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

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

continued

Vol. 9, No. 2—June 2015

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. 9, No. 2—June 2015

The Annals of Applied Statistics
2015, Vol. 9, No. 2, 547–571
DOI: 10.1214/15-AOAS824

MODELING SEA-LEVEL CHANGE USING ERRORS-IN-VARIABLES INTEGRATED GAUSSIAN PROCESSES[1]

BY NIAMH CAHILL*, ANDREW C. KEMP†, BENJAMIN P. HORTON‡,§ AND ANDREW C. PARNELL*

University College Dublin, Tufts University†, Rutgers University‡ and Nanyang Technological University§*

We perform Bayesian inference on historical and late Holocene (last 2000 years) rates of sea-level change. The input data to our model are tide-gauge measurements and proxy reconstructions from cores of coastal sediment. These data are complicated by multiple sources of uncertainty, some of which arise as part of the data collection exercise. Notably, the proxy reconstructions include temporal uncertainty from dating of the sediment core using techniques such as radiocarbon. The model we propose places a Gaussian process prior on the rate of sea-level change, which is then integrated and set in an errors-in-variables framework to take account of age uncertainty. The resulting model captures the continuous and dynamic evolution of sea-level change with full consideration of all sources of uncertainty. We demonstrate the performance of our model using two real (and previously published) example data sets. The global tide-gauge data set indicates that sea-level rise increased from a rate with a posterior mean of 1.13 mm/yr in 1880 AD (0.89 to 1.28 mm/yr 95% credible interval for the posterior mean) to a posterior mean rate of 1.92 mm/yr in 2009 AD (1.84 to 2.03 mm/yr 95% credible interval for the posterior mean). The proxy reconstruction from North Carolina (USA) after correction for land-level change shows the 2000 AD rate of rise to have a posterior mean of 2.44 mm/yr (1.91 to 3.01 mm/yr 95% credible interval). This is unprecedented in at least the last 2000 years.

REFERENCES

ABRAMOWITZ, M. and STEGUN, I. (1965). *Handbook of Mathematical Functions*. Dover, New York.

ATWATER, B. F. (1987). Evidence for great holocene earthquakes along the outer coast of Washington state. *Science* **236** 942–944.

BANERJEE, S. and FUENTES, M. (2012). Bayesian modeling for large spatial datasets. *Wiley Interdisciplinary Reviews*: *Computational Statistics* **4** 59–66.

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

BARNETT, T. P. (1984). The estimation of global sea level change: A problem of uniqueness. *Journal of Geophysical Research*: *Oceans* **89** 7980–7988.

BIRKS, H. J. B. (1995). Quantitative palaeoenvironmental reconstructions. In *Technical Guide* **5** 161–254. Quaternary Research Association, Cambridge.

Key words and phrases. Bayesian statistics, integrated Gaussian processes, errors-in-variables, proxy reconstruction, tide gauge.

BOON, J. D. (2012). Evidence of sea level acceleration at U.S. and Canadian tide stations, Atlantic Coast, North America. *Journal of Coastal Research* **28** 1437–1445.

CAHILL, N., KEMP, A. C., HORTON, B. P. and PARNELL, A. C. (2015). Supplement to "Modeling sea-level change using errors-in-variables integrated Gaussian processes." DOI:10.1214/15-AOAS824SUPP.

CAZENAVE, A. and LLOVEL, W. (2010). Contemporary sea level rise. *Annual Review of Marine Science* **2** 145–173.

CHANIOTIS, A. K. and POULIKAKOS, D. (2004). High order interpolation and differentiation using B-splines. *J. Comput. Phys.* **197** 253–274. MR2061246

CHURCH, J. A. and WHITE, N. J. (2006). A 20th century acceleration in global sea-level rise. *Geophysical Research Letters* **33**.

CHURCH, J. A. and WHITE, N. J. (2011). Sea-level rise from the late 19th to the early 21st century. *Surveys in Geophysics* **32** 585–602.

CRAMÉR, H. and LEADBETTER, M. R. (1967). *Stationary and Related Stochastic Processes. Sample Function Properties and Their Applications*. Wiley, New York. MR0217860

DEY, D. K., GHOSH, S. K. and MALLICK, B. K., eds. (2000). *Generalized Linear Models: A Bayesian Perspective. Biostatistics* **5**. Dekker, New York. MR1893779

DONNELLY, J. P., CLEARY, P., NEWBY, P. and ETTINGER, R. (2004). Coupling instrumental and geological records of sea-level change: Evidence from southern New England of an increase in the rate of sea-level rise in the late 19th century. *Geophysical Research Letters* **31** L05203.

DOUGLAS, B. C., KEARNEY, M. S. and LEATHERMAN, S. P. (2001). *Sea-Level Rise: History and Consequences*. Academic Press, San Diego, CA.

ENGELHART, S. E., HORTON, B. P., DOUGLAS, B. C., PELTIER, W. R. and TORNQVIST, T. E. (2009). Spatial variability of late Holocene and 20th century sea level rise along the Atlantic coast of the United States. *Geology* **37** 1115–1118.

GEHRELS, W. R. and WOODWORTH, P. L. (2013). When did modern rates of sea-level rise start? *Global and Planetary Change* **100** 263–277.

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

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

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

GORNITZ, V., LEBEDEFF, S. and HANSEN, J. (1982). Global sea level trend in the past century. *Science* **215** 1611–1614.

HASLETT, J. and PARNELL, A. C. (2008). A simple monotone process with application to radiocarbon-dated depth chronologies. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **57** 399–418. MR2526125

HAY, C. C., MORROW, E., KOPP, R. E. and MITROVICA, J. X. (2015). Probabilistic reanalysis of twentieth-century sea-level rise. *Nature* **517** 481–484.

HEIDELBERGER, P. and WELCH, P. D. (1983). Simulation run length control in the presence of an initial transient. *Oper. Res.* **31** 1109–1144.

HOLGATE, S. J. and WOODWORTH, P. L. (2004). Evidence for enhanced coastal sea level rise during the 1990s. *Geophysical Research Letters* **31** L07305.

HOLSCLAW, T., SANSÓ, B., LEE, H. K. H., HEITMANN, K., HABIB, S., HIGDON, D. and ALAM, U. (2013). Gaussian process modeling of derivative curves. *Technometrics* **55** 57–67. MR3038485

HORTON, B. P. and EDWARDS, R. J. (2006). Quantifying Holocene sea-level change using intertidal foraminifera: Lessons from the British Isles. *Cushman Foundation for Foraminiferal Research, Special Publication* **40** 97.

HORTON, B. P., EDWARDS, J. M. and LLOYD, R. J. (1999). UK intertidal foraminiferal distributions: Implications for sea-level studies. *Marine Micropaleontology* **36** 205–223.

HOUSTON, J. R. and DEAN, R. G. (2011). Sea-level acceleration based on U.S. tide gauges and extensions of previous global-gauge analyses. *Journal of Coastal Research* **27** 409–417.

IPCC (2013). Climate change 2013: The physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change. Cambridge Univ. Press, Cambridge.

JEVREJEVA, S., GRINSTED, A., MOORE, J. C. and HOLGATE, S. (2006). Nonlinear trends and multiyear cycles in sea level records. *Journal of Geophysical Research*: *Oceans* **111** C09012.

JEVREJEVA, S., MOORE, J. C., GRINSTED, A. and WOODWORTH, P. L. (2008). Recent global sea level acceleration started over 200 years ago? *Geophysical Research Letters* **35**.

JEVREJEVA, S., MOORE, J. C., GRINSTED, A., MATTHEWS, A. and SPADA, G. (2014). Trends and acceleration in global and regional sea levels since 1807. *Global and Planetary Change* **113** 11–22.

JUGGINS, S. and BIRKS, H. J. B. (2012). Quantiative environmental reconstructions from biological data. In *Tracking Environmental Change Using Lake Sediments* 5 431–494. Springer, Berlin.

KEMP, A. C., HORTON, B. P., CULVER, S. J., CORBETT, D. R., VAN DE PLASSCHE, O., GEHRELS, W. R., DOUGLAS, B. C. and PARNELL, A. C. (2009). Timing and magnitude of recent accelerated sea-level rise (North Carolina, United States). *Geology* **37** 1035–1038.

KEMP, A. C., HORTON, B. P., DONNELLY, J. P., MANN, M. E., VERMEER, M. and RAHMSTORF, S. (2011). Climate related sea-level variations over the past two millennia. *Proc. Natl. Acad. Sci. USA* **108** 11017–11022.

KEMP, A. C., HORTON, B. P., VANE, C. H., CORBETT, D. R., BERNHARDT, C. E., ENGELHART, S. E., ANISFELD, S. C., PARNELL, A. C. and CAHILL, N. (2013). Sea-level change during the last 2500 years in New Jersey, USA. *Quaternary Science Reviews* **81** 90–104.

KOPP, R. E. (2013). Does the mid-atlantic united states sea level acceleration hot spot reflect ocean dynamic variability? *Geophysical Research Letters* **40** 3981–3985.

LIANG, H. and WU, H. (2008). Parameter estimation for differential equation models using a framework of measurement error in regression models. *J. Amer. Statist. Assoc.* **103** 1570–1583. MR2504205

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

LONG, A. J., BARLOW, N. L. M., GEHRELS, W. R., SAHER, M. H., WOODWORTH, P. L., SCAIFE, R. G., BRAIN, M. J. and CAHILL, N. (2014). Contrasting records of sea-level change in the eastern and western North Atlantic during the last 300 years. *Earth and Planetary Science Letters* **388** 110–122.

MANN, M. E., ZHANG, Z., HUGHES, M. K., BRADLEY, R. S., MILLER, S. K., RUTHERFORD, S. and NI, F. (2008). Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia. *Proc. Natl. Acad. Sci. USA* **105** 13252–13257.

MARDIA, K. V., KENT, J. T., GOODALL, C. R. and LITTLE, J. A. (1996). Kriging and splines with derivative information. *Biometrika* **83** 207–221. MR1399165

MORRIS, J. T., SUNDARESHWAR, P. V., NIETCH, C. T., KJERFVE, B. and CAHOON, D. R. (2002). Response of coastal wetlands to rising sea level. *Ecology* **83** 2869–2877.

NEREM, R. S., CHAMBERS, D., CHOE, C. and MITCHUM, G. T. (2010). Estimating mean sea level change from the TOPEX and Jason altimeter missions. *Marine Geodesy* **33** 435–446.

NICHOLLS, R. J. and CAZENAVE, A. (2010). Sea-level rise and its impact on coastal zones. *Science* **328** 1517–1520.

O'HAGAN, A. (1992). Some Bayesian numerical analysis. In *Bayesian Statistics*, 4 (*PeñíScola*, 1991) 345–363. Oxford Univ. Press, New York. MR1380285

PARNELL, A. C., BUCK, C. E. and DOAN, T. K. (2011). A review of statistical chronology models for high-resolution, proxy-based Holocene palaeoenvironmental reconstruction. *Quaternary Science Reviews* **30** 2948–2960.

PARNELL, A. C., HASLETT, J., ALLEN, J. R. M., BUCK, C. E. and HUNTLEY, B. (2008). A flexible approach to assessing synchroneity of past events using Bayesian reconstructions of sedimentation history. *Quaternary Science Reviews* **27** 1872–1885.

PELTIER, W. R. (2004). Global glacial isostasy and the surface of the ice-age Earth: The ICE-5G (VM2) model and GRACE. *Annual Review of Earth and Planetary Sciences* **32** 111–149.

PELTIER, W. R. and TUSHINGHAM, A. M. (1991). Influence of glacial isostatic adjustment on tide gauge measurements of secular sea level change. *Journal of Geophysical Research*: *Solid Earth* **96** 6779–6796.

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*, *TU Wien*.

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

PLUMMER, M., BEST, N., COWLES, K. and VINES, K. (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News* **6** 7–11.

RAHMSTORF, S. (2007). A semi empirical approach to projecting future sea-level rise. *Science* **315** 368–370.

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

SALLENGER, A. H., DORAN, K. S. and HOW'D, P. A. (2012). Hotspot of accelerated sea-level rise on the Atlantic coast of North America. *Nature Clim*. *Change* **2** 884–888.

SCOTT, D. B. and MEDIOLI, F. S. (1978). Vertical zonations of marsh foraminifera as accurate indicators of former sea levels. *Nature* **272** 528–531.

SHENNAN, I. and HORTON, B. P. (2002). Holocene land- and sea-level changes in Great Britain. *Journal of Quaternary Science* **17** 511–526.

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

WHITE, N. J., CHURCH, J. A. and GREGORY, J. M. (2005). Coastal and global averaged sea level rise for 1950 to 2000. *Geophysical Research Letters* **32** L01601.

WILLIAMS, C. K. I. and RASMUSSEN, C. E. (1996). *Gaussian Processes for Regression*. MIT Press, Cambridge.

WOODWORTH, P. L. and PLAYER, R. (2003). The permanent service for mean sea level: An update to the 21st century. *Journal of Coastal Research* **19** 287–295.

WOODWORTH, P. L., WHITE, N. J., JEVREJEVA, S., HOLGATE, S. J., CHURCH, J. A. and GEHRELS, W. R. (2009). Evidence for the accelerations of sea level on multi-decade and century timescales. *International Journal of Climatology* **29** 777–789.

YAGLOM, I. M. (2011). *Geometric Transformations*. *I*. Mathematical Association of America, Washington, DC. MR2538066

The Annals of Applied Statistics
2015, Vol. 9, No. 2, 572–596
DOI: 10.1214/15-AOAS809

SEX, LIES AND SELF-REPORTED COUNTS: BAYESIAN MIXTURE MODELS FOR HEAPING IN LONGITUDINAL COUNT DATA VIA BIRTH–DEATH PROCESSES[1]

BY FORREST W. CRAWFORD[2,*], ROBERT E. WEISS[3,†] AND MARC A. SUCHARD[4,†,‡]

Yale School of Public Health[*], *UCLA Fielding School of Public Health*[†] *and David Geffen School of Medicine at UCLA*[‡]

Surveys often ask respondents to report nonnegative counts, but respondents may misremember or round to a nearby multiple of 5 or 10. This phenomenon is called heaping, and the error inherent in heaped self-reported numbers can bias estimation. Heaped data may be collected cross-sectionally or longitudinally and there may be covariates that complicate the inferential task. Heaping is a well-known issue in many survey settings, and inference for heaped data is an important statistical problem. We propose a novel reporting distribution whose underlying parameters are readily interpretable as rates of misremembering and rounding. The process accommodates a variety of heaping grids and allows for quasi-heaping to values nearly but not equal to heaping multiples. We present a Bayesian hierarchical model for longitudinal samples with covariates to infer both the unobserved true distribution of counts and the parameters that control the heaping process. Finally, we apply our methods to longitudinal self-reported counts of sex partners in a study of high-risk behavior in HIV-positive youth.

REFERENCES

BAILEY, N. T. J. (1964). *The Elements of Stochastic Processes with Applications to the Natural Sciences*. Wiley, New York. MR0165572

BAR, H. Y. and LILLARD, D. R. (2012). Accounting for heaping in retrospectively reported event data—A mixture-model approach. *Stat. Med.* **31** 3347–3365. MR3041816

BROWN, R. A., BURGESS, E. S., SALES, S. D., WHITELEY, J. A., EVANS, D. M. and MILLER, I. W. (1998). Reliability and validity of a smoking timeline follow-back interview. *Psychology of Addictive Behaviors* **12** 101–112.

BROWNING, M., CROSSLEY, T. F. and WEBER, G. (2003). Asking consumption questions in general purpose surveys. *The Economic Journal* **113** F540–F567.

CRAWFORD, F. W., MININ, V. N. and SUCHARD, M. A. (2014). Estimation for general birth–death processes. *J. Amer. Statist. Assoc.* **109** 730–747. MR3223746

CRAWFORD, F. W. and SUCHARD, M. A. (2012). Transition probabilities for general birth–death processes with applications in ecology, genetics, and evolution. *J. Math. Biol.* **65** 553–580. MR2960857

CRAWFORD, F. W., WEISS, R. E. and SUCHARD, M. A. (2015). Supplement to "Sex, lies and self-reported counts: Bayesian mixture models for heaping in longitudinal count data via birth–death processes." DOI:10.1214/15-AOAS809SUPP.

Key words and phrases. Bayesian hierarchical model, coarse data, continuous-time Markov chain, heaping, mixture model, rounding.

CROCKETT, A. and CROCKETT, R. (2006). Consequences of data heaping in the British religious census of 1851. *Historical Methods*: *A Journal of Quantitative and Interdisciplinary History* **39** 24–46.

FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications*. Wiley, New York.

