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

THE ANNALS *of* APPLIED STATISTICS

An Official Journal of the Institute of Mathematical Statistics

Articles

continued

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

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles—*Continued from front cover*

Vol. 11, No. 1—March 2017

The Annals of Applied Statistics
2017, Vol. 11, No. 1, 1–20
DOI: 10.1214/16-AOAS969

RANDOMIZATION INFERENCE FOR STEPPED-WEDGE CLUSTER-RANDOMIZED TRIALS: AN APPLICATION TO COMMUNITY-BASED HEALTH INSURANCE

BY XINYAO JI[*], GUNTHER FINK[†], PAUL JACOB ROBYN[‡] AND DYLAN S. SMALL[*]

University of Pennsylvania[*], Harvard University[†] and The World Bank[‡]

National health insurance schemes are generally impractical in low-income countries due to limited resources and low organizational capacity. In response to such obstacles, community-based health insurance (CBHI) schemes have emerged over the past 20 years. CBHIs are designed to reduce the financial burden generated by unanticipated treatment cost among individuals falling sick, and thus are expected to make health care more affordable. In this paper, we investigate whether CBHI schemes effectively protect individuals against large financial shocks using a stepped-wedge cluster-randomized design on data from a CBHI program rolled out in rural Burkina Faso. We investigate statistical properties of the stepped-wedge design following the parametric mixed model approach proposed by Hussey and Hughes in 2007. We find that testing for the treatment effect is generally sensitive to specification of the parametric model. For instance, if we fail to account for cluster-by-time interactions present in the data, the Type I error rate is severely inflated. We develop a more robust and efficient strategy—randomization inference. We demonstrate how to apply randomization inference to test for constant treatment effects and discuss test statistics suitable for the stepped-wedge design. Randomization inference guarantees a valid Type I error rate; simulation studies show that randomization inference test statistics also have power that is comparable to the currently used procedures that do not guarantee a valid Type I error rate. Finally, we apply our proposed method to the Burkina Faso CBHI dataset. We conclude that the insurance had limited effects on reducing the likelihood of low to moderate levels of catastrophic health expenditure, but substantially reduced the likelihood of extremely high health expenditure that exceeds half of a person's monthly income.

REFERENCES

ASENSO-OKYERE, W. K., OSEI-AKOTO, I., ANUM, A. and APPIAH, E. N. (1997). Willingness to pay for health insurance in a developing economy. A pilot study of the informal sector of Ghana using contingent valuation. *Health Policy* **42** 223–237.

BELLAN, S. E., PULLIAM, J. R., PEARSON, C. A., CHAMPREDON, D., FOX, S. J., SKRIP, L., GALVANI, A. P., GAMBHIR, M., LOPMAN, B. A., PORCO, T. C. et al. (2015). Statistical power and validity of Ebola vaccine trials in Sierra Leone: A simulation study of trial design and analysis. *Lancet*, *Infect*. *Dis*. **15** 703–710.

Key words and phrases. Randomization inference, stepped-wedge cluster-randomized trials, community-based health insurance.

BRAUN, T. M. and FENG, Z. (2001). Optimal permutation tests for the analysis of group randomized trials. *J. Amer. Statist. Assoc.* **96** 1424–1432. MR1946587

BROWN, C. A. and LILFORD, R. J. (2006). The stepped wedge trial design: A systematic review. *BMC Med. Res. Methodol.* **6** 54.

DEVADASAN, N., RANSON, K., DAMME, W. V., ACHARYA, A. and CRIEL, B. (2006). The landscape of community health insurance in India: An overview based on 10 case studies. *Health Policy* **78** 224–234.

DE ALLEGRI, M. D., KOUYATÉ, B., BECHER, H., GBANGOU, A., POKHREL, S., SANON, M. and SAUERBORN, R. (2006). Understanding enrolment in community health insurance in sub-Saharan Africa: A population-based case-control study in rural Burkina Faso. *Bull. World Health Organ.* **84** 852–858.

DE ALLEGRI, M. D., POKHREL, S., BECHER, H., DONG, H., MANSMANN, U., KOUYATÉ, B., KYNAST-WOLF, G., GBANGOU, A., SANON, M., BRIDGES, J. and SAUERBORN, R. (2008). Step-wedge cluster-randomised community-based trials: An application to the study of the impact of community health insurance. *Health Res. Policy Syst.* **6** 10.

DIMAIRO, M., BRADBURN, M. and WALTERS, S. J. (2011). Sample size determination through power simulation; practical lessons from a stepped wedge cluster randomised trial (SW CRT). *Trials* **12 (Suppl 1)** A26.

EKMAN, B. (2004). Community-based health insurance in low-income countries: A systematic review of the evidence. *Health Policy Plan.* **19** 249–270.

FINK, G., ROBYN, P. J., SIÉ, A. and SAUERBORN, R. (2013). Does health insurance improve health? Evidence from a randomized community-based insurance rollout in rural Burkina Faso. *J. Health Econ.* **32** 1043–1056.

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

GAIL, M. H., MARK, S. D., CARROLL, R. J., GREEN, S. B. and PEE, D. (1996). On design considerations and randomization-based inference for community intervention trials. *Stat. Med.* **15** 1069–1092.

GREENE, W. H. (2003). *Econometric Analysis*. Pearson Education India.

GREEVY, R., SILBER, J. H., CNAAN, A. and ROSENBAUM, P. R. (2004). Randomization inference with imperfect compliance in the ACE-inhibitor after anthracycline randomized trial. *J. Amer. Statist. Assoc.* **99** 7–15. MR2061884

HALL, A. J., INSKIP, H. M., LOIK, F., DAY, N. E., O'CONOR, G., BOSCH, X. and MUIR, C. S. (1987). The Gambia hepatitis intervention study. *Cancer Res.* **47** 5782–5787.

HANSEN, B. B. and BOWERS, J. (2009). Attributing effects to a cluster-randomized get-out-the-vote campaign. *J. Amer. Statist. Assoc.* **104** 873–885. MR2562000

HO, D. E. and IMAI, K. (2006). Randomization inference with natural experiments: An analysis of ballot effects in the 2003 California recall election. *J. Amer. Statist. Assoc.* **101** 888–900. MR2324090

HOAGLIN, D. C., MOSTELLER, F. and TUKEY, J. W. (2000). *Understanding Robust and Exploratory Data Analysis*. Wiley, New York. MR1800901

HUSSEY, M. A. and HUGHES, J. P. (2007). Design and analysis of stepped wedge cluster randomized trials. *Contemp. Clin. Trials* **28** 182–191.

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

JACQMIN-GADDA, H., SIBILLOT, S., PROUST, C., MOLINA, J.-M. and THIÉBAUT, R. (2007). Robustness of the linear mixed model to misspecified error distribution. *Comput. Statist. Data Anal.* **51** 5142–5154. MR2370713

MDEGE, N. D., MAN, M. S., TAYLOR, C. A. and TORGERSON, D. J. (2011). Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. *J. Clin. Epidemiol.* **64** 936–948.

MOULTON, L. H., GOLUB, J. E., DUROVNI, B., CAVALCANTE, S. C., PACHECO, A. G., SARACENI, V., KING, B. and CHAISSON, R. E. (2007). Statistical design of THRio: A phased implementation clinic-randomized study of a tuberculosis preventive therapy intervention. *Clin. Trials* **4** 190–199.

NEYMAN, J. (1990). On the application of probability theory to agricultural experiments. *Statist. Sci.* **5** 463–464.

RAZ, J. (1990). Testing for no effect when estimating a smooth function by nonparametric regression: A randomization approach. *J. Amer. Statist. Assoc.* **85** 132–138. MR1137359

RHODA, D. A., MURRAY, D. M., ANDRIDGE, R. R., PENNELL, M. L. and HADE, E. M. (2011). Studies with staggered starts: Multiple baseline designs and group-randomized trials. *Am. J. Publ. Health* **101** 2164–2169.

ROBYN, P. J., FINK, G., SIÉ, A. and SAUERBORN, R. (2012). Health insurance and health-seeking behavior: Evidence from a randomized community-based insurance rollout in rural Burkina Faso. *Soc. Sci. Med.* **75** 595–603.

ROSENBAUM, P. R. (2002a). Covariance adjustment in randomized experiments and observational studies. *Statist. Sci.* **17** 286–327. MR1962487

ROSENBAUM, P. R. (2002b). *Observational Studies*. Springer, New York.

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

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *J. Educ. Psychol.* **66** 688–701.

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

SEN, P. K. (1968). On a class of aligned rank order tests in two-way layouts. *Ann. Math. Stat.* **39** 1115–1124. MR0226774

SMALL, D. S., TEN HAVE, T. R. and ROSENBAUM, P. R. (2008). Randomization inference in a group-randomized trial of treatments for depression: Covariate adjustment, noncompliance, and quantile effects. *J. Amer. Statist. Assoc.* **103** 271–279. MR2420232

TUKEY, J. W. (1993). Tightening the clinical trial. *Control. Clin. Trials* **14** 266–285.

VAN DER TWEEL, I. and VAN DER GRAAF, R. (2013). Issues in the use of stepped wedge cluster and alternative designs in the case of pandemics. *Am. J. Bioeth.* **13** 23–24.

WANG, H., YIP, W., ZHANG, L. and HSIAO, W. C. (2009). The impact of rural mutual health care on health status: Evaluation of a social experiment in rural China. *Health Econ.* **18** S65–S82.

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

WOERTMAN, W., DE HOOP, E., MOERBEEK, M., ZUIDEMA, S. U., GERRITSEN, D. L. and TEERENSTRA, S. (2013). Stepped wedge designs could reduce the required sample size in cluster randomized trials. *J. Clin. Epidemiol.* **66** 752–758.

The Annals of Applied Statistics
2017, Vol. 11, No. 1, 21–40
DOI: 10.1214/16-AOAS989

MODELLING INDIVIDUAL MIGRATION PATTERNS USING A BAYESIAN NONPARAMETRIC APPROACH FOR CAPTURE–RECAPTURE DATA

BY ELENI MATECHOU AND FRANÇOIS CARON[1]

University of Kent and University of Oxford

We present a Bayesian nonparametric approach for modelling wildlife migration patterns using capture–recapture (CR) data. Arrival times of individuals are modelled in continuous time and assumed to be drawn from a Poisson process with unknown intensity function, which is modelled via a flexible nonparametric mixture model. The proposed CR framework allows us to estimate the following: (i) the total number of individuals that arrived at the site, (ii) their times of arrival and departure, and hence their stopover duration, and (iii) the density of arrival times, providing a smooth representation of the arrival pattern of the individuals at the site. We apply the model to data on breeding great crested newts (*Triturus cristatus*) and on migrating reed warblers (*Acrocephalus scirpaceus*). For the former, the results demonstrate the staggered arrival of individuals at the breeding ponds and suggest that males tend to arrive earlier than females. For the latter, they demonstrate the arrival of migrating flocks at the stopover site and highlight the considerable difference in stopover duration between caught and not-caught individuals.

REFERENCES

BASU, S. (1998). Capture–recapture and nonparametric Bayes. Technical report, Division of Statistics, Northern Illinois Univ., DeKalb, IL.

BASU, S. and EBRAHIMI, N. (2001). Bayesian capture–recapture methods for error detection and estimation of population size: Heterogeneity and dependence. *Biometrika* **88** 269–279. MR1841274

BAUER, S., VAN DINTHER, M., HØGDA, K.-A., KLAASSEN, M. and MADSEN, J. (2008). The consequences of climate-driven stop-over sites changes on migration schedules and fitness of Arctic geese. *J. Anim. Ecol.* **77** 654–660.

BESBEAS, P., FREEMAN, S. N., MORGAN, B. J. T. and CATCHPOLE, E. A. (2002). Integrating mark–recapture-recovery and census data to estimate animal abundance and demographic parameters. *Biometrics* **58** 540–547. MR1933532

BOTH, C., VAN TURNHOUT, C. A., BIJLSMA, R. G., SIEPEL, H., VAN STRIEN, A. J. and FOPPEN, R. P. (2009). Avian population consequences of climate change are most severe for long-distance migrants in seasonal habitats. *Proc. R. Soc. Lond., B Biol. Sci.* **277** 1259–1266.

BRIX, A. (1999). Generalized gamma measures and shot-noise Cox processes. *Adv. in Appl. Probab.* **31** 929–953. MR1747450

BRIX, A. and DIGGLE, P. J. (2001). Spatiotemporal prediction for log-Gaussian Cox processes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 823–841. MR1872069

CHARMANTIER, A. and GIENAPP, P. (2014). Climate change and timing of avian breeding and migration: Evolutionary versus plastic changes. *Evol. Appl.* **7** 15–28.

Key words and phrases. Chinese restaurant process, great crested newts, Poisson–Gamma process, reed warblers, shot-noise Cox process, stopover data.

CORMACK, R. (1964). Estimates of survival from the sighting of marked animals. *Biometrika* **51** 429–438.

DALEY, D. J. and VERE-JONES, D. (2003). *An Introduction to the Theory of Point Processes. Vol. I: Elementary Theory and Methods*, Springer, New York. MR1950431

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

ERNI, B., LIECHTI, F., UNDERHILL, L. G. and BRUDERER, B. (2002). Wind and rain govern the intensity of nocturnal bird migration in central Europe: A log-linear regression analysis. *Ardea* **90** 155–166.

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

GILBERT, S. L., LINDBERG, M. S., HUNDERTMARK, K. J. and PERSON, D. K. (2014). Dead before detection: Addressing the effects of left truncation on survival estimation and ecological inference for neonates. *Methods Ecol. Evol.* **5** 992–1001.

GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. MR1380810

ISHWARAN, H. and JAMES, L. F. (2004). Computational methods for multiplicative intensity models using weighted gamma processes: Proportional hazards, marked point processes, and panel count data. *J. Amer. Statist. Assoc.* **99** 175–190. MR2054297

JEHLE, R., THIESMEIER, B. and FOSTER, J. (2011). *The Crested Newt: A Dwindling Pond-Dweller*. Laurenti, Bielefeld.

JOLLY, G. M. (1965). Explicit estimates from capture–recapture data with both death and immigration-stochastic model. *Biometrika* **52** 225–247. MR0210227

KINGMAN, J. F. C. (1993). *Poisson Processes*. Oxford Univ. Press, New York. MR1207584

KOTTAS, A. and SANSÓ, B. (2007). Bayesian mixture modeling for spatial Poisson process intensities, with applications to extreme value analysis. *J. Statist. Plann. Inference* **137** 3151–3163. MR2365118

KOVÁCS, S., FEHÉRVÁRI, P., NAGY, K., HARNOS, A. and CSÖRGŐ, T. (2012). Changes in migration phenology and biometrical traits of reed, marsh and sedge warblers. *Cent. Eur. J. Biol.* **7** 115–125.

KUO, L. and GHOSH, S. K. (1997). Bayesian nonparametric inference for nonhomogeneous Poisson processes. Technical report, Dept. Statistics, Univ. Connecticut, Storrs, CT.

LEBRETON, J.-D., BURNHAM, K. P., CLOBERT, J. and ANDERSON, D. R. (1992). Modeling survival and testing biological hypotheses using marked animals: A unified approach with case studies. *Ecol. Monogr.* **62** 67–118.

LEWIS, B. (2012). An evaluation of mitigation actions for great crested newts at development sites. Ph.D. thesis, Durrell Institute of Conservation and Ecology, School of Anthropology & Conservation, University of Kent.

LO, A. Y. and WENG, C.-S. (1989). On a class of Bayesian nonparametric estimates. II. Hazard rate estimates. *Ann. Inst. Statist. Math.* **41** 227–245. MR1006487

LYONS, J. E., KENDALL, W. L., ROYLE, J. A., CONVERSE, S. J., ANDRES, B. A. and BUCHANAN, J. B. (2016). Population size and stopover duration estimation using mark-resight data and Bayesian analysis of a superpopulation model. *Biometrics* **72** 262–271.

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

MANRIQUE-VALLIER, D. (2016). Bayesian population size estimation using Dirichlet process mixtures. *Biometrics*. DOI:10.1111/biom.12502.

MATECHOU, E. and CARON, F. (2017a). Supplement to "Modelling individual migration patterns using a Bayesian nonparametric approach for capture–recapture data". DOI:10.1214/16-AOAS989SUPPA.

MATECHOU, E. and CARON, F. (2017b). Supplement to "Modelling individual migration patterns using a Bayesian nonparametric approach for capture–recapture data". DOI:10.1214/16-AOAS989SUPPB.
MATECHOU, E., MORGAN, B. J. T., PLEDGER, S., COLLAZO, J. A. and LYONS, J. E. (2013a). Integrated analysis of capture–recapture-resighting data and counts of unmarked birds at stopover sites. *J. Agric. Biol. Environ. Stat.* **18** 120–135. MR3067331
MATECHOU, E., PLEDGER, S., EFFORD, M., MORGAN, B. J. T. and THOMSON, D. L. (2013b). Estimating age-specific survival when age is unknown: Open population capture–recapture models with age structure and heterogeneity. *Methods Ecol. Evol.* **4** 654–664.
MCCLINTOCK, B. T., BAILEY, L. L., DREHER, B. P. and LINK, W. A. (2014). Probit models for capture–recapture data subject to imperfect detection, individual heterogeneity and misidentification. *Ann. Appl. Stat.* **8** 2461–2484. MR3292505
MCCREA, R. S., MORGAN, B. J. T., GIMENEZ, O., BESBEAS, P., LEBRETON, J.-D. and BREGNBALLE, T. (2010). Multi-site integrated population modelling. *J. Agric. Biol. Environ. Stat.* **15** 539–561. MR2788639
MØLLER, J. (2003). Shot noise Cox processes. *Adv. in Appl. Probab.* **35** 614–640. MR1990607
MØLLER, J., SYVERSVEEN, A. R. and WAAGEPETERSEN, R. P. (1998). Log Gaussian Cox processes. *Scand. J. Stat.* **25** 451–482. MR1650019
MØLLER, J. and WAAGEPETERSEN, R. P. (2003). *Statistical Inference and Simulation for Spatial Point Processes. Monographs on Statistics and Applied Probability* **100**. Chapman & Hall/CRC, Boca Raton, FL. MR2004226
NIETO-BARAJAS, L. E. and WALKER, S. G. (2004). Bayesian nonparametric survival analysis via Lévy driven Markov processes. *Statist. Sinica* **14** 1127–1146. MR2126344
PITMAN, J. (1996). Some developments of the Blackwell–MacQueen urn scheme. In *Statistics, Probability and Game Theory. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **30** 245–267. IMS, Hayward, CA. MR1481784
PLEDGER, S., EFFORD, M., POLLOCK, K. H., COLLAZO, J. A. and LYONS, J. E. (2009). Stopover duration analysis with departure probability dependent on unknown time since arrival. In *Modeling Demographic Processes in Marked Populations* (D. L. Thomson, E. G. Cooch and M. J. Conroy, eds.). *Environmental and Ecological Statistics* **3** 349–363.
PLUMMER, M., BEST, N., COWLES, K. and VINES, K. (2006). Coda: Convergence diagnosis and output analysis for MCMC. *R News* **6** 7–11.
PRADEL, R. (1996). Utilization of capture–mark–recapture for the study of recruitment and population growth rate. *Biometrics* **52** 703–709. MR1395003
ROCCHETTI, I., BUNGE, J. and BÖHNING, D. (2011). Population size estimation based upon ratios of recapture probabilities. *Ann. Appl. Stat.* **5** 1512–1533. MR2849784
ROYLE, J. A., DORAZIO, R. M. and LINK, W. A. (2007). Analysis of multinomial models with unknown index using data augmentation. *J. Comput. Graph. Statist.* **16** 67–85. MR2345748
ROYLE, J. A. and YOUNG, K. V. (2008). A hierarchical model for spatial capture–recapture data. *Ecology* **89** 2281–2289.
ROYLE, J. A., KARANTH, K. U., GOPALASWAMY, A. M. and KUMAR, N. S. (2009). Bayesian inference in camera trapping studies for a class of spatial capture–recapture models. *Ecology* **90** 3233–3244.
SCHAUB, M., LIECHTI, F. and JENNI, L. (2004). Departure of migrating European robins, *Erithacus rubecula*, from a stopover site in relation to wind and rain. *Anim. Behav.* **67** 229–237.
SCHAUB, M., PRADEL, R., JENNI, L. and LEBRETON, J.-D. (2001). Migrating birds stop over longer than usually thought: An improved capture–recapture analysis. *Ecology* **82** 852–859.
SCHWARZ, C. J. and ARNASON, A. N. (1996). A general methodology for the analysis of capture–recapture experiments in open populations. *Biometrics* **52** 860–873. MR1411736
SEBER, G. A. F. (1965). A note on the multiple-recapture census. *Biometrika* **52** 249–259. MR0210228

SEEBACHER, F. and POST, E. (2015). Climate change impacts on animal migration. *Clim. Change Responses* **2** 1.

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

SULLIVAN, A. R., FLASPOHLER, D. J., FROESE, R. E. and FORD, D. (2015). Climate variability and the timing of spring raptor migration in eastern North America. *J. Avian Biol.* **47** 208–218.

TADDY, M. A. and KOTTAS, A. (2012). Mixture modeling for marked Poisson processes. *Bayesian Anal.* **7** 335–361. MR2934954

R CORE TEAM (2014). *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

VAN BUSKIRK, J., MULVIHILL, R. S. and LEBERMAN, R. C. (2009). Variable shifts in spring and autumn migration phenology in North American songbirds associated with climate change. *Glob. Change Biol.* **15** 760–771.

WOLPERT, R. L. and ICKSTADT, K. (1998). Poisson/gamma random field models for spatial statistics. *Biometrika* **85** 251–267. MR1649114

The Annals of Applied Statistics
2017, Vol. 11, No. 1, 41–68
DOI: 10.1214/16-AOAS990

GENE NETWORK RECONSTRUCTION USING GLOBAL-LOCAL SHRINKAGE PRIORS[1]

BY GWENAËL G. R. LEDAY*, MATHISCA C. M. DE GUNST†, GINO B. KPOGBEZAN‡, AAD W. VAN DER VAART‡, WESSEL N. VAN WIERINGEN†,§ AND MARK A. VAN DE WIEL†,§

University of Cambridge, Vrije Universiteit Amsterdam†, Leiden University‡ and VU University Medical Center§*

Reconstructing a gene network from high-throughput molecular data is an important but challenging task, as the number of parameters to estimate easily is much larger than the sample size. A conventional remedy is to regularize or penalize the model likelihood. In network models, this is often done *locally* in the neighborhood of each node or gene. However, estimation of the many regularization parameters is often difficult and can result in large statistical uncertainties. In this paper we propose to combine local regularization with *global* shrinkage of the regularization parameters to borrow strength between genes and improve inference. We employ a simple Bayesian model with nonsparse, conjugate priors to facilitate the use of fast variational approximations to posteriors. We discuss empirical Bayes estimation of hyperparameters of the priors, and propose a novel approach to rank-based posterior thresholding. Using extensive model- and data-based simulations, we demonstrate that the proposed inference strategy outperforms popular (sparse) methods, yields more stable edges, and is more reproducible. The proposed method, termed `ShrinkNet`, is then applied to Glioblastoma to investigate the interactions between genes associated with patient survival.

REFERENCES

ALLEN, G. I. and LIU, Z. (2013). A local Poisson graphical model for inferring networks from sequencing data. *IEEE Trans. Nanobiosci.* **12** 189–198.

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

BONDELL, H. D. and REICH, B. J. (2012). Consistent high-dimensional Bayesian variable selection via penalized credible regions. *J. Amer. Statist. Assoc.* **107** 1610–1624. MR3036420

BRAUN, M. and MCAULIFFE, J. (2010). Variational inference for large-scale models of discrete choice. *J. Amer. Statist. Assoc.* **105** 324–335. MR2757203

CAMBY, I., MERCIER, M. L., LEFRANC, F. and KISS, R. (2006). Galectin-1: A small protein with major functions. *Glycobiology* **16** 137R–157R.

CERAMI, E. G., GROSS, B. E., DEMIR, E., RODCHENKOV, I., BABUR, Ö., ANWAR, N., SCHULTZ, N., BADER, G. D. and SANDER, C. (2011). Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res.* **39** D685–D690.

Key words and phrases. Undirected gene network, Bayesian inference, shrinkage, variational approximation, empirical Bayes.

CERAMI, E., GAO, J., DOGRUSOZ, U., GROSS, B. E., SUMER, S. O., AKSOY, B. A., JACOBSEN, A., BYRNE, C. J., HEUER, M. L., LARSSON, E., ANTIPIN, Y., REVA, B., GOLDBERG, A. P., SANDER, C. and SCHULTZ, N. (2012). The cBio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery* **2** 401–404.

