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

Vol. 9, No. 4—December 2015

The Annals of Applied Statistics
2015, Vol. 9, No. 4, 1709–1725
DOI: 10.1214/15-AOAS866

CUSTOMIZED TRAINING WITH AN APPLICATION TO MASS SPECTROMETRIC IMAGING OF CANCER TISSUE

BY SCOTT POWERS[1], TREVOR HASTIE[2] AND ROBERT TIBSHIRANI[3]

Stanford University

We introduce a simple, interpretable strategy for making predictions on test data when the features of the test data are available at the time of model fitting. Our proposal—*customized training*—clusters the data to find training points close to each test point and then fits an ℓ_1-regularized model (lasso) separately in each training cluster. This approach combines the local adaptivity of k-nearest neighbors with the interpretability of the lasso. Although we use the lasso for the model fitting, any supervised learning method can be applied to the customized training sets. We apply the method to a mass-spectrometric imaging data set from an ongoing collaboration in gastric cancer detection which demonstrates the power and interpretability of the technique. Our idea is simple but potentially useful in situations where the data have some underlying structure.

REFERENCES

BACHE, K. and LICHMAN, M. (2013). UCI machine learning repository. Univ. California Irvine School of Information and Computer Science, Irvine, CA.

BIEN, J. and TIBSHIRANI, R. (2011). Hierarchical clustering with prototypes via minimax linkage. *J. Amer. Statist. Assoc.* **106** 1075–1084. MR2894765

BOTTOU, L. and VAPNIK, V. (1992). Local learning algorithms. *Neural Comput.* **4** 888–900.

CORTES, C. and MOHRI, M. (2007). On transductive regression. In *Advances in Neural Information Processing Systems* 19. Vancouver, BC, Canada.

EBERLIN, L. S., TIBSHIRANI, R. J., ZHANG, J., LONGACRE, T. A., BERRY, G. J., BINGHAM, D. B., NORTON, J. A., ZARE, R. N. and POULTSIDES, G. A. (2014). Molecular assessment of surgical-resection margins of gastric cancer by mass-spectrometric imaging. *Proc. Natl. Acad. Sci. USA* **111** 2436–2441.

FU, Z., ROBLES-KELLY, A. and ZHOU, J. (2010). Mixing linear SVMs for nonlinear classification. *IEEE Trans. Neural Netw.* **21** 1963–1975.

GIL, D., GIRELA, J. L., DE JUAN, J., GOMEZ-TORRES, M. J. and JOHNSSON, M. (2012). Predicting seminal quality with artificial intelligence methods. *Expert Syst. Appl.* **39** 12564–12573.

GU, Q. and HAN, J. (2013). Clustered support vector machines. In *Proceedings of the* 16*th International Conference on Artificial Intelligence and Statistics* 307–315. Scottsdale, AZ. Available at DOI:10.1186/1475-925X-6-23.

HAMBURG, M. A. and COLLINS, F. S. (2010). The path to personalized medicine. *N. Engl. J. Med.* **363** 301–304.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. Springer, New York. MR2722294

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

Key words and phrases. Transductive learning, local regression, classification, clustering.

KAHRAMAN, H. T., SAGIROGLU, S. and COLAK, I. (2013). The development of intuitive knowledge classifier and the modeling of domain dependent data. *Knowledge-Based Systems* **37** 283–295.

LADICKY, L. and TORR, P. (2011). Locally linear support vector machines. In *Proceedings of the* 28*th International Conference on Machine Learning* 985–992. Bellevue, WA.

LITTLE, M. A., MCSHARRY, P. E., ROBERTS, S. J., COSTELLO, D. A. and MOROZ, I. M. (2007). Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *Biomed. Eng. Online* **6** 23.

MA, T. M. (2012). Local and personalised modelling for renal medical decision support system. Ph.D. thesis, Auckland Univ. Technology.

MANSOURI, K., RINGSTED, T., BALLABIO, D., TODESCHINI, R. and CONSONNI, V. (2013). Quantitative structure-activity relationship models for ready biodegradability of chemicals. *J. Chem. Inf. Model.* **53** 867–878.

SHAHBABA, B. and NEAL, R. (2009). Nonlinear models using Dirichlet process mixtures. *J. Mach. Learn. Res.* **10** 1829–1850. MR2540778

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

TORGO, L. and DACOSTA, J. P. (2003). Clustered partial linear regression. *Mach. Learn.* **50** 303–319.

TSANAS, A., LITTLE, M. A., FOX, C. and RAMIG, L. O. (2014). Objective automatic assessment of rehabilitative speech treatment in Parkinson's disease. *IEEE Trans. Neural Syst. Rehabil. Eng.* **22** 181–190.

WITTEN, D. M. and TIBSHIRANI, R. (2010). A framework for feature selection in clustering. *J. Amer. Statist. Assoc.* **105** 713–726. MR2724855

WORLD HEALTH ORGANIZATION (2013). Cancer. WHO Fact Sheet No. 297. Available at http://www.who.int/mediacentre/factsheets/fs297/en/index.html.

WU, M. and SCHÖLKOPF, B. (2007). Transductive classification via local learning regularization. In *Proceedings of the* 11*th International Conference on Artificial Intelligence and Statistics* 628–635. San Juan, Puerto Rico.

YU, K., ZHANG, T. and GONG, Y. (2009). Nonlinear learning using local coordinate coding. *Adv. Neural Inf. Process. Syst.* **21** 2223–2231.

ZHOU, D., BOUSQUET, O., LAL, T. N., WESTON, J. and SCHÖLKOPF, B. (2004). Learning with local and global consistency. *Adv. Neural Inf. Process. Syst.* **16** 321–328.

ZHU, X. (2007). Semi-supervised learning literature survey. Technical Report No. 1530, Dept. Computer Science, Univ. Wisconsin-Madison, Madison, WI.

ZHU, J., CHEN, N. and XING, E. P. (2011). Infinite SVM: A Dirichlet process mixture of large-margin kernel machines. In *Proceedings of the* 28*th International Conference on Machine Learning* 617–624. Bellevue, WA.

The Annals of Applied Statistics
2015, Vol. 9, No. 4, 1726–1760
DOI: 10.1214/15-AOAS861

THE DISCRIMINATIVE FUNCTIONAL MIXTURE MODEL FOR A COMPARATIVE ANALYSIS OF BIKE SHARING SYSTEMS

BY CHARLES BOUVEYRON, ETIENNE CÔME AND JULIEN JACQUES

Université Paris Descartes, IFSTTAR and Université Lumiére Lyon 2

Bike sharing systems (BSSs) have become a means of sustainable intermodal transport and are now proposed in many cities worldwide. Most BSSs also provide open access to their data, particularly to real-time status reports on their bike stations. The analysis of the mass of data generated by such systems is of particular interest to BSS providers to update system structures and policies. This work was motivated by interest in analyzing and comparing several European BSSs to identify common operating patterns in BSSs and to propose practical solutions to avoid potential issues. Our approach relies on the identification of common patterns between and within systems. To this end, a model-based clustering method, called FunFEM, for time series (or more generally functional data) is developed. It is based on a functional mixture model that allows the clustering of the data in a discriminative functional subspace. This model presents the advantage in this context to be parsimonious and to allow the visualization of the clustered systems. Numerical experiments confirm the good behavior of FunFEM, particularly compared to state-of-the-art methods. The application of FunFEM to BSS data from JCDecaux and the Transport for London Initiative allows us to identify 10 general patterns, including pathological ones, and to propose practical improvement strategies based on the system comparison. The visualization of the clustered data within the discriminative subspace turns out to be particularly informative regarding the system efficiency. The proposed methodology is implemented in a package for the R software, named `funFEM`, which is available on the CRAN. The package also provides a subset of the data analyzed in this work.

REFERENCES

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

BAUDRY, J.-P., MAUGIS, C. and MICHEL, B. (2012). Slope heuristics: Overview and implementation. *Stat. Comput.* **22** 455–470. MR2865029

BIRGÉ, L. and MASSART, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields* **138** 33–73. MR2288064

BORGNAT, P., ROBARDET, C., ROUQUIER, J. B., PARICE, A., FLEURY, E. and FLANDRIN, P. (2011). Shared bicycles in a city: A signal processing and data analysis perspective. *Adv. Complex Syst.* **14** 1–24.

BOUVEYRON, C. and BRUNET, C. (2012). Simultaneous model-based clustering and visualization in the Fisher discriminative subspace. *Stat. Comput.* **22** 301–324. MR2865072

Key words and phrases. Model-based clustering, functional data, dimension reduction, open data, bike sharing systems.

BOUVEYRON, C. and BRUNET, C. (2014). Discriminative variable selection for clustering with the sparse Fisher-EM algorithm. *Comput. Statist.* **29** 489–513. MR3261825

BOUVEYRON, C., GIRARD, S. and SCHMID, C. (2007). High-dimensional data clustering. *Comput. Statist. Data Anal.* **52** 502–519. MR2409998

BOUVEYRON, C. and JACQUES, J. (2011). Model-based clustering of time series in group-specific functional subspaces. *Adv. Data Anal. Classif.* **5** 281–300. MR2860102

CADIMA, J. and JOLLIFFE, I. T. (1995). Loadings and correlations in the interpretation of principal components. *J. Appl. Stat.* **22** 203–214. MR1342655

CÔME, E. and OUKHELLOU, L. (2014). Model-based count series clustering for bike-sharing system usage mining, a case study with the Vélib system of Paris. *Transportation Research—Part C Emerging Technologies* **22** 88.

DELL'OLIO, L., IBEAS, A. and MOURA, J. L. (2011). Implementing bike-sharing systems. In *ICE—Municipal Engineer* **164** 89–101. ICE publishing, London.

DUDA, R. O., HART, P. E. and STORK, D. G. (2001). *Pattern Classification*, 2nd ed. Wiley, New York. MR1802993

ESCABIAS, M., AGUILERA, A. M. and VALDERRAMA, M. J. (2005). Modeling environmental data by functional principal component logistic regression. *Environmetrics* **16** 95–107. MR2146901

FERRATY, F. and VIEU, P. (2003). Curves discrimination: A nonparametric functional approach. *Comput. Statist. Data Anal.* **44** 161–173. MR2020144

FISHER, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7** 179–188.

FRALEY, C. and RAFTERY, A. (1999). MCLUST: Software for model-based cluster analysis. *J. Classification* **16** 297–306.

FROEHLICH, J., NEUMANN, J. and OLIVER, N. (2008). Measuring the pulse of the city through shared bicycle programs. In *International Workshop on Urban, Community, and Social Applications of Networked Sensing Systems. UrbanSense*08 16–20. Raleigh, NC.

FROEHLICH, J., NEUMANN, J. and OLIVER, N. (2009). Sensing and predicting the pulse of the city through shared bicycling. In 21*st International Joint Conference on Artificial Intelligence, IJCAI*'09 1420–1426. AAAI Press, Menlo Park, CA.

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

FUKUNAGA, K. (1990). *Introduction to Statistical Pattern Recognition*, 2nd ed. Academic Press, Boston, MA. MR1075415

GIACOFCI, M., LAMBERT-LACROIX, S., MAROT, G. and PICARD, F. (2013). Wavelet-based clustering for mixed-effects functional models in high dimension. *Biometrics* **69** 31–40. MR3058049

HEARD, N. A., HOLMES, C. C. and STEPHENS, D. A. (2006). A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: An application of Bayesian hierarchical clustering of curves. *J. Amer. Statist. Assoc.* **101** 18–29. MR2252430

IEVA, F., PAGANONI, A. M., PIGOLI, D. and VITELLI, V. (2013). Multivariate functional clustering for the morphological analysis of electrocardiograph curves. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **62** 401–418. MR3060623

JACQUES, J. and PREDA, C. (2013). Funclust: A curves clustering method using functional random variable density approximation. *Neurocomputing* **112** 164–171.

JACQUES, J. and PREDA, C. (2014). Model-based clustering for multivariate functional data. *Comput. Statist. Data Anal.* **71** 92–106. MR3131956

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

KAHLE, D. and WICKHAM, H. (2013). ggmap: Spatial visualization with ggplot2. *The R Journal* **5** 144–161.

LATHIA, N., SANIUL, A. and CAPRA, L. (2012). Measuring the impact of opening the London shared bicycle scheme to casual users. *Transportation Research Part C*: *Emerging Technologies* **22** 88–102.

LÉVÉDER, C., ABRAHAM, P. A., CORNILLON, E., MATZNER-LOBER, E. and MOLINARI, N. (2004). Discrimination de courbes de prÈtrissage. In *ChimiomÈtrie* 2004 37–43.

LIN, J. R. and YANG, T. (2011). Strategic design of public bicycle sharing systems with service level constraints. *Transportation Research Part E*: *Logistics and Transportation Review* **47** 284–294.

LINDSAY, B. G. (1995). *Mixture Models*: *Theory, Geometry and Applications*. IMS, Hayward, CA.

OLSZEWSKI, R. T. (2001). Generalized feature extraction for structural pattern recognition in time-series data. Ph.D. thesis, Carnegie Mellon Univ., Pittsburgh, PA.

PREDA, C. (2007). Regression models for functional data by reproducing kernel Hilbert spaces methods. *J. Statist. Plann. Inference* **137** 829–840. MR2301719

PREDA, C., SAPORTA, G. and LÉVÉDER, C. (2007). PLS classification of functional data. *Comput. Statist.* **22** 223–235. MR2318457

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

RAY, S. and LINDSAY, B. G. (2008). Model selection in high dimensions: A quadratic-risk-based approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 95–118. MR2412633

RAY, S. and MALLICK, B. (2006). Functional clustering by Bayesian wavelet methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 305–332. MR2188987

SAMÉ, A., CHAMROUKHI, F., GOVAERT, G. and AKNIN, P. (2011). Model-based clustering and segmentation of time series with changes in regime. *Adv. Data Anal. Classif.* **5** 301–321. MR2860103

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

VOGEL, P., GREISER, T. and MATTFELD, D. C. (2011). Understanding bike-sharing systems using data mining: Exploring activity patterns. *Procedia—Social and Behavioral Sciences* **20** 514–523.

VOGEL, P. and MATTFELD, D. C. (2011). Strategic and operational planning of bike-sharing systems by data mining—A case study. In *ICCL* 127–141. Springer, Berlin.

XI, X., KEOGH, E., SHELTON, C., WEI, L. and RATANAMAHATANA, C. A. (2006). Fast time series classification using numerosity reduction. In 23*rd International Conference on Machine Learning* (*ICML* 2006) 1033–1040.

The Annals of Applied Statistics
2015, Vol. 9, No. 4, 1761–1791
DOI: 10.1214/15-AOAS862

MULTIVARIATE SPATIO-TEMPORAL MODELS FOR HIGH-DIMENSIONAL AREAL DATA WITH APPLICATION TO LONGITUDINAL EMPLOYER-HOUSEHOLD DYNAMICS[1]

BY JONATHAN R. BRADLEY, SCOTT H. HOLAN AND CHRISTOPHER K. WIKLE

University of Missouri

Many data sources report related variables of interest that are also referenced over geographic regions and time; however, there are relatively few general statistical methods that one can readily use that incorporate these multivariate spatio-temporal dependencies. Additionally, many multivariate spatio-temporal areal data sets are extremely high dimensional, which leads to practical issues when formulating statistical models. For example, we analyze Quarterly Workforce Indicators (QWI) published by the US Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) program. QWIs are available by different variables, regions, and time points, resulting in millions of tabulations. Despite their already expansive coverage, by adopting a fully Bayesian framework, the scope of the QWIs can be extended to provide estimates of missing values along with associated measures of uncertainty. Motivated by the LEHD, and other applications in federal statistics, we introduce the multivariate spatio-temporal mixed effects model (MSTM), which can be used to efficiently model high-dimensional multivariate spatio-temporal areal data sets. The proposed MSTM extends the notion of Moran's I basis functions to the multivariate spatio-temporal setting. This extension leads to several methodological contributions, including extremely effective dimension reduction, a dynamic linear model for multivariate spatio-temporal areal processes, and the reduction of a high-dimensional parameter space using a novel parameter model.

REFERENCES

ABOWD, J., SCHNEIDER, M. and VILHUBER, L. (2013). Differential privacy applications to Bayesian and linear mixed model estimation. *Journal of Privacy and Confidentiality* **5** 73–105.

ABOWD, J., STEPHENS, B., VILHUBER, L., ANDERSSON, F., MCKINNEY, K., ROEMER, M. and WOODCOCK, S. (2009). The LEHD infrastructure files and the creation of the Quarterly Workforce Indicators. In *Producer Dynamics*: *New Evidence from Micro Data* (T. Dunne, J. Jensen and M. Roberts, eds.) 149–230. Univ. Chicago Press, Chicago.

ALDWORTH, J. and CRESSIE, N. (1999). Sampling designs and prediction methods for Gaussian spatial processes. In *Multivariate Analysis*, *Design of Experiments*, *and Survey Sampling*. *Statist. Textbooks Monogr.* **159** 1–54. Dekker, New York. MR1719054

Key words and phrases. Bayesian hierarchical model, Longitudinal Employer-Household Dynamics (LEHD) program, Kalman filter, Markov chain Monte Carlo, multivariate spatio-temporal data, Moran's I basis.

ALLEGRETTO, S., DUBE, A., REICH, M. and ZIPPERER, B. (2013). Credible research designs for minimum wage studies. Working paper series 1–63, Institute for Research on Labor and Employment.

BANERJEE, S., CARLIN, B. P. and GELFAND, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall, London, UK.

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

BANERJEE, S., FINLEY, A. O., WALDMANN, P. and ERICSSON, T. (2010). Hierarchical spatial process models for multiple traits in large genetic trials. *J. Amer. Statist. Assoc.* **105** 506–521. MR2724841

BELL, W. and HILLMER, S. (1990). The time series approach to estimation for repeated surveys. *Surv. Methodol.* **16** 195–215.

BRADLEY, J. R., CRESSIE, N. and SHI, T. (2011). Selection of rank and basis functions in the spatial random effects model. In *Proceedings of the* 2011 *Joint Statistical Meetings* 3393–3406. American Statistical Association, Alexandria, VA.

BRADLEY, J. R., CRESSIE, N. and SHI, T. (2014). A comparison of spatial predictors when datasets could be very large. Preprint. Available at arXiv:1410.7748.

BRADLEY, J. R., CRESSIE, N. and SHI, T. (2015). Comparing and selecting spatial predictors using local criteria. *TEST* **24** 1–28 (Rejoinder, pp. 54–60). MR3314567

CARLIN, B. P. and BANERJEE, S. (2003). Hierarchical multivariate CAR models for spatio-temporally correlated survival data. In *Bayesian Statistics*, 7 (*Tenerife*, 2002) 45–63. Oxford Univ. Press, New York. MR2003166

CARTER, C. K. and KOHN, R. (1994). On Gibbs sampling for state space models. *Biometrika* **81** 541–553. MR1311096

CONGDON, P. (2002). A multivariate model for spatio-temporal health outcomes with an application to suicide mortality. *Geogr. Anal.* **36** 235–258.

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

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

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

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

DANIELS, M. J., ZHOU, Z. and ZOU, H. (2006). Conditionally specified space-time models for multivariate processes. *J. Comput. Graph. Statist.* **15** 157–177. MR2269367

DAVIS, E., FREEDMAN, M., LANE, J., MCCALL, B., NESTORIAK, N. and PARK, T. (2006). Supermarket human resource practices and competition from mass merchandisers. *Am. J. Agric. Econ.* **88** 1289–1295.

DUBE, A., LESTER, T. and REICH, M. (2013). Minimum wage, labor market flows, job turnover, search frictions, monopsony, unemployment. Working paper series 1–63, Institute for Research on Labor and Employment.

FEDER, M. (2001). Time series analysis of repeated surveys: The state-space approach. *Stat. Neerl.* **55** 182–199. MR1862486

FINLEY, A. O., SANG, H., BANERJEE, S. and GELFAND, A. E. (2009). Improving the performance of predictive process modeling for large datasets. *Comput. Statist. Data Anal.* **53** 2873–2884. MR2667597

FRÜHWIRTH-SCHNATTER, S. (1994). Data augmentation and dynamic linear models. *J. Time Series Anal.* **15** 183–202. MR1263889

GELMAN, A. and RUBIN, D. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 473–511.

GNEITING, T. (1999). Correlation functions for atmospheric data analysis. *Q. J. R. Meteorol. Soc.* **125** 2449–2464.
GRIFFITH, D. (2000). A linear regression solution to the spatial autocorrelation problem. *J. Geogr. Syst.* **2** 141–156.
GRIFFITH, D. A. (2002). A spatial filtering specification for the auto-Poisson model. *Statist. Probab. Lett.* **58** 245–251. MR1920751
GRIFFITH, D. (2004). A spatial filtering specification for the auto-logistic model. *Environ. Plann. A* **36** 1791–1811.
GRIFFITH, D. and TIEFELSDORF, M. (2007). Semiparametric filtering of spatial autocorrelation: The eigenvector approach. *Environ. Plann. A* **39** 1193–1221.
HIGHAM, N. J. (1988). Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra Appl.* **103** 103–118. MR0943997
HUGHES, J. and HARAN, M. (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 139–159. MR3008275
JONES, R. G. (1980). Best linear unbiased estimators for repeated surveys. *J. Roy. Statist. Soc. Ser. B* **42** 221–226. MR0583360
JONES, G. L., HARAN, M., CAFFO, B. S. and NEATH, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *J. Amer. Statist. Assoc.* **101** 1537–1547. MR2279478
KANG, E. L., CRESSIE, N. and SHI, T. (2010). Using temporal variability to improve spatial mapping with application to satellite data. *Canad. J. Statist.* **38** 271–289. MR2682762
KATZFUSS, M. and CRESSIE, N. (2012). Bayesian hierarchical spatio-temporal smoothing for very large datasets. *Environmetrics* **23** 94–107. MR2873787
LINDGREN, F., RUE, H. and LINDSTRÖM, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 423–498. MR2853727
OEHLERT, G. W. (1992). A note on the delta method. *Amer. Statist.* **46** 27–29. MR1149146
PETTITT, A. N., WEIR, I. S. and HART, A. G. (2002). A conditional autoregressive Gaussian process for irregularly spaced multivariate data with application to modelling large sets of binary data. *Stat. Comput.* **12** 353–367. MR1951708
PORTER, A. T., HOLAN, S. H. and WIKLE, C. K. (2015). Bayesian semiparametric hierarchical empirical likelihood spatial models. *J. Statist. Plann. Inference* **165** 78–90. MR3350260
PORTER, A. T., WIKLE, C. K. and HOLAN, S. H. (2015). Small area estimation via multivariate Fay-Herriot models with latent spatial dependence. *Aust. N. Z. J. Stat.* **57** 15–29. MR3335325
RAVISHANKER, N. and DEY, D. K. (2002). *A First Course in Linear Model Theory*. Chapman & Hall/CRC, Boca Raton, FL.
REICH, B. J., HODGES, J. S. and ZADNIK, V. (2006). Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics* **62** 1197–1206. MR2307445
ROBERTS, G. O. (1996). Markov chain concepts related to sampling algorithms. In *Markov Chain Monte Carlo in Practice* (W. Gilks, S. Richardson and D. Spiegelhalter, eds.). *Interdiscip. Statist.* 45–57. Chapman & Hall, London. MR1397967
ROYLE, J., BERLINER, M., WIKLE, C. and MILLIFF, R. (1999). A hierarchical spatial model for constructing wind fields from scatterometer data in the Labrador sea. In *Case Studies in Bayesian Statistics* (R. Kass, B. Carlin, A. Carriquiry, A. Gelman, I. Verdinelli and M. West, eds.) 367–382. Springer, New York.
SAMPSON, P. and GUTTORP, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *J. Amer. Statist. Assoc.* **87** 108–119.
SENGUPTA, A., CRESSIE, N., FREY, R. and KAHN, B. (2012). Statistical modeling of MODIS cloud data using the spatial random effects model. In *Proceedings of the Joint Statistical Meetings* 3111–3123. American Statistical Association, Alexandria, VA.
SHUMWAY, R. H. and STOFFER, D. S. (2006). *Time Series Analysis and Its Applications*: *With R Examples*, 2nd ed. Springer, New York. MR2228626

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

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

SUN, Y. and LI, B. (2012). Geostatistics for large datasets. In *Space-Time Processes and Challenges Related to Environmental Problems* (E. Porcu, J. M. Montero and M. Schlather, eds.) 55–77. Springer, Berlin.

THOMPSON, J. (2009). Using local labor market data to re-examine the employment effects of the minimum wage. *Ind. Labor Relat. Rev.* **63** 343–366.

TZALA, E. and BEST, N. (2008). Bayesian latent variable modelling of multivariate spatio-temporal variation in cancer mortality. *Stat. Methods Med. Res.* **17** 97–118. MR2420192

WIKLE, C. K. (2010). Low-rank representations for spatial processes. In *Handbook of Spatial Statistics* 107–118. CRC Press, Boca Raton, FL. MR2730946

ZHU, J., EICKHOFF, J. C. and YAN, P. (2005). Generalized linear latent variable models for repeated measures of spatially correlated multivariate data. *Biometrics* **61** 674–683. MR2196155

The Annals of Applied Statistics
2015, Vol. 9, No. 4, 1792–1822
DOI: 10.1214/15-AOAS884

RANDOM PARTITION MODELS AND COMPLEMENTARY CLUSTERING OF ANGLO-SAXON PLACE-NAMES

BY GIACOMO ZANELLA[1]

University of Warwick

Common cluster models for multi-type point processes model the aggregation of points of the same type. In complete contrast, in the study of Anglo-Saxon settlements it is hypothesized that administrative clusters involving complementary names tend to appear. We investigate the evidence for such a hypothesis by developing a Bayesian Random Partition Model based on clusters formed by points of different types (complementary clustering).

As a result, we obtain an intractable posterior distribution on the space of matchings contained in a k-partite hypergraph. We apply the Metropolis–Hastings (MH) algorithm to sample from this posterior. We consider the problem of choosing an efficient MH proposal distribution and we obtain consistent mixing improvements compared to the choices found in the literature. Simulated Tempering techniques can be used to overcome multimodality and a multiple proposal scheme is developed to allow for parallel programming. Finally, we discuss results arising from the careful use of convergence diagnostic techniques.

This allows us to study a data set including locations and place-names of 1316 Anglo-Saxon settlements dated approximately around 750–850 AD. Without strong prior knowledge, the model allows for explicit estimation of the number of clusters, the average intra-cluster dispersion and the level of interaction among place-names. The results support the hypothesis of organization of settlements into administrative clusters based on complementary names.