FENTON, K. A., JOHNSON, A. M., MCMANUS, S. and ERENS, B. (2001). Measuring sexual behaviour: Methodological challenges in survey research. *Sexually Transmitted Infections* **77** 84–92.

GHOSH, P. and TU, W. (2009). Assessing sexual attitudes and behaviors of young women: A joint model with nonlinear time effects, time varying covariates, and dropouts. *J. Amer. Statist. Assoc.* **104** 474–485. MR2751432

GOLUBJATNIKOV, R., PFISTER, J. and TILLOTSON, T. (1983). Homosexual promiscuity and the fear of AIDS. *The Lancet* **322** 681.

GRUNWALD, G. K., BRUCE, S. L., JIANG, L., STRAND, M. and RABINOVITCH, N. (2011). A statistical model for under- or overdispersed clustered and longitudinal count data. *Biom. J.* **53** 578–594. MR2829179

HEITJAN, D. F. (1989). Inference from grouped continuous data: A review. *Statist. Sci.* **4** 164–179.

HEITJAN, D. F. and RUBIN, D. B. (1990). Inference from coarse data via multiple imputation with application to age heaping. *J. Amer. Statist. Assoc.* **85** 304–314.

HEITJAN, D. F. and RUBIN, D. B. (1991). Ignorability and coarse data. *Ann. Statist.* **19** 2244–2253. MR1135174

HOBSON, R. (1976). Properties preserved by some smoothing functions. *J. Amer. Statist. Assoc.* **71** 763–766.

HUTTENLOCHER, J., HEDGES, L. V. and BRADBURN, N. M. (1990). Reports of elapsed time: Bounding and rounding processes in estimation. *Journal of Experimental Psychology*: *Learning*, *Memory*, *and Cognition* **16** 196–213.

JACOBSEN, M. and KEIDING, N. (1995). Coarsening at random in general sample spaces and random censoring in continuous time. *Ann. Statist.* **23** 774–786. MR1345200

KARLIN, S. and MCGREGOR, J. L. (1957). The differential equations of birth-and-death processes, and the Stieltjes moment problem. *Trans. Amer. Math. Soc.* **85** 489–546. MR0091566

KLAR, B., PARTHASARATHY, P. R. and HENZE, N. (2010). Zipf and Lerch limit of birth and death processes. *Probab. Engrg. Inform. Sci.* **24** 129–144. MR2575846

KLOVDAHL, A. S., POTTERAT, J. J., WOODHOUSE, D. E., MUTH, J. B., MUTH, S. Q. and DARROW, W. W. (1994). Social networks and infectious disease: The Colorado Springs study. *Social Science & Medicine* **38** 79–88.

LANGE, K. (2010). *Applied Probability*, 2nd ed. Springer, New York. MR2680838

LEE, J., WEISS, R. E. and SUCHARD, M. A. (2014). Using a birth–death process to account for reporting errors in longitudinal self-reported counts of behavior. Available at arXiv:1410.6870.

LINDLEY, D. V. (1950). Grouping corrections and maximum likelihood equations. *Math. Proc. Cambridge Philos. Soc.* **46** 106–110. MR0032141

MCLAIN, A. C., SUNDARAM, R., THOMA, M., LOUIS, B. and GERMAINE, M. (2014). Semiparametric modeling of grouped current duration data with preferential reporting. *Stat. Med.* **33** 3961–3972. MR3261055

MURPHY, J. A. and O'DONOHOE, M. R. (1975). Some properties of continued fractions with applications in Markov processes. *J. Inst. Math. Appl.* **16** 57–71. MR0393922

MYERS, R. J. (1954). Accuracy of age reporting in the 1950 United States census. *J. Amer. Statist. Assoc.* **49** 826–831.

MYERS, R. J. (1976). An instance of reverse heaping of ages. *Demography* **13** 577–580.

NOVOZHILOV, A. S., KAREV, G. P. and KOONIN, E. V. (2006). Biological applications of the theory of birth-and-death processes. *Brief. Bioinformatics* **7** 70–85.

RENSHAW, E. (2011). *Stochastic Population Processes*: *Analysis*, *Approximations*, *Simulations*. Oxford Univ. Press, Oxford. MR2865609

ROBERTS, J. M. JR. and BREWER, D. D. (2001). Measures and tests of heaping in discrete quantitative distributions. *J. Appl. Stat.* **28** 887–896. MR1863441

ROTHERAM-BORUS, M. J., LEE, M. B., MURPHY, D. A., FUTTERMAN, D., DUAN, N., BIRNBAUM, J. M. and LIGHTFOOT, M. (2001). Efficacy of a preventive intervention for youths living with HIV. *American Journal of Public Health* **91** 400–405.

ROWLAND, M. L. (1990). Self-reported weight and height. *Am. J. Clin. Nutr.* **52** 1125–1133.

SCHAEFFER, N. C. (1999). Asking questions about threatening topics: A selective overview. In *The Science of Self-Report*: *Implications for Research and Practice* (A. A. Stone, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman and V. S. Cain, eds.). Lawrence Erlbaum Associates, Mahwah, NJ.

SCHNEEWEISS, H. and AUGUSTIN, T. (2006). Some recent advances in measurement error models and methods. *Allg. Stat. Arch.* **90** 183–197. MR2255581

SCHNEEWEISS, H. and KOMLOS, J. (2009). Probabilistic rounding and Sheppard's correction. *Stat. Methodol.* **6** 577–593. MR2565305

SCHNEEWEISS, H., KOMLOS, J. and AHMAD, A. S. (2010). Symmetric and asymmetric rounding: A review and some new results. *AStA Adv. Stat. Anal.* **94** 247–271. MR2733174

SHEPPARD, W. F. (1897). On the calculation of the most probable values of frequency-constants, for data arranged according to equidistant division of a scale. *Proc. Lond. Math. Soc.* (3) **1** 353–380. MR1576445

SINGH, K. K., SUCHINDRAN, C. M. and SINGH, R. S. (1994). Smoothed breastfeeding durations and waiting time to conception. *Biodemography and Social Biology* **41** 229–239.

STOCKWELL, E. G. and WICKS, J. W. (1974). Age heaping in recent national censuses. *Biodemography and Social Biology* **21** 163–167.

TALLIS, G. M. (1967). Approximate maximum likelihood estimates from grouped data. *Technometrics* **9** 599–606. MR0224201

WANG, H. and HEITJAN, D. F. (2008). Modeling heaping in self-reported cigarette counts. *Stat. Med.* **27** 3789–3804. MR2526609

WANG, H., SHIFFMAN, S., GRIFFITH, S. D. and HEITJAN, D. F. (2012). Truth and memory: Linking instantaneous and retrospective self-reported cigarette consumption. *Ann. Appl. Stat.* **6** 1689–1706. MR3058680

WEINHARDT, L. S., FORSYTH, A. D., CAREY, M. P., JAWORSKI, B. C. and DURANT, L. E. (1998). Reliability and validity of self-report measures of HIV-related sexual behavior: Progress since 1990 and recommendations for research and practice. *Archives of Sexual Behavior* **27** 155–180.

WESTOFF, C. F. (1974). Coital frequency and contraception. *Family Planning Perspectives* **6** 136–141.

WIEDERMAN, M. W. (1997). The truth must be in here somewhere: Examining the gender discrepancy in self-reported lifetime number of sex partners. *Journal of Sex Research* **34** 375–386.

WRIGHT, D. E. and BRAY, I. (2003). A mixture model for rounded data. *The Statistician* **52** 3–13. MR1973879

The Annals of Applied Statistics
2015, Vol. 9, No. 2, 597–620
DOI: 10.1214/15-AOAS811

REGRESSION BASED PRINCIPAL COMPONENT ANALYSIS FOR SPARSE FUNCTIONAL DATA WITH APPLICATIONS TO SCREENING GROWTH PATHS

BY WENFEI ZHANG AND YING WEI

Columbia University

Growth charts are widely used in pediatric care for assessing childhood body size measurements (e.g., height or weight). The existing growth charts screen one body size at a single given age. However, when a child has multiple measures over time and exhibits a growth path, how to assess those measures jointly in a rigorous and quantitative way remains largely undeveloped in the literature. In this paper, we develop a new method to construct growth charts for growth paths. A new estimation algorithm using alternating regressions is developed to obtain principal component representations of growth paths (sparse functional data). The new algorithm does not rely on strong distribution assumptions and is computationally robust and easily incorporates subject level covariates, such as parental information. Simulation studies are conducted to investigate the performance of our proposed method, including comparisons to existing methods. When the proposed method is applied to monitor the puberty growth among a group of Finnish teens, it yields interesting insights.

REFERENCES

ABDOUS, B. and THEODORESCU, R. (1992). Note on the spatial quantile of a random vector. *Statist. Probab. Lett.* **13** 333–336. MR1160756

CHAKRABORTY, B. (2003). On multivariate quantile regression. *J. Statist. Plann. Inference* **110** 109–132. MR1944636

CHAUDHURI, P. (1996). On a geometric notion of quantiles for multivariate data. *J. Amer. Statist. Assoc.* **91** 862–872. MR1395753

CHEN, C., HE, X. and WEI, Y. (2008). Lower rank approximation of matrices based on fast and robust alternating regression. *J. Comput. Graph. Statist.* **17** 186–200. MR2424801

CHEN, K. and MÜLLER, H.-G. (2012). Conditional quantile analysis when covariates are functions, with application to growth data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 67–89. MR2885840

COLE, T. J. (1988). Fitting smoothed centile curves to reference data. *J. Roy. Statist. Soc. Ser. A* **151** 385–418.

COLE, T. J. and GREEN, P. J. (1992). Smoothing reference centile curves: The LMS method and penalized likelihood. *Stat. Med.* **11** 1305–1319.

CROUX, C., FILZMOSER, P., PISON, G. and ROUSSEEUW, P. J. (2003). Fitting multiplicative models by robust alternating regressions. *Stat. Comput.* **13** 23–36. MR1973864

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

Key words and phrases. Growth charts, sparse functional data, longitudinal data, principal component analysis.

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

GRAVES, S., HOOKER, G. and RAMSAY, J. (2009). Functional data analysis with R and MATLAB.

HAN, B. and LIM, N. (2010). Estimating conditional proportion curves by regression residuals. *Stat. Med.* **29** 1443–1454. MR2758127

HANSEN, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory* **24** 726–748. MR2409261

HETTMANSPERGER, T. P., NYBLOM, J. and OJA, H. (1992). On multivariate notions of sign and rank. In L_1*-Statistical Analysis and Related Methods* (*Neuchâtel*, 1992) 267–278. North-Holland, Amsterdam. MR1214838

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

KOLTCHINSKII, V. I. (1997). M-estimation, convexity and quantiles. *Ann. Statist.* **25** 435–477. MR1439309

LEGLER, J. D. and ROSE, L. C. (1998). Assessment of abnormal growth curves. *Am. Fam. Phys.* **58** 153–158.

LIU, R. Y., PARELIUS, J. M. and SINGH, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *Ann. Statist.* **27** 783–858. MR1724033

LOEVE, M. (1978). Probability theory, Vol. II. *Grad. Texts in Math.* **46** 0–387.

MCDERMOTT, J. P. and LIN, D. K. (2007). Quantile contours and multivariate density estimation for massive datasets via sequential convex hull peeling. *IIE Trans.* **39** 581–591.

PARZEN, E. (1979). Nonparametric statistical data modeling. *J. Amer. Statist. Assoc.* **74** 105–131. MR0529528

PENG, J. and PAUL, D. (2009). A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data. *J. Comput. Graph. Statist.* **18** 995–1015. MR2598035

PENG, J. and PAUL, D. (2011). fpca: Restricted MLE for functional principal components analysis. R package version 0.2-1.

PERE, A. (2000). Comparison of two methods for transforming height and weight to normality. *Annals of Human Biology* **27** 35–45.

SCHEIKE, T. H., ZHANG, M.-J. and JUUL, A. (1999). Comparing reference charts. *Biom. J.* **41** 679–687.

SERFLING, R. (2002). Quantile functions for multivariate analysis: Approaches and applications. *Stat. Neerl.* **56** 214–232. MR1916321

THOMAS, G. B., FINNEY, R. L. and WEIR, M. D. (1988). *Calculus and Analytic Geometry* **7**. Addison-Wesley, Reading, MA.

THOMPSON, M. L. and FATTI, L. (1997). Construction of multivariate centile charts for longitudinal measurements. *Stat. Med.* **16** 333–345.

TREFETHEN, L. N. and BAU, D. III (1997). *Numerical Linear Algebra*. SIAM, Philadelphia, PA. MR1444820

WEI, Y. (2008). An approach to multivariate covariate-dependent quantile contours with application to bivariate conditional growth charts. *J. Amer. Statist. Assoc.* **103** 397–409. MR2420242

WEI, Y., PERE, A., KOENKER, R. and HE, X. (2006). Quantile regression methods for reference growth charts. *Stat. Med.* **25** 1369–1382. MR2226792

WOLD, H. (1966). Nonlinear estimation by iterative least square procedures. In *Research Papers in Statistics* (*Festschrift J. Neyman*) 411–444. Wiley, New York. MR0210250

YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100** 577–590. MR2160561

ZHANG, W. (2012). Regression based principal component analysis for sparse functional data with applications to screening pubertal growth paths. Ph.D. thesis, Columbia Univ., New York.

ZHANG, W. and WEI, Y. (2015). Supplement to “Regression based principal component analysis for sparse functional data with applications to screening growth paths.” DOI:10.1214/15-AOAS811SUPP.

ZUO, Y. and SERFLING, R. (2000). General notions of statistical depth function. *Ann. Statist.* **28** 461–482. MR1790005

The Annals of Applied Statistics
2015, Vol. 9, No. 2, 621–639
DOI: 10.1214/15-AOAS817

A BAYESIAN FEATURE ALLOCATION MODEL FOR TUMOR HETEROGENEITY

BY JUHEE LEE[*], PETER MÜLLER[†,1], KAMALAKAR GULUKOTA[‡] AND YUAN JI[‡,§,1]

University of California Santa Cruz[*], University of Texas, Austin[†], NorthShore University HealthSystem[‡] and University of Chicago[§]

We develop a feature allocation model for inference on genetic tumor variation using next-generation sequencing data. Specifically, we record single nucleotide variants (SNVs) based on short reads mapped to human reference genome and characterize tumor heterogeneity by latent haplotypes defined as a scaffold of SNVs on the same homologous genome. For multiple samples from a single tumor, assuming that each sample is composed of some sample-specific proportions of these haplotypes, we then fit the observed variant allele fractions of SNVs for each sample and estimate the proportions of haplotypes. Varying proportions of haplotypes across samples is evidence of tumor heterogeneity since it implies varying composition of cell subpopulations. Taking a Bayesian perspective, we proceed with a prior probability model for all relevant unknown quantities, including, in particular, a prior probability model on the binary indicators that characterize the latent haplotypes. Such prior models are known as feature allocation models. Specifically, we define a simplified version of the Indian buffet process, one of the most traditional feature allocation models. The proposed model allows overlapping clustering of SNVs in defining latent haplotypes, which reflects the evolutionary process of subclonal expansion in tumor samples.

REFERENCES

BRODERICK, T., PITMAN, J. and JORDAN, M. I. (2013). Feature allocations, probability functions, and paintboxes. *Bayesian Anal*. **8** 801–836. MR3150470

BRODERICK, T., JORDAN, M. I. and PITMAN, J. (2013). Clusters and features from combinatorial stochastic processes. *Statist. Sci*. **28** 289–312.

CASELLA, G. and MORENO, E. (2006). Objective Bayesian variable selection. *J. Amer. Statist. Assoc*. **101** 157–167. MR2268035

CHURCH, D. M., SCHNEIDER, V. A., GRAVES, T., AUGER, K., CUNNINGHAM, F., BOUK, N., CHEN, H.-C., AGARWALA, R., MCLAREN, W. M., RITCHIE, G. R. S. et al. (2011). Modernizing reference genome assemblies. *PLoS Biol*. **9** e1001091.

ENGLE, L. J., SIMPSON, C. L. and LANDERS, J. E. (2006). Using high-throughput SNP technologies to study cancer. *Oncogene* **25** 1594–1601.

ERICHSEN, H. and CHANOCK, S. (2004). SNPs in cancer research and treatment. *British Journal of Cancer* **90** 747–751.

Key words and phrases. Haplotypes, feature allocation models, Indian buffet process, Markov chain Monte Carlo, next-generation sequencing, random binary matrices, variant calling.

