRAMMER, L., SMOLINSKI, M. S. and BRILLIANT, L. (2008). Detecting influenza epidemics using search engine query data. *Nature* **457** 1012–1014.

HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109.

HE, D., IONIDES, E. L. and KING, A. A. (2010). Plug-and-play inference for disease dynamics: Measles in large and small populations as a case study. *J. R. Soc. Interface* **7** 271–283.

HELD, L., HÖHLE, M. and HOFMANN, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Stat. Model.* **5** 187–199. MR2210732

HUQ, A., COLWELL, R. R., RAHMAN, R., ALI, A., CHOWDHURY, M. A., PARVEEN, S., SACK, D. A. and RUSSEK-COHEN, E. (1990). Detection of Vibrio cholerae O1 in the aquatic environment by fluorescent-monoclonal antibody and culture methods. *Appl. Environ. Microbiol.* **56** 2370–2373.

HUQ, A., SACK, R. B., NIZAM, A., LONGINI, I. M., NAIR, G. B., ALI, A., MORRIS, J. G. JR, KHAN, M. N., SIDDIQUE, A. K., YUNUS, M., ALBERT, M. J., SACK, D. A. and COLWELL, R. R. (2005). Critical factors influencing the occurrence of Vibrio cholerae in the environment of Bangladesh. *Appl. Environ. Microbiol.* **71** 4645–4654.

INTERNATIONAL VACCINE INSTITUTE (2012). Country investment case study on cholera vaccination: Bangladesh. International Vaccine Institute, Seoul.

IONIDES, E. L., BRETÓ, C. and KING, A. A. (2006). Inference for nonlinear dynamical systems. *Proc. Natl. Acad. Sci. USA* **103** 18438–18443.

KEELING, M. J. and ROHANI, P. (2008). *Modeling Infectious Diseases in Humans and Animals*. Princeton Univ. Press, Princeton, NJ. MR2354763

KEELING, M. J. and ROSS, J. V. (2008). On methods for studying stochastic disease dynamics. *J. R. Soc. Interface* **5** 171–181.

KING, A. A., IONIDES, E. L., PASCUAL, M. and BOUMA, M. J. (2008). Inapparent infections and cholera dynamics. *Nature* **454** 877–880.

KOELLE, K. and PASCUAL, M. (2004). Disentangling extrinsic from intrinsic factors in disease dynamics: A nonlinear time series approach with an application to cholera. *Amer. Nat.* **163** 901–913.

KOELLE, K., RODÓ, X., PASCUAL, M., YUNUS, M. and MOSTAFA, G. (2005). Refractory periods and climate forcing in cholera dynamics. *Nature* **436** 696–700.

KOEPKE, A. A., LONGINI, JR., I. M., HALLORAN, M., WAKEFIELD, J. and MININ, V. N. (2016). Supplement to "Predictive modeling of cholera outbreaks in Bangladesh." DOI:10.1214/16-AOAS908SUPP.

KOMOROWSKI, M., FINKENSTÄDT, B., HARPER, C. V. and RAND, D. A. (2009). Bayesian inference of biochemical kinetic parameters using the linear noise approximation. *BMC Bioinformatics* **10** 343.

LONGINI, I. M., YUNUS, M., ZAMAN, K., SIDDIQUE, A. K., SACK, R. B. and NIZAM, A. (2002). Epidemic and endemic cholera trends over a 33-year period in Bangladesh. *J. Infect. Dis.* **186** 246–251.

LONGINI, I. M. JR., NIZAM, A., ALI, M., YUNUS, M., SHENVI, N. and CLEMENS, J. D. (2007). Controlling endemic cholera with oral vaccines. *PLoS Med.* **4** e336.

MAY, R. and ANDERSON, R. M. (1991). *Infectious Diseases of Humans*: *Dynamics and Control*. Oxford Univ. Press, London.

MCKINLEY, T., COOK, A. R. and DEARDON, R. (2009). Inference in epidemic models without likelihoods. *Int. J. Biostat.* **5** Art. 24, 39. MR2533810

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

RASMUSSEN, D. A., RATMANN, O. and KOELLE, K. (2011). Inference for nonlinear epidemiological models using genealogies and time series. *PLoS Comput. Biol.* **7** e1002136, 11. MR2845064

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

SACK, R. B., SIDDIQUE, A. K., LONGINI, I. M., NIZAM, A., YUNUS, M., ISLAM, S., MORRIS, J. G., ALI, A., HUQ, A., NAIR, G. B., QADRI SHAH, F., FARUQUE, M., SACK, D. A. and COLWELL, R. R. (2003). A 4-year study of the epidemiology of Vibrio cholerae in four rural areas of Bangladesh. *J. Infect. Dis.* **187** 96–101.

TAYLOR, H. M. and KARLIN, S. (1998). *An Introduction to Stochastic Modeling*, 3rd ed. Academic Press, San Diego, CA. MR1627763

TIEN, J. H. and EARN, D. J. D. (2010). Multiple transmission pathways and disease dynamics in a waterborne pathogen model. *Bull. Math. Biol.* **72** 1506–1533. MR2671584

TONI, T., WELCH, D., STRELKOWA, N., IPSEN, A. and STUMPF, M. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface* **6** 187–202.

VAN KAMPEN, N. G. (1992). *Stochastic Processes in Physics and Chemistry* **1**. Elsevier, Amsterdam.

The Annals of Applied Statistics
2016, Vol. 10, No. 2, 596–617
DOI: 10.1214/16-AOAS912
© Institute of Mathematical Statistics, 2016

FEATURE SCREENING FOR TIME-VARYING COEFFICIENT MODELS WITH ULTRAHIGH-DIMENSIONAL LONGITUDINAL DATA

BY WANGHUAN CHU[1], RUNZE LI[2] AND MATTHEW REIMHERR

Pennsylvania State University

Motivated by an empirical analysis of the Childhood Asthma Management Project, CAMP, we introduce a new screening procedure for varying coefficient models with ultrahigh-dimensional longitudinal predictor variables. The performance of the proposed procedure is investigated via Monte Carlo simulation. Numerical comparisons indicate that it outperforms existing ones substantially, resulting in significant improvements in explained variability and prediction error. Applying these methods to CAMP, we are able to find a number of potentially important genetic mutations related to lung function, several of which exhibit interesting nonlinear patterns around puberty.

REFERENCES

CHU, W., LI, R. and REIMHERR, M. (2016). Supplement to "Feature screening for time-varying coefficient models with ultrahigh-dimensional longitudinal data." DOI:10.1214/16-AOAS912SUPP.

FAN, J., FENG, Y. and SONG, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J. Amer. Statist. Assoc.* **106** 544–557. MR2847969

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 849–911. MR2530322

FAN, J., MA, Y. and DAI, W. (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *J. Amer. Statist. Assoc.* **109** 1270–1284. MR3265696

FAN, J. and SONG, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.* **38** 3567–3604. MR2766861

HE, X., WANG, L. and HONG, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Ann. Statist.* **41** 342–369. MR3059421

HUANG, J. Z., WU, C. O. and ZHOU, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statist. Sinica* **14** 763–788. MR2087972

LI, R., ZHONG, W. and ZHU, L. (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc.* **107** 1129–1139. MR3010900

LIANG, K. Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22. MR0836430

LIU, J., LI, R. and WU, R. (2014). Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *J. Amer. Statist. Assoc.* **109** 266–274. MR3180562

REIMHERR, M. and NICOLAE, D. (2014). A functional data analysis approach for genetic association studies. *Ann. Appl. Stat.* **8** 406–429. MR3191996

Key words and phrases. Feature selection, time-varying coefficient models, ultrahigh-dimensional longitudinal data, genome-wide association study, functional linear model.

SONG, R., YI, F. and ZOU, H. (2014). On varying-coefficient independence screening for high-dimensional varying-coefficient models. *Statist. Sinica* **24** 1735–1752. MR3308660

THE CHILDHOOD ASTHMA MANAGEMENT PROGRAM RESEARCH GROUP (1999). The Childhood Asthma Management Program (CAMP): Design, rationale, and methods. *Control. Clin. Trials* **20** 91–120.

THE CHILDHOOD ASTHMA MANAGEMENT PROGRAM RESEARCH GROUP (2000). Long-term effects of budesonide or nedocromil in children with asthma. *N. Engl. J. Med.* **343** 1054–1063.

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

ZHU, L.-P., LI, L., LI, R. and ZHU, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc.* **106** 1464–1475. MR2896849

The Annals of Applied Statistics
2016, Vol. 10, No. 2, 618–637
DOI: 10.1214/16-AOAS927
© Institute of Mathematical Statistics, 2016

PSEUDO-VALUE APPROACH FOR CONDITIONAL QUANTILE RESIDUAL LIFETIME ANALYSIS FOR CLUSTERED SURVIVAL AND COMPETING RISKS DATA WITH APPLICATIONS TO BONE MARROW TRANSPLANT DATA[1]

BY KWANG WOO AHN AND BRENT R. LOGAN

Medical College of Wisconsin

Quantile residual lifetime analysis is conducted to compare remaining lifetimes among groups for survival data. Evaluating residual lifetimes among groups after adjustment for covariates is often of interest. The current literature is limited to comparing two groups for independent data. We propose a pseudo-value approach to compare quantile residual lifetimes given covariates between multiple groups for independent and clustered survival data. The proposed method considers clustered event times and clustered censoring times in addition to independent event times and censoring times. We show that the method can also be used to compare multiple groups on the cause-specific residual life distribution in the competing risk setting, for which there are no current methods which account for clustering. The empirical Type I errors and statistical power of the proposed study are examined in a simulation study, which shows that the proposed method controls Type I errors very well and has higher power than an existing method. The proposed method is illustrated by a bone marrow transplant data set.

REFERENCES

AHN, K. W. and LOGAN, B. R. (2016). Supplement to "Pseudo-value approach for conditional quantile residual lifetime analysis for clustered survival and competing risks data with applications to bone marrow transplant data." DOI:10.1214/16-AOAS927SUPP.

AHN, K. W. and MENDOLIA, F. (2014). Pseudo-value approach for comparing survival medians for dependent data. *Stat. Med.* **33** 1531–1538. MR3240767

ANDERSEN, P. K., KLEIN, J. P. and ROSTHØJ, S. (2003). Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika* **90** 15–27. MR1966547

BINDER, N., GERDS, T. A. and ANDERSEN, P. K. (2014). Pseudo-observations for competing risks with covariate dependent censoring. *Lifetime Data Anal.* **20** 303–315. MR3181016

BRESLOW, N. E. (1972). Discussion of the paper by D. R. Cox. *J. Roy. Statist. Soc. Ser. B* **34** 216–217.

COMMENGES, D. and ANDERSEN, P. K. (1995). Score test of homogeneity for survival data. *Lifetime Data Anal.* **1** 145–159. MR1353846

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

EAPEN, M., LOGAN, B. R., APPELBAUM, F. R., ANTIN, J. H., ANASETTI, C., COURIEL, D. R., CHEN, J., MAZIARZ, R. T., MCCARTHY, P. L., NAKAMURA, R., RATANATHARATHORN, V., VIJ, R. and CHAMPLIN, R. E. (2015). Long-term survival after transplantation of unrelated

Key words and phrases. Pseudo-value, residual lifetime, clustered data.

donor peripheral blood or bone marrow hematopoietic cells for hematologic malignancy. *Biol. Blood Marrow Transplant*. **21** 55–59.

FERRY, C., GEMAYEL, G., ROCHA, V., LABOPIN, M., ESPEROU, H., ROBIN, M., DE LATOUR, R. P., RIBAUD, P., DEVERGIE, A., LEBLANC, T., BARUCHEL, E. G. A. and SOCIE, G. (2007). Long-term outcomes after allogeneic stem cell transplantation for children with hematological malignancies. *Bone Marrow Transplant*. **40** 219–224.

HE, P., ERIKSSON, F., SCHEIKE, T. H. and ZHANG, M. J. (2016). A proportional hazards regression model for the subdistribution with covariates–adjusted censoring weight for competing risks data. *Scand. J. Stat*. **43** 103–122.

JACOBSEN, M. and MARTINUSSEN, T. (2014). A note on the large sample properties of estimators based on generalized linear models for correlated pseudo-observations. Research Report 14/4. Dept. Biostatistics, Univ. Copenhagen.

JACOBSOHN, D. A. (2015). Outcomes of pediatric bone marrow transplantation for leukemia and myelodysplasia using matched sibling, mismatched related, or matched unrelated donor. *Bone Marrow Transplant*. **50** 749–750.

JEONG, J.-H. and FINE, J. P. (2009). A note on cause-specific residual life. *Biometrika* **96** 237–242. MR2482149

JEONG, J.-H. and FINE, J. P. (2013). Nonparametric inference on cause-specific quantile residual life. *Biom. J.* **55** 68–81. MR3042385

KIM, M.-O., ZHOU, M. and JEONG, J.-H. (2012). Censored quantile regression for residual lifetimes. *Lifetime Data Anal*. **18** 177–194. MR2903719

KLEIN, J. P. and ANDERSEN, P. K. (2005). Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics* **61** 223–229. MR2135864

LEE, E. W., WEI, L. J. and AMATO, D. A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In *Survival Analysis*: *State of the Art* (*Columbus*, *OH*, 1991). *NATO Adv. Sci. Inst. Ser. E Appl. Sci*. **211** 237–247. Kluwer Academic, Dordrecht. MR1175646

LIANG, K. Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22. MR0836430

LIN, D. Y. (2007). On the Breslow estimator. *Lifetime Data Anal*. **13** 471–480. MR2416534

LIN, C., ZHANG, L. and ZHOU, Y. (2015). Conditional quantile residual lifetime models for right censored data. *Lifetime Data Anal*. **21** 75–96.

LOGAN, B. R., ZHANG, M.-J. and KLEIN, J. P. (2011). Marginal models for clustered time-to-event data with competing risks using pseudovalues. *Biometrics* **67** 1–7. MR2898811

MA, Y. and WEI, Y. (2012). Analysis on censored quantile residual life model via spline smoothing. *Statist. Sinica* **22** 47–68. MR2933167

MAJHAIL, N. S. and RIZZO, J. D. (2013). Surviving the cure: Long term followup of hematopoietic cell transplant recipients. *Bone Marrow Transplant*. **48** 1145–1151.

MARTIN, P. J., COUNTS, G. W., APPELBAUM, F. R., LEE, S. J., SANDERS, J. E., DEEG, H. J., FLOWERS, M. E. D., SYRJALA, K. L., HANSEN, J. A., STORB, R. F. and STORER, B. E. (2010). Life expectancy in patients surviving more than 5 years after hematopoietic cell transplantation. *J. Clin. Oncol*. **28** 1011–1016.

PRENTICE, R. L., KALBFLEISCH, J. D., PETERSON, A. V. JR., FLOURNOY, N., FAREWELL, V. T. and BRESLOW, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics* **34** 541–554.

SHAW, P. J., KAN, F., AHN, K. W., SPELLMAN, S. R., ALJURF, M., AYAS, M., BURKE, M., CAIRO, M. S., CHEN, A. R., DAVIES, S. M., FRANGOUL, H., GAJEWSKI, J., GALE, R. P., GODDER, K., HALE, G. A., HEEMSKERK, M. B. A., HORAN, J., KAMANI, N., KASOW, K. A., CHAN, K. W., LEE, S. J., LEUNG, W. H., LEWIS, V. A., MIKLOS, D., OUDSHOORN, M., PETERSDORF, E. W., RINGDÉN, O., SANDERS, J., SCHULTZ, K. R., SEBER, A., SETTERHOLM, M., WALL, D. A., YU, L. and PULSIPHER, M. A. (2010). Outcomes of pediatric bone

marrow transplantation for leukemia and myelodysplasia using matched sibling, mismatched related, or matched unrelated donors. *Blood* **116** 4007–4015.

SPIEKERMAN, C. F. and LIN, D. Y. (1998). Marginal regression models for multivariate failure time data. *J. Amer. Statist. Assoc.* **93** 1164–1175. MR1649210

YIN, G. and CAI, J. (2004). Additive hazards model with multivariate failure time data. *Biometrika* **91** 801–818. MR2126034

ZEGER, S. L., LIANG, K.-Y. and ALBERT, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics* **44** 1049–1060. MR0980999

ZENG, D. and LIN, D. Y. (2008). Efficient resampling methods for nonsmooth estimating functions. *Biostat.* **9** 355–363.

ZHAO, Y. Q., ZENG, D., LABER, E. B., SONG, R., YUAN, M. and KOSOROK, M. R. (2015). Doubly robust learning for estimating individualized treatment with censored data. *Biometrika* **102** 151–168. MR3335102

ZHOU, B., FINE, J., LATOUCHE, A. and LABOPIN, M. (2012). Competing risks regression for clustered data. *Biostat.* **13** 371–383.

The Annals of Applied Statistics
2016, Vol. 10, No. 2, 638–666
DOI: 10.1214/16-AOAS926
© Institute of Mathematical Statistics, 2016

A SPATIOTEMPORAL NONPARAMETRIC BAYESIAN MODEL OF MULTI-SUBJECT FMRI DATA

BY LINLIN ZHANG*, MICHELE GUINDANI†, FRANCESCO VERSACE‡, JEFFREY M. ENGELMANN§ AND MARINA VANNUCCI*

Rice University, MD Anderson Cancer Center†, University of Oklahoma Health Sciences Center‡ and MD Anderson Cancer Center§*

In this paper we propose a unified, probabilistically coherent framework for the analysis of task-related brain activity in multi-subject fMRI experiments. This is distinct from two-stage "group analysis" approaches traditionally considered in the fMRI literature, which separate the inference on the individual fMRI time courses from the inference at the population level. In our modeling approach we consider a spatiotemporal linear regression model and specifically account for the between-subjects heterogeneity in neuronal activity via a spatially informed multi-subject nonparametric variable selection prior. For posterior inference, in addition to Markov chain Monte Carlo sampling algorithms, we develop suitable variational Bayes algorithms. We show on simulated data that variational Bayes inference achieves satisfactory results at more reduced computational costs than using MCMC, allowing scalability of our methods. In an application to data collected to assess brain responses to emotional stimuli our method correctly detects activation in visual areas when visual stimuli are presented.

REFERENCES

BANERJEE, S., CARLIN, B. P. and GELFAND, A. E. (2015). *Hierarchical Modeling and Analysis for Spatial Data*, 2nd ed. *Monographs on Statistics and Applied Probability* **135**. CRC Press, Boca Raton, FL. MR3362184

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

BISHOP, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York. MR2247587

BLEI, D. M. and JORDAN, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Anal.* **1** 121–143 (electronic). MR2227367

BOWMAN, F., CAFFO, B., BASSETT, S. and KILTS, C. (2008). A Bayesian hierarchical framework for spatial modeling of fMRI data. *NeuroImage* **39** 146–156.

BUXTON, R. and FRANK, L. (1997). A model for the coupling between cerebral blood flow and oxygen metabolism during neural stimulation. *J. Cereb. Blood Flow Metab.* **17** 64–72.

CARBONETTO, P. and STEPHENS, M. (2012). Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Anal.* **7** 73–107. MR2896713

DAUBECHIES, I. (1992). *Ten Lectures on Wavelets*. *CBMS-NSF Regional Conference Series in Applied Mathematics* **61**. SIAM, Philadelphia, PA. MR1162107

Key words and phrases. Multi-subject fMRI, spatiotemporal linear regression, variable selection priors, variational Bayes.

EFRON, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.* **23** 1–22. MR2431866

FADILI, M. J. and BULLMORE, E. T. (2002). Wavelet-generalised least squares: A new BLU estimator of linear regression models with $1/f$ errors. *NeuroImage* **15** 217–232.

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

FLANDIN, G. and PENNY, W. D. (2007). Bayesian fMRI data analysis with sparse spatial basis function priors. *NeuroImage* **34** 1108–1125.

FRISTON, K. J. (1994). Functional and effective connectivity in neuroimaging: A synthesis. *Hum. Brain Mapp.* **2** 56–78.

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

FRISTON, K. J., JEZZARD, P. and TURNER, R. (1994). Analysis of functional MRI time-series. *Hum. Brain Mapp.* **1** 153–171.

FRISTON, K. J. and PENNY, W. (2003). Posterior probability maps and SPMs. *NeuroImage* **19** 1240–1249.

FRISTON, K. J., HOLMES, A. P., POLINE, J. B., GRASBY, P. J., WILLIAMS, S. C. R., FRACKOWIAK, R. S. J. and TURNER, R. (1995). Analysis of fMRI time-series revisited. *NeuroImage* **2** 45–53.

FRISTON, K. J., PENNY, W., PHILLIPS, C., KIEBEL, S., HINTON, G. and ASHBURNER, J. (2002). Classical and Bayesian inference in neuroimaging: Theory. *NeuroImage* **16** 465–483.

HARRISON, L. M. and GREEN, G. G. R. (2010). A Bayesian spatiotemporal model for very large data sets. *NeuroImage* **50** 1126–1141.

HARTVIG, N. V. and JENSEN, J. L. (2000). Spatial mixture modeling of fMRI data. *Hum. Brain Mapp.* **11** 233–248.

HOLMES, A. P. and FRISTON, K. J. (1998). Generalisability, random effects & population inference. *Neuroimage* **7** S754.

ISHWARAN, H. and JAMES, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.* **96** 161–173. MR1952729

JBABDI, S., WOOLRICH, M. W. and BEHRENS, T. E. J. (2009). Multiple-subjects connectivity-based parcellation using hierarchical Dirichlet process mixture models. *NeuroImage* **44** 373–384.

JEONG, J., VANNUCCI, M. and KO, K. (2013). A wavelet-based Bayesian approach to regression models with long memory errors and its application to fMRI data. *Biometrics* **69** 184–196. MR3058065

JOHNSON, T. D., LIU, Z., BARTSCH, A. J. and NICHOLS, T. E. (2013). A Bayesian non-parametric Potts model with application to pre-surgical FMRI data. *Stat. Methods Med. Res.* **22** 364–381. MR3190664

JOSET, A. E., GAZZOLA, V. and KEYSERS, C. (2009). An introduction to anatomical ROI-based fMRI classification analysis. *Brain Res.* **1282** 114–125.

KALUS, S., SÄMANN, P. G. and FAHRMEIR, L. (2014). Classification of brain activation via spatial Bayesian variable selection in fMRI regression. *Adv. Data Anal. Classif.* **8** 63–83. MR3168680

KIM, S., SMYTH, P. and STERN, H. (2006). A nonparametric Bayesian approach to detecting spatial activation patterns in fMRI data. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI* 2006 217–224.