CHEN, J. and CHEN, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95** 759–771. MR2443189

CHEN, S., WITTEN, D. M. and SHOJAIE, A. (2015). Selection and estimation for mixed graphical models. *Biometrika* **102** 47–64. MR3335095

CORDES, C., BARTLING, B., SIMM, A., AFAR, D., LAUTENSCHLAGER, C., HANSEN, G., SILBER, R.-E., BURDACH, S. and HOFMANN, H.-S. (2009). Simultaneous expression of Cathepsins B and K in pulmonary adenocarcinomas and squamous cell carcinomas predicts poor recurrence-free and overall survival. *Lung Cancer* **64** 79–85.

DOBRA, A., LENKOSKI, A. and RODRIGUEZ, A. (2011). Bayesian inference for general Gaussian graphical models with application to multivariate lattice data. *J. Amer. Statist. Assoc.* **106** 1418–1433. MR2896846

DOBRA, A., HANS, C., JONES, B., NEVINS, J. R., YAO, G. and WEST, M. (2004). Sparse graphical models for exploring gene expression data. *J. Multivariate Anal.* **90** 196–212. MR2064941

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

EFRON, B. (2010). *Large-Scale Inference*: *Empirical Bayes Methods for Estimation*, *Testing*, *and Prediction*. *Institute of Mathematical Statistics* (*IMS*) *Monographs* **1**. Cambridge Univ. Press, Cambridge. MR2724758

FORTIN, S., MERCIER, M. L., CAMBY, I., SPIEGL-KREINECKER, S., BERGER, W., LEFRANC, F. and KISS, R. (2010). Galectin-1 is implicated in the protein kinase C epsilon/vimentin-controlled trafficking of integrin-beta1 in glioblastoma cells. *Brain Pathol.* **20** 39–49.

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.

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

GAO, X., PU, D. Q., WU, Y. and XU, H. (2012). Tuning parameter selection for penalized likelihood estimation of Gaussian graphical model. *Statist. Sinica* **22** 1123–1146. MR2987486

GEIGER, D. and HECKERMAN, D. (2002). Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *Ann. Statist.* **30** 1412–1440. MR1936324

GIRAUD, C. (2008). Estimation of Gaussian graphs by model selection. *Electron. J. Stat.* **2** 542–563. MR2417393

GOLE, B., HUSZTHY, P. C., POPOVIĆ, M., JERUC, J., ARDEBILI, Y. S., BJERKVIG, R. and LAH, T. T. (2012). The regulation of cysteine cathepsins and cystatins in human gliomas. *Int. J. Cancer* **131** 1779–1789.

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

JACOBSEN, A. (2013). cgdsr: R-Based API for accessing the MSKCC Cancer Genomics Data Server (CGDS). R package version 1.1.30.

KALLUNKI, T., OLSEN, O. D. and JÄÄTTELÄ, M. (2013). Cancer-associated lysosomal changes: Friends or foes? *Oncogene* **32** 1995–2004.

KASS, R. E. and WASSERMAN, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Amer. Statist. Assoc.* **90** 928–934. MR1354008

KRÄMER, N., SCHÄFER, J. and BOULESTEIX, A.-L. (2009). Regularized estimation of large-scale gene association networks using graphical Gaussian models. *BMC Bioinformatics* **10** 384.

LEDAY, G. G. R., DE GUNST, M. C. M., KPOGBEZAN, G. B., VAN DER VAART, A. W., VAN WIERINGEN, W. N., and VAN DE WIEL, M. A. (2017). Supplement to "Gene network reconstruction using global-local shrinkage priors." DOI:10.1214/16-AOAS990SUPP.

LEDOIT, O. and WOLF, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.* **88** 365–411. MR2026339

LEWIS, C. A., BRAULT, C., PECK, B., BENSAAD, K., GRIFFITHS, B., MITTER, R., CHAKRAVARTY, P., EAST, P., DANKWORTH, B., ALIBHAI, D. et al. (2015). SREBP maintains lipid biosynthesis and viability of cancer cells under lipid-and oxygen-deprived conditions and defines a gene signature associated with poor survival in glioblastoma multiforme. *Oncogene*.

LIAN, H. (2011). Shrinkage tuning parameter selection in precision matrices estimation. *J. Statist. Plann. Inference* **141** 2839–2848. MR2787749

LIANG, F., PAULO, R., MOLINA, G., CLYDE, M. A. and BERGER, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *J. Amer. Statist. Assoc.* **103** 410–423. MR2420243

LIM, K. S., LIM, K. J., PRICE, A. C., ORR, B. A., EBERHART, C. G. and BAR, E. E. (2013). Inhibition of monocarboxylate transporter-4 depletes stem-like glioblastoma cells and inhibits HIF transcriptional response in a lactate-independent manner. *Oncogene*.

LUO, S., SONG, R. and WITTEN, D. (2014). Sure screening for Gaussian graphical models. Preprint. Available at arXiv:1407.7819.

MADHANKUMAR, A. B., SLAGLE-WEBB, B., MINTZ, A., SHEEHAN, J. M. and CONNOR, J. R. (2006). Interleukin-13 receptor-targeted nanovesicles are a potential therapy for glioblastoma multiforme. *Mol. Cancer Ther.* **5** 3162–3169.

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

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

MOHAMMADI, A. and WIT, E. C. (2015). Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Anal.* **10** 109–138. MR3420899

OATES, C. J. and MUKHERJEE, S. (2012). Network inference and biological dynamics. *Ann. Appl. Stat.* **6** 1209–1235. MR3012527

ORMEROD, J. T. and WAND, M. P. (2010). Explaining variational approximations. *Amer. Statist.* **64** 140–153. MR2757005

PARK, T. and CASELLA, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.* **103** 681–686. MR2524001

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

PORSTMANN, T., SANTOS, C. R., GRIFFITHS, B., CULLY, M., WU, M., LEEVERS, S., GRIFFITHS, J. R., CHUNG, Y.-L. and SCHULZE, A. (2008). SREBP activity is regulated by mTORC1 and contributes to akt-dependent cell growth. *Cell Metabolism* **8** 224–236.

RAJAGOPALAN, M. and BROEMELING, L. (1983). Bayesian inference for the variance components in general mixed linear models. *Comm. Statist. Theory Methods* **12** 701–723. MR0696816

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

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

SCHAEFER, J., OPGEN-RHEIN, R. and STRIMMER, K. (2006). Reverse engineering genetic networks using the GeneNet package. *R News* **6/5** 50–53.

SCHÄFER, J. and STRIMMER, K. (2005a). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* **4** Art. 32, 28 pp. (electronic). MR2183942

SCHÄFER, J. and STRIMMER, K. (2005b). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics* **21** 754–764.

SCUTARI, M. (2013). On the prior and posterior distributions used in graphical modelling. *Bayesian Anal*. **8** 505–532. MR3102220

VALPOLA, H. and HONKELA, A. (2006). Hyperparameter adaptation in variational Bayes for the gamma distribution. Technical report, Helsinki University of Technology.

VAN WIERINGEN, W. N. and PEETERS, C. F. W. (2016). Ridge estimation of inverse covariance matrices from high-dimensional data. *Comput*. *Statist*. *Data Anal*. **103** 284–303.

VAN DE WIEL, M. A., LEDAY, G. G. R., PARDO, L., RUE, H., VAN DER VAART, A. W. and VAN WIERINGEN, W. N. (2013). Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics* **14** 113–128.

WANG, H. and LI, S. Z. (2012). Efficient Gaussian graphical model determination under G-Wishart prior distributions. *Electron*. *J*. *Stat*. **6** 168–198. MR2879676

WARTON, D. I. (2008). Penalized normal likelihood and ridge regularization of correlation and covariance matrices. *J*. *Amer*. *Statist*. *Assoc*. **103** 340–349. MR2394637

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

YAJIMA, M., TELESCA, D., JI, Y. and MULLER, P. (2012). Differential patterns of interaction and Gaussian graphical models. *COBRA Preprint Series* **91**.

YANG, E., RAVIKUMAR, P., ALLEN, G. and LIU, Z. (2012). Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems* 25 (P. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger, eds.) 1367–1375.

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

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

YUAN, Y., CURTIS, C., CALDAS, C. and MARKOWETZ, F. (2012). A sparse regulatory network of copy-number driven gene expression reveals putative breast cancer oncogenes. *IEEE/ACM Trans Comput Biol Bioinform* **9** 947–954.

ZELLNER, A. and SIOW, A. (1980). Posterior odds ratios for selected regression hypotheses. In *Bayesian Statistics*: *Proceedings of the First International Meeting Held in Valencia* (*Spain*), *Vol*. 1 (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.). University Press, Valencia.

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., RÜTIMANN, P., XU, M. and BÜHLMANN, P. (2011). High-dimensional covariance estimation based on Gaussian graphical models. *J*. *Mach*. *Learn*. *Res*. **12** 2975–3026. MR2854354

The Annals of Applied Statistics
2017, Vol. 11, No. 1, 69–92
DOI: 10.1214/16-AOAS991

MULTIVARIATE SPATIAL MAPPING OF SOIL WATER HOLDING CAPACITY WITH SPATIALLY VARYING CROSS-CORRELATIONS[1]

BY RACHEL M. MESSICK, MATTHEW J. HEATON AND NEIL HANSEN

Brigham Young University

Irrigation in agriculture mitigates the adverse effects of drought and improves crop production and yield. Still, water scarcity remains a persistent issue and water resources need to be used responsibly. To improve water use efficiency, precision irrigation is emerging as an approach where farmers can vary the application of water according to within field variation in soil and topographic conditions. As a precursor, methods to characterize spatial variation of soil hydraulic properties are needed. One such property is soil water holding capacity (WHC). This analysis develops a multivariate spatial model for predicting WHC across a field at various soil depths using sparse WHC observations and covariates such as soil electrical conductivity. To capture spatially varying cross-correlations in an efficient manner, we propose to extend the conditional specification of a multivariate Gaussian process by using spatially varying coefficients. Because data is already sparse, our analysis fully utilizes incomplete observations by imputing missing values that we treat as not missing at random. Additionally, due to the high cost of measuring WHC, we use a multivariate integrated mean square error criterion to choose a new observation location that, after sampling, will result in the least predictive uncertainty across the entire field.

REFERENCES

ALLEN, R. G., PEREIRA, L. S., RAES, D., SMITH, M. et al. (1998). Crop evapotranspiration—guidelines for computing crop water requirements—FAO irrigation and drainage paper 56. *FAO, Rome* **300** D05109.

APANASOVICH, T. and GENTON, M. G. (2010). Cross-covariance functions for multivariate random fields based on latent dimensions. *Biometrika* **97** 15–30. MR2594414

APANASOVICH, T. V., GENTON, M. G. and SUN, Y. (2012). A valid Matérn class of cross-covariance functions for multivariate random fields with any number of components. *J. Amer. Statist. Assoc.* **107** 180–193. MR2949350

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

BRUCE, R. R. and LUXMOORE, R. J. (1986). Water retention: Field methods. In *Methods of Soil Analysis*: *Part* 1—*Physical and Mineralogical Methods*, 2nd ed. (A. Klute, ed.). 663–686. Soil Science Society of America, American Society of Agronomy, Madison, WI.

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 ZAMMIT-MANGION, A. (2015). Multivariate spatial covariance models: A conditional approach. Available at arXiv:1504.01865v1.

Key words and phrases. Multivariate spatial processes, conditional specification, spatial design, Gaussian process, not missing at random.

CURRIN, C., MITCHELL, T., MORRIS, M. and YLVISAKER, D. (1991). Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *J. Amer. Statist. Assoc.* **86** 953–963. MR1146343

DIGGLE, P. J. and RIBEIRO, P. J. (2002). Bayesian inference in Gaussian model-based geostatistics. *Geogr. Environ. Model.* **6** 129–146.

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

FUENTES, M. and REICH, B. (2013). Multivariate spatial nonparametric modelling via kernel processes mixing. *Statist. Sinica* **23** 75–97. MR3076159

GELFAND, A. E. and BANERJEE, S. (2010). Multivariate spatial process models. In *Handbook of Spatial Statistics* 495–515. CRC Press, Boca Raton, FL. MR2730963

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

GELFAND, A. E., SCHMIDT, A. M., BANERJEE, S. and SIRMANS, C. F. (2004). Nonstationary multivariate process modeling through spatially varying coregionalization. *TEST* **13** 263–312. With discussion by Montserrat Fuentes, Dave Higdon and Bruno Sansó and a rejoinder by the authors. MR2154003

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

GENTON, M. G. and KLEIBER, W. (2015). Cross-covariance functions for multivariate geostatistics. *Statist. Sci.* **30** 147–163. MR3353096

GNEITING, T., KLEIBER, W. and SCHLATHER, M. (2010). Matérn cross-covariance functions for multivariate random fields. *J. Amer. Statist. Assoc.* **105** 1167–1177. MR2752612

GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc.* **102** 359–378. MR2345548

GUHANIYOGI, R., FINLEY, A. O., BANERJEE, S. and KOBE, R. K. (2013). Modeling complex spatial dependencies: Low-rank spatially varying cross-covariances with application to soil nutrient data. *J. Agric. Biol. Environ. Stat.* **18** 274–298. MR3110894

HEATON, M. J., CHRISTENSEN, W. F. and TERRES, M. A. (2017). Nonstationary Gaussian process models using spatial hierarchical clustering from finite differences. *Technometrics* **59** 93–101. MR3604192

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

JOHNSON, M. E., MOORE, L. M. and YLVISAKER, D. (1990). Minimax and maximin distance designs. *J. Statist. Plann. Inference* **26** 131–148. MR1079258

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. and CRESSIE, N. (2011). Bayesian inference for the spatial random effects model. *J. Amer. Statist. Assoc.* **106** 972–983. MR2894757

KITCHEN, N. R., DRUMMOND, S. T., LUND, E. D., SUDDUTH, K. A. and BUCHLEITER, G. W. (2003). Soil electrical conductivity and topography related to yield for three contrasting soil–crop systems. *Agron. J.* **95** 483–495.

KLEIBER, W. and GENTON, M. G. (2013). Spatially varying cross-correlation coefficients in the presence of nugget effects. *Biometrika* **100** 213–220. MR3034334

KLEIBER, W., SAIN, S. R., HEATON, M. J., WILTBERGER, M., REESE, C. S. and BINGHAM, D. (2013). Parameter tuning for a multi-fidelity dynamical model of the magnetosphere. *Ann. Appl. Stat.* **7** 1286–1310. MR3127948

KLUTE, A. (1986). Water retention: Laboratory methods. In *Methods of Soil Analysis*: *Part* 1—*Physical and Mineralogical Methods*, 2nd ed. (A. Klute, ed.) 635–662. Soil Science Society of America, American Society of Agronomy, Madison, WI.

KRAUSE, A., SINGH, A. and GUESTRIN, C. (2008). Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *J. Mach. Learn. Res.* **9** 235–284.

LEMOS, R. T. and SANSÓ, B. (2009). A spatio-temporal model for mean, anomaly, and trend fields of North Atlantic sea surface temperature. *J. Amer. Statist. Assoc.* **104** 5–18. MR2662306

LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ. MR1925014

LONGCHAMPS, L., KHOSLA, R., REICH, R. and GUI, D. W. (2015). Spatial and temporal variability of soil water content in leveled fields. *Soil Sci. Soc. Amer. J.* **79** 1446–1454.

MAJUMDAR, A., PAUL, D. and BAUTISTA, D. (2010). A generalized convolution model for multivariate nonstationary spatial processes. *Statist. Sinica* **20** 675–695. MR2682636

MZUKU, M., KHOSLA, R., REICH, R., INMAN, D., SMITH, F. and MACDONALD, L. (2005). Spatial variability of measured soil properties across site-specific management zones. *Soil Sci. Soc. Amer. J.* **69** 1572–1579.

NATURAL RESOURCES CONSERVATION SERVICE (1997). National Engineering Handbook: Irrigation guide. U.S. Department of Agriculture.

NATURAL RESOURCES CONSERVATION SERVICE (2016). National Soil Survey Handbook. U.S. Department of Agriculture.

NYCHKA, D., BANDYOPADHYAY, S., HAMMERLING, D., LINDGREN, F. and SAIN, S. (2015). A multiresolution Gaussian process model for the analysis of large spatial datasets. *J. Comput. Graph. Statist.* **24** 579–599. MR3357396

PLASTER, E. (2013). *Soil Science and Management*. Cengage Learning, Independence, KY.

RANJAN, P., LU, W., BINGHAM, D., REESE, S., WILLIAMS, B. J., CHOU, C.-C., DOSS, F., GROSSKOPF, M. and HOLLOWAY, J. P. (2011). Follow-up experimental designs for computer models and physical processes. *J. Stat. Theory Pract.* **5** 119–136. MR2829827

ROYLE, J. A. and BERLINER, L. M. (1999). A hierarchical approach to multivariate spatial modeling and prediction. *J. Agric. Biol. Environ. Stat.* **4** 29–56. MR1812239

SACKS, J., SCHILLER, S. B. and WELCH, W. J. (1989). Designs for computer experiments. *Technometrics* **31** 41–47. MR0997669

SADLER, E. J., EVANS, R., STONE, K. C. and CAMP, C. R. (2005). Opportunities for conservation with precision irrigation. *J. Soil Water Conserv.* **60** 371–378.

SANG, H., JUN, M. and HUANG, J. Z. (2011). Covariance approximation for large multivariate spatial data sets with an application to multiple climate model errors. *Ann. Appl. Stat.* **5** 2519–2548. MR2907125

SANTNER, T. J., WILLIAMS, B. J. and NOTZ, W. I. (2003). *The Design and Analysis of Computer Experiments*. Springer, New York. MR2160708

SEAGER, R., HOERLING, M., SCHUBERT, S., WANG, H., LYON, B., KUMAR, A., NAKAMURA, J. and HENDERSON, N. (2014). Causes and predictability of the 2011–14 California drought. Assessment report, National Oceanic and Atmospheric Administration, Silver Spring, MD.

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

The Annals of Applied Statistics
2017, Vol. 11, No. 1, 93–113
DOI: 10.1214/16-AOAS992

COVARIATE-ADAPTIVE CLUSTERING OF EXPOSURES FOR AIR POLLUTION EPIDEMIOLOGY COHORTS[1]

BY JOSHUA P. KELLER*, MATHIAS DRTON*, TIMOTHY LARSON*, JOEL D. KAUFMAN*, DALE P. SANDLER† AND ADAM A. SZPIRO*

University of Washington and National Institute of Environmental Health Sciences†*

Cohort studies in air pollution epidemiology aim to establish associations between health outcomes and air pollution exposures. Statistical analysis of such associations is complicated by the multivariate nature of the pollutant exposure data as well as the spatial misalignment that arises from the fact that exposure data are collected at regulatory monitoring network locations distinct from cohort locations. We present a novel clustering approach for addressing this challenge. Specifically, we present a method that uses geographic covariate information to cluster multi-pollutant observations and predict cluster membership at cohort locations. Our predictive k-means procedure identifies centers using a mixture model and is followed by multiclass spatial prediction. In simulations, we demonstrate that predictive k-means can reduce misclassification error by over 50% compared to ordinary k-means, with minimal loss in cluster representativeness. The improved prediction accuracy results in large gains of 30% or more in power for detecting effect modification by cluster in a simulated health analysis. In an analysis of the NIEHS Sister Study cohort using predictive k-means, we find that the association between systolic blood pressure (SBP) and long-term fine particulate matter ($PM_{2.5}$) exposure varies significantly between different clusters of $PM_{2.5}$ component profiles. Our cluster-based analysis shows that, for subjects assigned to a cluster located in the Midwestern U.S., a 10 $\mu g/m^3$ difference in exposure is associated with 4.37 mmHg (95% CI, 2.38, 6.35) higher SBP.

REFERENCES

ADAR, S. D., KLEIN, R., KLEIN, B. E. K., SZPIRO, A. A., COTCH, M. F., WONG, T. Y., O'NEILL, M. S., SHRAGER, S., BARR, R. G., SISCOVICK, D. S., DAVIGLUS, M. L., SAMPSON, P. D. and KAUFMAN, J. D. (2010). Air pollution and the microvasculature: A cross-sectional assessment of in vivo retinal images in the population-based multi-ethnic study of atherosclerosis (MESA). *PLoS Medicine* **7** e1000372.

AUSTIN, E., COULL, B., THOMAS, D. and KOUTRAKIS, P. (2012). A framework for identifying distinct multipollutant profiles in air pollution data. *Environment International* **45** 112–121.

AUSTIN, E., COULL, B. A., ZANOBETTI, A. and KOUTRAKIS, P. (2013). A framework to spatially cluster air pollution monitoring sites in US based on the PM2.5 composition. *Environment International* **59** 244–254.

BELL, M. L., DOMINICI, F., EBISU, K., ZEGER, S. L. and SAMET, J. M. (2007). Spatial and temporal variation in PM2.5 chemical composition in the United States for health effects studies. *Environmental Health Perspectives* **115** 989–995.

Key words and phrases. Air pollution, clustering, dimension reduction, particulate matter.

BERGEN, S. and SZPIRO, A. A. (2015). Mitigating the impact of measurement error when using penalized regression to model exposure in two-stage air pollution epidemiology studies. *Environ. Ecol. Stat.* **22** 601–631. MR3384910

BERK, R., BROWN, L. and ZHAO, L. (2010). Statistical inference after model selection. *Journal of Quantitative Criminology* **26** 217–236.

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

BLANCHARD, C. L. and HIDY, G. M. (2003). Effects of changes in sulfate, ammonia, and nitric acid on particulate nitrate concentrations in the southeastern United States. *Journal of the Air & Waste Management Association* **53** 283–290.

BRAUER, M. (2010). How much, how long, what, and where: Air pollution exposure assessment for epidemiologic studies of respiratory disease. *Proc. Am. Thorac. Soc.* **7** 111–115.

BRAUER, M., HOEK, G., VAN VLIET, P., MELIEFSTE, K., FISCHER, P., GEHRING, U., HEINRICH, J., CYRYS, J., BELLANDER, T., LEWNE, M. and BRUNEKREEF, B. (2003). Estimating long-term average particulate air pollution concentrations: Application of traffic indicators and geographic information systems. *Epidemiology* **14** 228–239.

BROOK, R. D., RAJAGOPALAN, S., POPE, C. A., BROOK, J. R., BHATNAGAR, A., DIEZ-ROUX, A. V., HOLGUIN, F., HONG, Y., LUEPKER, R. V., MITTLEMAN, M. A., PETERS, A., SISCOVICK, D., SMITH, S. C., WHITSEL, L. and KAUFMAN, J. D. (2010). Particulate matter air pollution and cardiovascular disease: An update to the scientific statement from the American Heart Association. *Circulation* **121** 2331–2378.

CHAN, S. H., VAN HEE, V. C., BERGEN, S., SZPIRO, A. A., DEROO, L. A., LONDON, S. J., MARSHALL, J. D., KAUFMAN, J. D. and SANDLER, D. P. (2015). Long-term air pollution exposure and blood pressure in the Sister Study. *Environmental Health Perspectives* **123** 951–958.

COHEN, M. A., ADAR, S. D., ALLEN, R. W., AVOL, E., CURL, C. L., GOULD, T., HARDIE, D., HO, A., KINNEY, P., LARSON, T. V., SAMPSON, P., SHEPPARD, L., STUKOVSKY, K. D., SWAN, S. S., LIU, L. J. S. and KAUFMAN, J. D. (2009). Approach to estimating participant pollutant exposures in the multi-ethnic study of atherosclerosis and air pollution (MESA Air). *Environmental Science & Technology* **43** 4687–4693.

DOMINICI, F., SHEPPARD, L. and CLYDE, M. (2003). Health effects of air pollution: A statistical review. *Int. Stat. Rev.* **71** 243–276.

FRANKLIN, M., KOUTRAKIS, P. and SCHWARTZ, J. (2008). The role of particle composition on the association between PM2.5 and mortality. *Epidemiology* **19** 680–689.

HARTIGAN, J. and WONG, M. (1979). Algorithm AS 136: A k-means clustering algorithm. *Applied Statistics* **28** 100–108.

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

KELLER, J. P., OLIVES, C., KIM, S.-Y., SHEPPARD, L., SAMPSON, P. D., SZPIRO, A. A., ORON, A. P., LINDSTRÖM, J., VEDAL, S. and KAUFMAN, J. D. (2015). A unified spatiotemporal modeling approach for predicting concentrations of multiple air pollutants in the multi-ethnic study of atherosclerosis and air pollution. *Environmental Health Perspectives* **123** 301–309.

KELLER, J. P., DRTON, M., LARSON, T. V., KAUFMAN, J. D., SANDLER, D. P. and SZPIRO, A. A. (2017). Supplement to "Covariate-adaptive clustering of exposures for air pollution epidemiology cohorts." DOI:10.1214/16-AOAS992SUPP.

KIOUMOURTZOGLOU, M.-A., AUSTIN, E., KOUTRAKIS, P., DOMINICI, F., SCHWARTZ, J. and ZANOBETTI, A. (2015). PM2.5 and survival among older adults: Effect modification by particulate composition. *Epidemiology* **26** 321–327.