REFERENCES

BADDELEY, A. (2010). Multivariate and marked point processes. In *Handbook of Spatial Statistics*. 371–402. CRC Press, Boca Raton, FL. MR2730956

BADDELEY, A. J., MØLLER, J. and WAAGEPETERSEN, R. (2000). Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Stat. Neerl.* **54** 329–350. MR1804002

BADDELEY, A. and TURNER, R. (2005). Spatstat: an R package for analyzing spatial point patterns. *J. Stat. Softw.* **12** 1–42.

BADDELEY, A. J. and VAN LIESHOUT, M. N. M. (1995). Area-interaction point processes. *Ann. Inst. Statist. Math.* **47** 601–619. MR1370279

BECKER, R. A., WILKS, A. R. and BROWNRIGG, R. (2013). Mapdata: Extra Map Databases. R package version 2.2-2.

BERGE, C. (1973). *Graphs and Hypergraphs*. North-Holland, Amsterdam. MR0357172

Key words and phrases. Random partition models, complementary clustering, data association problems, Metropolis–Hastings algorithm, efficient proposal distribution, K-cross function, kernel smoothing, bandwidth, Anglo-Saxon place-names locations.

BROOKS, S. P. and GELMAN, A. (1998). General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Statist.* **7** 434–455. MR1665662

BROOKS, S. P. and ROBERTS, G. O. (1998). Assessing convergence of Markov chain Monte Carlo algorithms. *Stat. Comput.* **8** 319–335.

CHIU, S. N., STOYAN, D., KENDALL, W. S. and MECKE, J. (2013). *Stochastic Geometry and Its Applications*, 3rd ed. Wiley, Chichester. MR3236788

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

DELLAERT, F., SEITZ, S. M., THORPE, C. E. and THRUN, S. (2003). EM, MCMC, and chain flipping for structure from motion with unknown correspondence. *Mach. Learn.* **50** 45–71.

DICE, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology* **26** 297–302.

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

DIGGLE, P. J. (2003). *Statistical Analysis of Spatial Point Patterns*. Edward Arnold, London.

DIGGLE, P. J., EGLEN, S. J. and TROY, J. B. (2006). Modelling the bivariate spatial distribution of amacrine cells. In *Case Studies in Spatial Point Process Modeling*. *Lecture Notes in Statist.* **185** 215–233. Springer, New York. MR2232131

GELLING, M. and COLE, A. (2000). *The Landscape of Place-names*. Shaun Tyas, Stamford, Lincolnshire.

GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Anal.* **1** 515–534.

GELMAN, A. and RUBIN, D. (1992). Inference from Iterative Simulation using Multiple Sequences. *Statist. Sci.* **4** 457–511.

GEYER, C. J. and THOMPSON, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Amer. Statist. Assoc.* **90** 909–920.

GRABARNIK, P., MYLLYMÄKI, M. and STOYAN, D. (2011). Correct testing of mark independence for marked point patterns. *Ecol. Model.* **222** 3888–3894.

JANSON, S. and VEGELIUS, J. (1981). Measures of ecological association. *Oecologia* **49** 371–376.

JERRUM, M. (2003). *Counting, Sampling and Integrating: Algorithms and Complexity*. Birkhäuser, Basel. MR1960003

JERRUM, M. and SINCLAIR, A. (1996). The Markov chain Monte Carlo method: An approach to approximate counting and integration. In *Approximation Algorithms for NP-hard Problems* (D. S. Hochbaum, ed.) 482–520. PWS Publishing, Boston, MA.

JONES, R. and SEMPLE, S. (2012). *Sense of Place in Anglo-Saxon England*. Shaun Tyas, Donington.

KARPINSKI, M., RUCINSKI, A. and SZYMANSKA, E. (2012). Approximate Counting of Matchings in Sparse Uniform Hypergraphs. Preprint. Available at arXiv:1204.5335.

KUHN, H. W. (1955). The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **2** 83–97. MR0075510

LAU, J. W. and GREEN, P. J. (2007). Bayesian model-based clustering procedures. *J. Comput. Graph. Statist.* **16** 526–558. MR2351079

LAWSON, A. B. and DENISON, D. G. T. (2010). *Spatial Cluster Modelling*. CRC press, Boca Raton.

LOIZEAUX, M. A. and MCKEAGUE, I. W. (2001). Perfect sampling for posterior landmark distributions with an application to the detection of disease clusters. In *Selected Proceedings of the Symposium on Inference for Stochastic Processes* (*Athens, GA*, 2000). *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **37** 321–331. IMS, Beachwood, OH. MR2002518

MARINARI, E. and PARISI, G. (1992). Simulated tempering: A new Monte Carlo scheme. *EPL* (*Europhysics Letters*) **19** 451-458.

MÜLLER, P. and QUINTANA, F. (2010). Random partition models with regression on covariates. *J. Statist. Plann. Inference* **140** 2801–2808. MR2651966

OH, S., RUSSELL, S. and SASTRY, S. (2009). Markov chain Monte Carlo data association for multi-target tracking. *IEEE Trans. Automat. Control* **54** 481–497. MR2191542

PLUMMER, M., BEST, N., COWLES, K. and VINES, K. (2005). Output analysis and diagnostics for MCMC. R Package Version 0.10-3, URL: http://cran.rproject.org.

ROBERTS, G. O. (1998). Optimal Metropolis algorithms for product measures on the vertices of a hypercube. *Stoch. Stoch. Rep.* **62** 275–283. MR1613256

ROBERTS, G. O., GELMAN, A. and GILKS, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.* **7** 110–120. MR1428751

VALIANT, L. G. (1979). The complexity of enumeration and reliability problems. *SIAM J. Comput.* **8** 410–421. MR0539258

ZANELLA, G. (2015a). Supplement to "Bayesian complementary clustering, MCMC and Anglo-Saxon place-names." DOI:10.1214/15-AOAS884SUPPA.

ZANELLA, G. (2015b). Supplement to "Bayesian complementary clustering, MCMC and Anglo-Saxon place-names." DOI:10.1214/15-AOAS884SUPPB.

ZANELLA, G. (2015c). Supplement to "Bayesian complementary clustering, MCMC and Anglo-Saxon place-names." DOI:10.1214/15-AOAS884SUPPC.

ZANELLA, G. (2015d). Supplement to "Bayesian complementary clustering, MCMC and Anglo-Saxon place-names." DOI:10.1214/15-AOAS884SUPPD.

ZANELLA, G. (2015e). Supplement to "Bayesian complementary clustering, MCMC and Anglo-Saxon place-names." DOI:10.1214/15-AOAS884SUPPE.

ZANELLA, G. (2015f). Supplement to "Bayesian complementary clustering, MCMC and Anglo-Saxon place-names." DOI:10.1214/15-AOAS884SUPPF.

The Annals of Applied Statistics
2015, Vol. 9, No. 4, 1823–1863
DOI: 10.1214/15-AOAS859

THE LATENT STATE HAZARD MODEL, WITH APPLICATION TO WIND TURBINE RELIABILITY

BY RAMIN MOGHADDASS AND CYNTHIA RUDIN

Massachusetts Institute of Technology

We present a new model for reliability analysis that is able to distinguish the *latent* internal vulnerability state of the equipment from the vulnerability caused by temporary *external* sources. Consider a wind farm where each turbine is running under the external effects of temperature, wind speed and direction, etc. The turbine might fail because of the external effects of a spike in temperature. If it does not fail during the temperature spike, it could still fail due to internal degradation, and the spike could cause (or be an indication of) this degradation. The ability to identify the underlying latent state can help better understand the effects of external sources and thus lead to more robust decision-making. We present an experimental study using SCADA sensor measurements from wind turbines in Italy.

REFERENCES

ANDERSEN, P. K. and VAETH, M. (1989). Simple parametric and nonparametric models for excess and relative mortality. *Biometrics* **45** 523–535.

BANJEVIC, D. and JARDINE, A. K. S. (2006). Calculation of reliability function and remaining useful life for a Markov failure time process. *IMA J. Manag. Math.* **17** 115–130. MR2216398

BANJEVIC, D., JARDINE, A. K. S., MAKIS, V. and ENNIS, M. (2001). A control-limit policy and software for condition-based maintenance optimization. *INFOR* **39** 32–50.

BIAN, L. and GEBRAEEL, N. (2012). Computing and updating the first-passage time distribution for randomly evolving degradation signals. *IIE Transactions* (*Institute of Industrial Engineers*) **44** 974–987.

BIAN, L. and GEBRAEEL, N. (2013). Stochastic methodology for prognostics under continuously varying environmental profiles. *Stat. Anal. Data Min.* **6** 260–270. MR3062269

BOUTROS, T. and LIANG, M. (2011). Detection and diagnosis of bearing and cutting tool faults using hidden Markov models. *Mech. Syst. Signal Process.* **25** 2102–2124.

COLLETT, D. (2003). *Modelling Binary Data*, 2nd ed. *Chapman & Hall/CRC Texts in Statistical Science Series*. Chapman & Hall/CRC, Boca Raton, FL. MR1999899

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

DECASTRO, J. A., LITT, J. S. and FREDERICK, D. K. (2008). A modular aero-propulsion system simulation of a large commercial aircraft engine. In *AIAA*-2008-4579, 44*th AIAA/ASME/SAE/ASEE Joint Propulsion Conference & Exhibit*. Hartford, USA.

FISHER, L. D. and LIN, D. Y. (1999). Time-dependent covariates in the Cox proportional-hazards regression model. *Annu. Rev. Public Health* **20** 145–157.

FLORY, J. A., KHAROUFEH, J. P. and GEBRAEEL, N. Z. (2014). A switching diffusion model for lifetime estimation in randomly varying environments. *IIE Transactions* (*Institute of Industrial Engineers*) **46** 1227–1241.

Key words and phrases. Performance monitoring, reliability, maintenance, decision-making, big data.

GEBRAEEL, N. and PAN, J. (2008). Prognostic degradation models for computing and updating residual life distributions in a time-varying environment. *IEEE Trans. Reliab.* **57** 539–550.

GHASEMI, A., YACOUT, S. and OUALI, M. S. (2007). Optimal condition based maintenance with imperfect information and the proportional hazards model. *Int. J. Prod. Res.* **45** 989–1012.

GHASEMI, A., YACOUT, S. and OUALI, M. S. (2010). Evaluating the reliability function and the mean residual life for equipment with unobservable states. *IEEE Trans. Reliab.* **59** 45–54.

GORJIAN, N., MA, L., MITTINTY, M., YARLAGADDA, P. and SUN, Y. (2009). A review on reliability models with covariates. In 4*th World Congress on Engineering Asset Management*, *WCEAM* 2009 385–397. Athens, Greece.

GUO, H., WATSON, S., TAVNER, P. and XIANG, J. (2009). Reliability analysis for wind turbines with incomplete failure data collected from after the date of initial installation. *Reliab. Eng. Syst. Saf.* **94** 1057–1063.

HAMEED, Z., HONG, Y. S., CHO, Y. M., AHN, S. H. and SONG, C. K. (2009). Condition monitoring and fault detection of wind turbines and related algorithms: A review. *Renew. Sustain. Energy Rev.* **13** 1–39.

HÖHLE, M. (2009). Additive-multiplicative regression models for spatio-temporal epidemics. *Biom. J.* **51** 961–978. MR2744450

HONTELEZ, J. A. M., BURGER, H. H. and WIJNMALEN, D. J. D. (1996). Optimum condition-based maintenance policies for deteriorating systems with partial information. *Reliab. Eng. Syst. Saf.* **51** 267–274.

JARDINE, A. K. S., ANDERSON, P. M. and MANN, D. S. (1987). Application of the Weibull proportional hazards model to aircraft and marine engine failure data. *Qual. Reliab. Eng. Int.* **3** 77–82.

KALBFLEISCH, J. D. and PRENTICE, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd ed. Wiley, Hoboken, NJ. MR1924807

KHAROUFEH, J. P. (2003). Explicit results for wear processes in a Markovian environment. *Oper. Res. Lett.* **31** 237–244. MR1967296

KHAROUFEH, J. P. and COX, S. M. (2005). Stochastic models for degradation-based reliability. *IIE Transactions* (*Institute of Industrial Engineers*) **37** 533–542.

KHAROUFEH, J. P., FINKELSTEIN, D. E. and MIXON, D. G. (2006). Availability of periodically inspected systems with Markovian wear and shocks. *J. Appl. Probab.* **43** 303–317. MR2248566

KUSIAK, A., ZHANG, Z. and VERMA, A. (2013). Prediction, operations, and condition monitoring in wind energy. *Energy* **60** 1–12.

LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation*, 2nd ed. Springer, New York. MR1639875

LIN, D. Y. and YING, Z. (1995). Semiparametric analysis of general additive-multiplicative hazard models for counting processes. *Ann. Statist.* **23** 1712–1734. MR1370304

LU, B., LI, Y., WU, X. and YANG, Z. (2009). A review of recent advances in wind turbine condition monitoring and fault diagnosis. In 2009 *IEEE Power Electronics and Machines in Wind Applications*. Lincoln, NE.

MAKIS, V. and JARDINE, A. K. S. (1991). Computation of optimal policies in replacement models. *IMA J. Manag. Math.* **3** 169–175.

MARQUEZ, F. P. G., TOBIAS, A. M., PREZ, J. M. P. and PAPAELIAS, M. (2012). Condition monitoring of wind turbines: Techniques and methods. *Renew. Energy* **46** 169–178.

MARTINUSSEN, T. and SCHEIKE, T. H. (2002). A flexible additive multiplicative hazard model. *Biometrika* **89** 283–298. MR1913959

MARVUGLIA, A. and MESSINEO, A. (2012). Monitoring of wind farms' power curves using machine learning techniques. *Appl. Energy* **98** 574–583.

MOGHADDASS, R. and RUDIN, C. (2015). Supplement to "The latent state hazard model, with application to wind turbine reliability." DOI:10.1214/15-AOAS859SUPP.

MOGHADDASS, R. and ZUO, M. J. (2012). A parameter estimation method for a condition-monitored device under multi-state deterioration. *Reliab. Eng. Syst. Saf.* **106** 94–103.

PENG, Y. and DONG, M. (2011). A prognosis method using age-dependent hidden semi-Markov model for equipment health prediction. *Mech. Syst. Signal Process.* **25** 237–252.

PIJNENBURG, M. (1991). Additive hazards models in repairable systems reliability. *Reliab. Eng. Syst. Saf.* **31** 369–390.

QIAN, X. and WU, Y. (2014). Condition based maintenance optimization for the hydro generating unit with dynamic economic dependence. *International Journal of Control and Automation* **7** 317–326.

QIU, Y. N., FENG, Y. H., TAVNER, P. J., RICHARDSON, P., ERDOS, G. and CHEN, B. (2012). Wind turbine SCADA alarm analysis for improving reliability. *Wind Energy* **15** 951–966.

RASHID, M. and SHIFA, N. (2009). Consistency of the maximum likelihood estimator in logistic regression model: A different approach. *Journal of Statistics* **16** 1–11.

RUDIN, C. and VAHN, G.-Y. (2014). The big data newsvendor: Practical insights from machine learning. Working paper.

SAXENA, A., GOEBEL, K., SIMON, D. and EKLUND, N. (2008). Damage propagation modeling for aircraft engine run-to-failure simulation. In 2008 *International Conference on Prognostics and Health Management*, *PHM* 2008. Denver, USA.

SI, X.-S., WANG, W., HU, C.-H. and ZHOU, D.-H. (2011). Remaining useful life estimation—A review on the statistical data driven approaches. *European J. Oper. Res.* **213** 1–14. MR2795805

WU, X. and RYAN, S. M. (2011). Optimal replacement in the proportional hazards model with semi-Markovian covariate process and continuous monitoring. *IEEE Trans. Reliab.* **60** 580–589.

YANG, W., COURT, R. and JIANG, J. (2013). Wind turbine condition monitoring by the approach of SCADA data analysis. *Renewable Energy* **53** 365–376.

YANG, W., TAVNER, P. J., CRABTREE, C. J., FENG, Y. and QIU, Y. (2014). Wind turbine condition monitoring: Technical and commercial challenges. *Wind Energy* **17** 673–693.

ZAHER, A., MCARTHUR, S. D. J., INFIELD, D. G. and PATEL, Y. (2009). Online wind turbine fault detection through automated SCADA data analysis. *Wind Energy* **12** 574–593.

ZHAO, X., FOULADIRAD, M., BÉRENGUER, C. and BORDES, L. (2010). Condition-based inspection/replacement policies for non-monotone deteriorating systems with environmental covariates. *Reliab. Eng. Syst. Saf.* **95** 921–934.

ZHOU, R. R., SERBAN, N. and GEBRAEEL, N. (2011). Degradation modeling applied to residual lifetime prediction using functional data analysis. *Ann. Appl. Stat.* **5** 1586–1610. MR2849787

The Annals of Applied Statistics
2015, Vol. 9, No. 4, 1864–1888
DOI: 10.1214/15-AOAS853

BAYESIAN ANALYSIS OF TRAFFIC FLOW ON INTERSTATE I-55: THE LWR MODEL

BY NICHOLAS POLSON AND VADIM SOKOLOV

University of Chicago and Argonne National Laboratory

Transportation departments take actions to manage traffic flow and reduce travel times based on estimated current and projected traffic conditions. Travel time estimates and forecasts require information on traffic density which are combined with a model to project traffic flow such as the Lighthill–Whitham–Richards (LWR) model. We develop a particle filtering and learning algorithm to estimate the current traffic density state and the LWR parameters. These inputs are related to the so-called fundamental diagram, which describes the relationship between traffic flow and density. We build on existing methodology by allowing real-time updating of the posterior uncertainty for the critical density and capacity parameters. Our methodology is applied to traffic flow data from interstate highway I-55 in Chicago. We provide a real-time data analysis of how to learn the drop in capacity as a result of a major traffic accident. Our algorithm allows us to accurately assess the uncertainty of the current traffic state at shock waves, where the uncertainty is a mixture distribution. We show that Bayesian learning can correct the estimation bias that is present in the model with fixed parameters.

REFERENCES

ANACLETO, O., QUEEN, C. and ALBERS, C. J. (2013). Multivariate forecasting of road traffic flows in the presence of hetroscedasticity and measurement errors. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **62** 251–270. MR3045876

ARNOTT, R., DE PALMA, A. and LINDSEY, R. (1991). Does providing information to drivers reduce traffic congestion? *Transportation Research Part A*: *General* **25** 309–318.

BENGTSSON, T., BICKEL, P. and LI, B. (2008). Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems. In *Probability and Statistics*: *Essays in Honor of David A. Freedman. Inst. Math. Stat. Collect.* **2** 316–334. IMS, Beachwood, OH. MR2459957

BRILON, W., GEISTEFELDT, J. and REGLER, M. (2005). Reliability of freeway traffic flow: A stochastic concept of capacity. In *Proceedings of the* 16*th International Symposium on Transportation and Traffic Theory* **125143**. College Park, MD.

CARPENTER, J., CLIFFORD, P. and FEARNHEAD, P. (1999). Improved particle filter for nonlinear problems. *IEE Proc. Radar Sonar Navig.* **146** 2–7.

CARVALHO, C. M., JOHANNES, M. S., LOPES, H. F. and POLSON, N. G. (2010). Particle learning and smoothing. *Statist. Sci.* **25** 88–106. MR2741816

CBS CHICAGO (2014). Big delay on interstate 55 after truck fire in Romeoville. Romeoville.

CHIOU, J.-M. (2012). Dynamical functional prediction and classification, with application to traffic flow prediction. *Ann. Appl. Stat.* **6** 1588–1614. MR3058676

Key words and phrases. Traffic flow, intelligent transportation system, LWR model, particle filtering, Bayesian posterioor, traffic prediction.

CHIOU, Y.-C., LAN, L. W. and TSENG, C.-M. (2014). A novel method to predict traffic features based on rolling self-structured traffic patterns. *J. Intell. Transp. Syst.* **18** 352–366.

CHORUS, C. G., MOLIN, E. J. and VAN WEE, B. (2006). Use and effects of advanced traveller information services (ATIS): A review of the literature. *Transp. Rev.* **26** 127–149.

CHU, K.-C., YANG, L., SAIGAL, R. and SAITOU, K. (2011). Validation of stochastic traffic flow model with microscopic traffic simulation. In *IEEE Conference on Automation Science and Engineering* (*CASE*) 672–677. IEEE, New York.

CLAUDEL, C. G. and BAYEN, A. M. (2010). Lax–Hopf based incorporation of internal boundary conditions into Hamilton–Jacobi equation. Part I: Theory. *IEEE Trans. Automat. Control* **55** 1142–1157. MR2642079

COCLITE, G. M., GARAVELLO, M. and PICCOLI, B. (2005). Traffic flow on a road network. *SIAM J. Math. Anal.* **36** 1862–1886. MR2178224

COURANT, R., FRIEDRICHS, K. and LEWY, H. (1928). Über die partiellen Differenzengleichungen der mathematischen Physik. *Math. Ann.* **100** 32–74. MR1512478

DAGANZO, C. F. (1995). The cell transmission model, part II: Network traffic. *Transp. Res., Part B: Methodol.* **29** 79–93.

DERVISOGLU, G., GOMES, G., KWON, J., HOROWITZ, R. and VARAIYA, P. (2009). Automatic calibration of the fundamental diagram and empirical observations on capacity. In *Transportation Research Board* 88*th Annual Meeting* **15**. Washington, DC.

DOUCET, A., GODSILL, S. and ANDRIEU, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Stat. Comput.* **10** 197–208.

GARAVELLO, M. and PICCOLI, B. (2006). *Traffic Flow on Networks*. American Institute of Mathematical Sciences (AIMS), Springfield, MO. MR2328174

GAZIS, D. C. and KNAPP, C. H. (1971). On-line estimation of traffic densities from time-series of flow and speed data. *Transp. Sci.* **5** 283–301.

GODSILL, S. J., DOUCET, A. and WEST, M. (2004). Monte Carlo smoothing for nonlinear times series. *J. Amer. Statist. Assoc.* **99** 156–168. MR2054295

GODUNOV, S. K. (1959). A difference method for numerical calculation of discontinuous solutions of the equations of hydrodynamics. *Mat. Sb.* **47 (89)** 271–306. MR0119433

GORDON, N. J., SALMOND, D. J. and SMITH, A. F. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEE Proceedings F* (*Radar and Signal Processing*) **140** 107–113. IET.

HOLDEN, H. and RISEBRO, N. H. (1995). A mathematical model of traffic flow on a network of unidirectional roads. *SIAM J. Math. Anal.* **26** 999–1017. MR1338371

KNOOP, V. L., HOOGENDOORN, S. P. and VAN ZUYLEN, H. J. (2008). Capacity reduction at incidents: Empirical data collected from a helicopter. *Transportation Research Record: Journal of the Transportation Research Board* **2071** 19–25.

LEBACQUE, J.-P. (2005). First-order macroscopic traffic flow models: Intersection modeling, network modeling. In *Transportation and Traffic Theory. Flow, Dynamics and Human Interaction*. 16*th International Symposium on Transportation and Traffic Theory*. College Park, MD.

LEVEQUE, R. J. (2002). *Finite Volume Methods for Hyperbolic Problems*. Cambridge Univ. Press, Cambridge. MR1925043

LIGHTHILL, M. J. and WHITHAM, G. B. (1955). On kinematic waves. II. A theory of traffic flow on long crowded roads. *Proc. Roy. Soc. London. Ser. A.* **229** 317–345. MR0072606

LIU, J. and WEST, M. (2001). Combined parameter and state estimation in simulation-based filtering. In *Sequential Monte Carlo Methods in Practice* 197–223. Springer, New York. MR1847793

MAY, A. D. (1990). *Traffic Flow Fundamentals*. Prentice Hall, New York.

MIHAYLOVA, L., BOEL, R. and HEGYI, A. (2007). Freeway traffic estimation within particle filtering framework. *Automatica J. IFAC* **43** 290–300. MR2281833

MURALIDHARAN, A. and HOROWITZ, R. (2009). Imputation of ramp flow data for freeway traffic simulation. *Transportation Research Record*: *Journal of the Transportation Research Board* **2099** 58–64.

PITT, M. K. and SHEPHARD, N. (1999). Filtering via simulation: Auxiliary particle filters. *J. Amer. Statist. Assoc.* **94** 590–599. MR1702328

RICHARDS, P. I. (1956). Shock waves on the highway. *Operations Res.* **4** 42–51. MR0075522

SCHREITER, T., VAN HINSBERGEN, C., ZUURBIER, F., VAN LINT, H. and HOOGENDOORN, S. (2010). Data-model synchronization in extended Kalman filters for accurate online traffic state estimation. In *Proceedings of the Traffic Flow Theory Conference* **86**. Annecy, France.

SNYDER, C. (2011). Particle filters, the "optimal" proposal and high-dimensional systems. In *Proceedings of the ECMWF Seminar on Data Assimilation for Atmosphere and Ocean*. Shinfield Park, Reading.

STORVIK, G. (2002). Particle filters for state-space models with the presence of unknown static parameters. *IEEE Trans. Signal Process.* **50** 281–289.

SUN, X., MUÑOZ, L. and HOROWITZ, R. (2003). Highway traffic state estimation using improved mixture Kalman filters for effective ramp metering control. In *Proceedings of* 42*nd IEEE Conference on Decision and Control* **6** 6333–6338. IEEE, New York.

TEBALDI, C. and WEST, M. (1998). Bayesian inference on network traffic using link count data. *J. Amer. Statist. Assoc.* **93** 557–576. MR1631325

TRANSPORTATION RESEARCH BOARD (2010). Highway capacity manual. National Academies of Sciences, Engineering, and Medicine, Washington, DC.

WANG, Y. and PAPAGEORGIOU, M. (2005). Real-time freeway traffic state estimation based on extended Kalman filter: A general approach. *Transp. Res., Part B*: *Methodol.* **39** 141–167.

WESTGATE, B. S., WOODARD, D. B., MATTESON, D. S. and HENDERSON, S. G. (2013). Travel time estimation for ambulances using Bayesian data augmentation. *Ann. Appl. Stat.* **7** 1139–1161. MR3113504

WORK, D. B., TOSSAVAINEN, O.-P., BLANDIN, S., BAYEN, A. M., IWUCHUKWU, T. and TRACTON, K. (2008). An ensemble Kalman filtering approach to highway traffic estimation using GPS enabled mobile devices. In *CDC* 2008 *of* 47*th IEEE Conference on Decision and Control* 5062–5068. IEEE, New York.