LEE, K.-J., JONES, G. L., CAFFO, B. S. and BASSETT, S. S. (2014). Spatial Bayesian variable selection models on functional magnetic resonance imaging time-series data. *Bayesian Anal.* **9** 699–731. MR3256061

LI, F., ZHANG, T., WANG, Q., GONZALEZ, M. Z., MARESH, E. L. and COAN, J. A. (2015). Spatial Bayesian variable selection and grouping for high-dimensional scalar-on-image regression. *Ann. Appl. Stat.* **9** 687–713. MR3371331

LINDQUIST, M. A. (2008). The statistical analysis of fMRI data. *Statist. Sci.* **23** 439–464. MR2530545

MEYER, F. G. (2003). Wavelet-based estimation of a semiparametric generalized linear model of fMRI time-series. *IEEE Trans. Med. Imag.* **22** 315–322.

MÜLLER, P., PARMIGIANI, G. and RICE, K. (2007). FDR and Bayesian multiple comparisons rules. In *Bayesian Statistics* 8 (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.). *Oxford Sci. Publ.* 349–370. Oxford Univ. Press, Oxford. MR2433200

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.

PENNY, W., KIEBEL, S. and FRISTON, K. J. (2003). Variational Bayesian inference for fmri time series. *NeuroImage* **19** 727–741.

PENNY, W. D., TRUJILLO-BARRETO, N. and FRISTON, K. J. (2005). Bayesian fMRI time series analysis with spatial priors. *NeuroImage* **24** 350–362.

PROPP, J. G. and WILSON, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. In *Proceedings of the Seventh International Conference on Random Structures and Algorithms* (*Atlanta*, *GA*, 1995) **9** 223–252. Random Structures Algorithms, 1-2. MR1611693

QUIRÓS, A., DIEZ, R. M. and GAMERMAN, D. (2010). Bayesian spatiotemporal model of fMRI data. *NeuroImage* **49** 442–456.

RAFTERY, A. E. and LEWIS, S. M. (1992). One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo. *Statist. Sci.* **7** 493–497.

RODRÍGUEZ, A., DUNSON, D. B. and GELFAND, A. E. (2008). The nested Dirichlet process. *J. Amer. Statist. Assoc.* **103** 1131–1144. MR2528831

ROSENBLATT, J. D., VINK, M. and BENJAMINI, Y. (2014). Revisiting multi-subject random effects in fMRI: Advocating prevalence estimation. *NeuroImage* **84** 113–121.

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

SANYAL, N. and FERREIRA, M. A. (2012). Bayesian hierarchical multi-subject multiscale analysis of functional MRI data. *NeuroImage* **63** 1519–1531.

SAVITSKY, T. and VANNUCCI, M. (2010). Spiked Dirichlet process priors for Gaussian process models. *J. Probab. Stat.* Art. ID 201489, 14. MR2745498

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

SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650. MR1309433

SMITH, M. and FAHRMEIR, L. (2007). Spatial Bayesian variable selection with application to functional magnetic resonance imaging. *J. Amer. Statist. Assoc.* **102** 417–431. MR2370843

SU, S., CAFFO, B., GARRETT-MAYER, E. and BASSETT, S. (2009). Modified test statistics by inter-voxel variance shrinkage with an application to fMRI. *Biostatistics* **10** 219–227.

SUN, W., REICH, B. J., CAI, T. T., GUINDANI, M. and SCHWARTZMAN, A. (2015). False discovery control in large-scale spatial multiple testing. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 59–83. MR3299399

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

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.

VERSACE, F., ENGELMANN, J. M., JACKSON, E. F., SLAPIN, A., CORTESE, K. M., BEVERS, T. B. and SCHOVER, L. R. (2013). Brain responses to erotic and other emotional stimuli

in breast cancer survivors with and without distress about low sexual desire: A preliminary fmri study. *Brain Imaging Behav.* **7** 533–542.

WANG, C., PAISLEY, J. W. and BLEI, D. M. (2011). Online variational inference for the hierarchical Dirichlet process. In *International Conference on Artificial Intelligence and Statistics* 752–760.

WOOLRICH, M. W., BEHRENS, T. and SMITH, S. (2004). Constrained linear basis sets for HRF modelling using variational Bayes. *NeuroImage* **21** 1748–1761.

WOOLRICH, M. W., JENKINSON, M., BRADY, J. M. and SMITH, S. M. (2004). Fully Bayesian spatio-temporal modeling of fMRI data. *IEEE Trans. Med. Imag.* **23** 213–231.

WORNELL, G. W. and OPPENHEIM, A. V. (1992). Estimation of fractal signals from noisy measurements using wavelets. *IEEE Trans. Signal Process.* **40** 611–623.

WORSLEY, K. J. and FRISTON, K. J. (1995). Analysis of fMRI time-series revisited-again. *NeuroImage* **2** 173–181.

XIA, J., LIANG, F. and WANG, Y. (2009). FMRI analysis through Bayesian variable selection with a spatial prior. *IEEE Int. Symp. on Biomedical Imaging* 714–717.

XU, L., JOHNSON, T. D., NICHOLS, T. E. and NEE, D. E. (2009). Modeling inter-subject variability in fMRI activation location: A Bayesian hierarchical spatial model. *Biometrics* **65** 1041–1051. MR2756491

YAN, F., XU, N. and QI, Y. (2009). Parallel inference for latent dirichlet allocation on graphics processing units. In *Advances in Neural Information Processing Systems* 2134–2142.

ZHANG, L., GUINDANI, M., VERSACE, F. and VANNUCCI, M. (2014). A spatio-temporal nonparametric Bayesian variable selection model of fMRI data for clustering correlated time courses. *NeuroImage* **95** 162–175.

ZHANG, L., GUINDANI, M. and VANNUCCI, M. (2015). Bayesian models for fMRI data analysis. *Wiley Interdiscip. Rev.: Comput. Stat.* **7** 21–41.

ZHANG, L., GUINDANI, M., VERSACE, F., ENGELMANN, J. and VANNUCCI, M. (2016). Supplement to "A spatiotemporal nonparametric Bayesian model of multi-subject fMRI data." DOI:10.1214/16-AOAS926SUPP.

The Annals of Applied Statistics
2016, Vol. 10, No. 2, 667–689
DOI: 10.1214/15-AOAS851
© Institute of Mathematical Statistics, 2016

NONPARAMETRIC MULTI-LEVEL CLUSTERING OF HUMAN EPILEPSY SEIZURES[1]

BY DRAUSIN F. WULSIN, SHANE T. JENSEN AND BRIAN LITT

University of Pennsylvania

Understanding neuronal activity in the human brain is an extremely difficult problem both in terms of measurement and statistical modeling. We address a particular research question in this area: the analysis of human intracranial electroencephalogram (iEEG) recordings of epileptic seizures from a collection of patients. In these data, each seizure of each patient is defined by the activities of many individual recording channels. The modeling of epileptic seizures is challenging due the large amount of heterogeneity in iEEG signal between channels within a particular seizure, between seizures within an individual, and across individuals. We develop a new nonparametric hierarchical Bayesian model that simultaneously addresses these multiple levels of heterogeneity in our epilepsy data. Our approach, which we call a multi-level clustering hierarchical Dirichlet process (MLC-HDP), clusters over channel activities within a seizure, over seizures of a patient and over patients. We demonstrate the advantages of our methodology over alternative approaches in human EEG seizure data and show that its seizure clustering is close to manual clustering by a physician expert. We also address important clinical questions like "to which seizures of other patients is this seizure similar?"

REFERENCES

ADELI, H., ZHOU, Z. and DADMEHR, N. (2003). Analysis of EEG records in an epileptic patient using wavelet transform. *J. Neurosci. Methods* **123** 69–87.

BARTOLOMEI, F., COSANDIER-RIMELE, D., MCGONIGAL, A., AUBERT, S., RÉGIS, J., GAVARET, M., WENDLING, F. and CHAUVEL, P. (2010). From mesial temporal lobe to temporoperisylvian seizures: A quantified study of temporal lobe seizure networks. *Epilepsia* **51** 2147–2158.

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

CHAN, A. M., SUN, F. T., BOTO, E. H. and WINGEIER, B. M. (2008). Automated seizure onset detection for accurate onset time determination in intracranial EEG. *Clin. Neurophysiol.* **119** 2687–2696.

CHAOVALITWONGSE, W. A. (2008). Novel quadratic programming approach for time series clustering with biomedical application. *J. Comb. Optim.* **15** 225–241. MR2383309

DE TISI, J., BELL, G. S., PEACOCK, J. L., MCEVOY, A. W., HARKNESS, W. F. J., SANDER, J. W. and DUNCAN, J. S. (2011). The long-term outcome of adult epilepsy surgery, patterns of seizure remission, and relapse: A cohort study. *Lancet* **378** 1388–1395.

Key words and phrases. Epilepsy, seizures, intracranial electroencephalogram (iEEG), Dirichlet process, nonparametric Bayes, clustering.

ENGEL, J. JR. and PEDLEY, T. A., eds. (2008). *Epilepsy*: *A Comprehensive Textbook*, 2nd ed. Lippincot Williams & Wilkins, Philadelphia, PA.

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

FRENCH, J. A. (2007). Refractory epilepsy: Clinical overview. *Epilepsia* **48** 3–7.

GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409. MR1141740

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

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6** 721–741.

GHOSH-DASTIDAR, S., ADELI, H. and DADMEHR, N. (2008). Principal component analysis-enhanced cosine radial basis function neural network for robust epilepsy and seizure detection. *IEEE Trans. Biomed. Eng.* **55** 512–518.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*. Springer, New York. MR1851606

HEGDE, A., ERDOGMUS, D., SHIAU, D. S., PRINCIPE, J. C. and SACKELLARES, C. J. (2007). Clustering approach to quantify long-term spatio-temporal interactions in epileptic intracranial electroencephalography. *Comput. Intell. Neurosci.* **2007** 83416.

KLATCHKO, A., RAVIV, G., WEBBER, W. R. and LESSER, R. P. (1998). Enhancing the detection of seizures with a clustering algorithm. *Electroencephalogr. Clin. Neurophysiol.* **106** 52–63.

LIU, J. S., WONG, W. H. and KONG, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81** 27–40. MR1279653

MACKAY, D. J. C. (2003). *Information Theory*, *Inference and Learning Algorithms*. Cambridge Univ. Press, New York. MR2012999

OSSADTCHI, A., GREENBLATT, R. E., TOWLE, V. L., KOHRMAN, M. H. and KAMADA, K. (2010). Inferring spatiotemporal network patterns from intracranial EEG data. *Clin. Neurophysiol.* **121** 823–835.

PARAMANATHAN, P. and UTHAYAKUMAR, R. (2008). Application of fractal theory in analysis of human electroencephalographic signals. *Comput. Biol. Med.* **38** 372–378.

PITMAN, J. (2002). Poisson–Dirichlet and GEM invariant distributions for split-and-merge transformation of an interval partition. *Combin. Probab. Comput.* **11** 501–514. MR1930355

QUYEN, M. L. V., SOSS, J., NAVARRO, V., ROBERTSON, R., CHAVEZ, M., BAULAC, M. and MARTINERIE, J. (2005). Preictal state identification by synchronization changes in long-term intracranial EEG recordings. *Clin. Neurophysiol.* **116** 559–568.

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

REIJNEVELD, J. C., PONTEN, S. C., BERENDSE, H. W. and STAM, C. J. (2007). The application of graph theoretical analysis to complex networks in the brain. *Clinical Neurophysiology* **118** 2317–2331.

RODRÍGUEZ, A., DUNSON, D. B. and GELFAND, A. E. (2008). The nested Dirichlet process. *J. Amer. Statist. Assoc.* **103** 1131–1144. MR2528831

SARACENO, B., AVANZINI, G. and LEE, P. (2005). Atlas: Epilepsy care in the world. Technical report, World Health Organization, Geneva.

SCHIFF, S. J., SAUER, T., KUMAR, R. and WEINSTEIN, S. L. (2005). Neuronal spatiotemporal pattern discrimination: The dynamical evolution of seizures. *NeuroImage* **28** 1043–1055.

SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650. MR1309433

SHANNON, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* **27** 379–423, 623–656. MR0026286

SRINIVASAN, V., ESWARAN, C. and SRIRAAM, N. (2007). Approximate entropy-based epileptic EEG detection using artificial neural networks. *IEEE Trans. Inf. Technol. Biomed.* **11** 288–295.

STAM, C. J. (2005). Nonlinear dynamical analysis of EEG and MEG: Review of an emerging field. *Clin. Neurophysiol.* **116** 2266–2301.

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

WULSIN, D. F., JENSEN, S. T. and LITT, B. (2016). Supplement to "Nonparametric multi-level clustering of human epilepsy seizures." DOI:10.1214/15-AOAS851SUPP.

The Annals of Applied Statistics
2016, Vol. 10, No. 2, 690–725
DOI: 10.1214/15-AOAS871
© Institute of Mathematical Statistics, 2016

SHRINKAGE OF DISPERSION PARAMETERS IN THE BINOMIAL FAMILY, WITH APPLICATION TO DIFFERENTIAL EXON SKIPPING[1]

BY SEAN RUDDY, MARLA JOHNSON AND ELIZABETH PURDOM

University of California, Berkeley

The prevalence of sequencing experiments in genomics has led to an increased use of methods for count data in analyzing high-throughput genomic data to perform analyses. The importance of shrinkage methods in improving the performance of statistical methods remains. A common example is gene expression data, where the counts per gene are often modeled as some form of an overdispersed Poisson. Shrinkage estimates of the per-gene dispersion parameter have led to improved estimation of dispersion, particularly in the case of a small number of samples.

We address a different count setting introduced by the use of sequencing data: comparing differential proportional usage via an overdispersed binomial model. We are motivated by our interest in testing for differential exon skipping in mRNA-Seq experiments. We introduce a novel shrinkage method that models the overdispersion with the double binomial distribution proposed by Efron [*J. Amer. Statist. Assoc.* **81** (1986) 709–721].

Our method (WEB-Seq) is an empirical Bayes strategy for producing a shrunken estimate of dispersion and effectively detects differential proportional usage, and has close ties to the weighted-likelihood strategy of edgeR developed for gene expression data [*Bioinformatics* **23** (2007) 2881–2887, *Bioinformatics (Oxford, England)* **26** (2010) 139–140]. We analyze its behavior on simulated data sets as well as real data and show that our method is fast, powerful and gives accurate control of the FDR compared to alternative approaches. We provide implementation of our methods in the R package `DoubleExpSeq` available on CRAN.

REFERENCES

ANDERS, S. and HUBER, W. (2010). Differential expression analysis for sequence count data. *Genome Biol*. **11** 106.

ANDERS, S., REYES, A. and HUBER, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Res*. **22** 2008–2017.

BARBOSA-MORAIS, N. L., IRIMIA, M., PAN, Q., XIONG, H. Y., GUEROUSSOV, S., LEE, L. J., SLOBODENIUC, V., KUTTER, C., WATT, S., COLAK, R., KIM, T., MISQUITTA-ALI, C. M., WILSON, M. D., KIM, P. M., ODOM, D. T., FREY, B. J. and BLENCOWE, B. J. (2012). The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338** 1587–1593.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

Key words and phrases. Empirical Bayes, dispersion estimation, over-dispersed binomial, alternative splicing, mRNA-Seq.

BOURGON, R., GENTLEMAN, R. and HUBER, W. (2010). Independent filtering increases detection power for high-throughput experiments. *Proc. Natl. Acad. Sci. USA* **107** 9546–9551.

BROOKS, A. N., YANG, L., DUFF, M. O., HANSEN, K. D., PARK, J. W., DUDOIT, S., BRENNER, S. E. and GRAVELEY, B. R. (2011). Conservation of an RNA regulatory map between drosophila and mammals. *Genome Res.* **21** 193–202.

BROOKS, A. N., CHOI, P. S., DE WAAL, L., SHARIFNIA, T., IMIELINSKI, M., SAKSENA, G., PEDAMALLU, C. S., SIVACHENKO, A., ROSENBERG, M., CHMIELECKI, J., LAWRENCE, M. S., DELUCA, D. S., GETZ, G. and MEYERSON, M. (2014). A pan-cancer analysis of transcriptome changes associated with somatic mutations in U2AF1 reveals commonly altered splicing events. *PLoS ONE* **9** e87361.

CANCER GENOME ATLAS RESEARCH NETWORK (2011). Integrated genomic analyses of ovarian carcinoma. *Nature* **474** 609–615.

DENOEUD, F., AURY, J.-M., SILVA, C. D., NOEL, B., ROGIER, O., DELLEDONNE, M., MORGANTE, M., VALLE, G., WINCKER, P., SCARPELLI, C., JAILLON, O. and ARTIGUENAVE, F. (2008). Annotating genomes with massive-scale RNA sequencing. *Genome Biol.* **9** R175.

DOLZHENKO, E. and SMITH, A. D. (2014). Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC Bioinformatics* **15** 215.

EFRON, B. (1986). Double exponential families and their use in generalized linear regression. *J. Amer. Statist. Assoc.* **81** 709–721. MR0860505

FENG, H., CONNEELY, K. N. and WU, H. (2014). A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Res.* **42** e69–e69.

GUTTMAN, M., GARBER, M., LEVIN, J. Z., DONAGHEY, J., ROBINSON, J., ADICONIS, X., FAN, L., KOZIOL, M. J., GNIRKE, A., NUSBAUM, C., RINN, J. L., LANDER, E. S. and REGEV, A. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* **28** 503–510.

HARDCASTLE, T. J. and KELLY, K. A. (2010). BaySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* **11** 422.

HARDCASTLE, T. J. and KELLY, K. A. (2013). Empirical Bayesian analysis of paired high-throughput sequencing data with a beta-binomial distribution. *BMC Bioinformatics* **14** 135.

HU, Y., HUANG, Y., DU, Y., ORELLANA, C. F., SINGH, D., JOHNSON, A. R., MONROY, A., KUAN, P. F., HAMMOND, S. M., MAKOWSKI, L., RANDELL, S. H., CHIANG, D. Y., HAYES, D. N., JONES, C., LIU, Y., PRINS, J. F. and LIU, J. (2013). DiffSplice: The genome-wide detection of differential splicing events with RNA-seq. *Nucleic Acids Res.* **41** e39.

JIANG, H. and WONG, W. H. (2009). Statistical inferences for isoform expression in RNA-seq. *Bioinformatics* **25** 1026–1032.

JØRGENSEN, B. (1997). *The Theory of Dispersion Models. Monographs on Statistics and Applied Probability* **76**. Chapman & Hall, London. MR1462891

KATZ, Y., WANG, E. T., AIROLDI, E. M. and BURGE, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7** 1009–1015.

LAW, C. W., CHEN, Y., SHI, W. and SMYTH, G. K. (2014). Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15** R29.

LENG, N., DAWSON, J. A., THOMSON, J. A., RUOTTI, V., RISSMAN, A. I., SMITS, B. M. G., HAAG, J. D., GOULD, M. N., STEWART, R. M. and KENDZIORSKI, C. (2013). EBSeq: An empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics* **29** 1035–1043.

MARIONI, J. C., MASON, C. E., MANE, S. M., STEPHENS, M. and GILAD, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18** 1509–1517.

McCarthy, D. J., Chen, Y. and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. *Nucleic Acids Res*. **40** 4288–4297.

National Human Genome Research Institute (2014). Alternative splicing. Available at www.genome.gov.

Pan, Q., Shai, O., Lee, L. J., Frey, B. J. and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet*. **40** 1413–1415.

Pawitan, Y. (2001). *In All Likelihood*: *Statistical Modelling and Inference Using Likelihood*. Oxford Univ Press, London.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Richard, H., Schulz, M. H., Sultan, M., Nürnberger, A., Schrinner, S., Balzereit, D., Dagand, E., Rasche, A., Lehrach, H., Vingron, M., Haas, S. A. and Yaspo, M.-L. (2010). Prediction of alternative isoforms from exon expression levels in RNA-seq experiments. *Nucleic Acids Res*. **38** e112.

Robinson, M. D., Mccarthy, D. J. and Smyth, G. K. (2010). edgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* (*Oxford, England*) **26** 139–140.

Robinson, M. D. and Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **23** 2881–2887.

Robinson, M. D. and Smyth, G. K. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* **9** 321–332.

Ruddy, S., Johnson, M. and Purdom, E. (2015a). Supplement A to "Shrinkage of dispersion parameters in the binomial family, with application to differential exon skipping." DOI:10.1214/15-AOAS871SUPPA.

Ruddy, S., Johnson, M. and Purdom, E. (2015b). Supplement B to "Shrinkage of dispersion parameters in the binomial family, with application to differential exon skipping." DOI:10.1214/15-AOAS871SUPPB.

Ruddy, S., Johnson, M. and Purdom, E. (2015c). Supplement C to "Shrinkage of dispersion parameters in the binomial family, with application to differential exon skipping." DOI:10.1214/15-AOAS871SUPPC.

Salzman, J., Jiang, H. and Wong, W. H. (2010). Statistical modeling of RNA-Seq data. Technical Report No. BIO-252, Division of Biostatistics, Stanford Univ., Palo Alto.

Shen, S., Park, J. W., Huang, J., Dittmar, K. A., Lu, Z.-x., Zhou, Q., Carstens, R. P. and Xing, Y. (2012). MATS: A Bayesian framework for flexible detection of differential alternative splicing from RNA-seq data. *Nucleic Acids Res*. **40** e61.

Shi, Y. and Jiang, H. (2013). rSeqDiff: Detecting differential isoform expression from RNA-seq data using hierarchical likelihood ratio test. *PloS One* **8** e79448.

Smyth, G. K. (2005). Limma: Linear models for microarray data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry and S. Dudoit, eds.) 397–420. Springer, New York.

Sun, D., Xi, Y., Rodriguez, B., Park, H. J., Tong, P., Meong, M., Goodell, M. A. and Li, W. (2014). MOABS: Model based analysis of bisulfite sequencing data. *Genome Biol*. **15** R38.

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.

VENABLES, J. P., KLINCK, R., KOH, C., GERVAIS-BIRD, J., BRAMARD, A., INKEL, L., DURAND, M., COUTURE, S., FROEHLICH, U., LAPOINTE, E., LUCIER, J.-F., THIBAULT, P., RANCOURT, C., TREMBLAY, K., PRINOS, P., CHABOT, B. and ELELA, S. A. (2009). Cancer-associated regulation of alternative splicing. *Nature Publishing Group* **16** 670–676.

WANG, X. (2006). Approximating Bayesian inference by weighted likelihood. *Canad. J. Statist.* **34** 279–298. MR2323997

WILLIAMS, D. A. (1982). Extrabinomial variation in logistic linear models. *J. Roy. Statist. Soc. Ser. C* **31** 144–148. MR0673714

WU, T. D. and NACU, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* (*Oxford*, *England*) **26** 873–881.

WU, H., WANG, C. and WU, Z. (2013). A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics* **14** 232–243.