KÜNZLI, N., MEDINA, S. and KAISER, R. (2001). Assessment of deaths attributable to air pollution: Should we use risk estimates based on time series or on cohort studies? *Am. J. Epidemiol.* **153** 1050–1055.

LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927. MR3485948

OAKES, M., BAXTER, L. and LONG, T. C. (2014). Evaluating the application of multipollutant exposure metrics in air pollution health studies. *Environment International* **69** 90–99.

PELTIER, R. E., HSU, S.-I., LALL, R. and LIPPMANN, M. (2009). Residual oil combustion: A major source of airborne nickel in New York city. *Journal of Exposure Science & Environmental Epidemiology* **19** 603–612.

SAMPSON, P. D., RICHARDS, M., SZPIRO, A. A., BERGEN, S., SHEPPARD, L., LARSON, T. V. and KAUFMAN, J. D. (2013). A regionalized national universal kriging model using partial least squares regression for estimating annual PM2.5 concentrations in epidemiology. *Atmospheric Environment* **75** 383–392.

SHACKLETTE, H. T. and BOERNGEN, J. (1984). Element concentrations in soils and other surficial materials of the conterminous United States. Technical report.

THURSTON, G. D., ITO, K., LALL, R., BURNETT, R. T., TURNER, M. C., KREWSKI, D., SHI, Y., JERRETT, M., GAPSTUR, S. M., DIVER, W. R. and POPE, C. A. (2013). NPACT study 4. mortality and long-term exposure to PM2.5 and its components in the American cancer society's cancer prevention study II cohort. In *National Particle Component Toxicity* (*NPACT*) *Initiative*: *Integrated Epidemiologic and Toxicologic Studies of the Health Effects of Particulate Matter Components*. *Research Report* 177. Health Effects Institute, Boston, MA.

U.S. EPA (2003). Compilation of existing studies on source apportionment for PM2.5. Technical report, Office of Air Quality Planning and Standards, Washington, DC.

U.S. EPA (2006). Chapter 4: Air quality impacts. In *Regulatory Impact Analysis, 2006 National Ambient Air Quality Standards for Particle Pollution*. Research Triangle Park, NC.

WILSON, J. G., KINGHAM, S., PEARCE, J. and STURMAN, A. P. (2005). A review of intraurban variations in particulate air pollution: Implications for epidemiological research. *Atmospheric Environment* **39** 6444–6462.

ZANOBETTI, A., FRANKLIN, M., KOUTRAKIS, P. and SCHWARTZ, J. (2009). Fine particulate air pollution and its components in association with cause-specific emergency admissions. *Environmental Health* **8** 58.

The Annals of Applied Statistics
2017, Vol. 11, No. 1, 114–138
DOI: 10.1214/16-AOAS994

A MIXED-EFFECTS MODEL FOR INCOMPLETE DATA FROM LABELING-BASED QUANTITATIVE PROTEOMICS EXPERIMENTS

BY LIN S. CHEN[*,1,2], JIEBIAO WANG[*,1,2], XIANLONG WANG[†,3] AND PEI WANG[‡,1,3,4]

University of Chicago[*], Fred Hutchinson Cancer Research Center[†] and Icahn School of Medicine at Mount Sinai[‡]

In mass spectrometry (MS) based quantitative proteomics research, the emerging iTRAQ (isobaric tag for relative and absolute quantitation) and TMT (tandem mass tags) techniques have been widely adopted for high throughput protein profiling. In a typical iTRAQ/TMT proteomics study, samples are grouped into batches, and each batch is processed by one multiplex experiment, in which the abundances of thousands of proteins/peptides in a batch of samples can be measured simultaneously. The multiplex labeling technique greatly enhances the throughput of protein quantification. However, the technical variation across different iTRAQ/TMT multiplex experiments is often large due to the dynamic nature of MS instruments. This leads to strong batch effects in the iTRAQ/TMT data. Moreover, the iTRAQ/TMT data often contain substantial batch-level nonignorable missing entries. Specifically, the abundance measures of a given protein/peptide are often either observed or missing altogether in all the samples from the same batch, with the missing probability depending on the combined batch-level abundances. We term this unique missing-data mechanism as the Batch-level Abundance-Dependent Missing-data Mechanism (BADMM). We introduce a new method—`mixEMM`—for analyzing iTRAQ/TMT data with batch effects and batch-level nonignorable missingness. The `mixEMM` method employs a linear mixed-effects model and explicitly models the batch effects and the BADMM. With simulation studies, we showed that, compared with existing approaches that utilize relative abundances and ignore the missing batches under the missing-completely-at-random assumption, the `mixEMM` method achieves more accurate parameter estimation and inference. We applied the method to an iTRAQ proteomics data from a breast cancer study and identified phosphopeptides differentially expressed between different breast cancer subtypes. The method can be applied to general clustered data with cluster-level nonignorable missing-data mechanisms.

REFERENCES

BERNARDO, G. M., BEBEK, G., GINTHER, C. L., SIZEMORE, S. T., LOZADA, K. L., MIEDLER, J. D., ANDERSON, L. A., GODWIN, A. K., ABDUL-KARIM, F. W., SLAMON, D. J. and KERI, R. A. (2013). FOXA1 represses the molecular phenotype of basal breast cancer cells. *Oncogene* **32** 554–563.

Key words and phrases. Mixed-effects models, the expectation-conditional-maximization (`ECM`) algorithm, Batch-level Abundance-Dependent Missing-data Mechanism (BADMM).

CHANG, C.-Y., PICOTTI, P., HÜTTENHAIN, R., HEINZELMANN-SCHWARZ, V., JOVANOVIC, M., AEBERSOLD, R. and VITEK, O. (2012). Protein significance analysis in selected reaction monitoring (SRM) measurements. *Mol. Cell. Proteomics* **11** M111.014662.

CHEN, L. S., PRENTICE, R. L. and WANG, P. (2014). A penalized EM algorithm incorporating missing data mechanism for Gaussian parameter estimation. *Biometrics* **70** 312–322. MR3258036

CIMINO-MATHEWS, A., SUBHAWONG, A. P., ELWOOD, H., WARZECHA, H. N., SHARMA, R., PARK, B. H., TAUBE, J. M., ILLEI, P. B. and ARGANI, P. (2013). Neural crest transcription factor Sox10 is preferentially expressed in triple-negative and metaplastic breast carcinomas. *Human Pathol.* **44** 959–965.

CLOUGH, T., KEY, M., OTT, I., RAGG, S., SCHADOW, G. and VITEK, O. (2009). Protein quantification in label-free LC-MS experiments. *J. Proteome Res.* **8** 5275–5284.

DOBBIN, K. and SIMON, R. (2002). Comparison of microarray designs for class comparison and class discovery. *Bioinformatics* **18** 1438–1445.

ELLIS, M., GILLETTE, M., CARR, S., PAULOVICH, A., SMITH, R., RODLAND, K., TOWNSEND, R., KINSINGER, C., MESRI, M., RODRIGUEZ, H., LIEBLER, D. and CPTAC (2013). Connecting genomic alterations to cancer biology with proteomics: The NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer Discovery* **3** 1108–1112.

FRANKEN, H., MATHIESON, T., CHILDS, D., SWEETMAN, G. M. A., WERNER, T., TÖGEL, I., DOCE, C., GADE, S., BANTSCHEFF, M., DREWES, G., REINHARD, F. B. M., HUBER, W. and SAVITSKI, M. M. (2015). Thermal proteome profiling for unbiased identification of direct and indirect drug targets using multiplexed quantitative mass spectrometry. *Nat. Protoc.* **10** 1567–1593.

HILL, E. G., SCHWACKE, J. H., COMTE-WALTERS, S., SLATE, E. H., OBERG, A. L., ECKEL-PASSOW, J. E., THERNEAU, T. M. and SCHEY, K. L. (2008). A statistical model for iTRAQ data analysis. *J. Proteome Res.* **7** 3091–3101.

IBRAHIM, J. G. and MOLENBERGHS, G. (2009). Missing data methods in longitudinal studies: A review. *TEST* **18** 1–43. MR2495958

IVANOV, S. V., PANACCIONE, A., NONAKA, D., PRASAD, M. L., BOYD, K. L., BROWN, B., GUO, Y., SEWELL, A. and YARBROUGH, W. G. (2013). Diagnostic SOX10 gene signatures in salivary adenoid cystic and breast basal-like carcinomas. *Br. J. Cancer* **109** 444–451.

KARP, N. A., HUBER, W., SADOWSKI, P. G., CHARLES, P. D., HESTER, S. V. and LILLEY, K. S. (2010). Addressing accuracy and precision issues in iTRAQ quantitation. *Mol. Cell. Proteomics* **9** 1885–1897.

LAIRD, N. M. and WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38** 963–974.

LIEBLER, D., ZHANG, B., WANG, J., WANG, X., ZHU, J., LIU, Q., SHI, Z., CHAMBERS, M. C. et al. (2014). Proteogenomic characterization of human colon and rectal cancer. *Nature* **513** 382–387.

LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley, Hoboken, NJ. MR1925014

LUO, R., COLANGELO, C. M., SESSA, W. C. and ZHAO, H. (2009). Bayesian analysis of iTRAQ data with nonrandom missingness: Identification of differentially expressed proteins. *Statistics in Biosciences* **1** 228–245.

MCALISTER, G. C., NUSINOW, D. P., JEDRYCHOWSKI, M. P., WÜHR, M., HUTTLIN, E. L., ERICKSON, B. K., RAD, R., HAAS, W. and GYGI, S. P. (2014). MultiNotch MS3 enables accurate, sensitive, and multiplexed detection of differential expression across cancer cell line proteomes. *Analytical Chemistry* **86** 7150–7158.

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

MERTINS, P., MANI, D. R., RUGGLES, K. V., GILLETTE, M. A., CLAUSER, K. R., WANG, P. et al. (2016). Proteogenomic connects somatic mutations to signaling in breast cancer. *Nature* **534** 55–62.

MEYER, K. B. and CARROLL, J. S. (2012). FOXA1 and breast cancer risk. *Nat. Genet.* **44** 1176–1177.

THE CANCER GENOME ATLAS NETWORK (2012). Comprehensive molecular portraits of human breast tumours. *Nature* **490** 61–70.

OBERG, A. L., MAHONEY, D. W., ECKEL-PASSOW, J. E., MALONE, C. J., WOLFINGER, R. D., HILL, E. G., COOPER, L. T., ONUMA, O. K., SPIRO, C., THERNEAU, T. M. and BERGEN, H. (2008). Statistical analysis of relative labeled mass spectrometry data from complex samples using ANOVA. *J. Proteome Res.* **7** 225–233.

PAULO, J. A., MCALLISTER, F. E., EVERLEY, R. A., BEAUSOLEIL, S. A., BANKS, A. S. and GYGI, S. P. (2014). Effects of MEK inhibitors GSK1120212 and PD0325901 in vivo using 10-plex quantitative proteomics and phosphoproteomics. *Proteomics* **15** 462–473.

PAULOVICH, A. G., BILLHEIMER, D., HAM, A. J., VEGA-MONTOTO, L., RUDNICK, P. A., TABB, D. L., WANG, P., BLACKMAN, R. K., BUNK, D. M. and CARDASIS, H. ET AL. (2010). Interlaboratory study characterizing a yeast performance standard for benchmarking LC-MS platform performance. *Mol. Cell. Proteomics* **9** 242–254.

PINHEIRO, J. C. and BATES, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *J. Comput. Graph. Statist.* **4** 12–35.

RAUNIYAR, N. and YATES III, J. R. (2014). Isobaric labeling-based relative quantification in shotgun proteomics. *J. Proteome Res.* **13** 5293–5309.

ROSS, P. L., HUANG, Y. N., MARCHESE, J. N., WILLIAMSON, B., PARKER, K., HATTAN, S., KHAINOVSKI, N., PILLAI, S., DEY, S., DANIELS, S., PURKAYASTHA, S., JUHASZ, P., MARTIN, S., BARTLET-JONES, M., HE, F., JACOBSON, A. and PAPPIN, D. J. (2004). Multiplexed protein quantitation in saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell. Proteomics* **3** 1154–1169.

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

SZKLARCZYK, D., FRANCESCHINI, A., WYDER, S., FORSLUND, K., HELLER, D., HUERTA-CEPAS, J., SIMONOVIC, M., ROTH, A., SANTOS, A., TSAFOU, K. P. et al. (2014). STRING v10: Protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* gku1003.

WANG, P., TANG, H., ZHANG, H., WHITEAKER, J., PAULOVICH, A. G. and MCINTOSH, M. (2006). Normalization regarding non-random missing values in high-throughput mass spectrometry data. *Pacific Symposium on Biocomputing* 315–326.

WERNER, T., SWEETMAN, G., SAVITSKI, M. F., MATHIESON, T., BANTSCHEFF, M. and SAVITSKI, M. M. (2014). Ion coalescence of neutron encoded TMT 10-plex reporter ions. *Anal. Chem.* **86** 3594–3601.

WIESE, S., REIDEGELD, K. A., MEYER, H. E. and WARSCHEID, B. (2007). Protein labeling by iTRAQ: A new tool for quantitative mass spectrometry in proteome research. *Proteomics* **7** 340–350.

ZHANG, H., LIU, T., ZHANG, Z., PAYNE, S. H., ZHANG, B. and MCDERMOTT, J. E. et al. (2016). Integrated proteogenomic characterization of human high-grade serous ovarian cancer. *Cell* **166** 755–765.

The Annals of Applied Statistics
2017, Vol. 11, No. 1, 139–160
DOI: 10.1214/16-AOAS995

STATIC AND ROVING SENSOR DATA FUSION FOR SPATIO-TEMPORAL HAZARD MAPPING WITH APPLICATION TO OCCUPATIONAL EXPOSURE ASSESSMENT[1]

BY GUILHERME LUDWIG*, TINGJIN CHU†, JUN ZHU*, HAONAN WANG‡ AND KIRSTEN KOEHLER§

*University of Wisconsin-Madison**, *Renmin University of China*†, *Colorado State University*‡ *and Johns Hopkins University*§

Rapid technological advances have drastically improved the data collection capacity in occupational exposure assessment. However, advanced statistical methods for analyzing such data and drawing proper inference remain limited. The objectives of this paper are (1) to provide new spatio-temporal methodology that combines data from both roving and static sensors for data processing and hazard mapping across space and over time in an indoor environment, and (2) to compare the new method with the current industry practice, demonstrating the distinct advantages of the new method and the impact on occupational hazard assessment and future policy making in environmental health as well as occupational health. A novel spatio-temporal model with a continuous index in both space and time is proposed, and a profile likelihood-based model fitting procedure is developed that allows fusion of the two types of data. To account for potential differences between the static and roving sensors, we extend the model to have nonhomogenous measurement error variances. Our methodology is applied to a case study conducted in an engine test facility, and dynamic hazard maps are drawn to show features in the data that would have been missed by existing approaches, but are captured by the new method.

REFERENCES

CARAGEA, P. C. and SMITH, R. L. (2007). Asymptotic properties of computationally efficient alternative estimators for a class of multivariate normal models. *J. Multivariate Anal.* **98** 1417–1440. MR2364128

CHERRIE, J. W. (2003). Commentary: The beginning of the science underpinning occupational hygiene. *Ann. Occup. Hyg.* **47** 179–185.

CHU, T., WANG, H. and ZHU, J. (2014). On semiparametric inference of geostatistical models via local Karhunen–Loève expansion. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 817–832. MR3248678

CHU, T., ZHU, J. and WANG, H. (2011). Penalized maximum likelihood estimation and variable selection in geostatistics. *Ann. Statist.* **39** 2607–2625. MR2906880

COWLES, M. K., ZIMMERMAN, D. L., CHRIST, A. and MCGINNIS, D. L. (2002). Combining snow water equivalent data from multiple sources to estimate spatio-temporal trends and compare measurement systems. *J. Agric. Biol. Environ. Stat.* **7** 536–557.

Key words and phrases. Geostatistics, kriging, semiparametric methods, spatial statistics, spatio-temporal statistics.

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

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

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

EVANS, D. E., HEITBRINK, W. A., SLAVIN, T. J. and PETERS, T. M. (2008). Ultrafine and respirable particles in an automotive grey iron foundry. *Ann. Occup. Hyg.* **52** 9–21.

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

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

GROMENKO, O. and KOKOSZKA, P. (2013). Nonparametric inference in small data sets of spatially indexed curves with application to ionospheric trend determination. *Comput. Statist. Data Anal.* **59** 82–94. MR3000043

HALL, D. L. and MCMULLEN, S. A. (2004). *Mathematical Techniques in Multisensor Data Fusion*. Artech House, Boston.

ISAACSON, J. D. and ZIMMERMAN, D. L. (2000). Combining temporally correlated environmental data from two measurement systems. *J. Agric. Biol. Environ. Stat.* **5** 398–416. MR1812083

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

KOEHLER, K. A. and PETERS, T. M. (2013). Influence of analysis methods on interpretation of hazard maps. *Ann. Occup. Hyg.* **57** 558–570.

KOEHLER, K. A. and VOLCKENS, J. (2011). Prospects and pitfalls of occupational hazard mapping: 'between these lines there be dragons'. *Ann. Occup. Hyg.* **55** 829–840.

LAKE, K., ZHU, J., WANG, H., VOLCKENS, J. and KOEHLER, K. A. (2015). Effects of data sparsity and spatiotemporal variability on hazard maps of workplace noise. *J. Occup. Environ. Hyg.* **12** 256–265.

LUDWIG, G., CHU, T., KOEHLER, K., WANG, H. and ZHU, J. (2017a). Supplement to "Static and Roving Sensor Data Fusion for Spatio-Temporal Hazard Mapping with Application to Occupational Exposure Assessment." DOI:10.1214/16-AOAS995SUPPA.

LUDWIG, G., CHU, T., KOEHLER, K., WANG, H. and ZHU, J. (2017b). Supplement to "Static and Roving Sensor Data Fusion for Spatio-Temporal Hazard Mapping with Application to Occupational Exposure Assessment." DOI:10.1214/16-AOAS995SUPPB.

MA, C. (2003). Families of spatio-temporal stationary covariance models. *J. Statist. Plann. Inference* **116** 489–501. MR2000096

O'BRIEN, D. M. (2003). Aerosol mapping of a facility with multiple cases of hypersensitivity pneumonitis: Demonstration of mist reduction and a possible dose/response relationship. *Appl. Occup. Environ. Hyg.* **18** 947–952.

OLOGE, F. E., AKANDE, T. M. and OLAJIDE, T. G. (2006). Occupational noise exposure and sensorineural hearing loss among workers of a steel rolling Mill. *Eur. Arch. Oto-Rhino-Laryngol.* **263** 618–621.

PETERS, T. M., HEITBRINK, W. A., EVANS, D. E., SLAVIN, T. J. and MAYNARD, A. D. (2006). The mapping of fine and ultrafine particle concentrations in an engine machining and assembly facility. *Ann. Occup. Hyg.* **50** 249–257.

PETERS, T. M., ANTHONY, T. R., TAYLOR, C., ALTMAIER, R., ANDERSON, K. and T O'SHAUGHNESSY, P. (2012). Distribution of particle and gas concentrations in swine gestation confined animal feeding operations. *Ann. Occup. Hyg.* **56** 1080–1090.

QUICK, H., BANERJEE, S. and CARLIN, B. P. (2015). Bayesian modeling and analysis for gradients in spatiotemporal processes. *Biometrics* **71** 575–584. MR3402593

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

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

SMITH, B. J. and COWLES, M. K. (2007). Correlating point-referenced Radon and areal uranium data arising from a common spatial process. *J. Roy. Statist. Soc. Ser. C* **56** 313–326. MR2370992

STEIN, M. L. (1999). *Interpolation of Spatial Data*: *Some Theory for Kriging*. Springer, New York. MR1697409

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

TORNERO-VELEZ, R., SYMANSKI, E., KROMHOUT, H., YU, R. C. and RAPPAPORT, S. M. (1997). Compliance versus risk in assessing occupational exposures. *Risk Anal.* **17** 279–292.

TRACEY, J. A., SHEPPARD, J., ZHU, J., WEI, F., SWAISGOOD, R. R. and FISHER, R. N. (2014). Movement-based estimation and visualization of space use in 3D for wildlife ecology and conservation. *PLoS ONE* **9** e101205.

WAHBA, G. (1990). *Spline Models for Observational Data*. *CBMS-NSF Regional Conference Series in Applied Mathematics* **59**. SIAM, Philadelphia, PA. MR1045442

YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005). Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33** 2873–2903. MR2253106

The Annals of Applied Statistics
2017, Vol. 11, No. 1, 161–184
DOI: 10.1214/16-AOAS998

A STATISTICAL FRAMEWORK FOR DATA INTEGRATION THROUGH GRAPHICAL MODELS WITH APPLICATION TO CANCER GENOMICS

BY YUPING ZHANG[1], ZHENGQING OUYANG[2] AND HONGYU ZHAO[3]

University of Connecticut, The Jackson Laboratory for Genomic Medicine and Yale University

Recent advances in high-throughput biotechnologies have generated various types of genetic, genomic, epigenetic, transcriptomic and proteomic data across different biological conditions. It is likely that integrating data from diverse experiments may lead to a more unified and global view of biological systems and complex diseases. We present a coherent statistical framework for integrating various types of data from distinct but related biological conditions through graphical models. Specifically, our statistical framework is designed for modeling multiple networks with shared regulatory mechanisms from heterogeneous high-dimensional datasets. The performance of our approach is illustrated through simulations and its applications to cancer genomics.

REFERENCES

ALBERT, R., JEONG, H. and BARABÁSI, A.-L. (2000). Error and attack tolerance of complex networks. *Nature* **406** 378–382.

AUSLENDER, A. and TEBOULLE, M. (2006). Interior gradient and proximal methods for convex and conic optimization. *SIAM J. Optim.* **16** 697–725 (electronic). MR2197553

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

BECK, A. and TEBOULLE, M. (2009). Gradient-based algorithms with applications to signal recovery. *Convex Optim. Signal Process. Commun.* 42–88.

BOYD, S., PARIKH, N., CHU, E., PELEATO, B. and ECKSTEIN, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **3** 1–122.

CHEN, X., SLACK, F. J. and ZHAO, H. (2013). Joint analysis of expression profiles from multiple cancers improves the identification of microRNA–gene interactions. *Bioinformatics* **29** 2137–2145.

CHEN, S., WITTEN, D. M. and SHOJAIE, A. (2015). Selection and estimation for mixed graphical models. *Biometrika* **102** 47–64.

CHENG, J., LEVINA, E. and ZHU, J. (2013). High-dimensional mixed graphical models. Preprint. Available at arXiv:1304.2810.

CHUN, H., CHEN, M., LI, B. and ZHAO, H. (2013). Joint conditional Gaussian graphical models with multiple sources of genomic data. *Front. Genet.* **4** Article ID 294. DOI:10.3389/fgene.2013.00294.

Key words and phrases. Cancer genomics, data integration, graphical models.

CIRIELLO, G., MILLER, M. L., AKSOY, B. A., SENBABAOGLU, Y., SCHULTZ, N. and SANDER, C. (2013). Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* **45** 1127–1133.

DANAHER, P., WANG, P. and WITTEN, D. M. (2013). The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 373–397. MR3164871

FELLINGHAUER, B., BÜHLMANN, P., RYFFEL, M., VON RHEIN, M. and REINHARDT, J. D. (2013). Stable graphical model estimation with random forests for discrete, continuous, and mixed variables. *Comput. Statist. Data Anal.* **64** 132–152. MR3061894

FENG, Z., ZHANG, H., LEVINE, A. J. and JIN, S. (2005). The coordinate regulation of the p53 and mTOR pathways in cells. *Proc. Natl. Acad. Sci. USA* **102** 8204–8209.

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

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2009). Glmnet: Lasso and elastic-net regularized generalized linear models. R Package Version 1.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). A note on the group lasso and a sparse group lasso. Technical report, Dept. Statistics, Stanford Univ., Stanford.

GE, H., WALHOUT, A. J. and VIDAL, M. (2003). Integrating 'omic' information: A bridge between genomics and systems biology. *Trends Genet.* **19** 551–560.

GOVINDAN, R. and TANGMUNARUNKIT, H. (2000). Heuristics for Internet map discovery. In *Proceedings IEEE INFOCOM* 2000. *Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies* **3** 1371–1380. IEEE, New York.

GUO, J., LEVINA, E., MICHAILIDIS, G. and ZHU, J. (2010). Joint structure estimation for categorical Markov networks. Technical report, Dept. Statistics, Univ. of Michigan, Ann Arbor.