The Annals of Applied Statistics
2015, Vol. 9, No. 4, 1889–1905
DOI: 10.1214/15-AOAS872

SPACE–TIME SMOOTHING OF COMPLEX SURVEY DATA: SMALL AREA ESTIMATION FOR CHILD MORTALITY[1]

BY LAINA D. MERCER[*], JON WAKEFIELD[*], ATHENA PANTAZIS[*], ANGELINA M. LUTAMBI[†], HONORATI MASANJA[†] AND SAMUEL CLARK[*,‡,§,¶,‖,2]

University of Washington[*], Ifakara Health Institute[†], University of Colorado[‡], University of the Witwatersrand[§], INDEPTH Network[¶] and ALPHA Network[‖]

Many people living in low- and middle-income countries are not covered by civil registration and vital statistics systems. Consequently, a wide variety of other types of data, including many household sample surveys, are used to estimate health and population indicators. In this paper we combine data from sample surveys and demographic surveillance systems to produce small area estimates of child mortality through time. Small area estimates are necessary to understand geographical heterogeneity in health indicators when full-coverage vital statistics are not available. For this endeavor spatio-temporal smoothing is beneficial to alleviate problems of data sparsity. The use of conventional hierarchical models requires careful thought since the survey weights may need to be considered to alleviate bias due to nonrandom sampling and nonresponse. The application that motivated this work is an estimation of child mortality rates in five-year time intervals in regions of Tanzania. Data come from Demographic and Health Surveys conducted over the period 1991–2010 and two demographic surveillance system sites. We derive a variance estimator of under five years child mortality that accounts for the complex survey weighting. For our application, the hierarchical models we consider include random effects for area, time and survey and we compare models using a variety of measures including the conditional predictive ordinate (CPO). The method we propose is implemented via the fast and accurate integrated nested Laplace approximation (INLA).

REFERENCES

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

ALKEMA, L., NEW, J. R., PEDERSEN, J., YOU, D. et al. (2014). Child mortality estimation 2013: An overview of updates in estimation methods by the United Nations inter-agency group for child mortality estimation. *PloS ONE* **9** e101112.

ALLISON, P. (1984). *Event History Analysis*: *Regression for Longitudinal Event Data*. Number 46. Sage, Thousand Oaks, CA.

BESAG, J., YORK, J. and MOLLIÉ, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Statist. Math.* **43** 1–59. MR1105822

Key words and phrases. Bayesian smoothing, infant mortality, small area estimation, survey sampling.

BINDER, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Int. Stat. Rev.* **51** 279–292. MR0731144

BYASS, P., WORKU, A., EMMELIN, A. and BERHANE, Y. (2007). Dss and dhs: Longitudinal and cross-sectional viewpoints on child and adolescent mortality in Ethiopia. *Population Health Metrics* **5** 12.

CLARK, S. J., WAKEFIELD, J., MCCORMICK, T. and MICHELLE, R. (2012). Hyak mortality monitoring system innovative sampling and estimation methods: Proof of concept by simulation. Technical Report 118, Center for Statistics and the Social Sciences (CSSS), Univ. Washington.

CLARK, S. J., KAHN, K., HOULE, B., ARTECHE, A., COLLINSON, M. A., TOLLMAN, S. M. and STEIN, A. (2013). Young children's probability of dying before and after their mother's death: A rural South African population-based surveillance study. *PLoS Med.* **10** e1001409.

DEMOGRAPHIC AND HEALTH SURVEYS (1992). Demographic Health Survey 1991/1992. Bureau of Statistics Planning Commission.

DEMOGRAPHIC AND HEALTH SURVEYS (1997). Tanzania Demographic and Health Survey 1996. Bureau of Statistics Tanzania and Macro International Inc.

DEMOGRAPHIC AND HEALTH SURVEYS (2000). Tanzania Demographic and Health Survey 1999. Bureau of Statistics Tanzania and Macro International Inc.

DEMOGRAPHIC AND HEALTH SURVEYS (2005). Tanzania Demographic and Health Survey 2004–05. National Bureau of Statistics (NBS) Tanzania and ORC Macro.

DEMOGRAPHIC AND HEALTH SURVEYS (2010). Tanzania Demographic and Health Survey 2010. National Bureau of Statistics (NBS) Tanzania and ICF Macro.

DWYER-LINDGREN, L., KAKUNGU, F., HANGOMA, P., NG, M., WANG, H., FLAXMAN, A. D., MASIYE, F. and GAKIDOU, E. (2014). Estimation of district-level under-5 mortality in Zambia using birth history data, 1980–2010. *Spat. Spatiotemporal Epidemiol.* **11** 89–107.

FONG, Y., RUE, H. and WAKEFIELD, J. (2010). Bayesian inference for generalized linear mixed models. *Biostatistics* **11** 397–412.

FOTTRELL, E., ENQUSELASSIE, F. and BYASS, P. (2009). The distribution and effects of child mortality risk factors in Ethiopia: A comparison of estimates from dss and dhs. *Ethiopian Journal of Health Development* **23** 163–168.

GELMAN, A. (2007). Struggles with survey weighting and regression modeling. *Statist. Sci.* **22** 153–164. MR2408951

HAMMER, G. P., KOUYATÉ, B., RAMROTH, H. and BECHER, H. (2006). Risk factors for childhood mortality in sub-Saharan Africa. A comparison of data from a demographic and health survey and from a demographic surveillance system. *Acta Trop.* **98** 212–218.

HELD, L., SCHRÖDLE, B. and RUE, H. (2010). Posterior and cross-validatory predictive checks: A comparison of MCMC and INLA. In *Statistical Modelling and Regression Structures* 91–110. Physica-Verlag/Springer, Heidelberg. MR2664630

HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663–685. MR0053460

INDEPTH NETWORK (2014). Health and demographic surveillance systems. Available at http://www.indepth-network.org/index.php?option=com_content&task=view&id=1798&Itemid=501.

JENKINS, S. P. (1995). Easy estimation methods for discrete-time duration models. *Oxford Bulletin of Economics and Statistics* **57** 129–136.

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

LOHR, S. L. (2010). *Sampling*: *Design and Analysis*, 2nd ed. Brooks/Cole, Cengage Learning, Boston, MA. MR3057878

LUMLEY, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software* **9** 1–19.

MERCER, L., WAKEFIELD, J., CHEN, C. and LUMLEY, T. (2014). A comparison of spatial smoothing methods for small area estimation with sampling weights. *Spat. Stat.* **8** 69–85. MR3326822

MERCER, L. D., WAKEFIELD, J., PANTAZIS, A., LUTAMBI, A., MASANJA, H. and CLARK, S. (2015). Supplement to "Space–time smoothing of complex survey data: Small area estimation for child mortality." DOI:10.1214/15-AOAS872SUPP.

PARIS21 (2014). Paris21: Partnership for statistics in development in the 21st century. Available at http://www.paris21.org.

PEDERSEN, J. and LIU, J. (2012). Child mortality estimation: Appropriate time periods for child mortality estimates from full birth histories. *PLoS Med*. **9** e1001289.

PLUMMER, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics* **9** 523–539.

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

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

RUTSTEIN, S. O. and ROJAS, G. (2006). *Tanzania Demographic and Health Survey* 1996. ORC Macro, Calverton, MD.

SCHRÖDLE, B. and HELD, L. (2011). Spatio-temporal disease mapping using INLA. *Environmetrics* **22** 725–734. MR2843139

SØRBYE, S. H. and RUE, H. (2014). Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spat. Stat*. **8** 39–51. MR3326820

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

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2014). The deviance information criterion: 12 years on. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **76** 485–493. MR3210727

UN (2000). Millennium development goals. Available at http://www.un.org/millenniumgoals/.

UN (2014a). Civil registration and vital statistics coverage. Available at http://unstats.un.org/unsd/demographic/CRVS/CR_coverage.htm.

UN (2014b). Data revolution for sustainable development. Available at http://www.un.org/apps/news/story.asp?NewsID=48594#.VEVQpoctuvJ.

UN (2014c). Millennium development goal number 4: Reduce by two thirds, between 1990 and 2015, the under-five mortality rate. Available at http://www.un.org/millenniumgoals/childhealth.shtml.

UN (2014d). The post-2015 development agenda. Available at http://www.post2015hlp.org/the-report/.

UN (2014e). Sustainable development goals. Available at http://sustainabledevelopment.un.org/owg.html.

USAID (2014). Demographic and health surveys. United States Agency for International Development. Available at http://www.dhsprogram.com.

WAKEFIELD, J. (2009). Multi-level modelling, the ecologic fallacy, and hybrid study designs. *Int. J. Epidemiol*. **38** 330–336.

WANG, H., LIDDELL, C. A., COATES, M. M., MOONEY, M. D., LEVITZ, C. E., SCHUMACHER, A. E., APFEL, H., IANNARONE, M., PHILLIPS, B., LOFGREN, K. T. et al. (2014). Global, regional, and national levels of neonatal, infant, and under-5 mortality during 1990–2013: A systematic analysis for the global burden of disease study 2013. *The Lancet* **384** 957–979.

WORLD BANK AND WORLD HEALTH ORGANIZATION (2014). Global civil registration and vital statistics scaling up investment plan 2015–2024. Available at http://www.worldbank.org/en/topic/health/publication/global-civil-registration-vital-statistics-scaling-up-investment.

YE, Y., WAMUKOYA, M., EZEH, A., EMINA, J. B. and SANKOH, O. (2012). Health and demographic surveillance systems: A step towards full civil registration and vital statistics system in sub-Sahara Africa? *BMC Public Health* **12** 741.

The Annals of Applied Statistics
2015, Vol. 9, No. 4, 1906–1931
DOI: 10.1214/15-AOAS881

EVALUATING THE CAUSAL EFFECT OF UNIVERSITY GRANTS ON STUDENT DROPOUT: EVIDENCE FROM A REGRESSION DISCONTINUITY DESIGN USING PRINCIPAL STRATIFICATION

BY FAN LI[*,1], ALESSANDRA MATTEI[†,2] AND FABRIZIA MEALLI[†]

Duke University[*] and University of Florence[†]

Regression discontinuity (RD) designs are often interpreted as locally randomized experiments for units with a realized value of a pretreatment variable falling around a threshold. Motivated by the evaluation of Italian university grants, we consider a fuzzy RD design where the treatment status is based on both eligibility criteria and a voluntary application status. Resting on the fact that grant application and grant receipt statuses are post-assignment (post-eligibility) intermediate variables, we use the principal stratification framework to define causal estimands within the Rubin Causal Model. We propose a probabilistic formulation of the assignment mechanism underlying RD designs, by reformulating the Stable Unit Treatment Value Assumption (SUTVA) and making an explicit local overlap assumption for a subpopulation around the threshold. We invoke a local randomization assumption instead of the more standard continuity assumptions. We also develop a Bayesian approach to select the target subpopulation(s) with adjustment for multiple comparisons, and to draw inference for the target causal estimands within this framework. Applying the method to the data from two Italian universities, we find evidence that university grants are effective in preventing students from low-income families from dropping out of higher education.

REFERENCES

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

ANGRIST, J. D., IMBENS, G. W. and RUBIN, D. B. (1996). Identification of causal effects using instrumental variables (with discussion). *J. Amer. Statist. Assoc.* **91** 444–472.

BATTISTIN, E. and RETTORE, E. (2008). Ineligibles and eligible non-participants as a double comparison group in regression discontinuity designs. *J. Econometrics* **142** 715–730.

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

BERK, R. A. and DE LEUUW, J. (1999). An evaluation of California's inmate classification system using a generalized regression discontinuity design. *J. Amer. Statist. Assoc.* **94** 1045–1052.

BERRY, S. M. and BERRY, D. A. (2004). Accounting for multiplicities in assessing drug safety: A three-level hierarchical mixture model. *Biometrics* **60** 418–426. MR2066276

CATTANEO, M. D., FRANDSEN, B. and TITIUNIK, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the U.S. Senate. *Journal of Causal Inference* **3** 1–24.

Key words and phrases. Bayesian, causal effects, intermediate variables, principal stratification, randomization, regression discontinuity, university grants.

CELLINI, S. R., FERREIRA, F. and ROTHSTEIN, J. (2010). The value of school facility investments: Evidence from a dynamic regression discontinuity design. *Q. J. Econ.* **125** 215–261.

CHIB, S. and GREENBERG, E. (2014). Nonparametric Bayes analysis of the sharp and fuzzy regression discontinuity designs. Technical report, Washington Univ. St Louis, Olin School of Business.

CHIB, S. and JACOBI, L. (2015). Bayesian fuzzy regression discontinuity analysis and returns to compulsory schooling. *J. Appl. Econometrics*. Published online in Wiley Online Library (wileyonlinelibrary.com), DOI:10.1002/jae.2481.

COOK, T. D. (2008). "Waiting for life to arrive": A history of the regression-discontinuity design in psychology, statistics and economics. *J. Econometrics* **142** 636–654. MR2416822

DE FINETTI, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Ann. Inst. Henri Poincaré* **7** 1–68.

DINARDO, J. and LEE, D. S. (2011). Program evaluation and research designs. In *Handbook of Labor Economics* **4A** 463–536. Elsevier, Philadelphia, PA.

ELLIOTT, M. R., RAGHUNATHAN, T. E. and LI, Y. (2010). Bayesian inference for causal mediation effects using principal stratification with dichotomous mediators and outcomes. *Biostatistics* **11** 353–372.

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

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

GARIBALDI, P., GIAVAZZI, F., ICHINO, A. and RETTORE, E. (2012). College cost and time to complete a degree: Evidence from tuition discontinuities. *Rev. Econ. Stat.* **94** 699–711.

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

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

GOSSELIN, F. (2011). A new calibrated Bayesian internal goodness-of-fit method: Sampled posterior p-values as simple and general p-values that allow double use of the data. *PLoS ONE* **6** 1–10.

HAHN, J., TODD, P. E. and VAN DER KLAAUW, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* **69** 201–209.

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

IMBENS, G. W. and ANGRIST, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62** 467–476.

IMBENS, G. and KALYANARAMAN, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *Rev. Econ. Stud.* **79** 933–959. MR2986387

IMBENS, G. W. and LEMIEUX, T. (2008). Regression discontinuity designs: A guide to practice. *J. Econometrics* **142** 615–635. MR2416821

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

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

IMBENS, G. W. and ZAJONC, T. (2011). Regression discontinuity design with multiple forcing variables. Technical report, Harvard Univ., Dept. Economics.

JOHNSON, V. E. (2007). Bayesian model assessment using pivotal quantities. *Bayesian Anal.* **2** 719–733. MR2361972

LEE, D. S. (2008). Randomized experiments from non-random selection in U.S. House elections. *J. Econometrics* **142** 675–697. MR2416824

LEE, D. S. and LEMIEUX, T. (2010). Regression discontinuity designs in economics. *J. Econ. Lit.* **485** 281–355.

LI, F., MATTEI, A. and MEALLI, F. (2015). Supplement to "Evaluating the causal effect of university grants on student dropout: Evidence from a regression discontinuity design using principal stratification." DOI:10.1214/15-AOAS881SUPP.

LUDWIG, J. and MILLER, D. L. (2007). Does head start improve children's life chances? Evidence from a regression discontinuity design. *Q. J. Econ.* **122** 15981–208.

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

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

MEALLI, F. and RAMPICHINI, C. (2012). Evaluating the effects of university grants by using regression discontinuity designs. *J. Roy. Statist. Soc. Ser. A* **175** 775–798. MR2948374

MEALLI, F. and RUBIN, B. D. (2002). Assumptions when analyzing randomized experiments with noncompliance and missing outcomes. *Health Serv. Outcomes Res. Methodol.* **3** 225–232.

MERCATANTI, A. (2013). A likelihood-based analysis for relaxing the exclusion restriction in randomized experiments with noncompliance. *Aust. N. Z. J. Stat.* **55** 129–153. MR3079024

MERCATANTI, A., LI, F. and MEALLI, F. (2015). Improving inference of Gaussian mixtures using auxiliary variables. *Stat. Anal. Data Min.* **8** 34–48. MR3315979

MURPHY, S. A. (2003). Optimal dynamic treatment regimes. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65** 331–366. MR1983752

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psycology* **66** 688–701.

RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58. MR0472152

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

SALES, A. and HANSEN, B. (2014). Limitless regression discontinuity: Causal inference for a population surrounding a threshold. Available at arXiv:1403.5478.

SCHWARTZ, S. L., LI, F. and MEALLI, F. (2011). A Bayesian semiparametric approach to intermediate variables in causal inference. *J. Amer. Statist. Assoc.* **31** 949–962.

SCOTT, J. G. and BERGER, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *J. Statist. Plann. Inference* **136** 2144–2162. MR2235051

THISTLETHWAITE, D. and CAMPBELL, D. (1960). Regression-discontinuity analysis: An alternative to the ex-post facto experiment. *J. Educ. Psychol.* **51** 309–317.

VAN DER KLAAUW, W. (2002). Estimating the effect of financial aid offers on college enrollment: A regression-discontinuity approach. *Internat. Econom. Rev.* **43** 1249–1287.

VAN DER KLAAUW, W. (2008). Regression-discontinuity analysis: A survey of recent development in economics. *Labour* **22** 219–245.

ZAJONC, T. (2012). Bayesian inference for dynamic treatment regimes: Mobility, equity, and efficiency in student tracking. *J. Amer. Statist. Assoc.* **107** 80–92. MR2949343

The Annals of Applied Statistics
2015, Vol. 9, No. 4, 1932–1949
DOI: 10.1214/15-AOAS867

LYMPHANGIOGENESIS AND CARCINOMA IN THE UTERINE CERVIX: JOINT AND HIERARCHICAL MODELS FOR RANDOM CLUSTER SIZES AND CONTINUOUS OUTCOMES

BY T. R. FANSHAWE*, C. M. CHAPMAN† AND T. CRICK†

*University of Oxford** *and Royal Lancaster Infirmary*†

Although the lymphatic system is clearly linked to the metastasis of most human carcinomas, the mechanisms by which lymphangiogenesis occurs in response to the presence of carcinoma remain unclear. Hierarchical models are presented to investigate the properties of lymphatic vessel production in 2997 fields taken from 20 individuals with invasive carcinoma, 21 individuals with cervical intraepithelial neoplasia and 21 controls. Such data demonstrate a high degree of correlation within tumour samples from the same individual. Joint hierarchical models utilising shared random effects are discussed and fitted in a Bayesian framework to allow for the correlation between two key outcome measures: a random cluster size (the number of lymphatic vessels in a tissue sample) and a continuous outcome (vessel size). Results show that invasive carcinoma samples are associated with increased production of smaller and more irregularly-shaped lymphatic vessels and suggest a mechanistic link between carcinoma of the cervix and lymphangiogenesis.

REFERENCES

ALITALO, K., TAMMELA, T. and PETROVA, T. V. (2005). Lymphangiogenesis in development and human disease. *Nature* **438** 946–953.

CARTER, B. (2010). Cluster size variability and imbalance in cluster randomized controlled trials. *Stat. Med.* **29** 2984–2993. MR2758393

CATALANO, P. J. and RYAN, L. M. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *J. Amer. Statist. Assoc.* **87** 651–658.

CHAPMAN, C. M., FANSHAWE, T. R. and CRICK, T. (2013). An investigation into the changes of lymphatic vessel density due to invasive carcinoma and cervical intraepithelial lesions of the uterine cervix. *Morecambe Bay Medical Journal* **6** 355–359.

CONG, X. J., YIN, G. and SHEN, Y. (2007). Marginal analysis of correlated failure time data with informative cluster sizes. *Biometrics* **63** 663–672. MR2395702

DAY, S. J. and GRAHAM, D. F. (1989). Sample size and power for comparing two or more treatment groups in clinical trials. *BMJ* **299** 663–665.

DEAN, C. B. (1992). Testing for overdispersion in Poisson and Negative Binomial regression models. *J. Amer. Statist. Assoc.* **87** 451–457.

DUNSON, D. B., CHEN, Z. and HARRY, J. (2003). A Bayesian approach for joint modeling of cluster size and subunit-specific outcomes. *Biometrics* **59** 521–530. MR2004257

FERLAY, J., SHIN, H. R., BRAY, F., FORMAN, D., MATHERS, C. and PARKIN, D. M. (2008). GLOBOCAN 2008 v2.0, Cancer incidence and mortality worldwide: IARC CancerBase no. 10 [Internet] (accessed 4th March 2013).

Key words and phrases. Cervical carcinoma, informative cluster size, hierarchical model, joint model, lymphangiogenesis, random effect.

FITZMAURICE, G. M. and LAIRD, N. M. (1995). Regression models for a bivariate discrete and continuous outcome with clustering. *J. Amer. Statist. Assoc.* **90** 845–852. MR1354003

FRIEDL, P. and WOLF, K. (2003). Tumour-cell invasion and migration: Diversity and escape mechanisms. *Nat. Rev. Cancer* **3** 362–374.

GAO, P., ZHOU, G. Y., YIN, G., LIU, Y., LIU, Z. Y., ZHANG, J. and HAO, C. Y. (2006). Lymphatic vessel density as a prognostic indicator for patients with stage I cervical carcinoma. *Human Pathology* **37** 719–725.

GOLDSTEIN, H., YANG, M., OMAR, R., TURNER, R. and THOMPSON, S. (2000). Meta-analysis using multilevel models with an application to the study of class size effects. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **49** 399–412. MR1824548

GOMBOS, Z., XU, X., CHU, C. S., ZHANG, P. J. and ACS, G. (2005). Peritumoral lymphatic vessel density and vascular endothelial growth factor C expression in early-stage squamous cell carcinoma of the uterine cervix. *Clinical Cancer Research* **11** 8364–8371.

GUEORGUIEVA, R. V. (2005). Comments about joint modeling of cluster size and binary and continuous subunit-specific outcomes. *Biometrics* **61** 862–867. MR2196176

GUEORGUIEVA, R. V. and AGRESTI, A. (2001). A correlated probit model for joint modeling of clustered binary and continuous reponses. *J. Amer. Statist. Assoc.* **96** 1102–1112. MR1947258

HIBBS, A. M., BLACK, D., PALERMO, L., CNAAN, A., LUAN, X., TRUOG, W. E., WALSH, M. C. and BALLARD, R. A. (2010). Accounting for multiple births in neonatal and perinatal trials: Systematic review and case study. *J. Pediatr.* **156** 202–208.

IBRAHIM, J. G., CHU, H. and CHEN, L. M. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *J. Clin. Oncol.* **28** 2796–2801.

LINDSEY, J. K. (1999). On the use of corrections for overdispersion. *Applied Statistics* **48** 553–561.

LONGATTO-FILHO, A., PINHEIRO, C., PEREIRA, S. M. M., ETLINGER, D., MOREIRA, M. A. R., JUBÉ, L. F., QUIEROZ, G. S., BALTAZAR, F. and SCHMITT, F. C. (2007). Lymphatic vessel density and epithelial D2-40 immunoreactivity in pre-invasive and invasive lesions of the uterine cervix. *Gynecologic Oncology* **107** 45–51.

LUNN, D. J., THOMAS, A., BEST, N. and SPIEGELHALTER, D. (2000). WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. *Stat. Comput.* **10** 325–337.

MA, R., JØRGENSEN, B. and WILLMS, J. D. (2009). Clustered binary data with random cluster sizes: A dual Poisson modelling approach. *Stat. Model.* **9** 137–150. MR2750122

NEUHAUS, J. M. and MCCULLOCH, C. E. (2011). Estimation of covariate effects in generalized linear mixed models with informative cluster sizes. *Biometrika* **98** 147–162. MR2804216

PANAGEAS, K. S., SCHRAG, D., LOCALIO, A. R., VENKATRAMAN, E. S. and BEGG, C. B. (2007). Properties of analysis methods that account for clustering in volume-outcome studies when the primary predictor is cluster size. *Stat. Med.* **26** 2017–2035. MR2364289

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

PINHEIRO, J., BATES, D., DEBROY, S., SARKAR, D. and THE R CORE TEAM (2008). nlme: Linear and nonlinear mixed effects models. R package version 3.1-89.

R DEVELOPMENT CORE TEAM (2008). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

REGAN, M. M. and CATALANO, P. J. (1999). Likelihood models for clustered binary and continuous outcomes: Application to developmental toxicology. *Biometrics* **55** 760–768.

TEN HAVE, T. R. and CHINCHILLI, V. M. (1998). Two-stage negative binomial and overdispersed Poisson models for clustered developmental toxicity data with random cluster size. *J. Agric. Biol. Environ. Stat.* **3** 75–98. MR1817034

VENABLES, W. N. and RIPLEY, B. D. (2002). *Modern Applied Statistics with S*, 4th ed. Springer, New York.

WICKSELL, S. D. (1925). The corpuscle problem: A mathematical study of a biometric problem. *Biometrika* **17** 84–99.

WILLIAMSON, J. M., DATTA, S. and SATTEN, G. A. (2003). Marginal analyses of clustered data when cluster size is informative. *Biometrics* **59** 36–42. MR1978471

ZHANG, S., YU, H. and ZHANG, L. (2009). Clinical implications of increased lymph vessel density in the lymphatic metastasis of early stage invasive cervical carcinoma: A clinical immunohistochemical method study. *BMC Cancer* **9** 1–6.

The Annals of Applied Statistics
2015, Vol. 9, No. 4, 1950–1972
DOI: 10.1214/15-AOAS858

ANALYSIS OF MULTIVIEW LEGISLATIVE NETWORKS WITH STRUCTURED MATRIX FACTORIZATION: DOES TWITTER INFLUENCE TRANSLATE TO THE REAL WORLD?

BY SHAWN MANKAD AND GEORGE MICHAILIDIS

Cornell University and University of Michigan

The rise of social media platforms has fundamentally altered the public discourse by providing easy to use and ubiquitous forums for the exchange of ideas and opinions. Elected officials often use such platforms for communication with the broader public to disseminate information and engage with their constituencies and other public officials. In this work, we investigate whether Twitter conversations between legislators reveal their real-world position and influence by analyzing multiple Twitter networks that feature different types of link relations between the Members of Parliament (MPs) in the United Kingdom and an identical data set for politicians within Ireland. We develop and apply a matrix factorization technique that allows the analyst to emphasize nodes with contextual local network structures by specifying network statistics that guide the factorization solution. Leveraging only link relation data, we find that important politicians in Twitter networks are associated with real-world leadership positions, and that rankings from the proposed method are correlated with the number of future media headlines.

REFERENCES

BARRAT, A., BARTHÉLEMY, M., PASTOR-SATORRAS, R. and VESPIGNANI, A. (2004). The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. USA* **101** 3747–3752.