WU, J., AKERMAN, M., SUN, S., MCCOMBIE, W. R., KRAINER, A. R. and ZHANG, M. Q. (2011). SpliceTrap: A method to quantify alternative splicing under single cellular conditions. *Bioinformatics* **27** 3010–3016.

YANG, X., TODD, J. A., CLAYTON, D. and WALLACE, C. (2012). Extra-binomial variation approach for analysis of pooled DNA sequencing data. *Bioinformatics* **28** 2898–2904.

YU, D., HUBER, W. and VITEK, O. (2013). Shrinkage estimation of dispersion in negative binomial models for RNA-seq experiments with small sample size. *Bioinformatics* **29** 1275–1282.

ZHOU, Y. H., XIA, K. and WRIGHT, F. A. (2011). A powerful and flexible approach to the analysis of RNA sequence count data. *Bioinformatics* (*Oxford*, *England*) **27** 2672–2678.

The Annals of Applied Statistics
2016, Vol. 10, No. 2, 726–755
DOI: 10.1214/15-AOAS892
© Institute of Mathematical Statistics, 2016

SCAN STATISTICS ON POISSON RANDOM FIELDS WITH APPLICATIONS IN GENOMICS

BY NANCY R. ZHANG[2,3], BENJAMIN YAKIR[1], LI C. XIA[2] AND DAVID SIEGMUND[1]

University of Pennsylvania, Hebrew University of Jerusalem, Stanford University School of Medicine and Stanford University

The detection of local genomic signals using high-throughput DNA sequencing data can be cast as a problem of scanning a Poisson random field for local changes in the rate of the process. We propose a likelihood-based framework for such scans, and derive formulas for false positive rate control and power calculations. The framework can also accommodate modified processes that involve overdispersion. As a specific, detailed example, we consider the detection of insertions and deletions by paired-end DNA-sequencing. We propose several statistics for this problem, compare their power under current experimental designs, and illustrate their application on an Illumina Platinum Genomes data set.

REFERENCES

ABYZOV, A., URBAN, A. E., SNYDER, M. and GERSTEIN, M. (2011). CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res*. **21** 974–984.

ADLER, R. J. and TAYLOR, J. E. (2007). *Random Fields and Geometry*. Springer, New York. MR2319516

BRAUN, J., DUMM, J., DE PALMA, F., FINLEY, C., KARLE, A. and MONTARULI, T. (2008). Methods for point source analysis in high energy neutrino telescopes. *Astroparticle Physics* **29** 299–305.

CAMPBELL, P. J., STEPHENS, P. J., PLEASANCE, E. D., O'MEARA, S., LI, H., SANTARIUS, T., STEBBINGS, L. A., LEROY, C., EDKINS, S., HARDY, C., TEAGUE, J. W., MENZIES, A., GOODHEAD, I., TURNER, D. J., CLEE, C. M., QUAIL, M. A., COX, A., BROWN, C., DURBIN, R., HURLES, M. E., EDWARDS, P. A. W., BIGNELL, G. R., STRATTON, M. R. and FUTREAL, P. A. (2008). Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature Genetics* **40** 722–729.

CHAISSON, M. J. P., HUDDLESTON, J., DENNIS, M. Y., SUDMANT, P. H., MALIG, M., HORMOZDIARI, F., ANTONACCI, F., SURTI, U., SANDSTROM, R., BOITANO, M., LANDOLIN, J. M., STAMATOYANNOPOULOS, J. A., HUNKAPILLER, M. W., KORLACH, J. and EICHLER, E. E. (2015). Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517** 608–611.

CHAN, H. P. and ZHANG, N. R. (2007). Scan statistics with weighted observations. *J. Amer. Statist. Assoc*. **102** 595–602. MR2370856

CHEN, K., WALLIS, J. W., MCLELLAN, M. D., LARSON, D. E., KALICKI, J. M., POHL, C. S., MCGRATH, S. D., WENDL, M. C., ZHANG, Q., LOCKE, D. P., SHI, X., FULTON, R. S., LEY,

Key words and phrases. Scan statistics, Poisson processes, change-point detection, next-generation sequencing, structural variation.

T. J., WILSON, R. K., DING, L. and MARDIS, E. R. (2009). BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **6** 677–681.

CHIANG, D. Y., GETZ, G., JAFFE, D. B., O'KELLY, M. J. T., ZHAO, X., CARTER, S. L., RUSS, C., NUSBAUM, C., MEYERSON, M. and LANDER, E. S. (2009). High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods* **6** 99–103.

EBERLE, M. A., KALLBERG, M., CHUANG, H.-Y., TEDDER, P., HUMPHRAY, S., BENTLEY, D. and MARGULIES, E. H. (2014). Platinum Genomes: A systematic assessment of variant accuracy using a large family pedigree. In *ASHG 2013 Annual Meeting*.

FEINGOLD, E., BROWN, P. O. and SIEGMUND, D. (1993). Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. *Am. J. Hum. Genet.* **53** 234–251.

FLICEK, P. and BIRNEY, E. (2009). Sense from sequence reads: Methods for alignment and assembly. *Nat. Methods* **6** S6–S12.

THE GENOMES PROJECT CONSORTIUM (2015). A global reference for human genetic variation. *Nature* **526** 68–74.

GROSS, E. and VITELLS, O. (2010). Trial factors for the look elsewhere effect in high energy physics. *The European Physical Journal C* **70** 525–530.

KARLIN, S., DEMBO, A. and KAWABATA, T. (1990). Statistical composition of high-scoring segments from molecular sequences. *Ann. Statist.* **18** 571–581. MR1056327

KULLDORFF, M. (1997). A spatial scan statistic. *Comm. Statist. Theory Methods* **26** 1481–1496. MR1456844

KULLDORFF, M. (1999). Spatial scan statistics: Models, calculations, and applications. In *Scan Statistics and Applications* (J. Glaz and N. Balakrishnan, eds.) 303–322. Birkhäuser, Boston, MA. MR1697758

KULLDORFF, M. and NAGARWALLA, N. (1995). Spatial disease clusters: Detection and inference. *Stat. Med.* **14** 799–810.

LANDER, E. S. and BOTSTEIN, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121** 185–199.

LI, Y. et al. (2011). Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly. *Nat. Biotech.* **29** 723–730.

MEDVEDEV, P., STANCIU, M. and BRUDNO, M. (2009). Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods* **6** S13–S20.

PENG, J. and SIEGMUND, D. (2005). The admixture model in linkage analysis. *J. Statist. Plann. Inference* **130** 317–324. MR2128010

PEPKE, S., WOLD, B. and MORTAZAVI, A. (2009). Computation for ChIP-seq and RNA-seq studies. *Nat. Methods* **6** S22–S32.

RABINOWITZ, D. and SIEGMUND, D. (1997). The approximate distribution of the maximum of a smoothed Poisson random field. *Statist. Sinica* **7** 167–180. MR1441152

RAUSCH, T., ZICHNER, T., SCHLATTL, A., STÜTZ, A. M., BENES, V. and KORBEL, J. O. (2012). DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28** i333–i339.

SCHWARTZMAN, A., JAFFE, A., GAVRILOV, Y. and MEYER, C. A. (2013). Multiple testing of local maxima for detection of peaks in ChIP-Seq data. *Ann. Appl. Stat.* **7** 471–494. MR3086427

SHEN, J. J. and ZHANG, N. R. (2012). Change-point model on nonhomogeneous Poisson processes with application in copy number profiling by next-generation DNA sequencing. *Ann. Appl. Stat.* **6** 476–496. MR2976479

SIEGMUND, D. (1985). *Sequential Analysis*: *Tests and Confidence Intervals*. Springer, New York. MR0799155

SIEGMUND, D. O. and WORSLEY, K. J. (1995). Testing for a signal with unknown location and scale in a stationary Gaussian random field. *Ann. Statist.* **23** 608–639. MR1332585

SIEGMUND, D. and YAKIR, B. (2007). *The Statistics of Gene Mapping*. Springer, New York. MR2301277

SIEGMUND, D., YAKIR, B. and ZHANG, N. (2010). Tail approximations for maxima of random fields by likelihood ratio transformations. *Sequential Anal.* **29** 245–262. MR2747524

SIEGMUND, D., YAKIR, B. and ZHANG, N. R. (2011). Detecting simultaneous variant intervals in aligned sequences. *Ann. Appl. Stat.* **5** 645–668. MR2840169

SIEGMUND, D. O., ZHANG, N. R. and YAKIR, B. (2011). False discovery rate for scanning statistics. *Biometrika* **98** 979–985. MR2860337

SONG, K., REN, J., ZHAI, Z., LIU, X., DENG, M. and SUN, F. (2013). Alignment-free sequence comparison based on next-generation sequencing reads. *J. Comput. Biol.* **20** 64–79. MR3021670

TANG, H. K. and SIEGMUND, D. (2001). Mapping quantitative trait loci in oligogenic models. *Biostatistics* **2** 147–162.

TOTH, B., LILLO, F. and FARMER, J. D. (2010). Segmentation algorithm for non-stationary compound Poisson processes. *Eur. Phys. J. B* **78** 235–243.

WORSLEY, K. J., EVANS, A. C., MARRETT, S. and NEELIN, P. (1992). A three-dimensional statistical analysis for CBF activation studies in human brain. *J. Cereb. Blood Flow Metab.* **12** 900–918.

YAKIR, B. (2013). *Extremes in Random Fields*: *A Theory and Its Applications*. Wiley, Chichester. MR3241226

YE, K., SCHULZ, M. H., LONG, Q., APWEILER, R. and NING, Z. (2009). Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25** 2865–2871.

ZHANG, N. R., YAKIR, B., XIA, L. C. and SIEGMUND, D. (2016). Supplement to "Scan statistics on Poisson random fields with applications in genomics." DOI:10.1214/15-AOAS892SUPP.

The Annals of Applied Statistics
2016, Vol. 10, No. 2, 756–785
DOI: 10.1214/15-AOAS901
© Institute of Mathematical Statistics, 2016

A STATISTICAL MODELING APPROACH FOR AIR QUALITY DATA BASED ON PHYSICAL DISPERSION PROCESSES AND ITS APPLICATION TO OZONE MODELING

BY XIAO LIU*, KYONGMIN YEO*, YOUNGDEOK HWANG*, JITENDRA SINGH† AND JAYANT KALAGNANAM*

*IBM T.J. Watson Research Center** *and IBM Research Collaboratory Singapore*†

For many complex environmental processes such as air pollution, the underlying physical mechanism usually provides valuable insights into the statistical modeling. In this paper, we propose a statistical air quality model motivated by a commonly used physical dispersion model, called the scalar transport equation. The emission of a pollutant is modeled by covariates such as land use, traffic pattern and meteorological conditions, while the transport and decay of a pollutant are modeled through a convolution approach which takes into account the dynamic wind field. This approach naturally establishes a nonstationary random field with a space–time nonseparable and anisotropic covariance structure. Note that, due to the extremely complex interactions between the pollutant and environmental conditions, the space–time covariance structure of pollutant concentration data is often dynamic and can hardly be specified or envisioned directly. The relationship between the proposed spatial-temporal model and the physics model is also shown, and the approach is applied to model the hourly ozone concentration data in Singapore.

REFERENCES

BANERJEE, S., CARLIN, B. P. and GELFAND, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*, 2nd ed. *Monographs on Statistics and Applied Probability* **135**. CRC Press, Boca Raton, FL. MR3362184

BERLINER, L. M. (2003). Physical-statistical modeling in geophysics. *Journal of Geophysical Research-Atmospheres* **108** STS 3-1–STS 3-10.

BROWN, P. E., KÅRESEN, K. F., ROBERTS, G. O. and TONELLATO, S. (2000). Blur-generated nonseparable space–time models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 847–860. MR1796297

BYUN, D. and SCHERE, K. L. (2006). Review of the governing equations, computational algorithms, and other components of the models-3 community multiscale air quality (CMAQ) modeling system. *Applied Mechanics Reviews* **59** 51–77.

CALDER, C. A. (2007). Dynamic factor process convolution models for multivariate space–time data with application to air quality assessment. *Environ. Ecol. Stat.* **14** 229–247. MR2405328

CAMELETTI, M., LINDGREN, F., SIMPSON, D. and RUE, H. (2013). Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Adv. Stat. Anal.* **97** 109–131. MR3045763

Key words and phrases. Spatial-temporal modeling, air quality model, partial differential equation, space–time nonseparable and anisotropic random field.

CARROLL, R., CHEN, E., LI, T., NEWTON, H., SCHMIEDICHE, H. and WANG, N. (1997). Ozone exposure and population density in Harris county. *Texas*. *Journal of the American Statistical Association* **92** 392–404.

CHRISTAKOS, G. and VYAS, V. (1998). A composite space–time approach to studying ozone distribution over eastern United States. *Atmospheric Environment* **32** 2845–2857.

COATS, C. (1996). High performance algorithms in the sparse matrix operator kernel emissions modelling system. In *Proceedings of the Ninth Joint Conference on Applications of Air Pollution Meteorology of the American Meteorological Society and the Air and Waste Management Association*. Atlanta, GA.

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

DOU, Y., LE, N. D. and ZIDEK, J. V. (2010). Modeling hourly ozone concentration fields. *Ann. Appl. Stat*. **4** 1183–1213. MR2751338

FUENTES, M. (2009). Statistical issues in health impact assessment at the state and local levels. *Air Quality, Atmosphere and Health* **2** 47–55.

FUENTES, M., CHEN, L., DAVIS, J. M. and LACKMANN, G. M. (2005). Modeling and predicting complex space–time structures and patterns of coastal wind fields. *Environmetrics* **16** 449–464. MR2147536

GHOSH, S. K., BHAVE, P. V., DAVIS, J. M. and LEE, H. (2010). Spatio-temporal analysis of total nitrate concentrations using dynamic statistical models. *J. Amer. Statist. Assoc*. **105** 538–551. MR2759930

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

HAN, S., BIAN, H., FENG, Y., LIU, A., LI, X., ZENG, F. and ZHANG, X. (2011). Analysis of the relationship between O3, NO and NO2 in tianjin. *China*. *Aerosol and Air Quality Research* **11** 128–139.

HASLETT, J. and RAFTERY, A. (1989). Space–time modelling with long-memory dependence: Assessing Ireland's wind power resource (with discussion). *Applied Statistics* **38** 1–50.

HIGDON, D. (2002). Space and space–time modeling using process convolutions. In *Quantitative Methods for Current Environmental Issues* (C. Anderson, V. Barnett, P. Chatwind and A. El-Shaarawi, eds.) 37–56. Springer, London. MR2059819

HIGDON, D. (2007). A process-convolution approach to modeling temperatures in the North Atlantic Ocean. *Environ. Ecol. Stat*. **5** 173–190.

HUANG, H. and HSU, N. (2004). Modeling transport effects on ground-level ozone using a nonstationary space–time model. *Environmetrics* **15** 251–268.

LIU, X., YEO, K., HWANG, Y., SINGH, J. and KALAGNANAM, J. (2016). Supplement to "A statistical modeling approach for air quality data based on physical dispersion processes and its application to ozone modeling." DOI:10.1214/15-AOAS901SUPP.

MALMBERG, A., ARELLANO, A., EDWARDS, D. P., FLYER, N., NYCHKA, D. and WIKLE, C. (2008). Interpolating fields of carbon monoxide data using a hybrid statistical-physical model. *Ann. Appl. Stat*. **2** 1231–1248. MR2655657

REICH, B. J. and FUENTES, M. (2007). A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields. *Ann. Appl. Stat*. **1** 249–264. MR2393850

REICH, B. J., EIDSVIK, J., GUINDANI, M., NAIL, A. J. and SCHMIDT, A. M. (2011). A class of covariate-dependent spatiotemporal covariance functions for the analysis of daily ozone concentration. *Ann. Appl. Stat*. **5** 2425–2447. MR2907121

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

SAHU, S. K., GELFAND, A. E. and HOLLAND, D. M. (2007). High-resolution space–time ozone modeling for assessing trends. *J. Amer. Statist. Assoc*. **102** 1221–1234. MR2412545

SCHABENBERGER, O. and GOTWAY, C. A. (2005). *Statistical Methods for Spatial Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL. MR2134116

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

SIGRIST, F., KÜNSCH, H. R. and STAHEL, W. A. (2015). Stochastic partial differential equation based modelling of large space–time data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 3–33. MR3299397

SKAMAROCK, W. C., KLEMP, J. B., DUDHIA, J., GILL, D. O., BARKER, D. M., DUDA, M. G., HUANG, X. Y., WANG, W. and POWERS, J. G. (2008). A description of the advanced research WRF version 3, Boulder, Colorado, USA. Ncar Technical Note: NCAR/TN–475+STR.

SMITH, L., FUENTES, M., REICH, B. and EDER, B. (2013). Prediction of speciated particulate matter and bias assessment of numerical output data. *International Journal of Environmental Science and Engineering Research* **4** 8–17.

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

STEIN, M. L. (2009). Spatial interpolation of high-frequency monitoring data. *Ann. Appl. Stat.* **3** 272–291. MR2668708

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

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY (1998). EPA third-generation air quality modeling system, models-3 (EPA-600/R-98/069a). U.S. Environmental Protection Agency, Research Triangle Park, NC.

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY (2003). Air Quality index—A guide to air quality and your health, EPA-454/K-03-002. U.S. Environmental Protection Agency, Research Triangle Park, NC.

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY (2012). Clean air act: Title I—Air pollution prevention and control. Available at http://epa.gov/oar/caa/title1.html.

WIKLE, C. K., MILLIFF, R. F., NYCHKA, D. and BERLINER, L. M. (2001). Spatiotemporal hierarchical Bayesian modeling: Tropical ocean surface winds. *J. Amer. Statist. Assoc.* **96** 382–397. MR1939342

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

WORLD HEALTH ORGANIZATION (2005). WHO Air Quality Guidelines for Paticularte Matter, Ozone, Nitrogen Dioxide and Sulfur Dioxide–Global Update (WHO/SDE/PHE/OEH/06.02), World Health Organization.

XU, Y., VIZUETE, W. and SERRE, M. (2012). Characterization of air quality ozone model performance using land use regression model: An application in exposure assessment for epidemology studies. In *The* 11*th Annual CMAS Conference*, Chapel Hill, NC.

ZHANG, H. and WANG, Y. (2010). Kriging and cross-validation for massive spatial data. *Environmetrics* **21** 290–304. MR2842244

The Annals of Applied Statistics
2016, Vol. 10, No. 2, 786–811
DOI: 10.1214/16-AOAS909
© Institute of Mathematical Statistics, 2016

A BAYESIAN GRAPHICAL MODEL FOR GENOME-WIDE ASSOCIATION STUDIES (GWAS)[1]

BY LAURENT BRIOLLAIS[*,†], ADRIAN DOBRA[‡], JINNAN LIU[*], MATT FRIEDLANDER[*], HILMI OZCELIK[*] AND HÉLÈNE MASSAM[§]

Lunenfeld-Tanenbaum Research Institute[*], University of Toronto[†], University of Washington[‡] and York University[§]

The analysis of GWAS data has long been restricted to simple models that cannot fully capture the genetic architecture of complex human diseases. As a shift from standard approaches, we propose here a general statistical framework for multi-SNP analysis of GWAS data based on a Bayesian graphical model. Our goal is to develop a general approach applicable to a wide range of genetic association problems, including GWAS and fine-mapping studies, and, more specifically, be able to: (1) Assess the joint effect of multiple SNPs that can be linked or unlinked and interact or not; (2) Explore the multi-SNP model space efficiently using the Mode Oriented Stochastic Search (MOSS) algorithm and determine the best models. We illustrate our new methodology with an application to the CGEM breast cancer GWAS data. Our algorithm selected several SNPs embedded in multi-locus models with high posterior probabilities. Most of the SNPs selected have a biological relevance. Interestingly, several of them have never been detected in standard single-SNP analyses. Finally, our approach has been implemented in the open source *R* package genMOSS.

REFERENCES

ANGLIAN BREAST CANCER STUDY GROUP (2000). Prevalence and penetrance of BRCA1 and BRCA2 in a population based series of breast cancer cases. *The British Journal of Cancer* **83** 1301–1308.

BARRETT, J. C., FRY, B., MALLER, J. and DALY, M. J. (2005). Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* **15** 263–265.

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

BRIOLLAIS L., DOBRA, A., LIU, J., FRIEDLANDER, M., OZCELIK, H. and MASSAM, H. (2016a). Supplement to "A Bayesian graphical model for genome-wide association studies (GWAS)." DOI:10.1214/16-AOAS909SUPPA.

BRIOLLAIS L., DOBRA, A., LIU, J., FRIEDLANDER, M., OZCELIK, H. and MASSAM, H. (2016b). Supplement to "A Bayesian graphical model for genome-wide association studies (GWAS)." DOI:10.1214/16-AOAS909SUPPB.

BRIOLLAIS L., DOBRA, A., LIU, J., FRIEDLANDER, M., OZCELIK, H. and MASSAM, H. (2016c). Supplement to "A Bayesian graphical model for genome-wide association studies (GWAS)." DOI:10.1214/16-AOAS909SUPPC.

Key words and phrases. Graphical model, Bayesian, stochastic search, GWAS, SNP, breast cancer.

BRIOLLAIS L., DOBRA, A., LIU, J., FRIEDLANDER, M., OZCELIK, H. and MASSAM, H. (2016d). Supplement to "A Bayesian graphical model for genome-wide association studies (GWAS)." DOI:10.1214/16-AOAS909SUPPD.

BRIOLLAIS L., DOBRA, A., LIU, J., FRIEDLANDER, M., OZCELIK, H. and MASSAM, H. (2016e). Supplement to "A Bayesian graphical model for genome-wide association studies (GWAS)." DOI:10.1214/16-AOAS909SUPPE.

COLLABORATIVE GROUP ON HORMONAL FACTORS IN BREAST CANCER (2002). Breast cancer and breastfeeding: Collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50302 women with breast cancer and 96973 women without the disease. *Lancet* **360** 187–195.

THE BREAST CANCER LINKAGE CONSORTIUM (1999). Cancer risks in BRCA2 mutation carriers. *J. Natl. Cancer Inst.* **91** 1310–1316.

DELLAPORTAS, P. and FORSTER, J. J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika* **86** 615–633. MR1723782

DEVLIN, B. and RISCH, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29** 311–322.

DIAMANDIS, E. P. and YOUSSEF, G. M. (2002). Human tissue kallikreins: A family of new cancer biomarkers. *Clinical Chemistry* **48** 1196–1205.

DOBRA, A. and MASSAM, H. (2010). The mode oriented stochastic search (MOSS) algorithm for log-linear models with conjugate priors. *Stat. Methodol.* **7** 240–253. MR2643600

EDWARDS, D. and HAVRÁNEK, T. (1985). A fast procedure for model search in multidimensional contingency tables. *Biometrika* **72** 339–351. MR0801773

GAIL, M. H. (2008). Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. *J. Natl. Cancer Inst.* **100** 1037–1041.