GUO, J., LEVINA, E., MICHAILIDIS, G. and ZHU, J. (2011). Joint estimation of multiple graphical models. *Biometrika* **98** 1–15. MR2804206

HAWKINS, R. D., HON, G. C. and REN, B. (2010). Next-generation genomics: An integrative approach. *Nat. Rev. Genet.* **11** 476–486.

HECKER, M., LAMBECK, S., TOEPFER, S., VAN SOMEREN, E. and GUTHKE, R. (2009). Gene regulatory network inference: Data integration in dynamic models—A review. *Biosystems* **96** 86–103.

HESTENES, M. R. (1969). Multiplier and gradient methods. *J. Optim. Theory Appl.* **4** 303–320. MR0271809

HOEFLING, H. (2010). A path algorithm for the fused lasso signal approximator. *J. Comput. Graph. Statist.* **19** 984–1006. Supplementary materials available online. MR2791265

HÖFLING, H. and TIBSHIRANI, R. (2009). Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. *J. Mach. Learn. Res.* **10** 883–906. MR2505138

JEONG, H., MASON, S. P., BARABÁSI, A-L. and OLTVAI, Z. N. (2001). Lethality and centrality in protein networks. *Nature* **411** 41–42.

JOYCE, A. R. and PALSSON, B. Ø. (2006). The model organism as a system: Integrating "omics" data sets. *Nat. Rev., Mol. Cell Biol.* **7** 198–210.

LAURITZEN, S. L. (1996). *Graphical Models. Oxford Statistical Science Series* **17**. Oxford Univ. Press, New York. MR1419991

LEE, J. D. and HASTIE, T. J. (2012). Learning mixed graphical models. Preprint. Available at arXiv:1205.5012.

LI, B., CHUN, H. and ZHAO, H. (2012). Sparse estimation of conditional graphical models with application to gene networks. *J. Amer. Statist. Assoc.* **107** 152–167. MR2949348

MAZUMDER, R. and HASTIE, T. (2012). Exact covariance thresholding into connected components for large-scale graphical lasso. *J. Mach. Learn. Res.* **13** 781–794. MR2913718

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

MYERS, C. L. and TROYANSKAYA, O. G. (2007). Context-sensitive data integration and prediction of biological networks. *Bioinformatics* **23** 2322–2330.

MYERS, C. L., ROBSON, D., WIBLE, A., HIBBS, M. A., CHIRIAC, C., THEESFELD, C. L., DOLINSKI, K. and TROYANSKAYA, O. G. (2005). Discovery of biological networks from diverse functional genomic data. *Genome Biol.* **6** Article ID R114. DOI:10.1186/gb-2005-6-13-r114.

MYERS, C. L., BARRETT, D. R., HIBBS, M. A., HUTTENHOWER, C. and TROYANSKAYA, O. G. (2006). Finding function: Evaluation methods for functional genomic data. *BMC Genomics* **7** 187.

NETWORK, C. G. A. et al. (2012). Comprehensive molecular portraits of human breast tumours. *Nature* **490** 61–70.

NEWMAN, M. E. J. (2006). Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* (3) **74** Article ID 036104. MR2282139

OUYANG, Z., ZHOU, Q. and WONG, W. H. (2009). ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc. Natl. Acad. Sci. USA* **106** 21521–21526.

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

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

RITCHIE, M. D., HOLZINGER, E. R., LI, R., PENDERGRASS, S. A. and KIM, D. (2015). Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet.* **16** 85–97.

SHEN, K. and TSENG, G. C. (2010). Meta-analysis for pathway enrichment analysis when combining multiple genomic studies. *Bioinformatics* **26** 1316–1323.

TOMCZAK, K., CZERWIŃSKA, P. and WIZNEROWICZ, M. (2015). The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Contemp. Oncol.* **19** A68–A77.

TROYANSKAYA, O. G., DOLINSKI, K., OWEN, A. B., ALTMAN, R. B. and BOTSTEIN, D. (2003). A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proc. Natl. Acad. Sci. USA* **100** 8348–8353.

VARAMBALLY, S., YU, J., LAXMAN, B., RHODES, D. R., MEHRA, R., TOMLINS, S. A., SHAH, R. B., CHANDRAN, U., MONZON, F. A., BECICH, M. J. et al. (2005). Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. *Cancer Cell* **8** 393–406.

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

YANG, E., RAVIKUMAR, P., ALLEN, G. I. and LIU, Z. (2013). On graphical models via univariate exponential family distributions. Preprint. Available at arXiv:1301.4183.

YIN, J. and LI, H. (2011). A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *Ann. Appl. Stat.* **5** 2630–2650. MR2907129

YOOK, S.-H., OLTVAI, Z. N. and BARABÁSI, A.-L. (2004). Functional and topological characterization of protein interaction networks. *Proteomics* **4** 928–942.

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67. MR2212574

ZHANG, Y., OUYANG, Z. and ZHAO, H. (2017). Supplement to "A statistical framework for data integration through graphical models with application to cancer genomics." DOI:10.1214/16-AOAS998SUPP.

The Annals of Applied Statistics
2017, Vol. 11, No. 1, 185–201
DOI: 10.1214/16-AOAS999

A PENALIZED COX PROPORTIONAL HAZARDS MODEL WITH MULTIPLE TIME-VARYING EXPOSURES[1]

BY CHENKUN WANG, HAI LIU AND SUJUAN GAO

Indiana University and Gilead Sciences, Inc

In recent pharmacoepidemiology research, the increasing use of electronic medication dispensing data provides an unprecedented opportunity to examine various health outcomes associated with long-term medication usage. Often, patients may take multiple types of medications intended for the same medical condition and the medication exposure status and intensity may vary over time, posing challenges to the statistical modeling of such data. In this article, we propose a penalized Cox proportional hazards (PH) model with multiple functional covariates and potential interaction effects. We also consider constrained coefficient functions to ensure a diminishing medication effect over time. Hypothesis testing of interaction effect and main effect was discussed under the penalized Cox PH model setting. Our simulation studies demonstrate the adequate performance of the proposed methods for both parameter estimation and hypothesis testing. Application to a primary care depression cohort study was also illustrated to examine the effects of two common types of antidepressants on the risk of coronary artery disease.

REFERENCES

ABRAHAMOWICZ, M., BARTLETT, G., TAMBLYN, R. and DU BERGER, R. (2006). Modeling cumulative dose and exposure duration provided insights regarding the associations between benzodiazepines and injuries. *J. Clin. Epidemiol.* **59** 393–403.

BERGSTROM, R. F., LEMBERGER, L., FARID, N. A. and WOLEN, R. L. (1988). Clinical pharmacology and pharmacokinetics of fluoxetine: A review. *Br. J. Psychiatry Suppl.* **3** 47–50.

BERHANE, K., HAUPTMANN, M. and LANGHOLZ, B. (2008). Using tensor product splines in modeling exposure-time-response relationships: Application to the Colorado Plateau Uranium Miners cohort. *Stat. Med.* **27** 5484–5496. MR2542365

BHADRA, D., DANIELS, M. J., KIM, S., GHOSH, M. and MUKHERJEE, B. (2012). A Bayesian semiparametric approach for incorporating longitudinal information on exposure history for inference in case–control studies. *Biometrics* **68** 361–370. MR2959602

BRESLOW, N. E., LUBIN, J. H., MAREK, P. and LANGHOLZ, B. (1983). Multiplicative models and cohort analysis. *J. Amer. Statist. Assoc.* **78** 1–12.

CALLAHAN, C. M., HUI, S. L., NIENABER, N. A. and MUSICK, B. S. (1994). Longitudinal study of depression and health services use among elderly primary care patients. *Journal of the American Geriatrics Society* **42** 833–838.

DAMUSH, T. M., JIA, H., RIED, L. D., QIN, H., CAMEON, R., PLUE, L. and WILLIAMS, L. S. (2008). Case-finding algorithm for post-stroke depression in the veterans health administration. *International Journal of Geriatric Psychiatry* **23** 517–522.

Key words and phrases. Pharmacoepidemiology, time-varying exposure, interaction, penalized spline.

DE LA TORRE, B. R., DREHER, J., MALEVANY, I., BAGLI, M., KOLBINGER, M., OMRAN, H., LÜDERITZ, B. and RAO, M. L. (2001). Serum levels and cardiovascular effects of tricyclic antidepressants and selective serotonin reuptake inhibitors in depressed patients. *Therapeutic Drug Monitoring* **23** 435–440.

FERRATY, F. and VIEU, P. (2009). Additive prediction and boosting for functional data. *Comput. Statist. Data Anal.* **53** 1400–1413. MR2657100

FUCHS, K., SCHEIPL, F. and GREVEN, S. (2015). Penalized scalar-on-functions regression with interaction. *Comput. Statist. Data Anal.* **81** 38–51. MR3257399

GASPARRINI, A. (2014). Modeling exposure-lag-response associations with distributed lag nonlinear models. *Stat. Med.* **33** 881–899. MR3249094

GLASSMAN, A. H. (1984). Cardiovascular effects of tricyclic antidepressants. *Annual Review of Medicine* **35** 503–511.

GOEMAN, J., MEIJER, R. and CHATURVEDI, N. (2016). L1 and L2 penalized regression models. Available at http://cran.r-project.org/web/packages/penalized/vignettes/penalized.pdf.

GOLDSMITH, J., CRAINICEANU, C. M., CAFFO, B. and REICH, D. (2012). Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **61** 453–469. MR2914521

GRAY, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *J. Amer. Statist. Assoc.* **87** 942–951.

GRAY, R. J. (1994). Spline-based tests in survival analysis. *Biometrics* **50** 640–652. MR1309310

GRAY, S. L., ANDERSON, M. L., DUBLIN, S., HANLON, J. T., HUBBARD, R., WALKER, R., YU, O., CRANE, P. K. and LARSON, E. B. (2015). Cumulative use of strong anticholinergics and incident dementia: A prospective cohort study. *JAMA Internal Medicine* **175** 401–407.

HASTIE, T. and TIBSHIRANI, R. (1986). Generalized additive models. *Statist. Sci.* **1** 297–318. MR0858512

HAUPTMANN, M., WELLMANN, J., LUBIN, J. H., ROSENBERG, P. S. and KREIENBROCK, L. (2000). Analysis of exposure-time–response relationships using a spline weight function. *Biometrics* **56** 1105–1108. MR1815589

HURVICH, C. M., SIMONOFF, J. S. and TSAI, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 271–293. MR1616041

JEFFERSON, J. W. (1975). A review of the cardiovascular effects and toxicity of tricyclic antidepressants. *Psychosomatic Medicine* **37** 160–179.

LANGHOLZ, B., THOMAS, D., XIANG, A. and STRAM, D. (1999). Latency analysis in epidemiologic studies of occupational exposures: Application to the Colorado Plateau uranium miners cohort. *American Journal of Industrial Medicine* **35** 246–256.

LINDSLEY, C. W. (2012). The top prescription drugs of 2011 in the United States: Antipsychotics and antidepressants once again lead CNS therapeutics. *ACS Chemical Neuroscience* **3** 630–631.

MCDONALD, C. J., OVERHAGE, J. M., TIERNEY, W. M., DEXTER, P. R., MARTIN, D. K., SUICO, J. G., ZAFAR, A., SCHADOW, G., BLEVINS, L., GLAZENER, T. et al. (1999). The Regenstrief medical record system: A quarter century experience. *International Journal of Medical Informatics* **54** 225–253.

MUGGEO, V. M. (2008). Modeling temperature effects on mortality: Multiple segmented relationships with common break points. *Biostatistics* **9** 613–620.

OBERMEIER, V., SCHEIPL, F., HEUMANN, C., WASSERMANN, J. and KÜCHENHOFF, H. (2015). Flexible distributed lags for modelling earthquake data. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **64** 395–412. MR3302306

PACHER, P. and KECSKEMETI, V. (2004). Cardiovascular side effects of new antidepressants and antipsychotics: New drugs, old concerns? *Curr. Pharm. Des.* **10** 2463–2475.

PERPEROGLOU, A. (2014). Cox models with dynamic ridge penalties on time-varying effects of the covariates. *Stat. Med.* **33** 170–180. MR3141561

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

RICHARDSON, D. B. (2009). Latency models for analyses of protracted exposures. *Epidemiology* (*Cambridge*, *Mass*.) **20** 395.

SCHEIPL, F. and GREVEN, S. (2015). Identifiability in penalized function-on-function regression models. Preprint. Available at arXiv:1506.03627.

SCHIPPER, M., TAYLOR, J. M. G. and LIN, X. (2008). Generalized monotonic functional mixed models with application to modelling normal tissue complications. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **57** 149–163. MR2420434

SYLVESTRE, M.-P. and ABRAHAMOWICZ, M. (2008). Comparison of algorithms to generate event times conditional on time-dependent covariates. *Stat. Med.* **27** 2618–2634. MR2440055

SYLVESTRE, M.-P. and ABRAHAMOWICZ, M. (2009). Flexible modeling of the cumulative effects of time-dependent exposures on the hazard. *Stat. Med.* **28** 3437–3453. MR2744373

THERNEAU, T. M., GRAMBSCH, P. M. and PANKRATZ, V. S. (2003). Penalized survival models and frailty. *J. Comput. Graph. Statist.* **12** 156–175. MR1965213

THOMAS, D. C. (1988). Models for exposure-time-response relationships with applications to cancer epidemiology. *Annual Review of Public Health* **9** 451–482.

VACEK, P. M. (1997). Assessing the effect of intensity when exposure varies over time. *Stat. Med.* **16** 505–513.

WOOD, S. N. (2006). *Generalized Additive Models*: *An Introduction with R*. Chapman & Hall/CRC, Boca Raton, FL. MR2206355

WOOD, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 3–36. MR2797734

ZHANG, D., LIN, X. and SOWERS, M. (2007). Two-stage functional mixed models for evaluating the effect of longitudinal covariate profiles on a scalar outcome. *Biometrics* **63** 351–362. MR2370793

The Annals of Applied Statistics
2017, Vol. 11, No. 1, 202–224
DOI: 10.1214/16-AOAS1000

FORECASTING SEASONAL INFLUENZA WITH A STATE-SPACE SIR MODEL[1]

BY DAVE OSTHUS[*,†], KYLE S. HICKMANN[*], PETRUŢA C. CARAGEA[†], DAVE HIGDON[‡,*] AND SARA Y. DEL VALLE[*]

Los Alamos National Laboratory[*], Iowa State University[†], and Virginia Tech[‡]

Seasonal influenza is a serious public health and societal problem due to its consequences resulting from absenteeism, hospitalizations, and deaths. The overall burden of influenza is captured by the Centers for Disease Control and Prevention's influenza-like illness network, which provides invaluable information about the current incidence. This information is used to provide decision support regarding prevention and response efforts. Despite the relatively rich surveillance data and the recurrent nature of seasonal influenza, forecasting the timing and intensity of seasonal influenza in the U.S. remains challenging because the form of the disease transmission process is uncertain, the disease dynamics are only partially observed, and the public health observations are noisy. Fitting a probabilistic state-space model motivated by a deterministic mathematical model [a susceptible-infectious-recovered (SIR) model] is a promising approach for forecasting seasonal influenza while simultaneously accounting for multiple sources of uncertainty. A significant finding of this work is the importance of thoughtfully specifying the prior, as results critically depend on its specification. Our conditionally specified prior allows us to exploit known relationships between latent SIR initial conditions and parameters and functions of surveillance data. We demonstrate advantages of our approach relative to alternatives via a forecasting comparison using several forecast accuracy metrics.

REFERENCES

ANDERSON, J. L. (2001). An ensemble adjustment Kalman filter for data assimilation. *Mon. Weather Rev.* **129** 2884–2903.

BRAUER, F., VAN DEN DRIESSCHE, P. and WU, J., eds. (2008). *Mathematical Epidemiology*. *Lecture Notes in Math*. **1945**. Springer, Berlin. MR2452129

CAPALDI, A., BEHREND, S., BERMAN, B., SMITH, J., WRIGHT, J. and LLOYD, A. L. (2012). Parameter estimation and uncertainty quantification for an epidemic model. *Math. Biosci. Eng.* **9** 553–576. MR2957535

CDC.GOV (2017). Influenza (Flu) Past Pandemics. Available at http://www.cdc.gov/flu/pandemic-resources/basics/past-pandemics.html. Accessed: 02-06-2017.

CENTERS FOR DISEASE CONTROL AND PREVENTION (2014a). Free resources. Available at http://www.cdc.gov/flu/freeresources/. Accessed: 05-5-2015.

CENTERS FOR DISEASE CONTROL AND PREVENTION (2014b). Estimating Seasonal Influenza-Associated Deaths in the United States. Available at http://www.cdc.gov/flu/about/disease/us_flu-related_deaths.htm. Accessed: 02-06-2017.

Key words and phrases. Bayesian modeling, state-space modeling, SIR model, forecasting, influenza, time-series.

CENTERS FOR DISEASE CONTROL AND PREVENTION (2015). In Overview of influenza surveillance in the United States. Available at http://www.cdc.gov/flu/weekly/overview.htm. Accessed: 04-30-2015.

CHRETIEN, J.-P., GEORGE, D., SHAMAN, J., CHITALE, R. A. and MCKENZIE, F. E. (2014). Influenza forecasting in human populations: A scoping review. *PLoS ONE* **9** e94130.

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

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

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

GENEROUS, N., FAIRCHILD, G., DESHPANDE, A., VALLE, S. Y. D. and PRIEDHORSKY, R. (2014). Global disease monitoring and forecasting with Wikipedia. *PLoS Comput. Biol.* **10** e1003892.

GERMANN, T. C., KADAU, K., LONGINI, I. M. and MACKEN, C. A. (2006). Mitigation strategies for pandemic influenza in the United States. *Proc. Natl. Acad. Sci. USA* **103** 5935–5940.

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

HARRIS, K. M., MAURER, J. and KELLERMANN, A. L. (2010). Influenza vaccine. *N. Engl. J. Med.* **363** 2183–2185.

HEFFERNAN, J., SMITH, R. and WAHL, L. (2005). Perspectives on the basic reproductive ratio. *J. Roy. Soc. Interface* **2** 281–293.

HICKMANN, K. S., FAIRCHILD, G., PRIEDHORSKY, R., GENEROUS, N., HYMAN, J. M., DESHPANDE, A. and VALLE, S. Y. D. (2015). Forecasting the 2013–2014 influenza season using Wikipedia. *PLoS Comput. Biol.* **11** e1004239.

KERMACK, W. O. and MCKENDRICK, A. G. (1927). A contribution to the mathematical theory of epidemics. In *In Proceedings of the Royal Society of London A*: *Mathematical*, *Physical and Engineering Sciences* **115** 700–721. The Royal Society.

MILLS, C. E., ROBINS, J. M. and LIPSITCH, M. (2004). Transmissibility of 1918 pandemic influenza. *Nature* **432** 904–906.

NSOESIE, E., MARARTHE, M. and BROWNSTEIN, J. (2013). Forecasting peaks of seasonal influenza epidemics. *PLoS Curr.* **5**.

NSOESIE, E. O., BROWNSTEIN, J. S., RAMAKRISHNAN, N. and MARATHE, M. V. (2014). A systematic review of studies on forecasting the dynamics of influenza outbreaks. *Influenza and other Respiratory Viruses* **8** 309–316.

OSTHUS, D., HICKMANN, K. S., CARAGEA, P. C., HIGDON, D. and DEL VALLE, S. Y. (2017). Supplement to "Forecasting seasonal influenza with a state-space SIR model." DOI:10.1214/16-AOAS1000SUPP.

PLUMMER, M. (2014). rjags: Bayesian graphical models using MCMC. R package version 3-14.

PLUMMER, M. et al. (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* **124** 125. Vienna.

ROSS, R. (1911). Some quantitative studies in epidemiology. *Nature* **87** 466–467.

SHAMAN, J. and KARSPECK, A. (2012). Forecasting seasonal outbreaks of influenza. *Proc. Natl. Acad. Sci. USA* **109** 20425–20430.

SHAMAN, J., KARSPECK, A., YANG, W., TAMERIUS, J. and LIPSITCH, M. (2013). Real-time influenza forecasts during the 2012–2013 season. *Nat. Commun.* **4** 2837.

R CORE TEAM (2015). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

TOWERS, S., CHOWELL, G., HAMEED, R., JASTREBSKI, M., KHAN, M., MEEKS, J., MUBAYI, A. and HARRIS, G. (2013). Climate change and influenza: The likelihood of early and severe influenza seasons following warmer than average winters. *PLoS Curr.* **5**.

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES (2017). Regional Offices. Available at http://www.hhs.gov/about/agencies/regional-offices/. Accessed: 02-06-2017.

WEISS, H. H. (2013). The SIR model and the foundations of public health. In *Materials Matemàtics* 0001–17.

YANG, W., LIPSITCH, M. and SHAMAN, J. (2015). Inference of seasonal and pandemic influenza transmission dynamics. *Proc. Natl. Acad. Sci. USA* **112** 2723–2728.

The Annals of Applied Statistics
2017, Vol. 11, No. 1, 225–247
DOI: 10.1214/16-AOAS1001

SENSITIVITY ANALYSIS FOR AN UNOBSERVED MODERATOR IN RCT-TO-TARGET-POPULATION GENERALIZATION OF TREATMENT EFFECTS

BY TRANG QUYNH NGUYEN[*,1], CYRUS EBNESAJJAD[*,2], STEPHEN R. COLE[†,3] AND ELIZABETH A. STUART[*,2]

Johns Hopkins Bloomberg School of Public Health and University of North Carolina†*

In the presence of treatment effect heterogeneity, the average treatment effect (ATE) in a randomized controlled trial (RCT) may differ from the average effect of the same treatment if applied to a target population of interest. If all treatment effect moderators are observed in the RCT and in a dataset representing the target population, then we can obtain an estimate for the target population ATE by adjusting for the difference in the distribution of the moderators between the two samples. This paper considers sensitivity analyses for two situations: (1) where we cannot adjust for a specific moderator V observed in the RCT because we do not observe it in the target population; and (2) where we are concerned that the treatment effect may be moderated by factors not observed even in the RCT, which we represent as a composite moderator U. In both situations, the outcome is not observed in the target population. For situation (1), we offer three sensitivity analysis methods based on (i) an outcome model, (ii) full weighting adjustment and (iii) partial weighting combined with an outcome model. For situation (2), we offer two sensitivity analyses based on (iv) a bias formula and (v) partial weighting combined with a bias formula. We apply methods (i) and (iii) to an example where the interest is to generalize from a smoking cessation RCT conducted with participants of alcohol/illicit drug use treatment programs to the target population of people who seek treatment for alcohol/illicit drug use in the US who are also cigarette smokers. In this case a treatment effect moderator is observed in the RCT but not in the target population dataset.

REFERENCES

ARAH, O. A., CHIBA, Y. and GREENLAND, S. (2008). Bias formulas for external adjustment and sensitivity analysis of unmeasured confounders. *Ann. Epidemiol.* **18** 637–646.

COLE, S. R. and STUART, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *Am. J. Epidemiol.* **172** 107–115.

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

DING, P. and VANDERWEELE, T. J. (2014). Generalized Cornfield conditions for the risk difference. *Biometrika* **101** 971–977. MR3286930

Key words and phrases. Sensitivity analysis, generalization, treatment effect heterogeneity, unobserved moderator, unobserved effect modifier.

DING, P. and VANDERWEELE, T. J. (2015). The differential geometry of homogeneity spaces ccross effect scales. Available at arXiv:1510.08534v2.

DING, P. and VANDERWEELE, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology* **27** 368–377.

GASTWIRTH, J. L., KRIEGER, A. M. and ROSENBAUM, P. R. (1998). Dual and simultaneous sensitivity analysis for matched pairs. *Biometrika* **85** 907–920.

GREENHOUSE, J. B., KAIZAR, E. E., KELLEHER, K., SELTMAN, H. and GARDNER, W. (2008). Generalizing from clinical trial data: A case study. The risk of suicidality among pediatric antidepressant users. *Stat. Med.* **27** 1801–1813. MR2420346

GREENLAND, S. (1996). Basic methods for sensitivity analysis of biases. *Int. J. Epidemiol.* **25** 1107–1116.

HONG, G. (2010). Ratio of mediator probability weighting for estimating natural direct and indirect effects. *Proc. Am. Stat. Assoc. Biom. Sect.* 2401–2415.

KERN, H. L., STUART, E. A., HILL, J. L. and GREEN, D. P. (2016). Assessing methods for generalizing experimental impact estimates to target samples. *J. Res. Educ. Eff.* **9** 103–127.

NGUYEN, T. Q., EBNESAJJAD, C., COLE, S. R. and STUART, E. A. (2017). Supplement to "Sensitivity analysis for an unobserved moderator in RCT-to-target-population generalization of treatment effects." DOI:10.1214/16-AOAS1001SUPPA, DOI:10.1214/16-AOAS1001SUPPB.

OLSEN, R. B., ORR, L. L., BELL, S. H. and STUART, E. A. (2013). External validity in policy evaluations that choose sites purposively. *J. Policy Anal. Manage.* **32** 107–121.