BERRY, M. W., BROWNE, M., LANGVILLE, A. N., PAUCA, V. P. and PLEMMONS, R. J. (2007). Algorithms and applications for approximate nonnegative matrix factorization. *Comput. Statist. Data Anal.* **52** 155–173. MR2409971

BRANDES, U., FLEISCHER, D. and PUPPE, T. (2006). Dynamic spectral layout of small worlds. In *Graph Drawing* (P. Healy and N. Nikolov, eds.). *Lecture Notes in Computer Science* **3843** 25–36. Springer, Berlin. MR2244497

CHA, M., HADDADI, H., BENEVENUTO, F. and GUMMADI, P. K. (2010). Measuring user influence in Twitter: The million follower fallacy. *ICWSM* **10** 10–17.

CHEW, C. and EYSENBACH, G. (2010). Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak. *PLoS ONE* **5** e14118.

COAKLEY, J. and GALLAGHER, M. (2005). *Politics in the Republic of Ireland*. Psychology Press, New York.

DING, C., LI, T. and JORDAN, M. I. (2010). Convex and semi-nonnegative matrix factorizations. *IEEE Trans. Pattern Anal. Mach. Intell.* **32** 45–55.

FIENBERG, S. E. (2012). A brief history of statistical models for network analysis and open challenges. *J. Comput. Graph. Statist.* **21** 825–839. MR3005799

FOSDICK, B. K. and HOFF, P. D. (2013). Testing and modeling dependencies between a network and nodal attributes. Available at arXiv:1306.4708.

Key words and phrases. Matrix factorization, networks, influence, Twitter.

FOSDICK, B. K. and HOFF, P. D. (2014). Separable factor analysis with applications to mortality data. *Ann. Appl. Stat.* **8** 120–147. MR3191985

FREEMAN, L. C. (1979). Centrality in social networks conceptual clarification. *Social Networks* **1** 215–239.

GEMULLA, R., NIJKAMP, E., HAAS, P. J. and SISMANIS, Y. (2011). Large-scale matrix factorization with distributed stochastic gradient descent. In *Proceedings of the* 17*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 69–77. ACM, New York.

GILLIS, N. and GLINEUR, F. (2008). Nonnegative factorization and the maximum edge biclique problem. Available at arXiv:0810.4225.

GOLBECK, J., GRIMES, J. M. and ROGERS, A. (2010). Twitter use by the U.S. Congress. *J. Am. Soc. Inf. Sci. Technol.* **61** 1612–1621.

GREENE, D. and CUNNINGHAM, P. (2013). Producing a unified graph representation from multiple social network views. Available at arXiv:1301.5809.

GREENE, D., O'CALLAGHAN, D. and CUNNINGHAM, P. (2012). Identifying topical Twitter communities via user list aggregation. In 2*nd International Workshop on Mining Communities and People Recommenders* (*COMMPER* 2012) *at ECML* 2012. Bristol, UK.

HUBERMAN, B. A., ROMERO, D. M. and WU, F. (2008). Social networks that matter: Twitter under the microscope. *CoRR* **abs/0812.1045**.

JOLLIFFE, I. T. (1986). *Principal Component Analysis*. Springer, New York. MR0841268

KATAYAMA, J., TAKAHASHI, N. and TAKEUCHI, J. (2013). Boundedness of modified multiplicative updates for nonnegative matrix factorization. In *IEEE* 5*th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing* (*CAMSAP*) 252–255. St. Martin.

KLEINBERG, J. M. (1999). Authoritative sources in a hyperlinked environment. *J. ACM* **46** 604–632. MR1747649

KOREN, Y. (2005). Drawing graphs by eigenvectors: Theory and practice. *Comput. Math. Appl.* **49** 1867–1888. MR2154691

KROONENBERG, P. M. and DE LEEUW, J. (1980). Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika* **45** 69–97. MR0570771

LEE, D. D. and SEUNG, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* **401** 788–791.

LIN, Y.-R., CHI, Y., ZHU, S., SUNDARAM, H. and TSENG, B. L. (2008). Facetnet: A framework for analyzing communities and their evolutions in dynamic networks. In *Proceedings of the* 17*th International Conference on World Wide Web*. 685–694. ACM, New York.

MANKAD, S. and MICHAILIDIS, G. (2013a). Discovery of path-important nodes using structured semi-nonnegative matrix factorization. In *IEEE* 5*th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing* (*CAMSAP*) 288–291. St. Martin.

MANKAD, S. and MICHAILIDIS, G. (2013b). Structural and functional discovery in dynamic networks with non-negative matrix factorization. *Phys. Rev. E* **88** 042812.

MANKAD, S. and MICHAILIDIS, G. (2015). Supplement to "Analysis of multiview legislative networks with structured matrix factorization: Does Twitter influence translate to the real world?" DOI:10.1214/15-AOAS858SUPP.

MCKELVEY, K., DIGRAZIA, J. and ROJAS, F. (2014). Twitter publics: How online political communities signaled electoral outcomes in the 2010 US house election. *Information, Communication & Society* **17** 436–450.

NEWMAN, M. E. J. (2010). *Networks*. Oxford Univ. Press, Oxford. MR2676073

OWEN, A. B. and PERRY, P. O. (2009). Bi-cross-validation of the SVD and the nonnegative matrix factorization. *Ann. Appl. Stat.* **3** 564–594. MR2750673

PAGE, L., BRIN, S., MOTWANI, R. and WINOGRAD, T. (1999). The PageRank citation ranking: Bringing order to the web. Stanford InfoLab, Stanford, CA. Available at: http://ilpubs.stanford.edu:8090/422/.

PSORAKIS, I., ROBERTS, S., EBDEN, M. and SHELDON, B. (2011). Overlapping community detection using Bayesian non-negative matrix factorization. *Phys. Rev. E* **83** 066114.

RECHT, B. and RÉ, C. (2013). Parallel stochastic gradient algorithms for large-scale matrix completion. *Math. Program. Comput.* **5** 201–226. MR3069879

ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915. MR2893856

ROHE, K. and YU, B. (2012). Co-clustering for directed graphs; the stochastic co-blockmodel and a spectral algorithm. Available at arXiv:1204.2296.

ROMERO, D. M., MEEDER, B. and KLEINBERG, J. (2011). Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter. In *Proceedings of the* 20*th International Conference on World Wide Web*. 695–704. ACM, New York.

SALTER-TOWNSHEND, M. and MURPHY, T. B. (2015). Role analysis in networks using mixtures of exponential random graph models. *J. Comput. Graph. Statist.* **24** 520–538.

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

THE NEW YORK TIMES BLOGS (2011). Twitter Starts Selling Political Ads. Available at http://thecaucus.blogs.nytimes.com/2011/09/21/twitter-starts-selling-political-ads/. Accessed: 2013-11-13.

THE NEW YORK TIMES BLOGS (2012). Pepsi and Twitter Announce Partnership on Ad Campaign. Available at http://mediadecoder.blogs.nytimes.com/2012/05/30/pepsi-and-twitter-announce-partnership-on-ad-campaign. Accessed: 2013-11-13.

THE NEW YORK TIMES (2013). Using Twitter to Move the Markets. http://www.nytimes.com/2013/10/07/business/media/using-twitter-to-move-the-markets.html. Accessed: 2013-11-13.

TUMASJAN, A., SPRENGER, T. O., SANDNER, P. G. and WELPE, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* 178–185. Washington, DC.

TWITTER, INC. (2014). About Twitter, Inc. Available at https://about.twitter.com/company. Accessed: 2014-09-19.

UNANKARD, S., LI, X., SHARAF, M., ZHONG, J. and LI, X. (2014). Predicting elections from social networks based on sub-event detection and sentiment analysis. In *Web Information Systems Engineering—WISE* 2014 (B. Benatallah, A. Bestavros, Y. Manolopoulos, A. Vakali and Y. Zhang, eds.). *Lecture Notes in Computer Science* **8787** 1–16. Springer, Berlin.

WANG, F., LI, T., WANG, X., ZHU, S. and DING, C. (2011). Community discovery using nonnegative matrix factorization. *Data Min. Knowl. Discov.* **22** 493–521. MR2785131

XU, W., LIU, X. and GONG, Y. (2003). Document clustering based on non-negative matrix factorization. In *Proceedings of the* 26*th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval* 267–273. ACM, New York.

YANG, J., MCAULEY, J. and LESKOVEC, J. (2013). Community detection in networks with node attributes. In *IEEE* 13*th International Conference on Data Mining* (*ICDM*) 1151–1156. IEEE, New York.

The Annals of Applied Statistics
2015, Vol. 9, No. 4, 1973–1996
DOI: 10.1214/15-AOAS870

FEATURE EXTRACTION FOR PROTEOMICS IMAGING MASS SPECTROMETRY DATA

BY LYRON J. WINDERBAUM, INGE KOCH, OVE J. R. GUSTAFSSON, STEPHAN MEDING AND PETER HOFFMANN[1]

The University of Adelaide

Imaging mass spectrometry (IMS) has transformed proteomics by providing an avenue for collecting spatially distributed molecular data. Mass spectrometry data acquired with matrix assisted laser desorption ionization (MALDI) IMS consist of tens of thousands of spectra, measured at regular grid points across the surface of a tissue section. Unlike the more standard liquid chromatography mass spectrometry, MALDI-IMS preserves the spatial information inherent in the tissue.

Motivated by the need to differentiate cell populations and tissue types in MALDI-IMS data accurately and efficiently, we propose an integrated cluster and feature extraction approach for such data. We work with the derived binary data representing presence/absence of ions, as this is the essential information in the data. Our approach takes advantage of the spatial structure of the data in a noise removal and initial dimension reduction step and applies k-means clustering with the cosine distance to the high-dimensional binary data. The combined smoothing-clustering yields spatially localized clusters that clearly show the correspondence with cancer and various noncancerous tissue types.

Feature extraction of the high-dimensional binary data is accomplished with our difference in proportions of occurrence (DIPPS) approach which ranks the variables and selects a set of variables in a data-driven manner. We summarize the best variables in a single image that has a natural interpretation. Application of our method to data from patients with ovarian cancer shows good separation of tissue types and close agreement of our results with tissue types identified by pathologists.

REFERENCES

AEBERSOLD, R. and MANN, M. (2003). Mass spectrometry-based proteomics. *Nature* **422** 198–207.

ALEXANDROV, T. and BARTELS, A. (2013). Testing for presence of known and unknown molecules in imaging mass spectrometry. *Bioinformatics* **29** 2335–2342.

ALEXANDROV, T. and KOBARG, J. H. (2011). Efficient spatial segmentation of large imaging mass spectrometry datasets with spatially aware clustering. *Bioinformatics* **13** i230–i238.

ALEXANDROV, T., BECKER, M., DEININGER, S.-O., ERNST, G., WEHDER, L., GRASMAIR, M., VON EGGELING, F., THIELE, H. and MAASS, P. (2010). Spatial segmentation of imaging mass spectrometry data with edge-preserving image denoising and clustering. *J. Proteome Res.* **9** 6535–6546.

Key words and phrases. Proteomics, mass spectrometry data, high-dimensional, binary data, MALDI-IMS, unsupervised feature extraction.

ALEXANDROV, T., CHERNYAVSKY, I., BECKER, M., VON EGGELING, F. and NIKOLENKO, S. (2013). Analysis and interpretation of imaging mass spectrometry data by clustering mass-to-charge images according to their spatial similarity. *Analytical Chemistry* **85** 11189–11195.

AMERICA, A. H. and CORDEWENER, J. H. (2008). Comparative LC-MS: A landscape of peaks and valleys. *Proteomics* **8** 731–749.

AOKI, Y., TOYAMA, A., SHIMADA, T., SUGITA, T., AOKI, C., UMINO, Y., SUZUKI, A., AOKI, D., DAIGO, Y., NAKAMURA, Y. et al. (2007). A novel method for analyzing formalin-fixed paraffin embedded (FFPE) tissue sections by mass spectrometry imaging. *Proceedings of the Japan Academy. Series B, Physical and Biological Sciences* **83** 205–214.

BONNEL, D., LONGUESPEE, R., FRANCK, J., ROUDBARAKI, M., GOSSET, P., DAY, R., SALZET, M. and FOURNIER, I. (2011). Multivariate analyses for biomarkers hunting and validation through on-tissue bottom-up or in-source decay in MALDI-MSI: Application to prostate cancer. *Anal. Bioanal. Chem.* **401** 149–165.

CASADONTE, R. and CAPRIOLI, R. M. (2011). Proteomic analysis of formalin-fixed paraffin-embedded tissue by MALDI imaging mass spectrometry. *Nat. Protoc.* **6** 1695–1709.

CORNETT, D. S., REYZER, M. L., CHAURAND, P. and CAPRIOLI, R. M. (2007). MALDI imaging mass spectrometry: Molecular snapshots of biochemical systems. *Nat. Methods* **4** 828–833.

DEININGER, S.-O., EBERT, M. P., FÜTTERER, A., GERHARD, M. and RÖCKEN, C. (2008). MALDI imaging combined with hierarchical clustering as a new tool for the interpretation of complex human cancers. *J. Proteome Res.* **7** 5230–5236. PMID: 19367705.

DEININGER, S.-O., CORNETT, D. S., PAAPE, R., BECKER, M., PINEAU, C., RAUSER, S., WALCH, A. and WOLSKI, E. (2011). Normalization in MALDI-TOF imaging datasets of proteins: Practical considerations. *Anal. Bioanal. Chem.* **401** 167–181.

DEUTSKENS, F., YANG, J. and CAPRIOLI, R. M. (2011). High spatial resolution imaging mass spectrometry and classical histology on a single tissue section. *J. Mass Spectrom.* **46** 568–571.

DU, P., KIBBE, W. A. and LIN, S. M. (2006). Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics* **22** 2059–2065.

GARDEN, R. W. and SWEEDLER, J. V. (2000). Heterogeneity within MALDI samples as revealed by mass spectrometric imaging. *Analytical Chemistry* **72** 30–36.

GARDNER, M. (1970). Mathematical games: The fantastic combinations of John Conway's new solitaire game "life". *Scientific American* **223** 120–123.

GESSEL, M. M., NORRIS, J. L. and CAPRIOLI, R. M. (2014). MALDI imaging mass spectrometry: Spatial molecular analysis to enable a new age of discovery. *Journal of Proteomics* **107** 71–82. Special Issue: 20 in memory of Vitaliano Pallini.

GORZOLKA, K. and WALCH, A. (2014). November. MALDI mass spectrometry imaging of formalin-fixed paraffin-embedded tissues in clinical research. *Histology and Histopathology* **29** 1365–1376.

GRAY, L. (2003). A mathematician looks at S. Wolfram's new kind of science. *Notices Amer. Math. Soc.* **50** 200–211. MR1951106

GROSECLOSE, M. R., ANDERSSON, M., HARDESTY, W. M. and CAPRIOLI, R. M. (2006). Identification of proteins directly from tissue: In situ tryptic digestions coupled with imaging mass spectrometry. *J. Mass Spectrom.* **42** 254–262.

GROSECLOSE, M. R., MASSION, P. P., CHAURAND, P. and CAPRIOLI, R. M. (2008). High-throughput proteomic analysis of formalin-fixed paraffin-embedded tissue microarrays using maldi imaging mass spectrometry. *Proteomics* **8** 3715–3724.

GUSTAFSSON, O. J. R. (2012). Molecular characterization of metastatic ovarian cancer by MALDI imaging mass spectrometry. Ph.D. thesis, School of Molecular and Biomedical Science, Univ. Adelaide.

GUSTAFSSON, J. O. R., OEHLER, M. K., RUSZKIEWICZ, A., MCCOLL, S. R. and HOFFMANN, P. (2011). MALDI imaging mass spectrometry (MALDI-IMS)—application of spatial proteomics for ovarian cancer classification and diagnosis. *Int. J. Mol. Sci.* **12** 773–794.

GUSTAFSSON, J. O., EDDES, J. S., MEDING, S., KOUDELKA, T., OEHLER, M. K., MCCOLL, S. R. and HOFFMANN, P. (2012). Internal calibrants allow high accuracy peptide matching between MALDI imaging MS and LC-MS/MS. *Journal of Proteomics* **75** 5093–5105. Special Issue: Imaging Mass Spectrometry: A Users Guide to a New Technique for Biological and Biomedical Research.

GYGI, S. P., CORTHALS, G. L., ZHANG, Y., ROCHON, Y. and AEBERSOLD, R. (2000). Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc. Natl. Acad. Sci. USA* **97** 9390–9395.

JACCARD, P. (1901). Distribution de la Flore Alpine: Dans le Bassin des dranses et dans quelques régions voisines. Rouge.

JEMAL, A., BRAY, F., CENTER, M. M., FERLAY, J., WARD, E. and FORMAN, D. (2011). Global cancer statistics. *CA: A Cancer Journal for Clinicians* **61** 69–90.

JONES, E. A., VAN REMOORTERE, A., VAN ZEIJL, R. J., HOGENDOORN, P. C., BOVÉE, J. V., DEELDER, A. M. and MCDONNELL, L. A. (2011). Multiple statistical analysis techniques corroborate intratumor heterogeneity in imaging mass spectrometry datasets of myxofibrosarcoma. *PloS One* **6** e24913.

JONES, E. A., DEININGER, S.-O., HOGENDOORN, P. C., DEELDER, A. M. and MCDONNELL, L. A. (2012). Imaging mass spectrometry statistical analysis. *Journal of Proteomics* **75** 4962–4989. Special Issue: Imaging Mass Spectrometry: A User's Guide to a New Technique for Biological and Biomedical Research.

KARPIEVITCH, Y. V., POLPITIYA, A. D., ANDERSON, G. A., SMITH, R. D. and DABNEY, A. R. (2010). Liquid chromatography mass spectrometry-based proteomics: Biological and technological aspects. *Ann. Appl. Stat.* **4** 1797–1823.

KOCH, I. (2013). *Analysis of Multivariate and High-Dimensional Data*. Cambridge Univ. Press, New York. MR3154467

KOENIG, T., MENZE, B. H., KIRCHNER, M., MONIGATTI, F., PARKER, K. C., PATTERSON, T., STEEN, J. J., HAMPRECHT, F. A. and STEEN, H. (2008). Robust prediction of the mascot score for an improved quality assessment in mass spectrometric proteomics. *J. Proteome Res.* **7** 3708–3717.

MEDING, S., MARTIN, K., GUSTAFSSON, O. J., EDDES, J. S., HACK, S., OEHLER, M. K. and HOFFMANN, P. (2012). Tryptic peptide reference data sets for MALDI imaging mass spectrometry on formalin-fixed ovarian cancer tissues. *J. Proteome Res.* **12** 308–315.

MORRIS, J. S. (2012). Statistical methods for proteomic biomarker discovery based on feature extraction or functional modeling approaches. *Stat. Interface* **5** 117–135. MR2896986

MORRIS, J. S. and CARROLL, R. J. (2006). Wavelet-based functional mixed models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 179–199. MR2188981

MORRIS, J. S., COOMBES, K. R., KOOMEN, J., BAGGERLY, K. A. and KOBAYASHI, R. (2005). Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics* **21** 1764–1775.

NORRIS, J. L., CORNETT, D. S., MOBLEY, J. A., ANDERSSON, M., SEELEY, E. H., CHAURAND, P. and CAPRIOLI, R. M. (2007). Processing MALDI mass spectra to improve mass spectral direct tissue analysis. *Int. J. Mass Spectrom. Ion Phys.* **260** 212–221.

ONG, S.-E. and MANN, M. (2005). Mass spectrometry-based proteomics turns quantitative. *Nature Chemical Biology* **1** 252–262.

RICCIARDELLI, C. and OEHLER, M. K. (2009). Diverse molecular pathways in ovarian cancer and their clinical significance. *Maturitas* **62** 270–275.

ROGOWSKA-WRZESINSKA, A., LE BIHAN, M.-C., THAYSEN-ANDERSEN, M. and ROEPSTORFF, P. (2013). 2D gels still have a niche in proteomics. *Journal of Proteomics* **88** 4–13.

SCHOBER, Y., GUENTHER, S., SPENGLER, B. and RÖMPP, A. (2012). Single cell matrix-assisted laser desorption/ionization mass spectrometry imaging. *Analytical Chemistry* **84** 6293–6297.

STEURER, S., BORKOWSKI, C., ODINGA, S., BUCHHOLZ, M., KOOP, C., HULAND, H., BECKER, M., WITT, M., TREDE, D., OMIDI, M. et al. (2013). MALDI mass spectrometric imaging based identification of clinically relevant signals in prostate cancer using large-scale tissue microarrays. *Int. J. Cancer* **133** 920–928.

STONE, G., CLIFFORD, D., GUSTAFSSON, J. O., MCCOLL, S. R. and HOFFMANN, P. (2012). Visualisation in imaging mass spectrometry using the minimum noise fraction transform. *BMC Research Notes* **5** 419.

TEKWE, C. D., CARROLL, R. J. and DABNEY, A. R. (2012). Application of survival analysis methodology to the quantitative analysis of LC-MS proteomics data. *Bioinformatics* **28** 1998–2003.

TOMASI, C. and MANDUCHI, R. (1998). Bilateral filtering for gray and color images. 839–846, cited by 2167.

WAND, M. P. and JONES, M. C. (1995). *Kernel Smoothing*. Chapman & Hall, London. MR1319818

WASINGER, V. C., CORDWELL, S. J., CERPA-POLJAK, A., YAN, J. X., GOOLEY, A. A., WILKINS, M. R., DUNCAN, M. W., HARRIS, R., WILLIAMS, K. L. and HUMPHERY-SMITH, I. (1995). Progress with gene-product mapping of the mollicutes: Mycoplasma genitalium. *Electrophoresis* **16** 1090–1094.

WILKINS, M. R., PASQUALI, C., APPEL, R. D., OU, K., GOLAZ, O., SANCHEZ, J.-C., YAN, J. X., GOOLEY, A. A., HUGHES, G., HUMPHERY-SMITH, I. et al. (1996). From proteins to proteomes: Large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Nature Biotechnology* **14** 61–65.

WINDERBAUM, L. J., KOCH, I., GUSTAFSSON, O., MEDING, S. and HOFFMANN, P. (2015a). Supplement to "Feature extraction for proteomics imaging mass spectrometry data." DOI:10.1214/15-AOAS870SUPPA.

WINDERBAUM, L. J., KOCH, I., GUSTAFSSON, O., MEDING, S. and HOFFMANN, P. (2015b). Supplement to "Feature extraction for proteomics imaging mass spectrometry data." DOI:10.1214/15-AOAS870SUPPB.

WINDERBAUM, L. J., KOCH, I., GUSTAFSSON, O., MEDING, S. and HOFFMANN, P. (2015c). Supplement to "Feature extraction for proteomics imaging mass spectrometry data." DOI:10.1214/15-AOAS870SUPPC.

WU, B., ABBOTT, T., FISHMAN, D., MCMURRAY, W., MOR, G., STONE, K., WARD, D., WILLIAMS, K. and ZHAO, H. (2003). Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics* **19** 1636–1643.

YU, W., WU, B., HUANG, T., LI, X., WILLIAMS, K. and ZHAO, H. (2006). Statistical methods in proteomics. In *Springer Handbook of Engineering Statistics* 623–638. Springer, Berlin.

The Annals of Applied Statistics
2015, Vol. 9, No. 4, 1997–2022
DOI: 10.1214/15-AOAS837

REGULARIZED BRAIN READING WITH SHRINKAGE AND SMOOTHING

BY LEILA WEHBE[*,1], AADITYA RAMDAS[*,2], REBECCA C. STEORTS[†,3] AND COSMA ROHILLA SHALIZI[‡,4]

University of California, Berkeley[*]*, Duke University*[†] *and Carnegie Mellon University*[‡]

Functional neuroimaging measures how the brain responds to complex stimuli. However, sample sizes are modest, noise is substantial, and stimuli are high dimensional. Hence, direct estimates are inherently imprecise and call for regularization. We compare a suite of approaches which regularize via *shrinkage*: ridge regression, the elastic net (a generalization of ridge regression and the lasso), and a hierarchical Bayesian model based on small area estimation (SAE). We contrast regularization with *spatial smoothing* and combinations of smoothing and shrinkage. All methods are tested on functional magnetic resonance imaging (fMRI) data from multiple subjects participating in two different experiments related to reading, for both predicting neural response to stimuli and decoding stimuli from responses. Interestingly, when the regularization parameters are chosen by cross-validation independently for every voxel, low/high regularization is chosen in voxels where the classification accuracy is high/low, indicating that the regularization intensity is a good tool for identification of relevant voxels for the cognitive task. Surprisingly, all the regularization methods work about equally well, suggesting that beating basic smoothing and shrinkage will take not only clever methods, but also careful modeling.

REFERENCES

ABBOTT, L. F. and SEJNOWSKI, T. J. EDS. (1998). *Neural Codes and Distributed Representations*: *Foundations of Neural Computation*. MIT Press, Cambridge, MA.

ASHBURNER, J., BARNES, G., CHEN, C.-C., DAUNIZEAU, J., FLANDIN, G., FRISTON, K., KIEBEL, S., KILNER, J., LITVAK, V., MORAN, R., PENNY, W., ROSA, M., STEPHAN, K., GITELMAN, D., HENSON, R., HUTTON, C., GLAUCHE, V., MATTOUT, J. and PHILLIPS, C. (2008). SPM8 Manual. Functional Imaging Laboratory, Wellcome Trust Centre for Neuroimaging, Institute of Neurology, UCL.

ASHBY, F. G. (2011). *Statistical Analysis of FMRI Data*. MIT Press, Cambridge, MA.

BALLARD, D. H., ZHANG, Z. and RAO, R. P. N. (2002). Distributed synchrony: A probabilistic model of neural signalling. In *Probabilistic Models of the Brain*: *Perception and Neural Function* (R. P. N. Rao, B. A. Olshausen and M. S. Lewicki, eds.). *Neural Information Processing Series* 273–284. MIT Press, Cambridge, MA.

BRODERICK, T., BOYD, N., WIBISONO, A., WILSON, A. C. and JORDAN, M. I. (2013). Streaming variational Bayes. In *Advances in Neural Information Processing Systems* 26 [*NIPS* 2013] (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger, eds.) 1727–1735.

Key words and phrases. fMRI, small area estimation, regularization, shrinkage, spatial smoothing.

CRISTIANINI, N. and SHAWE-TAYLOR, J. (2000). *An Introduction to Support Vector Machines*: *And Other Kernel-Based Learning Methods*. Cambridge Univ. Press, Cambridge, MA.