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

HAN, B., PARK, M. and CHEN, X. W. (2010). A Markov blanket-based method for detecting causal SNPs in GWAS. *BMC Bioinformatics* **11 Suppl. 3** S5.

HE, Q. and LIN, D.-Y. (2011). A variable selection method for genome-wide association studies. *Bioinformatics* **27** 1–8.

HINDORFF, L. A., SETHUPATHY, P., JUNKINS, H. A., RAMOS, E. M., MEHTA, J. P., COLLINS, F. S. and MANOLIO, T. A. (2009a). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **106** 9362–9367.

HINDORFF, L. A., JUNKINS, H. A., HALL, P. N., MEHTA, J. P. and MANOLIO, T. A. (2009b). A catalog of published genome-wide association studies. preprint. Available at www.genome.gov/gwastudies.

HIRSCHHORN, J. N. and DALY, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6** 95–108.

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

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., JOSEPH, F. F., JR., 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.

JIANG, X., BARMADA, M. M. and VISWESWARAN, S. (2010). Identifying genetic interactions in genome-wide data using Bayesian networks. *Genet. Epidemiol.* **34** 575–581.

KINGSMORE, S. F., LINQUIST, I. E., MUDGE, J., GESSLER, D. D. and BEAVIS, W. D. (2008). Genome-wide association studies: Progress and potential for drug discovery and development. *Nature Reviews* **7** 221–230.

KRUGLYAK, L. (2008). The road to genome-wide association studies. *Nature Genetics* **9** 314–318.

LETAC, G. and MASSAM, H. (2012). Bayes factors and the geometry of discrete hierarchical log-linear models. *Ann. Statist.* **40** 861–890. MR2985936

LI, Y., WILLER, C. J., DING, J., SCHEET, P. and ABECASIS, G. R. (2010). MaCH: Using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34** 816–834.

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.

MARCHINI, J., DONNELLY, P. and CARDON, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.* **37** 413–417.

MASSAM, H., LIU, J. and DOBRA, A. (2009). A conjugate prior for discrete hierarchical log-linear models. *Ann. Statist.* **37** 3431–3467. MR2549565

MCCARTHY, M. I. and HIRSCHHORN, J. N. (2008). Genome-wide association studies: Potential next steps on a genetic journey. *Hum. Mol. Genet.* **17** R156–R165.

PETO, J. and MACK, T. M. (2000). High constant incidence in twins and other relatives of women with breast cancer. *Nat. Genet.* **26** 411–414.

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

PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A. R., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I. W., DALY, M. J. and SHAM, P. C. (2007). PLINK: A toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* **81** 559–575.

RISCH, N. J. (2000). Searching for genetic determinants in the new millennium. *Nature* **405** 847–856.

RISCH, N. and MERIKANGAS, K. (1996). The future of genetic studies of complex human diseases. *Science* **273** 1516–1517.

SCHWARTZ, D. F., ZIEGLER, A. and KONIG, I. R. (2008). Beyond the results of genome-wide association studies. *Genet. Epidemiol.* **32** 671.

THOMAS, A. and CAMP, N. J. (2004). Graphical modelling of the joint distribution of alleles at associated loci. *Am. J. Hum. Genet.* **74** 1088–1101.

THOMPSON, D. and EASTON, D. F. (2004). The genetic epidemiology of breast cancer genes. *J. Mammary Gland Biol. Neoplasia* **9** 221–236.

THOMPSON, D., EASTON, D. F. and BREAST CANCER LINKAGE CONSORTIUM (2002). Cancer incidence in BRCA1 mutation carriers. *J. Natl. Cancer Inst.* **94** 1358–1365.

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

UNGVARI, I., HULLAM, G., ANTAL, P., KISZEL, P. S., GEZSI, A., HADADI, É., VIRÁG, V., HAJÓS, G., MILLINGHOFFER, A., NAGY, A., KISS, A., SEMSEI, Á. F., TEMESI, G., MELEGH, B., KISFALI, P., SZÉLL, M., BIKOV, A., GÁLFFY, G., TAMÁSI, L., FALUS, A. and SZALAI, C. (2012). Evaluation of a partial genome screening of two asthma susceptibility regions using Bayesian network based Bayesian multilevel analysis of relevance. *PLoS One* **7** e33573.

VERZILLI, C. J., STALLARD, N. and WHITTAKER, J. C. (2006). Bayesian graphical models for genomewide association studies. *Am. J. Hum. Genet.* **79** 100–112.

WACHOLDER, S., HARTGE, P., PRENTICE, R., GARCIA-CLOSAS, M., FEIGELSON, H. S., DIVER, W. R., THUN, M. J., COX, D. G., HANKINSON, S. E., KRAFT, P., ROSNER, B., BERG, C. D., BRINTON, L. A., LISSOWSKA, J., SHERMAN, M. E., CHLEBOWSKI, R.,

KOOPERBERG, C., JACKSON, R. D., BUCKMAN, D. W., HUI, P., PFEIFFER, R., JACOBS, K. B., THOMAS, G. D., HOOVER, R. N., GAIL, M. H., CHANOCK, S. J. and HUNTER, D. J. (2010). Performance of common genetic variants in breast-cancer risk models. *N. Engl. J. Med.* **362** 986–993.

WILSON, M. A., IVERSEN, E. S., CLYDE, M. A., SCHMIDLER, S. C. and SCHILDKRAUT, J. M. (2010). Bayesian model search and multilevel inference for SNP association studies. *Ann. Appl. Stat.* **4** 1342–1364. MR2758331

WU, Z. and ZHAO, H. (2009). Statistical power of model selection strategies for genome-wide association studies. *PLoS Genet.* **5** e1000582.

XING, H., MCDONAGH, P. D., BIENKOWSKA, J., CASHORALI, T., RUNGE, K., MILLER, R. E., DECAPRIO, D., CHURCH, B., ROUBENOFF, R., KHALIL, I. G. and CARULLI, J. (2011). Causal modeling using network ensemble simulations of genetic and gene expression data predicts genes involved in rheumatoid arthritis. *PLoS Comput. Biol.* **7** e1001105.

YEUNG, K. Y., BUMGARNER, R. E. and RAFTERY, A. E. (2005). Bayesian model averaging: Development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics* **21** 2394–2402.

ZHANG, Y. (2012). A novel Bayesian graphical model for genome-wide multi-SNP association mapping. *Genet. Epidemiol.* **36** 36–47.

ZHANG, Y. and LIU, J. S. (2007). Bayesian inference of epistatic interactions in case-control studies. *Nat. Genet.* **39** 1167–1173.

The Annals of Applied Statistics
2016, Vol. 10, No. 2, 812–833
DOI: 10.1214/16-AOAS913
© Institute of Mathematical Statistics, 2016

UNDERSTANDING RESIDENT MOBILITY IN MILAN THROUGH INDEPENDENT COMPONENT ANALYSIS OF *TELECOM ITALIA* MOBILE USAGE DATA

BY PAOLO ZANINI[1], HAIPENG SHEN[2] AND YOUNG TRUONG[3]

MOX at Politecnico di Milano, University of Hong Kong and University of North Carolina at Chapel Hill

We consider an urban planning application where *Telecom Italia* collected mobile-phone traffic data in the metropolitan area of Milan, Italy, aiming to retrieve meaningful information regarding working, residential, and mobility activities around the city. The independent component analysis (ICA) framework is used to model underlying spatial activities as spatial processes on a lattice independent of each other. To incorporate spatial dependence within the spatial sources, we develop a spatial colored ICA (scICA) method. The method models spatial dependence within each source in the frequency domain, exploiting the power of Whittle likelihood and local linear log-spectral density estimation. An iterative algorithm is derived to estimate the model parameters through maximum Whittle likelihood. We then apply scICA to the Italian mobile traffic application.

REFERENCES

AHAS, R. and MARK, U. (2005). Location based services—new challenges for planning and public administration. *Futures* **37** 547–561.

AMARI, S., CICHOCKI, A. and YANG, H. H. (1996). A new learning algorithm for blind signal separation. *Adv. Neural Inf. Process. Syst.* 757–763.

BOX, G. E. P. and JENKINS, G. M. (1970). *Times Series Analysis. Forecasting and Control*. Holden-Day, San Francisco. MR0272138

CALHOUN, V. D., ADALI, T., PEARLSON, G. D. and PEKAR, J. J. (2001). Spatial and temporal independent component analysis of functional MRI data containing a pair of task-related waveforms. *Hum. Brain Mapp.* **13** 43–53.

CHEN, A. and BICKEL, P. J. (2005). Consistent independent component analysis and prewhitening. *IEEE Trans. Signal Process.* **53** 3625–3632. MR2239886

CHEN, A. and BICKEL, P. J. (2006). Efficient independent component analysis. *Ann. Statist.* **34** 2825–2855. MR2329469

COMON, P. and JUTTEN, C. (2010). *Handbook of Blind Source Separation*: *Independent Component Analysis and Applications*. Academic Press, San Diego.

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

CRUJEIRAS, R. M. (2006). Contribution to spectral spatial statistics. Ph.D. dissertation, Universidade de Santiago de Compostela.

CRUJEIRAS, R. M. and FERNÁNDEZ-CASAL, R. (2010). On the estimation of the spectral density for continuous spatial processes. *Statistics* **44** 587–600. MR2739414

Key words and phrases. Spatial stochastic processes, periodogram, Whittle likelihood, mobile phone traffic, urban planning.

EDDELBUETTEL, D. and FRANÇOIS, R. (2011). Rcpp: Seamless R and C++ integration. *J. Stat. Softw.* **40** 1–18.

ELOYAN, A., CRAINICEANU, C. M. and CAFFO, B. S. (2013). Likelihood-based population independent component analysis. *Biostatistics* **14** 514–527.

FAN, J. and KREUTZBERGER, E. (1998). Automatic local smoothing for spectral density estimation. *Scand. J. Statist.* **25** 359–369. MR1649039

GEBIZLIOĞLU, Ö. L. (1988). Spatial processes: Modelling, estimation, and hypothesis testing. *Comm. Fac. Sci. Univ. Ankara Ser. A_1 Math. Statist.* **37** 67–94 (1991). MR1203412

GONZÁLEZ, M. C., HIDALGO, C. A. and BARABÁSI, A.-L. (2008). Understanding individual human mobility patterns. *Nature* **453** 779–782.

GUO, Y. (2011). A general probabilistic model for group independent component analysis and its estimation methods. *Biometrics* **67** 1532–1542. MR2872404

HASTIE, T. and TIBSHIRANI, R. (2010). ProDenICA: Product Density Estimation for ICA using tilted Gaussian density estimates. R package version 1.0.

HYVARINEN, A., KARHUNEN, J. and OJA, E. (2001). *Independent Component Analysis*. Wiley, New York.

HYVÄRINEN, A. and OJA, E. (2000). Independent component analysis: Algorithms and applications. *Neural Netw.* **13** 411–430.

KAUFMANN, V. (2012). *Re-Thinking Mobility*: *Contemporary Sociology*. Ashgate, Aldershot.

LEE, S., SHEN, H., TRUONG, Y., LEWIS, M. and HUANG, X. (2011). Independent component analysis involving autocorrelated sources with an application to functional magnetic resonance imaging. *J. Amer. Statist. Assoc.* **106** 1009–1024. MR2894760

LUÈ, A., COLORNI, A., NOCERINO, R. and PARUSCIO, V. (2012). Green move: An innovative electric vehicle-sharing system. *Procedia—Social and Behavioral Sciences* **48** 2978–2987.

MANFREDINI, F., PUCCI, P., SECCHI, P., TAGLIOLATO, P., VANTINI, S. and VITELLI, V. (2015). Treelet decomposition of mobile phone data for deriving city usage and mobility pattern in the Milan urban region. In *Advances in Complex Data Modeling and Computational Methods in Statistics* (A. Paganoni and P. Secchi, eds.) 133–147. Springer, Berlin.

MATTESON, D. S. and TSAY, R. S. (2016). Independent component analysis via distance covariance. *J. Amer. Statist. Assoc.* To appear. DOI:10.1080/01621459.2016.1150851.

OECD (2006). *Territorial Reviews*: *Milan*, *Italy*. OECD Publishing.

R CORE TEAM (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

RATTI, C., PULSELLI, R. M., WILLIAMS, S. and FRENCHMAN, D. (2006). Mobile landscapes: Using location data from cell phones for urban analysis. *Environ. Plann. B, Plann. Des.* **33** 727–748.

SECCHI, P., VANTINI, S. and VITELLI, V. (2015). Analysis of spatio-temporal mobile phone data: A case study in the metropolitan area of Milan. *Stat. Methods Appl.* 1–22.

SHELLER, M. and URRY, J. (2006). The new mobilities paradigm. *Environ. Plann. A.* **38** 207–226.

VAN DE VEN, V. G., FORMISANO, E., PRVULOVIC, D., ROEDER, C. H. and LINDEN, D. E. J. (2004). Functional connectivity as revealed by spatial independent component analysis of fMRI measurements during rest. *Hum. Brain Mapp.* **22** 165–178.

WHITTLE, P. (1952). Some results in time series analysis. *Skand. Aktuarietidskr.* **35** 48–60. MR0049539

The Annals of Applied Statistics
2016, Vol. 10, No. 2, 834–863
DOI: 10.1214/16-AOAS914
© Institute of Mathematical Statistics, 2016

MULTILEVEL MODELING OF INSURANCE CLAIMS USING COPULAS

BY PENG SHI[*,1], XIAOPING FENG[*] AND JEAN-PHILIPPE BOUCHER[†]

University of Wisconsin-Madison[*] and Université du Québec à Montréal[†]

In property-casualty insurance, claims management is featured with the modeling of a semi-continuous insurance cost associated with individual risk transfer. This practice is further complicated by the multilevel structure of the insurance claims data, where a contract often contains a group of policyholders, each policyholder is insured under multiple types of coverage, and the contract is repeatedly observed over time. The data hierarchy introduces a complex dependence structure among claims and leads to diversification in the insurer's liability portfolio.

To capture the unique features of policy-level insurance costs, we propose a copula regression for the multivariate longitudinal claims. In the model, the Tweedie double generalized linear model is employed to examine the semi-continuous claim cost of each coverage type, and a Gaussian copula is specified to accommodate the cross-sectional and temporal dependence among the multilevel claims. Estimation and inference is based on the composite likelihood approach and the properties of parameter estimates are investigated through simulation studies. When applied to a portfolio of personal automobile policies from a Canadian insurer, we show that the proposed copula model provides valuable insights to an insurer's claims management process.

REFERENCES

AAS, K., CZADO, C., FRIGESSI, A. and BAKKEN, H. (2009). Pair-copula constructions of multiple dependence. *Insurance Math. Econom.* **44** 182–198. MR2517884

ARELLANO-VALLE, R. B., CASTRO, L. M., GONZÁLEZ-FARÍAS, G. and MUÑOZ-GAJARDO, K. A. (2012). Student-t censored regression model: Properties and inference. *Stat. Methods Appl.* **21** 453–473. MR2992913

CASTRO, L. M., LACHOS, V. H., FERREIRA, G. P. and ARELLANO-VALLE, R. B. (2014). Partially linear censored regression models using heavy-tailed distributions: A Bayesian approach. *Stat. Methodol.* **18** 14–31. MR3151861

COX, D. R. and REID, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91** 729–737. MR2090633

DUNN, P. K. and SMYTH, G. K. (2005). Series evaluation of Tweedie exponential dispersion model densities. *Stat. Comput.* **15** 267–280. MR2205390

DUNN, P. K. and SMYTH, G. K. (2008). Evaluation of Tweedie exponential dispersion model densities by Fourier inversion. *Stat. Comput.* **18** 73–86. MR2416440

FAHRMEIR, L. and TUTZ, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2nd ed. Springer, New York. MR1832899

Key words and phrases. Composite likelihood, insurance claims, longitudinal data, multivariate regression, property-casualty insurance, Tweedie distribution.

FREES, E. (2014). Frequency and severity models. In *Predictive Modeling Applications in Actuarial Sciences* (E. Frees, G. Meyers and R. Derrig, eds.) 138–166. Cambridge Univ. Press, Cambridge.
FREES, E. W., SHI, P. and VALDEZ, E. A. (2009). Actuarial applications of a hierarchical insurance claims model. *Astin Bull.* **39** 165–197. MR2749883
FREES, E. W. and VALDEZ, E. A. (2008). Hierarchical insurance claims modeling. *J. Amer. Statist. Assoc.* **103** 1457–1469. MR2655723
GAO, X. and SONG, P. X.-K. (2010). Composite likelihood Bayesian information criteria for model selection in high-dimensional data. *J. Amer. Statist. Assoc.* **105** 1531–1540. MR2796569
GARAY, A. M., LACHOS, V. H., BOLFARINE, H. and CABRAL, C. R. (2016). Linear censored regression models with scale mixtures of normal distributions. *Statistical Papers*. To appear.
GODAMBE, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Stat.* **31** 1208–1211. MR0123385
GREENE, W. (2007). *Econometric Analysis*. Prentice Hall, New Jersey.
HINTZE, J. L. and NELSON, R. D. (1998). Violin plots: A box plot-density trace synergism. *Amer. Statist.* **52** 181–184.
JOE, H. (2015). *Dependence Modeling with Copulas*. *Monographs on Statistics and Applied Probability* **134**. CRC Press, Boca Raton, FL. MR3328438
JOE, H. and LEE, Y. (2009). On weighting of bivariate margins in pairwise likelihood. *J. Multivariate Anal.* **100** 670–685. MR2478190
JØRGENSEN, B. (1987). Exponential dispersion models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **49** 127–162. MR0905186
JØRGENSEN, B. and PAES DE SOUZA, M. C. (1994). Fitting Tweedie's compound Poisson model to insurance claims data. *Scand. Actuar. J.* **1** 69–93. MR1286486
KLUGMAN, S., PANJER, H. and WILLMOT, G. (2012). *Loss Models*: *From Data to Decisions*, 4th ed. Wiley, New York.
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
MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman & Hall, London. MR3223057
MOLENBERGHS, G. and VERBEKE, G. (2005). *Models for Discrete Longitudinal Data*. Springer, New York. MR2171048
NELSEN, R. B. (2006). *An Introduction to Copulas*, 2nd ed. Springer, New York. MR2197664
OLSEN, M. K. and SCHAFER, J. L. (2001). A two-part random-effects model for semicontinuous longitudinal data. *J. Amer. Statist. Assoc.* **96** 730–745. MR1946438
PANAGIOTELIS, A., CZADO, C. and JOE, H. (2012). Pair copula constructions for multivariate discrete data. *J. Amer. Statist. Assoc.* **107** 1063–1072. MR3010894
PARKS, R. W. (1967). Efficient estimation of a system of regression equations when disturbances are both serially and contemporaneously correlated. *J. Amer. Statist. Assoc.* **62** 500–509. MR0216633
POURAHMADI, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika* **86** 677–690. MR1723786
POURAHMADI, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika* **87** 425–435. MR1782488
POURAHMADI, M. (2007). Cholesky decompositions and estimation of a covariance matrix: Orthogonality of variance-correlation parameters. *Biometrika* **94** 1006–1013. MR2376812
SHI, P. (2016). Insurance ratemaking using a copula-based multivariate Tweedie model. *Scand. Actuar. J.* **3** 198–215. MR3435180
SHI, P., ZHANG, W. and VALDEZ, E. A. (2012). Testing adverse selection with two-dimensional information: Evidence from the Singapore auto insurance market. *Journal of Risk and Insurance* **79** 1077–1114.

SMITH, M., MIN, A., ALMEIDA, C. and CZADO, C. (2010). Modeling longitudinal data using a pair-copula decomposition of serial dependence. *J. Amer. Statist. Assoc.* **105** 1467–1479. MR2796564

SMYTH, G. K. (1996). Regression analysis of quantity data with exact zeros. In *Proceedings of the Second Australia-Japan Workshop on Stochastic Models in Engineering, Technology and Management* (R. Wilson, S. Osaki and D. Murthy, eds.) 17–19. Gold Coast, Australia.

SMYTH, G. K. and JØRGENSEN, B. (2002). Fitting Tweedie's compound Poisson model to insurance claims data: Dispersion modelling. *Astin Bull.* **32** 143–157. MR1930491

SONG, P. X.-K., LI, M. and YUAN, Y. (2009). Joint regression analysis of correlated data using Gaussian copulas. *Biometrics* **65** 60–68. MR2665846

TOBIN, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica* **26** 24–36. MR0090462

TWEEDIE, M. C. K. (1984). An index which distinguishes between some important exponential families. In *Statistics*: *Applications and New Directions* (*Calcutta*, 1981) 579–604. Indian Statist. Inst., Calcutta. MR0786162

VARIN, C. (2008). On composite marginal likelihoods. *AStA nnv. Stat. Anal.* **92** 1–28. MR2414624

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

VARIN, C. and VIDONI, P. (2005). A note on composite likelihood inference and model selection. *Biometrika* **92** 519–528. MR2202643

ZHANG, Y. (2013). Likelihood-based and Bayesian methods for Tweedie compound Poisson linear mixed models. *Stat. Comput.* **23** 743–757. MR3247830

ZHAO, Y. and JOE, H. (2005). Composite likelihood estimation in multivariate data analysis. *Canad. J. Statist.* **33** 335–356. MR2193979

The Annals of Applied Statistics
2016, Vol. 10, No. 2, 864–883
DOI: 10.1214/16-AOAS916
© Institute of Mathematical Statistics, 2016

LEVEL-SCREENING DESIGNS FOR FACTORS WITH MANY LEVELS

BY PHILIP J. BROWN AND MARTIN S. RIDOUT

University of Kent

We consider designs for f factors each at m levels, where f is small but m is large. Main effect designs with mf experimental points are presented. For two factors, two types of designs are investigated, termed *sawtooth* and *dumbbell* designs, based on a graphical representation. For three factors, cyclic sawtooth designs are considered. The paper seeks optimal and near optimal designs which involve factors with many levels but few observations. It also investigates issues of robustness when as much as one third of the data is structurally missing. An important area of application is in screening for drug discovery and we compare our designs with others using a published data set with two factors each with fifty levels, where the dumbbell design outperforms others and is an example of an inherently unbalanced design dominating more balanced designs.

REFERENCES

ATKINSON, A. C. and DONEV, A. N. (1992). *Optimum Experimental Designs*. Oxford Univ. Press, Oxford.

BAILEY, R. A. (2007). Designs for two-colour microarray experiments. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **56** 365–394. MR2409757

BALANI, S. K., MIWA, G. T., GAN, L. S. and LEE, F. W. (2005). Strategy of utilizing in vitro and in vivo ADME tools for lead optimisation and drug candidate selection. *Current Topics in Medicinal Chemistry* **5** 1033–1038.