GERLINGER, M., ROWAN, A. J., HORSWELL, S., LARKIN, J., ENDESFELDER, D., GRONROOS, E., MARTINEZ, P., MATTHEWS, N., STEWART, A., TARPEY, P., VARELA, I., PHILLIMORE, B., BEGUM, S., MCDONALD, N. Q., BUTLER, A., JONES, D., RAINE, K., LATIMER, C., SANTOS, C. R., NOHADANI, M., EKLUND, A. C., SPENCER-DENE, B., CLARK, G., PICKERING, L., STAMP, G., GORE, M., SZALLASI, Z., DOWNWARD, J., FUTREAL, P. A. and SWANTON, C. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med*. **366** 883–892.

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

GRIFFITHS, T. and GHAHRAMANI, Z. (2005). Infinite latent feature models and the Indian buffet process. Technical Report 2005-001, Gatsby Computational Neuroscience Unit, 2005.

JI, Y., XU, Y., ZHANG, Q., TSUI, K.-W., YUAN, Y., NORRIS, C. JR., LIANG, S. and LIANG, H. (2011). BM-map: Bayesian mapping of multireads for next-generation sequencing data. *Biometrics* **67** 1215–1224. MR2872372

KANEHISA, M., GOTO, S., FURUMICHI, M., TANABE, M. and HIRAKAWA, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res*. **38** D355–D360.

LANDAU, D. A., CARTER, S. L., STOJANOV, P., MCKENNA, A., STEVENSON, K., LAWRENCE, M. S., SOUGNEZ, C., STEWART, C., SIVACHENKO, A., WANG, L., WAN, Y., ZHANG, W., SHUKLA, S. A., VARTANOV, A., FERNANDES, S. M., SAKSENA, G., CIBULSKIS, K., TESAR, B., GABRIEL, S., HACOHEN, N., MEYERSON, M., LANDER, E. S., NEUBERG, D., BROWN, J. R., GETZ, G. and WU, C. J. (2013). Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152** 714–726.

LARSON, N. B. and FRIDLEY, B. L. (2013). PurBayes: Estimating tumor cellularity and subclonality in next-generation sequencing data. *Bioinformatics* **29** 1888–1889.

LEE, J., MÜLLER, P., GULUKOTA, K. and JI, Y. (2015). Supplement to "A Bayesian feature allocation model for tumor heterogeneity." DOI:10.1214/15-AOAS817SUPP.

LI, H. and DURBIN, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25** 1754–1760.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G., DURBIN, R. and 1000 GENOME PROJECT DATA PROCESSING SUBGROUP (2009). The sequence Alignment/Map format and SAMtools. *Bioinformatics* **25** 2078–2079.

MARUSYK, A. and POLYAK, K. (2010). Tumor heterogeneity: Causes and consequences. *Biochim. Biophys. Acta*. **1085** 1.

MCKENNA, A., HANNA, M., BANKS, E., SIVACHENKO, A., CIBULSKIS, K., KERNYTSKY, A., GARIMELLA, K., ALTSHULER, D., GABRIEL, S., DALY, M. and DEPRISTO, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. **20** 1297–1303.

NAVIN, N., KRASNITZ, A., RODGERS, L., COOK, K., METH, J., KENDALL, J., RIGGS, M., EBERLING, Y., TROGE, J., GRUBOR, V. et al. (2010). Inferring tumor progression from genomic heterogeneity. *Genome Res*. **20** 68–80.

NG, P. C. and KIRKNESS, E. F. (2010). Whole genome sequencing. In *Genetic Variation* 215–226. Springer, New York.

O'HAGAN, A. (1995). Fractional Bayes factors for model comparison. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **57** 99–138. MR1325379

ROTH, A., KHATTRA, J., YAP, D., WAN, A., LAKS, E., BIELE, J., HA, G., APARICIO, S., BOUCHARD-CÔTÉ, A. and SHAH, S. P. (2014). Pyclone: Statistical inference of clonal population structure in cancer. *Nature Methods* **11** 396–398.

RUSSNES, H. G., NAVIN, N., HICKS, J. and BORRESEN-DALE, A.-L. (2011). Insight into the heterogeneity of breast cancer through next-generation sequencing. *J. Clin. Invest*. **121** 3810–3818.

SU, X., ZHANG, L., ZHANG, J., MERIC-BERNSTAM, F. and WEINSTEIN, J. N. (2012). PurityEst: Estimating purity of human tumor samples using next-generation sequencing data. *Bioinformatics* **28** 2265–2266.

TEH, Y. W., GÖRÜR, D. and GHAHRAMANI, Z. (2007). Stick-breaking construction for the Indian buffet process. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, Vol. 11. The Society for Artificial Intelligence and Statistics, NJ.

WERSTO, R. P., LIBLIT, R. L., DEITCH, D. and KOSS, L. G. (1991). Variability in DNA measurements in multiple tumor samples of human colonic carcinoma. *Cancer* **67** 106–115.

WHEELER, D. A., SRINIVASAN, M., EGHOLM, M., SHEN, Y., CHEN, L., MCGUIRE, A., HE, W., CHEN, Y.-J., MAKHIJANI, V., ROTH, G. T. et al. (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452** 872–876.

The Annals of Applied Statistics
2015, Vol. 9, No. 2, 640–664
DOI: 10.1214/15-AOAS808

BAYESIAN GROUP LASSO FOR NONPARAMETRIC VARYING-COEFFICIENT MODELS WITH APPLICATION TO FUNCTIONAL GENOME-WIDE ASSOCIATION STUDIES

BY JIAHAN LI[*], ZHONG WANG[†,1], RUNZE LI[‡,2] AND RONGLING WU[‡,†,3]

University of Notre Dame[*], Beijing Forestry University[†] and Pennsylvania State University[‡]

Although genome-wide association studies (GWAS) have proven powerful for comprehending the genetic architecture of complex traits, they are challenged by a high dimension of single-nucleotide polymorphisms (SNPs) as predictors, the presence of complex environmental factors, and longitudinal or functional natures of many complex traits or diseases. To address these challenges, we propose a high-dimensional varying-coefficient model for incorporating functional aspects of phenotypic traits into GWAS to formulate a so-called functional GWAS or *f*GWAS. The Bayesian group lasso and the associated MCMC algorithms are developed to identify significant SNPs and estimate how they affect longitudinal traits through time-varying genetic actions. The model is generalized to analyze the genetic control of complex traits using subject-specific sparse longitudinal data. The statistical properties of the new model are investigated through simulation studies. We use the new model to analyze a real GWAS data set from the Framingham Heart Study, leading to the identification of several significant SNPs associated with age-specific changes of body mass index. The *f*GWAS model, equipped with the Bayesian group lasso, will provide a useful tool for genetic and developmental analysis of complex traits or diseases.

REFERENCES

CHO, S., KIM, H., OH, S., KIM, K. and PARK, T. (2009). Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis. *BMC Proc.* **3 Suppl 7** S25.

CUI, Y., WU, R., CASELLA, G. and ZHU, J. (2008). Nonparametric functional mapping of quantitative trait loci underlying programmed cell death. *Stat. Appl. Genet. Mol. Biol.* **7** Art. 4, 32. MR2386321

DALY, A. K. (2010). Genome-wide association studies in pharmacogenomics. *Nat. Rev. Genet.* **11** 241–246.

DAS, K., LI, J., WANG, Z., FU, G., LI, Y., MAUGER, D., LI, R. and WU, R. (2011). A dynamic model for genome-wide association studies. *Hum. Genet.* **129** 629–639.

DAWBER, T. R., MEADORS, G. F. and MOORE, F. E. JR. (1951). Epidemiological approaches to heart disease: The framingham study. *Am. J. Publ. Health* **41** 279–286.

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

FILIAULT, D. L. and MALOOF, J. N. (2012). A genome-wide association study identifies variants underlying the arabidopsis thaliana shade avoidance response. *PLoS Genet.* **8** e1002589.

Key words and phrases. Bayesian approach, group variable selection, longitudinal data, GWAS.

FRAYLING, T. M. (2007). Genome-wide association studies provide new insights into type 2 diabetes aetiology. *Nat. Rev. Genet.* **8** 657–662.

FRAYLING, T. M., TIMPSON, N. J., WEEDON, M. N., ZEGGINI, E., FREATHY, R. M., LINDGREN, C. M., PERRY, J. R. B., ELLIOTT, K. S., LANGO, H., RAYNER, N. W., SHIELDS, B., HARRIES, L. W., BARRETT, J. C., ELLARD, S., GROVES, C. J., KNIGHT, B., PATCH, A.-M., NESS, A. R., EBRAHIM, S., LAWLOR, D. A., RING, S. M., BEN-SHLOMO, Y., JARVELIN, M.-R., SOVIO, U., BENNETT, A. J., MELZER, D., FERRUCCI, L., LOOS, R. J. F., BARROSO, I., WAREHAM, N. J., KARPE, F., OWEN, K. R., CARDON, L. R., WALKER, M., HITMAN, G. A., PALMER, C. N. A., DONEY, A. S. F., MORRIS, A. D., SMITH, G. D., HATTERSLEY, A. T. and MCCARTHY, M. I. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316** 889–894.

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

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

GORLOVA, O. Y., AMOS, C. I., WANG, N. W., SHETE, S., TURNER, S. T. and BOERWINKLE, E. (2003). Genetic linkage and imprinting effects on body mass index in children and young adults. *European Journal of Human Genetics* **11** 425–432.

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

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

JAQUISH, C. E. (2007). The framingham heart study, on its way to becoming the gold standard for cardiovascular genetic epidemiology? *BMC Med. Genet.* **8** 63.

JOOD, K., JERN, C., WILHELMSEN, L. and ROSENGREN, A. (2004). Body mass index in mid-life is associated with a first stroke in men: A prospective population study over 28 years. *Stroke* **35** 2764–2769.

LETTRE, G. (2011). Recent progress in the study of the genetics of height. *Human Genetics* **129** 465–472.

LI, J., DAS, K., FU, G., LI, R. and WU, R. (2012). Bayesian lasso for genome-wide association studies. *Bioinformatics* **27** 516–523.

LI, J., WANG, Z., LI, R. and WU, R. (2015). Supplement to "Bayesian group Lasso for nonparametric varying-coefficient models with application to functional genome-wide association studies." DOI:10.1214/15-AOAS808SUPP.

LIN, M. and WU, R. (2006). A joint model for nonparametric functional mapping of longitudinal trajectory and time-to-event. *BMC Bioinformatics* **7** 138.

LIN, Y. and ZHANG, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.* **34** 2272–2297. MR2291500

LYNCH, M. and WALSH, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer, Sunderland, MA.

MA, C. X., CASELLA, G. and WU, R. L. (2002). Functional mapping of quantitative trait loci underlying the character process: A theoretical framework. *Genetics* **161** 1751–1762.

MICHEL, S., LIANG, L., DEPNER, M., KLOPP, N., RUETHER, A., KUMAR, A., SCHEDEL, M., VOGELBERG, C., VON MUTIUS, E., VON BERG, A., BUFE, A., RIETSCHEL, E., HEINZMANN, A., LAUB, O., SIMMA, B., FRISCHER, T., GENUNEIT, J., GUT, I. G., SCHREIBER, S., LATHROP, M., ILLIG, T. and KABESCH, M. (2010). Unifying candidate gene and GWAS approaches in asthma. *PLoS ONE* **5** e13894.

MORGAN, A. R., THOMPSON, J. M., MURPHY, R., BLACK, P. N., LAM, W. J., FERGUSON, L. R. and MITCHELL, E. A. (2010). Obesity and diabetes genes are associated with being born small for gestational age: Results from the auckland birthweight collaborative study. *BMC Medical Genetics* **11** 125.

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

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

SANDHU, M. S., WEEDON, M. N., FAWCETT, K. A., WASSON, J., DEBENHAM, S. L., DALY, A., LANGO, H., FRAYLING, T. M., NEUMANN, R. J., SHERVA, R., BLECH, I., PHAROAH, P. D., PALMER, C. N. A., KIMBER, C., TAVENDALE, R., MORRIS, A. D., MCCARTHY, M. I., WALKER, M., HITMAN, G., GLASER, B., PERMUTT, M. A., HATTERSLEY, A. T., WAREHAM, N. J. and BARROSO, I. (2007). Common variants in WFS1 confer risk of type 2 diabetes. *Nat. Genet.* **39** 951–953.

SCOTT, L. J., MOHLKE, K. L., BONNYCASTLE, L. L., WILLER, C. J., LI, Y., DUREN, W. L., ERDOS, M. R., STRINGHAM, H. M., CHINES, P. S., JACKSON, A. U., PROKUNINA-OLSSON, L., DING, C.-J., SWIFT, A. J., NARISU, N., HU, T., PRUIM, R., XIAO, R., LI, X.-Y., CONNEELY, K. N., RIEBOW, N. L., SPRAU, A. G., TONG, M., WHITE, P. P., HETRICK, K. N., BARNHART, M. W., BARK, C. W., GOLDSTEIN, J. L., WATKINS, L., XIANG, F., SARAMIES, J., BUCHANAN, T. A., WATANABE, R. M., VALLE, T. T., KINNUNEN, L., ABECASIS, G. R., PUGH, E. W., DOHENY, K. F., BERGMAN, R. N., TUOMILEHTO, J., COLLINS, F. S. and BOEHNKE, M. (2007). A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316** 1341–1345.

SHULDINER, A. R. et al. (2009). Association of cytochrome P450 2C19 genotype with the antiplatelet effect and clinical efficacy of clopidogrel therapy. *J. Am. Med. Assoc.* **302** 849–857.

STEINTHORSDOTTIR, V., THORLEIFSSON, G., REYNISDOTTIR, I., BENEDIKTSSON, R., JONSDOTTIR, T., WALTERS, G. B., STYRKARSDOTTIR, U., GRETARSDOTTIR, S., EMILSSON, V., GHOSH, S., BAKER, A., SNORRADOTTIR, S., BJARNASON, H., NG, M. C. Y., HANSEN, T., BAGGER, Y., WILENSKY, R. L., REILLY, M. P., ADEYEMO, A., CHEN, Y., ZHOU, J., GUDNASON, V., CHEN, G., HUANG, H., LASHLEY, K., DOUMATEY, A., SO, W.-Y., MA, R. C. Y., ANDERSEN, G., BORCH-JOHNSEN, K., JORGENSEN, T., VAN VLIET-OSTAPTCHOUK, J. V., HOFKER, M. H., WIJMENGA, C., CHRISTIANSEN, C., RADER, D. J., ROTIMI, C., GURNEY, M., CHAN, J. C. N., PEDERSEN, O., SIGURDSSON, G., GULCHER, J. R., THORSTEINSDOTTIR, U., KONG, A. and STEFANSSON, K. (2007). A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nat. Genet.* **39** 770–775.

SUCHOCKI, T. and SZYDA, J. (2011). Statistical modelling of growth using a mixed model with orthogonal polynomials. *J. Appl. Genet.* **52** 95–100.

TAKEUCHI, F., MCGINNIS, R., BOURGEOIS, S., BARNES, C., ERIKSSON, N., SORANZO, N., WHITTAKER, P., RANGANATH, V., KUMANDURI, V., MCLAREN, W., HOLM, L., LINDH, J., RANE, A., WADELIUS, M. and DELOUKAS, P. (2009). A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. *PLoS Genet.* **5** e1000433.

TEICHERT, M., EIJGELSHEIM, M., RIVADENEIRA, F., UITTERLINDEN, A. G., VAN SCHAIK, R. H. N., HOFMAN, A., SMET, P. A. G. M. D., VAN GELDER, T., VISSER, L. E. and STRICKER, B. H. C. (2009). A genome-wide association study of acenocoumarol maintenance dosage. *Hum. Mol. Genet.* **18** 3758–3768.

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

VIDAL-PUIG, A. J., CONSIDINE, R. V., JIMENEZ-LIÑAN, M., WERMAN, A., PORIES, W. J., CARO, J. F. and FLIER, J. S. (1997). Peroxisome proliferator-activated receptor gene expression in human tissues. Effects of obesity, weight loss, and regulation by insulin and glucocorticoids. *J. Clin. Invest.* **99** 2416–2422.

WANG, L., LI, H. and HUANG, J. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *J. Amer. Statist. Assoc.* **103** 1556–1569. MR2504204

WANG, Z., LI, Y., LI, Q. and WU, R. (2009). Joint functional mapping of quantitative trait loci for HIV-1 and CD4$^+$ dynamics. *Int. J. Biostat.* **5** Art. 9, 26. MR2491436

WU, R. and LIN, M. (2006). Functional mapping—How to map and study the genetic architecture of dynamic complex traits. *Nature Review Genetics* **7** 229–237.