DATTA, G. S., GHOSH, M., STEORTS, R. and MAPLES, J. (2011). Bayesian benchmarking with applications to small area estimation. *TEST* **20** 574–588. MR2864715

ENGEL, A. K., FRIES, P. and SINGER, W. (2001). Dynamic predictions: Oscillations and synchrony in top-down processing. *Nat. Rev., Neurosci.* **2** 704–716.

FRIEDMAN, J., HASTIE, T., TIBSHIRANI, R. and JIANG, H. (2010). Glmnet for Matlab. Statistics Department, Stanford Univ., Stanford.

FRIES, P. (2009). Neuronal gamma-band synchronization as a fundamental process in cortical computation. *Annu. Rev. Neurosci.* **32** 209–224.

FRISTON, K. J., ROTHSHTEIN, P., GENG, J. J., STERZER, P. and HENSON, R. N. (2010). A critique of functional localizers. In *Foundational Issues in Human Brain Mapping* (S. J. Hanson and M. Bunzl, eds.) 3–24. MIT Press, Cambridge, MA.

GENOVESE, C. R. (2000). A Bayesian time-course model for functional magnetic resonance imaging data. *J. Amer. Statist. Assoc.* **95** 691–703.

GOLUB, G. H., HEATH, M. and WAHBA, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21** 215–223. MR0533250

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

HAUFE, S., MEINECKE, F., GÖRGEN, K., DÄHNE, S., HAYNES, J.-D., BLANKERTZ, B. and BIESSMANN, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage* **87** 96–110.

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

JAAKKOLA, T. and HAUSSLER, D. (1999). Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems* 11 [*NIPS* 1998] (M. J. Kearns, S. A. Solla and D. A. Cohn, eds.) 487–493. MIT Press, Cambridge, MA.

KYUNG, M., GILL, J., GHOSH, M. and CASELLA, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Anal.* **5** 369–411. MR2719657

LEE, K.-J., JONES, G. L., CAFFO, B. S. and BASSETT, S. S. (2011). Spatial Bayesian Variable Selection Models on Functional Magnetic Resonance Imaging Time-Series Data. Preprint.

LOGOTHETIS, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature* **453** 869–878.

LOUIS, T. A. (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods. *J. Amer. Statist. Assoc.* **79** 393–398. MR0755093

MITCHELL, T. M., SHINKAREVA, S. V., CARLSON, A., CHANG, K.-M., MALAVE, V. L., MASON, R. A. and JUST, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science* **320** 1191–1195.

NASELARIS, T., KAY, K. N., NISHIMOTO, S. and GALLANT, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage* **56** 400–410.

NEISWANGER, W., WANG, C. and XING, E. (2013). Asymptotically exact, Embarrassingly Parallel MCMC. Preprint. Available at arXiv:1311.4780.

NORMAN, K. A., POLYN, S. M., DETRE, G. J. and HAXBY, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends Cogn. Sci.* **10** 424–430.

PALATUCCI, M., POMERLEAU, D., HINTON, G. E. and MITCHELL, T. M. (2009). Zero-shot learning with semantic output codes. In *Advances in Neural Information Processing Systems* 22 [*NIPS* 2009] (Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams and A. Culotta, eds.) 1410–1418. MIT Press, Cambridge, MA.

PARK, M., KOYEJO, O., GHOSH, J., POLDRACK, R. A. and PILLOW, J. W. (2013). Bayesian structure learning for functional neuroimaging. In 16*th International Conference on Artificial Intelligence and Statistics* (C. M. Carlvaho and P. Ravikumar, eds.) 489–497.

PEREIRA, F., MITCHELL, T. and BOTVINICK, M. (2009). Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage* **45** S199–S209.

PFEFFERMANN, D. (2013). New important developments in small area estimation. *Statist. Sci.* **28** 40–68. MR3075338

POLDRACK, R. A. (2008). The role of fMRI in cognitive neuroscience: Where do we stand? *Curr. Opin. Neurobiol.* **18** 223–227.

RAO, J. N. K. (2003). *Small Area Estimation*. Wiley, Hoboken, NJ. MR1953089

RIEKE, F., WARLAND, D., DE RUYTER VAN STEVENINCK, R. and BIALEK, W. (1999). *Spikes: Exploring the Neural Code*. MIT Press, Cambridge, MA. MR1983010

ROWLING, J. K. (2012). *Harry Potter and the Sorcerer's Stone*. Pottermore Limited, London.

SCOTT, S. L., BLOCKER, A. W. and BONASSI, F. V. (2013). Bayes and big data: The consensus Monte Carlo algorithm. Presented at the "EFaBBayes 250" conference, 16 December 2013, Duke Univ.

SHEPHERD, G. M. (1994). *Neurobiology*, 3rd ed. Oxford Univ. Press, London.

SMITH, S. M. (2004). Overview of fMRI analysis. *Br. J. Radiol.* **77** S167–S175.

SUDRE, G., POMERLEAU, D., PALATUCCI, M., WEHBE, L., FYSHE, A., SALMELIN, R. and MITCHELL, T. (2012). Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage* **62** 451–463.

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

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.

VON LUXBURG, U. (2007). A tutorial on spectral clustering. *Stat. Comput.* **17** 395–416. MR2409803

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

WEHBE, L., MURPHY, B., TALUKDAR, P., FYSHE, A., RAMDAS, A. and MITCHELL, T. (2014). Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLoS ONE* **9** e112575.

WEHBE, L., RAMDAS, A., STEORTS, R. C. and SHALIZI, C. R. (2015). Supplement to "Regularized brain reading with shrinkage and smoothing." DOI:10.1214/15-AOAS837SUPP.

YARKONI, T., POLDRACK, R. A., NICHOLS, T. E., VAN ESSEN, D. C. and WAGER, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods* **8** 665–670.

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

The Annals of Applied Statistics
2015, Vol. 9, No. 4, 2023–2050
DOI: 10.1214/15-AOAS863

EXTREMES ON RIVER NETWORKS[1]

BY PEIMAN ASADI*, ANTHONY C. DAVISON† AND SEBASTIAN ENGELKE*,†

Université de Lausanne and Ecole Polytechnique Fédérale de Lausanne†*

Max-stable processes are the natural extension of the classical extreme-value distributions to the functional setting, and they are increasingly widely used to estimate probabilities of complex extreme events. In this paper we broaden them from the usual situation in which dependence varies according to functions of Euclidean distance to situations in which extreme river discharges at two locations on a river network may be dependent because the locations are flow-connected or because of common meteorological events. In the former case dependence depends on river distance, and in the second it depends on the hydrological distance between the locations, either of which may be very different from their Euclidean distance. Inference for the model parameters is performed using a multivariate threshold likelihood, which is shown by simulation to work well. The ideas are illustrated with data from the upper Danube basin.

REFERENCES

ASADI, P., DAVISON, A. C. and ENGELKE, S. (2015). Supplement to "Extremes on river networks." DOI:10.1214/15-AOAS863SUPP.

BIENVENÜE, A. and ROBERT, C. (2014). Likelihood based inference for high-dimensional extreme value distributions. Available at http://arxiv.org/abs/1403.0065.

BLANCHET, J. and DAVISON, A. C. (2011). Spatial modeling of extreme snow depth. *Ann. Appl. Stat.* **5** 1699–1725. MR2884920

BÖHM, O. and WETZEL, K.-F. (2006). Flood history of the Danube tributaries Lech and Isar in the Alpine foreland of Germany. *Hydrological Sciences Journal* **51** 784–798.

BROWN, B. M. and RESNICK, S. I. (1977). Extreme values of independent stochastic processes. *J. Appl. Probab.* **14** 732–739. MR0517438

CHANDLER, R. E. and BATE, S. (2007). Inference for clustered data using the independence log-likelihood. *Biometrika* **94** 167–183. MR2367830

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

COLES, S., HEFFERNAN, J. and TAWN, J. (1999). Dependence measures for extreme value analyses. *Extremes* **2** 339–365.

COLES, S. G. and TAWN, J. A. (1991). Modelling extreme multivariate events. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **53** 377–392. MR1108334

COOLEY, D., NAVEAU, P. and PONCET, P. (2006). Variograms for spatial max-stable random fields. In *Dependence in Probability and Statistics* (P. Bertail, P. Soulier and P. Doukhan, eds.). *Lecture Notes in Statist.* **187** 373–390. Springer, New York. MR2283264

CRESSIE, N., FREY, J., HARCH, B. and SMITH, M. (2006). Spatial prediction on a river network. *J. Agric. Biol. Environ. Stat.* **11** 127–150.

Key words and phrases. Extremal coefficient, hydrological distance, max-stable process, network dependence, threshold-based inference, upper Danube basin.

DAVISON, A. C. and GHOLAMREZAEE, M. M. (2012). Geostatistics of extremes. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **468** 581–608. MR2874052

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

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

DOMBRY, C., ENGELKE, S. and OESTING, M. (2016). Exact simulation of max-stable processes. *Biometrika* **103**. To appear.

EINMAHL, J., KIRILIOUK, A., KRAJINA, A. and SEGERS, J. (2015). An M-estimator of spatial tail dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77**. To appear.

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

ENGELKE, S., KABLUCHKO, Z. and SCHLATHER, M. (2011). An equivalent representation of the Brown–Resnick process. *Statist. Probab. Lett.* **81** 1150–1154. MR2803757

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

HEFFERNAN, J. E. and TAWN, J. A. (2004). A conditional approach for multivariate extreme values. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 497–546. MR2088289

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

HUSER, R., DAVISON, A. C. and GENTON, M. G. (2014). A comparative study of parametric estimators for multivariate extremes. *Extremes*. Under review.

HÜSLER, J. and REISS, R.-D. (1989). Maxima of normal random vectors: Between independence and complete dependence. *Statist. Probab. Lett.* **7** 283–286. MR0980699

KABLUCHKO, Z. (2011). Extremes of independent Gaussian processes. *Extremes* **14** 285–310. MR2824498

KABLUCHKO, Z., SCHLATHER, M. and DE HAAN, L. (2009). Stationary max-stable fields associated to negative definite functions. *Ann. Probab.* **37** 2042–2065. MR2561440

KALLACHE, M., RUST, H. W., LANGE, H. and KROPP, J. P. (2010). Extreme value analysis considering trends: Application to discharge data of the Danube river basin. In *Extremis*: *Disruptive Events and Trends in Climate and Hydrology* (J. Kropp and H. Schellnhuber, eds.) 167–184. Springer, Berlin.

KATZ, R. W., PARLANGE, M. B. and NAVEAU, P. (2002). Statistics of extremes in hydrology. *Advances in Water Resources* **25** 1287–1304.

KEEF, C., SVENSSON, C. and TAWN, J. A. (2009). Spatial dependence in extreme river flows and precipitation for Great Britain. *Journal of Hydrology* **378** 240–252.

KEEF, C., TAWN, J. A. and LAMB, R. (2013). Estimating the probability of widespread flood events. *Environmetrics* **24** 13–21. MR3042270

KEEF, C., TAWN, J. and SVENSSON, C. (2009). Spatial risk assessment for extreme river flows. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **58** 601–618. MR2750258

KUNDZEWICZ, Z. W., ULBRICH, U., BRÜCHER, T., GRACZYK, D., KRÜGER, A., LECKEBUSCH, G. C., MENZEL, L., PIŃSKWAR, I., RADZIEJEWSKI, M. and SZWED, M. (2005). Summer floods in central Europe–Climate change track? *Natural Hazards* **36** 165–189.

MERZ, R. and BLÖSCHL, G. (2005). Flood frequency regionalisation—Spatial proximity vs. catchment attributes. *Journal of Hydrology* **302** 283–306.

OESTING, M., KABLUCHKO, Z. and SCHLATHER, M. (2012). Simulation of Brown–Resnick processes. *Extremes* **15** 89–107. MR2891311

OPITZ, T. (2013). Extremal t processes: Elliptical domain of attraction and a spectral representation. *J. Multivariate Anal.* **122** 409–413. MR3189331

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

PALUTIKOF, J. P., BRABSON, B. B., LISTER, D. H. and ADCOCK, S. T. (1999). A review of methods to calculate extreme wind speeds. *Meteorol. Appl.* **6** 119–132.

RENARD, B. and LANG, M. (2007). Use of a Gaussian copula for multivariate extreme value analysis: Some case studies in hydrology. *Advances in Water Resources* **30** 897–912.

RESNICK, S. I. (1987). *Extreme Values, Regular Variation, and Point Processes. Applied Probability. A Series of the Applied Probability Trust* **4**. Springer, New York. MR0900810

ROOTZÉN, H. and TAJVIDI, N. (2006). Multivariate generalized Pareto distributions. *Bernoulli* **12** 917–930. MR2265668

SALVADORI, G. and DE MICHELE, C. (2010). Multivariate multiparameter extreme value models and return periods: A copula approach. *Water Resources Research* **46** W10501.

SCHLATHER, M. (2002). Models for stationary max-stable random fields. *Extremes* **5** 33–44. MR1947786

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

SKØIEN, J., MERZ, R. and BLÖSCHL, G. (2006). Top-kriging-geostatistics on stream networks. *Hydrol. Earth Syst. Sci.* **10** 277–287.

TAWN, J. A. (1988). An extreme-value theory model for dependent observations. *Journal of Hydrology* **101** 227–250.

THIBAUD, E. and OPITZ, T. (2015). Efficient inference and simulation for elliptical Pareto processes. *Biometrika* **102** 855–870.

VER HOEF, J. M. and PETERSON, E. E. (2010). A moving average approach for spatial statistical models of stream networks. *J. Amer. Statist. Assoc.* **105** 6–18. MR2757185

VER HOEF, J. M., PETERSON, E. and THEOBALD, D. (2006). Spatial statistical models that use flow and stream distance. *Environ. Ecol. Stat.* **13** 449–464. MR2297373

WADSWORTH, J. L. and TAWN, J. A. (2014). Efficient inference for spatial extreme value processes associated to log-Gaussian random functions. *Biometrika* **101** 1–15. MR3180654

WANG, Y. and STOEV, S. A. (2010). On the structure and representations of max-stable processes. *Adv. in Appl. Probab.* **42** 855–877. MR2779562

The Annals of Applied Statistics
2015, Vol. 9, No. 4, 2051
DOI: 10.1214/15-AOAS864

LETTER TO THE EDITOR[1]

BY MILAN STEHLÍK[*,†] AND PHILIPP HERMANN[†]

University of Valparaíso[*] and Johannes Kepler University Linz[†]

REFERENCES

GOERG, G. M. (2011). Lambert *W* random variables—A new family of generalized skewed distributions with applications to risk estimation. *Ann. Appl. Stat.* **5** 2197–2230. MR2884937

STEHLÍK, M. and HERMANN, P. (2015). Supplement to "Letter to the Editor." DOI:10.1214/15-AOAS864SUPP.

The Annals of Applied Statistics
2015, Vol. 9, No. 4, 2052–2072
DOI: 10.1214/15-AOAS865

IDENTIFYING HETEROGENEOUS TRANSGENERATIONAL DNA METHYLATION SITES VIA CLUSTERING IN BETA REGRESSION[1]

BY SHENGTONG HAN[*,2], HONGMEI ZHANG[*,3], GABRIELLE A. LOCKETT[†,2], NANDINI MUKHERJEE[*,2], JOHN W. HOLLOWAY[†,2] AND WILFRIED KARMAUS[*,4]

University of Memphis and University of Southampton* †

This paper explores the transgenerational DNA methylation pattern (DNA methylation transmitted from one generation to the next) via a clustering approach. Beta regression is employed to model the transmission pattern from parents to their offsprings at the population level. To facilitate this goal, an expectation maximization algorithm for parameter estimation along with a BIC criterion to determine the number of clusters is proposed. Applying our method to the DNA methylation data composed of 4063 CpG sites of 41 mother–father-infant triads, we identified a set of CpG sites in which DNA methylation transmission is dominated by fathers, while at a large number of CpG sites, DNA methylation is mainly maternally transmitted to the offspring.

REFERENCES

ARSHAD, S. H. and HIDE, D. W. (1992). Effect of environmental factors on the development of allergic disorders in infancy. *J. Allergy Clin. Immunol.* **90** 235–241.

ARSHAD, S. H., KARMAUS, W., RAZA, A., KURUKULAARATCHY, R. J., MATTHEWS, S. M., HOLLOWAY, J. W., SADEGHNEJAD, A., ZHANG, H., ROBERTS, G. and EWART, S. L. (2012). The effect of parental allergy on childhood allergic diseases depends on the sex of the child. *J. Allergy Clin. Immunol.* **130** 427–434.

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

BIELAWSKI, D. M., ZAHER, F. M., SVINARICH, D. M. and ABEL, E. L. (2002). Paternal alcohol exposure affects sperm cytosine methyltransferase messenger RNA levels. *Alcohol. Clin. Exp. Res.* **26** 347–351.

CICERO, T. J., ADAMS, M. L., GIORDANO, A., MILLER, B. T., O'CONNOR, L. and NOCK, B. (1991). Influence of morphine exposure during adolescence on the sexual maturation of male rats and the development of their offspring. *J. Pharmacol. Exp. Ther.* **256** 1086–1093.

FERRARI, S. L. P. and CRIBARI-NETO, F. (2004). Beta regression for modelling rates and proportions. *J. Appl. Stat.* **31** 799–815. MR2095753

HARTIGAN, J. A. and WONG, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **28** 100–108.

HE, F., LIDOW, I. A. and LIDOW, M. S. (2006). Consequences of paternal cocaine exposure in mice. *Neurotoxicol. Teratol.* **28** 198–209.

Key words and phrases. DNA Methylation transmission, EM, clustering, Beta regression.

HOUSEMAN, E. A., CHRISTENSEN, B., YEH, R.-F., MARSIT, C., KARAGAS, M., WRENSCH, M., NELSON, H., WIEMELS, J., ZHENG, S., WIENCKE, J. and KELSEY, K. (2008). Model-based clustering of DNA methylation array data: A recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC Bioinformatics* **9** 365.

JOHNSON, W. E., LI, C. and RABINOVIC, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8** 118–127.

KAATI, G., BYGREN, L. O., PEMBREY, M. and SJÖSTRÖM, M. (2007). Transgenerational response to nutrition, early life circumstances and longevity. *Eur. J. Hum. Genet.* **15** 784–790.

LEDIG, M., MISSLIN, R., VOGEL, E., HOLOWNIA, A., COPIN, J. C. and THOLEY, G. (1998). Paternal alcohol exposure: Developmental and behavioral effects on the offspring of rats. *Neuropharmacology* **37** 57–66.

LOCKETT, G. A. and HOLLOWAY, J. W. (2013). Genome-wide association studies in asthma; perhaps, the end of the beginning. *Curr. Opin. Allergy Clin. Immunol.* **13** 463–469.

LOCKETT, G. A., PATIL, V. K., SOTO-RAMIREZ, N., ZIYAB, A. H., HOLLOWAY, J. W. and KARMAUS, W. (2013). Epigenomics and allergic disease. *Epigenomics* **5** 685–699.

MACQUEEN, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability* (*Berkeley, Calif.*, 1965/66) 281–297. Univ. California Press, Berkeley, CA. MR0214227

MANOLIO, T. A., COLLINS, F. S., COX, N. J., GOLDSTEIN, D. B., HINDORFF, L. A., HUNTER, D. J., MCCARTHY, M. I., RAMOS, E. M., CARDON, L. R., CHAKRAVARTI, A., CHO, J. H., GUTTMACHER, A. E., KONG, A., KRUGLYAK, L., MARDIS, E., ROTIMI, C. N., SLATKIN, M., VALLE, D., WHITTEMORE, A. S., BOEHNKE, M., CLARK, A. G., EICHLER, E. E., GIBSON, G., HAINES, J. L., MACKAY, T. F. C., MCCARROLL, S. A. and VISSCHER, P. M. (2009). Finding the missing heritability of complex diseases. *Nature* **461** 747–753.

NESTOR, C. E., BARRENÄS, F., WANG, H., LENTINI, A., ZHANG, H., BRUHN, S., JÖRNSTEN, R., LANGSTON, M. A., ROGERS, G., GUSTAFSSON, M. and BENSON, M. (2014). DNA methylation changes separate allergic patients from healthy controls and may reflect altered CD4+ T-cell population structure. *PLOS Genetics* **10** e1004059.

OUKO, L. A., SHANTIKUMAR, K., KNEZOVICH, J., HAYCOCK, P., SCHNUGH, D. J. and RAMSAY, M. (2009). Effect of alcohol consumption on CpG methylation in the differentially methylated regions of H19 and IG-DMR in male gametes-implications for fetal alcohol spectrum disorders. *Alcohol. Clin. Exp. Res.* **33** 1615–1627.

PADMANABHAN, N., JIA, D., GEARY-JOO, C., WU, X., FERGUSON-SMITH, A. C., FUNG, E., BIEDA, M. C., SNYDER, F. F., GRAVEL, R. A., CROSS, J. C. and WATSONEMAIL, E. D. (2013). Mutation in folate metabolism causes epigenetic instability and transgenerational effects on development. *Cell* **155** 81–93.

PARK, H. S. and JUN, C. H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert Syst. Appl.* **36** 3336–3341.

PEMBREY, M. E., BYGREN, L. O., KAATI, G., EDVINSSON, S., NORTHSTONE, K., SJÖSTRÖM, M., GOLDING, J. and TEAM, T. A. S. (2006). Sex-specific, male-line transgenerational responses in humans. *Eur. J. Hum. Genet.* **14** 159–166.

QIN, L.-X. and SELF, S. G. (2006). The clustering of regression models method with applications in gene expression data. *Biometrics* **62** 526–533. MR2236835

RAKYAN, V. K., CHONG, S., CHAMP, M. E., CUTHBERT, P. C., MORGAN, H. D., LUU, K. V. K. and WHITELAW, E. (2003). Transgenerational inheritance of epigenetic states at the murine AxinFu allele occurs after maternal and paternal transmission. *Proc. Natl. Acad. Sci. USA* **100** 2538–2543.

ROMIEU, I., TORRENT, M., GARCIA-ESTEBAN, R., FERRER, C., RIBAS-FITÓ, N., ANTÓ, J. M. and SUNYER, J. (2007). Maternal fish intake during pregnancy and atopy and asthma in infancy. *Clilincal and Experimental Allergy* **37** 518–525.

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

SOTO-RAMIREZ, N., ARSHAD, S. H., HOLLOWAY, J. W., ZHANG, H., SCHAUBERGER, E., EWART, S., PATIL, V. and KARMAUS, W. (2013). The interaction of genetic variants and DNA methylation of the interleukin-4 receptor gene increase the risk of asthma at age 18 years. *Clinical Epigenetics* **5** 1–8.

SZYF, M. (2009). Epigenetics, DNA methylation, and chromatin modifying drugs. *Annu. Rev. Pharmacol. Toxicol.* **49** 243–264.

WANG, D., YAN, L., HU, Q., SUCHESTON, L. E., HIGGINS, M. J., AMBROSONE, C. B., JOHNSON, C. S., SMIRAGLIA, D. J. and LIU, S. (2012). IMA: An R package for high-throughput analysis of illumina's 450K infinium methylation data. *Bioinformatics* **28** 729–730.

YOUSEFI, M., KARMAUS, W., ZHANG, H., EWART, S., ARSHAD, H. and HOLLOWAY, J. W. (2013). The methylation of the LEPR/LEPROT genotype at the promoter and body regions influence concentrations of leptin in girls and BMI at age 18 years if their mother smoked during pregnancy. *International Journal of Molecular Epidemiology and Genetics* **4** 86–100.

ZHANG, H., TONG, X., HOLLOWAY, J. W., REZWAN, F. I., PATIL, V., RAY, M., EVERSON, T. M., SOTO-RAMÍREZ, N., ARSHAD, S. H. et al. (2014). The interplay of DNA methylation over time with Th2 pathway genetic variants on asthma risk and temporal asthma transition. *Clinical Epigenetics* **6** 8.

ZIYAB, A. H., KARMAUS, W., HOLLOWAY, J. W., ZHANG, H., EWART, S. and ARSHAD, S. H. (2012). DNA methylation of the filaggrin gene adds to the risk of eczema associated with loss-of-function variants. *J. Eur. Acad. Dermatol. Venereol.* **27** e420–e423.

The Annals of Applied Statistics
2015, Vol. 9, No. 4, 2073–2089
DOI: 10.1214/15-AOAS868

MODELING COMPETITION BETWEEN TWO PHARMACEUTICAL DRUGS USING INNOVATION DIFFUSION MODELS[1]

BY RENATO GUSEO AND CINZIA MORTARINO

University of Padova

The study of competition among brands in a common category is an interesting strategic issue for involved firms. Sales monitoring and prediction of competitors' performance represent relevant tools for management. In the pharmaceutical market, the diffusion of product *knowledge* plays a special role, different from the role it plays in other competing fields. This latent feature naturally affects the evolution of drugs' performances in terms of the number of packages sold. In this paper, we propose an innovation diffusion model that takes the spread of knowledge into account. We are motivated by the need of modeling competition of two antidiabetic drugs in the Italian market.

REFERENCES

ABRAMSON, G. and ZANETTE, D. H. (1998). Statistics of extinction and survival in Lotka–Volterra systems. *Phys. Rev. E* (3) **57** 4572–4577.

BASS, F. M. (1969). A new product growth model for consumer durables. *Management Science* **15** 215–227.

BASS, F., KRISHNAN, T. and JAIN, D. (1994). Why the Bass model fits without decision variables. *Marketing Science* **13** 203–223.

BEAUCHAMP, J. J. and CORNELL, R. G. (1966). Simultaneous nonlinear estimation. *Technometrics* **8** 319–326. MR0205364

BOSWIJK, H. P. and FRANSES, P. H. (2005). On the econometrics of the Bass diffusion model. *J. Bus. Econom. Statist.* **23** 255–268. MR2159678

CENTRONE, F., GOIA, A. and SALINELLI, E. (2007). Demographic processes in a model of innovation diffusion with dynamic market. *Technological Forecasting and Social Change* **74** 247–266.

FURLAN, C. and MORTARINO, C. (2012). Pleural mesothelioma: Forecasts of the death toll in the area of Casale Monferrato, Italy. *Stat. Med.* **31** 4114–4134. MR3041797

GUSEO, R. and GUIDOLIN, M. (2009). Modelling a dynamic market potential: A class of automata networks for diffusion of innovations. *Technological Forecasting and Social Change* **76** 806–820.