BORROTTI, M., MARCH, D. D., SLANZI, D. and POLI, I. (2014). Designing lead optimisation of MMP-12 inhibitors. *Comput. Math. Methods Med.* **2014** 258627.

BOX, G. E. P., HUNTER, J. S. and HUNTER, W. G. (2005). *Statistics for Experimenters*: *Design*, *Innovation*, *and Discovery*, 2nd ed. Wiley-Interscience, Hoboken, NJ. MR2140250

BROWN, P. J. and RIDOUT, M. S. (2016). Supplement to "Level-screening designs for factors with many levels." DOI:10.1214/16-AOAS916SUPP.

DANIEL, C. (1973). One-at-a-time plans. *J. Amer. Statist. Assoc.* **68** 353–360.

DETTE, H. and O'BRIEN, T. E. (1999). Optimality criteria for regression models based on predicted variance. *Biometrika* **86** 93–106. MR1688074

FRANCK, C. T., NIELSEN, D. M. and OSBORNE, J. A. (2013). A method for detecting hidden additivity in two-factor unreplicated experiments. *Comput. Statist. Data Anal.* **67** 95–104. MR3079590

HALL, W. B. and WILLIAMS, E. R. (1973). Cyclic superimposed designs. *Biometrika* **60** 47–53. MR0359203

JOHN, J. A. and WILLIAMS, E. R. (1995). *Cyclic and Computer Generated Designs*, 2nd ed. *Monographs on Statistics and Applied Probability* **38**. Chapman & Hall, London. MR1382127

Key words and phrases. Screening designs, lead optimization in drug discovery, main effects, microarray loop designs, connectivity, identifiability, prediction and contrast variance.

KERR, M. K. and CHURCHILL, G. A. (2001). Experimental design for gene expression microarrays. *Biostatistics* **2** 183–201.

MEYER, R. D., STEINBERG, D. M. and BOX, G. E. P. (1996). Follow-up designs to resolve confounding in multifactor experiments (with discussion). *Technometrics* **38** 303–313.

PICKETT, S. D., GREEN, D. V. S., HUNT, D. L., PARDOE, D. A. and HUGHES, I. (2011). Automated lead optimisation of MMP-12 inhibitors using a genetic algorithm. *ACS Medicinal Chemistry Letters* **2** 28–33.

THAYER, A. M. (1996). Combinatorial chemistry becoming a core technology at drug development companies. *Chem. Eng. News* **74** 57–64.

TJUR, T. (1991). Block designs and electrical networks. *Ann. Statist.* **19** 1010–1027. MR1105858

VINCIOTTI, V., KHANIN, R., D'ALEMONTE, D., DE JESUS, O., RASAIYAAH, J., SMITH, C. P., KELLAM, P. and WIT, E. (2005). An experimental evaluation of a loop versus a reference design for two channel microarrays. *Bioinformatics* **21** 492–501.

WYNN, H. P. (2008). Algebraic solutions to the connectivity problem for m-way layouts: Interaction-contrast aliasing. *J. Statist. Plann. Inference* **138** 259–271. MR2369631

The Annals of Applied Statistics
2016, Vol. 10, No. 2, 884–905
DOI: 10.1214/16-AOAS917
© Institute of Mathematical Statistics, 2016

A BAYESIAN HIERARCHICAL SPATIAL MODEL FOR DENTAL CARIES ASSESSMENT USING NON-GAUSSIAN MARKOV RANDOM FIELDS[1]

BY ICK HOON JIN, YING YUAN AND DIPANKAR BANDYOPADHYAY

University of Notre Dame, University of Texas and Virginia Commonwealth University

Research in dental caries generates data with two levels of hierarchy: that of a tooth overall and that of the different surfaces of the tooth. The outcomes often exhibit spatial referencing among neighboring teeth and surfaces, that is, the disease status of a tooth or surface might be influenced by the status of a set of proximal teeth/surfaces. Assessments of dental caries (tooth decay) at the tooth level yield binary outcomes indicating the presence/absence of teeth, and trinary outcomes at the surface level indicating healthy, decayed or filled surfaces. The presence of these mixed discrete responses complicates the data analysis under a unified framework. To mitigate complications, we develop a Bayesian two-level hierarchical model under suitable (spatial) Markov random field assumptions that accommodates the natural hierarchy within the mixed responses. At the first level, we utilize an autologistic model to accommodate the spatial dependence for the tooth-level binary outcomes. For the second level and conditioned on a tooth being nonmissing, we utilize a Potts model to accommodate the spatial referencing for the surface-level trinary outcomes. The regression models at both levels were controlled for plausible covariates (risk factors) of caries and remain connected through shared parameters. To tackle the computational challenges in our Bayesian estimation scheme caused due to the doubly-intractable normalizing constant, we employ a double Metropolis–Hastings sampler. We compare and contrast our model performances to the standard nonspatial (naive) model using a small simulation study, and illustrate via an application to a clinical dataset on dental caries.

REFERENCES

AFROUGHI, S., FAGHIHZADEH, S., KHALEDI, M. J. and MOTLAGH, M. G. (2010). Dental caries analysis in 3–5-years-old children: A spatial modelling. *Arch. Oral Biol.* **55** 374–378.

ALFÓ, M., NIEDDU, L. and VICARI, D. (2009). Finite mixture models for mapping spatially dependent disease counts. *Biom. J.* **51** 84–97. MR2667513

BADER, J. D., VOLLMER, W. M., SHUGARS, D. A., GILBERT, G. H., AMAECHI, B. T., BROWN, J. P., LAWS, R. L., FUNKHOUSER, K. A., MAKHIJA, S. K., RITTER, A. V. et al. (2013). Results from the Xylitol for Adult Caries Trial (X-ACT). *J. Am. Dent. Assoc.* **144** 21–30.

BANDYOPADHYAY, D., REICH, B. J. and SLATE, E. H. (2009). Bayesian modeling of multivariate spatial binary data with applications to dental caries. *Stat. Med.* **28** 3492–3508. MR2744378

Key words and phrases. Autologistic model, Bayesian inference, dental caries, Markov chain Monte Carlo, Potts model, spatial data analysis.

BESAG, J. E. (1972). Nearest-neighbour systems and the auto-logistic model for binary data. *J. Roy. Statist. Soc. Ser. B* **34** 75–83. MR0323276

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

BRÉMAUD, P. (1999). *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues. Texts in Applied Mathematics* **31**. Springer, New York. MR1689633

BURNSIDE, G., PINE, C. M. and WILLIAMSON, P. R. (2007). The application of multilevel modelling to dental caries data. *Stat. Med.* **26** 4139–4149. MR2405797

DARBY, M. L. and WALSH, M. M. (1995). *Dental Hygiene: Theory and Practice*. W. B. Saunders Company.

FEATHERSTONE, J. D. (2000). The science and practice of caries prevention. *J. Am. Dent. Assoc.* **131** 887–899.

FERNANDES, J., SLATE, E. H., WIEGAND, R. E., LONDON, S. D., GREWAL, J. S., WERNER, P., SANDERS, J. J., LOPES-VIRELLA, M. and SALINAS, C. F. (2007). Dental caries in type 2 Gullah diabetics. *J. Dent. Res.* **86** 1054.

GARCÍA-ZATTERA, M. J., JARA, A., LESAFFRE, E. and DECLERCK, D. (2007). Conditional independence of multivariate binary data with an application in caries research. *Comput. Statist. Data Anal.* **51** 3223–3234. MR2345637

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2013). *Bayesian Data Analysis*, 3nd ed. Chapman & Hall, New York, NY.

GELMAN, A., JAKULIN, A., PITTAU, M. G. and SU, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.* **2** 1360–1383. MR2655663

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

HIGDON, D. M. (1998). Auxiliary variable methods for Markov chain Monte Carlo with applications. *J. Amer. Statist. Assoc.* **93** 585–595.

HOETING, J. A., LEECASTER, M. and BOWDEN, D. (2000). An improved model for spatially correlated binary responses. *J. Agric. Biol. Environ. Stat.* **5** 102–114. MR1817027

JAMISON, D. T., BREMAN, J. G., MEASHAM, A. R., ALLEYNE, G., GLAESON, M., EVANS, D. B., JHA, P., MILLS, A. and MUSGROVE, P. (2006). *Disease Control Priorities in Developing Countries*, 2nd ed. Oxford Univ. Press, London.

JIN, I. H., YUAN, Y. and BANDYOPADHYAY, D. (2016). Supplement to "A Bayesian hierarchical spatial model for dental caries assessment using non-Gaussian Markov random fields." DOI:10.1214/16-AOAS917SUPP.

JOHNSON, T. D. and PIERT, M. (2009). A Bayesian analysis of dual autoradiographic images. *Comput. Statist. Data Anal.* **53** 4570–4583. MR2744348

JOHNSON, T. D., LIU, Z., BARTSCH, A. J. and NICHOLS, T. E. (2012). A Bayesian non-parametric Potts model with application to pre-surgical FMRI data. *Stat. Methods Med. Res.* **22** 364–381. DOI:10.1177/0962280212448970.

KIDD, E. A. M., SMITH, B. G. N. and WATSON, T. F. (2003). *Pickard's Manual of Operative Dentistry*, 8th ed. Oxford Univ. Press, London.

LIANG, F. (2010). A double Metropolis–Hastings sampler for spatial models with intractable normalizing constants. *J. Stat. Comput. Simul.* **80** 1007–1022. MR2742519

LIANG, F., LIU, C. and CARROLL, R. J. (2007). Stochastic approximation in Monte Carlo computation. *J. Amer. Statist. Assoc.* **102** 305–320. MR2345544

LIM, J., WANG, X. and SHERMAN, M. (2007). An adjustment for edge effects using an augmented neighborhood model in the spatial auto-logistic model. *Comput. Statist. Data Anal.* **51** 3679–3688. MR2364483

MURRAY, I., GHAHRAMANI, Z. and MACKAY, D. J. C. (2006). MCMC for doubly-intractable distributions. In *Proceedings of* 22*nd Annual Conference on Uncertainty in Artificial Intelligence* (*UAI*).

MUSTVARI, T., BANDYOPADHYAY, D., LESAFFRE, E. and DECLERCK, D. (2013). A multilevel model for spatially correlated binary data in the presence of misclassification: An application in oral health research. *Stat. Med.* DOI:10.1002/sim.5944.

NELSEN, R. B. (2006). *An Introduction to Copulas*, 2nd ed. Springer, New York. MR2197664

PITTS, N. B. (2004). Are we ready to move from operative to non-operative/preventive treatment of dental caries in clinical practice? *Caries Res.* **38** 294–304.

POTTS, R. B. (1952). Some generalized order-disorder transformations. *Proc. Cambridge Philos. Soc.* **48** 106–109. MR0047571

PREISLER, H. K. (1993). Modeling spatial patterns of trees attacked by Bark-beetles. *Appl. Stat.* **42** 501–514.

REICH, B. J. and BANDYOPADHYAY, D. (2010). A latent factor model for spatial data with informative missingness. *Ann. Appl. Stat.* **4** 439–459. MR2758179

REICH, B. J., BANDYOPADHYAY, D. and BONDELL, H. D. (2013). A nonparametric spatial model for periodontal data with nonrandom missingness. *J. Amer. Statist. Assoc.* **108** 820–831. MR3174665

REICH, B. J. and HODGES, J. S. (2008). Modeling longitudinal spatial periodontal data: A spatially adaptive model with tools for specifying priors and checking fit. *Biometrics* **64** 790–799. MR2526629

REICH, B. J., HODGES, J. S. and CARLIN, B. P. (2007). Spatial analyses of periodontal data using conditionally autoregressive priors having two classes of neighbor relations. *J. Amer. Statist. Assoc.* **102** 44–55. MR2345531

SELWITZ, R. H., ISMAIL, A. I. and PITTS, N. B. (2007). Dental caries. *Lancet* **369** 51–59.

SHERMAN, M., APANASOVICH, T. V. and CARROLL, R. J. (2006). On estimation in binary autologistic spatial models. *J. Stat. Comput. Simul.* **76** 167–179. MR2223145

SOAMES, J. V. and SOUTHAM, J. C. (1993). *Oral Pathology*, 2nd ed. Oxford Univ. Press, London.

WINKLER, G. (2003). *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods*, 2nd ed. *Applications of Mathematics* (*New York*) **27**. Springer, Berlin. MR1950762

WU, H. and HUFFER, F. W. (1997). Modeling the distribution of plant species using the autologistic regression model. *Ecological Statistics* **4** 49–64.

ZHANG, X., JOHNSON, T. D., LITTLE, R. J. A. and CAO, Y. (2010). A Bayesian image analysis of radiation induced changes in tumor vascular permeability. *Bayesian Anal.* **5** 189–212. MR2596441

ZHU, H., HE, F. and ZHOU, J. (2008). Auto-multicategorical regression model for the distribution of vegetation. *Stat. Interface* **1** 63–73. MR2425345

The Annals of Applied Statistics
2016, Vol. 10, No. 2, 906–924
DOI: 10.1214/16-AOAS918
© Institute of Mathematical Statistics, 2016

A BAYESIAN APPROACH TO THE SEMIPARAMETRIC ESTIMATION OF A MINIMUM INHIBITORY CONCENTRATION DISTRIBUTION[1]

BY STIJN JASPERS[*,2], PHILIPPE LAMBERT[†,‡] AND MARC AERTS[*]

Hasselt University,[*] Université de Liège[†] and Institut de Statistique, Biostatistique et Sciences Actuarielles (ISBA), Université Catholique de Louvain[‡]

Bacteria that have developed a reduced susceptibility against antimicrobials pose a major threat to public health. Hence, monitoring their distribution in the general population is of major importance. This monitoring is performed based on minimum inhibitory concentration (MIC) values, which are collected through dilution experiments. We present a semiparametric mixture model to estimate the MIC density on the full continuous scale. The wild-type first component is assumed to be of a parametric form, while the nonwild-type second component is modelled nonparametrically using Bayesian P-splines combined with the composite link model. A Metropolis within Gibbs strategy was used to draw a sample from the joint posterior. The newly developed method was applied to a specific bacterium–antibiotic combination, that is, *Escherichia coli* tested against ampicillin. After obtaining an estimate for the entire density, model-based classification can be performed to check whether or not an isolate belongs to the wild-type subpopulation. The performance of the new method, compared to two existing competitors, is assessed through a small simulation study.

REFERENCES

2013/652/EU (2013). Commission Implementing Decision of 12 November 2013 on the monitoring and reporting of antimicrobial resistance in zoonotic and commensal bacteria (notified under document C(2013) 7145). Text with EEA relevance.

ANDREWS, J. M. (2001). Determination of minimum inhibition concentrations. *J. Antimicrob. Chemother.* **48** S1.5–S1.16.

ANNIS, D. H. and CRAIG, B. A. (2005). Statistical properties and inference of the antimicrobial MIC test. *Stat. Med.* **24** 3631–3644. MR2212304

ATCHADÉ, Y. F. and ROSENTHAL, J. S. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli* **11** 815–828. MR2172842

AZZALINI, A. and CAPITANIO, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65** 367–389. MR1983753

BÖHNING, D. (1986). A vertex-exchange-method in D-optimal design theory. *Metrika* **33** 337–347. MR0868043

CRAIG, B. A. (2000). Modeling approach to diameter breakpoint determination. *Diagn. Microbiol. Infect. Dis.* **36** 193–202.

Key words and phrases. Antimicrobial resistance, Bayesian, composite link model, interval-censored, semiparametric.

EILERS, P. H. C. (2007). Ill-posed problems with counts, the composite link model and penalized likelihood. *Stat. Model.* **7** 239–254. MR2749992

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

FINCH, R. G., GREENWOOD, D., WHITLEY, R. J. and NORRBY, S. R. (2010). *Antibiotic and Chemotherapy*. Saunders, Elsevier.

HAARIO, H., SAKSMAN, E. and TAMMINEN, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7** 223–242. MR1828504

IGARASHI, T., INATOMI, J., WAKE, A., TAKAMIZAWA, M., KATAYAMA, H. and IWATA, T. (1999). Failure of pre-diarrheal antibiotics to prevent hemolytic uremic syndrome in serologically proven *Escherichia coli* O157:H7 gastrointestinal infection. *J. Pediatr.* **135** 768–769.

JASPERS, S., AERTS, M., VERBEKE, G. and BELOEIL, P.-A. (2014a). Estimation of the wild-type minimum inhibitory concentration value distribution. *Stat. Med.* **33** 289–303. MR3146764

JASPERS, S., AERTS, M., VERBEKE, G. and BELOEIL, P.-A. (2014b). A new semi-parametric mixture model for interval censored data, with applications in the field of antimicrobial resistance. *Comput. Statist. Data Anal.* **71** 30–42. MR3131952

JASPERS, S., VERBEKE, G., BÖHNING, D. and AERTS, M. (2016). Application of the Vertex Exchange Method to estimate a semi-parametric mixture model for the MIC density of *Escherichia coli* isolates tested for susceptibility against ampicillin. *Biostatistics* **17** 94–107. MR3449853

KAHLMETER, G., BROWN, D. F. J., GOLDSTEIN, F. W., MACGOWAN, A. P., MOUTON, J. W., OSTERLUND, A., RODLOFF, A., STEINBAKK, M., URBASKOVA, P. and VATOPOULOS, A. (2003). European harmonization of MIC breakpoints for antimicrobial susceptibility testing of bacteria. *Journal of Antimicrobial Chemotherapy* **52** 145–148.

KRONVALL, G. (2010). Antimicrobial resistance 1979–2009 at Karolinska hospital, Sweden: Normalized resistance interpretation during a 30-year follow-up on *Staphylococcus aureus* and *Escherichia coli* resistance development. *APMIS* **118** 621–639.

LAMBERT, P. and EILERS, P. H. C. (2009). Bayesian density estimation from grouped continuous data. *Comput. Statist. Data Anal.* **53** 1388–1399. MR2657099

LANG, S. and BREZGER, A. (2004). Bayesian P-splines. *J. Comput. Graph. Statist.* **13** 183–212. MR2044877

LEE, M. L. T. and WHITMORE, G. A. (1999). Statistical inference for serial dilution assay data. *Biometrics* **55** 1215–1220.

PALUMBI, S. R. (2001). Humans as the world's greatest evolutionary force. *Science* **293** 1786–1790.

SCHELLHASE, C. and KAUERMANN, G. (2012). Density estimation and comparison with a penalized mixture approach. *Comput. Statist.* **27** 757–777. MR3041856

TADESSE, D. A., ZHAO, S., TONG, E., AYERS, S., SINGH, A., BARTHOLOMEW, M. J. and MCDERMOTT, P. F. (2012). Antimicrobial drug resistance in *Escherichia coli* from humans and food Animals, United States, 1950–2002. *Emerg. Infect. Dis.* **18 (5)** 741–749.

THOMPSON, R. and BAKER, R. J. (1981). Composite link functions in generalized linear models. *J. Roy. Statist. Soc. Ser. C* **30** 125–131. MR0629493

TURNIDGE, J., KAHLMETER, G. and KRONVALL, G. (2006). Statistical characterisation of bacterial wild-type MIC value distributions and the determination of epidemiological cut-off values. *Clin. Microbiol. Infect.* **12** 418–425.

WIEGAND, I., HILPERT, K. and HANCOCK, R. E. W. (2008). Agar and broth dilution methods to determine the minimal inhibitory concentration (MIC) of antimicrobial substances. *Nat. Protoc.* **3** 163–175.

The Annals of Applied Statistics
2016, Vol. 10, No. 2, 925–945
DOI: 10.1214/16-AOAS919
© Institute of Mathematical Statistics, 2016

UNMIXING RASCH SCALES: HOW TO SCORE AN EDUCATIONAL TEST

BY MARIA BOLSINOVA[*,†], GUNTER MARIS[†,‡] AND HERBERT HOIJTINK[*]

Utrecht University[*], CITO Dutch Institute for Educational Measurement[†] and University of Amsterdam[‡]

One of the important questions in the practice of educational testing is how a particular test should be scored. In this paper we consider what an appropriate simple scoring rule should be for the Dutch as a second language test consisting of listening and reading items. As in many other applications, here the Rasch model which allows to score the test with a simple sumscore is too restrictive to adequately represent the data. In this study we propose an exploratory algorithm which clusters the items into subscales each fitting a Rasch model and thus provides a scoring rule based on observed data. The scoring rule produces either a weighted sumscore based on equal weights within each subscale or a set of sumscores (one for each of the subscales). An MCMC algorithm which enables to determine the number of Rasch scales constituting the test and to unmix these scales is introduced and evaluated in simulations. Using the results of unmixing, we conclude that the Dutch language test can be scored with a weighted sumscore with three different weights.

REFERENCES

ADAMS, R., WILSON, M. and WANG, W. (1997). The multidimensional random coefficients multinomial logit model. *Appl. Psychol. Meas.* **12** 261–280.

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

ANDERSEN, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika* **38** 123–140. MR0311064

ANDERSEN, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika* **42** 69–81. MR0483255

BIRNBAUM, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In *Statistical Theories of Mental Test Scores* (F. M. Lord and M. R. Novick, eds.) 395–479. Addison-Wesley, Reading, MA.

BOLSINOVA, M., MARIS, G. and HOIJTINK, H. (2016). Supplement to "Unmixing Rasch scales: How to score an educational test." DOI:10.1214/16-AOAS919SUPP.

CASELLA, G. and GEORGE, E. I. (1992). Explaining the Gibbs sampler. *Amer. Statist.* **46** 167–174. MR1183069

CELEUX, G., HURN, M. and ROBERT, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *J. Amer. Statist. Assoc.* **95** 957–970. MR1804450

COLLEGE VOOR TOETSEN EN EXAMENS: STAATSEXAMENS NT2 (n.d.). Retrieved September 25, 2015. Available at http://www.staatsexamensnt2.nl.

Key words and phrases. Educational testing, Markov chain Monte Carlo, mixture model, multidimensional IRT, one parameter logistic model, Rasch model, scoring rule.

COUNCIL OF EUROPE (2011). Common European Framework of Reference for Learning, Teaching, Assessment. Council of Europe.

DEBELAK, R. and ARENDASY, M. (2012). An algorithm for testing unidimensionality and clustering items in Rasch measurement. *Educ. Psychol. Meas.* **72** 375–387.

DIEBOLT, J. and ROBERT, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *J. Roy. Statist. Soc. Ser. B* **56** 363–375. MR1281940

FISCHER, G. H. (1995). Derivations of the Rasch model. In *Rasch Models* (*Vienna*, 1993) (G. H. Fisher and I. W. Molenaar, eds.) 15–38. Springer, New York. MR1367343

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

GAMERMAN, D. and LOPES, H. F. (2006). *Markov Chain Monte Carlo*: *Stochastic Simulation for Bayesian Inference*, 2nd ed. Chapman & Hall/CRC, Boca Raton, FL. MR2260716

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6** 721–741.