WU, R., MA, C.-X., LIN, M., WANG, Z. and CASELLA, G. (2004). Functional mapping of quantitative trait loci underlying growth trajectories using a transform-both-sides logistic model. *Biometrics* **60** 729–738. MR2089449

WU, T. T., CHEN, Y. F., HASTIE, T., SOBEL, E. and LANGE, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* **25** 714–721.

XU, Z. and TAYLOR, J. A. (2009). SNPinfo: Integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. *Nucleic Acids Res.* **37(suppl 2)** W600–W605.

YANG, R. and XU, S. (2007). Bayesian shrinkage analysis of quantitative trait loci for dynamic traits. *Genetics* **176** 1169–1185.

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

YAP, J. S., FAN, J. and WU, R. (2009). Nonparametric modeling of longitudinal covariance structure in functional mappings of quantitative trait loci. *Biometrics* **65** 1068–1077. MR2756494

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, H. H. and LIN, Y. (2006). Component selection and smoothing for nonparametric regression in exponential families. *Statist. Sinica* **16** 1021–1041. MR2281313

ZHAO, W., CHEN, Y. Q., CASELLA, G., CHEVERUD, J. M. and WU, R. L. (2005). A nonstationary model for functional mapping of complex traits. *Bioinformatics* **21** 2469–2477.

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

The Annals of Applied Statistics
2015, Vol. 9, No. 2, 665–686
DOI: 10.1214/14-AOAS776
© Institute of Mathematical Statistics, 2015

WAVELET-BASED GENETIC ASSOCIATION ANALYSIS OF FUNCTIONAL PHENOTYPES ARISING FROM HIGH-THROUGHPUT SEQUENCING ASSAYS[1]

BY HEEJUNG SHIM AND MATTHEW STEPHENS

University of Chicago

Understanding how genetic variants influence cellular-level processes is an important step toward understanding how they influence important organismal-level traits, or "phenotypes," including human disease susceptibility. To this end, scientists are undertaking large-scale genetic association studies that aim to identify genetic variants associated with molecular and cellular phenotypes, such as gene expression, transcription factor binding, or chromatin accessibility. These studies use high-throughput sequencing assays (e.g., RNA-seq, ChIP-seq, DNase-seq) to obtain high-resolution data on how the traits vary along the genome in each sample. However, typical association analyses fail to exploit these high-resolution measurements, instead aggregating the data at coarser resolutions, such as genes, or windows of fixed length. Here we develop and apply statistical methods that better exploit the high-resolution data. The key idea is to treat the sequence data as measuring an underlying "function" that varies along the genome, and then, building on wavelet-based methods for functional data analysis, test for association between genetic variants and the underlying function. Applying these methods to identify genetic variants associated with chromatin accessibility (dsQTLs), we find that they identify substantially more associations than a simpler window-based analysis, and in total we identify 772 novel dsQTLs not identified by the original analysis.

REFERENCES

ABRAMOVICH, F. and ANGELINI, C. (2006). Testing in mixed-effects FANOVA models. *J. Statist. Plann. Inference* **136** 4326–4348. MR2323419

ANTONIADIS, A. and SAPATINAS, T. (2007). Estimation and inference in functional mixed-effects models. *Comput. Statist. Data Anal.* **51** 4793–4813. MR2364541

BARSKI, A., CUDDAPAH, S., CUI, K., ROH, T.-Y., SCHONES, D. E., WANG, Z., WEI, G., CHEPELEV, I. and ZHAO, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell* **129** 823–837.

BENJAMINI, Y. and SPEED, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* **40** e72.

BESAG, J. and CLIFFORD, P. (1991). Sequential Monte Carlo p-values. *Biometrika* **78** 301–304. MR1131163

Key words and phrases. Wavelets, high-throughput sequencing assays, RNA-seq, DNase-seq, chromatin accessibility, ChIP-seq, genetic association analysis, hierarchical model, Bayesian inference, functional data.

BOYLE, A. P., DAVIS, S., SHULHA, H. P., MELTZER, P., MARGULIES, E. H., WENG, Z., FUREY, T. S. and CRAWFORD, G. E. (2008). High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132** 311–322.

CHEUNG, V. G., NAYAK, R. R., WANG, I. X., ELWYN, S., COUSINS, S. M., MORLEY, M. and SPIELMAN, R. S. (2010). Polymorphic cis- and trans-regulation of human gene expression. *PLoS Biol*. **8** e1000480.

CLEMENT, L., DE BEUF, K., THAS, O., VUYLSTEKE, M., IRIZARRY, R. A. and CRAINICEANU, C. M. (2012). Fast wavelet based functional models for transcriptome analysis with tiling arrays. *Stat*. *Appl*. *Genet*. *Mol*. *Biol*. **11** Art. 4, 38. MR2924207

CROUSE, M. S., NOWAK, R. D. and BARANIUK, R. G. (1998). Wavelet-based statistical signal processing using hidden Markov models. *IEEE Trans*. *Signal Process*. **46** 886–902. MR1665651

DABNEY, A., STOREY, J. D. and WARNES, G. R. (2015). qvalue: Q-value estimation for false discovery rate control. R package version 1.30.0.

DAY, N., HEMMAPLARDH, A., THURMAN, R. E., STAMATOYANNOPOULOS, J. A. and NOBLE, W. S. (2007). Unsupervised segmentation of continuous genomic data. *Bioinformatics* **23** 1424–1426.

DEGNER, J. F., PAI, A. A., PIQUE-REGI, R., VEYRIERAS, J.-B., GAFFNEY, D. J., PICKRELL, J. K., DE LEON, S., MICHELINI, K., LEWELLEN, N., CRAWFORD, G. E., STEPHENS, M., GILAD, Y. and PRITCHARD, J. K. (2012). DNaseI sensitivity QTLs are a major determinant of human expression variation. *Nature* **482** 390–394.

DONOHO, D. L. and JOHNSTONE, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J*. *Amer*. *Statist*. *Assoc*. **90** 1200–1224. MR1379464

FAN, J. and LIN, S.-K. (1998). Test of significance when data are curves. *J*. *Amer*. *Statist*. *Assoc*. **93** 1007–1021. MR1649196

FRAZEE, A. C., SABUNCIYAN, S., HANSEN, K. D., IRIZARRY, R. A. and LEEK, J. T. (2014). Differential expression analysis of RNA-seq data at single-base resolution. *Biostatistics* **15** 413–426.

FRYZLEWICZ, P. and NASON, G. P. (2004). A Haar–Fisz algorithm for Poisson intensity estimation. *J*. *Comput*. *Graph*. *Statist*. **13** 621–638. MR2087718

HESSELBERTH, J. R., CHEN, X., ZHANG, Z., SABO, P. J., SANDSTROM, R., REYNOLDS, A. P., THURMAN, R. E., NEPH, S., KUEHN, M. S., NOBLE, W. S., FIELDS, S. and STAMATOYANNOPOULOS, J. A. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature Methods* **6** 283–289.

JACKMAN, S. (2009). *Bayesian Analysis for the Social Sciences*. Wiley, Chichester. MR2584520

JOHNSON, D. S., MORTAZAVI, A., MYERS, R. M. and WOLD, B. (2007). Genome-wide mapping of in vivo protein–DNA interactions. *Science* **316** 1497–1502.

KARCZEWSKI, K. J., DUDLEY, J. T., KUKURBA, K. R., CHEN, R., BUTTE, A. J., MONTGOMERY, S. B. and SNYDER, M. (2013). Systematic functional regulatory assessment of disease-associated variants. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **110** 9607–9612.

KASOWSKI, M., GRUBERT, F., HEFFELFINGER, C., HARIHARAN, M., ASABERE, A., WASZAK, S. M., HABEGGER, L., ROZOWSKY, J., SHI, M., URBAN, A. E., HONG, M.-Y., KARCZEWSKI, K. J., HUBER, W., WEISSMAN, S. M., GERSTEIN, M. B., KORBEL, J. O. and SNYDER, M. (2010). Variation in transcription factor binding among humans. *Science* **328** 232–235.

KOLACZYK, E. D. (1999). Bayesian multiscale models for Poisson processes. *J*. *Amer*. *Statist*. *Assoc*. **94** 920–933. MR1723303

LEEK, J. T. and STOREY, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*. **3** 1724–1735.

MALLAT, S. G. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans*. *Pattern Anal*. *Mach*. *Intell*. **11** 674–693.

MANGRAVITE, L. M., ENGELHARDT, B. E., MEDINA, M. W., SMITH, J. D., BROWN, C. D., CHASMAN, D. I., MECHAM, B. H., HOWIE, B., SHIM, H., NAIDOO, D., FENG, Q., RIEDER, M. J., CHEN, Y.-D. I., ROTTER, J. I., RIDKER, P. M., HOPEWELL, J. C., PARISH, S., ARMITAGE, J., COLLINS, R., WILKE, R. A., NICKERSON, D. A., STEPHENS, M. and KRAUSS, R. M. (2013). A statin-dependent QTL for GATM expression is associated with statin-induced myopathy. *Nature* **502** 377–380.

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

MIKKELSEN, T. S., KU, M., JAFFE, D. B., ISSAC, B., LIEBERMAN, E., GIANNOUKOS, G., ALVAREZ, P., BROCKMAN, W., KIM, T.-K., KOCHE, R. P., LEE, W., MENDENHALL, E., O'DONOVAN, A., PRESSER, A., RUSS, C., XIE, X., MEISSNER, A., WERNIG, M., JAENISCH, R., NUSBAUM, C., LANDER, E. S. and BERNSTEIN, B. E. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448** 553–560.

MITRA, A. and SONG, J. (2012). WaveSeq: A novel data-driven method of detecting histone modification enrichments using wavelets. *PLoS ONE* **7** e45486.

MONTGOMERY, S. B., SAMMETH, M., GUTIERREZ-ARCELUS, M., LACH, R. P., INGLE, C., NISBETT, J., GUIGO, R. and DERMITZAKIS, E. T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464** 773–777.

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

MORRIS, J. S., BROWN, P. J., HERRICK, R. C., BAGGERLY, K. A. and COOMBES, K. R. (2008). Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models. *Biometrics* **64** 479–489. MR2432418

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

NICOLAE, D. L., GAMAZON, E., ZHANG, W., DUAN, S., DOLAN, M. E. and COX, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS. *PLoS Genet*. **6** e1000888.

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

PIQUE-REGI, R., DEGNER, J. F., PAI, A. A., BOYLE, A. P., SONG, L., LEE, B.-K., GAFFNEY, D. J., GILAD, Y. and PRITCHARD, J. K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res*. **21** 447–455.

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

SHIM, H. and STEPHENS, M. (2015). Supplement to "Wavelet-based genetic association analysis of functional phenotypes arising from high-throughput sequencing assays." DOI:10.1214/14-AOAS776SUPP.

SPENCER, C. C. A., DELOUKAS, P., HUNT, S., MULLIKIN, J., MYERS, S., SILVERMAN, B., DONNELLY, P., BENTLEY, D. and MCVEAN, G. (2006). The influence of recombination on human genetic diversity. *PLoS Genet*. **2** e148.

STEGLE, O., PARTS, L., DURBIN, R. and WINN, J. (2010). A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput. Biol*. **6** e1000770. MR2659818

TESLOVICH, T. M., MUSUNURU, K., SMITH, A. V., EDMONDSON, A. C., STYLIANOU, I. M., KOSEKI, M., PIRRUCCELLO, J. P., RIPATTI, S., CHASMAN, D. I., WILLER, C. J., JOHANSEN, C. T., FOUCHIER, S. W., ISAACS, A., PELOSO, G. M., BARBALIC, M., RICK-

ETTS, S. L. et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466** 707–713.

TIMMERMANN, K. E. and NOWAK, R. D. (1999). Multiscale modeling and estimation of Poisson processes with application to photon-limited imaging. *IEEE Trans. Inform. Theory* **45** 846–862. MR1682515

VAN DER WAERDEN, B. L. (1953). Order tests for the two-sample problem. II, III. *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen, Serie A* **564** 303–310, 311–316.

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

WTCCC (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447** 661–678.

WU, S., WANG, J., ZHAO, W., POUNDS, S. and CHENG, C. (2010). ChIP-PaM: An algorithm to identify protein-DNA interaction using ChIP-seq data. *Theor. Biol. Med. Model* **7** 18.

YANG, X. and NIE, K. (2008). Hypothesis testing in functional linear regression models with Neyman's truncation and wavelet thresholding for longitudinal data. *Stat. Med.* **27** 845–863. MR2420116

ZHANG, Y., SHIN, H., SONG, J. S., LEI, Y. and LIU, X. S. (2008). Identifying positioned nucleosomes with epigenetic marks in human from ChIP-seq. *BMC Genomics* **9** 537.

ZHAO, W. and WU, R. (2008). Wavelet-based nonparametric functional mapping of longitudinal curves. *J. Amer. Statist. Assoc.* **103** 714–725. MR2524004

ZHU, H., BROWN, P. J. and MORRIS, J. S. (2011). Robust, adaptive functional regression in functional mixed model framework. *J. Amer. Statist. Assoc.* **106** 1167–1179. MR2894772

The Annals of Applied Statistics
2015, Vol. 9, No. 2, 687–713
DOI: 10.1214/15-AOAS818

SPATIAL BAYESIAN VARIABLE SELECTION AND GROUPING FOR HIGH-DIMENSIONAL SCALAR-ON-IMAGE REGRESSION

BY FAN LI[*,1,4], TINGTING ZHANG[†,2,4], QUANLI WANG[*], MARLEN Z. GONZALEZ[†], ERIN L. MARESH[†] AND JAMES A. COAN[3,†]

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

Multi-subject functional magnetic resonance imaging (fMRI) data has been increasingly used to study the population-wide relationship between human brain activity and individual biological or behavioral traits. A common method is to regress the scalar individual response on imaging predictors, known as a scalar-on-image (SI) regression. Analysis and computation of such massive and noisy data with complex spatio-temporal correlation structure is challenging. In this article, motivated by a psychological study on human affective feelings using fMRI, we propose a joint Ising and Dirichlet Process (Ising-DP) prior within the framework of Bayesian stochastic search variable selection for selecting brain voxels in high-dimensional SI regressions. The Ising component of the prior makes use of the spatial information between voxels, and the DP component groups the coefficients of the large number of voxels to a small set of values and thus greatly reduces the posterior computational burden. To address the phase transition phenomenon of the Ising prior, we propose a new analytic approach to derive bounds for the hyperparameters, illustrated on 2- and 3-dimensional lattices. The proposed method is compared with several alternative methods via simulations, and is applied to the fMRI data collected from the KLIFF hand-holding experiment.

REFERENCES

ALLEN, J. P., PORTER, M., MCFARLAND, F. C., MCELHANEY, K. B. and MARSH, P. (2007). The relation of attachment security to adolescents' paternal and peer relationships, depression, and externalizing behavior. *Child Development* **78** 1222–1239.

ANTONIAK, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2** 1152–1174. MR0365969

BECKES, L. and COAN, J. A. (2011). Social baseline theory: The role of social proximity in emotion and economy of action. *Social and Personality Psychology Compass* **5** 976–988.

BOWMAN, F. D. (2007). Spatiotemporal models for region of interest analyses of functional neuroimaging data. *J. Amer. Statist. Assoc.* **102** 442–453. MR2370845

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

BRADLEY, M. M. and LANG, P. J. (1994). Measuring emotion: The self-assessment mankin and the semantic differential. *J. Behav. Ther. Exp. Psychiatry* **25** 49–59.

COAN, J. A. (2010). Adult attachment and the brain. *J. Soc. Pers. Relatsh.* **27** 210–217.

COAN, J. A. (2011). The social regulation of emotion. In *Oxford Handbook of Social Neuroscience* 614–623. Oxford Univ. Press, New York.

Key words and phrases. Bayesian, Dirichlet Process, fMRI, Ising model, phase transition, scalar-on-image regression, stochastic search, variable selection.

COAN, J. A., BECKES, L. and ALLEN, J. P. (2013). Childhood maternal support and social capital moderate the regulatory impact of social relationships in adulthood. *Int. J. Psychophysiol.* **88** 224–231.

COAN, J. A. and MARESH, E. L. (2014). Social baseline theory and the social regulation of emotion. In *The Handbook of Emotion Regulation*, 2nd ed. (J. Gross, ed.) 221–236. The Guilford Press, New York.

COAN, J. A., SCHAEFER, H. S. and DAVIDSON, R. J. (2006). Lending a hand: Social regulation of the neural response to threat. *Psychol. Sci.* **17** 1032–1039.

CRAIG, A. D. (2009). How do you fell now? The anterior insula and human awareness. *Nat. Rev. Neurosci.* **10** 59–70.