REID, M. S., FALLON, B., SONNE, S., FLAMMINO, F., NUNES, E. V., JIANG, H., KUONIOTIS, E., LIMA, J., BRADY, R., BURGESS, C., ARFKEN, C., PIHLGREN, E., GIORDANO, L., STAROSTA, A., ROBISON, J. and ROTROSEN, J. (2008). Smoking cessation treatment in community-based substance abuse rehabilitation programs. *J. Subst. Abuse Treat.* **35** 68–77.

ROSENBAUM, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika* **74** 13–26. MR0885915

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

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

SCHNEEWEISS, S. (2006). Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiol. Drug Saf.* **15** 291–303.

STUART, E. A., BRADSHAW, C. P. and LEAF, P. J. (2015). Assessing the generalizability of randomized trial results to target populations. *Prev. Sci.* **16** 475–485.

STUART, E. A. and RHODES, A. (2016). Generalizing treatment effect estimates from sample to population: A case study in the difficulties of finding sufficient data. *Eval. Rev.*

STUART, E. A., COLE, S. R., BRADSHAW, C. P. and LEAF, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *J. Roy. Statist. Soc. Ser. A* **174** 369–386.

SUSUKIDA, R., CRUM, R. M., STUART, E. A., EBNESAJJAD, C. and MOJTABAI, R. (2016). Assessing sample representativeness in randomized controlled trials: Application to the National Institute of Drug Abuse Clinical Trials Network. *Addiction* **111** 1226–1234.

TIPTON, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *J. Educ. Behav. Stat.* **38** 239–266.

VANDERWEELE, T. J. and ARAH, O. A. (2011). Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology* **22** 42–52.

WEISBERG, H. I., HAYDEN, V. C. and PONTES, V. P. (2009). Selection criteria and generalizability within the counterfactual framework: Explaining the paradox of antidepressant-induced suicidality? *Clin. Trials* **6** 109–118.

The Annals of Applied Statistics
2017, Vol. 11, No. 1, 248–273
DOI: 10.1214/16-AOAS1002

ELECTRICITY PRICE DEPENDENCE IN NEW YORK STATE ZONES: A ROBUST DETRENDED CORRELATION APPROACH[1]

BY DEBBIE J. DUPUIS

HEC Montréal

The cost of electricity varies across the zones of the New York State electric system. While fair and open access to the electrical grid is sought, we show that residents currently do not equally benefit, or suffer, from price changes. Upcoming major investments in the grid offer an opportunity to rectify these inequalities, but only if we understand the price-change propagation dynamics for the current underlying infrastructure. We study these dynamics, estimating the partial correlations between changes in electricity prices in connected zones. We develop and investigate a robust exponentially weighted correlation estimator that performs well in the presence of electricity price spikes and can track a rapidly changing time-varying correlation. We show that price-change partial correlations are mostly positive, but can also be negative, and provide new insight into price-change dynamics within the grid that cannot be extracted from the price-setting algorithm or obtained from available transmission capability data.

REFERENCES

ALEXANDER, C. (2001). *Market Models*. Wiley, Chichester.

BEN AMOR, M., BILLETTE DE VILLEMEUR, E., PELLAT, M. and PINEAU, P.-O. (2014). Influence of wind power on hourly electricity prices and GHG (greenhouse gas) emissions: Evidence that congestion matters from Ontario zonal data. *Energy* **66** 458–469.

CHRISTENSEN, T. M., HURN, A. S. and LINDSAY, K. A. (2012). Forecasting spikes in electricity prices. *International Journal of Forecasting* **28** 400–411.

CORMEN, T. H., LEISERSON, C. E. and RIVEST, R. L. (1990). *Introduction to Algorithms*. MIT Press, Cambridge, MA. MR1066870

EIA (2015). *Henry Hub Natural Gas Spot Price*. US Energy Information Administration, Washington. Web site accessed on January 15, 2015. http://www.eia.gov/dnav/ng/hist/rngwhhdd.htm.

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

EYDELAND, A. and WOLYNIEC, K. (2012). *Energy and Power Risk Management*, 2nd ed. Wiley, Hoboken, NJ.

FISHER, T. J. and GALLAGHER, C. M. (2012). WeightedPortTest: Weighted portmanteau tests for time series goodness-of-fit. R package version 1.0. Available at https://CRAN.R-project.org/package=WeightedPortTest.

FRANCQ, C. and SUCARRAT, G. (2017). An equation-by-equation estimator of a multivariate log-GARCH-X model of financial returns. *J. Multivariate Anal.* **153** 16–32. MR3578836

Key words and phrases. Exponentially weighted robust correlation estimator, least median of squares, log-GARCH-X, market efficiency.

GEMAN, H. and RONCORNI, A. (2006). Understanding the fine structure of electricity prices. *Journal of Business* **79** 1225–1261.

HELLSTRÖM, J., LUNDGREN, J. and YU, H. (2012). Why do electricity prices jump? Empirical evidence from the nordic electricity market. *Energy Economics* **34** 1774–1781.

HICKEY, E., LOOMIS, D. G. and MOHAMMADI, H. (2012). Forecasting hourly electricity prices using ARMAX-GARCH models: An application to MISO hubs. *Energy Economics* **34** 307–315.

JANCZURA, J., TRÜCK, S., WERON, R. and WOLFF, R. C. (2013). Identifying spikes and seasonal components in electricity spot price data: A guide to robust modeling. *Energy Economics* **38** 96–110.

KARAKATSANI, N. V. and BUNN, D. W. (2008). Forecasting electricity prices: The impact of fundamentals and time-varying coefficients. *International Journal of Forecasting* **24** 764–785.

KENDALL, M. G. and STUART, A. (1979). *The Advanced Theory of Statistics, Vol. 2. Inference and Relationship*. London, Griffin.

LANGFELDER, P. and HORVATH, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*.

LANGFELDER, P. and HORVATH, S. (2012). Fast R functions for robust correlations and hierarchical clustering. *J. Stat. Softw.* **46** 1–17.

LAX, D. A. (1985). Robust estimators of scale: Finite-sample performance in long-tailed symmetric distributions. *J. Amer. Statist. Assoc.* **80** 736–741.

LI, W. K. and MAK, T. K. (1994). On the squared residual autocorrelations in nonlinear time series with conditional heteroskedasticity. *J. Time Series Anal.* **15** 627–636. MR1312326

NEW YORK INDEPENDENT SYSTEM OPERATOR (2013). Day-Ahead Scheduling Manual. Schenectady, New York.

NEW YORK INDEPENDENT SYSTEM OPERATOR (2014). Power Trends 2014: Evolution of the Grid. Schenectady, New York.

NEWEY, W. K. and WEST, K. D. (1987). A simple, positive semidefinite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55** 703–708. MR0890864

NOWOTARSKI, J., TOMCZYK, J. and WERON, R. (2013). Robust estimation and forecasting of the long-term seasonal component of electricity spot prices. *Energy Economics* **39** 13–27.

PĚNA, J. I. (2012). A note on panel hourly electricity prices. *Journal of Energy Markets* **5** 81–97.

PINEAU, P.-O., DUPUIS, D. J. and CENESIZOGLU, T. (2015). Assessing the value of power interconnections under climate and natural gas price risks. *Energy* **82** 128–137.

POZZI, F., DI MATTEO, T. and ASTE, T. (2012). Exponential smoothing weighted correlations. *Eur. Phys. J. B* **85** 175.

SHOEMAKER, L. H. and HETTMANSPERGER, T. P. (1982). Robust estimates and tests for the one- and two-sample scale models. *Biometrika* **69** 47–53. MR0655669

SUCARRAT, G. (2014). lgarch: Simulation and estimation of log-GARCH models. R package version 0.5. Available at http://CRAN.R-project.org/package=lgarch.

SUCARRAT, G. and ESCRIBANO, Á. (2014). Unbiased Estimation of Log-GARCH Models in the Presence of Zero Return. MPRA Paper 59040.

SUCARRAT, G., GRØNNEBERG, S. and ESCRIBANO, A. (2016). Estimation and inference in univariate and multivariate log-GARCH-X models when the conditional density is unknown. *Comput. Statist. Data Anal.* **100** 582–594.

WILCOX, R. R. (2012). *Introduction to Robust Estimation and Hypothesis Testing*, 3rd ed. Academic Press, San Diego, CA.

The Annals of Applied Statistics
2017, Vol. 11, No. 1, 274–296
DOI: 10.1214/16-AOAS1003

PARTIALLY TIME-VARYING COEFFICIENT PROPORTIONAL HAZARDS MODELS WITH ERROR-PRONE TIME-DEPENDENT COVARIATES—AN APPLICATION TO THE AIDS CLINICAL TRIAL GROUP 175 DATA

BY XIAO SONG[1] AND LI WANG[2]

University of Georgia and Iowa State University

Due to cost and time considerations, interest has focused on identifying surrogate markers that could be substituted for the clinical endpoint, time to an event of interest, in evaluation of treatment efficacy. Joint models are often used to assess the effect of surrogate markers and treatment. Motivated by recent works studying the AIDS Clinical Trial Group (ACTG) 175 data, we propose a partially time-varying coefficient proportional hazards model for modeling the relationship between the hazard of failure and time-dependent and time-independent covariates. The time-varying coefficients are approximated by polynomial splines, and the corrected score and conditional score approaches are adopted to estimate the regression coefficients. The proposed estimators are consistent, and the asymptotic normality is established for the constant coefficients, which enables us to construct confidence intervals and permits joint inference. The finite-sample performance of the proposed method is assessed by Monte Carlo simulation studies. The proposed model is applied to ACTG 175 data to assess the temporal dynamics of the effect of treatment and CD4 count on time to AIDS or death.

REFERENCES

BYCOTT, P. and TAYLOR, J. (1998). A comparison of smoothing techniques for CD4 data measured with error in a time-dependent Cox proportional hazards model. *Stat. Med.* **17** 2061–2077.

CAI, J., FAN, J., JIANG, J. and ZHOU, H. (2008). Partially linear hazard regression with varying coefficients for multivariate survival data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 141–158. MR2412635

DAFNI, U. G. and TSIATIS, A. A. (1998). Evaluating surrogate markers of clinical outcome measured with error. *Biometrics* **54** 1445–1462.

DEGRUTTOLA, V. and TU, X. M. (1994). Modeling progression of CD-4 lymphocyte count and its relationship to survival time. *Biometrics* **50** 1003–1014.

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

FAUCETT, C. L. and THOMAS, D. C. (1996). Simultaneously modeling censored survival data and repeatedly measured covariates: A Gibbs sampling approach. *Stat. Med.* **15** 1663–1685.

GU, L., WANG, L., HÄRDLE, W. K. and YANG, L. (2014). A simultaneous confidence corridor for varying coefficient regression with sparse functional data. *TEST* **23** 806–843. MR3274476

Key words and phrases. Corrected score, conditional score, joint modeling, measurement error, polynomial spline, survival.

HAMMER, S. M., KATEZSTEIN, D. A., HUGHES, M. D., GUNDAKER, H., SCHOOLEY, R. T., HAUBRICH, R. H., HENRY, W. K., LEDERMAN, M. M., PHAIR, J. P., NIU, M., HIRSCH, M. S. and MERIGAN, T. C. (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *N. Engl. J. Med.* **335** 1081–1090.

HENDERSON, R., DIGGLE, P. and DOBSON, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostat.* **4** 465–480.

HSIEH, F., TSENG, Y.-K. and WANG, J.-L. (2006). Joint modeling of survival and longitudinal data: Likelihood approach revisited. *Biometrics* **62** 1037–1043. MR2297674

HUANG, J. (1999). Efficient estimation of the partly linear additive Cox model. *Ann. Statist.* **27** 1536–1563. MR1742499

HUANG, J. Z. and LIU, L. (2006). Polynomial spline estimation and inference of proportional hazards regression models with flexible relative risk form. *Biometrics* **62** 793–802. MR2247208

LIU, R. and YANG, L. (2010). Spline-backfitted kernel smoothing of additive coefficient model. *Econometric Theory* **26** 29–59.

NAN, B., LIN, X., LISABETH, L. D. and HARLOW, S. D. (2005). A varying-coefficient Cox model for the effect of age at a marker event on age at menopause. *Biometrics* **61** 576–583. MR2140931

PAWITAN, Y. and SELF, S. (1993). Modeling disease marker processes in AIDS. *J. Amer. Statist. Assoc.* **88** 719–726.

PRENTICE, R. L. (1982). Covariate measurement errors and parameter estimates in a failure time regression model. *Biometrika* **69** 331–342.

PRENTICE, R. L. (1989). Surrogate endpoints in clinical trials: Definition and operation criteria. *Stat. Med.* **8** 431–440.

SONG, X., DAVIDIAN, M. and TSIATIS, A. A. (2002a). An estimator for the proportional hazards model with multiple longitudinal covariates measured with error. *Biostat.* **3** 511–528.

SONG, X., DAVIDIAN, M. and TSIATIS, A. A. (2002b). A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics* **58** 742–753. MR1945011

SONG, X. and HUANG, Y. (2005). On corrected score approach for proportional hazards model with covariate measurement error. *Biometrics* **61** 702–714. MR2196158

SONG, X. and WANG, C. Y. (2008). Semiparametric approaches for joint modeling of longitudinal and survival data with time-varying coefficients. *Biometrics* **64** 557–566, 669. MR2432426

SONG, X. and WANG, L. (2017). Supplement to "Partially time-varying coefficient proportional hazards models with error-prone time-dependent covariates—an application to the AIDS Clinical Trial Group 175 data." DOI:10.1214/16-AOAS1003SUPP.

TSIATIS, A. A. and DAVIDIAN, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika* **88** 447–458. MR1844844

TSIATIS, A. A., DEGRUTTOLA, V. and WULFSOHN, M. S. (1995). Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *J. Amer. Statist. Assoc.* **90** 27–37.

WANG, C. Y. (2006). Corrected score estimator for joint modeling of longitudinal and failure time data. *Statist. Sinica* **16** 235–253. MR2256090

WANG, L. and YANG, L. (2007). Spline-backfitted kernel smoothing of nonlinear additive autoregression model. *Ann. Statist.* **35** 2474–2503. MR2382655

WULFSOHN, M. S. and TSIATIS, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53** 330–339. MR1450186

XU, J. and ZEGER, S. L. (2001). Joint analysis of longitudinal data comprising repeated measures and times to events. *J. Roy. Statist. Soc. Ser. C* **50** 375–387. MR1856332

XUE, L. and LIANG, H. (2010). Polynomial spline estimation for a generalized additive coefficient model. *Scand. J. Stat.* **37** 26–46. MR2675938

XUE, L. and YANG, L. (2006). Additive coefficient modeling via polynomial spline. *Statist. Sinica* **16** 1423–1446. MR2327498

The Annals of Applied Statistics
2017, Vol. 11, No. 1, 297–319
DOI: 10.1214/16-AOAS1004

FUNCTIONAL TIME SERIES MODELS FOR ULTRAFINE PARTICLE DISTRIBUTIONS[1]

BY HEIDI J. FISCHER*, QUNFANG ZHANG†, YIFANG ZHU†, AND ROBERT E. WEISS†

Kaiser Permanente Southern California and UCLA Fielding School of Public Health†*

We propose Bayesian functional mixed effect time series models to explain the impact of engine idling on ultrafine particle (UFP) counts inside school buses. UFPs are toxic to humans and school engines emit particles primarily in the UFP size range. As school buses idle at bus stops, UFPs penetrate into cabins through cracks, doors, and windows. Counts increase over time at a size dependent rate once the engine turns on. How UFP counts inside buses vary by particle size over time and under different idling conditions is not yet well understood. We model UFP counts at a given time using a mixed effect model with a cubic B-spline basis as a function of size. The log residual variance over size is modeled using a quadratic B-spline basis to account for heterogeneity in error across size bin, and errors are autoregressive over time. Model predictions are communicated graphically. These methods provide information needed to quantify UFP counts by size and possibly minimize UFP exposure in the future.

REFERENCES

ALESSANDRINI, F., SCHULZ, H., TAKENAKA, S., LENTNER, B., KARG, E., BEHRENDT, H. and JAKOB, T. (2006). Effects of ultrafine carbon particle inhalation on allergic inflammation of the lung. *Journal of Allergy and Clinical Immunology* **117** 824–830.

BALADANDAYUTHAPANI, V., MALLICK, B. K. and CARROLL, R. J. (2005). Spatially adaptive Bayesian penalized regression splines (P-splines). *J. Comput. Graph. Statist.* **14** 378–394. MR2160820

BENNETT, W. D. and ZEMAN, K. L. (1998). Deposition of fine particles in children spontaneously breathing at rest. *Inhalation Toxicology* **10** 831–842.

BERHANE, K. and MOLITOR, N. T. (2008). A Bayesian approach to functional-based multilevel modeling of longitudinal data: Applications to environmental epidemiology. *Biostatistics* **4** 686–699.

CHALONER, K. (1994). Residual analysis and outliers in Bayesian hierarchical models. In *Aspects of Uncertainty*. 149–157. Wiley, Chichester. MR1309691

CHALONER, K. and BRANT, R. (1988). A Bayesian approach to outlier detection and residual analysis. *Biometrika* **75** 651–659.

CRAINICEANU, C. M., RUPPERT, D., CARROLL, R. J., JOSHI, A. and GOODNER, B. (2007). Spatially adaptive Bayesian penalized splines with heteroscedastic errors. *J. Comput. Graph. Statist.* **16** 265–288. MR2370943

Key words and phrases. Bayesian statistics, hierarchical models, varying coefficient models, heteroskedasticity.

DE BOOR, C. (1978). *A Practical Guide to Splines*. *Applied Mathematical Sciences* **27**. Springer, New York. MR0507062

DELFINO, R. J., SIOUTAS, C. and MALIK, S. (2005). Potential role of ultrafine particles in associations between airborne particle mass and cardiovascular health. *Environmental Health Perspectives* **113** 934–946.

EPA (2002). Health assessment document for diesel engine exhaust. The National Technical Information Service, Springfield, VA.

EPA (2014). Clean school bus USA. Available at http://www.epa.gov/cleanschoolbus/csb-overview.htm.

FERIN, J., OBERDORSTER, G., PENNEY, D. P., SODERHOLM, S. C., GELEIN, R. and PIPER, H. C. (1990). Increased pulmonary toxicity of ultrafine particles? *Journal of Aerosol Science* **21** 384–387.

FISCHER, H. J, ZHANG, Q., ZHU, Y. and WEISS, R. E (2017). Supplement to "Functional time series models for ultrafine particle distributions." DOI:10.1214/16-AOAS1004SUPPA, DOI:10.1214/16-AOAS1004SUPPB.

FRAMPTON, M. W., STEWART, J. C., OBERDORSTER, G., MORROW, P. E., CHALUPA, D., PIETROPAOLI, A. P., FRASIER, L. M., SPEERS, D. M., COX, C., HUANG, L. S. and UTELL, M. J. (2006). Inhalation of ultrafine particles alters blood leukocyte expression of adhesion molecules in humans. *Environmental Health Perspectives* **114** 51–58.

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

GOLDSMITH, J. and KITAGO, K. (2013). Assessing systematic effects of stroke on motor control using hierarchical scalar-on-function regression. Techical report, Columbia Univ., New York, NY.

HADFIELD, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *J. Stat. Softw.* **33** 1–22.

HASTIE, T. and TIBSHIRANI, R. (1993). Varying-coefficient models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **55** 757–796. MR1229881

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

HUSSEIN, T., MASO, M. D., PETÄJÄ, T., KOPONEN, I. K., PAATERO, P., AALTO, P., HÄMERI, K. and KULMALA, M. (2005). Evaluation of an automatic algorithm for fitting the particle number size distributions. *Boreal Environment Research* **10** 337–355.

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

MORAWSKA, L., RISTOVSKI, Z., JAYARATNE, E. R., KEOGH, D. U. and LING, X. (2008). Ambient nano and ultrafine particles from motor vehicle emissions: Characteristics, ambient processing and implications on human exposure. *Atmospheric Environment* **42** 8113–8138.

MORRIS, J. (2015). Functional regression. *Annual Review of Statistics and Its Application* **2** 321–359.

MORRIS, J. S., VANNUCCI, M., BROWN, P. J. and CARROLL, R. J. (2003). Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis. *J. Amer. Statist. Assoc.* **98** 573–597. MR2011673

OBERDORSTER, G., SHARP, Z., ATUDOREI, V., ELDER, A., GELEIN, R., KREYLING, W. and COX, C. (2004). Translocation of inhaled ultrafine particles to the brain. *Inhalation Toxicology* **16** 437–445.

PALMER, J. and PETTIT, L. (1996). Risks of using improper priors with Gibbs sampling and autocorrelated errors. *J. Comput. Graph. Statist.* **5** 245–249.

PLUMMER, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, Vienna.

PRADO, R. and WEST, M. (2010). *Time Series*: *Modeling*, *Computation*, *and Inference*. CRC Press, Boca Raton, FL.

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

SAMET, J. M., RAPPOLD, A., GRAFF, D., CASCIO, W. E., BERNTSEN, J. H., HUANG, Y. C. T., HERBST, M., BASSETT, M., MONTILLA, T., HAZUCHA, M. J., BROMBERG, P. A. and DEVLIN, R. B. (2009). Concentrated ambient ultrafine particle exposure induces cardiac changes in young healthy volunteers. *American Journal of Respiratory and Critical Care Medicine* **179** 1034–1042.

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

WEISS, R. E. and LAZARO, C. G. (1992). Residual plots for repeated measures. *Stat. Med.* **11** 115–124.

WHITBY, P. H. (1978). Modal aerosol dynamics modelling. Technical report, U.S. Environment Protection Agency, Atmospheric Research and Exposure Assessment Laboratory, Washington, DC.

WHITBY, P. H., MCMURRY, P. H., SHANKER, U. and BINKOWSKI, F. S. (1991). Modal aerosol dynamics modeling. Technical report, U.S. Environment Protection Agency, Atmospheric Research and Exposure Assessment Laboratory, Washington, DC.

WRAITH, D., ALSTON, C., MENGERSEN, K. and HUSSEIN, T. (2009). Bayesian mixture model estimation of aerosol particle size distributions. *Environmetrics* **22** 23–34.

WRAITH, D., MENGERSEN, K., ALSTON, C., ROUSSEU, J. and HUSSEIN, T. (2014). Using informative priors in the estimation of mixtures over time with application to aerosol particle size distributions. *Ann. Appl. Stat.* **8** 232–258. MR3191989

ZHANG, Q., FISCHER, H. J., WEISS, R. E. and ZHU, Y. (2012). Ultrafine particle concentrations in and around idling school buses. *Atmospheric Environment* **69** 65–75.

The Annals of Applied Statistics
2017, Vol. 11, No. 1, 320–339
DOI: 10.1214/16-AOAS1006

EFFICIENT ESTIMATION OF AGE-SPECIFIC SOCIAL CONTACT RATES BETWEEN MEN AND WOMEN

BY JAN VAN DE KASSTEELE, JAN VAN EIJKEREN AND JACCO WALLINGA

National Institute for Public Health and the Environment (RIVM)

Social contact patterns reveal with whom individuals tend to socialize, and therefore to whom they transmit respiratory infections. We infer highly detailed age-specific contact rates between the sexes using a hierarchical Bayesian model that smooths while simultaneously guaranteeing the inherent reciprocity of contact rates. Application of this approach to social contact data from a large prospective survey confirms a tendency that people, especially children and adolescents, mostly contact other people of their own age and sex, and reveals that women have more contact with children than men. These findings imply different exposure patterns between the two sexes for specific age groups, which agrees with available observations.

REFERENCES

BORGDORFF, M. W., NAGELKERKE, N. J. D., DYE, C. and NUNN, P. (2000). Gender and tuberculosis: A comparison of prevalence surveys with notification data to explore sex differences in case detection. *Int. J. Tuberc. Lung Dis.* **4** 123–132.

BROWN, A. C. and MOSS, W. J. (2010). Sex, pregnancy and measles. In *Sex Hormones and Immunity to Infection* (S. L. Klein and C. Roberts, eds.) 281–302. Springer, Berlin, Heidelberg.

CAUCHEMEZ, S., BHATTARAI, A., MARCHBANKS, T. L., FAGAN, R. P., OSTROFF, S., FERGUSON, N. M., SWERDLOW, D., SODHA, S. V., MOLL, M. E., ANGULO, F. J., PALEKAR, R., ARCHER, W. R. and FINELLI, L. (2011). Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza. *Proc. Natl. Acad. Sci. USA* **108** 2825–2830.

CONLAN, A. J. K., EAMES, K. T. D., GAGE, J. A., VON KIRCHBACH, J. C., ROSS, J. V., SAENZ, R. A. and GOG, J. R. (2011). Measuring social networks in British primary schools through scientific engagement. *Proc. R. Soc. Lond., B Biol. Sci.* **278** 1467–1475.