GHOSH, M., GHOSH, A., CHEN, M. and AGRESTI, A. (2000). Noninformative priors for one-parameter item response models. *J. Statist. Plann. Inference* **88** 99–115.

HARDOUIN, J.-B. and MESBAH, M. (2004). Clustering binary variables in subscales using an extended Rasch model and Akaike information criterion. *Comm. Statist. Theory Methods* **33** 1277–1294. MR2069568

HOFF, P. D. (2009). *A First Course in Bayesian Statistical Methods*. *Springer Texts in Statistics*. Springer, New York. MR2648134

HUMPHRY, S. (2011). The role of the unit in physics and psychometrics. *Measurement*: *Interdisciplinary Research and Perspective* **9** 1–24.

HUMPHRY, S. (2012). Item set discrimination and the unit in the Rasch model. *J. Appl. Meas.* **13** 165–224.

HUMPHRY, S. and ANDRICH, D. (2008). Understanding the unit in the Rasch model. *J. Appl. Meas.* **9** 249–264.

LORD, F. M. and NOVICK, M. R. (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA.

MAIR, P. and HATZINGER, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *J. Stat. Softw.* **20** 1–20.

MCLACHLAN, G. and PEEL, D. (2000). *Finite Mixture Models*. Wiley, New York. MR1789474

RASCH, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*, expanded edition. The Univ. Chicago Press, Chicago.

RECKASE, M. (2008). *Multidimensional Item Response Theory*. Springer, New York, NY.

ROST, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement* **14** 271–282.

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

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82** 528–550. MR0898357

VERHELST, N. D. and GLAS, C. A. W. (1995). The one parameter logistic model: OPLM. In *Rasch Models*: *Foundations*, *Recent Developments and Applications* (G. H. Fischer and I. W. Molenaar, eds.) 215–238. Springer, New York.

ZEGER, K. and KARIM, M. (1991). Generalized linear models with random effects: A Gibbs sampling approach. *J. Amer. Statist. Assoc.* **86** 79–86.

The Annals of Applied Statistics
2016, Vol. 10, No. 2, 946–963
DOI: 10.1214/16-AOAS920
© Institute of Mathematical Statistics, 2016

ROBUST HYPERPARAMETER ESTIMATION PROTECTS AGAINST HYPERVARIABLE GENES AND IMPROVES POWER TO DETECT DIFFERENTIAL EXPRESSION[1]

BY BELINDA PHIPSON*, STANLEY LEE[†,‡], IAN J. MAJEWSKI[†,‡], WARREN S. ALEXANDER[†,‡] AND GORDON K. SMYTH[†,‡]

Murdoch Childrens Research Institute, The Walter and Eliza Hall Institute of Medical Research*† *and The University of Melbourne*‡

One of the most common analysis tasks in genomic research is to identify genes that are differentially expressed (DE) between experimental conditions. Empirical Bayes (EB) statistical tests using moderated genewise variances have been very effective for this purpose, especially when the number of biological replicate samples is small. The EB procedures can, however, be heavily influenced by a small number of genes with very large or very small variances. This article improves the differential expression tests by robustifying the hyperparameter estimation procedure. The robust procedure has the effect of decreasing the informativeness of the prior distribution for outlier genes while increasing its informativeness for other genes. This effect has the double benefit of reducing the chance that hypervariable genes will be spuriously identified as DE while increasing statistical power for the main body of genes. The robust EB algorithm is fast and numerically stable. The procedure allows exact small-sample null distributions for the test statistics and reduces exactly to the original EB procedure when no outlier genes are present. Simulations show that the robustified tests have similar performance to the original tests in the absence of outlier genes but have greater power and robustness when outliers are present. The article includes case studies for which the robust method correctly identifies and downweights genes associated with hidden covariates and detects more genes likely to be scientifically relevant to the experimental conditions. The new procedure is implemented in the limma software package freely available from the Bioconductor repository.

REFERENCES

BALDI, P. and LONG, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes. *Bioinformatics* **17** 509–519.

BARLOW, R. E., BARTHOLOMEW, D. J., BREMNER, J. M. and BRUNK, H. D. (1972). *Statistical Inference Under Order Restrictions*. Wiley, London. MR0326887

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

BERGER, J. O. (1984). The robust Bayesian viewpoint. In *Robustness of Bayesian Analyses* (J. Kadane, ed.). *Stud. Bayesian Econometrics* **4** 63–144. North-Holland, Amsterdam. With comments and with a reply by the author. MR0785367

Key words and phrases. Empirical Bayes, outliers, robustness, gene expression, microarrays.

BERGER, J. O. (1990). Robust Bayesian analysis: Sensitivity to the prior. *J. Statist. Plann. Inference* **25** 303–328. MR1064429

BOLSTAD, B. M., IRIZARRY, R. A., ÅSTRAND, M. and SPEED, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19** 185–193.

BRENT, R. P. (1973). *Algorithms for Minimization Without Derivatives*. Prentice-Hall, Englewood Cliffs, NJ.

CASELLA, G. (1985). An introduction to empirical Bayes data analysis. *Amer. Statist.* **39** 83–87. MR0789118

CHEN, Y., LUN, A. T. L. and SMYTH, G. K. (2014). Differential expression analysis of complex RNA-seq experiments using edgeR. In *Statistical Analysis of Next Generation Sequence Data* (S. Datta and D. S. Nettleton, eds.) 51–74. Springer, New York.

CLEVELAND, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74** 829–836. MR0556476

EFRON, B. and MORRIS, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators. II. The empirical Bayes case. *J. Amer. Statist. Assoc.* **67** 130–139. MR0323015

EFRON, B. and MORRIS, C. (1973). Stein's estimation rule and its competitors—An empirical Bayes approach. *J. Amer. Statist. Assoc.* **68** 117–130. MR0388597

GAVER, D. P. and O'MUIRCHEARTAIGH, I. G. (1987). Robust empirical Bayes analyses of event rates. *Technometrics* **29** 1–15. MR0876882

GOOD-JACOBSON, K. L., CHEN, Y., VOSS, A. K., SMYTH, G. K., THOMAS, T. and TARLINTON, D. (2014). Regulation of germinal center responses and B-cell memory by the chromatin modifier MOZ. *Proc. Natl. Acad. Sci. USA* **111** 9585–9590.

GOTTARDO, R., RAFTERY, A. E., YEUNG, K. Y. and BUMGARNER, R. E. (2006). Bayesian robust inference for differential gene expression in microarrays with multiple samples. *Biometrics* **62** 10–18. MR2226551

INSUA, D. R. and RUGGERI, F., eds. (2000). *Robust Bayesian Analysis*. *Lecture Notes in Statistics* **152**. Springer, New York. MR1795206

JEANMOUGIN, M., DE REYNIES, A., MARISA, L., PACCARD, C., NUEL, G. and GUEDJ, M. (2010). Should we abandon the t-test in the analysis of gene expression microarray data: A comparison of variance modeling strategies. *PLoS ONE* **5** e12336.

JI, H. and LIU, X. S. (2010). Analyzing 'omics data using hierarchical models. *Nature Biotechnology* **28** 337.

KOOPERBERG, C., ARAGAKI, A., STRAND, A. D. and OLSON, J. M. (2005). Significance testing for small microarray experiments. *Stat. Med.* **24** 2281–2298. MR2151706

LAW, C. W., CHEN, Y., SHI, W. and SMYTH, G. K. (2014). Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology* **15** R29.

LIAO, J. G., MCMURRY, T. and BERG, A. (2014). Prior robust empirical Bayes inference for large-scale data by conditioning on rank with application to microarray data. *Biostatistics* **15** 60–73.

LUN, A. T. L., CHEN, Y. and SMYTH, G. K. (2016). It's DE-licious: A recipe for differential expression analyses of RNA-seq experiments using quasi-likelihood methods in edgeR. *Methods in Molecular Biology* **1418** 391–416.

LUN, A. T. L. and SMYTH, G. K. (2015a). diffHic: A bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics* **16** 258.

LUN, A. T. L. and SMYTH, G. K. (2015b). From reads to regions: A Bioconductor workflow to detect differential binding in ChIP-seq data. *F1000Research* **4** 1080.

LUN, A. T. L. and SMYTH, G. K. (2016). csaw: A bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. *Nucleic Acids Res.* **44** e45.

MAJEWSKI, I. J., BLEWITT, M. E., DE GRAAF, C. A., MCMANUS, E. J., BAHLO, M., HILTON, A. A., HYLAND, C. D., SMYTH, G. K., CORBIN, J. E., METCALF, D. et al. (2008).

Polycomb repressive complex 2 (PRC2) restricts hematopoietic stem cell activity. *PLOS Biology* **6** e93.

MAJEWSKI, I. J., RITCHIE, M. E., PHIPSON, B., CORBIN, J., PAKUSCH, M., EBERT, A., BUSSLINGER, M., KOSEKI, H., HU, Y., SMYTH, G. K. et al. (2010). Opposing roles of polycomb repressive complexes in hematopoietic stem and progenitor cells. *Blood* **116** 731–739.

MCCARTHY, D. J. and SMYTH, G. K. (2009). Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics* **25** 765–771.

MORRIS, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *J. Amer. Statist. Assoc.* **78** 47–65. With discussion. MR0696849

MURIE, C., WOODY, O., LEE, A. Y. and NADON, R. (2009). Comparison of small *n* statistical tests of differential expression applied to microarrays. *BMC Bioinformatics* **10** 45.

PHIPSON, B., LEE, S., MAJEWSKI, I. J., ALEXANDER, W. S. and SMYTH, G. K. (2016). Supplement to "Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression." DOI:10.1214/16-AOAS920SUPP.

PICKRELL, J. K., MARIONI, J. C., PAI, A. A., DEGNER, J. F., ENGELHARDT, B. E., NKADORI, E., VEYRIERAS, J.-B., STEPHENS, M., GILAD, Y. and PRITCHARD, J. K. (2010a). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464** 768–772.

PICKRELL, J. K., PAI, A. A., GILAD, Y. and PRITCHARD, J. K. (2010b). Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet.* **6** e1001236.

RITCHIE, M. E., PHIPSON, B., WU, D., HU, Y., LAW, C. W., SHI, W. and SMYTH, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43** e47.

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

SARTOR, M. A., TOMLINSON, C. R., WESSELKAMPER, S. C., SIVAGANESAN, S., LEIKAUF, G. D. and MEDVEDOVIC, M. (2006). Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments. *BMC Bioinformatics* **7** 538.

SHEIKH, B. N., DOWNER, N. L., PHIPSON, B., VANYAI, H. K., KUEH, A. J., MCCARTHY, D. J., SMYTH, G. K., THOMAS, T. and VOSS, A. K. (2015). MOZ and BMI1 play opposing roles during Hox gene activation in ES cells and in body segment identity specification in vivo. *Proc. Natl. Acad. Sci. USA* **112** 5437–5442.

SHI, W., OSHLACK, A. and SMYTH, G. K. (2010). Optimizing the noise versus bias trade-off for Illumina whole genome expression BeadChips. *Nucleic Acids Res.* **38** e204.

SMYTH, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3** Article 3. MR2101454

TUKEY, J. W. (1962). The future of data analysis. *Ann. Math. Stat.* **33** 1–67. MR0133937

WRIGHT, G. W. and SIMON, R. M. (2003). A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* **19** 2448–2455.

ZHOU, X., LINDSAY, H. and ROBINSON, M. D. (2014). Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Res.* **42** e91.

The Annals of Applied Statistics
2016, Vol. 10, No. 2, 964–988
DOI: 10.1214/16-AOAS923
© Institute of Mathematical Statistics, 2016

CLUSTERING CHLOROPHYLL-A SATELLITE DATA USING QUANTILES

BY CARLO GAETAN*, PAOLO GIRARDI*,
ROBERTO PASTRES* AND ANTOINE MANGIN†

*DAIS, Università Ca' Foscari—Venezia** *and ACRI-ST, Sophia-Antipolis*†

The use of water quality indicators is of crucial importance to identify risks to the environment, society and human health. In particular, the Chlorophyll type A (Chl-a) is a shared indicator of trophic status and for monitoring activities it may be useful to discover local dangerous behaviours (for example, the anoxic events). In this paper we consider a comprehensive data set, covering the whole Adriatic Sea, derived from Ocean Colour satellite data, during the period 2002–2012, with the aim of identifying homogeneous areas. Such zonation is becoming extremely relevant for the implementation of European policies, such the Marine Strategy Framework Directive. As an alternative to clustering based on an "average" value over the whole period, we propose a new clustering procedure for the time series. The procedure shares some similarities with the functional data clustering and combines nonparametric quantile regression with an agglomerative clustering algorithm. This approach permits to take into account some features of the time series as nonstationarity in the marginal distribution and the presence of missing data. A small simulation study is also presented for illustrating the relative merits of the procedure.

REFERENCES

ABRAHAM, C., CORNILLON, P. A., MATZNER-LØBER, E. and MOLINARI, N. (2003). Unsupervised curve clustering using B-splines. *Scand. J. Stat.* **30** 581–595. MR2002229

ANTONIADIS, A., BROSSAT, X., CUGLIARI, J. and POGGI, J.-M. (2013). Clustering functional data using wavelets. *Int. J. Wavelets Multiresolut. Inf. Process.* **11** 1350003, 30. MR3038615

BEHRENFELD, M. J. and FALKOWSKI, P. G. (1997). Photosynthetic rates derived from satellite-based chlorophyll concentration. *Limnology and Oceanography* **42** 1–20.

BONDELL, H. D., REICH, B. J. and WANG, H. (2010). Noncrossing quantile regression curve estimation. *Biometrika* **97** 825–838. MR2746154

CAMPBELL, J. W. (1995). The lognormal distribution as a model for bio-optical variability in the sea. *Journal of Geophysical Research*: *Oceans* **100** 13237–13254.

CHENG, K. F. (1983). Nonparametric estimators for percentile regression functions. *Comm. Statist. Theory Methods* **12** 681–692. MR0696815

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

D'ORTENZIO, F. and RIBERA D'ALCALÀ, M. (2009). On the trophic regimes of the Mediterranean Sea: A satellite analysis. *Biogeosciences* **6** 139–148.

Key words and phrases. Functional data clustering, quantile sheet, nonparametric regression, clustering methods, surface water classification, satellite data.

DJAKOVAC, T., DEGOBBIS, D., SUPIĆ, N. and PRECALI, R. (2012). Marked reduction of eutrophication pressure in the northeastern Adriatic in the period 2000–2009. *Estuarine, Coastal and Shelf Science* **115** 25–32.

EILERS, P. H. C., CURRIE, I. D. and DURBÁN, M. (2006). Fast and compact smoothing on large multidimensional grids. *Comput. Statist. Data Anal.* **50** 61–76. MR2196222

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

EILERS, P. H. C., GAMPE, J., MARX, B. D. and RAU, R. (2008). Modulation models for seasonal time series and incidence tables. *Stat. Med.* **27** 3430–3441. MR2523924

FRÜHWIRTH-SCHNATTER, S. and KAUFMANN, S. (2008). Model-based clustering of multiple time series. *J. Bus. Econom. Statist.* **26** 78–89. MR2422063

GIANI, M., DJAKOVAC, T., DEGOBBIS, D., COZZI, S., SOLIDORO, C. and UMANI, S. F. (2012). Recent changes in the marine ecosystems of the northern Adriatic Sea. *Estuarine, Coastal and Shelf Science* **115** 1–13.

GIRALDO, R., DELICADO, P. and MATEU, J. (2012). Hierarchical clustering of spatially correlated functional data. *Stat. Neerl.* **66** 403–421. MR2983302

HAGGARTY, R. A., MILLER, C. A. and SCOTT, E. M. (2015). Spatially weighted functional clustering of river network data. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **64** 491–506. MR3325460

HAGGARTY, R. A., MILLER, C. A., SCOTT, E. M., WYLLIE, F. and SMITH, M. (2012). Functional clustering of water quality data in Scotland. *Environmetrics* **23** 685–695. MR3019060

HE, X. (1997). Quantile curves without crossing. *Amer. Statist.* **51** 186–192.

HENDERSON, B. (2006). Exploring between site differences in water quality trends: A functional data analysis approach. *Environmetrics* **17** 65–80. MR2222034

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

HUNTER, D. R. and LANGE, K. (2000). Quantile regression via an MM algorithm. *J. Comput. Graph. Statist.* **9** 60–77. MR1819866

HUOT, Y., BABIN, M., BRUYANT, F., GROB, C., TWARDOWSKI, M. S. and CLAUSTRE, H. (2007). Does chlorophyll a provide the best index of phytoplankton biomass for primary productivity studies? *Biogeosciences Discussions* **4** 707–745.

JACQUES, J. and PREDA, C. (2014). Functional data clustering: A survey. *Adv. Data Anal. Classif.* **8** 231–255. MR3253859

JAMES, G. M. and SUGAR, C. A. (2003). Clustering for sparsely sampled functional data. *J. Amer. Statist. Assoc.* **98** 397–408. MR1995716

JIANG, H. and SERBAN, N. (2012). Clustering random curves under spatial interdependence with application to service accessibility. *Technometrics* **54** 108–119. MR2929427

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

KOENKER, R. (2005). *Quantile Regression. Econometric Society Monographs* **38**. Cambridge Univ. Press, Cambridge. MR2268657

KOENKER, R., NG, P. and PORTNOY, S. (1994). Quantile smoothing splines. *Biometrika* **81** 673–680. MR1326417

LIAO, T. W. (2005). Clustering of time series data—A survey. *Pattern Recognition* **38** 1857–1874.

MARINI, M., GRILLI, F., GUARNIERI, A., JONES, B. H., KLAJIC, Z., PINARDI, N. and SANXHAKU, M. (2010). Is the southeastern Adriatic Sea coastal strip an eutrophic area? *Estuarine, Coastal and Shelf Science* **88** 395–406.

MARITORENA, S., D'ANDON, O. H. F., MANGIN, A. and SIEGEL, D. A. (2010). Merged satellite ocean color data products using a bio-optical model: Characteristics, benefits and issues. *Remote Sensing of Environment* **114** 1791–1804.

MÉLIN, F., VANTREPOTTE, V., CLERICI, M., D'ALIMONTE, D., ZIBORDI, G., BERTHON, J.-F. and CANUTI, E. (2011). Multi-sensor satellite time series of optical properties and chlorophyll-a concentration in the Adriatic Sea. *Progress in Oceanography* **91** 229–244.

NIETO-BARAJAS, L. E. and CONTRERAS-CRISTÁN, A. (2014). A Bayesian nonparametric approach for time series clustering. *Bayesian Anal*. **9** 147–169. MR3188303

PASTRES, R., PASTORE, A. and TONELLATO, S. F. (2011). Looking for similar patterns among monitoring stations. Venice Lagoon application. *Environmetrics* **22** 712–724. MR2843138

PETITJEAN, F., INGLADA, J. and GANÇARSKI, P. (2012). Satellite image time series analysis under time warping. *Geoscience and Remote Sensing*, *IEEE Transactions on* **50** 3081–3095.

PICCOLO, D. (1990). A distance measure for classifying ARMA models. *J. Time Series Anal*. **2** 153–163.

R CORE TEAM (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

RAMOS, E., JUANES, J. A., GALVÁN, C., NETO, J. M., MELO, R., PEDERSEN, A., SCANLAN, C., WILKES, R., VAN DEN BERGH, E., BLOMQVIST, M., KARUP, H. P., HEIBER, W., REITSMA, J. M., XIMENES, M. C., SILIÓ, A., MÉNDEZ, F. and GONZÁLEZ, B. (2012). Coastal waters classification based on physical attributes along the NE Atlantic region. An approach for rocky macroalgae potential distribution. *Estuarine*, *Coastal and Shelf Science* **112** 105–114.

RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. Springer, New York. MR2168993

REICH, B. J. (2012). Spatiotemporal quantile regression for detecting distributional changes in environmental processes. *J. R. Stat. Soc. Ser. C. Appl. Stat*. **61** 535–553. MR2960737

SCHLOSSMACHER, E. J. (1973). An iterative technique for absolute deviations curve fitting. *J. Amer. Statist. Assoc*. **68** 857–859.

SCHNABEL, S. K. and EILERS, P. H. C. (2013). Simultaneous estimation of quantile curves using quantile sheets. *AStA Adv. Stat. Anal*. **97** 77–87. MR3030781

TIBSHIRANI, R., WALTHER, G. and HASTIE, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **63** 411–423. MR1841503

WANG, X., SMITH, K. and HYNDMAN, R. (2006). Characteristic-based clustering for time series data. *Data Min. Knowl. Discov.* **13** 335–364. MR2252381

YODER, J. A., MCCLAIN, C. R., FELDMAN, G. C. and ESAIAS, W. E. (1993). Annual cycles of phytoplankton chlorophyll concentrations in the global ocean: A satellite view. *Global Biogeochemical Cycles* **7** 181–193.

YUAN, M. (2006). GACV for quantile smoothing splines. *Comput. Statist. Data Anal*. **50** 813–829. MR2207010

The Annals of Applied Statistics
2016, Vol. 10, No. 2, 989–1018
DOI: 10.1214/16-AOAS924
© Institute of Mathematical Statistics, 2016

ASYMMETRIC CONDITIONAL CORRELATIONS IN STOCK RETURNS

BY HUI JIANG*, PATRICK W. SAART† AND YINGCUN XIA*,‡,[1]

National University of Singapore, Newcastle University† and University of Electronic Science and Technology of China‡*

Modeling and estimation of correlation coefficients is a fundamental step in risk management, especially with the aftermath of the financial crisis in 2008, which challenged the traditional measuring of dependence in the financial market. Because of the serial dependence and small signal-to-noise ratio, patterns of the dependence in the data cannot be easily detected and modeled. This paper introduces a common factor analysis into the conditional correlation coefficients to extract the features of dependence. While statistical properties are thoroughly derived, extensive empirical analysis provides us with common patterns for the conditional correlation coefficients that give new insight into a number of important questions in financial data, especially the asymmetry of cross-correlations and the factors that drive the cross-correlations.

REFERENCES

AIELLI, G. P. (2013). Dynamic conditional correlation: On properties and estimation. *J. Bus. Econom. Statist.* **31** 282–299. MR3173682

AMIRA, K., TAAMOUTI, A. and TSAFACK, G. (2011). What drives international equity correlations? Volatility or market direction? *J. Int. Money Financ.* **30** 1234–1263.

ANDERSEN, T. G., BOLLERSLEV, T., DIEBOLD, F. X. and LABYS, P. (2001). The distribution of realized exchange rate volatility. *J. Amer. Statist. Assoc.* **96** 42–55. MR1952727

ANG, A. and BEKAERT, G. (2002). International asset allocation with regime shifts. *Rev. Financ. Stud.* **15** 1137–1187.