CRITCHLEY, H. D., CORFIELD, D. R., CHANDLER, M. P., MATHIAS, C. J. and DOLAN, R. J. (2000). Cerebral correlates of autonomic cardiovascular arousal: A functional neuroimaging investigation in humans. *J. Physiol.* (*Lond.*) **523** 259–270.

DERADO, G., BOWMAN, F. D. and KILTS, C. D. (2010). Modeling the spatial and temporal dependence in fMRI data. *Biometrics* **66** 949–957. MR2758231

DUNSON, D. B., HERRING, A. H. and ENGEL, S. M. (2008). Bayesian selection and clustering of polymorphisms in functionally related genes. *J. Amer. Statist. Assoc.* **103** 534–546. MR2523991

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

FERGUSON, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2** 615–629. MR0438568

FRISTON, K. J., HOLMES, A. P., WORSLEY, K., POLINE, P. J., FRITH, C. and FRACKOWIAK, R. (1995). Statistical parametric maps in functional imaging: A general linear approach. *Hum. Brain Mapp.* **2** 189–210.

GE, T., MÜLLER-LENKE, N., BENDFELDT, K., NICHOLS, T. E. and JOHNSON, T. D. (2014). Analysis of multiple sclerosis lesions via spatially varying coefficients. *Ann. Appl. Stat.* **8** 1095–1118. MR3262547

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

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

GEORGE, E. and MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7** 339–373.

GOLDSMITH, J., HUANG, L. and CRAINICEANU, C. M. (2014). Smooth scalar-on-image regression via spatial Bayesian variable selection. *J. Comput. Graph. Statist.* **23** 46–64. MR3173760

GÖSSL, C., AUER, D. P. and FAHRMEIR, L. (2001). Bayesian spatiotemporal inference in functional magnetic resonance imaging. *Biometrics* **57** 554–562. MR1855691

HUANG, L., GOLDSMITH, J., REISS, P. T., REICH, D. S. and CRAINICEANU, C. M. (2013). Bayesian scalar-on-image regression with application to association between intracranial DTI and cognitive outcomes. *NeuroImage* **83** 210–223.

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

ISHWARAN, H. and ZAREPOUR, M. (2000). Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika* **87** 371–390. MR1782485

JENKINSON, M., BANNISTER, P., BRADY, M. and SMITH, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage* **17** 825–841.

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

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

KANG, J., JOHNSON, T. D., NICHOLS, T. E. and WAGER, T. D. (2011). Meta analysis of functional neuroimaging data via Bayesian spatial point processes. *J. Amer. Statist. Assoc.* **106** 124–134. MR2816707

KIM, S., TADESSE, M. G. and VANNUCCI, M. (2006). Variable selection in clustering via Dirichlet process mixture models. *Biometrika* **93** 877–893. MR2285077

LANG, P. J., GREENWALD, M. K., BRADLEY, M. M. and HAMM, A. O. (1993). Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology* **30** 261–273.

LANGE, K. (2008). *Optimization. Springer Texts in Statistics* **95**. Springer, New York.

LEWIS, P. A., CRITCHLEY, H. D., ROTSHTEIN, P. and DOLAN, R. J. (2007). Neural correlates of processing valence and arousal in affective words. *Cereb. Cortex* **17** 742–748.

LI, F. and ZHANG, N. R. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *J. Amer. Statist. Assoc.* **105** 1202–1214. MR2752615

LI, F., ZHANG, T., WANG, Q., GONZALEZ, M., MARESH, E. L. and COAN, J. A. (2015). Supplement to “Spatial Bayesian variable selection and grouping for high-dimensional scalar-on-image regression.” DOI:10.1214/15-AOAS818SUPP.

MARESH, E. L., BECKES, L. and COAN, J. A. (2013). The social regulation of threat-related attentional disengagement in highly anxious individuals. *Front. Human Neurosci.* **7** 515.

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

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

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

RAFTERY, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika* **83** 251–266. MR1439782

REISS, P. T., MENNES, M., PETKOVA, E., HUANG, L., HOPTMAN, M. J., BISWAL, B. B., COLCOMBE, S. J., ZUO, X.-N. and MILHAM, M. P. (2011). Extracting information from functional connectivity maps via function-on-scalar regression. *NeuroImage* **56** 140–148.

REISS, P. T., HUO, L., ZHAO, Y., KELLY, C. and OGDEN, R. T. (2015). Wavelet-domain regression and predictive inference in psychiatric neuroimaging. *Ann. Appl. Stat.* **9** 1076–1101.

RUSSELL, J. (1980). A circumplex model of affect. *J. Pers. Soc. Psychol.* **39** 1161–1178.

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

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

SMITH, M. and KOHN, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics* **75** 317–343.

SMITH, M., PÜTZ, B., AUER, D. and FAHRMEIR, L. (2003). Assessing brain activity through spatial Bayesian variable selection. *NeuroImage* **20** 802–815.

SMITH, S. M., JENKINSON, M., WOOLRICH, M. W., BECKMANN, C. F., BEHRENS, T. E. J., JOHANSEN-BERG, H., BANNISTER, P. R., DE LUCA, M., DROBNJAK, I., FLITNEY, D. E., NIAZY, R., SAUNDERS, J., VICKERS, J., ZHANG, Y., DE STEFANO, N., BRADY, J. M. and MATTHEWS, P. M. (2004). In advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage* **23(S1)** 208–219.

STANLEY, H. E. (1987). *Introduction to Phase Transitions and Critical Phenomena*. Oxford Univ. Press, New York.

STINGO, F. C., CHEN, Y. A., TADESSE, M. G. and VANNUCCI, M. (2011). Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes. *Ann. Appl. Stat.* **5** 1978–2002. MR2884929

SUCHARD, M. A., WANG, Q., CHAN, C., FRELINGER, J., CRON, A. and WEST, M. (2010). Understanding GPU programming for statistical computation: Studies in massively parallel massive mixtures. *J. Comput. Graph. Statist.* **19** 419–438. MR2758309

TADESSE, M. G., SHA, N. and VANNUCCI, M. (2005). Bayesian variable selection in clustering high-dimensional data. *J. Amer. Statist. Assoc.* **100** 602–617. MR2160563

VANNUCCI, M. and STINGO, F. C. (2011). Bayesian models for variable selection that incorporate biological information. In *Bayesian Statistics* 9 (J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith and M. West, eds.) 659–678. Oxford Univ. Press, Oxford. MR3204022

WEST, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. In *Bayesian Statistics* 7 (*Tenerife*, 2002) (J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, eds.) 733–742. Oxford Univ. Press, New York. MR2003537

WIECH, K., PLONER, M. and TRACEY, I. (2008). Neurocognitive aspects of pain perception. *Trends Cogn. Sci.* **12** 306–313.

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

YUE, Y. R., LINDQUIST, M. A. and LOH, J. M. (2012). Meta-analysis of functional neuroimaging data using Bayesian nonparametric binary regression. *Ann. Appl. Stat.* **6** 697–718. MR2976488

ZHANG, T., LI, F., BECKES, L. and COAN, J. A. (2013). A semi-parametric model of the hemodynamic response for multi-subject fMRI data. *NeuroImage* **75** 136–145.

The Annals of Applied Statistics
2015, Vol. 9, No. 2, 714–730
DOI: 10.1214/15-AOAS810

SEMIPARAMETRIC TIME TO EVENT MODELS IN THE PRESENCE OF ERROR-PRONE, SELF-REPORTED OUTCOMES—WITH APPLICATION TO THE WOMEN'S HEALTH INITIATIVE[1]

BY XIANGDONG GU*, YUNSHENG MA† AND RAJI BALASUBRAMANIAN*,[2]

University of Massachusetts Amherst and*
University of Massachusetts Medical School†

The onset of several silent, chronic diseases such as diabetes can be detected only through diagnostic tests. Due to cost considerations, self-reported outcomes are routinely collected in lieu of expensive diagnostic tests in large-scale prospective investigations such as the Women's Health Initiative. However, self-reported outcomes are subject to imperfect sensitivity and specificity. Using a semiparametric likelihood-based approach, we present time to event models to estimate the association of one or more covariates with a error-prone, self-reported outcome. We present simulation studies to assess the effect of error in self-reported outcomes with regard to bias in the estimation of the regression parameter of interest. We apply the proposed methods to prospective data from 152,830 women enrolled in the Women's Health Initiative to evaluate the effect of statin use with the risk of incident diabetes mellitus among postmenopausal women. The current analysis is based on follow-up through 2010, with a median duration of follow-up of 12.1 years. The methods proposed in this paper are readily implemented using our freely available R software package *icensmis*, which is available at the Comprehensive R Archive Network (CRAN) website.

REFERENCES

ANDERSON, G., CUMMINGS, S., FREEDMAN, L. S., FURBERG, C., HENDERSON, M., JOHNSON, S. R., KULLER, L., MANSON, J., OBERMAN, A., PRENTICE, R. L., ROSSOUW, J. E. and GRP, W. H. I. S. (1998). Design of the women's health initiative clinical trial and observational study. *Control. Clin. Trials* **19** 61–109.

BALASUBRAMANIAN, R. and LAGAKOS, S. W. (2001). Estimation of the timing of perinatal transmission of HIV. *Biometrics* **57** 1048–1058. MR1950420

BALASUBRAMANIAN, R. and LAGAKOS, S. W. (2003). Estimation of a failure time distribution based on imperfect diagnostic tests. *Biometrika* **90** 171–182. MR1966558

CHEN, H. H., DUFFY, S. W. and TABAR, L. (1996). A Markov chain method to estimate the tumour progression rate from preclinical to clinical phase, sensitivity and positive predictive value for mammography in breast cancer screening. *Statistician* **45** 307–317.

COOK, T. D. (2000). Adjusting survival analysis for the presence of unadjudicated study events. *Control. Clin. Trials* **21** 208–222.

COOK, T. D. and KOSOROK, M. R. (2004). Analysis of time-to-event data with incomplete event adjudication. *J. Amer. Statist. Assoc.* **99** 1140–1152. MR2109502

Key words and phrases. Measurement error, panel data, interval censoring, time to event outcomes.

COX, D. R. and HINKLEY, D. V. (1979). *Theoretical Statistics*. Chapman & Hall, London.

CULVER, A. L., OCKENE, I. S., BALASUBRAMANIAN, R., OLENDZKI, B. C., SEPAVICH, D. M., WACTAWSKI-WENDE, J., MANSON, J. E., QIAO, Y. X., LIU, S. M., MERRIAM, P. A., RAHILLY-TIERNY, C., THOMAS, F., BERGER, J. S., OCKENE, J. K., CURB, J. D. and MA, Y. S. (2012). Statin use and risk of diabetes mellitus in postmenopausal women in the women's health initiative. *Arch. Intern. Med.* **172** 144–152.

FINKELSTEIN, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* **42** 845–854. MR0872963

GARCÍA-ZATTERA, M. J., JARA, A., LESAFFRE, E. and MARSHALL, G. (2012). Modeling of multivariate monotone disease processes in the presence of misclassification. *J. Amer. Statist. Assoc.* **107** 976–989. MR3010884

GU, X. and BALASUBRAMANIAN, R. (2013). icensmis: Study Design and Data Analysis in the presence of error-prone diagnostic tests and self-reported outcomes. R package version 1.1.

GU, X., MA, Y. and BALASUBRAMANIAN, R. (2015). Supplement to "Semiparametric time to event models in the presence of error-prone, self-reported outcomes—With application to the women's health initiative." DOI:10.1214/15-AOAS810SUPP.

GUIHENNEUC-JOUYAUX, C., RICHARDSON, S. and LONGINI, I. M. JR. (2000). Modeling markers of disease progression by a hidden Markov process: Application to characterizing CD4 cell decline. *Biometrics* **56** 733–741.

HE, C. Y., ZHANG, C. L., HUNTER, D. J., HANKINSON, S. E., LOUIS, G. M. B., HEDIGER, M. L. and HU, F. B. (2010). Age at menarche and risk of type 2 diabetes: Results from 2 large prospective cohort studies. *Am. J. Epidemiol.* **171** 334–344.

HU, F. B., MANSON, J. E., STAMPFER, M. J., COLDITZ, G., LIU, S., SOLOMON, C. G. and WILLETT, W. C. (2001). Diet, lifestyle, and the risk of type 2 diabetes mellitus in women. *N. Engl. J. Med.* **345** 790–797.

JACKSON, C. H. and SHARPLES, L. D. (2002). Hidden Markov models for the onset and progression of bronchiolitis obliterans syndrome in lung transplant recipients. *Stat. Med.* **21** 113–128.

JACKSON, C. H., SHARPLES, L. D., THOMPSON, S. G., DUFFY, S. W. and COUTO, E. (2003). Multistate Markov models for disease progression with classification error. *The Statistician* **52** 193–209. MR1977260

JACKSON, J. M., DEFOR, T. A., CRAIN, A. L., KERBY, T. J., STRAYER, L. S., LEWIS, C. E., WHITLOCK, E. P., WILLIAMS, S. B., VITOLINS, M. Z., RODABOUGH, R. J., LARSON, J. C., HABERMANN, E. B. and MARGOLIS, K. L. (2014). Validity of diabetes self-reports in the women's health initiative. *Menopause* **8** 861–868.

KIRBY, A. J. and SPIEGELHALTER, D. J. (1994). *Statistical Modelling for the Precursors of Cervical Cancer*. Wiley, New York.

LYLES, R. H., TANG, L., SUPERAK, H. M., KING, C. C., CELENTANO, D. D., LO, Y. and SOBEL, J. D. (2011). Validation data-based adjustments for outcome misclassification in logistic regression: An illustration. *Epidemiology* **22** 589–597.

MARGOLIS, K. L., QI, L. H., BRZYSKI, R., BONDS, D. E., HOWARD, B. V., KEMPOINEN, S., LIU, S. M., ROBINSON, J. G., SAFFORD, M. M., TINKER, L. T., PHILLIPS, L. S. and WOMENS HLTH, I. (2008). Validity of diabetes self-reports in the women's health initiative: Comparison with medication inventories and fasting glucose measurements. *Clinical Trials* **5** 240–247.

MCKEOWN, K. and JEWELL, N. P. (2010). Misclassification of current status data. *Lifetime Data Anal.* **16** 215–230. MR2608286

MEIER, A. S., RICHARDSON, B. A. and HUGHES, J. P. (2003). Discrete proportional hazards models for mismeasured outcomes. *Biometrics* **59** 947–954. MR2025118

NEUHAUS, J. M. (1999). Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika* **86** 843–855. MR1741981

OKSANEN, T., KIVIMÄKI, M., PENTTI, J., VIRTANEN, M., KLAUKKA, T. and VAHTERA, J. (2010). Self-report as an indicator of incident disease. *Ann. Epidemiol.* **20** 547–554.

SATTEN, G. A. and LONGINI, I. M. (1996). Markov chains with measurement error: Estimating the 'true' course of a marker of the progression of human immunodeficiency virus disease. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **45** 275–295.

SHAW, P. A. and PRENTICE, R. L. (2012). Hazard ratio estimation for biomarker-calibrated dietary exposures. *Biometrics* **68** 397–407. MR2959606

SNAPINN, S. M. (1998). Survival analysis with uncertain endpoints. *Biometrics* **54** 209–218.

SPIEGELMAN, D., ROSNER, B. and LOGAN, R. (2000). Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs. *J. Amer. Statist. Assoc.* **95:449** 51–61.

TURNBULL, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. Roy. Statist. Soc. Ser. B* **38** 290–295. MR0652727

The Annals of Applied Statistics
2015, Vol. 9, No. 2, 731–753
DOI: 10.1214/15-AOAS826

BIASED SAMPLING DESIGNS TO IMPROVE RESEARCH EFFICIENCY: FACTORS INFLUENCING PULMONARY FUNCTION OVER TIME IN CHILDREN WITH ASTHMA[1]

BY JONATHAN S. SCHILDCROUT*, PAUL J. RATHOUZ†, LEILA R. ZELNICK‡, SHAWN P. GARBETT* AND PATRICK J. HEAGERTY‡

*Vanderbilt University School of Medicine *, University of Wisconsin School of Medicine and Public Health† and University of Washington School of Public Health‡*

Substudies of the Childhood Asthma Management Program [*Control. Clin. Trials* **20** (1999) 91–120; *N. Engl. J. Med.* **343** (2000) 1054–1063] seek to identify patient characteristics associated with asthma symptoms and lung function. To determine if genetic measures are associated with trajectories of lung function as measured by forced vital capacity (FVC), children in the primary cohort study retrospectively had candidate loci evaluated. Given participant burden and constraints on financial resources, it is often desirable to target a subsample for ascertainment of costly measures. Methods that can leverage the longitudinal outcome on the full cohort to selectively measure informative individuals have been promising, but have been restricted in their use to analysis of the targeted subsample. In this paper we detail two multiple imputation analysis strategies that exploit outcome and partially observed covariate data on the nonsampled subjects, and we characterize alternative design and analysis combinations that could be used for future studies of pulmonary function and other outcomes. Candidate predictor (e.g., IL10 cytokine polymorphisms) associations obtained from targeted sampling designs can be estimated with very high efficiency compared to standard designs. Further, even though multiple imputation can dramatically improve estimation efficiency for covariates available on all subjects (e.g., gender and baseline age), relatively modest efficiency gains were observed in parameters associated with predictors that are exclusive to the targeted sample. Our results suggest that future studies of longitudinal trajectories can be efficiently conducted by use of outcome-dependent designs and associated full cohort analysis.