GUSEO, R. and MORTARINO, C. (2012). Sequential market entries and competition modelling in multi-innovation diffusions. *European J. Oper. Res.* **216** 658–667. MR2845865

GUSEO, R. and MORTARINO, C. (2014a). Multivariate nonlinear least squares: Robustness and efficiency of standard versus Beauchamp and Cornell methodologies. *Comput. Statist.* **29** 1609–1636. MR3279009

GUSEO, R. and MORTARINO, C. (2014b). Within-brand and cross-brand word-of-mouth for sequential multi-innovation diffusions. *IMA J. Manag. Math.* **25** 287–311. MR3226506

GUSEO, R. and MORTARINO, C. (2015). Supplement to "Modeling competition between two pharmaceutical drugs using innovation diffusion models." DOI:10.1214/15-AOAS868SUPP.

Key words and phrases. Competition, innovation diffusion, dynamic market potential, communication network, nonlinear regression.

HANDCOCK, M. S. and GILE, K. J. (2010). Modeling social networks from sampled data. *Ann. Appl. Stat.* **4** 5–25. MR2758082

JHA, P. C., CHAUDHARY, K. and GUTPA, A. (2011). On the development of adoption of newer successive technologies using stochastic differential equation. In *IEEE International Conference on Industrial Engineering and Engineering Management* 1853–1858. IEEE, Singapore.

KIM, N., BRIDGES, E. and SRIVASTAVA, R. K. (1999). A simultaneous model for innovative product category sales diffusion and competitive dynamics. *International Journal of Research in Marketing* **16** 95–111.

KRISHNAN, T. V., BASS, F. M. and KUMAR, V. (2000). Impact of a late entrant on the diffusion of a new product/service. *Journal of Marketing Research* **XXXVII** 269–278.

LIBAI, B., MULLER, E. and PERES, R. (2009). The role of within-brand and cross-brand communications in competitive growth. *Journal of Marketing* **73** 19–34.

MEADE, N. and ISLAM, T. (2006). Modelling and forecasting the diffusion of innovation—A 25-year review. *International Journal of Forecasting* **22** 519–545.

MEYER, P. S. and AUSUBEL, J. H. (1999). Carrying capacity: A model with logistically varying limits. *Technological Forecasting and Social Change* **61** 209–214.

PERES, R., MULLER, E. and MAHAJAN, V. (2010). Innovation diffusion and new product growth models: A critical review and research directions. *International Journal of Research in Marketing* **27** 91–106.

PUTSIS, W. P. (1996). Temporal aggregation in diffusion models of first-time purchase: Does choice of frequency matter? *Technological Forecasting and Social Change* **51** 265–279.

SAVIN, S. and TERWIESCH, C. (2005). Optimal product launch times in a duopoly: Balancing life-cycle revenues with product cost. *Oper. Res.* **53** 26–47. MR2131098

SHARIF, M. N. and RAMANATHAN, K. (1981). Binomial innovation diffusion models with dynamic potential adopter population. *Technological Forecasting and Social Change* **20** 63–87.

SYDOW, J. and SCHREYÖGG, G. (2013). *Self-Reinforcing Processes in and Among Organizations*. Palgrave MacMillan, New York.

The Annals of Applied Statistics
2015, Vol. 9, No. 4, 2090–2109
DOI: 10.1214/15-AOAS869

"VIRUS HUNTING" USING RADIAL DISTANCE WEIGHTED DISCRIMINATION[1]

BY JIE XIONG, D. P. DITTMER AND J. S. MARRON

University of North Carolina at Chapel Hill

Motivated by the challenge of using DNA-seq data to identify viruses in human blood samples, we propose a novel classification algorithm called "Radial Distance Weighted Discrimination" (or Radial DWD). This classifier is designed for binary classification, assuming one class is surrounded by the other class in very diverse radial directions, which is seen to be typical for our virus detection data. This separation of the 2 classes in multiple radial directions naturally motivates the development of Radial DWD. While classical machine learning methods such as the Support Vector Machine and linear Distance Weighted Discrimination can sometimes give reasonable answers for a given data set, their generalizability is severely compromised because of the linear separating boundary. Radial DWD addresses this challenge by using a more appropriate (in this particular case) spherical separating boundary. Simulations show that for appropriate radial contexts, this gives much better generalizability than linear methods, and also much better than conventional kernel based (nonlinear) Support Vector Machines, because the latter methods essentially use much of the information in the data for determining the shape of the separating boundary. The effectiveness of Radial DWD is demonstrated for real virus detection.

REFERENCES

ALIZADEH, F. and GOLDFARB, D. (2003). Second-order cone programming. *Math. Program.* **95** 3–51. MR1971381

BURGES, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.* **2** 955–974.

FAN, J. and LV, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica* **20** 101–148. MR2640659

GOLDSTEIN, D. B., ALLEN, A., KEEBLER, J., MARGULIES, E. H., PETROU, S., PETROVSKI, S. and SUNYAEV, S. (2013). Sequencing studies in human genetics: Design and interpretation. *Nat. Rev. Genet.* **14** 460–470.

GRADA, A. and WEINBRECHT, K. (2013). Next-generation sequencing: Methodology and application. *J. Invest. Dermatol.* **133** e11.

HALL, P., MARRON, J. S. and NEEMAN, A. (2005). Geometric representation of high dimension, low sample size data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 427–444. MR2155347

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. Springer, New York. MR2722294

JIANG, J., MARRON, J. S. and JIANG, X. (2009). Robust centroid based classification with minimum error rates for high dimension, low sample size data. *J. Statist. Plann. Inference* **139** 2571–2580. MR2523649

Key words and phrases. Virus hunting, nonlinear classification, high-dimension low-sample size data analysis, DNA sequencing.

JUNG, S. and MARRON, J. S. (2009). PCA consistency in high dimension, low sample size context. *Ann. Statist.* **37** 4104–4130. MR2572454

LIU, Y., HAYES, D. N., NOBEL, A. and MARRON, J. S. (2008). Statistical significance of clustering for high-dimension, low-sample size data. *J. Amer. Statist. Assoc.* **103** 1281–1293. MR2528840

MARRON, J. S., TODD, M. J. and AHN, J. (2007). Distance-weighted discrimination. *J. Amer. Statist. Assoc.* **102** 1267–1271. MR2412548

METZKER, M. L. (2010). Sequencing technologies—the next generation. *Nat. Rev. Genet.* **11** 31–46.

MWENIFUMBO, J. C. and MARRA, M. A. (2013). Cancer genome-sequencing study design. *Nat. Rev. Genet.* **14** 321–332.

QIAO, X., ZHANG, H. H., LIU, Y., TODD, M. J. and MARRON, J. S. (2010). Weighted distance weighted discrimination and its asymptotic properties. *J. Amer. Statist. Assoc.* **105** 401–414. MR2656058

REHM, H. L. (2013). Disease-targeted sequencing: A cornerstone in the clinic. *Nat. Rev. Genet.* **14** 295–300.

SCHÖLKOPF, B. and SMOLA, A. J. (2002). *Learning with Kernels*. MIT Press, Cambridge, MA.

SHAWE-TAYLOR, J. and CRISTIANINI, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge Univ. Press, Cambridge, MA.

SHEN, D., SHEN, H. and MARRON, J. S. (2013). Consistency of sparse PCA in high dimension, low sample size contexts. *J. Multivariate Anal.* **115** 317–333. MR3004561

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

TUTUNCU, R. H., TOH, K. C. and TODD, M. J. (2001). SDPT3—a MATLAB software package for semidefinite-quadratic-linear programming. Available at http://www.math.cmu.edu/users/reha/home.html.

VAPNIK, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer, New York. MR1367965

WORLD HEALTH ORGANIZATION WHO (2014). Middle East respiratory syndrome coronavirus (MERS-CoV) summary and literature update-as of 9 May 2014. Available at http://www.who.int/csr/disease/coronavirus_infections/MERS_CoV_Update_09_May_2014.pdf.

XIONG, J., DITTMER, D. P. and MARRON, J. S. (2015). Supplement to: "Virus hunting" using Radial Distance Weighted Discrimination. DOI:10.1214/15-AOAS869SUPP.

YATA, K. and AOSHIMA, M. (2013). PCA consistency for the power spiked model in high-dimensional settings. *J. Multivariate Anal.* **122** 334–354. MR3189327

The Annals of Applied Statistics
2015, Vol. 9, No. 4, 2110–2132
DOI: 10.1214/15-AOAS875

A STOCHASTIC SPACE-TIME MODEL FOR INTERMITTENT PRECIPITATION OCCURRENCES[1]

BY YING SUN AND MICHAEL L. STEIN

King Abdullah University of Science and Technology and University of Chicago

Modeling a precipitation field is challenging due to its intermittent and highly scale-dependent nature. Motivated by the features of high-frequency precipitation data from a network of rain gauges, we propose a threshold space-time t random field (tRF) model for 15-minute precipitation occurrences. This model is constructed through a space-time Gaussian random field (GRF) with random scaling varying along time or space and time. It can be viewed as a generalization of the purely spatial tRF, and has a hierarchical representation that allows for Bayesian interpretation. Developing appropriate tools for evaluating precipitation models is a crucial part of the model-building process, and we focus on evaluating whether models can produce the observed conditional dry and rain probabilities given that some set of neighboring sites all have rain or all have no rain. These conditional probabilities show that the proposed space-time model has noticeable improvements in some characteristics of joint rainfall occurrences for the data we have considered.

REFERENCES

AILLIOT, P., THOMPSON, C. and THOMSON, P. (2009). Space-time modelling of precipitation by using a hidden Markov model and censored Gaussian distributions. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **58** 405–426. MR2750013

AIYYER, A. R. and THORNCROFT, T. (2006). Climatology of vertical wind shear in the tropical Atlantic. *J. Climate* **19** 2969–2983.

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (*Tsahkadsor*, 1971) (B. N. Petrov and F. Csake, eds.) 267–281. Akadémiai Kiadó, Budapest. MR0483125

BÁRDOSSY, A. and PLATE, E. J. (1992). Space-time model for daily rainfall using atmospheric circulation patterns. *Water Resour. Res.* **28** 1247–1259.

BELL, T. L. (1987). A space-time stochastic model of rainfall for satellite remote-sensing studies. *J. Geophys. Res.* **92** 9631–9643.

BELL, T. L. and KUNDU, P. K. (1996). A study of the sampling error in satellite rainfall estimates using optimal averaging of data and a stochastic model. *J. Climate* **9** 1251–1268.

BELL, T. L. and KUNDU, P. K. (2003). Comparing satellite rainfall estimates with rain gauge data: Optimal strategies suggested by a spectral model. *J. Geophys. Res.* **108** 4121.

BERROCAL, V. J., RAFTERY, A. E. and GNEITING, T. (2008). Probabilistic quantitative precipitation field forecasting using a two-stage spatial model. *Ann. Appl. Stat.* **2** 1170–1193. MR2655654

COWPERTWAIT, P. S. P. (1994). A generalized point process model for rainfall. *Proc. Roy. Soc. London Ser. A* **447** 23–37. MR1303321

Key words and phrases. Binary random field, Gaussian random field, Monte Carlo methods, random scaling, spatio-temporal dependence, t random field.

COX, D. R. and ISHAM, V. (1988). A simple spatial-temporal model of rainfall. *Proc. Roy. Soc. London Ser. A* **415** 317–328. MR0932924

GLASBEY, C. A. and NEVISON, I. M. (1997). Rainfall modelling using a latent Gaussian variable. In *Modelling Longitudinal and Spatially Correlated Data*: *Methods*, *Applications*, *and Future Directions* (Gregoire, T. G., Brillinger, D. R., Diggle, P. J., Russek-Cohen, E., Warren, W. G. and Wolfinger, R. D., eds.) 233–242. *Lecture Notes in Statistics* **122**. Springer, New York.

HELGASON, H., PIPIRAS, V. and ABRY, P. (2011). Fast and exact synthesis of stationary multivariate Gaussian time series using circulant embedding. *Signal Process.* **91** 1123–1133.

HERNÁNDEZ, A., GUENNI, L. and SANSÓ, B. (2009). Extreme limit distribution of truncated models for daily rainfall. *Environmetrics* **20** 962–980. MR2838498

HUGHES, J. P. and GUTTORP, P. (1999). A non-homogeneous hidden Markov model for precipitation occurrence. *Appl. Stat.* **48** 15–30.

KATZ, R. W. (1977). Precipitation as a chain-dependent process. *J. Appl. Meteorol.* **16** 671–676.

KATZ, R. W. (1996). Use of conditional stochastic models to generate climate change scenarios. *Clim. Change* **32** 237–255.

KLEIBER, W., KATZ, R. W. and RAJAGOPALAN, B. (2012). Daily spatiotemporal precipitation simulation using latent and transformed Gaussian processes. *Water Resour. Res.* **48** W01523.

KUNDU, P. K. and SIDDANI, R. K. (2007). A new class of probability distributions for describing the spatial statistics of area-averaged rainfall. *J. Geophys. Res. D* **18113** 112.

KUNDU, P. K. and SIDDANI, R. K. (2011). Scale dependence of spatiotemporal intermittence of rain. *Water Resour. Res.* **47** 318–340.

LE CAM, L. (1961). A stochastic description of precipitation. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. III* (J. Newman, ed.) 165–186. Univ. California Press, Berkeley, CA. MR0135598

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

MARAUN, D. et al. (2010). Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Rev. Geophys.* **48** 3003.

MARSAN, D., SCHERTZER, D. and LOVEJOY, S. (1996). Causal space-time multi-fractal processes: Predictability and forecasting of rain fields. *J. Geophys. Res.* **101** 26333–26346.

OVER, T. M. and GUPTA, V. K. (1996). A space-time theory of mesoscale rainfall using random cascades. *J. Geophys. Res.* **101** 26319–26331.

R CORE TEAM (2013). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org/.

RICHARDSON, C. W. (1981). Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resour. Res.* **17** 182–190.

RICHARDSON, C. W. and WRIGHT, D. A. (1984). WGEN: A model for generating daily weather variables. USDA, ARS-8, NTIS, Springfield, VA.

RODRIGUEZ-ITURBE, I., COX, D. R. and ISHAM, V. (1987). Some models for rainfall based on stochastic point processes. *Proc. Roy. Soc. London Ser. A* **410** 269–288. MR0887878

RODRIGUEZ-ITURBE, I., COX, D. R. and ISHAM, V. (1988). A point process model for rainfall: Further developments. *Proc. Roy. Soc. London Ser. A* **417** 283–298. MR0952338

RØISLIEN, J. and OMRE, H. (2006). T-distributed random fields: A parametric model for heavy-tailed well-log data. *Math. Geol.* **38** 821–849.

SANSÓ, B. and GUENNI, L. (1999). Venezuelan rainfall data analysis using a Bayesian space-time model. *J. R. Stat. Soc. Ser. C Appl. Stat.* **48** 345–362.

SIGRIST, F., KÜNSCH, H. R. and STAHEL, W. A. (2012). A dynamic nonstationary spatio-temporal model for short term prediction of precipitation. *Ann. Appl. Stat.* **6** 1452–1477. MR3058671

STEIN, M. L. (1992). Prediction and inference for truncated spatial data. *J. Comput. Graph. Statist.* **1** 91–110.

STEIN, M. L. (2005). Statistical methods for regular monitoring data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 667–687. MR2210686

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

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

SUN, Y. and GENTON, M. G. (2012). Adjusted functional boxplots for spatio-temporal data visualization and outlier detection. *Environmetrics* **23** 54–64. MR2873783

SUN, Y., GENTON, M. G. and NYCHKA, D. (2012). Exact fast computation of band depth for large functional datasets: How quickly can one million curves be ranked? *Stat* **1** 68–74.

SUN, Y., BOWMAN, K. P., GENTON, M. G. and TOKAY, A. (2015). A Matérn model of the spatial covariance structure of point rain rates. *Stoch. Environ. Res. Risk Assess.* **29** 411–416.

TOKAY, A., BASHOR, P. G. and MCDOWELL, V. L. (2010). Comparison of rain gauge measurements in the mid-Atlantic region. *J. Hydrometeorol.* **11** 553–565.

WAYMIRE, E. D., GUPTA, V. K. and RODRÍGUEZ-ITURBE, I. (1984). Spectral theory of rainfall intensity at the meso-β scale. *Water Resour. Res.* **20** 1453–1465.

WILKS, D. S. (2010). Use of stochastic weather generators for precipitation downscaling. *Wiley Interdiscip. Rev.: Clim. Change* **1** 898–907.

WOOD, A. T. A. and CHAN, G. (1994). Simulation of stationary Gaussian processes in $[0, 1]^d$. *J. Comput. Graph. Statist.* **3** 409–432. MR1323050

ZHENG, X. and KATZ, R. W. (2008). Simulation of spatial dependence in daily rainfall using multisite generators. *Water Resour. Res.* **44** W09403.

ZHENG, X., RENWICK, J. and CLARK, A. (2010). Simulation of multisite precipitation using an extended chain-dependent process. *Water Resour. Res.* **46** W01504.

The Annals of Applied Statistics
2015, Vol. 9, No. 4, 2133–2152
DOI: 10.1214/15-AOAS877

CORRECTING FOR MEASUREMENT ERROR IN LATENT VARIABLES USED AS PREDICTORS[1]

BY LYNNE STEUERLE SCHOFIELD

Swarthmore College

This paper represents a methodological-substantive synergy. A new model, the Mixed Effects Structural Equations (MESE) model which combines structural equations modeling and item response theory, is introduced to attend to measurement error bias when using several latent variables as predictors in generalized linear models. The paper investigates racial and gender disparities in STEM retention in higher education. Using the MESE model with 1997 National Longitudinal Survey of Youth data, I find prior mathematics proficiency and personality have been previously underestimated in the STEM retention literature. Pre-college mathematics proficiency and personality explain large portions of the racial and gender gaps. The findings have implications for those who design interventions aimed at increasing the rates of STEM persistence among women and underrepresented minorities.

REFERENCES

AYERS, E. and JUNKER, B. (2008). IRT modeling of tutor performance to predict end-of-year exam scores. *Educ. Psychol. Meas.* **68** 972–987. MR2516788

BOLLEN, K. A. (1989). *Structural Equations with Latent Variables*. Wiley, New York. MR0996025

BORSBOOM, D., MELLENBERGH, G. J. and VAN HEERDEN, J. (2003). The theoretical status of latent variables. *Psychol. Rev.* **110** 203–219.

CHANG, M. J., EAGAN, M. K., LIN, M. H. and HURTADO, S. (2011). Considering the impact of racial stigmas and science identity: Persistence among biomedical and behavioral science aspirants. *J. High. Educ.* **82** 564–596.

COSTA, P. T., JR. and MCCRAE, R. R. (1992). *Revised NEO Personality Inventory* (*NEO-PI-R*) *and NEO Five-Factor Inventory* (*NEO-FFI*) *Manual*. Psychological Assessment Resources, Odessa, FL.

DAVENPORT, B. M. (1976). A comparison of the Peabody Individual Achievement Test, the Metropolitan Achievement Test, and the Otis-Lennon Mental Ability Test. *Psychol. Sch.* **13** 291–297.

DOMINICI, F., ZEGER, S. L. and SAMET, J. M. (2000). A measurement error model for time-series studies of air pollution and mortality. *Biostatistics* **1** 157–175.

DRESHER, A. (2006). Results from NAEP marginal estimation research. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

DUNSON, D. B. (2001). Commentary: Practical advantages of Bayesian analysis of epidemiologic data. *Am. J. Epidemiol.* **153** 1222–1226.

ESPINOSA, L. L. (2011). Pipelines and pathways: Women of color in undergraduate STEM majors and the college experiences that contribute to persistence. *Harv. Educ. Rev.* **81** 209–240.

Key words and phrases. Structural equations models, item response theory, STEM retention, higher education.

FELDER, R. M., FELDER, G. N. and DIETZ, E. J. (2002). The effects of personality type on engineering student performance and attitudes. *Journal of Engineering Education* **91** 3–17.

FOX, J.-P. and GLAS, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika* **66** 271–288. MR1836937

FULLER, W. A. (2006). *Measurement Error Models*. Wiley, Hoboken, NJ. MR2301581

GELMAN, A. and HILL, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge Univ. Press, New York.

GOSLING, S. D., RENTFROW, P. J. and SWANN, W. B., JR. (2003). A very brief measure of the big five personality domains. *J. Res. Pers.* **37** 504–528.

GRIFFITH, A. (2010). Persistence of women and minorities in STEM field majors: Is it the school that matters? *Econ. Educ. Rev.* **29** 911–922.

HAINING, R., LI, G., MAHESWARAN, R., BLANGIARDO, M., LAW, J., BEST, N. and RICHARDSON, S. (2010). Inference from ecological models: Estimating the relative risk of stroke from air pollution exposure using small area data. *Spat Spatiotemporal Epidemiol.* **1** 123–131.

HECKMAN, J. J., STIXRUD, J. and URZUA, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics* **24** 411–482.

HOLLAND, J. L. (1997). *Making Vocational Choices*: *A Theory of Vocational Personality and Work Environments*, 3rd ed. ed. Psychological Assessment Resources, Odessa.

HUSMAN, J., LYNCH, C., HILPERT, J. and DUGGAN, M. A. (2007). Validating measures of future time perspective for engineering students: Steps toward improving engineering education. In *Proceedings of the American Society for Engineering Education Annual Conference & Exposition*. Honolulu, HI.

JÖRESKOG, K. G. and GOLDBERGER, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *J. Amer. Statist. Assoc.* **70** 631–639. MR0395057

JUNKER, B. W., SCHOFIELD, L. S. and TAYLOR, L. (2012). The use of cognitive ability measures as explanatory variables in regression analysis. *IZA Journal of Labor Economics* **1**.

KLEPPER, S. and LEAMER, E. E. (1984). Consistent sets of estimates for regressions with errors in all variables. *Econometrica* **52** 163–183. MR0729214

KORPERSHOEK, H., KUYPER, H. and VAN DER WERF, M. P. C. (2012). The role of personality in relation to gender differences in school subject choices in pre-university education. *Sex Roles* **67** 630–645.

KRISHNAKUMAR, J. and NADAR, A. (2008). On exact statistical properties of multidimensional indices based on principal components, factor analysis, MIMIC and structural equation models. *Social Indicators Research* **86** 481–496.

LESLIE, L. L., MCCLURE, G. T. and OAXACA, R. L. (1998). Women and minorities in science and engineering: A life sequence analysis. *J. High. Educ.* **69** 239–276.

LOCKWOOD, J. R. and MCCAFFREY, D. F. (2014). Correcting for test score measurement error in ANCOVA models for estimating treatment effects. *J. Educ. Behav. Stat.* **39** 22–52.

MAJOR, D. A., HOLLAND, J. M. and OBORN, K. L. (2012). The influence of proactive personality and coping on commitment to STEM majors. *Career Dev. Q.* **60** 16–24.

MALTESE, A. V. and TAI, R. H. (2011). Pipeline persistence: Examining the association of educational experiences with earned degrees in STEM among U.S. students. *Science Education* **95** 877–907.

MARKWARDT, F. C. (1998). Peabody Individual Achievement Test—revised manual, Minneapolis, MN, Pearson.

MISLEVY, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika* **56** 177–196.

MUTHÉN, L. K. and MUTHÉN, B. O. (1998–2011). *Mplus User's Guide*, 6th ed. Muthén & Muthén, Los Angeles, CA.

NATIONAL CENTER FOR EDUCATION STATISTICS (NCES) (2009). Students who study science, technology, engineering, and mathematics (STEM) in postsecondary education. NCES Stats in Brief (NCES 2009-161). Available at http://nces.ed.gov/pubs2009/2009161.pdf.

PALMER, R. T., DAVIS, R. J. and MARAMBA, D. (2011). *Racial and Ethnic Minority Student Success in STEM Education*: *ASHE Higher Education Report*. Wiley, New York.

PATZ, R. and JUNKER, B. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *J. Educ. Behav. Stat.* **24** 146–178.

PLUMMER, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the* 3*rd International Workshop on Distributed Statistical Computing*. Vienna, Austria.

RABE-HESKETH, S., SKRONDAL, A. and PICKLES, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika* **69** 167–190. MR2272445

RABE-HESKETH, S., SKRONDAL, A. and PICKLES, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *J. Econometrics* **128** 301–323. MR2189555

RICHARDSON, S. and GILKS, W. R. (1993). Conditional independence models for epidemiological studies with covariate measurement error. *Stat. Med.* **12** 1703–1722.

RICHARDSON, S., LEBLOND, L., JAUSSENT, I. and GREEN, P. J. (2002). Mixture models in measurement error problems, with reference to epidemiological studies. *J. Roy. Statist. Soc. Ser. A* **165** 549–566. MR1934339

RIEGLE-CRUMB, C., KING, B., GRODSKY, E. and MULLER, C. (2012). The more things change, the more they stay the same? Prior achievement fails to explain gender inequality in entry in STEM college majors over time. *Am. Educ. Res. J.* **49** 1048–1073.

SAMEJIMA, F. (1969). *Estimation of Latent Ability Using a Response Pattern of Graded Scores*. Psychometric Society, Richmond, VA.

SCHMITT, D. P., REALO, A., VORACEK, M. and ALLIK, J. (2008). Why can't a man be more like a woman? Sex differences in big five personality traits across 55 cultures. *J. Pers. Soc. Psychol.* **94** 168–182.

SCHOFIELD, L. S. (2008). Modeling measurement error when using cognitive test scores in social science research. Dissertation, Carnegie Mellon Univ.

SCHOFIELD, L. S. (2014). Measurement error in the AFQT in the NLSY79. *Econom. Lett.* **123** 262–265. MR3202249

SCHOFIELD, L. S., JUNKER, B., TAYLOR, L. J. and BLACK, D. A. (2015). Predictive Inference Using Latent Variables with Covariates. *Psychometrika* **80** 727–747. MR3392027

SEYMOUR, E. and HEWITT, N. M. (1997). *Talking About Leaving*: *Why Undergraduates Leave the Sciences*. Westview Press, Boulder, CO.

SKRONDAL, A. and RABE-HESKETH, S. (2004). *Generalized Latent Variable Modeling*: *Multilevel*, *Longitudinal*, *and Structural Equation Models*. Chapman & Hall/CRC, Boca Raton, FL. MR2059021

SPIEGELHALTER, D. J., THOMAS, A. and BEST, N. G. (2000). WinBUGS version 1.3 user manual. Medical Research Council Biostatistics Unit, Cambridge, MA.

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

STAIGER, D. and STOCK, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica* **65** 557–586. MR1445622

STEFANSKI, L. A. (2000). Measurement error models. *J. Amer. Statist. Assoc.* **95** 1353–1358. MR1825293

SULLINS, E. S., HERNANDEZ, D. and FULLER, C. (1995). Predicting who will major in a science discipline: Expectancy-value theory as part of an ecological model for studying academic communities. *J. Res. Sci. Teach.* **32** 99–119.