ANG, A. and CHEN, J. (2002). Asymmetric correlations of equity portfolios. *J. Financ. Econ.* **63** 443–494.

ASLANIDIS, N. and CASAS, I. (2013). Noparametric correlation models for portfolio allocation. *J. Bank. Financ.* **37** 2268–2283.

BAI, J. and NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221. MR1926259

BOLLERSLEV, T. (1990). Modeling the coherence in short-run nominal exchange rates: A multivariate generalized ARCH model. *Rev. Econ. Stat.* **72** 498–505.

BONEVA, L., LINTON, O. and VOGT, M. (2015). A semiparametric model for heterogeneous panel data with fixed effects. *J. Econometrics* **188** 327–345. MR3383213

CHESNAY, F. and JONDEAU, E. (2001). Does correlation between stock returns really increase during turbulent periods? *Economic Notes by Banca Monte dei Paschi di Siena SpA* **30** 53–80.

ENGLE, R. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *J. Bus. Econom. Statist.* **20** 339–350. MR1939905

Key words and phrases. Conditional cross-correlation coefficient, kernel smoothing, reduced rank model, semiparametric models.

GLOSTEN, L. R., JAGANNATHAN, R. and RUNKLE, D. E. (1993). On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. *J. Finance* **48** 1779–1801.

ENGLE, R. F., NG, V. K. and ROTHSCHILD, M. (1990). Asset pricing with a factor-ARCH covariance structure: Empirical estimates for treasury bills. *J. Econometrics* **45** 213–237.

ERB, C. B., HARVEY, C. R. and VISKANTA, T. E. (1994). Forecasting international equity correlations. *Financ. Anal. J.* **50** 32–45.

FAN, J. and YAO, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* **85** 645–660. MR1665822

HÄRDLE, W. and MARRON, J. S. (1990). Semiparametric comparison of regression curves. *Ann. Statist.* **18** 63–89. MR1041386

JAMES, G. M., HASTIE, T. J. and SUGAR, C. A. (2000). Principal component models for sparse functional data. *Biometrika* **87** 587–602. MR1789811

JIANG, H., SAART, P. W. and XIA Y. (2016). Supplement to "Asymmetric conditional correlations in stock returns." DOI:10.1214/16-AOAS924SUPP.

KAPLANIS, E. (1988). Stability and forecasting of the comovement measures of international stock market returns. *J. Int. Money Financ.* **7** 63–75.

KRING, S., RACHEV, S. T., HOCHSTOTTER, M. and FABOZZI, F. J. (2007). Composed and factor composed multivariate GARCH models. Technical report, University of Karlsruhe.

LANNE, M. and SAIKKONEN, P. (2007). A multivariate generalized orthogonal factor GARCH model. *J. Bus. Econom. Statist.* **25** 61–75. MR2338871

LI, Y., WANG, N. and CARROLL, R. J. (2013). Selecting the number of principal components in functional data. *J. Amer. Statist. Assoc.* **108** 1284–1294. MR3174708

LONGIN, F. and SOLNIK, B. (1995). Is the correlation in international equity returns constant: 1960–1990? *J. Int. Money Financ.* **14** 3–26.

LONGIN, F. and SOLNIK, B. (2001). Extreme correlations in international equity markets. *J. Finance* **56** 649–676.

MUNK, A. and DETTE, H. (1998). Nonparametric comparison of several regression functions: Exact and asymptotic theory. *Ann. Statist.* **26** 2339–2368. MR1700235

PELLETIER, D. (2006). Regime switching for dynamic correlations. *J. Econometrics* **131** 445–473. MR2276007

RAMCHAND, L. and SUSMEL, R. (1998). Volatility and cross correlation across major stock markets. *J. Empir. Finance* **5** 397–416.

SANTOS, A. A. P. and MOURA, G. V. (2014). Dynamic factor multivariate GARCH model. *Comput. Statist. Data Anal.* **76** 606–617. MR3209460

SHEPPARD, K. and XU, W. (2014). Factor high-frequency based volatility (heavy) models. Working paper, Univ. Oxford.

SILVENNOINEN, A. and TERÄSVIRTA, T. (2015). Modeling conditional correlations of asset returns: A smooth transition approach. *Econometric Rev.* **34** 174–197. MR3268917

TSE, Y. K. and TSUI, A. K. C. (2002). A multivariate generalized autoregressive conditional heteroscedasticity model with time-varying correlations. *J. Bus. Econom. Statist.* **20** 351–362. MR1939906

VRONTOS, I. D., DELLAPORTAS, P. and POLITIS, D. N. (2003). A full-factor multivariate GARCH model. *Econom. J.* **6** 312–334. MR2028238

The Annals of Applied Statistics
2016, Vol. 10, No. 2, 1019–1040
DOI: 10.1214/16-AOAS928
© Institute of Mathematical Statistics, 2016

REGRESSION ANALYSIS FOR MICROBIOME COMPOSITIONAL DATA

BY PIXU SHI[*,1], ANRU ZHANG[†] AND HONGZHE LI[*1]

University of Pennsylvania[*] and University of Wisconsin-Madison[†]

One important problem in microbiome analysis is to identify the bacterial taxa that are associated with a response, where the microbiome data are summarized as the composition of the bacterial taxa at different taxonomic levels. This paper considers regression analysis with such compositional data as covariates. In order to satisfy the subcompositional coherence of the results, linear models with a set of linear constraints on the regression coefficients are introduced. Such models allow regression analysis for subcompositions and include the log-contrast model for compositional covariates as a special case. A penalized estimation procedure for estimating the regression coefficients and for selecting variables under the linear constraints is developed. A method is also proposed to obtain debiased estimates of the regression coefficients that are asymptotically unbiased and have a joint asymptotic multivariate normal distribution. This provides valid confidence intervals of the regression coefficients and can be used to obtain the p-values. Simulation results show the validity of the confidence intervals and smaller variances of the debiased estimates when the linear constraints are imposed. The proposed methods are applied to a gut microbiome data set and identify four bacterial genera that are associated with the body mass index after adjusting for the total fat and caloric intakes.

REFERENCES

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

AITCHISON, J. (2003). *The Statistical Analysis of Compositional Data*. Blackburn Press, Cadwell, NJ.

AITCHISON, J. and BACON-SHONE, J. (1984). Log contrast models for experiments with mixtures. *Biometrika* **71** 323–330.

BERTSEKAS, D. P. (1996). *Constrained Optimization and Lagrange Multiplier Methods*. Athena Scientific, Belmont.

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

CORNELL, J. A. (2002). *Experiments with Mixtures*: *Designs*, *Models*, *and the Analysis of Mixture Data*, 3rd ed. Wiley, New York. MR1882356

EFRON, B. (2014). Estimation and accuracy after model selection. *J. Amer. Statist. Assoc.* **109** 991–1007. MR3265671

GRANT, M. and BOYD, S. (2013). CVX: Matlab software for disciplined convex programming, version 2.0 beta. Technical report. Available at http://cvxr.com/cvx.

Key words and phrases. Compositional coherence, coordinate descent method of multipliers, high dimension, log-contrast model, model selection, regularization.

HUSON, D. H., AUCH, A. F., QI, J. and SCHUSTER, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Res*. **17** 377–386.

JAMES, G. M., PAULSON, C. and RUSMEVICHIENTONG, P. (2015). Penalized and constrained regression. Unpublished manuscript.

JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res*. **15** 2869–2909. MR3277152

KURTZ, Z. D., MÜLLER, C. L., MIRALDI, E. R., LITTMAN, D. R., BLASER, M. J. and BONNEAU, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS Computational Biolology* **11** e1004226.

LAM, Y. Y., HA, C. W. Y., CAMPBELL, C. R., MITCHELL, A. J., DINUDOM, A., OSCARSSON, J., COOK, D. I., HUNT, N. H., CATERSON, I. D., HOLMES, A. J. and STORLIEN, L. H. (2012). Increased gut permeability and microbiota change associate with mesenteric fat inflammation and metabolic dysfunction in diet-induced obese mice. *PLoS ONE* **7** e34233.

LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist*. **44** 907–927. MR3485948

LEY, R. E., BÄCKHED, F., TURNBAUGH, P., LOZUPONE, C. A., KNIGHT, R. D. and GORDON, J. I. (2005). Obesity alters gut microbial ecology. *Proc. Natl. Acad. Sci. USA* **102** 11070–11075.

LEY, R. E., TURNBAUGH, P. J., KLEIN, S. and GORDON, J. I. (2006). Microbial ecology: Human gut microbes associated with obesity. *Nature* **444** 1022–1023.

LIN, W., SHI, P., FENG, R. and LI, H. (2014). Variable selection in regression with compositional covariates. *Biometrika* **101** 785–797. MR3286917

MANICHANH, C., BORRUEL, N., CASELLAS, F. and GUARNER, F. (2012). The gut microbiota in IBD. *Nat. Rev. Gastroenterol. Hepatol*. **9** 599–608.

QIN, J., LI, R., RAES, J., ARUMUGAM, M., BURGDORF, K. S., MANICHANH, C., NIELSEN, T., PONS, N., LEVENEZ, F., YAMADA, T. et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464** 59–65.

QIN, J., LI, Y., CAI, Z., LI, S., ZHU, J., ZHANG, F., LIANG, S., ZHANG, W., GUAN, Y., SHEN, D. et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490** 55–60.

SEGATA, N., WALDRON, L., BALLARINI, A., NARASIMHAN, V., JOUSSON, O. and HUTTENHOWER, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **9** 811–814.

SHI, P., ZHANG, A. and LI (2016). Supplement to "Regression analysis for microbiome compositional data." DOI:10.1214/16-AOAS928SUPP.

SNEE, R. D. (1973). Techniques for the analysis of mixture data. *Technometrics* **15** 517–528.

SUN, T. and ZHANG, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898. MR2999166

TURNBAUGH, P. J., LEY, R. E., MAHOWALD, M. A., MAGRINI, V., MARDIS, E. R. and GORDON, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444** 1027–1031.

TURNBAUGH, P. J., LEY, R. E., HAMADY, M., FRASER-LIGGETT, C. M., KNIGHT, R. and GORDON, J. I. (2007). The human microbiome project. *Nature* **449** 804–810.

VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist*. **42** 1166–1202. MR3224285

WALKER, A. W., INCE, J., DUNCAN, S. H., WEBSTER, L. M., HOLTROP, G., ZE, X., BROWN, D., STARES, M. D., SCOTT, P., BERGERAT, A., LOUIS, P., MCINTOSH, F., JOHNSTONE, A. M., LOBLEY, G. E., PARKHILL, J. and FLINT, H. J. (2011). Dominant and diet-responsive groups of bacteria within the human colonic microbiota. *ISME J*. **5** 220–230.

WU, G. D., CHEN, J., HOFFMANN, C., BITTINGER, K., CHEN, Y.-Y., KEILBAUGH, S. A., BEWTRA, M., KNIGHTS, D., WALTERS, W. A., KNIGHT, R., SINHA, R., GILROY, E., GUPTA, K., BALDASSANO, R., NESSEL, L., LI, H., BUSHMAN, F. D. and LEWIS, J. D. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334** 105–108.

ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940

The Annals of Applied Statistics
2016, Vol. 10, No. 2, 1041–1062
DOI: 10.1214/16-AOAS929
© Institute of Mathematical Statistics, 2016

LATENT SPATIAL MODELS AND SAMPLING DESIGN FOR LANDSCAPE GENETICS[1]

BY EPHRAIM M. HANKS[*], MEVIN B. HOOTEN[†,‡], STEVEN T. KNICK[§], SARA J. OYLER-MCCANCE[¶], JENNIFER A. FIKE[¶], TODD B. CROSS[‖,**] AND MICHAEL K. SCHWARTZ[‖]

Pennsylvania State University[*], U.S. Geological Survey, Colorado Cooperative Fish and Wildlife Research Unit[†], Colorado State University[‡], U.S. Geological Survey, Forest and Rangeland Ecosystem Science Center[§], U.S. Geological Survey, Fort Collins Science Center[¶], U.S. Forest Service, Rocky Mountain Research Station[‖] and University of Montana[**]

We propose a spatially-explicit approach for modeling genetic variation across space and illustrate how this approach can be used to optimize spatial prediction and sampling design for landscape genetic data. We propose a multinomial data model for categorical microsatellite allele data commonly used in landscape genetic studies, and introduce a latent spatial random effect to allow for spatial correlation between genetic observations. We illustrate how modern dimension reduction approaches to spatial statistics can allow for efficient computation in landscape genetic statistical models covering large spatial domains. We apply our approach to propose a retrospective spatial sampling design for greater sage-grouse (*Centrocercus urophasianus*) population genetics in the western United States.

REFERENCES

CUSHMAN, S. A. and LANDGUTH, E. L. (2010). Scale dependent inference in landscape genetics. *Landsc. Ecol.* **25** 967–979.

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

BERLINER, L. M., WIKLE, C. K. and CRESSIE, N. (2000). Long-lead prediction of Pacific SSTs via Bayesian dynamic modeling. *J. Climate* **13** 3953–3968.

BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **36** 192–236. With discussion by D. R. Cox, A. G. Hawkes, P. Clifford, P. Whittle, K. Ord, R. Mead, J. M. Hammersley, and M. S. Bartlett and with a reply by the author. MR0373208

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

CALDER, C. A. and CRESSIE, N. (2007). Some topics in convolution-based spatial modeling. In *Proceedings of the* 56*th Session of the International Statistics Institute* 22–29. Instituto Nacional de Estatistica, Lisbon, Portugal.

CONNELLY, J. W., HAGEN, C. A. and SCHROEDER, M. A. (2011). Characteristics and dynamics of greater sage-grouse populations. In *Greater Sage-Grouse*: *Ecology and Conservation of a*

Key words and phrases. Landscape genetics, sage grouse, optimal sampling.

Landscape Species and Its Habitats. *Studies in Avian Biology* **38** 53–67. Univ. California Press, Oakland.

CONNELLY, J. W., SCHROEDER, M. A., SANDS, A. R. and BRAUN, C. E. (2000). Guidelines to manage sage-grouse populations and their habitats. *Wildl. Soc. Bull.* **28** 967–985.

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

CROOKS, K. R. and SANJAYAN, M. A. (2006). *Connectivity Conservation*. Cambridge Univ. Press, Cambridge.

CUSHMAN, S. A., MCKELVEY, K. S., HAYDEN, J. and SCHWARTZ, M. K. (2006). Gene flow in complex landscapes: Testing multiple hypotheses with causal modeling. *Amer. Nat.* **168** 486–499.

DALKE, P. D., PYRAH, D. B., STANTON, D. C., CRAWFORD, J. E. and SCHLATTERER, E. F. (1963). Ecology, productivity, and management of sage-grouse in Idaho. *J. Wildl. Manag.* **27** 811–841.

DIGGLE, P. J. and RIBEIRO, P. J. JR. (2007). *Model-Based Geostatistics*. Springer, New York. MR2293378

DURAND, E., JAY, F., GAGGIOTTI, O. E. and FRANÇOIS, O. (2009). Spatial inference of admixture proportions and secondary contact zones. *Mol. Biol. Evol.* **26** 1963–1973.

FLEGAL, J. M., HARAN, M. and JONES, G. L. (2008). Markov chain Monte Carlo: Can we trust the third significant figure? *Statist. Sci.* **23** 250–260. MR2516823

GARTON, E. O., CONNELLY, J. W., HORNE, J. S., HAGEN, C. A., MOSER, A. and SCHROEDER, M. A. (2011). Greater sage-grouse population dynamics and probability of persistence. In *Greater Sage-Grouse*: *Ecology and Conservation of a Landscape Species and Its Habitats*. *Studies in Avian Biology* **38** 293–381. Univ. California Press, Oakland.

GENTLE, J. E. (2007). *Matrix Algebra*: *Theory*, *Computations*, *and Applications in Statistics*. Springer, New York. MR2337395

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

GUILLOT, G., ESTOUP, A., MORTIER, F. and COSSON, J. F. (2005). A spatial statistical model for landscape genetics. *Genetics* **170** 1261–1280.

GUILLOT, G., LEBLOIS, R., COULON, A. and FRANTZ, A. C. (2009). Statistical methods in spatial genetics. *Mol. Ecol.* **18** 4734–4756.

HANKS, E. M. and HOOTEN, M. B. (2013). Circuit theory and model-based inference for landscape connectivity. *J. Amer. Statist. Assoc.* **108** 22–33. MR3174600

HANKS, E. M., HOOTEN, M. B., KNICK, S. T., OYLER-MCCANCE, S. J., FIKE, J. A., CROSS, T. B. and SCHWARTZ, M. K. (2016). Supplement to "Latent spatial models and sampling design for landscape genetics." DOI:10.1214/00-AOAS929SUPP.

HARVILLE, D. A. (2008). *Matrix Algebra from a Statistician's Perspective*. Springer, New York.

HOOTEN, M. B., WIKLE, C. K., SHERIFF, S. L. and RUSHIN, J. W. (2009). Optimal spatio-temporal hybrid sampling designs for ecological monitoring. *J. Veg. Sci.* **20** 639–649.

KLEIN, D. J. and RANDIĆ, M. (1993). Resistance distance. *J. Math. Chem.* **12** 81–95. Applied graph theory and discrete mathematics in chemistry (Saskatoon, SK, 1991). MR1219566

KNICK, S. T. and HANSER, S. E. (2011). Connecting pattern and process in greater sage-grouse populations and sagebrush landscapes. In *Greater Sage-Grouse*: *Ecology and Conservation of a Landscape Species and Its Habitats*. *Studies in Avian Biology* **38** 383–406. Univ. California Press, Oakland.

MARSAGLIA, G. (1963). Expressing the normal distribution with covariance matrix $A + B$ in terms of one with covariance matrix A. *Biometrika* **50** 535–538. MR0181061

MCKELVEY and SCHWARTZ (2005). dropout: A program to identify problem loci and samples for noninvasive genetic samples in a capture-mark-recapture framework. *Molecular Ecology Notes* **5** 716–718.

MCRAE, B. H. (2006). Isolation by resistance. *Evolution* **60** 1551–1561.

McRAE, B. H. and BEIER, P. (2007). Circuit theory predicts gene flow in plant and animal populations. *Proc. Natl. Acad. Sci. USA* **104** 19885–19890.

McRAE, B. H., DICKSON, B. G., KEITT, T. H. and SHAH, V. B. (2008). Using circuit theory to model connectivity in ecology, evolution, and conservation. *Ecology* **89** 2712–2724.

PATTERSON, R. L. (1952). *The Sage-Grouse in Wyoming*. Sage Books, Los Angeles.

ROYLE, J. (1998). An algorithm for the construction of spatial coverage designs with implementation in S-PLUS. *Comput. Geosci.* **24** 479–488.

SCHMIDT, A. M. and O'HAGAN, A. (2003). Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 743–758. MR1998632

SCHROEDER, M. A., ALDRIDGE, C. L., APA, A. D., BOHNE, J. R., BRAUN, C. E., BUNNELL, S. D., CONNELLY, J. W., DEIBERT, P. A., GARDNER, S. C., HILLIARD, M. A. et al. (2004). Distribution of sage-grouse in North America. *Condor* **106** 363–376.

SCHWARTZ, M. K., MILLS, L. S., ORTEGA, Y., RUGGIERO, L. F. and ALLENDORF, F. W. (2003). Landscape location affects genetic variation of Canada lynx (*Lynx canadensis*). *Mol. Ecol.* **12** 1807–1816.

SELANDER, R. K. (1970). Behavior and genetic variation in natural populations. *Am. Zool.* **10** 53–66.

SLATKIN, M. (1987). Gene flow and the geographic structure of natural populations. *Science* **236** 787–792.

SMITH, J. T., FLAKE, L. D., HIGGINS, K. F., KOBRIGER, G. D. and HOMER, C. G. (2005). Evaluating lek occupancy of greater sage-grouse in relation to landscape cultivation in the Dakotas. *West. N. Am. Nat.* **65** 310–320.

SMOUSE, P. E. and PEAKALL, R. (1999). Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity* **82** 561–573.

SOKAL, R. R., ODEN, N. L. and WILSON, C. (1991). Genetic evidence for the spread of agriculture in Europe by demic diffusion. *Nature* **351** 143–145.

SPEAR, S. F., BALKENHOL, N., FORTIN, M. J., McRAE, B. H. and SCRIBNER, K. (2010). Use of resistance surfaces for landscape genetic studies: Considerations for parameterization and analysis. *Mol. Ecol.* **19** 3576–3591.

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

TAYLOR, P. D., FAHRIG, L., HENEIN, K. and MERRIAM, G. (1993). Connectivity is a vital element of landscape structure. *Oikos* **68** 571–573.

WIKLE, C. K. and CRESSIE, N. (1999). A dimension-reduced approach to space-time Kalman filtering. *Biometrika* **86** 815–829. MR1741979

WIKLE, C. K. and ROYLE, J. A. (2005). Dynamic design of ecological monitoring networks for non-Gaussian spatio-temporal data. *Environmetrics* **16** 507–522. MR2147540

WILEY, R. H. (1973). Territoriality and non-random mating in sage-grouse, *Centrocercus urophasianus*. *Anim. Behav. Monogr.* **6** 87–169.

ZIMMERMAN, D. L. (2006). Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. *Environmetrics* **17** 635–652. MR2247174

The Annals of Applied Statistics
2016, Vol. 10, No. 2, 1063–1085
DOI: 10.1214/16-AOAS930
© Institute of Mathematical Statistics, 2016

AN IMPUTATION APPROACH FOR HANDLING MIXED-MODE SURVEYS

BY SEUNGHWAN PARK[1], JAE KWANG KIM[2] AND SANGUN PARK[3]

Seoul National University, Iowa State University and Yonsei University

Mixed-mode surveys are becoming more popular recently because of their convenience for users, but different mode effects can complicate the comparability of the survey results. Motivated by the Private Education Expenditure Survey (PEES) of Korea, we propose a novel application of fractional imputation to handle mixed-mode survey data. The proposed method is applied to create imputed values of the unobserved counterfactual outcome variables in the mixed-mode surveys. The proposed method is directly applicable when the choice of survey mode is self-selected. Variance estimation using Taylor linearization is developed. Results from a limited simulation study are also presented.

REFERENCES

AMEMIYA, T. (1973). Regression analysis when the dependent variable is truncated normal. *Econometrica* **41** 997–1016. MR0440773

BIEMER, P. P. (2001). Nonresponse bias and measurement bias in a comparison of face to face and telephone interviewing. *J. Off. Stat.* **17** 295–320.

BOOTH, J. G. and HOBERT, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **61** 265–285.

BUELENS, B. and VAN DEN BRAKEL, J. (2013). On the necessity to include personal interviewing in mixed-mode surveys. *Survey Practice* **3** 1–7.