REFERENCES

BATES, D. and MAECHLER, M. (2010). lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-34.

BRESLOW, N. E., LUMLEY, T., BALLANTYNE, C. M., CHAMBLESS, L. E. and KULICH, M. (2009a). Improved Horvitz–Thompson estimation of model parameters from two-phase stratified samples: Applications in epidemiology. *Stat. Biosci.* **1** 32.

Key words and phrases. Biased sampling, childhood asthma, conditional likelihood, epidemiological study design, forced vital capacity, linear mixed effect models, longitudinal data analysis, multiple imputation, outcome dependent sampling, time-dependent covariates.

BRESLOW, N. E., LUMLEY, T., BALLANTYNE, C. M., CHAMBLESS, L. E. and KULICH, M. (2009b). Using the whole cohort in the analysis of case-cohort data. *Am. J. Epidemiol.* **169** 1398–1405.

BŮŽKOVÁ, P. and LUMLEY, T. (2009). Semiparametric modeling of repeated measurements under outcome-dependent follow-up. *Stat. Med.* **28** 987–1003. MR2518361

CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A. and CRAINICEANU, C. M. (2006). *Measurement Error in Nonlinear Models*: *A Modern Perspective*, 2nd ed. *Monographs on Statistics and Applied Probability* **105**. Chapman & Hall/CRC, Boca Raton, FL. MR2243417

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

CAMP RESEARCH GROUP (1999). The childhood asthma management program (CAMP): Design, rationale, and methods. Childhood asthma management program research group. *Control. Clin. Trials* **20** 91–120.

CAMP RESEARCH GROUP (2000). Long-term effects of budesonide or nedocrimil in children with asthma. *N. Engl. J. Med.* **343** 1054–1063.

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

KISH, L. (1965). *Survey Sampling*. Wiley, New York.

KORN, E. L. and GRAUBARD, B. I. (2011). *Analysis of Health Surveys*. Wiley, New York.

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

LAWLESS, J. F., KALBFLEISCH, J. D. and WILD, C. J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **61** 413–438. MR1680310

LIN, D. Y. and YING, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data. *J. Amer. Statist. Assoc.* **96** 103–126. MR1952726

LIPSITZ, S. R., FITZMAURICE, G. M., IBRAHIM, J. G., GELBER, R. and LIPSHULTZ, S. (2002). Parameter estimation in longitudinal studies with outcome-dependent follow-up. *Biometrics* **58** 621–630. MR1933535

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

LITTLE, R. J. A. and SCHLUCHTER, M. D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika* **72** 497–512. MR0817564

LYON, H., LANGE, C., LAKE, S., SILVERMAN, E. K., RANDOLPH, A. G., KWIATKOWSKI, D., RABY, B. A., LAZARUS, R., WEILAND, K. M., LAIRD, N. and WEISS, S. T. (2004). IL10 gene polymorphisms are associated with asthma phenotypes in children. *Genet. Epidemiol.* **26** 155–165.

MARTI, H. and CHAVANCE, M. (2011). Multiple imputation analysis of case-cohort studies. *Stat. Med.* **30** 1595–1607. MR2828892

NEUHAUS, J., SCOTT, A. J. and WILD, C. J. (2002). The analysis of retrospective family studies. *Biometrika* **89** 23–37. MR1888343

NEUHAUS, J. M., SCOTT, A. J. and WILD, C. J. (2006). Family-specific approaches to the analysis of case–control family data. *Biometrics* **62** 488–494. MR2236831

NEUHAUS, J. M., SCOTT, A. J., WILD, C. J., JIANG, Y., MCCULLOCH, C. E. and BOYLAN, R. (2014). Likelihood-based analysis of longitudinal data from outcome-related sampling designs. *Biometrics* **70** 44–52. MR3251665

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

RAGHUNATHAN, T. E., LEPKOWSKI, J. M., HOEWYK, J. V. and SOLENBERGER, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv. Methodol.* **27** 85–95.

ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866. MR1294730

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

SCHAFER, J. L. (2010). *Analysis of Incomplete Multivariate Data*. CRC Press, Boca Raton, FL.

SCHAFER, J. L. and GRAHAM, J. W. (2002). Missing data: Our view of the state of the art. *Psychol. Methods* **7** 147–177.

SCHILDCROUT, J. S., GARBETT, S. P. and HEAGERTY, P. J. (2013). Outcome vector dependent sampling with longitudinal continuous response data: Stratified sampling based on summary statistics. *Biometrics* **69** 405–416. MR3071059

SCHILDCROUT, J. S. and HEAGERTY, P. J. (2008). On outcome-dependent sampling designs for longitudinal binary response data with time-varying covariates. *Biostatistics* **9** 735–749.

SCHILDCROUT, J. S. and HEAGERTY, P. J. (2011). Outcome-dependent sampling from existing cohorts with longitudinal binary response data: Study planning and analysis. *Biometrics* **67** 1583–1593. MR2872409

SCHILDCROUT, J. S. and RATHOUZ, P. J. (2010). Longitudinal studies of binary response data following case–control and stratified case–control sampling: Design and analysis. *Biometrics* **66** 365–373. MR2758816

SCHILDCROUT, J. S., MUMFORD, S. L., CHEN, Z., HEAGERTY, P. J. and RATHOUZ, P. J. (2012). Outcome-dependent sampling for longitudinal binary response data based on a time-varying auxiliary variable. *Stat. Med.* **31** 2441–2456. MR2972258

SCHILDCROUT, J. S., RATHOUZ, P. J., ZELNICK, L. R., GARBETT, S. P. and HEAGERTY, P. J. (2015). Supplement to "Biased sampling designs to improve research efficiency: Factors influencing pulmonary function over time in children with asthma." DOI:10.1214/15-AOAS826SUPPA, DOI:10.1214/15-AOAS826SUPPB.

VAN BUUREN, S. (2012). *Flexible Imputation of Missing Data*. CRC Press, Boca Raton, FL.

WEAVER, M. A. and ZHOU, H. (2005). An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. *J. Amer. Statist. Assoc.* **100** 459–469. MR2160550

WHITE, I. R., ROYSTON, P. and WOOD, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Stat. Med.* **30** 377–399. MR2758870

ZHOU, H., WEAVER, M. A., QIN, J., LONGNECKER, M. P. and WANG, M. C. (2002). A semiparametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome. *Biometrics* **58** 413–421. MR1908182

ZHOU, H., CHEN, J., RISSANEN, T. H., KORRICK, S. A., HU, H., SALONEN, J. T. and LONGNECKER, M. P. (2007). Outcome-dependent sampling: An efficient sampling and inference procedure for studies with a continuous outcome. *Epidemiology* **18** 461–468.

ZHOU, H., WU, Y., LIU, Y. and CAI, J. (2011). Semiparametric inference for a 2-stage outcome-auxiliary-dependent sampling design with continuous outcome. *Biostatistics* **12** 521–534.

The Annals of Applied Statistics
2015, Vol. 9, No. 2, 754–776
DOI: 10.1214/15-AOAS828
© Institute of Mathematical Statistics, 2015

GOODNESS OF FIT IN NONLINEAR DYNAMICS: MISSPECIFIED RATES OR MISSPECIFIED STATES?

BY GILES HOOKER[1] AND STEPHEN P. ELLNER[2]

Cornell University

This paper introduces diagnostic tests for the nature of lack of fit in ordinary differential equation models (ODEs) proposed for data. We present a hierarchy of three possible sources of lack of fit: unaccounted-for stochastic variation, misspecification of functional forms in rate equations, and omission of dynamic variables in the description of the system. We represent lack of fit by allowing a parameter vector to vary over time, and propose generic testing procedures that do not rely on specific alternative models. Instead, different sources for lack of fit are characterized in terms of nonparametric relationships among latent variables. The tests are carried out through a combination of residual bootstrap and permutation methods. We demonstrate the effectiveness of these tests on simulated data and on real data from laboratory ecological experiments and electro-cardiogram data.

REFERENCES

ABARBANEL, H. D. I. (1996). *Analysis of Observed Chaotic Data*. Springer, New York. MR1363486

ARORA, N. and BIEGLER, L. T. (2004). A trust region SQP algorithm for equality constrained parameter estimation with simple parameter bounds. *Comput. Optim. Appl.* **28** 51–86. MR2049675

BATES, D. M. and WATTS, D. G. (1988). *Nonlinear Regression Analysis and Its Applications*. Wiley, New York. MR1060528

BECKS, L., ELLNER, S. P., JONES, L. E. and HAIRSTON, N. G. (2010). Reduction of adaptive genetic diversity radically alters eco-evolutionary community dynamics. *Ecol. Lett.* **13** 989–997.

BELLMAN, R. and ROTH, R. S. (1971). The use of splines with unknown end points in the identification of systems. *J. Math. Anal. Appl.* **34** 26–33. MR0277269

BOCK, H. G. (1983). Recent advances in parameter identification techniques for ODE. In *Numerical Treatment of Inverse Problems in Differential and Integral Equations* (*Heidelberg*, 1982) (P. Deuflhard and E. Harrier, eds.). *Progr. Sci. Comput.* **2** 95–121. Birkhäuser, Boston, MA. MR0714563

DATTNER, I. and KLAASSEN, C. A. J. (2013). Optimal rate of direct estimators in systems of ordinary differential equations linear in functions of the parameters. Preprint. Available at arXiv:1305.4126.

ELLNER, S. P., SEIFU, Y. and SMITH, R. H. (2002). Fitting population dynamic models to time-series data by gradient matching. *Ecology* **83** 2256–2270.

FAN, J. and YAO, Q. (2003). *Nonlinear Time Series*: *Nonparametric and Parametric Methods*. Springer, New York. MR1964455

GIROLAMI, M. and CALDERHEAD, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 123–214. MR2814492

Key words and phrases. Differential equation, diagnostics, goodness of fit, attractor reconstruction, bootstrap.

GOLDBERGER, A. L., AMARAL, L. A., GLASS, L., HAUSDORFF, J. M., IVANOV, P. C., MARK, R. G., MIETUS, J. E., PENG M. G.B., C.-K. and STANLEY, H. E. (2000). Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation* **101** e215–e220.

GOLIGHTLY, A. and WILKINSON, D. J. (2011). Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface Focus* **1** 1–14.

GOURIÉROUX, C. and MONFORT, A. (1997). *Simulation-Based Econometric Methods*. Oxford Univ. Press, Oxford.

HAMILTON, J. D. (1994). *Time Series Analysis*. Princeton Univ. Press, Princeton, NJ. MR1278033

HILTUNEN, T., HAIRSTON, N. G. JR., HOOKER, G., JONES, L. E. and ELLNER, S. P. (2014). AUG. A newly discovered role of evolution in previously published consumer-resource dynamics. *Ecology Letters* **17** 915–923.

HOOKER, G. (2009). Forcing function diagnostics for nonlinear dynamics. *Biometrics* **65** 928–936. MR2649866

HOOKER, G. and ELLNER, S. P. (2015). Supplement to "Goodness of fit in nonlinear dynamics: Misspecified rates or misspecified states?" DOI:10.1214/15-AOAS828SUPP.

HOOKER, G., LIN, K. K. and ROGERS, B. (2015). Control theory and experimental design in diffusion processes. Under review. *Journal on Uncertainty Quantification*.

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

KANTZ, H. and SCHREIBER, T. (2005). *Nonlinear Time Series Analysis*, Cambridge Univ. Press, Cambridge. MR2040330

MOODY, G. B. and MARK, R. G. (2001). The impact of the MIT-BIH arrhythmia database. *IEEE Eng. Med. Biol. Mag.* **20** 45–50.

PASCUAL, M. and ELLNER, S. P. (2000). Linking ecological patterns to environmental forcing via nonlinear time series models. *Ecology* **81** 2767–2780.

RAMSAY, J. O., HOOKER, G. and GRAVES, S. (2009). *Functional Data Analysis in R and Matlab*. Springer, New York.

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

RATMANN, O., ANDRIEU, C., WIUF, C. and RICHARDSON, S. (2009). Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proc. Nat. Acad. Sci. USA* **106** 10576–10581.

REUMAN, D. C., DESHARNAIS, R. A., COSTANTINO, R. F., AHMAD, O. S. and COHEN, J. E. (2006). Power spectra reveal the influence of stochasticity on nonlinear population dynamics. *Proceedings of the National Academies of Sciences* **103** 18660–18665.

RÖSSLER, O. E. (1976). An equation for continuous chaos. *Physics Letters* **57A(5)** 397–398.

SRIVASTAVA, R. K. (2014). An exact two-sample test in high dimensions using random projections. Preprint. Available at arXiv:1405.1792.

STARK, J., BROOMHEAD, D. S., DAVIES, M. E. and HUKE, J. (1997). Takens embedding theorems for forced and stochastic systems. *Nonlinear. Anal.* **30** 5303–5314. MR1726033

TAKENS, F. (1981). Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence*, *Warwick* 1980 (*Coventry*, 1979/1980). *Lecture Notes in Math.* **898** 366–381. Springer, Berlin. MR0654900

THORBERGSSON, L. and HOOKER, G. (2013). Experimental design for partially observed Markov decision processes. Preprint. Available at arXiv:1209.4019.

TIEN, J. H. and GUCKENHEIMER, J. (2008). Parameter estimation for bursting neural models. *J. Comput. Neurosci.* **24** 358–373. MR2399636

VAN DER POL, B. (1927). On relaxation-oscillations. *The London*, *Edinburgh and Dublin Philosophical Magazine and Journal of Science* **2** 978–992.

VARAH, J. M. (1982). A spline least squares method for numerical parameter estimation in differential equations. *SIAM J. Sci. Statist. Comput.* **3** 28–46. MR0651865

WILSON, H. R. (1999). *Spikes, Decisions, and Actions: The Dynamical Foundations of Neuroscience*. Oxford Univ. Press, New York. MR1972484

WOOD, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature* **466** 1102–U113.

WOOD, S. (2013). mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation. R package version 1.7-27.

WU, H., XUE, H. and KUMAR, A. (2012). Numerical discretization-based estimation methods for ordinary differential equation models via penalized spline smoothing with applications in biomedical research. *Biometrics* **68** 344–352. MR2959600

YOSHIDA, T., JONES, L. E., ELLNER, S. P., FUSSMANN, G. F. and HAIRSTON, N. G. (2003). Rapid evolution drives ecological dynamics in a predator–prey system. *Nature* **424** 303–306.

The Annals of Applied Statistics
2015, Vol. 9, No. 2, 777–800
DOI: 10.1214/15-AOAS816

COVARIANCE PATTERN MIXTURE MODELS FOR THE ANALYSIS OF MULTIVARIATE HETEROGENEOUS LONGITUDINAL DATA[1]

BY LAURA ANDERLUCCI AND CINZIA VIROLI

University of Bologna

We propose a novel approach for modeling multivariate longitudinal data in the presence of unobserved heterogeneity for the analysis of the Health and Retirement Study (HRS) data. Our proposal can be cast within the framework of linear mixed models with discrete individual random intercepts; however, differently from the standard formulation, the proposed Covariance Pattern Mixture Model (CPMM) does not require the usual local independence assumption. The model is thus able to simultaneously model the heterogeneity, the association among the responses and the temporal dependence structure.

We focus on the investigation of temporal patterns related to the cognitive functioning in retired American respondents. In particular, we aim to understand whether it can be affected by some individual socio-economical characteristics and whether it is possible to identify some homogenous groups of respondents that share a similar cognitive profile. An accurate description of the detected groups allows government policy interventions to be opportunely addressed.

Results identify three homogenous clusters of individuals with specific cognitive functioning, consistent with the class conditional distribution of the covariates. The flexibility of CPMM allows for a different contribution of each regressor on the responses according to group membership. In so doing, the identified groups receive a global and accurate phenomenological characterization.