CURRIE, I. D., DURBAN, M. and EILERS, P. H. (2004). Smoothing and forecasting mortality rates. *Stat. Model.* **4** 279–298.

CZADO, C., GNEITING, T. and HELD, L. (2009). Predictive model assessment for count data. *Biometrics* **65** 1254–1261. MR2756513

DANON, L., READ, J. M., HOUSE, T. A., VERNON, M. C. and KEELING, M. J. (2013). Social encounter networks: Characterizing Great Britain. *Proc. R. Soc. Lond., B Biol. Sci.* **280** 20131037.

DAVIS, N. F., MCGUIRE, B. B., MAHON, J. A., SMYTH, A. E., O'MALLEY, K. J. and FITZPATRICK, J. M. (2010). The increasing incidence of mumps orchitis: A comprehensive review. *BJU Int.* **105** 1060–1065.

DAWID, A. P. (1984). Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *J. R. Stat. Soc. A, General* **147** 278–292.

Key words and phrases. Social contact patterns, hierarchical Bayesian model, Gaussian Markov random field, integrated nested Laplace approximations, infectious disease transmission.

DODD, P. J., LOOKER, C., PLUMB, I. D., BOND, V., SCHAAP, A., SHANAUBE, K., MUYOYETA, M., VYNNYCKY, E., GODFREY-FAUSSETT, P., CORBETT, E. L., BEYERS, N., AYLES, H. and WHITE, R. G. (2016). Age- and sex-specific social contact patterns and incidence of mycobacterium tuberculosis infection. *Am. J. Epidemiol.* **183** 156–166.

EAMES, K., BANSAL, S., FROST, S. and RILEY, S. (2015). Six challenges in measuring contact networks for use in modelling. *Epidemics* **10** 72–77.

EDMUNDS, W. J., O'CALLAGHAN, C. J. and NOKES, D. J. (1997). Who mixes with whom? A method to determine the contact patterns of adults that may lead to the spread of airborne infections. *Proc. R. Soc. Lond., B Biol. Sci.* **264** 949–957.

EILERS, P. H. C. and MARX, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statist. Sci.* **11** 89–121. With comments and a rejoinder by the authors. MR1435485

FALAGAS, M. E., MOURTZOUKOU, E. G. and VARDAKAS, K. Z. (2007). Sex differences in the incidence and severity of respiratory tract infections. *Respir. Med.* **101** 1845–1863.

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

GOEYVAERTS, N., HENS, N., OGUNJIMI, B., AERTS, M., SHKEDY, Z., DAMME, P. V. and BEUTELS, P. (2010). Estimating infectious disease parameters from data on social contacts and serological status. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **59** 255–277. MR2744473

HENS, N., GOEYVAERTS, N., AERTS, M., SHKEDY, Z., DAMME, P. V. and BEUTELS, P. (2009). Mining social mixing patterns for infectious disease models based on a two-day population survey in Belgium. *BMC Infect. Dis.* **9** 5.

HENS, N., SHKEDY, Z., AERTS, M., FAES, C., VAN DAMME, P. and BEUTELS, P. (2012). *Modeling Infectious Disease Parameters Based on Serological and Social Contact Data*: *A Modern Statistical Perspective*. Springer, New York. MR2986066

HOLMES, C. B., HAUSLER, H. and NUNN, P. (1998). A review of sex differences in the epidemiology of tuberculosis. *Int. J. Tuberc. Lung Dis.* **2** 96–104.

KEELING, M. J. and WHITE, P. J. (2011). Targeting vaccination against novel infections: Risk, age and spatial structure for pandemic influenza in Great Britain. *J. R. Soc. Interface* **8** 661–670.

KLEIN, S. L., PASSARETTI, C., ANKER, M., OLUKOYA, P. and PEKOSZ, A. (2010). The impact of sex, gender and pregnancy on 2009 H1N1 disease. *Biology Sex Differ.* **1** 1–12.

KUCHARSKI, A. J., KWOK, K. O., WEI, V. W. I., COWLING, B. J., READ, J. M., LESSLER, J., CUMMINGS, D. A. and RILEY, S. (2014). The contribution of social behaviour to the transmission of influenza A in a human population. *PLoS Pathog.* **10** e1004206.

KWOK, K. O., COWLING, B. J., WEI, V. W. I., WU, K. M., READ, J. M., LESSLER, J., CUMMINGS, D. A., PEIRIS, J. S. M. and RILEY, S. (2014). Social contacts and the locations in which they occur as risk factors for influenza infection. *Proc. R. Soc. Lond., B Biol. Sci.* **281** 20140709.

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

MEDLOCK, J. and GALVANI, A. P. (2009). Optimizing influenza vaccine distribution. *Science* **325** 1705–1708.

MEHL, M. R., VAZIRE, S., RAMÍREZ-ESPARZA, N., SLATCHER, R. B. and PENNEBAKER, J. W. (2007). Are women really more talkative than men? *Science* **317** 82.

MILLER, E., HOSCHLER, K., HARDELID, P., STANFORD, E., ANDREWS, N. and ZAMBON, M. (2010). Incidence of 2009 pandemic influenza A H1N1 infection in England: A cross-sectional serological study. *Lancet* **375** 1100–1108.

MOSSONG, J., HENS, N., JIT, M., BEUTELS, P., AURANEN, K., MIKOLAJCZYK, R., MASSARI, M., SALMASO, S., TOMBA, G. S., WALLINGA, J., HEIJNE, J., SADKOWSKA-TODYS, M., ROSINSKA, M. and EDMUNDS, W. J. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med.* **5** e74.

NEYROLLES, O. and QUINTANA-MURCI, L. (2009). Sexual inequality in tuberculosis. *PLoS Med.* **6** e1000199.

PASS, R. F., ZHANG, C., EVANS, A., SIMPSON, T., ANDREWS, W., HUANG, M.-L., COREY, L., HILL, J., DAVIS, E., FLANIGAN, C. and CLOUD, G. (2009). Vaccine prevention of maternal cytomegalovirus infection. *N. Engl. J. Med.* **360** 1191–1199.

POLAK, M. F. (1959). Influenzasterfte in de herfst van 1957. *Ned. Tijdschr. Geneeskd.* **103** 1098–109.

READ, J. M., EDMUNDS, W. J., RILEY, S., LESSLER, J. and CUMMINGS, D. A. T. (2012). Close encounters of the infectious kind: Methods to measure social mixing behaviour. *Epidemiol. Infect.* **140** 2117–2130.

ROHANI, P., ZHONG, X. and KING, A. A. (2010). Contact network structure explains the changing epidemiology of pertussis. *Science* **330** 982–985.

ROOS, M., MARTINS, T. G., HELD, L. and RUE, H. (2015). Sensitivity analysis for Bayesian hierarchical models. *Bayesian Anal.* **10** 321–349.

RUE, H. and HELD, L. (2005). *Gaussian Markov Random Fields*: *Theory and Applications*, 1st ed. Chapman & Hall/CRC, Boca Raton, FL.

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

SALATHÉ, M., KAZANDJIEVA, M., LEE, J. W., LEVIS, P., FELDMAN, M. W. and JONES, J. H. (2010). A high-resolution human contact network for infectious disease transmission. *Proc. Natl. Acad. Sci. USA* **107** 22020–22025.

STATLINE (2015). Population on 1 January by age and sex. Available at http://statline.cbs.nl/Statweb/search/?Q=population&LA=EN.

VAN DE KASSTEELE, J., VAN EIJKEREN, J. and WALLINGA, J. (2017). Supplement to "Efficient estimation of age-specific social contact rates between men and women." DOI:10.1214/16-AOAS1006SUPP.

VYNNYCKY, E. and WHITE, R. (2010). *An Introduction to Infectious Disease Modelling*. Oxford Univ. Press, Oxford.

WALLINGA, J., TEUNIS, P. and KRETZSCHMAR, M. (2006). Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. *Am. J. Epidemiol.* **164** 936–944.

WALLINGA, J., VAN BOVEN, M. and LIPSITCH, M. (2010). Optimizing infectious disease interventions during an emerging epidemic. *Proc. Natl. Acad. Sci. USA* **107** 923–928.

WATANABE, S. (2013). A widely applicable Bayesian information criterion. *J. Mach. Learn. Res.* **14** 867–897.

WOOD, S. N. (2006). *Generalized Additive Models*: *An Introduction with R*. Chapman & Hall/CRC, Boca Raton, FL. MR2206355

YOUNG, N. S. and BROWN, K. E. (2004). Parvovirus B19. *N. Engl. J. Med.* **350** 586–597.

The Annals of Applied Statistics
2017, Vol. 11, No. 1, 340–361
DOI: 10.1214/16-AOAS1007

BIOMASS PREDICTION USING A DENSITY-DEPENDENT DIAMETER DISTRIBUTION MODEL[1]

BY ERIN M. SCHLIEP*, ALAN E. GELFAND†, JAMES S. CLARK† AND BRADLEY J. TOMASEK†

*University of Missouri * and Duke University†*

Prediction of aboveground biomass, particularly at large spatial scales, is necessary for estimating global-scale carbon sequestration. Since biomass can be measured only by sacrificing trees, total biomass on plots is never observed. Rather, allometric equations are used to convert individual tree diameter to individual biomass, perhaps with noise. The values for all trees on a plot are then summed to obtain a *derived* total biomass for the plot. Then, with derived total biomasses for a collection of plots, regression models, using appropriate environmental covariates, are employed to attempt explanation and prediction. Not surprisingly, when out-of-sample validation is examined, such a model will predict total biomass well for holdout data because it is obtained using exactly the same derived approach.

Apart from the somewhat circular nature of the regression approach, it also fails to employ the actual observed plot level response data. At each plot, we observe a *random* number of trees, each with an associated diameter, producing a sample of diameters. A model based on this random number of tree diameters provides understanding of how environmental regressors explain abundance of individuals, which in turn explains individual diameters.

We incorporate density dependence because the distribution of tree diameters over a plot of fixed size depends upon the number of trees on the plot. After fitting this model, we can obtain predictive distributions for individual-level biomass and plot-level total biomass. We show that predictive distributions for plot-level biomass obtained from a density-dependent model for diameters will be much different from predictive distributions using the regression approach. Moreover, they can be more informative for capturing uncertainty than those obtained from modeling derived plot-level biomass directly.

We develop a density-dependent diameter distribution model and illustrate with data from the national Forest Inventory and Analysis (FIA) database. We also describe how to scale predictions to larger spatial regions. Our predictions agree (in magnitude) with available wisdom on mean and variation in biomass at the hectare scale.

REFERENCES

BACCINI, A., GOETZ, S., WALKER, W., LAPORTE, N., SUN, M., SULLA-MENASHE, D., HACKLER, J., BECK, P., DUBAYAH, R., FRIEDL, M., SAMANTA, S. and HOUGHTON, R. (2012). Es-

Key words and phrases. Allometry, big O behavior, hierarchical models, Markov chain Monte Carlo, Poisson process.

timated carbon dioxide emissions from tropical deforestation improved by carbon-density maps. *Nature Climate Change* **2** 182–185.

BECHTOLD, W. A. and PATTERSON, P. L. (2005). The enhanced forest inventory and analysis program: National sampling design and estimation procedures. Technical report, US Department of Agriculture Forest Service, Southern Research Station Asheville, NC.

BLACKARD, J., FINCO, M., HELMER, E., HOLDEN, G., HOPPUS, M., JACOBS, D., LISTER, A., MOISEN, G., NELSON, M., RIEMANN, R., RUEFENACHT, B., SALAJANU, D., WEYERMANN, D., WINTERBERGER, K., BRANDEIS, T., CZAPLEWSKI, R., MCROBERTS, R., PATTERSON, P. and TYMCIO, R. (2008). Mapping US forest biomass using nationwide forest inventory data and moderate resolution information. *Remote Sensing of Environment* **112** 1658–1677.

BRZOSTEK, E. R., DRAGONI, D., SCHMID, H. P., RAHMAN, A. F., SIMS, D., WAYSON, C. A., JOHNSON, D. J. and PHILLIPS, R. P. (2014). Chronic water stress reduces tree growth and the carbon sink of deciduous hardwood forests. *Glob. Change Biol.* **20** 2531–2539.

CAMARERO, J. J., GAZOL, A., GALVÁN, J. D., SANGÜESA-BARREDA, G. and GUTIÉRREZ, E. (2015). Disparate effects of global-change drivers on mountain conifer forests: Warming-induced growth enhancement in young trees vs. CO2 fertilization in old trees from wet sites. *Glob. Change Biol.* **21** 738–749.

CHAVE, J., CONDIT, R., AGUILAR, S., HERNANDEZ, A., LAO, S. and PEREZ, R. (2004). Error propagation and scaling for tropical forest biomass estimates. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **359** 409–420.

CHAVE, J., RÉJOU-MÉCHAIN, M., BÚRQUEZ, A., CHIDUMAYO, E., COLGAN, M. S., DELITTI, W. B., DUQUE, A., EID, T., FEARNSIDE, P. M., GOODMAN, R. C., HENRY, M., MARTINEZ-YRIZAR, A., MULLER-LANDAU, H. C., MENCUCCINI, M., NELSON, B. W., NGOMANDA, A., NOGUEIRA, E. M., ORTIZ-MALAVASSI, E., PÉLISSIER, R., PLOTON, P., RYAN, C. M., SALDARRIAGA, J. G. and VIEILLEDENT, G. (2014). Improved allometric models to estimate the aboveground biomass of tropical trees. *Glob. Change Biol.* **20** 3177–3190.

DONG, J., KAUFMANN, R. K., MYNENI, R. B., TUCKER, C. J., KAUPPI, P. E., LISKI, J., BUERMANN, W., ALEXEYEV, V. and HUGHES, M. K. (2003). Remote sensing estimates of boreal and temperate forest woody biomass: Carbon pools, sources, and sinks. *Remote Sensing of Environment* **84** 393–410.

GELFAND, A. E., SAHU, S. K. and CARLIN, B. P. (1995). Efficient parameterisations for normal linear mixed models. *Biometrika* **82** 479–488. MR1366275

HENRY, M., BOMBELLI, A., TROTTA, C., ALESSANDRINI, A., BIRIGAZZI, L., SOLA, G., VIEILLEDENT, G., SANTENOISE, P., LONGUETAUD, F., VALENTINI, R., PICARD, N. and SAINT-ANDRÉ, L. (2013). Globallometree: International platform for tree allometric equations to support volume, biomass and carbon assessment. *IForest-Biogeosciences and Forestry* **6** 326.

JENKINS, J. C., CHOJNACKY, D. C., HEATH, L. S. and BIRDSEY, R. A. (2003). National-scale biomass estimators for United States tree species. *Forest Science* **49** 12–35.

LAMBERT, M., UNG, C. and RAULIER, F. (2005). Canadian national tree aboveground biomass equations. *Canadian Journal of Forest Research* **35** 1996–2018.

MALHI, Y., WOOD, D., BAKER, T. R., WRIGHT, J., PHILLIPS, O. L., COCHRANE, T., MEIR, P., CHAVE, J., ALMEIDA, S., ARROYO, L., HIGUCHI, N., KILLEEN, T. J., LAURANCE, S. G., LAURANCE, W. F., LEWIS, S. L., MONTEAGUDO, A., NEILL, D. A., VARGAS, P. N., PITMAN, N. C. A., QUESADA, C. A., SALOMÃO, R., SILVA, J. N. M., LEZAMA, A. T., TERBORGH, J., MARTÍNES, R. V. and VINCETI, B. (2006). The regional variation of aboveground live biomass in old-growth Amazonian forests. *Glob. Change Biol.* **12** 1107–1138.

MCROBERTS, R. E., MOSER, P., ZIMERMANN OLIVEIRA, L. and VIBRANS, A. C. (2015). A general method for assessing the effects of uncertainty in individual-tree volume model predictions on large-area volume estimates with a subtropical forest illustration. *Canadian Journal of Forest Research* **45** 44–51.

MOLTO, Q., ROSSI, V. and BLANC, L. (2013). Error propagation in biomass estimation in tropical forests. *Methods Ecol. Evol.* **4** 175–183.

PAN, Y., BIRDSEY, R. A., FANG, J., HOUGHTON, R., KAUPPI, P. E., KURZ, W. A., PHILLIPS, O. L., SHVIDENKO, A., LEWIS, S. L., CANADELL, J. G., CIAIS, P., JACKSON, R. B., PACALA, S. W., MCGUIRE, A. D., PIAO, S., RAUTIAINEN, A., SITCH, S. and HAYES, D. (2011). A large and persistent carbon sink in the worlds forests. *Science* **333** 988–993.

PICARD, N., SAINT-ANDRÉ, L. and HENRY, M. (2012). Manual for building tree volume and biomass allometric equations: From field measurement to prediction. Technical report, FAO/CIRAD.

SAATCHI, S. S., HARRIS, N. L., BROWN, S., LEFSKY, M., MITCHARD, E. T., SALAS, W., ZUTTA, B. R., BUERMANN, W., LEWIS, S. L., HAGEN, S., PETROVA, S., WHITE, L., SILMAN, M. and MOREL, A. (2011). Benchmark map of forest carbon stocks in tropical regions across three continents. *Proc. Natl. Acad. Sci. USA* **108** 9899–9904.

SCHLIEP, E. M., GELFAND, A. E., CLARK, J. S. and ZHU, K. (2016). Modeling change in forest biomass across the eastern US. *Environ. Ecol. Stat.* **23** 23–41. MR3458607

SHUGART, H., SAATCHI, S. and HALL, F. (2010). Importance of structure and its measurement in quantifying function of forest ecosystems. *Journal of Geophysical Research*: *Biogeosciences* **115**.

THURNER, M., BEER, C., SANTORO, M., CARVALHAIS, N., WUTZLER, T., SCHEPASCHENKO, D., SHVIDENKO, A., KOMPTER, E., AHRENS, B., LEVICK, S. R. and SCHMULLIUS, C. (2014). Carbon stock and density of northern boreal and temperate forests. *Global Ecology and Biogeography* **23** 297–310.

UNG, C.-H., BERNIER, P. and GUO, X.-J. (2008). Canadian national biomass equations: New parameter estimates that include British Columbia data. *Canadian Journal of Forest Research* **38** 1123–1132.

WHITTAKER, R. H. and WOODWELL, G. M. (1968). Dimension and production relations of trees and shrubs in the Brookhaven Forest, New York. *The Journal of Ecology* 1–25.

WILSON, B. T., WOODALL, C. W. and GRIFFITH, D. M. (2013). Imputing forest carbon stock estimates from inventory plots to a nationally continuous coverage. *Carbon Balance and Management* **8** 1–15.

The Annals of Applied Statistics
2017, Vol. 11, No. 1, 362–392
DOI: 10.1214/16-AOAS1008

A MULTIVARIATE MIXED HIDDEN MARKOV MODEL FOR BLUE WHALE BEHAVIOUR AND RESPONSES TO SOUND EXPOSURE[1]

BY STACY L. DERUITER[*,†], ROLAND LANGROCK[‡,†], TOMAS SKIRBUTAS[†], JEREMY A. GOLDBOGEN[§], JOHN CALAMBOKIDIS[¶], ARI S. FRIEDLAENDER[‖,**] AND BRANDON L. SOUTHALL[**]

Calvin College[*], University of St Andrews[†], Bielefeld University[‡], Stanford University[§], Cascadia Research Collective[¶], Oregon State University[‖] and SEA, Inc.[**]

Characterization of multivariate time series of behaviour data from animal-borne sensors is challenging. Biologists require methods to objectively quantify baseline behaviour, and then assess behaviour changes in response to environmental stimuli. Here, we apply hidden Markov models (HMMs) to characterize blue whale movement and diving behaviour, identifying latent states corresponding to three main underlying behaviour states: shallow feeding, travelling, and deep feeding. The model formulation accounts for inter-whale differences via a computationally efficient discrete random effect, and measures potential effects of experimental acoustic disturbance on between-state transition probabilities. We identify clear differences in blue whale disturbance response depending on the behavioural context during exposure, with whales less likely to initiate deep foraging behaviour during exposure. Findings are consistent with earlier studies using smaller samples, but the HMM approach provides a more nuanced characterization of behaviour changes.

REFERENCES

ALTMAN, R. M. (2007). Mixed hidden Markov models: An extension of the hidden Markov model to the longitudinal data setting. *J. Amer. Statist. Assoc.* **102** 201–210. MR2345538

ANTUNES, R., KVADSHEIM, P. H., LAM, F. P. A., TYACK, P. L., THOMAS, L., WENSVEEN, P. J. and MILLER, P. J. O. (2014). High thresholds for avoidance of sonar by free-ranging long-finned pilot whales (*Globicephala melas*). *Mar. Pollut. Bull.* **83** 165–180.

BAGNIEWSKA, J. M., HART, T., HARRINGTON, L. A. and MACDONALD, D. W. (2013). Hidden Markov analysis describes dive patterns in semiaquatic animals. *Behav. Ecol.* **24** 659–667.

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

DEAN, B., FREEMAN, R., KIRK, H., LEONARD, K., PHILLIPS, R. A., PERRINS, C. M. and GUILFORD, T. (2012). Behavioural mapping of a pelagic seabird: Combining multiple sensors and a hidden Markov model reveals the distribution of at-sea behaviour. *J. R. Soc. Interface* 20120570. DOI:10.1098/rsif.2012.0570.

Key words and phrases. Forward algorithm, hidden Markov model, multivariate time series, numerical maximum likelihood, random effects, blue whales.

DeRuiter, S. L. (2010). Marine animal acoustics. In *An Introduction to Underwater Acoustics* (X. Lurton, ed.) 425–474. Praxis Publishing Limited, Chichester, UK.

DeRuiter, S. L., Southall, B. L., Calambokidis, J., Zimmer, W. M. X., Sadykova, D., Falcone, E. A., Friedlaender, A. S., Joseph, J. E., Moretti, D., Schorr, G. S., Thomas, L. and Tyack, P. L. (2013). First direct measurements of behavioural responses by Cuvier's beaked whales to mid-frequency active sonar. *Biol. Lett.* **9** 20130223. DOI:10.1098/rsbl.2013.0223.

DeRuiter, S. L., Langrock, R., Skirbutas, T., Goldbogen, J. A., Calambokidis, J., Friedlaender, A. S. and Southall, B. L. (2017a). Supplement to "A multivariate mixed hidden Markov model for blue whale behaviour and responses to sound exposure." DOI:10.1214/16-AOAS1008SUPPA.

DeRuiter, S. L., Langrock, R., Skirbutas, T., Goldbogen, J. A., Calambokidis, J., Friedlaender, A. S. and Southall, B. L. (2017b). Supplement to "A multivariate mixed hidden Markov model for blue whale behaviour and responses to sound exposure." DOI:10.1214/16-AOAS1008SUPPB.

Donovan, C. R., Harris, C., Harwood, J. and Milazzo, L. (2012). A simulation-based method for quantifying and mitigating the effects of anthropogenic sound on marine mammals. In *Proceedings of Meetings on Acoustics* **17** 070043. Acoustical Society of America, Melville, NY.

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

Ellison, W. T., Southall, B. L., Clark, C. W. and Frankel, A. S. (2012). A new context-based approach to assess marine mammal behavioral responses to anthropogenic sounds. *Conserv. Biol.* **26** 21–28.

Friedlaender, A. S., Goldbogen, J. A., Hazen, E. L., Calambokidis, J. and Southall, B. L. (2015). Feeding performance by sympatric blue and fin whales exploiting a common prey resource. *Mar. Mamm. Sci.* **31** 345–354.

Friedlaender, A. S., Hazen, E. L., Goldbogen, J. A., Stimpert, A. K., Calambokidis, J. and Southall, B. L. (2016). Prey-mediated behavioral responses of feeding blue whales in controlled sound exposure experiments. *Ecol. Appl.* **26** 1075–1085.

Goldbogen, J. A., Southall, B. L., DeRuiter, S. L., Calambokidis, J., Friedlaender, A. S., Hazen, E. L., Falcone, E. A., Schorr, G. S., Douglas, A., Moretti, D. J., Kyburg, C., McKenna, M. F. and Tyack, P. L. (2013a). Blue whales respond to simulated mid-frequency military sonar. *Proc. Biol. Sci.* **280** 20130657.

Goldbogen, J. A., Friedlaender, A. S., Calambodkis, J., McKenna, M. F., Simon, M. and Nowacek, D. P. (2013b). Integrative approaches to the study of baleen whale diving behavior, feeding performance, and foraging ecology. *Bioscience* **63** 90–100.

Goldbogen, J. A., Hazen, E. L., Friedlaender, A. S., Calambokidis, J., DeRuiter, S. L., Stimpert, A. K. and Southall, B. L. (2015). Prey density and distribution drive the three-dimensional foraging strategies of the largest filter feeder. *Funct. Ecol.* **29** 951–961.