BURGETTE, L. F. and REITER, J. P. (2012). Nonparametric Bayesian multiple imputation for missing data due to mid-study switching of measurement methods. *J. Amer. Statist. Assoc.* **107** 439–449. MR2980056

CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A. and CRAINICEANU, C. M. (2006). *Measurement Error in Nonlinear Models*: *A Modern Perspective*, 2nd ed. *Monographs on Statistics and Applied Probability* **105**. Chapman & Hall/CRC, Boca Raton, FL. MR2243417

CHEN, M.-H., SHAO, Q.-M. and IBRAHIM, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer, New York. MR1742311

DE LEEUW, E. D. (2005). To mix or not to mix data collection modes in surveys. *J. Off. Stat.* **21** 233–255.

DILLMAN, D. A. and CHRISTIAN, L. M. (2003). Survey mode as a source of instability in responses across surveys. *Field Methods* **15** 1–22.

DILLMAN, D., PHELPS, G., TORTORA, R., SWIFT, K., KOHRELL, J., BERCK, J. and MESSER, B. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response and the Internet. *Soc. Sci. Res.* **39** 1–18.

DURRANT, G. B. and SKINNER, C. (2006). Using missing data methods to correct for measurement error in a distribution function. *Surv. Methodol.* **32** 25–36.

Key words and phrases. Counterfactual outcome, fractional imputation, measurement error model, missing data, survey sampling.

KIM, J. K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika* **98** 119–132. MR2804214

KLAUSCH, T., SCHOUTEN, B. and HOX, J. (2014). The use of within-subject experiments for estimating measurement effects in mixed-mode surveys. Statistics Netherlands Discussion Paper, 201406.

KOLENIKOV, S. and KENNEDY, C. (2014). Evaluating three approaches to statistically adjust for mode effects. *Journal of Survey Statistics and Methodology* **2** 126–158.

KROSNICK, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Appl. Cogn. Psychol.* **5** 213–236.

KROSNICK, J. A. (1999). Survey research. *Annu. Rev. Psychol.* **50** 537–567.

MORGAN, S. L. and WINSHIP, C. (2007). *Counterfactuals and Causal Inference*: *Methods and Principles for Social Research*. Cambridge Univ. Press, Cambridge.

POWERS, J. R., MISHRA, G. and YOUNG, A. F. (2005). Differences in mail and telephone responses to self-rated health: Use of multiple imputation in correcting for response bias. *Aust. N. Z. J. Public Health* **29** 149–154.

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

SCHNEDLER, W. (2005). Likelihood estimation for censored random vectors. *Econometric Rev.* **24** 195–217. MR2190316

VANNIEUWENHUYZE, J. T., LOOSVELDT, G. and MOLENBERGHS, G. (2010). A method for evaluating mode effects in mixed-mode surveys. *Public Opin. Q.* **74** 1027–1045.

VOOGT, R. and SARIS, W. (2005). Mixed mode designs: Finding the balance between nonresponse bias and mode effects. *J. Off. Stat.* **21** 367–387.

WEI, G. C. and TANNER, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the Poor Man's data augmentation algorithms. *J. Amer. Statist. Assoc.* **85** 699–704.

The Annals of Applied Statistics
2016, Vol. 10, No. 2, 1086–1106
DOI: 10.1214/16-AOAS932
© Institute of Mathematical Statistics, 2016

HOW STRONG IS STRONG ENOUGH? STRENGTHENING INSTRUMENTS THROUGH MATCHING AND WEAK INSTRUMENT TESTS

BY LUKE KEELE AND JASON W. MORGAN

Georgetown University and Ohio State University

In a natural experiment, treatment assignments are made through a haphazard process that is thought to be as-if random. In one form of the natural experiment, encouragement to accept treatment rather than treatments themselves are assigned in this haphazard process. This encouragement to accept treatment is often referred to as an instrument. Instruments can be characterized by different levels of strength depending on the amount of encouragement. Weak instruments that provide little encouragement may produce biased inferences, particularly when assignment of the instrument is not strictly randomized. A specialized matching algorithm can be used to strengthen instruments by selecting a subset of matched pairs where encouragement is strongest. We demonstrate how weak instrument tests can guide the matching process to ensure that the instrument has been sufficiently strengthened. Specifically, we combine a matching algorithm for strengthening instruments and weak instrument tests in the context of a study of whether turnout influences party vote share in US elections. It is thought that when turnout is higher, Democratic candidates will receive a higher vote share. Using excess rainfall as an instrument, we hope to observe an instance where unusually wet weather produces lower turnout in an as-if random fashion. Consistent with statistical theory, we find that strengthening the instrument reduces sensitivity to bias from an unobserved confounder.

REFERENCES

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. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *J. Econ. Perspect.* **24** 3–30.

BAIOCCHI, M., SMALL, D. S., LORCH, S. and ROSENBAUM, P. R. (2010). Building a stronger instrument in an observational study of perinatal care for premature infants. *J. Amer. Statist. Assoc.* **105** 1285–1296. MR2796550

BAIOCCHI, M., SMALL, D. S., YANG, L., POLSKY, D. and GROENEVELD, P. W. (2012). Near/far matching: A study design approach to instrumental variables. *Health Serv. Outcomes Res. Methodol.* **12** 237–253.

BOUND, J., JAEGER, D. A. and BAKER, R. M. (1995). Problems with intrustmental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *J. Amer. Statist. Assoc.* **90** 443–450.

BRUMBACK, B. A., HERNÁN, M. A., HANEUSE, S. J. and ROBINS, J. M. (2004). Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Stat. Med.* **23** 749–767.

Key words and phrases. Causal inference, instrumental variables, matching, weak instruments.

CHAMBERLAIN, G. and IMBENS, G. (2004). Random effects estimators with many instrumental variables. *Econometrica* **72** 295–306. MR2031020

CHAO, J. C. and SWANSON, N. R. (2005). Consistent estimation with a large number of weak instruments. *Econometrica* **73** 1673–1692. MR2156676

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

DUNNING, T. (2012). *Natural Experiments in the Social Sciences*: *A Design-Based Approach*. Cambridge Univ. Press, Cambridge, UK.

GOMEZ, B. T., HANSFORD, T. G. and KRAUSE, G. A. (2007). The republicans should pray for rain: Weather turnout, and voting in U.S. presidential elections. *J. Polit.* **69** 649–663.

HANSFORD, T. G. and GOMEZ, B. T. (2010). Estimating the electoral effects of voter turnout. *Am. Polit. Sci. Rev.* **104** 268–288.

HODGES, J. L. JR. and LEHMANN, E. L. (1963). Estimates of location based on rank tests. *Ann. Math. Stat.* **34** 598–611. MR0152070

HOLLAND, P. W. (1988). Causal inference, path analysis, and recursive structural equation models. *Sociol. Method.* **18** 449–484.

IMBENS, G. W. and ANGRIST, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62** 467–476.

IMBENS, G. W. and ROSENBAUM, P. R. (2005). Robust, accurate confidence intervals with a weak instrument: Quarter of birth and education. *J. Roy. Statist. Soc. Ser. A* **168** 109–126. MR2113230

KEELE, L. J. and MINOZZI, W. (2012). How much is Minnesota like Wisconsin? Assumptions and counterfactuals in causal inference with observational data. *Polit. Anal.* **21** 193–216.

KEELE, L., TITIUNIK, R. and ZUBIZARRETA, J. R. (2015). Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout. *J. Roy. Statist. Soc. Ser. A* **178** 223–239. MR3291769

LIN, D. Y., PSATY, B. M. and KRONMAL, R. A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics* **54** 948–963.

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.

LORCH, S. A., BAIOCCHI, M., AHLBERG, C. E. and SMALL, D. S. (2012). The differential impact of delivery hospital on the outcomes of premature infants. *Pediatrics* **130** 270–278.

LU, B., ZANUTTO, E., HORNIK, R. and ROSENBAUM, P. R. (2001). Matching with doses in an observational study of a media campaign against drug abuse. *J. Amer. Statist. Assoc.* **96** 1245–1253. MR1973668

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

NAGLER, J. (1991). The effect of registration laws and education on united-states voter turnout. *Am. Polit. Sci. Rev.* **85** 1393–1405.

ROBINS, J. M., ROTNITZKY, A. and SCHARFSTEIN, D. O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials* (*Minneapolis, MN*, 1997) (E. Halloran and D. Berry, eds.). *IMA Vol. Math. Appl.* **116** 1–94. Springer, New York. MR1731681

ROSENBAUM, P. R. (1984). The consequences of adjusting for a concomitant variable that has been affected by the treatment. *J. Roy. Statist. Soc. Ser. A* **147** 656–666.

ROSENBAUM, P. R. (1996). Identification of causal effects using instrumental variables: Comment. *J. Amer. Statist. Assoc.* **91** 465–468.

ROSENBAUM, P. R. (1999). Using quantile averages in matched observational studies. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **48** 63–78.

ROSENBAUM, P. R. (2002a). *Observational Studies*, 2nd ed. Springer, New York. MR1899138

ROSENBAUM, P. R. (2002b). Covariance adjustment in randomized experiments and observational studies. *Statist. Sci.* **17** 286–327. MR1962487

ROSENBAUM, P. R. (2005). Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *Amer. Statist.* **59** 147–152. MR2133562

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

ROSENZWEIG, M. R. and WOLPIN, K. I. (2000). Natural "natural experiments" in economics. *J. Econ. Lit.* **38** 827–874.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **6** 688–701.

RUBIN, D. B. (1986). Which ifs have causal answers. *J. Amer. Statist. Assoc.* **81** 961–962.

SMALL, D. S. (2007). Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *J. Amer. Statist. Assoc.* **102** 1049–1058. MR2411664

SMALL, D. S. and ROSENBAUM, P. R. (2008). War and wages: The strength of instrumental variables and their sensitivity to unobserved biases. *J. Amer. Statist. Assoc.* **103** 924–933. MR2528819

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

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

STAIGER, D. and STOCK, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica* **65** 557–586. MR1445622

STOCK, J. H. and YOGO, M. (2005). Testing for weak instruments in linear IV regression. In *Identification and Inference in Econometric Models*: *Essays in Honor of Thomas J. Rothenberg* (D. W. K. Andrews and J. H. Stock, eds.) Cambridge Univ. Press, Cambridge. MR2229140

WALD, A. (1940). The fitting of straight lines if both variables are subject to error. *Ann. Math. Stat.* **11** 285–300. MR0002739

WOLFINGER, R. E. and ROSENSTONE, S. J. (1980). *Who Votes?* Yale Univ. Press, New Haven.

ZIGLER, C. M., DOMINICI, F. and WANG, Y. (2012). Estimating causal effects of air quality regulations using principal stratification for spatially correlated multivariate intermediate outcomes. *Biostatistics* **13** 289–302.

ZUBIZARRETA, J. R., SMALL, D. S. and ROSENBAUM, P. R. (2014). Isolation in the construction of natural experiments. *Ann. Appl. Stat.* **8** 2096–2121. MR3292490

ZUBIZARRETA, J. R., SMALL, D. S., GOYAL, N. K., LORCH, S. and ROSENBAUM, P. R. (2013). Stronger instruments via integer programming in an observational study of late preterm birth outcomes. *Ann. Appl. Stat.* **7** 25–50. MR3086409

The Annals of Applied Statistics
2016, Vol. 10, No. 2, 1107–1134
DOI: 10.1214/16-AOAS933
© Institute of Mathematical Statistics, 2016

DETECTING ABRUPT CHANGES IN THE SPECTRA OF HIGH-ENERGY ASTROPHYSICAL SOURCES[1]

BY RAYMOND K. W. WONG*, VINAY L. KASHYAP†,2, THOMAS C. M. LEE‡,3 AND DAVID A. VAN DYK§,4

Iowa State University, Harvard Smithsonian Center for Astrophysics*†*, University of California, Davis*‡* and Imperial College London*§*

Variable-intensity astronomical sources are the result of complex and often extreme physical processes. Abrupt changes in source intensity are typically accompanied by equally sudden spectral shifts, that is, sudden changes in the wavelength distribution of the emission. This article develops a method for modeling photon counts collected from observation of such sources. We embed change points into a marked Poisson process, where photon wavelengths are regarded as marks and both the Poisson intensity parameter and the distribution of the marks are allowed to change. To the best of our knowledge, this is the first effort to embed change points into a *marked* Poisson process. Between the change points, the spectrum is modeled nonparametrically using a mixture of a smooth radial basis expansion and a number of local deviations from the smooth term representing spectral emission lines. Because the model is over-parameterized, we employ an ℓ_1 penalty. The tuning parameter in the penalty and the number of change points are determined via the minimum description length principle. Our method is validated via a series of simulation studies and its practical utility is illustrated in the analysis of the ultra-fast rotating yellow giant star known as FK Com.

REFERENCES

AKMAN, V. E. and RAFTERY, A. E. (1986). Asymptotic inference for a change-point Poisson process. *Ann. Statist.* **14** 1583–1590. MR0868320

AUE, A. and LEE, T. C. M. (2011). On image segmentation using information theoretic criteria. *Ann. Statist.* **39** 2912–2935. MR3012396

AUE, A., CHEUNG, R. C. Y., LEE, T. C. M. and ZHONG, M. (2014). Segmented model selection in quantile regression using the minimum description length principle. *J. Amer. Statist. Assoc.* **109** 1241–1256.

BOPP, B. W. and STENCEL, R. E. (1981). The FK comae stars. *Astrophys. J.* **247** L131–L134.

BUHMANN, M. D. (2003). *Radial Basis Functions*: *Theory and Implementations*. *Cambridge Monographs on Applied and Computational Mathematics* **12**. Cambridge Univ. Press, Cambridge. MR1997878

CARLIN, B. P., GELFAND, A. E. and SMITH, A. F. (1992). Hierarchical Bayesian analysis of changepoint problems. *Appl. Stat.* **41** 389–405.

CHAN, H. P. and ZHANG, N. R. (2007). Scan statistics with weighted observations. *J. Amer. Statist. Assoc.* **102** 595–602. MR2370856

Key words and phrases. Astronomy, change point estimation, FK Comae Berenices, marked poisson process, minimum description length principle, semi-parametric modeling, stars, stellar coronae, stellar flare, X-ray astronomy.

CHEN, J. and CHEN, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95** 759–771. MR2443189

CHIB, S. (1998). Estimation and comparison of multiple change-point models. *J. Econometrics* **86** 221–241. MR1649222

COHEN, O., DRAKE, J. J., KASHYAP, V. L., KORHONEN, H., ELSTNER, D. and GOMBOSI, T. I. (2010). Magnetic structure of rapidly rotating FK comae-type coronae. *Astrophys. J.* **719** 299–306.

DAVIS, R. A. and YAU, C. Y. (2013). Consistency of minimum description length model selection for piecewise stationary time series models. *Electron. J. Stat.* **7** 381–411. MR3020426

DRAKE, J. J., CHUNG, S. M., KASHYAP, V., KORHONEN, H., VAN BALLEGOOIJEN, A. and ELSTNER, D. (2008). X-ray spectroscopic signatures of the extended corona of FK comae. *Astrophys. J.* **679** 1522–1530.

ELSTNER, D. and KORHONEN, H. (2005). Flip-flop phenomenon: Observations and theory. Preprint. Available at arXiv:Astro-ph/0501343.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. MR1380810

GRÜNWALD, P. D., MYUNG, I. J. and PITT, M. A. (2005). *Advances in Minimum Description Length*: *Theory and Applications*. MIT press, Cambridge.

GÜDEL, M. (2004). X-ray astronomy of stellar coronae. *Astron. Astrophys. Rev.* **12** 71–237.

KASHYAP, V. L., DRAKE, J. J., GÜDEL, M. and AUDARD, M. (2002). Flare heating in stellar coronae. *Astrophys. J.* **580** 1118–1132.

KORHONEN, H., BERDYUGINA, S. V., HACKMAN, T., DUEMMLER, R., ILYIN, I. V. and TUOMINEN, I. (1999). Study of FK Comae Berenices. I. Surface images for 1994 and 1995. *Astron. Astrophys.* **346** 101–110.

LAI, T. L. and XING, H. (2011). A simple Bayesian approach to multiple change-points. *Statist. Sinica* **21** 539–569. MR2829846

LECLERC, Y. G. (1989). Constructing simple stable descriptions for image partitioning. *Int. J. Comput. Vis.* **3** 73–102.

LEE, T. C. M. (1997). Some models and methods in image segmentation. Ph.D. thesis, Macquarie Univ., Sydney, Australia.

LEE, T. C. M. (1998). Segmenting images corrupted by correlated noise. *IEEE Trans. Pattern Anal. Mach. Intell.* **20** 481–492.

LEE, T. C. M. (2002a). Automatic smoothing for discontinuous regression functions. *Statist. Sinica* **12** 823–842. MR1929966

LEE, T. C. M. (2002b). On algorithms for ordinary least squares regression spline fitting: A comparative study. *J. Stat. Comput. Simul.* **72** 647–663. MR1930486

LEE, H., KASHYAP, V. L., VAN DYK, D. A., CONNORS, A., DRAKE, J. J., IZEM, R., MENG, X. L., MIN, S., PARK, T., RATZLAFF, P., SIEMIGINOWSKA, A. and ZEZAS, A. (2011). Accounting for calibration uncertainties in X-ray analysis: Effective areas in spectral fitting. *Astrophys. J.* **731** 126–144.

LOADER, C. R. (1992). A log-linear model for a Poisson process change point. *Ann. Statist.* **20** 1391–1411. MR1186255

MACKAY, D. J. C. (2003). *Information Theory*, *Inference*, *and Algorithms*. Cambridge Univ. Press, Cambridge, UK.

MCKEE, M. (2012). Superflares' erupt on some Sun-like stars. *Nature News*, May 16, 2012.

MEI, Y., HAN, S. W. and TSUI, K.-L. (2011). Early detection of a change in Poisson rate after accounting for population size effects. *Statist. Sinica* **21** 597–624. MR2829848

MORENO, E., CASELLA, G. and GARCIA-FERRER, A. (2005). An objective Bayesian analysis of the change point problem. *Stoch. Environ. Res. Risk Assess.* **19** 191–204.

NINOMIYA, Y. (2015). Change-point model selection via AIC. *Ann. Inst. Statist. Math.* **67** 943–961.
PARK, J. H. (2010). Structural change in US presidents' use of force. *Amer. J. Polit. Sci.* **54** 766–782.
PARK, T., KRAFTY, R. T. and SÁNCHEZ, A. I. (2012). Bayesian semi-parametric analysis of Poisson change-point regression models: Application to policy-making in Cali, Colombia. *J. Appl. Stat.* **39** 2285–2298. MR2968025
PARK, T., KASHYAP, V., SIEMIGINOWSKA, A., VAN DYK, D. A., ZEZAS, A., HEINKE, C. and WARGELIN, B. J. (2006). Bayesian estimation of hardness ratios: Modeling and computations. *Astrophys. J.* **652** 610–628.
RAFTERY, A. E. and AKMAN, V. E. (1986). Bayesian analysis of a Poisson process with a change-point. *Biometrika* **73** 85–89. MR0836436
RISSANEN, J. (1989). *Stochastic Complexity in Statistical Inquiry*. World Scientific, Teaneck, NJ. MR1082556
RISSANEN, J. (2007). *Information and Complexity in Statistical Modeling*. Springer, New York. MR2287233
RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003). *Semiparametric Regression*. Cambridge Univ. Press, Cambridge. MR1998720
SCARGLE, J. D. (1998). Studies in astronomical time series analysis. V. Bayesian blocks, a new method to analyze structure in photon counting data. *Astrophys. J.* **504** 405–418.
SCARGLE, J. D., NORRIS, J. P., JACKSON, B. and CHIANG, J. (2013). Studies in astronomical time series analysis. VI. Bayesian block representations. *Astrophys. J.* **764** 167–192.
SHEN, J. J. and ZHANG, N. R. (2012). Change-point model on nonhomogeneous Poisson processes with application in copy number profiling by next-generation DNA sequencing. *Ann. Appl. Stat.* **6** 476–496. MR2976479
STRASSMEIER, K. G. (2009). Starspots. *Astron. Astrophys. Rev.* **17** 251–308.
TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242
VAN DYK, D. A. and KANG, H. (2004). Highly structured models for spectral analysis in high-energy astrophysics. *Statist. Sci.* **19** 275–293. MR2140542
VAN DYK, D. A., CONNORS, A., KASHYAP, V. and SIEMIGINOWSKA, A. (2001). Analysis of energy spectra with low photon counts via Bayesian posterior simulation. *Astrophys. J.* **548** 224–243.
VAN DYK, D., CONNORS, A., ESCH, D. N., FREEMAN, P., KANG, H., KAROVSKA, M., KASHYAP, V., SIEMIGINOWSKA, A. and ZEZAS, A. (2006). Deconvolution in high-energy astrophysics: Science, instrumentation, and methods. *Bayesian Anal.* **1** 189–235. MR2221261
WORSLEY, K. J. (1986). Confidence regions and test for a change-point in a sequence of exponential family random variables. *Biometrika* **73** 91–104. MR0836437
XU, J., VAN DYK, D. A., KASHYAP, V. L., SIEMIGINOWSKA, A., CONNORS, A., DRAKE, J. J., MENG, X. L., RATZLAFF, P. and YU, Y. (2014). A fully Bayesian method for jointly fitting instrumental calibration and X-ray spectral models. *Astrophys. J.* **794** 97 (21 pages).
YAO, Y.-C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statist. Probab. Lett.* **6** 181–189. MR0919373
YAU, C. Y., TANG, C. M. and LEE, T. C. M. (2015). Estimation of multiple-regime threshold autoregressive models with structural breaks. *J. Amer. Statist. Assoc.* **110** 1175–1186. MR3420693
ZHANG, N. R. and SIEGMUND, D. O. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* **63** 22–32. MR2345571

The Annals of Applied Statistics
2016, Vol. 10, No. 2, 1135–1136
DOI: 10.1214/16-AOAS922
Main article DOI: 10.1214/15-AOAS823
© Institute of Mathematical Statistics, 2016

CORRECTION

BAYESIAN STRUCTURED ADDITIVE DISTRIBUTIONAL REGRESSION WITH AN APPLICATION TO REGIONAL INCOME INEQUALITY IN GERMANY

BY NADJA KLEIN*, THOMAS KNEIB*,
STEFAN LANG† AND ALEXANDER SOHN*

*Georg-August-University Göttingen** *and University of Innsbruck*†

REFERENCES

KLEIN, N., KNEIB, T., LANG, S. and SOHN, A. (2015). Bayesian structured additive distributional regression with an application to regional income inequality in Germany. *Ann. Appl. Stat.* **9** 1024–1052. MR3371346

KLEIN, N., KNEIB, T., LANG, S. and SOHN, A. (2016). Supplement C to 'Correction: Bayesian structured additive distributional regression with an application to regional income inequality in Germany'. DOI:10.1214/16-AOAS922SUPP.