REFERENCES

ANDERLUCCI, L. and VIROLI, C. (2015). Supplement to "Covariance pattern mixture models for the analysis of multivariate heterogeneous longitudinal data." DOI:10.1214/15-AOAS816SUPP.

BANDYOPADHYAY, S., GANGULI, B. and CHATTERJEE, A. (2011). A review of multivariate longitudinal data analysis. *Stat. Methods Med. Res.* **20** 299–330. MR2829153

BANFIELD, J. D. and RAFTERY, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49** 803–821. MR1243494

BARTOLUCCI, F., BACCI, S. and PENNONI, F. (2014). Longitudinal analysis of self-reported health status by mixture latent auto-regressive models. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **63** 267–288. MR3234343

BARTOLUCCI, F., FARCOMENI, A. and PENNONI, F. (2012). *Latent Markov Models for Longitudinal Data*. Chapman & Hall/CRC, London.

BROWN, E. R. and IBRAHIM, J. G. (2003). A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. *Biometrics* **59** 221–228. MR1987388

CELEUX, G. and GOVAERT, G. (1995). Gaussian parsimonious clustering models. *Pattern Recogn.* **28** 781–793.

Key words and phrases. Mixture models, temporal dependence, random effects model.

CHI, E. M. and REINSEL, G. C. (1989). Models for longitudinal data with random effects and AR(1) errors. *J. Amer. Statist. Assoc.* **84** 452–459. MR1010333

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

DE LA CRUZ-MESÍA, R., QUINTANA, F. A. and MARSHALL, G. (2008). Model-based clustering for longitudinal data. *Comput. Statist. Data Anal.* **52** 1441–1457. MR2422747

DUTILLEUL, P. (1999). The MLE algorithm for the matrix normal distribution. *J. Stat. Comput. Simul.* **64** 105–123.

EROSHEVA, E. A., MATSUEDA, R. L. and TELESCA, D. (2014). Breaking bad: Two decades of life-course data analysis in criminology, developmental phsycology, and beyond. *Annual Review of Statistics and Its Application* **1** 301–332.

FERRER, E. and MCARDLE, J. J. (2003). Alternative structural models for multivariate longitudinal data analysis. *Struct. Equ. Model.* **10** 493–524. MR2011191

FITZMAURICE, G., DAVIDIAN, M., VERBEKE, G. and MOLENBERGHS, G., eds. (2009). *Longitudinal Data Analysis*. CRC Press, Boca Raton, FL. MR1500110

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

GOLDSTEIN, H. (1995). *Multilevel Statistical Models*. Wiley, New York.

GRÜN, B. and LEISCH, F. (2007). Fitting finite mixtures of generalized linear regressions in R. *Comput. Statist. Data Anal.* **51** 5247–5252. MR2370869

HEERINGA, S. G., FISHER, G. G., HURD, M. D., LANGA, K. M., OFSTEDAL, M. B., PLASSMAN, B. L., RODGERS, W. and WEIR, D. R. (2007). Aging, demographics and memory study (ADAMS). Sample design, weights, and analysis for ADAMS. Available at http://hrsonline.isr.umich.edu/meta/adams/desc/AdamsSampleWeights.pdf.

JUSTER, F. T. and SUZMAN, R. (1995). An overview of the health and retirement study. *J. Hum. Resour.* **30** 135–145.

KLEINMAN, K. and IBRAHIM, J. (1998). A semi-parametric Bayesian approach to the random effects model. *Biometrics* **54** 921–938.

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

LANGA, K. M., PLASSMAN, B. L., WALLACE, R. B., HERZOG, A. R., HEERINGA, S. G., OFSTEDAL, M. B., BURKE, J. R., FISHER, G. G., FULTZ, N. H., HURD, M. D., POTTER, G. G., RODGERS, W. L., STEFFENS, D. C., WEIR, D. R. and WILLIS, R. J. (2005). The aging, demographics, and memory study: Study design and methods. *Neuroepidemiology* **25** 181–191.

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

LEIBY, B. E., SAMMEL, M. D., TEN HAVE, T. R. and LYNCH, K. G. (2009). Identification of multivariate responders and non-responders by using Bayesian growth curve latent class models. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **58** 505–524. MR2750090

MANRIQUE-VALLIER, D. (2014). Longitudinal mixed membership trajectory models for disability survey data. *Ann. Appl. Stat.* **8** 2268–2291. MR3292497

MCARDLE, J. J., FISHER, G. G. and KADLEC, K. M. (2007). Latent variable analysis of age trends in tests of cognitive ability in the health and retirement survey, 1992–2004. *Psychol. Aging* **22** 525–545.

MCCULLOCH, C. (2008). Joint modelling of mixed outcome types using latent variables. *Stat. Methods Med. Res.* **17** 53–73. MR2420190

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

MCNICHOLAS, P. D. and MURPHY, T. B. (2010). Model-based clustering of longitudinal data. *Canad. J. Statist.* **38** 153–168. MR2676935

MÜLLER, P. and ROSNER, G. (1997). A Bayesian population model with hierarchical mixture priors applied to blood count data. *J. Amer. Statist. Assoc.* **92** 1279–1292.

MÜLLER, P., ROSNER, G. L., DE IORIO, M. and MACEACHERN, S. (2005). A nonparametric Bayesian model for inference in related longitudinal studies. *J. Roy. Statist. Soc. Ser. C* **54** 611–626. MR2137257

MUTHÉN, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika* **29** 81–117. MR1894462

MUTHÉN, B. and ASPAROUHOV, T. (2009). Growth mixture modeling: Analysis with non-Gaussian random effects. In *Longitudinal Data Analysis* 143–165. CRC Press, Boca Raton, FL. MR1500116

NAIK, D. N. and RAO, S. S. (2001). Analysis of multivariate repeated measures data with a Kronecker product structured covariance matrix. *J. Appl. Stat.* **28** 91–105. MR1834425

NEWTON, H. J. (1988). *TIMESLAB*: *A Time Series Analysis Laboratory*. Wadsworth & Brooks/Cole, Pacific Grove, CA.

PLASSMAN, B. L., LANGA, K. M., FISHER, G. G., HEERINGA, S. G., WEIR, D. R., OFSTEDAL, M. B., BURKE, J. R., HURD, M. D., POTTER, G. G., RODGERS, W. L., STEFFENS, D. C., MCARDLE, J. J., WILLIS, R. J. and WALLACE, R. B. (2008). Prevalence of cognitive impairment without dementia in the United States. *Ann. Intern. Med.* **148** 427–434.

POURAHMADI, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika* **86** 677–690. MR1723786

PROUST-LIMA, C., AMIEVA, H. and JACQMIN-GADDA, H. (2013). Analysis of multivariate mixed longitudinal data: A flexible latent process approach. *Br. J. Math. Stat. Psychol.* **66** 470–486. MR3120963

PROUST-LIMA, C. and JACQMIN-GADDA, H. (2005). Estimation of linear mixed models with a mixture of distribution for the random-effects. *Comput. Methods Programs Biomed.* **78** 165–173.

QUANDT, R. E. and RAMSEY, J. B. (1978). Estimating mixtures of normal distributions and switching regressions. *J. Amer. Statist. Assoc.* **73** 730–752. MR0521324

REINSEL, G. (1984). Estimation and prediction in a multivariate random effects generalized linear model. *J. Amer. Statist. Assoc.* **79** 406–414. MR0755095

SKRONDAL, A. and RABE-HESKETH, S. (2004). *Generalized Latent Variable Modeling*: *Multilevel*, *Longitudinal*, *and Structural Equation Models*. Chapman & Hall/CRC, Boca Raton, FL. MR2059021

STEFFENS, D. C., FISHER, G. G., LANGA, K. M., POTTER, G. G. and PLASSMAN, L. G. (2009). Prevalence of depression among older americans: The aging, demographics and memory study. *Int. Psychogeriatr.* **21** 879–888.

TIMMERMAN, M. E. and KIERS, H. A. L. (2003). Four simultaneous component models for the analysis of multivariate time series from more than one subject to model intraindividual and interindividual differences. *Psychometrika* **68** 105–121. MR2272373

VASDEKIS, V. G. S., CAGNONE, S. and MOUSTAKI, I. (2012). A composite likelihood inference in latent variable models for ordinal longitudinal responses. *Psychometrika* **77** 425–441. MR2943106

VERBEKE, G. and LESAFFRE, E. (1996). A linear mixed-effects model with heterogeneity in the random effects population. *J. Amer. Statist. Assoc.* **91** 217–221.

VERBEKE, G., FIEUWS, S., MOLENBERGHS, G. and DAVIDIAN, M. (2014). The analysis of multivariate longitudinal data: A review. *Stat. Methods Med. Res.* **23** 42–59. MR3190686

VERMUNT, J. K. and MAGIDSON, J. (2003). Latent class models for classification. *Comput. Statist. Data Anal.* **41** 531–537. MR1968068

VIROLI, C. (2011). Finite mixtures of matrix normal distributions for classifying three-way data. *Stat. Comput.* **21** 511–522. MR2826689

VIROLI, C. (2012). On matrix-variate regression analysis. *J. Multivariate Anal.* **111** 296–309. MR2944423

The Annals of Applied Statistics
2015, Vol. 9, No. 2, 801–820
DOI: 10.1214/15-AOAS812

WEAKLY SUPERVISED CLUSTERING: LEARNING FINE-GRAINED SIGNALS FROM COARSE LABELS

BY STEFAN WAGER[*,1], ALEXANDER BLOCKER[†] AND NIALL CARDIN[†]

Stanford University[*] and Google, Inc.[†]

Consider a classification problem where we do not have access to labels for individual training examples, but only have average labels over subpopulations. We give practical examples of this setup and show how such a classification task can usefully be analyzed as a *weakly supervised clustering problem*. We propose three approaches to solving the weakly supervised clustering problem, including a latent variables model that performs well in our experiments. We illustrate our methods on an analysis of aggregated elections data and an industry data set that was the original motivation for this research.

REFERENCES

AGRESTI, A. (2002). *Categorical Data Analysis*, 2nd ed. Wiley, New York. MR1914507

BISHOP, C. M. (1995). Training with noise is equivalent to Tikhonov regularization. *Neural Comput*. **7** 108–116.

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

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

BUCKLIN, R. E. and SISMEIRO, C. (2009). Click here for Internet insight: Advances in clickstream data analysis in marketing. *J. Interact. Market* **23** 35–48.

COPAS, J. B. (1988). Binary regression models for contaminated data. *J. Roy. Statist. Soc. Ser. B* **50** 225–265. MR0964178

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

EFRON, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Amer. Statist. Assoc*. **78** 316–331. MR0711106

EFRON, B. and TIBSHIRANI, R. (1993). *An Introduction to the Bootstrap*. CRC press, Boca Raton, FL.

FRALEY, C., RAFTERY, A. E., MURPHY, T. B. and SCRUCCA, L. (2012). MCLUST version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation. Technical report.

GORDON, A. D. (1999). *Classification*. Chapman & Hall, London.

HOFMANN, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn*. **42** 177–196.

HUNTER, D. R. and LANGE, K. (2004). A tutorial on MM algorithms. *Amer. Statist*. **58** 30–37. MR2055509

KÜCK, H. and DE FREITAS, N. (2005). Learning about individuals from group statistics. In *Proceedings of the* 21*st Conference on Uncertainty in Artificial Intelligence* 332–339. AUAI Press, Arlington, VA.

Key words and phrases. Latent variables model, uncertain class label.

LANGE, K., HUNTER, D. R. and YANG, I. (2000). Optimization transfer using surrogate objective functions. *J. Comput. Graph. Statist.* **9** 1–59. MR1819865

LEVY, S. (2011). *In the Plex*: *How Google Thinks*, *Works*, *and Shapes Our Lives*. Simon and Schuster, New York.

MAGDER, L. S. and HUGHES, J. P. (1997). Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology* **146** 195–203.

POLITIS, D. N., ROMANO, J. P. and WOLF, M. (1999). *Subsampling*. Springer, New York. MR1707286

QUADRIANTO, N., SMOLA, A. J., CAETANO, T. S. and LE, Q. V. (2009). Estimating labels from label proportions. *J. Mach. Learn. Res.* **10** 2349–2374. MR2563985

RUEPING, S. (2010). SVM classifier estimation from group probabilities. In *Proceedings of the* 27*th International Conference on Machine Learning* 911–918.

SCULLEY, D., MALKIN, R. G., BASU, S. and BAYARDO, R. J. (2009). Predicting bounce rates in sponsored search advertisements. In *Proceedings of the* 15*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1325–1334. ACM, New York.

SURDEANU, M., TIBSHIRANI, J., NALLAPATI, R. and MANNING, C. D. (2012). Multi-instance multi-label learning for relation extraction. In *Proceedings of the* 2012 *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* 455–465. Association for Computational Linguistics, Stroudsburg, PA.

TÄCKSTRÖM, O. and MCDONALD, R. (2011a). Discovering fine-grained sentiment with latent variable structured prediction models. In *Advances in Information Retrieval* 368–374. Springer, Berlin.

TÄCKSTRÖM, O. and MCDONALD, R. (2011b). Semi-supervised latent variable models for sentence-level sentiment analysis. In *Proceedings of the* 49*th Annual Meeting of the Association for Computational Linguistics*: *Human Language Technologies*. *Short Papers*, *Volume* 2 569–574. Association for Computational Linguistics, Stroudsburg, PA.

TOUTANOVA, K. and JOHNSON, M. (2007). A Bayesian LDA-based model for semi-supervised part-of-speech tagging. In *Advances in Neural Information Processing Systems* 1521–1528. Curran Associates, Red Hook, NY.

VAN DER MAATEN, L., CHEN, M., TYREE, S. and WEINBERGER, K. Q. (2013). Learning with marginalized corrupted features. In *Proceedings of the* 30*th International Conference on Machine Learning* 410–418.

WAGER, S., WANG, S. and LIANG, P. (2013). Dropout training as adaptive regularization. In *Advances in Neural Information Processing Systems*. Curran Associates, Red Hook, NY.

XING, E. P., JORDAN, M. I., RUSSELL, S. and NG, A. (2002). Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems* 505–512. Curran Associates, Red Hook, NY.

XU, G., YANG, S.-H. and LI, H. (2009). Named entity mining from click-through data using weakly supervised latent Dirichlet allocation. In *Proceedings of the* 15*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1365–1374. ACM, New York.

YASUI, Y., PEPE, M., HSU, L., ADAM, B.-L. and FENG, Z. (2004). Partially supervised learning using an EM-boosting algorithm. *Biometrics* **60** 199–206. MR2044116

The Annals of Applied Statistics
2015, Vol. 9, No. 2, 821–848
DOI: 10.1214/13-AOAS700

ESTIMATING HETEROGENEOUS GRAPHICAL MODELS FOR DISCRETE DATA WITH AN APPLICATION TO ROLL CALL VOTING

BY JIAN GUO*, JIE CHENG†, ELIZAVETA LEVINA†,1, GEORGE MICHAILIDIS†,2 AND JI ZHU†,3

Harvard University and University of Michigan*†

We consider the problem of jointly estimating a collection of graphical models for discrete data, corresponding to several categories that share some common structure. An example for such a setting is voting records of legislators on different issues, such as defense, energy, and healthcare. We develop a Markov graphical model to characterize the heterogeneous dependence structures arising from such data. The model is fitted via a joint estimation method that preserves the underlying common graph structure, but also allows for differences between the networks. The method employs a group penalty that targets the common zero interaction effects across all the networks. We apply the method to describe the internal networks of the U.S. Senate on several important issues. Our analysis reveals individual structure for each issue, distinct from the underlying well-known bipartisan structure common to all categories which we are able to extract separately. We also establish consistency of the proposed method both for parameter estimation and model selection, and evaluate its numerical performance on a number of simulated examples.

REFERENCES

AIROLDI, E. M. (2007). Getting started in probabilistic graphical models. *PLoS Comput. Biol.* **3** e252.

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

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

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

BESAG, J. (1986). On the statistical analysis of dirty pictures. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **48** 259–302. MR0876840

CLINTON, J., JACKMAN, S. and RIVERS, D. (2004). The statistical analysis of roll call data. *American Political Science Review* **98** 355–370.

DANAHER, P., WANG, P. and WITTEN, D. M. (2011). The joint graphical lasso for inverse covariance estimation across multiple classes. Available at arXiv:1111.0324.

DE LEEUW, J. (2006). Principal component analysis of senate voting patterns. In *Real Data Analysis* (S. S. Sawilowski, ed.) 405–411. Information Age Publishing, Charlotte, NC.

Key words and phrases. Graphical models, group penalty, high-dimensional data, ℓ_1 penalty, Markov network, binary data.