Hart, T., Mann, R., Coulson, T., Pettorelli, N. and Trathan, P. (2010). Behavioural switching in a central place forager: Patterns of diving behaviour in the macaroni penguin (*Eudyptes chrysolophus*). *Mar. Biol.* **157** 1543–1553.

Hazen, E. L., Friedlaender, A. S. and Goldbogen, J. A. (2015). Blue whales (*Balaenoptera musculus*) optimize foraging efficiency by balancing oxygen use and energy gain as a function of prey density. *Sci. Adv.* **1** e1500469.

Holyoak, M., Casagrandi, R., Nathan, R., Revilla, E. and Spiegel, O. (2008). Trends and missing parts in the study of movement ecology. *Proc. Natl. Acad. Sci. USA* **105** 19060–19065.

Houser, D. S. (2006). A method for modeling marine mammal movement and behavior for environmental impact assessment. *IEEE J. Oceanic Eng.* **31** 76–81.

ISOJUNNO, S. and MILLER, P. J. O. (2015). Sperm whale response to tag boat presence: Biologically informed hidden state models quantify lost feeding opportunities. *Ecosphere* **6** 1–46.

JAMES, F. C. and MCCULLOCH, C. E. (1990). Multivariate analysis in ecology and systematics: Panacea or Pandora's box?. *Ann. Rev. Ecolog. Syst.* **21** 129–166.

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

JOHNSON, M. P., HICKMOTT, L. S., SOTO, N. A. and MADSEN, P. T. (2008). Echolocation behaviour adapted to prey in foraging Blainville's beaked whale (*Mesoplodon densirostris*). *Proceedings of the Royal Society of London B: Biological Sciences* **275** 133–139.

KING, S. L., SCHICK, R. S., DONOVAN, C., BOOTH, C. G., BURGMAN, M., THOMAS, L. and HARWOOD, J. (2015). An interim framework for assessing the population consequences of disturbance. *Methods Ecol. Evol.* **6** 1150–1158.

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

LANGROCK, R., MARQUES, T. A., BAIRD, R. W. and THOMAS, L. (2014). Modeling the diving behavior of whales: A latent-variable approach with feedback and semi-Markovian components. *J. Agric. Biol. Environ. Stat.* **19** 82–100. MR3257903

LANGROCK, R., KNEIB, T., SOHN, A. and DERUITER, S. L. (2015). Nonparametric inference in hidden Markov models using P-splines. *Biometrics* **71** 520–528.

MACDONALD, I. L. (2014). Numerical maximisation of likelihood: A neglected alternative to EM? *Int. Stat. Rev.* **82** 296–308. MR3248241

MARUOTTI, A. and RYDÉN, T. (2009). A semiparametric approach to hidden Markov models under longitudinal observations. *Stat. Comput.* **19** 381–393.

MCCLINTOCK, B. T., KING, R., THOMAS, L., MATTHIOPOULOS, J., MCCONNELL, B. J. and MORALES, J. M. (2012). A general discrete-time modeling framework for animal movement using multistate random walks. *Ecol. Monogr.* **82** 335–349.

MCCLINTOCK, B. T., RUSSELL, D. J. F., MATTHIOPOULOS, J. and KING, R. (2013). Combining individual animal movement and ancillary biotelemetry data to investigate population-level activity budgets. *Ecology* **94** 838–849.

MCKELLAR, A. E., LANGROCK, R., WALTERS, J. R. and KESLER, D. C. (2014). Using mixed hidden Markov models to examine behavioral states in a cooperatively breeding bird. *Behav. Ecol.* **26** 148–157.

MILLER, P. J. O., KVADSHEIM, P. H., LAM, F.-P. A., WENSVEEN, P. J., ANTUNES, R., ALVES, A. C., VISSER, F., KLEIVANE, L., TYACK, P. L. and SIVLE, L. D. (2012). The severity of behavioral changes observed during experimental exposures of killer (*Orcinus orca*), long-finned pilot (*Globicephala melas*), and sperm (*Physeter macrocephalus*) whales to naval sonar. *Aquat. Mamm.* **38** 362–401.

MILLER, P. J. O., KVADSHEIM, P. H., LAM, F. P. A., TYACK, P. L., CURE, C., DERUITER, S. L., KLEIVANE, L., SIVLE, L. D., VAN IJSSELMUIDE, S. P., VISSER, F., WENSVEEN, P. J., VON BENDA-BECKMANN, A. M., MARTIN LOPEZ, L. M., NARAZAKI, T. and HOOKER, S. K. (2015). First indications that northern bottlenose whales are sensitive to behavioural disturbance from anthropogenic noise. *R. Soc. Open Sci.* **2** 140484.

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

NATIONAL RESEARCH COUNCIL (2005). *Marine Mammal Populations and Ocean Noise: DeterminIng When Noise Causes Biologically Significant Effects*. National Academies Press, Washington, DC.

NEW, L. F., MORETTI, D. J., HOOKER, S. K., COSTA, D. P. and SIMMONS, S. E. (2013). Using energetic models to investigate the survival and reproduction of beaked whales (family Ziphiidae). *PLoS ONE* **8** e68725.

NEW, L. F., CLARK, J. S., COSTA, D. P., FLEISHMAN, E., HINDELL, M. A., KLANJŠČEK, T., LUSSEAU, D., KRAUS, S., MCMAHON, C. R., ROBINSON, P. W., SCHICK, R. S., SCHWARZ, L. K., SIMMONS, S. E., THOMAS, L., TYACK, P. and HARWOOD, J. (2014). Using short-term measures of behaviour to estimate long-term fitness of southern elephant seals. *Mar. Ecol. Prog. Ser.* **496** 99–108.

NOWACEK, D. P., JOHNSON, M. P. and TYACK, P. L. (2004). Proceedings of the Royal Society of London B: Biological Sciences. *Proceedings of the Royal Society B—Biological Sciences* **271** 227–231.

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

R CORE TEAM (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at https://www.R-project.org/.

SCHICK, R. S., NEW, L. F., THOMAS, L., COSTA, D. P., HINDELL, M. A., MCMAHON, C. R., ROBINSON, P. W., SIMMONS, S. E., THUMS, M., HARWOOD, J. and CLARK, J. S. (2013). Estimating resource acquisition and at-sea body condition of a marine predator. *J. Anim. Ecol.* **82** 1300–1315.

SHANNON, G., MCKENNA, M. F., ANGELONI, L. M., CROOKS, K. R., FRISTRUP, K. M., BROWN, E., WARNER, K. A., NELSON, M. D., WHITE, C., BRIGGS, J., MCFARLAND, S. and WITTEMYER, G. (2015). A synthesis of two decades of research documenting the effects of noise on wildlife. *Biol. Rev. Camb. Philos. Soc.* **91** 982–1005.

SIVLE, L. D., KVADSHEIM, P. H., CURÉ, C., ISOJUNNO, S., WENSVEEN, P. J., LAM, F.-P. A., VISSER, F., KLEIVANE, L., TYACK, P. L., HARRIS, C. M. and MILLER, P. J. O. (2015). Severity of expert-identified behavioural responses of humpback whale, minke whale and northern bottlenose whale to naval sonar. *Aquat. Mamm.* **41** 469–502.

SOUTHALL, B. L., BOWLES, A. E., ELLISON, W. T., FINNERAN, J. J., GENTRY, R. L., GREENE, C. R., KASTAK, D., KETTEN, D. R., MILLER, J. H., NACHTIGALL, P. E., RICHARDSON, W. J., THOMAS, J. A. and TYACK, P. L. (2007). Marine mammal noise exposure criteria: Initial scientific recommendations. *Aquat. Mamm.* **33** 411–521.

SOUTHALL, B. L., MORETTI, D., ABRAHAM, B., CALAMBOKIDIS, J., DERUITER, S. L. and TYACK, P. L. (2012). Marine mammal behavioral response studies in southern California: Advances in technology and experimental methods. *Mar. Technol. Soc. J.* **46** 48–59.

STIMPERT, A. K., DERUITER, S. L., SOUTHALL, B. L., MORETTI, D. J., FALCONE, E. A., GOLDBOGEN, J. A., FRIEDLAENDER, A., SCHORR, G. S. and CALAMBOKIDIS, J. (2014). Acoustic and foraging behavior of a Baird's beaked whale, *Berardius bairdii*, exposed to simulated sonar. *Sci. Rep.* **4** 7031.

TOWNER, A., LEOS-BARAJAS, V., LANGROCK, R., SCHICK, R. S., SMALE, M. J., JEWELL, O., KASCHKE, T. and PAPASTAMATIOU, Y. P. (2016). Sex-specific and individual preferences for hunting strategies in white sharks. *Funct. Ecol.* **30** 1397–1407.

TYACK, P., GORDON, J. and THOMPSON, D. (2003). Controlled exposure experiments to determine the effects of noise on marine mammals. *Mar. Technol. Soc. J.* **37** 41–53.

TYACK, P. L., ZIMMER, W. M. X., MORETTI, D., SOUTHALL, B. L., CLARIDGE, D. E., DURBAN, J. W., CLARK, C. W., D'AMICO, A., DIMARZIO, N., JARVIS, S., MCCARTHY, E., MORRISSEY, R., WARD, J. and BOYD, I. L. (2011). Beaked whales respond to simulated and actual navy sonar. *PLoS ONE* **6** e17009.

VAN DE KERK, M., ONORATO, D. P., CRIFFIELD, M. A., BOLKER, B. M., AUGUSTINE, B. C., MCKINLEY, S. A. and OLI, M. K. (2015). Hidden semi-Markov models reveal multiphasic movement of the endangered Florida panther. *J. Anim. Ecol.* **84** 576–585.

VENZON, D. J. and MOOLGAVKAR, S. H. (1988). A method for computing profile-likelihood-based confidence intervals. *Appl. Stat.* **37** 87.

WOOD, S. N. (2001). Minimizing model fitting objectives that contain spurious local minima by bootstrap restarting. *Biometrics* **57** 240–244.

ZUCCHINI, W., MACDONALD, I. L. and LANGROCK, R. (2016). *Hidden Markov Models for Time Series*: *An Introduction Using R* (*Second Edition*). Chapman & Hall, London.

ZUCCHINI, W., RAUBENHEIMER, D. and MACDONALD, I. L. (2008). Modeling time series of animal behavior by means of a latent-state model with feedback. *Biometrics* **64** 807–815.

The Annals of Applied Statistics
2017, Vol. 11, No. 1, 393–426
DOI: 10.1214/16-AOAS1009

BAYESIAN NONHOMOGENEOUS MARKOV MODELS VIA PÓLYA-GAMMA DATA AUGMENTATION WITH APPLICATIONS TO RAINFALL MODELING[1]

BY TRACY HOLSCLAW*, ARTHUR M. GREENE†, ANDREW W. ROBERTSON† AND PADHRAIC SMYTH*

*University of California, Irvine** *and Columbia University*†

Discrete-time hidden Markov models are a broadly useful class of latent-variable models with applications in areas such as speech recognition, bioinformatics, and climate data analysis. It is common in practice to introduce temporal nonhomogeneity into such models by making the transition probabilities dependent on time-varying exogenous input variables via a multinomial logistic parametrization. We extend such models to introduce additional nonhomogeneity into the emission distribution using a generalized linear model (GLM), with data augmentation for sampling-based inference. However, the presence of the logistic function in the state transition model significantly complicates parameter inference for the overall model, particularly in a Bayesian context. To address this, we extend the recently-proposed Pólya-Gamma data augmentation approach to handle nonhomogeneous hidden Markov models (NHMMs), allowing the development of an efficient Markov chain Monte Carlo (MCMC) sampling scheme. We apply our model and inference scheme to 30 years of daily rainfall in India, leading to a number of insights into rainfall-related phenomena in the region. Our proposed approach allows for fully Bayesian analysis of relatively complex NHMMs on a scale that was not possible with previous methods. Software implementing the methods described in the paper is available via the R package NHMM.

REFERENCES

AILLIOT, P. and MONBET, V. (2012). Markov-switching autoregressive models for wind time series. *Environ. Model. Softw.* **30** 92–101.

AILLIOT, P., ALLARD, D., MONBET, V. and NAVEAU, P. (2015). Stochastic weather generators: An overview of weather type models. *J. SFdS* **156** 101–113. MR3338244

AITCHISON, J. and BENNETT, J. (1970). Polychotomous quantal response by maximum indicant. *Biometrika* **57** 253–262.

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

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

BELLONE, E., HUGHES, J. P. and GUTTORP, P. (2000). A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts. *Clim. Res.* **15** 1–12.

BERROCAL, V. J., GELFAND, A. E. and HOLLAND, D. M. (2010). A bivariate space–time downscaler under space and time misalignment. *Ann. Appl. Stat.* **4** 1942–1975. MR2829942

Key words and phrases. Nonhomogenous hidden Markov model, multivariate time series, Pólya-Gamma latent variables, probit and logit link.

CAREY-SMITH, T., SANSOM, J. and THOMSON, P. (2014). A hidden seasonal switching model for multisite daily rainfall. *Water Resour. Res*. **50** 257–272.

CHALLINOR, A. J., EWERT, F., ARNOLD, S., SIMELTON, E. and FRASER, E. (2009). Crops and climate change: Progress, trends, and challenges in simulating impacts and informing adaptation. *J. Exp. Bot*. **60** 2775–2789.

CHARLES, S. P., BATES, B. C. and HUGHES, J. P. (1999). A spatiotemporal model for downscaling precipitation occurrence and amounts. *J. Geophys. Res*. **104** 31657–31669.

CHARLES, S. P., BATES, B. C., SMITH, I. N. and HUGHES, J. P. (2004). Statistical downscaling of daily precipitation from observed and modelled atmospheric fields. *Hydrol. Process*. **18** 1373–1394.

CHIB, S. and GREENBURG, E. (1998). Analysis of multivariate probit models. *Biometrika* **85** 347–361.

COX, D. R. (1970). *The Analysis of Binary Data*. Methuen & Co., Ltd., London. MR0282453

DEMPSTER, A. P. (1997). The direct use of likelihood for significance testing. *Stat. Comput*. **7** 247–252.

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

DIEBOLD, F. X. and LEE, J. H. (1994). Regime switching with time-varying transition probabilities. In *Nonstationary Time Series Analysis and Cointegrations* (C. W. J. Granger and G. Mixon, eds.) 283–302. Oxford Univ. Press, London.

FILARDO, A. J. and GORDON, S. F. (1998). Business cycle durations. *J. Econometrics* **85** 99–123.

FORNEY, G. D. JR. (1973). The Viterbi algorithm. *Proc. IEEE* **61** 268–278. MR0439384

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

FRÜHWIRTH-SCHNATTER, S. (2006). *Finite Mixture and Markov Switching Models*. Springer Science & Business Media, Berlin.

FRÜHWIRTH-SCHNATTER, S. and FRÜHWIRTH, R. (2007). Auxiliary mixture sampling with applications to logistic models. *Comput. Statist. Data Anal*. **51** 3509–3528.

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

FURRER, E. M. and KATZ, R. W. (2007). Generalized linear modeling approach to stochastic weather generators. *Clim. Res*. **34** 129–144.

GADGIL, S. (2003). The Indian monsoon and its variability. *Annu. Rev. Earth Planet. Sci*. **31** 429–467.

GERMAIN, S. (2010). Bayesian spatio-temporal modelling of rainfall through non-homogenous hidden Markov models. Ph.D. thesis, Newcastle University, Newcastle, UK.

GERSHUNOV, A., SCHNEIDER, N. and BARNET, T. (2001). Low-frequency modulation of the ENSO-Indian monsoon rainfall relationship: Signal or noise? *J. Climate* **14** 2486–2492.

GHIL, M. and ROBERTSON, A. W. (2002). "Waves" vs. "particles" in the atmosphere's phase space: A pathway to long-range forecasting? *Proc. Natl. Acad. Sci. USA* **99** 2493–2500.

GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Amer. Statist. Assoc*. **102** 359–378.

GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. MR1380810

GREENE, A. M., ROBERTSON, A. W. and KIRSHNER, S. (2008). Analysis of Indian monsoon daily rainfall on subseasonal to multidecadal time-scales using a hidden Markov model. *Q. J. R. Meteorol. Soc*. **134** 875–887.

GREENE, A. M., ROBERTSON, A. W., SMYTH, P. and TRIGLIA, S. (2011). Downscaling projectios of the Indian monsoon rainfall using a non-homogeneous hidden Markov model. *Q. J. R. Meteorol. Soc*. **137** 347–359.

HANSEN, J. W., CHALLINOR, A., INES, A., WHEELER, T. and MORON, V. (2006). Translating climate forecasts into agricultural terms: Advances and challenges. *Clim. Res.* **33** 27–41.

HAY, L. E., MCCABE, G. J., WOLOCK, D. M. and AYERS, M. A. (1991). Simulation of precipitation by weather type analysis. *Water Resour. Res.* **27** 493–501.

HEAPS, S. E., BOYS, R. J. and FARROW, M. (2015). Bayesian modelling of rainfall data by using non-homogeneous hidden Markov models and latent Gaussian variables. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **64** 543–568.

HOLMES, C. C. and HELD, L. (2006a). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Anal.* **1** 145–168. MR2227368

HOLMES, C. and HELD, L. (2006b). Response to van der Lans. *Bayesian Anal.* **6** 357–358.

HOLSCLAW, T., GREENE, A. M., ROBERTSON, A. W. and SMYTH, P. (2016). A Bayesian hidden Markov model of daily precipitation over South and East Asia. *J. Hydrometeorol.* **17** 3–25.

HOLSCLAW, T., GREENE, A. M., ROBERTSON, A. W. and SMYTH, P. (2017). Supplement to "Bayesian nonhomogeneous Markov models via Pólya-Gamma data augmentation with applications to rainfall modeling." DOI:10.1214/16-AOAS1009SUPP.

HOOTEN, M. B. and WIKLE, C. K. (2010). Statistical agent-based models for discrete spatio-temporal systems. *J. Amer. Statist. Assoc.* **105** 236–248. MR2757201

HUGHES, J. P. and GUTTORP, P. (1994). A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena. *Water Resour. Res.* **30** 1535–1546.

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

IMAI, K. and VAN DYK, D. A. (2005). MNP: R package for fitting the multinomial probit model. *J. Stat. Softw.* **14** 1–32.

IMMERZEEL, W. W., VAN BEEK, L. P. H. and BIERKENS, M. F. P. (2010). Climate change will affect the Asian water towers. *Science* **328** 1382–1385.

JASRA, A., HOLMES, C. C. and STEPHENS, D. A. (2005). Markov chain Monte Carlo and the label switching problem in Bayesian mixture modelling. *J. Statist. Plann. Inference* **20** 2305–2315.

JOHNDROW, J. E., LUM, K. and DUNSON, D. (2013). Diagonal orthant multinomial probit models. *J. Mach. Learn. Res. Workshop Conf. Proc.* **31** 29–38.

JOSEPH, P. V., GOKULAPALAN, B., NAIR, A. and WILSON, S. S. (2013). Variability of summer monsoon rainfall in India on inter-annual and decadal time scales. *Atmos. Ocean. Sci. Lett.* **6** 398–403.

JURAFSKY, D. and MARTIN, J. H. (2014). *Speech and Language Processing*. Prentice Hall, New York.

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

KATZ, R. and PARLANGE, M. (1995). Generalization of chain-dependent processes: Application to hourly precipitation. *Water Resour. Res.* **31** 1331–1341.

KIM, C.-J., PIGER, J. and STARTZ, R. (2008). Estimation of Markov regime-switching regression models with endogenous switching. *J. Econometrics* **143** 263–273. MR2423067

KIRSHNER, S. (2010). Modeling of multivariate time series using hidden Markov models. Ph.D. thesis, University of California, Irvine.

KIRSHNER, S., SMYTH, P. and ROBERTSON, A. W. (2004). Conditional Chow-Liu tree structures for modeling discrete-valued vector time series. In *Proc. 20th Conf. UAI* 317–324.

LAU, K.-M. and CHAN, P. H. (1986). Aspects of the 40–50 day oscillation during the northern summer as inferred from outgoing longwave radiation. *Mon. Weather Rev.* **114** 1354–1367.

LEE, J. Y., WANG, B., WHEELER, M. C., FU, X., WALISER, D. E. and KANG, I. S. (2013). Real-time multivariate indices for the boreal summer intraseasonal oscillation over the Asian summer monsoon region. *Clim. Dyn.* **40** 493–509.

MACDONALD, I. L. and ZUCCHINI, W. (1997). *Hidden Markov and Other Models for Discrete-Valued Time Series*. *Monographs on Statistics and Applied Probability* **70**. Chapman & Hall, London. MR1692202

MAMON, R. S. and ELLIOTT, R. J., eds. (2007). *Hidden Markov Models in Finance*. *International Series in Operations Research & Management Science* **104**. Springer, New York. MR2407726

MARAUN, D., WETTERHALL, F., IRESON, A. M., CHANDLER, R. E., KENDON, E. J., WIDMANN, M., BRIENEN, S., RUST, H. W., SAUTER, T., THEMESSL, M. et al. (2010). Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Rev. Geophys*. **48** 1–34.

MCCULLAGH, P. and NELDER, J. (1989). *Generalized Linear Models*. Chapman & Hall, New York.

MCCULLOCH, R., POLSON, N. G. and ROSSI, P. E. (2000). A Bayesian analysis of the multinomial probit model with fully identified parameters. *J. Econometrics* **99** 173–193.

MELIGKOTSIDOU, L. and DELLAPORTAS, P. (2011). Forecasting with non-homogeneous hidden Markov models. *Stat. Comput*. **21** 439–449.

MORON, V., ROBERTSON, A. W. and GHIL, M. (2012). Impact of the modulated annual cycle and intraseasonal oscillation on daily-to-interannual rainfall variability across monsoonal India. *Clim. Dyn*. **38** 2409–2435.

NEAL, R. M. (1997). Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. Technical Report No. 9702, Department of Statistics, University of Toronto.

O'BRIEN, S. M. and DUNSON, D. B. (2004). Bayesian multivariate logistic regression. *Biometrics* **60** 739–746. MR2089450

PAAP, R. and FRANCES, P. H. (2000). A dynamic multinomial probit model for brand choices with different short-run effects of marketing mix variables. *J. Appl. Econometrics* **15** 717–744.

PAROLI, R. and SPEZIA, L. (2008). Bayesian inference in non-homogeneous Markov mixtures of periodic autoregressions with state-dependent exogenous variables. *Comput. Statist. Data Anal*. **52** 2311–2330.

PATTERSON, T. A., PARTON, A., LANGROCK, R., BLACKWELL, P. G., THOMAS, L. and KING, R. (2016). Statistical modelling of animal movement: A myopic review and a discussion of good practice. Available at http://arxiv.org/abs/0901.4804.

PIANI, C., WEEDON, G. P., BEST, M., GOMES, S. M., VITERBO, P., HAGEMANN, S. and HAERTER, J. O. (2010). Statistical bias correction of global simulated daily precipitation and temperature for the application of hydrological models. *J. Hydrol*. **395** 199–215.

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

RAJAGOPALAN, B., LALL, U. and TARBOTON, D. G. (1996). Nonhomogeneous Markov model for daily precipitation. *J. Hydrol. Eng*. **1** 33–40.

RAPHAEL, C. (1999). Automatic segmentation of acoustic musical signals using hidden Markov models. *IEEE Trans. Pattern Anal. Mach. Intell*. **21** 360–370.

RIIHIMAKI, J., JYLANKI, P. and VEHTARI, A. (2013). Nested expectation propagation for Gaussian process classification with a multinomial probit likelihood. *J. Mach. Learn. Res*. **14** 75–109.

ROBERT, C. P., RYDÉN, T. and TITTERINGTON, D. M. (2000). Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **62** 57–75.

ROBERTSON, A. W. (2009). Seasonal predictability of daily rainfall statistics over indramayu district, Indonesia. *Int. J. Climatol*. **29** 1449–1462.

RYDÉN, T. (2008). EM versus Markov chain Monte Carlo for estimation of hidden Markov models: A computational perspective. *Bayesian Anal*. **3** 659–688. MR2469793

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

SCOTT, S. L. (2002). Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *J. Amer. Statist. Assoc*. **97** 337–351.

SCOTT, S. L. (2011). Data augmentation, frequentist estimation, and the Bayesian analysis of multinomial logit models. *Statist. Papers* **52** 87–109. MR2777069

SHUKLA, J. and PAOLINO, D. A. (1983). The southern oscillation and long-range forecasting of the summer monsoon rainfall over India. *Mon. Weather Rev.* **111** 1830–1837.

SIEPEL, A. and HAUSSLER, D. (2004). Combining phylogenetic and hidden Markov models in biosequence analysis. *J. Comput. Biol.* **11** 413–428.

SMITH, T. M., REYNOLDS, R. W., PETERSON, T. C. and LAWRIMORE, J. (2008). Improvements to NOAA's historical merged land-ocean surface temperature analysis (1880–2006). *J. Climate* **21** 2283–2296.