VAN DER LINDEN, W. J. and HAMBLETON, R. K., eds. (1997). *Handbook of Modern Item Response Theory*. Springer, New York. MR1601043

VAN LANGEN, A. (2005). *Unequal Participation in Mathematics and Science Education*. ITS, Nijmegen.

WALSH, W. B. (2001). The changing nature of the science of vocational psychology. *J. Vocat. Behav.* **59** 262–274.

WEINBERGER, C. J. (2012). Is the science and engineering workforce drawn from the far upper tail of the math ability distribution? Unpublished manuscript.

WIKOFF, R. L. (1978). Correlational and factor analysis of the peabody individual achievement test and the WISC-R. *J. Consult. Clin. Psychol.* **46** 322–325.

XIE, Y. and SHAUMAN, K. A. (2003). *Women in Science Career Processes and Outcomes*. Harvard Univ. Press, Cambridge, MA.

The Annals of Applied Statistics
2015, Vol. 9, No. 4, 2153–2178
DOI: 10.1214/15-AOAS879

BFLCRM: A BAYESIAN FUNCTIONAL LINEAR COX REGRESSION MODEL FOR PREDICTING TIME TO CONVERSION TO ALZHEIMER'S DISEASE[1]

BY EUNJEE LEE[*], HONGTU ZHU[2,*], DEHAN KONG[*], YALIN WANG[†], KELLY SULLIVAN GIOVANELLO[*], JOSEPH G. IBRAHIM[2,*] AND FOR THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE

University of North Carolina at Chapel Hill[*] and Arizona State University[†]

The aim of this paper is to develop a Bayesian functional linear Cox regression model (BFLCRM) with both functional and scalar covariates. This new development is motivated by establishing the likelihood of conversion to Alzheimer's disease (AD) in 346 patients with mild cognitive impairment (MCI) enrolled in the Alzheimer's Disease Neuroimaging Initiative 1 (ADNI-1) and the early markers of conversion. These 346 MCI patients were followed over 48 months, with 161 MCI participants progressing to AD at 48 months. The functional linear Cox regression model was used to establish that functional covariates including hippocampus surface morphology and scalar covariates including brain MRI volumes, cognitive performance (ADAS-Cog) and APOE-ε4 status can accurately predict time to onset of AD. Posterior computation proceeds via an efficient Markov chain Monte Carlo algorithm. A simulation study is performed to evaluate the finite sample performance of BFLCRM.

REFERENCES

ALBERT, M. S., DEKOSKY, S. T., DICKSON, D., DUBOIS, B., FELDMAN, H. H., FOX, N. C., GAMST, A., HOLTZMAN, D. M., JAGUST, W. J., PETERSEN, R. C., SNYDER, P. J., CARRILLO, M. C., THIES, B. and PHELPS, C. H. (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the national institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement*. **7** 270–279.

ANDERSON, N. D., EBERT, P. L., JENNINGS, J. M., GRADY, C. L., CABEZA, R. and GRAHAM, S. J. (2008). Recollection- and familiarity-based memory in healthy aging and amnestic mild cognitive impairment. *Neuropsychology* **22** 177–187.

APOSTOLOVA, L. G., DINOV, I. D., DUTTON, R. A., HAYASHI, K. M., TOGA, A. W., CUMMINGS, J. L. and THOMPSON, P. M. (2006a). 3D comparison of hippocampal atrophy in amnestic mild cognitive impairment and Alzheimer's disease. *Brain* **129** 2867–2873.

APOSTOLOVA, L. G., DUTTON, R. A., DINOV, I. D., HAYASHI, K. M., TOGA, A. W., CUMMINGS, J. L. and THOMPSON, P. M. (2006b). Conversion of mild cognitive impairment to Alzheimer disease predicted by hippocampal atrophy maps. *Arch. Neurol*. **63** 693–699.

BISWAS, A., DATTA, S., FINE, J. P. and SEGAL, M. R. (2008). *Statistical Advances in the Biomedical Sciences*: *Clinical Trials*, *Epidemiology*, *Survival Analysis*, *and Bioinformatics*. Wiley, Hoboken, NJ. MR2407832

Key words and phrases. Alzheimer's disease, functional principal component analysis, hippocampus surface morphology, mild cognitive impairment, proportional hazard model.

BRYANT, C., GIOVANELLO, K. S., IBRAHIM, J. G., CHANG, J., SHEN, D., PETERSON, B. S., ZHU, H. and ADNI (2013). Mapping the genetic variation of regional brain volumes as explained by all common SNPs from the ADNI study. *PloS One* **8** e71723.

CHEN, K. H., CHUAH, L. Y., SIM, S. K. and CHEE, M. W. (2010). Hippocampal region-specific contributions to memory performance in normal elderly. *Brain and Cognition* **72** 400–407.

COLOM, R., STEIN, J. L., RAJAGOPALAN, P., MARTINEZ, K., HERMEL, D., WANG, Y., ÁLVAREZ-LINERA, J., BURGALETA, M., QUIROGA, M., SHIH, P. C. and THOMPSON, P. M. (2013). Hippocampal structure and human cognition: Key role of spatial processing and evidence supporting the efficiency hypothesis in females. *Intelligence* **41** 129–140.

CORDER, E. H., SAUNDERS, A. M., STRITTMATTER, W. J., SCHMECHEL, D. E., GASKELL, P. C., SMALL, G., ROSES, A. D., HAINES, J. L. and PERICAK-VANCE, M. A. (1993). Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261** 921–923.

COSTAFREDA, S. G., DINOV, I. D., TU, Z., SHI, Y., LIU, C.-Y., KLOSZEWSKA, I., MECOCCI, P., SOININEN, H., TSOLAKI, M., VELLAS, B., WAHLUND, L.-O., SPENGER, C., TOGA, A. W., LOVESTONE, S. and SIMMONS, A. (2011). Automated hippocampal shape analysis predicts the onset of dementia in mild cognitive impairment. *Neuroimage* **56** 212–219.

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

CUI, Y., LIU, B., LUO, S., ZHEN, X., FAN, M., LIU, T., ZHU, W., PARK, M., JIANG, T., JIN, J. S. and ADNI (2011). Identification of conversion from mild cognitive impairment to Alzheimer's disease using multivariate predictors. *PloS One* **6** e21896.

DA, X., TOLEDO, J. B., ZEE, J., WOLK, D. A., XIE, S. X., OU, Y., SHACKLETT, A., PARMPI, P., SHAW, L., TROJANOWSKI, J. Q., DAVATZIKOS, C. and ALZHEIMER'S NEUROIMAGING INITIATIVE (2014). Integration and relative value of biomarkers for prediction of MCI to AD progression: Spatial patterns of brain atrophy, cognitive scores, APOE genotype and CSF biomarkers. *Neuroimage Clin.* **4** 164–173.

DESIKAN, R. S., CABRAL, H. J., FISCHL, B., GUTTMANN, C. R. G., BLACKER, D., HYMAN, B. T., ALBERT, M. S. and KILLIANY, R. J. (2009). Temporoparietal MR imaging measures of atrophy in subjects with mild cognitive impairment that predict subsequent diagnosis of Alzheimer disease. *American Journal of Neuroradiology* **30** 532–538.

DEVANAND, D. P., PRADHABAN, G., LIU, X., KHANDJI, A., DE SANTI, S., SEGAL, S., RUSINEK, H., PELTON, G. H., HONIG, L. S., MAYEUX, R., STERN, Y., TABERT, M. H. and DE LEON, M. J. (2007). Hippocampal and entorhinal atrophy in mild cognitive impairment: Prediction of Alzheimer disease. *Neurology* **68** 828–836.

DE LA TORRE, J. C. (2010). Alzheimer's disease is incurable but preventable. *J. Alzheimers Dis.* **20** 861–870.

DE LEON, M. J., GEORGE, A. E., GOLOMB, J., TARSHISH, C., CONVIT, A., KLUGER, A., DE SANTI, S., MC RAE, T., FERRIS, S. H., REISBERG, B., INCE, C., RUSINEK, H., BOBINSKI, M., QUINN, B., MILLER, D. C. and WISNIEWSKI, H. M. (1997). Frequency of hippocampal formation atrophy in normal aging and Alzheimer's disease. *Neurobiol. Aging* **18** 1–11.

DICKERSON, B. C., WOLK, D. A. and ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2013). Biomarker-based prediction of progression in MCI: Comparison of AD signature and hippocampal volume with spinal fluid amyloid-β and tau. *Front Aging Neurosci.* **5** 55.

DICKERSON, B. C., GONCHAROVA, I., SULLIVAN, M. P., FORCHETTI, C., WILSON, R. S., BENNETT, D. A., BECKETT, L. A. and DETOLEDO-MORRELL, L. (2001). MRI-derived entorhinal and hippocampal atrophy in incipient and very mild Alzheimer's disease. *Neurobiol. Aging* **22** 747–754.

FAN, J. and LV, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica* **20** 101–148. MR2640659

FAN, Y., BATMANGHELICH, N., CLARK, C. M., DAVATZIKOS, C. and ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2008). Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. *Neuroimage* **39** 1731–1743.

FENNEMA-NOTESTINE, C., HAGLER, D. J. JR, MCEVOY, L. K., FLEISHER, A. S., WU, E. H., KAROW, D. S., DALE, A. M. and ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2009). Structural MRI biomarkers for preclinical and mild Alzheimer's disease. *Hum. Brain Mapp.* **30** 3238–3253.

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

FLEMING, T. R. and HARRINGTON, D. P. (2011). *Counting Processes and Survival Analysis* **169**. Wiley, Hoboken, NJ.

FRANKÓ, E., JOLY, O. and ADNI (2013). Evaluating Alzheimer's disease progression using rate of regional hippocampal atrophy. *PloS One* **8** e71354.

GAUTHIER, S., REISBERG, B., ZAUDIG, M., PETERSEN, R. C., RITCHIE, K., BROICH, K., BELLEVILLE, S., BRODATY, H., BENNETT, D., CHERTKOW, H., CUMMINGS, J. L., DE LEON, M., FELDMAN, H., GANGULI, M., HAMPEL, H., SCHELTENS, P., TIERNEY, M. C., WHITEHOUSE, P. and WINBLAD, B. (2006). Mild cognitive impairment. *The Lancet* **367** 1262–1270.

GOMAR, J. J., BOBES-BASCARAN, M. T., CONEJERO-GOLDBERG, C., DAVIES, P., GOLDBERG, T. E. and ADNI (2011). Utility of combinations of biomarkers, cognitive markers, and risk factors to predict conversion from mild cognitive impairment to Alzheimer disease in patients in the Alzheimer's disease neuroimaging initiative. *Archives of General Psychiatry* **68** 961–969.

GRUNDMAN, M., SENCAKOVA, D., JACK, C. R., PETERSEN, R. C., KIM, H. T., SCHULTZ, A., WEINER, M. F., DECARLI, C., DEKOSKY, S. T., VAN DYCK, C., THOMAS, R. G., THAL, L. J. and ADCS (2002). Brain MRI hippocampal volume and prediction of clinical status in a mild cognitive impairment trial. *Journal of Molecular Neuroscience* **19** 23–27.

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models* **43**. CRC Press, Boca Raton, FL.

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

HUANG, J., SUN, T., YING, Z., YU, Y. and ZHANG, C.-H. (2013). Oracle inequalities for the LASSO in the Cox model. *Ann. Statist.* **41** 1142–1165. MR3113806

HUNG, H. and CHIANG, C.-T. (2010). Estimation methods for time-dependent AUC models with survival data. *Canad. J. Statist.* **38** 8–26. MR2676927

IBRAHIM, J. G., CHEN, M.-H. and KIM, S. (2008). Bayesian variable selection for the Cox regression model with missing covariates. *Lifetime Data Anal.* **14** 496–520. MR2464772

IBRAHIM, J. G., CHEN, M.-H. and SINHA, D. (2005). *Bayesian Survival Analysis*. Wiley Online Library.

JACK, C. R., PETERSEN, R. C., O'BRIEN, P. C. and TANGALOS, E. G. (1992). MR-based hippocampal volumetry in the diagnosis of Alzheimer's disease. *Neurology* **42** 183–183.

JACK, C. R., PETERSEN, R. C., XU, Y. C., WARING, S. C., O'BRIEN, P. C., TANGALOS, E. G., SMITH, G. E., IVNIK, R. J. and KOKMEN, E. (1997). Medial temporal atrophy on MRI in normal aging and very mild Alzheimer's disease. *Neurology* **49** 786–794.

JACK, C. R. JR., KNOPMAN, D. S., JAGUST, W. J., SHAW, L. M., AISEN, P. S., WEINER, M. W., PETERSEN, R. C. and TROJANOWSKI, J. Q. (2010). Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *The Lancet Neurology* **9** 119–128.

JAMES, G. M. (2002). Generalized linear models with functional predictors. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 411–432. MR1924298

KALBFLEISCH, J. D. and PRENTICE, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd ed. Wiley, Hoboken, NJ. MR1924807

KAYE, J. A., MOORE, M. M., DAME, A., QUINN, J., CAMICIOLI, R., HOWIESON, D., CORBRIDGE, E., CARE, B., NESBIT, G. and SEXTON, G. (2005). Asynchronous regional brain volume losses in presymptomatic to moderate AD. *J. Alzheimers Dis*. **8** 51–56.

KESSLAK, J. P., NALCIOGLU, O. and COTMAN, C. W. (1991). Quantification of magnetic resonance scans for hippocampal and parahippocampal atrophy in Alzheimer's disease. *Neurology* **41** 51–54.

LEE, E., ZHU, H., KONG, D., WANG, Y., GIOVANELLO, K. S., IBRAHIM, J. G. and ADNI (2016). Supplement to "BFLCRM: A Bayesian functional linear Cox regression model for predicting time to conversion to Alzheimer's disease." DOI:10.1214/15-AOAS879SUPP.

LI, J. and MA, S. (2013). *Survival Analysis in Medicine and Genetics*. Chapman & Hall/CRC, Boca Raton, FL.

LI, S., OKONKWO, O., ALBERT, M. and WANG, M.-C. (2013). Variation in variables that predict progression from MCI to AD dementia over duration of follow-up. *Am. J. Alzheimers Dis.* (*Columbia*) **2** 12–28.

LORENSEN, W. E. and CLINE, H. E. (1987). Marching cubes: A high resolution 3D surface construction algorithm. In *ACM Siggraph Computer Graphics* **21** 163–169. ACM, New York.

LUDERS, E., THOMPSON, P. M., KURTH, F., HONG, J. Y., PHILLIPS, O. R., WANG, Y., GUTMAN, B. A., CHOU, Y. Y., NARR, K. L. and TOGA, A. W. (2013). Global and regional alterations of hippocampal anatomy in long-term meditation practitioners. *Human Brain Mapping* **34** 3369–3375.

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. and TELLER, E. (2004). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* **21** 1087–1092.

MISRA, C., FAN, Y. and DAVATZIKOS, C. (2009). Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: Results from ADNI. *Neuroimage* **44** 1415–1422.

MIZUNO, K., WAKAI, M., TAKEDA, A. and SOBUE, G. (2000). Medial temporal atrophy and memory impairment in early stage of Alzheimer's disease: An MRI volumetric and memory assessment study. *Journal of the Neurological Sciences* **173** 18–24.

MONJE, M., THOMASON, M. E., RIGOLO, L., WANG, Y., WABER, D. P., SALLAN, S. E. and GOLBY, A. J. (2013). Functional and structural differences in the hippocampus associated with memory deficits in adult survivors of acute lymphoblastic leukemia. *Pediatric Blood & Cancer* **60** 293–300.

MURPHY, K. R., LANDAU, S. M., CHOUDHURY, K. R., HOSTAGE, C. A., SHPANSKAYA, K. S., SAIR, H. I., PETRELLA, J. R., WONG, T. Z., DORAISWAMY, P. M. and ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2013). Mapping the effects of ApoE4, age and cognitive status on 18F-florbetapir PET measured regional cortical patterns of beta-amyloid density and growth. *Neuroimage* **78** 474–480.

OKUIZUMI, K., ONODERA, O., TANAKA, H., KOBAYASHI, H., TSUJI, S., TAKAHASHI, H., OYANAGI, K., SEKI, K., TANAKA, M., NARUSE, S., MIYATAKE, T., MIZUSAWA, H. and KANAZAWA, I. (1994). ApoE-ε4 and early-onset Alzheimer's. *Nature Genetics* **7** 10–11.

PATENAUDE, B., SMITH, S. M., KENNEDY, D. N. and JENKINSON, M. (2011). A Bayesian model of shape and appearance for subcortical brain segmentation. *Neuroimage* **56** 907–922.

PENNANEN, C., KIVIPELTO, M., TUOMAINEN, S., HARTIKAINEN, P., HÄNNINEN, T., LAAKSO, M. P., HALLIKAINEN, M., VANHANEN, M., NISSINEN, A., HELKALA, E.-L., VAINIO, P., VANNINEN, R., PARTANEN, K. and SOININEN, H. (2004). Hippocampus and entorhinal cortex in mild cognitive impairment and early AD. *Neurobiol. Aging* **25** 303–310.

PERRI, R., SERRA, L., CARLESIMO, G. A. and CALTAGIRONE, C. (2007). Amnestic mild cognitive impairment: Difference of memory profile in subjects who converted or did not convert to Alzheimer's disease. *Neuropsychology* **21** 549–558.

PETERSEN, R. C., THOMAS, R. G., GRUNDMAN, M., BENNETT, D., DOODY, R., FERRIS, S., GALASKO, D., JIN, S., KAYE, J., LEVEY, A., PFEIFFER, E., SANO, M., VAN DYCK, C. H., THAL, L. J. and ALZHEIMER'S DISEASE COOPERATIVE STUDY GROUP (2005). Vitamin E and donepezil for the treatment of mild cognitive impairment. *N. Engl. J. Med.* **352** 2379–2388.

POULIN, S. P., DAUTOFF, R., MORRIS, J. C., BARRETT, L. F., DICKERSON, B. C. and ADNI (2011). Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity. *Psychiatry Research: Neuroimaging* **194** 7–13.

PRESTIA, A., CAROLI, A., VAN DER FLIER, W. M., OSSENKOPPELE, R., VAN BERCKEL, B., BARKHOF, F., TEUNISSEN, C. E., WALL, A. E., CARTER, S. F., SCHÖLL, M., CHOO, I. H., NORDBERG, A., SCHELTENS, P. and FRISONI, G. B. (2013). Prediction of dementia in MCI patients based on core diagnostic markers for Alzheimer disease. *Neurology* **80** 1048–1056.

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

REISS, P. T. and OGDEN, R. T. (2010). Functional generalized linear models with images as predictors. *Biometrics* **66** 61–69. MR2756691

RISACHER, S. L., SAYKIN, A. J., WES, J. D., SHEN, L., FIRPI, H. A. and MCDONALD, B. C. (2009). Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort. *Current Alzheimer Research* **6** 347–361.

ROBERTS, G. O., GELMAN, A. and GILKS, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.* **7** 110–120. MR1428751

ROSEN, W. G., MOHS, R. C. and DAVIS, K. L. (1984). A new rating scale for Alzheimer's disease. *Am. J. Psychiatry* **141** 1356–1364.

SAUNDERS, A. M., STRITTMATTER, W. J., SCHMECHEL, D., GEORGE-HYSLOP, P. S., PERICAK-VANCE, M. A., JOO, S. H., ROSI, B. L., GUSELLA, J. F., CRAPPER-MACLACHLAN, D. R., ALBERTS, M. J., HULETTE, C., CRAIN, B., GOLDGABER, D. and ROSES, A. D. (1993). Association of apolipoprotein E allele $\varepsilon 4$ with late-onset familial and sporadic Alzheimer's disease. *Neurology* **43** 1467–1467.

SCHELTENS, P. H., LEYS, D., BARKHOF, F., HUGLO, D., WEINSTEIN, H. C., VERMERSCH, P., KUIPER, M., STEINLING, M., WOLTERS, E. C. and VALK, J. (1992). Atrophy of medial temporal lobes on MRI in "probable" Alzheimer's disease and normal ageing: Diagnostic value and neuropsychological correlates. *Journal of Neurology, Neurosurgery & Psychiatry* **55** 967–972.

SHAW, L. M., VANDERSTICHELE, H., KNAPIK-CZAJKA, M., CLARK, C. M., AISEN, P. S., PETERSEN, R. C., BLENNOW, K., SOARES, H., SIMON, A., LEWCZUK, P., DEAN, R., SIEMERS, E., POTTER, W., LEE, V. M., TROJANOWSKI, J. Q. and ADNI (2009). Cerebrospinal fluid biomarker signature in Alzheimer's disease neuroimaging initiative subjects. *Annals of Neurology* **65** 403–413.

SHI, J., THOMPSON, P. M., GUTMAN, B., WANG, Y. and ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2013a). Surface fluid registration of conformal representation: Application to detect disease burden and genetic influence on hippocampus. *Neuroimage* **78** 111–134.

SHI, J., WANG, Y., CESCHIN, R., AN, X., LAO, Y., VANDERBILT, D., NELSON, M. D., THOMPSON, P. M., PANIGRAHY, A. and LEPORÉ, N. (2013b). A multivariate surface-based analysis of the putamen in premature newborns: Regional differences within the ventral striatum. *PloS One* **8** e66736.

SHI, J., LEPORÉ, N., GUTMAN, B. A., THOMPSON, P. M., BAXTER, L. C., CASELLI, R. L., WANG, Y. and ADNI (2014). Genetic influence of apolipoprotein E4 genotype on hippocampal morphometry: An $N = 725$ surface-based Alzheimer's disease neuroimaging initiative study. *Human Brain Mapping* **35** 3903–3918.

SINHA, D., CHEN, M.-H. and GHOSH, S. K. (1999). Bayesian analysis and model selection for interval-censored survival data. *Biometrics* **55** 585–590. MR1705161

SINHA, D., IBRAHIM, J. G. and CHEN, M.-H. (2003). A Bayesian justification of Cox's partial likelihood. *Biometrika* **90** 629–641. MR2006840

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

STRITTMATTER, W. J., SAUNDERS, A. M., SCHMECHEL, D., PERICAK-VANCE, M., ENGHILD, J., SALVESEN, G. S. and ROSES, A. D. (1993). Apolipoprotein E: High-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proc. Natl. Acad. Sci. USA* **90** 1977–1981.

TABERT, M. H., MANLY, J. J., LIU, X., PELTON, G. H., ROSENBLUM, S., JACOBS, M., ZAMORA, D., GOODKIND, M., BELL, K., STERN, Y. and DEVANAND, D. P. (2006). Neuropsychological prediction of conversion to Alzheimer disease in patients with mild cognitive impairment. *Arch. Gen. Psychiatry* **63** 916–924.

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

VEMURI, P., GUNTER, J. L., SENJEM, M. L., WHITWELL, J. L., KANTARCI, K., KNOPMAN, D. S., BOEVE, B. F., PETERSEN, R. C. and JACK JR, C. R. (2008). Alzheimer's disease diagnosis in individual subjects using structural MR images: Validation studies. *Neuroimage* **39** 1186–1197.

WANG, Y., ZHANG, J., GUTMAN, B., CHAN, T. F., BECKER, J. T., AIZENSTEIN, H. J., LOPEZ, O. L., TAMBURO, R. J., TOGA, A. W. and THOMPSON, P. M. (2010). Multivariate tensor-based morphometry on surfaces: Application to mapping ventricular abnormalities in HIV/AIDS. *NeuroImage* **49** 2141–2157.

WANG, Y., SONG, Y., RAJAGOPALAN, P., AN, T., LIU, K., CHOU, Y.-Y., GUTMAN, B., TOGA, A. W. and THOMPSON, P. M. (2011). Surface-based TBM boosts power to detect disease effects on the brain: An $N = 804$ ADNI study. *Neuroimage* **56** 1993–2010.

YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005). Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33** 2873–2903. MR2253106

YOUNG, J., MODAT, M., CARDOSO, M. J., MENDELSON, A., CASH, D. and OURSELIN, S. (2013). Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. *NeuroImage*: *Clinical* **2** 735–745.

ZHANG, D., SHEN, D. and ADNI (2012). Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers. *PloS One* **7** e33182.

The Annals of Applied Statistics
2015, Vol. 9, No. 4, 2179–2214
DOI: 10.1214/15-AOAS882

A FOCUSED INFORMATION CRITERION FOR GRAPHICAL MODELS IN FMRI CONNECTIVITY WITH HIGH-DIMENSIONAL DATA[1]

BY EUGEN PIRCALABELU*, GERDA CLAESKENS*, SARA JAHFARI† AND LOURENS J. WALDORP‡

KU Leuven, Vrije Universiteit Amsterdam† and University of Amsterdam‡*

Connectivity in the brain is the most promising approach to explain human behavior. Here we develop a focused information criterion for graphical models to determine brain connectivity tailored to specific research questions. All efforts are concentrated on high-dimensional settings where the number of nodes in the graph is larger than the number of samples. The graphical models may include autoregressive times series components, they can relate graphs from different subjects or pool data via random effects. The proposed method selects a graph with a small estimated mean squared error for a user-specified focus. The performance of the proposed method is assessed on simulated data sets and on a resting state functional magnetic resonance imaging (fMRI) data set where often the number of nodes in the estimated graph is equal to or larger than the number of samples.

REFERENCES

ABEGAZ, F. and WIT, E. (2013). Sparse time series chain graphical models for reconstructing genetic networks. *Biostatistics* **14** 586–599.

ACHARD, S., SALVADOR, R., WHITCHER, B., SUCKLING, J. and BULLMORE, E. (2006). A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. *J. Neurosci.* **26** 63–72.

ALLEN, E. A., DAMARAJU, E., PLIS, S. M., ERHARDT, E. B., EICHELE, T. and CALHOUN, V. D. (2014). Tracking whole-brain connectivity dynamics in the resting state. *Cereb. Cortex* **24** 663–676.

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

BASSETT, D. S., BULLMORE, E., VERCHINSKI, B. A., MATTAY, V. S., WEINBERGER, D. R. and MEYER-LINDENBERG, A. (2008). Hierarchical organization of human cortical networks in health and schizophrenia. *J. Neurosci.* **28** 9239–9248.

BUCKNER, R. L., ANDREWS-HANNA, J. R. and SCHACTER, D. L. (2008). The brain's default network: Anatomy, function, and relevance to disease. *Ann. N. Y. Acad. Sci.* **1124** 1–38.

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

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

Key words and phrases. fMRI connectivity, focused information criterion, model selection, Gaussian graphical model, penalization, high-dimensional data.