SPEZIA, L. (2009). Reversible jump and the label switching problem in hidden Markov models. *Statist. Sci.* **139** 50–67.

SPEZIA, L., COOKSLEY, S. L., BREWER, M. J., DONNELLY, D. and TREE, A. (2014). Modelling species abundance in a river by Negative Binomial hidden Markov models. *Comput. Statist. Data Anal.* **71** 599–614.

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

STERN, R. D. and COE, R. (1984). A model fitting analysis of daily rainfall data. *J. Roy. Statist. Soc. Ser. A* **147** 1–34.

VERMEULEN, S. J., CHALLINOR, A. J., THORNTON, P. K., CAMPBELL, B. M., ERIYAGAMA, N., VERVOORT, J. M., KINYANGI, J., JARVIS, A., LÄDERACH, P., RAMIREZ-VILLEGAS, J. et al. (2013). Addressing uncertainty in adaptation planning for agriculture. *Proc. Natl. Acad. Sci. USA* **110** 8357–8362.

WANG, B. and FAN, Z. (1999). Choice of South Asian summer monsoon indices. *Bull. Am. Meteorol. Soc.* **80** 629–638.

WILKS, D. S. (1998). Multisite generalization of a daily stochastic precipitation generation model. *J. Hydrol.* **210** 178–191.

WILKS, D. S. (1999a). Interannual variability and extreme-value characteristics of several stochastic daily precipitation models. *Agric. For. Meteorol.* **93** 153–170.

WILKS, D. S. (1999b). Multisite downscaling of daily precipitation with a stochastic weather generator. *Clim. Res.* **11** 125–136.

WILKS, D. S. and WILBY, R. L. (1999). The weather generation game: A review of stochastic weather models. *Prog. Phys. Geogr.* **23** 329–357.

WOOLHISER, D. A. and ROLDAN, J. (1982). Stochastic daily precipitation models 2. A comparison of distributions of amounts. *Water Resour. Res.* **18** 1461–1468.

YOO, J. H., ROBERTSON, A. W. and KANG, I.-S. (2010). Analysis of intraseasonal and interannual variability of the Asian summer monsoon using a hidden Markov model. *J. Climate* **23** 5498–5516.

ZHANG, X., BOSCARDIN, W. J. and BELIN, T. R. (2008). Bayesian analysis of multivariate nominal measures using multivariate multinomial probit models. *Comput. Statist. Data Anal.* **52** 3697–3708. MR2427374

ZHANG, Y., WALLACE, J. M. and BATTISTI, D. S. (1997). ENSO-like interdecadal variability: 1900–93. *J. Climate* **10** 1004–1020.

ZUCCHINI, W. and GUTTORP, P. (1991). A hidden Markov model for space–time precipitation. *Water Resour. Res.* **27** 1917–1923.

ZUCCHINI, W., MACDONALD, I. and LANGROCK, R. (2016). *Hidden Markov Models for Time Series*: *An Introduction Using R*. Chapman & Hall, Boca Raton.

The Annals of Applied Statistics
2017, Vol. 11, No. 1, 427–455
DOI: 10.1214/16-AOAS1010

INFERENCE FOR SOCIAL NETWORK MODELS FROM EGOCENTRICALLY SAMPLED DATA, WITH APPLICATION TO UNDERSTANDING PERSISTENT RACIAL DISPARITIES IN HIV PREVALENCE IN THE US

BY PAVEL N. KRIVITSKY[1,2] AND MARTINA MORRIS[1]

University of Wollongong and University of Washington

Egocentric network sampling observes the network of interest from the point of view of a set of sampled actors, who provide information about themselves and anonymized information on their network neighbors. In survey research, this is often the most practical, and sometimes the only, way to observe certain classes of networks, with the sexual networks that underlie HIV transmission being the archetypal case. Although methods exist for recovering some descriptive network features, there is no rigorous and practical statistical foundation for estimation and inference for network models from such data. We identify a subclass of exponential-family random graph models (ERGMs) amenable to being estimated from egocentrically sampled network data, and apply pseudo-maximum-likelihood estimation to do so and to rigorously quantify the uncertainty of the estimates. For ERGMs parametrized to be invariant to network size, we describe a computationally tractable approach to this problem. We use this methodology to help understand persistent racial disparities in HIV prevalence in the US. We also discuss some extensions, including how our framework may be applied to triadic effects when data about ties among the respondent's neighbors are also collected.

REFERENCES

ADMIRAAL, R. (2009). Dynamic network models based on revealed preference for observed relations and egocentric data. Ph.D. thesis, Univ. Washington, Seattle, WA.

AIROLDI, E., BLEI, D., FIENBERG, S., GOLDENBERG, A., XING, E. and ZHENG, A. (2008). *Statistical Network Analysis*: *Models*, *Issues*, *and New Directions*. Springer, Berlin.

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

BROWN, L. D. (1986). *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **9**. IMS, Hayward, CA. MR0882001

BURT, R. S. (1984). Network items and the general social survey. *Soc. Netw.* **6** 293–339.

BUTTS, C. T. (2008). Social network analysis with `sna`. *J. Stat. Softw.* **24** 1–51.

DHANJAL, C., CLÉMENÇON, S., ARAZOZA, H. D., ROSSI, F. and TRAN, V. C. (2011). The Evolution of the Cuban HIV/AIDS Network. Preprint. Available at arXiv:1109.2499.

FELLOWS, I. and HANDCOCK, M. S. (2012). Exponential-family random network models. Preprint. Available at arXiv:1208.0121.

Key words and phrases. Social network, ERGM, random graph, egocentrically sampled data, pseudo maximum likelihood, pseudolikelihood.

FIRTH, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80** 27–38.

FRANK, O. and STRAUSS, D. (1986). Markov graphs. *J. Amer. Statist. Assoc.* **81** 832–842. MR0860518

FULLER, W. A. (2011). *Sampling Statistics*. *Wiley Series in Survey Methodology* **560**. Wiley.

GEYER, C. J. and THOMPSON, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. Roy. Statist. Soc. Ser. B* **54** 657–699. MR1185217

GJOKA, M., SMITH, E. and BUTTS, C. (2014). Estimating clique composition and size distributions from sampled network data. In *Sixth IEEE International Workshop on Network Science for Communication Networks*.

GJOKA, M., SMITH, E. and BUTTS, C. T. (2015). Estimating subgraph frequencies with or without attributes from egocentrically sampled data. Preprint. Available at arXiv:1510.08119.

GOODREAU, S. M., KITTS, J. A. and MORRIS, M. (2009). Birds of a feather, or friend of a friend? Using exponential random graph models to investigate adolescent social networks. *Demography* **46** 103–125.

GOODREAU, S., CASSELS, S., KASPRZYK, D., MONTAÑO, D., GREEK, A. and MORRIS, M. (2012). Concurrent partnerships, acute infection and HIV epidemic dynamics among young adults in Zimbabwe. *AIDS Behav.* **16** 1–11.

GUPTA, S., ANDERSON, R. M. and MAY, R. M. (1989). Networks of sexual contacts: Implications for the pattern of spread of HIV. *AIDS* **3** 807–818.

HÁJEK, J. (1971). Comment on an essay on the logical foundations of survey sampling by Basu, Debabrata. In *Foundations of Statistical Inference*: *Proceedings of the Symposium on the Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.). René Descartes Foundation, Holt McDougal, Dept. Statistics, Univ. Waterloo, Ont., Canada, March 31 to April 9, 1970.

HALLFORS, D. D., IRITANI, B. J., MILLER, W. C. and BAUER, D. J. (2007). Sexual and drug behavior patterns and HIV and STD racial disparities: The need for new directions. *Am. J. Public Health* **97** 125–132.

HAMILTON, D. T. and MORRIS, M. (2010). Consistency of self-reported sexual behavior in surveys. *Arch. Sex. Behav.* **39** 842–860.

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

HANDCOCK, M. S., HUNTER, D. R., BUTTS, C. T., GOODREAU, S. M., KRIVITSKY, P. N. and MORRIS, M. (2014). `ergm`: Fit, simulate and diagnose exponential-family models for networks. The Statnet Project (http://www.statnet.org). R package version 3.1.2.

HUMMEL, R. M., HUNTER, D. R. and HANDCOCK, M. S. (2012). Improving simulation-based algorithms for fitting ERGMs. *J. Comput. Graph. Statist.* **21** 920–939. MR3005804

HUNTER, D. R., GOODREAU, S. M. and HANDCOCK, M. S. (2008a). Goodness of fit for social network models. *J. Amer. Statist. Assoc.* **103** 248–258.

HUNTER, D. R. and HANDCOCK, M. S. (2006). Inference in curved exponential family models for networks. *J. Comput. Graph. Statist.* **15** 565–583. MR2291264

HUNTER, D. R., HANDCOCK, M. S., BUTTS, C. T., GOODREAU, S. M. and MORRIS, M. (2008b). `ergm`: A package to fit, simulate and diagnose exponential-family models for networks. *J. Stat. Softw.* **24** 1–29.

ILLENBERGER, J. and FLÖTTERÖ, G. (2012). Estimating network properties from snowball sampled data. *Soc. Netw.* **34** 701–711.

KOSKINEN, J. H., ROBINS, G. L. and PATTISON, P. E. (2010). Analysing exponential random graph (p-star) models with missing data using Bayesian data augmentation. *Stat. Methodol.* **7** 366–384.

KRIVITSKY, P. N. (2012). Modeling of Dynamic Networks based on Egocentric Data with Durational Information. Technical Report 2012-01, Dept. Statistics, Pennsylvania State Univ., State College, PA.

KRIVITSKY, P. N. and HANDCOCK, M. S. (2014). A separable model for dynamic networks. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 29–46. MR3153932

KRIVITSKY, P. N., HANDCOCK, M. S. and MORRIS, M. (2011). Adjusting for network size and composition effects in exponential-family random graph models. *Stat. Methodol.* **8** 319–339.

KRIVITSKY, P. N. and KOLACZYK, E. D. (2015). On the question of effective sample size in network modeling: An asymptotic inquiry. *Statist. Sci.* **30** 184–198. MR3353102

KRIVITSKY, P. N. and MORRIS, M. (2017). Supplement to "Inference for social network models from egocentrically sampled data, with application to understanding persistent racial disparities in HIV prevalence in the US." DOI:10.1214/16-AOAS1010SUPP.

LAUMANN, E. O., GAGNON, J. H., MICHAEL, R. T. and MICHAELS, S. (1992). National health and social life survey. Univ. Chicago and National Opinion Research Center [producer], Chicago, IL, USA. 1995. Inter-university Consortium for Political and Social Research [distributor], Ann Arbor, MI, USA. 2008-04-17. DOI:10.3886/ICPSR06647.

LAUMANN, E. O., GAGNON, J. H., MICHAEL, R. T. and MICHAELS, S. (1994). *The Social Organization of Sexuality*. Univ. Chicago Press, Chicago, IL.

MARSDEN, P. V. (1981). Models and methods for characterizing the structural parameters of groups. *Soc. Netw.* **3** 1–27.

MARSDEN, P. V. (1987). Core discussion networks of Americans. *Am. Sociol. Rev.* **52** 122–131.

MEASURE DHS (2000–2014). *Demographic and Health Surveys*. ICF International, Fairfax, VA.

MORRIS, M. (1991). A log-linear modeling framework for selective mixing. *Math. Biosci.* **107** 349–377.

MORRIS, M. (1993a). Epidemiology and social networks: Modeling structured diffusion. *Sociol. Methods Res.* **22** 99–126.

MORRIS, M. (1993b). Telling tails explain the discrepancy in sexual partner reports. *Nature* **365** 437–440.

MORRIS, M. and KRETZSCHMAR, M. (1997). Concurrent partnerships and the spread of HIV. *AIDS* **11** 641–648.

MORRIS, M., HANDCOCK, M. S., MILLER, W. C., FORD, C. A., SCHMITZ, J. L., HOBBS, M. M., COHEN, M. S., HARRIS, K. M. and UDRY, J. R. (2006). Prevalence of HIV infection among young adults in the U.S.: Results from the ADD health study. *Am. J. Public Health* **96** 1091–1097.

MORRIS, M., KURTH, A. E., HAMILTON, D. T., MOODY, J. and WAKEFIELD, S. (2009). Concurrent partnerships and HIV prevalence disparities by race: Linking science and public health practice. *Am. J. Public Health* **99** 1023–1031.

NATIONAL CENTER FOR HIV/AIDS, VIRAL HEPATITIS, STD, AND TB PREVENTION (NCHHSTP) (2012). Estimated HIV incidence in the United States, 2007–2010. HIV surveillance supplemental report 17(4), Centers for Disease Control and Prevention. Online. Available at https://www.cdc.gov/hiv/library/reports/hiv-surveillance.html. Retrieved January 8, 2015.

NATIONAL CENTER FOR HIV/AIDS, VIRAL HEPATITIS, STD, AND TB PREVENTION (NCHHSTP) (2013). Diagnoses of HIV infection among adults aged 50 years and older in the United States and dependent areas, 2007–2010. HIV surveillaince supplemental report 18(3), Centers for Disease Control and Prevention. Online. Available at https://www.cdc.gov/hiv/library/reports/hiv-surveillance.html. Retrieved January 8, 2015.

NATIONAL COMMUNICABLE DISEASE CENTER (NCDC) (1967). Morbidity and mortality weekly report: Reported incidence of notifiable diseases in the United States, 1966. Annual Supplement 15(53), U.S. Dept. Health, Education, and Welfare, Atlanta, GA. Online. Available at https://stacks.cdc.gov/view/cdc/615. Retrieved January 8, 2015.

NATIONAL SURVEY OF FAMILY GROWTH STAFF (2002, 2006–2011). National Survey of Family Growth (NSFG). Division of Vital Statistics, National Center for Health Statistics. Available at https://www.cdc.gov/nchs/nsfg/.

PATTISON, P. E., ROBINS, G. L., SNIJDERS, T. A. B. and WANG, P. (2013). Conditional estimation of exponential random graph models from snowball sampling designs. *J. Math. Psych.* **57** 284–296. MR3137882

PFEFFERMANN, D. (1993). The role of sampling weights when modeling survey data. *Int. Stat. Rev.* **61** 317–337.

POPULATION ESTIMATES PROGRAM (2001). Resident population estimates of the United States by age and sex: April 1, 1990 to July 1, 1999, with short-term projection to November 1, 2000. Population Division, U.S. Census Bureau. Online. Available at https://www.census.gov/population/estimates/nation/intfile3-1.txt. Retrieved June 9, 2009.

PUTNAM, R. D. (2000). *Bowling Alone*: *The Collapse and Revival of American Community*. Simon & Schuster, New York.

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

SALGANIK, M. J. and HECKATHORN, D. D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociol. Method.* **34** 193–239.

SHALIZI, C. R. and RINALDO, A. (2013). Consistency under sampling of exponential random graph models. *Ann. Statist.* **41** 508–535.

SMITH, J. A. (2012). Macrostructure from microstructure: Generating whole systems from ego networks. *Sociol. Method.* **42** 155–205.

SNIJDERS, T. A. (2010). Conditional marginalization for exponential random graph models. *J. Math. Sociol.* **34** 239–252.

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

TANFER, K. (1991). National survey of women. In *AIDS/STD Data Archive* (E. A. McKean, K. L. Muller and E. L. Lang, eds.) 17–19. Sociometrics Corporation, Los Altos, CA.

THOMPSON, S. K. and FRANK, O. (2000). Model-based estimation with link-tracing sampling designs. *Surv. Methodol.* **26** 87–98.

TOMAS, A. and GILE, K. J. (2011). The effect of differential recruitment, non-response and non-recruitment on estimators for respondent-driven sampling. *Electron. J. Stat.* **5** 899–934. MR2831520

TROTTER, R. T. II, BALDWIN, J. A. and BOWEN, A. M. (1995). Network structure and proxy network measures of HIV, drug and incarceration risks for active drug users. *Connections* **18** 88–103.

UDRY, J. R. (2003). The National Longitudinal Study of Adolescent Health (Add Health), Waves I & II, 1994–1996; Wave III, 2001–2002. Carolina Population Center, Univ. North Carolina at Chapel Hill. Online. Available at http://www.cpc.unc.edu/projects/addhealth/design/wave3. Retrieved January 8, 2015.

UNAIDS (2014). HIV Estimates with Uncertainty Bounds 1990–2013. Tech. Rep., United Nations.

VAN DUIJN, M. A. J., VAN BUSSCHBACH, J. T. and SNIJDERS, T. A. B. (1999). Multilevel analysis of personal networks as dependent variables. *Soc. Netw.* **21** 187–210.

VOLZ, E. and HECKATHORN, D. D. (2008). Probability based estimation theory for respondent driven sampling. *J. Off. Stat.* **24** 79–97.

WASSERMAN, S. S. and PATTISON, P. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p^*. *Psychometrika* **61** 401–425.

The Annals of Applied Statistics
2017, Vol. 11, No. 1, 456–479
DOI: 10.1214/16-AOAS951

ASSESSING DIFFERENCES IN LEGISLATORS' REVEALED PREFERENCES: A CASE STUDY ON THE 107TH U.S. SENATE

BY CHELSEA L. LOFLAND*, ABEL RODRÍGUEZ* AND SCOTT MOSER†

*University of California, Santa Cruz** *and University of Nottingham*†

Roll call data are widely used to assess legislators' preferences and ideology, as well as test theories of legislative behavior. In particular, roll call data is often used to determine whether the revealed preferences of legislators are affected by outside forces such as party pressure, minority status or procedural rules. This paper describes a Bayesian hierarchical model that extends existing spatial voting models to test sharp hypotheses about differences in preferences using posterior probabilities associated with such hypotheses. We use our model to investigate the effect of the change of party majority status during the 107th U.S. Senate on the revealed preferences of senators. This analysis provides evidence that change in party affiliation might affect the revealed preferences of legislators, but provides no evidence about the effect of majority status on the revealed preferences of legislators.

REFERENCES

ABDI, H. (2007). Bonferroni and Šidák corrections for multiple comparisons. In *Encyclopedia of Measurement and Statistics* (N. J. Salkind, ed.) 103–107. Sage, Thousand Oaks, CA.

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

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

CARROLL, R., LEWIS, J. B., LO, J., POOLE, K. T. and ROSENTHAL, H. (2009). Measuring bias and uncertainty in DW-NOMINATE ideal point estimates via the parametric bootstrap. *Polit. Anal.* **17** 261–275.

CARSON, J. L., CRESPIN, M. H., JENKINS, J. A. and VANDER WIELEN, R. J. (2004). Shirking in the contemporary Congress: A reappraisal. *Polit. Anal.* **12** 176–179.

CLAUSEN, A. R. and WILCOX, C. (1987). Policy partisanship in legislative leadership recruitment and behavior. *Legis. Stud. Q.* **12** 243–263.

CLINTON, J., JACKMAN, S. and RIVERS, D. (2004). The statistical analysis of roll call data. *Am. Polit. Sci. Rev.* **98** 355–370.

CONVERSE, P. E. (1964). *The Nature of Belief Systems in Mass Publics*. Free Press, New York.

ENELOW, J. M. and HINICH, M. J. (1984). *The Spatial Theory of Voting*: *An Introduction*. Cambridge Univ. Press, Cambridge, MA.

FOX, J.-P. (2010). *Bayesian Item Response Modeling*. *Theory and Applications*. Springer, New York. MR2657265

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

Key words and phrases. Spatial voting model, hypothesis testing, spike-and-slab prior, revealed preferences, factor analysis.

GHOSH, J. and DUNSON, D. B. (2009). Default prior distributions and efficient posterior computation in Bayesian factor analysis. *J. Comput. Graph. Statist.* **18** 306–320. MR2749834

HAHN, P. R., CARVALHO, C. M. and SCOTT, J. G. (2012). A sparse factor analytic probit model for congressional voting patterns. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **61** 619–635. MR2960741

JACKMAN, S. (2001). Multidimensional analysis of roll call data via Bayesian simulation: Identification, estimation, inference, and model checking. *Polit. Anal.* **9** 227–241.

JENKINS, J. A. (2000). Examining the robustness of ideological voting: Evidence from the confederate house of representatives. *Amer. J. Polit. Sci.* **44** 811–822.

JESSEE, S. A. (2012). *Ideology and Spatial Voting in American Elections*. Cambridge Univ. Press, Cambridge, MA.

JESSEE, S. A. and THERIAULT, S. M. (2014). The two faces of congressional roll-call voting. *Party Polit.* **20** 836–848.

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

LEWIS, J. B. and POOLE, K. T. (2004). Measuring bias and uncertainty in ideal point estimates via the parametric bootstrap. *Polit. Anal.* **12** 105–127.

LIU, C., RUBIN, D. B. and WU, Y. N. (1998). Parameter expansion to accelerate EM: The PX–EM algorithm. *Biometrika* **85** 755–770. MR1666758

LOFLAND, C. L., RODRÍGUEZ, A. and MOSER, S. (2017). Supplement to "Assessing differences in legislators' revealed preferences: A case study on the 107th U.S. Senate." DOI:10.1214/16-AOAS951SUPP.

MARTIN, A. D. and QUINN, K. M. (2002). Dynamic ideal point estimation via Markov chain Monte Carlo for the US Supreme Court, 1953–1999. *Polit. Anal.* **10** 134–153.

MAY, J. D. (1973). Opinion structure of political parties: The special law of curvilinear disparity. *Polit. Stud.* **21** 135–151.

MCCARTY, N., POOLE, K. T. and ROSENTHAL, H. (2006). *Polarized America*: *The Dance of Ideology and Unequal Riches*. MIT Press, Cambridge, MA.

MERRILL, S., GROFMAN, B., BRUNELL, T. and KOETZLE, W. (1999). The power of ideologically concentrated minorities. *J. Theor. Polit.* **11** 57–74.

NOKKEN, T. P. (2000). Dynamics of congressional loyalty: Party defection and roll-call behavior, 1947–97. *Legis. Stud. Q.* **25** 417–444.

NOKKEN, T. P. and POOLE, K. T. (2004). Congressional party defection in American history. *Legis. Stud. Q.* **29** 545–568.

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

POOLE, K. T. and ROSENTHAL, H. (1984). The polarization of American politics. *J. Polit.* **46** 1061–1079.

POOLE, K. T. and ROSENTHAL, H. (1985). A spatial model for legislative roll call analysis. *Amer. J. Polit. Sci.* **29** 357–384.

POOLE, K. T. and ROSENTHAL, H. (1987). Analysis of congressional coalition patterns: A unidimensional spatial model. *Legis. Stud. Q.* **12(1)** 55–75.

POOLE, K. T. and ROSENTHAL, H. (1991). Patterns of congressional voting. *Amer. J. Polit. Sci.* **35(1)** 228–278.

POOLE, K. and ROSENTHAL, H. (1997). *Congress*: *A Political-Economic History of Roll-Call Voting*. Oxford Univ. Press, Oxford.

RICHTER, M. K. (1966). Revealed preference theory. *Econometrica* **34** 635–645.

RIVERS, D. (2003). Identification of multidimensional item-response models. Technical report, Dept. Political Science, Stanford Univ., Stanford, CA.

ROBERT, C. P. and CASELLA, G. (2004). *Monte Carlo Statistical Methods*, 2nd ed. Springer, New York. MR2080278

ROBERTS, J. M. (2007). The statistical analysis of roll-call data: A cautionary tale. *Legis. Stud. Q.* **32** 341–360.

RODRÍGUEZ, A. and MOSER, S. (2015). Measuring and accounting for strategic abstentions in the US Senate, 1989–2012. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **64** 779–797. MR3415950

ROSENTHAL, H. and VOETEN, E. (2004). Analyzing roll calls with perfect spatial voting: France 1946–1958. *Amer. J. Polit. Sci.* **48** 620–632.

ROTHENBERG, L. S. and SANDERS, M. S. (2000). Severing the electoral connection: Shirking in the contemporary Congress. *Amer. J. Polit. Sci.* 316–325.

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

SCOTT, J. G. and BERGER, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* **38** 2587–2619. MR2722450

SHOR, B., BERRY, C. and MCCARTY, N. (2010). A bridge to somewhere: Mapping state and congressional ideology on a cross-institutional common space. *Legis. Stud. Q.* **35** 417–448.

SHOR, B. and MCCARTY, N. (2011). The ideological mapping of American legislatures. *Am. Polit. Sci. Rev.* **105** 530–551.

SNYDER JR., J. M. and GROSECLOSE, T. (2000). Estimating party influence in congressional roll-call voting. *Amer. J. Polit. Sci.* **44** 193–211.

VARIAN, H. R. (2006). Revealed preference. In *Samuelsonian Economics and the Twenty-First Century* (M. Szenberg, L. Ramrattan and A. A. Gottesman, eds.) 99–115. Oxford Univ. Press, Oxford.