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

BUNEA, F., SHE, Y., OMBAO, H., GONGVATANA, A., DEVLIN, K. and COHEN, R. (2011). Penalized least squares regression methods and applications to neuroimaging. *Neuroimage* **55** 1519–1527.

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

CAMMOUN, L., GIGANDET, X., MESKALDJI, D., THIRAN, J. P., SPORNS, O., DO, K. Q., MAEDER, P., MEULI, R. and HAGMANN, P. (2012). Mapping the human connectome at multiple scales with diffusion spectrum MRI. *J. Neurosci. Methods* **203** 386–397.

CHAI, X. J., WHITFIELD-GABRIELI, S., SHINN, A. K., GABRIELI, J. D. E., CASTAÑÓN, A. N., MCCARTHY, J. M., COHEN, B. M. and ONGÜR, D. (2011). Abnormal medial prefrontal cortex resting-state connectivity in bipolar disorder and schizophrenia. *Neuropsychopharmacology* **36** 2009–2017.

CLAESKENS, G. (2012). Focused estimation and model averaging with penalization methods: An overview. *Stat. Neerl.* **66** 272–287. MR2955420

CLAESKENS, G. and HJORT, N. L. (2003). The focused information criterion. *J. Amer. Statist. Assoc.* **98** 900–945. MR2041482

CRAVEN, P. and WAHBA, G. (1978/79). Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31** 377–403. MR0516581

CRIBBEN, I., HARALDSDOTTIR, R., ATLAS, L. Y., WAGER, T. D. and LINDQUIST, M. A. (2012). Dynamic connectivity regression: Determining state-related changes in brain connectivity. *NeuroImage* **61** 907–920.

DAHLHAUS, R. and EICHLER, M. (2003). Causality and graphical models in time series analysis. In *Highly Structured Stochastic Systems*. *Oxford Statist. Sci. Ser.* **27** 115–144. Oxford Univ. Press, Oxford. MR2082408

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

DESHPANDE, G., SANTHANAM, P. and HU, X. (2011). Instantaneous and causal connectivity in resting state brain networks derived from functional MRI data. *Neuroimage* **54** 1043–1052.

DESIKAN, R. S., SÈGONNE, F., FISCHL, B., QUINN, B. T., DICKERSON, B. C., BLACKER, D., BUCKNER, R. L., DALE, A. M., MAGUIRE, R. P., HYMAN, B. T., ALBERT, M. S. and KILLIANY, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* **31** 968–980.

FAN, J., FENG, Y. and WU, Y. (2009). Network exploration via the adaptive lasso and SCAD penalties. *Ann. Appl. Stat.* **3** 521–541. MR2750671

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, T., YAO, L. and WU, X. (2012). Independent component analysis of the resting-state brain functional MRI study in adults with bipolar depression. In *Proceedings of* 2012 *International Conference on Complex Medical Engineering* 38–42. IEEE.

FOYGEL, R. and DRTON, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. In *Advances in Neural Information Processing Systems* 23 (J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel and A. Culotta, eds.) 604–612. MIT Press, Cambridge, MA.

FRANK, M. J. (2011). Computational models of motivated action selection in corticostriatal circuits. *Curr. Opin. Neurobiol.* **21** 381–386.

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

FRISTON, K. J., KAHAN, J., BISWAL, B. and RAZI, A. (2014). A DCM for resting state fMRI. *Neuroimage* **94** 396–407.

FU, W. J. (1998). Penalized regressions: The bridge versus the lasso. *J. Comput. Graph. Statist.* **7** 397–416. MR1646710

GAO, W. and TIAN, Z. (2010). Latent ancestral graph of structure vector autoregressive models. *J. Syst. Eng. Electron.* **21** 233–238.

GERHARD, S., DADUCCI, A., LEMKADDEM, A., MEULI, R., THIRAN, J.-P. and HAGMANN, P. (2011). The connectome viewer toolkit: An open source framework to manage, analyze, and visualize connectomes. *Front Neuroinform* **5** 3.

HAGMANN, P., CAMMOUN, L., GIGANDET, X., MEULI, R., HONEY, C. J., WEDEEN, J. V. and SPORNS, O. (2008). Mapping the structural core of human cerebral cortex. *PLoS Biology* **6** e159.

HONEY, C. J., SPORNS, O., CAMMOUN, L., GIGANDET, X., THIRAN, J. P., MEULI, R. and HAGMANN, P. (2009). Predicting human resting-state functional connectivity from structural connectivity. *Proc. Natl. Acad. Sci. USA* **106** 2035–2040.

HUMPHRIES, M. D. and GURNEY, K. (2008). Network "small-world-ness": A quantitative method for determining canonical network equivalence. *PLoS ONE* **3** e0002051.

HUMPHRIES, M. D., GURNEY, K. and PRESCOTT, T. J. (2006). The brainstem reticular formation is a small-world, not scale-free, network. *Proceedings of the Royal Society B* **273** 503–511.

HUNTER, D. R. and LI, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* **33** 1617–1642. MR2166557

ISODA, M. and HIKOSAKA, O. (2007). Switching from automatic to controlled action by monkey medial frontal cortex. *Nat. Neurosci.* **10** 240–248.

JAHFARI, S., WALDORP, L. J., VAN DEN WILDENBERG, W. P. M., SCHOLTE, H. S., RIDDERINKHOF, K. R. and FORSTMANN, B. U. (2011). Effective connectivity reveals important roles for both the hyperdirect (fronto-subthalamic) and the indirect (fronto-striatal-pallidal) fronto-basal ganglia pathways during response inhibition. *J. Neurosci.* **31** 6891–6899.

JAHFARI, S., VERBRUGGEN, F., FRANK, M. J., WALDORP, L. J., COLZATO, L., RIDDERINKHOF, K. R. and FORSTMANN, B. U. (2012). How preparation changes the need for top-down control of the basal ganglia when inhibiting premature actions. *J. Neurosci.* **32** 10870–10878.

JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I* 361–379. Univ. California Press, Berkeley, CA. MR0133191

JAMES, G. A., KELLEY, M. E., CRADDOCK, R. C., HOLTZHEIMER, P. E., DUNLOP, B., NEMEROFF, C. and HU, X. P. (2009). Exploratory structural equation modeling of resting-state fMRI: Applicability of group models to individual subjects. *Neuroimage* **45** 778–787.

JENKINSON, M. and SMITH, S. (2001). A global optimisation method for robust affine registration of brain images. *Med. Image Anal.* **5** 143–156.

JENKINSON, M., BANNISTER, P., BRADY, M. and SMITH, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* **17** 825–841.

KOLAR, M., SONG, L., AHMED, A. and XING, E. P. (2010). Estimating time-varying networks. *Ann. Appl. Stat.* **4** 94–123. MR2758086

KOYAMA, M. S., MARTINO, A. D., ZUO, X.-N., KELLY, C., MENNES, M., JUTAGIR, D. R., CASTELLANOS, F. X. and MILHAM, M. P. (2011). Resting-state functional connectivity indexes reading competence in children and adults. *J. Neurosci.* **31** 8617–8624.

KRISHNAMURTHY, V., AHIPAŞAOĞLU, S. D. and D'ASPREMONT, A. (2012). A pathwise algorithm for covariance selection. In *Optimization for Machine Learning* (S. Sra, S. Nowozin and S. J. Wright, eds.) 479–494. MIT Press, Cambridge, MA.

LAIRD, N., LANGE, N. and STRAM, D. (1987). Maximum likelihood computations with repeated measures: Application of the EM algorithm. *J. Amer. Statist. Assoc.* **82** 97–105. MR0883338

LAURITZEN, S. L. (1996). *Graphical Models. Oxford Statistical Science Series* **17**. The Clarendon Press, Oxford Univ. Press, New York. MR1419991

LEI, Y., TONG, L. and YAN, B. (2013). A mixed L2 norm regularized HRF estimation method for rapid event-related fMRI experiments. *Comput. Math. Methods Med.* **2013** 643129.

LEONARDI, N., RICHIARDI, J., GSCHWIND, M., SIMIONI, S., ANNONI, J. M., SCHLUEP, M. and VAN DE VILLE, D. (2013). Principal components of functional connectivity: A new approach to study dynamic brain connectivity during rest. *NeuroImage* **83** 937–950.

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

LI, X., ZHAO, T. and LIU, H. (2013). camel: Calibrated machine learning. R package version 0.2.0.

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

LIU, H. and WANG, L. (2012). TIGER: A tuning-insensitive approach for optimally estimating large undirected graphs. Technical report.

MAZUMDER, R. and HASTIE, T. (2012). The graphical lasso: New insights and alternatives. *Electron. J. Stat.* **6** 2125–2149. MR3020259

MCLACHLAN, G. J. and KRISHNAN, T. (2008). *The EM Algorithm and Extensions*, 2nd ed. Wiley, Hoboken, NJ. MR2392878

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

MOHAMMADI, A. and WIT, E. (2015). Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Anal.* **10** 109–138.

MOUSSA, M. N., STEEN, M. R., LAURIENTI, P. J. and HAYASAKA, S. (2012). Consistency of network modules in resting-state FMRI connectome data. *PLoS ONE* **7** e44428.

O'NEIL, E. B., HUTCHISON, R. M., MCLEAN, D. A. and KÖHLER, S. (2014). Resting-state fMRI reveals functional connectivity between face-selective perirhinal cortex and the fusiform face area related to face inversion. *Neuroimage* **92** 349–355.

PIRCALABELU, E., CLAESKENS, G. and WALDORP, L. (2015). A focused information criterion for graphical models. *Stat. Comput.* **25** 1071–1092. MR3401874

RAICHLE, M. E., MACLEOD, A. M., SNYDER, A. Z., POWERS, W. J., GUSNARD, D. A. and SHULMAN, G. L. (2001). A default mode of brain function. *Proc. Natl. Acad. Sci. USA* **98** 676–682.

RAVIKUMAR, P. D., RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2008). Model selection in Gaussian graphical models: High-dimensional consistency of l_1-regularized MLE. In *Proceedings of the* 22*nd Annual Conference on Neural Information Processing Systems* (D. Koller, D. Schuurmans, Y. Bengio and L. Bottou, eds.) 1329–1336. MIT Press, Cambridge, MA.

RIDDERINKHOF, K. R., ULLSPERGER, M., CRONE, E. A. and NIEUWENHUIS, S. (2004). The role of the medial frontal cortex in cognitive control. *Science* **306** 443–447.

RYALI, S., SUPEKAR, K., ABRAMS, D. A. and MENON, V. (2010). Sparse logistic regression for whole-brain classification of fMRI data. *Neuroimage* **51** 752–764.

RYALI, S., CHEN, T., SUPEKAR, K. and MENON, V. (2012). Estimation of functional connectivity in fMRI data using stability selection-based sparse partial correlation with elastic net penalty. *Neuroimage* **59** 3852–3861.

SCHEINBERG, K. and RISH, I. (2010). Learning sparse Gaussian Markov networks using a greedy coordinate ascent approach. In *Proceedings of the* 2010 *European Conference on Machine Learning and Knowledge Discovery in Databases*: *Part III* 196–212. Springer, Berlin.

SCHMIDT, M., NICULESCU-MIZIL, A. and MURPHY, K. (2007). Learning graphical model structure using ℓ_1-regularization paths. In *Proceedings of the* 22*nd National Conference on Artificial Intelligence* **2** 1278–1283. AAAI Press, Menlo Park, CA.

SMITH, S. M. (2002). Fast robust automated brain extraction. *Hum. Brain Mapp.* **17** 143–155.

SPORNS, O. and HONEY, C. J. (2006). Small worlds inside big brains. *Proc. Natl. Acad. Sci. USA* **103** 19219–19220.

THOMPSON, P. M., CANNON, T. D., NARR, K. L., VAN ERP, T., POUTANEN, V. P., HUTTUNEN, M., LÖNNQVIST, J., STANDERTSKJÖLD-NORDENSTAM, C. G., KAPRIO, J., KHALEDY, M., DAIL, R., ZOUMALAN, C. I. and TOGA, A. W. (2001). Genetic influences on brain structure. *Nat. Neurosci.* **4** 1253–1258.

WAINWRIGHT, M. J., RAVIKUMAR, P. and LAFFERTY, J. D. (2007). High-dimensional graphical model selection using ℓ_1-regularized logistic regression. In *Advances in Neural Information Processing Systems* 19 (B. Schölkopf, J. Platt and T. Hoffman, eds.) 1465–1472. MIT Press, Cambridge, MA.

WALDORP, L. J. (2009). Robust and unbiased variance of GLM coefficients for misspecified autocorrelation and hemodynamic response models in fMRI. *Int. J. Biomed. Imaging* **2009** 1–11.

WEEDA, W. D., WALDORP, L. J., CHRISTOFFELS, I. and HUIZENGA, H. M. (2010). Activated region fitting: A robust high-power method for fMRI analysis using parameterized regions of activation. *Hum. Brain Mapp.* **30** 2595–2605.

WINK, A. M. and ROERDINK, J. B. T. M. (2006). BOLD noise assumptions in fMRI. *Int. J. Biomed. Imaging* **2006** 1–11.

WITTEN, D. M., FRIEDMAN, J. H. and SIMON, N. (2011). New insights and faster computations for the graphical lasso. *J. Comput. Graph. Statist.* **20** 892–900. MR2878953

WOODWARD, N. D., ROGERS, B. and HECKERS, S. (2011). Functional resting-state networks are differentially affected in schizophrenia. *Schizophr. Res.* **130** 86–93.

WORSLEY, K. J. (2001). Statistical analysis of activation images. In *Functional MRI*: *An Introduction to Methods* (P. Jezzard, P. Matthews and S. M. Smith, eds.) 251–270. Oxford Univ. Press, London.

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

ZHANG, X. and LIANG, H. (2011). Focused information criterion and model averaging for generalized additive partial linear models. *Ann. Statist.* **39** 174–200. MR2797843

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

ZHOU, S., LAFFERTY, J. and WASSERMAN, L. (2010). Time varying undirected graphs. *Mach. Learn.* **80** 295–319. MR3108169

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

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

ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533. MR2435443

The Annals of Applied Statistics
2015, Vol. 9, No. 4, 2215–2236
DOI: 10.1214/15-AOAS860

ON THE ANALYSIS OF TUBERCULOSIS STUDIES WITH INTERMITTENT MISSING SPUTUM DATA[1]

BY DANIEL SCHARFSTEIN[*], ANDREA ROTNITZKY[†], MARIA ABRAHAM[‡], AIDAN MCDERMOTT[*], RICHARD CHAISSON[*] AND LAWRENCE GEITER[§]

Johns Hopkins University[*], Universidad Torcuato Di Tella[†], Statistics Collaborative[‡] and Otsuka Novel Products[§]

In randomized studies evaluating treatments for tuberculosis (TB), individuals are scheduled to be routinely evaluated for the presence of TB using sputum cultures. One important endpoint in such studies is the time of culture conversion, the first visit at which a patient's sputum culture is negative and remains negative. This article addresses how to draw inference about treatment effects when sputum cultures are intermittently missing on some patients. We discuss inference under a novel benchmark assumption and under a class of assumptions indexed by a treatment-specific sensitivity parameter that quantify departures from the benchmark assumption. We motivate and illustrate our approach using data from a randomized trial comparing the effectiveness of two treatments for adult TB patients in Brazil.

REFERENCES

AMERICAN THORACIC SOCIETY (2000). Diagnostic standards and classification of tuberculosis in adults and children. *Am. J. Respir. Crit. Care Med.* **161** 1376–1395.

BARNDORFF-NIELSEN, O. E. and COX, D. R. (1994). *Inference and Asymptotics*. Chapman & Hall, London. MR1317097

CONDE, M. B., EFRON, A., LOREDO, C., SOUZA, G. R. M. D., GRAÇA, N. P., CEZAR, M. C., RAM, M., CHAUDHARY, M. A., BISHAI, W. R., KRITSKI, A. L. and CHAISSON, R. E. (2009). Moxifloxacin versus ethambutol in the initial treatment of tuberculosis: A double-blind, randomised, controlled phase II trial. *Lancet* **373** 1183–1189.

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

EUROPEAN MEDICINES AGENCY, COMMITTEE FOR MEDICINAL PRODUCTS FOR HUMAN USE (2010). *Addendum to the Note for Guidance on Evaluation of Medicinal Products Indicated for Treatment of Bacterial Infections to Specifically Address the Clinical Development of New Agents to Treat Disease Due to Mycobacterium Tuberculosis*. European Medicines Agency, London.

GILL, R. D., VAN DER LAAN, M. J. and ROBINS, J. M. (1997). Coarsening at random: Characterizations, conjectures and counter-examples. In *Proceedings of the First Seattle Symposium in Biostatistics*: *Survival Analysis* (D. Y. Lin and T. R. Fleming, eds.) 255–294. Springer, Berlin.

HEITJAN, D. F. (1993). Ignorability and coarse data: Some biomedical examples. *Biometrics* **49** 1099–1109.

HEITJAN, D. F. (1994). Ignorability in general incomplete-data models. *Biometrika* **81** 701–708. MR1326420

Key words and phrases. Culture conversion, curse of dimensionality, exponential tilting, reverse-time hazard, sensitivity analysis.

HEITJAN, D. F. and RUBIN, D. B. (1991). Ignorability and coarse data. *Ann. Statist.* **19** 2244–2253. MR1135174

NELSON, D. R., ZEUZEM, S., ANDREONE, P., FERENCI, P., HERRING, R., JENSEN, D. M., MARCELLIN, P., POCKROS, P. J., RODRÍGUEZ-TORRES, M., ROSSARO, L. et al. (2012). Balapiravir plus peginterferon alfa-2a (40KD)/ribavirin in a randomized trial of hepatitis C genotype 1 patients. *Annals of Hepatology* **11** 15.

NEWEY, W. K. and MCFADDEN, D. (1994). Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, *Vol. IV* (R. F. Engle and D. L. McFadden, eds.) 2111–2245. North-Holland, Amsterdam. MR1315971

ROBINS, J. M. and GILL, R. D. (1997). Non-response models for the analysis of non-monotone ignorable missing data. *Stat. Med.* **16** 39–56.

The Annals of Applied Statistics
2015, Vol. 9, No. 4, 2237–2265
DOI: 10.1214/15-AOAS878

ASSESSING NONRESPONSE BIAS IN A BUSINESS SURVEY: PROXY PATTERN-MIXTURE ANALYSIS FOR SKEWED DATA

BY REBECCA ANDRIDGE AND KATHERINE JENNY THOMPSON

Ohio State University and U.S. Census Bureau

The Service Annual Survey (SAS) is a business survey conducted annually by the U.S. Census Bureau that collects aggregate and detailed revenues and expenses data. Typical of many business surveys, the SAS population is highly positively skewed, with large companies comprising a large proportion of the published totals. When alternative data are not available, missing data are handled with ratio imputation models that assume missingness is at random. We propose a proxy pattern-mixture (PPM) model that provides a simple framework for assessing nonresponse bias with respect to different nonresponse mechanisms. PPM models were first introduced in this context by Andridge and Little [*Journal of Official Statistics* **27** (2011) 153–180], but their model assumed the characteristic of interest and the predicted proxy have a bivariate normal distribution, conditional on the missingness indicator. Although often appropriate for large demographic surveys, the normality assumption is less justifiable for the highly skewed SAS data. We propose an alternative PPM model using a bivariate gamma distribution more appropriate for the SAS data. We compare the two PPM models through application to data from six years of data collection in three industries in the health care and transportation sectors of the SAS. Finally, we illustrate properties of the method through simulation.

REFERENCES

ANDRIDGE, R. R. and LITTLE, R. J. A. (2011). Proxy pattern-mixture analysis for survey nonresponse. *Journal of Official Statistics* **27** 153–180.

ANDRIDGE, R. R. and THOMPSON, K. J. (2015a). Using the fraction of missing information to identify auxiliary variables for imputation procedures via proxy pattern-mixture models. *Int. Stat. Rev.* **83** 472–492.

ANDRIDGE, R. R. and THOMPSON, K. J. (2015b). Supplement to "Assessing nonresponse bias in a business survey: Proxy pattern-mixture analysis for skewed data." DOI:10.1214/15-AOAS878.

BAVDAŽ, M. (2010). The multidimensional integral business survey response model. *Survey Methodology* **1** 81–93.

BEAUMONT, J.-F., HAZIZA, D. and BOCCI, C. (2011). On variance estimation under auxiliary value imputation in sample surveys. *Statist. Sinica* **21** 515–537. MR2829845

DEVROYE, L. (2002). Simulating Bessel random variables. *Statist. Probab. Lett.* **57** 249–257. MR1912083

EFRON, B. (1994). Missing data, imputation, and the bootstrap. *J. Amer. Statist. Assoc.* **89** 463–479. MR1294072

Key words and phrases. Missing data, nonresponse bias analysis, nonignorable missingness, multiple imputation, skewed data, business surveys, proxy pattern-mixture models.

FAY, R. E. III and HERRIOT, R. A. (1979). Estimates of income for small places: An application of James–Stein procedures to census data. *J. Amer. Statist. Assoc.* **74** 269–277. MR0548019

FELLER, W. (1966). *An Introduction to Probability Theory and Its Applications. Vol. II*. Wiley, New York. MR0210154

HAREL, O. (2007). Inferences on missing information under multiple imputation and two-stage multiple imputation. *Stat. Methodol.* **4** 75–89. MR2339010

HAZIZA, D., THOMPSON, K. J. and YUNG, W. (2010). The effect of nonresponse adjustments on variance estimation. *Survey Methodology* **36** 35–43.

ILIOPOULOS, G., KARLIS, D. and NTZOUFRAS, I. (2005). Bayesian estimation in Kibble's bivariate gamma distribution. *Canad. J. Statist.* **33** 571–589. MR2232381

IZAWA, T. (1965). Two or multi-dimensional gamma-type distribution and its application to rainfall data. *Papers in Meteorology and Geophysics* **15** 167–200.

KIBBLE, W. F. (1941). A two-variate gamma type distribution. *Sankhyā* **5** 137–150. MR0007218

KREUTER, F., OLSON, K., WAGNER, J., YAN, T., EZZATI-RICE, T. M., CASAS-CORDERO, C., LEMAY, M., PEYTCHEV, A., GROVES, R. M. and RAGHUNATHAN, T. E. (2010). Using proxy measures and other correlates of survey outcomes to adjust for non-response: Examples from multiple surveys. *J. Roy. Statist. Soc. Ser. A* **173** 389–407. MR2751883

KREWSKI, D. and RAO, J. N. K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *Ann. Statist.* **9** 1010–1019. MR0628756

LITTLE, R. J. A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika* **81** 471–483. MR1311091

LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley, Hoboken, NJ. MR1925014

LOHR, S. L. (2010). *Sampling: Design and Analysis*, 2nd ed. Brooks/Cole, Boston, MA. MR3057878

MAKAROV, R. N. and GLEW, D. (2010). Exact simulation of Bessel diffusions. *Monte Carlo Methods Appl.* **16** 283–306. MR2747817

ONG, S. H. (1992). The computer generation of bivariate binomial variables with given marginals and correlations. *Comm. Statist. Simulation Comput.* **21** 285–299.

PEYTCHEVA, E. and GROVES, R. M. (2009). Using variation in response rates of demographic subgroups as evidence of nonresponse bias in survey estimates. *Journal of Official Statistics* **25** 193–201.

R CORE TEAM (2012). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org/.

RAO, J. N. K. (2003). *Small Area Estimation*. Wiley, Hoboken, NJ. MR1953089

RAO, J. N. K. and SCOTT, A. J. (1992). A simple method of the analysis of clustered binary data. *Biometrika* **74** 577–585.

ROBERTS, G., RAO, J. N. K. and KUMAR, S. (1987). Logistic regression analysis of sample survey data. *Biometrika* **74** 1–12. MR0885914

ROYALL, R. M. (1992). The model based (prediction) approach to finite population sampling theory. In *Current Issues in Statistical Inference: Essays in Honor of D. Basu. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **17** 225–240. IMS, Hayward, CA. MR1194420

SNIJKERS, G., HARALDSEN, G., JONES, J. and WILLIMACK, D. K. (2013). *Designing and Conducting Business Surveys*. Wiley, New York.

THOMPSON, K. J. (2005). An empirical investigation into the effects of replicate reweighting on variance estimates for the annual capital expenditures survey. In *Proceedings of the Federal Committee on Statistical Methods Research Conference*. U.S. Office of Management and Budget, Washington, DC.

THOMPSON, K. J. and OLIVER, B. E. (2012). Response rates in business surveys: Going beyond the usual performance measure. *Journal of Official Statistics* **28** 221–237.

THOMPSON, K. J. and WASHINGTON, K. T. (2013). Challenges in the treatment of unit nonresponse for selected business surveys: A case study. *Survey Methods: Insights from the Field*. Retrieved from http://surveyinsights.org/?p=2991.

WAGNER, J. (2010). The fraction of missing information as a tool for monitoring the quality of survey data. *Public Opinion Quarterly* **74** 223–243.

WAGNER, J. (2012). A comparison of alternative indicators for the risk of nonresponse bias. *Public Opinion Quarterly* **76** 555–575.

WILLIMACK, D. K. and NICHOLS, E. (2010). A hybrid response process model for business surveys. *Journal of Official Statistics* **1** 3–24.

YUAN, L. and KALBFLEISCH, J. D. (2000). On the Bessel distribution and related problems. *Ann. Inst. Statist. Math.* **52** 438–447. MR1794244

The Annals of Applied Statistics
2015, Vol. 9, No. 4, 2266–2267
DOI: 10.1214/15-AOAS895
Main article DOI: 10.1214/11-AOAS504

CORRECTION

EFFICIENT REGULARIZED ISOTONIC REGRESSION WITH APPLICATION TO GENE–GENE INTERACTION SEARCH

BY RONNY LUSS*, SAHARON ROSSET† AND MONI SHAHAR†

*IBM Research** *and Tel Aviv University*†

REFERENCES

HOCHBAUM, D. S. and QUEYRANNE, M. (2003). Minimizing a convex cost closure set. *SIAM J. Discrete Math*. **16** 192–207 (electronic). MR1982135

LUSS, R., ROSSET, S. and SHAHAR, M. (2012). Efficient regularized isotonic regression with application to gene–gene interaction search. *Ann. Appl. Stat.* **6** 253–283. MR2951537

MAXWELL, W. L. and MUCKSTADT, J. A. (1985). Establishing consistent and realistic reorder intervals in production–distribution systems. *Oper. Res*. **33** 1316–1341.

SPOUGE, J., WAN, H. and WILBUR, W. J. (2003). Least squares isotonic regression in two dimensions. *J. Optim. Theory Appl*. **117** 585–605. MR1989929