DIACONIS, P., GOEL, S. and HOLMES, S. (2008). Horseshoes in multidimensional scaling and local kernel methods. *Ann. Appl. Stat.* **2** 777–807. MR2516794

ENELOW, J. M. and HINICH, M. J. (1984). *The Spatial Theory of Voting: An Introduction*. Cambridge Univ. Press, Cambridge.

GERRISH, S. M. (2011). Predicting legislative roll calls from text. In *Proc. 28th Internat. Conf. on Machine Learning (ICML-11)*. Omnipress, Madison, WI.

GUO, J., LEVINA, E., MICHAILIDIS, G. and ZHU, J. (2009). Joint structure estimation of Markov network. Technical report, Dept. Statistics, Univ. Michigan, Ann Arbor, MI.

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

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

HAN, J. H. (2007). Analysing roll calls of the European Parliament: A Bayesian application. *European Union Politics* **8** 479–507.

HARA, S. and WASHIO, T. (2013). Learning a common substructure of multiple graphical Gaussian models. *Neural Networks* **38** 23–38.

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

HØJSGAARD, S. (2004). Statistical inference in context specific interaction models for contingency tables. *Scand. J. Stat.* **31** 143–158. MR2042604

JUNG, S. Y., PARK, Y. C., CHOI, K. S. and KIM, Y. (1996). Markov random field based English part-of-speech tagging system. In *Proceedings of the 16th Conference on Computational Linguistics* 236–242. Association for Computational Linguistics, Stroudsburg, PA.

KOLAR, M. and XING, E. P. (2008). Improved estimation of high-dimensional Ising models. Available at arXiv:0811.1239.

LEEB, H. and PÖTSCHER, B. M. (2008). Sparse estimators and the oracle property, or the return of Hodges' estimator. *J. Econometrics* **142** 201–211. MR2394290

LI, S. Z. (2001). *Markov Random Field Modeling in Image Analysis*. Springer, New York.

LI, H. and GUI, J. (2006). Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics* **7** 302–317.

MATTHEWS, D. R. and STIMSON, J. A. (1975). *Yeas and Nays: Normal Decision-Making in the U.S. House of Representatives*. Wiley, New York.

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

MORTON, R. B. (1999). *Methods and Models: A Guide to the Empirical Analysis of Formal Models in Political Science*. Cambridge Univ. Press, Cambridge.

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

POOLE, K. T. and ROSENTHAL, H. (1997). *Congress: A Political-Economic History of Roll-Call Voting*. Oxford Univ. Press, Oxford.

PÖTSCHER, B. M. and LEEB, H. (2009). On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *J. Multivariate Anal.* **100** 2065–2082. MR2543087

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

RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980. MR2836766

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

WAINWRIGHT, M. J. and JORDAN, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning* **1** 1–305.

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

YANG, S., PAN, Z., SHEN, X., WONKA, P. and YE, J. (2012). Fused multiple graphical lasso. Available at arXiv:1209.2139.

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

ZHOU, N. and ZHU, J. (2007). Group variable selection via a hierarchical lasso and its oracle property. Technical report, Dept. Statistics, Univ. Michigan, Ann Arbor, MI.

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

The Annals of Applied Statistics
2015, Vol. 9, No. 2, 849–865
DOI: 10.1214/14-AOAS765
In the Public Domain

A TWO-STATE MIXED HIDDEN MARKOV MODEL FOR RISKY TEENAGE DRIVING BEHAVIOR

BY JOHN C. JACKSON*, PAUL S. ALBERT[†,1] AND ZHIWEI ZHANG[†,1]

United States Military Academy and Eunice Kennedy Shriver National Institute of Child Health and Human Development*[†]

This paper proposes a joint model for longitudinal binary and count outcomes. We apply the model to a unique longitudinal study of teen driving where risky driving behavior and the occurrence of crashes or near crashes are measured prospectively over the first 18 months of licensure. Of scientific interest is relating the two processes and predicting crash and near crash outcomes. We propose a two-state mixed hidden Markov model whereby the hidden state characterizes the mean for the joint longitudinal crash/near crash outcomes and elevated g-force events which are a proxy for risky driving. Heterogeneity is introduced in both the conditional model for the count outcomes and the hidden process using a shared random effect. An estimation procedure is presented using the *forward–backward* algorithm along with adaptive Gaussian quadrature to perform numerical integration. The estimation procedure readily yields hidden state probabilities as well as providing for a broad class of predictors.

REFERENCES

ALBERT, P. S. and FOLLMANN, D. A. (2007). Random effects and latent processes approaches for analyzing binary longitudinal data with missingness: A comparison of approaches using opiate clinical trial data. *Stat. Methods Med. Res.* **16** 417–439. MR2405479

ALFÒ, M., MARUOTTI, A. and TROVATO, G. (2011). A finite mixture model for multivariate counts under endogenous selectivity. *Stat. Comput.* **21** 185–202. MR2774851

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

ALTMAN, R. M. (2008). A variance component test for mixed hidden Markov models. *Statist. Probab. Lett.* **78** 1885–1893. MR2528559

BARTOLUCCI, F., LUPPARELLI, M. and MONTANARI, G. E. (2009). Latent Markov model for longitudinal binary data: An application to the performance evaluation of nursing homes. *Ann. Appl. Stat.* **3** 611–636. MR2750675

BARTOLUCCI, F. and PENNONI, F. (2007). A class of latent Markov models for capture-recapture data allowing for time, heterogeneity, and behavior effects. *Biometrics* **63** 568–578. MR2370816

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

JACKSON, J., ALBERT, P. and ZHANG, Z. (2015). Supplement to "A two-state mixed hidden Markov model for risky teenage driving behavior." DOI:10.1214/14-AOAS765SUPP.

Key words and phrases. Adaptive quadrature, hidden Markov model, joint model, random effects.

JACKSON, J. C., ALBERT, P. S., ZHANG, Z. and SIMONS-MORTON, B. (2013). Ordinal latent variable models and their application in the study of newly licensed teenage drivers. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **62** 435–450. MR3060625

LANGEHEINE, R. and VAN DE POL, F. (1994). Discrete-time mixed Markov latent class models. In *Analyzing Social and Political Change: A Casebook of Methods* (A. Dale and R. B. Davies, eds.). Sage, London.

LIU, Q. and PIERCE, D. A. (1994). A note on Gauss–Hermite quadrature. *Biometrika* **81** 624–629. MR1311107

MARUOTTI, A. (2011). Mixed hidden Markov models for longitudinal data: An overview. *International Statistical Review* **79** 427–454.

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

MCCULLOCH, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *J. Amer. Statist. Assoc.* **92** 162–170. MR1436105

SCOTT, S. L. (2002). Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *J. Amer. Statist. Assoc.* **97** 337–351. MR1963393

SHIRLEY, K. E., SMALL, D. S., LYNCH, K. G., MAISTO, S. A. and OSLIN, D. W. (2010). Hidden Markov models for alcoholism treatment trial data. *Ann. Appl. Stat.* **4** 366–395. MR2758176

SIMONS-MORTON, B. G., OUIMET, M. C., ZHANG, Z., KLAUER, S. E., LEE, S. E., WANG, J., ALBERT, P. S. and DINGUS, T. A. (2011). Crash and risky driving involvement among novice adolescent drivers and their parents. *Am. J. Public Health* **101** 2362–2367.

SMITH, M. D. and MOFFATT, P. G. (1999). Fisher's information on the correlation coefficient in bivariate logistic models. *Aust. N. Z. J. Stat.* **41** 315–330. MR1718026

WEI, G. C. G. and TANNER, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Amer. Statist. Assoc.* **85** 699–704.

The Annals of Applied Statistics
2015, Vol. 9, No. 2, 866–882
DOI: 10.1214/15-AOAS813

MULTI-SPECIES DISTRIBUTION MODELING USING PENALIZED MIXTURE OF REGRESSIONS

BY FRANCIS K. C. HUI[*,1], DAVID I. WARTON[*,2] AND SCOTT D. FOSTER[†,3]

University of New South Wales[*] and CSIRO[†]

Multi-species distribution modeling, which relates the occurrence of multiple species to environmental variables, is an important tool used by ecologists for both predicting the distribution of species in a community and identifying the important variables driving species co-occurrences. Recently, Dunstan, Foster and Darnell [*Ecol. Model.* **222** (2011) 955–963] proposed using finite mixture of regression (FMR) models for multi-species distribution modeling, where species are clustered based on their environmental response to form a small number of "archetypal responses." As an illustrative example, they applied their mixture model approach to a presence–absence data set of 200 marine organisms, collected along the Great Barrier Reef in Australia. Little attention, however, was given to the problem of model selection—since the archetypes (mixture components) may depend on different but likely overlapping sets of covariates, a method is needed for performing variable selection on all components simultaneously. In this article, we consider using penalized likelihood functions for variable selection in FMR models. We propose two penalties which exploit the grouped structure of the covariates, that is, each covariate is represented by a group of coefficients, one for each component. This leads to an attractive form of shrinkage that allows a covariate to be removed from all components simultaneously. Both penalties are shown to possess specific forms of variable selection consistency, with simulations indicating they outperform other methods which do not take into account the grouped structure. When applied to the Great Barrier Reef data set, penalized FMR models offer more insight into the important variables driving species co-occurrence in the marine community (compared to previous results where no model selection was conducted), while offering a computationally stable method of modeling complex species–environment relationships (through regularization).

REFERENCES

CLARK, J. S. (2010). Individuals and the variation needed for high species diversity in forest trees. *Science* **327** 1129–1132.

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

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

DUNSTAN, P. K., FOSTER, S. D., HUI, F. K. C. and WARTON, D. I. (2013). Finite mixture of regression modeling for high-dimensional count and biomass data in ecology. *J. Agric. Biol. Environ. Stat.* **18** 357–375. MR3110898

Key words and phrases. Community level models, finite mixture models, penalized likelihood, regularization, species archetype models, variable selection.

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

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

FITHIAN, W. and HASTIE, T. (2013). Finite-sample equivalence in statistical models for presence-only data. *Ann. Appl. Stat.* **7** 1917–1939. MR3161707

FOLLMANN, D. A. and LAMBERT, D. (1991). Identifiability of finite mixtures of logistic regression models. *J. Statist. Plann. Inference* **27** 375–381. MR1108557

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

GRÜN, B. and LEISCH, F. (2008). Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. *J. Classification* **25** 225–247. MR2476114

HENNIG, C. (2000). Identifiability of models for clusterwise linear regression. *J. Classification* **17** 273–296. MR1814811

HUI, F. K. C., WARTON, D. I., FOSTER, S. D. and DUNSTAN, P. K. (2013). To mix or not to mix: Comparing the predictive performance of mixture models versus separate species distribution models. *Ecology* **94** 1913–1919.

HUI, F. K. C., WARTON, D. I. and FOSTER, S. D. (2015a). Tuning parameter selection for the adaptive lasso using ERIC. *J. Amer. Statist. Assoc.* **110** 262–269. MR3338501

HUI, F. K. C., WARTON, D. I. and FOSTER, S. D. (2015b). Supplement to "Multi-species distribution modeling using penalized mixture of regressions." DOI:10.1214/15-AOAS813SUPP.

KHALILI, A. and CHEN, J. (2007). Variable selection in finite mixture of regression models. *J. Amer. Statist. Assoc.* **102** 1025–1038. MR2411662

KHALILI, A. and LIN, S. (2013). Regularization in finite mixture of regression models with diverging number of parameters. *Biometrics* **69** 436–446. MR3071062

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, London. MR3223057

MCLACHLAN, G. and PEEL, D. (2004). *Finite Mixture Models*. Wiley, New York.

OVASKAINEN, O., HOTTOLA, J. and SIITONEN, J. (2010). Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology* **91** 2514–2521.

OVASKAINEN, O. and SOININEN, J. (2011). Making more out of sparse data: Hierarchical modeling of species communities. *Ecology* **92** 289–295.

PITCHER, R. C., DOHERTY, P. P., ARNOLD, P. P., HOOPER, J. J., GRIBBLE, N. N. et al. (2007). *Seabed Biodiversity on the Continental Shelf of the Great Barrier Reef World Heritage Area*. CSIRO Marine and Atmospheric Research, Queensland, Australia.

POLLOCK, L. J., TINGLEY, R., MORRIS, W. K., GOLDING, N., O'HARA, R. B., PARRIS, K. M., VESK, P. A. and MCCARTHY, M. A. (2014). Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (JSDM). *Methods in Ecology and Evolution* **5** 397–406.

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

SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2013). A sparse-group lasso. *J. Comput. Graph. Statist.* **22** 231–245. MR3173712

STÄDLER, N., BÜHLMANN, P. and VAN DE GEER, S. (2010). ℓ_1-penalization for mixture regression models. *TEST* **19** 209–256. MR2677722

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

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

WARTON, D. I. and SHEPHERD, L. C. (2010). Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. *Ann. Appl. Stat.* **4** 1383–1402. MR2758333

WEDEL, M. and DESARBO, W. S. (1995). A mixture likelihood approach for generalized linear models. *J. Classification* **12** 21–55.

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., LI, R. and TSAI, C.-L. (2010). Regularization parameter selections via generalized information criterion. *J. Amer. Statist. Assoc.* **105** 312–323. MR2656055

ZHAO, P., ROCHA, G. and YU, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Ann. Statist.* **37** 3468–3497. MR2549566

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

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

The Annals of Applied Statistics
2015, Vol. 9, No. 2, 883–900
DOI: 10.1214/15-AOAS814

JUMP DETECTION IN GENERALIZED ERROR-IN-VARIABLES REGRESSION WITH AN APPLICATION TO AUSTRALIAN HEALTH TAX POLICIES[1]

BY YICHENG KANG[*], XIAODONG GONG[§], JITI GAO[¶] AND PEIHUA QIU[*]

University of Florida[*], *University of Canberra*[§], *Australian National University*[§], *IZA*[§] *and Monash University*[¶]

Without measurement errors in predictors, discontinuity of a nonparametric regression function at unknown locations could be estimated using a number of existing approaches. However, it becomes a challenging problem when the predictors contain measurement errors. In this paper, an error-in-variables jump point estimator is suggested for a nonparametric generalized error-in-variables regression model. A major feature of our method is that it does not impose any parametric distribution on the measurement error. Its performance is evaluated by both numerical studies and theoretical justifications. The method is applied to studying the impact of Medicare Levy Surcharge on the private health insurance take-up rate in Australia.

REFERENCES

BUCHMUELLER, T., DIDARDO, J. and VALLETTA, R. (2011). The effect of an employer health insurance mandate on health insurance coverage and the demand for labor: Evidence from Hawaii. *American Economic Journal Economic Policy* **3** 25–51.

BUTLER, J. R. G. (2002). Policy change and private health insurance: Did the cheapest policy do the trick? *Aust. Health Rev.* **25** 33–41.

CARROLL, R. J., MACA, J. D. and RUPPERT, D. (1999). Nonparametric regression in the presence of measurement error. *Biometrika* **86** 541–554. MR1723777

CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A. and CRAINICEANU, C. M. (2006). *Measurement Error in Nonlinear Models*: *A Modern Perspective*, 2nd ed. Chapman & Hall, Boca Raton, FL. MR2243417

CHENG, K. F. and LIN, P. E. (1981). Nonparametric estimation of a regression function. *Z. Wahrsch. Verw. Gebiete* **57** 223–233. MR0626817

COMTE, F. and TAUPIN, M.-L. (2007). Adaptive estimation in a nonparametric regression model with errors-in-variables. *Statist. Sinica* **17** 1065–1090. MR2397388

COOK, J. R. and STEFANSKI, L. A. (1994). Simulation–extrapolation estimation in parametric measurement error models. *J. Amer. Statist. Assoc.* **89** 1314–1328.

DELAIGLE, A. (2008). An alternative view of the deconvolution problem. *Statist. Sinica* **18** 1025–1045. MR2440402

DELAIGLE, A. and MEISTER, A. (2007). Nonparametric regression estimation in the heteroscedastic errors-in-variables problem. *J. Amer. Statist. Assoc.* **102** 1416–1426. MR2372541

FAN, J. and MASRY, E. (1992). Multivariate regression estimation with errors-in-variables: Asymptotic normality for mixing processes. *J. Multivariate Anal.* **43** 237–271. MR1193614

Key words and phrases. Bandwidth selection, demand for private health insurance, exponential family, generalized regression, kernel smoothing, measurement errors.

FAN, J. and TRUONG, Y. K. (1993). Nonparametric regression with errors in variables. *Ann. Statist.* **21** 1900–1925. MR1245773

FINKELSTEIN, A. (2002). The effect of tax subsidies to employer-provided supplementary health insurance: Evidence from Canada. *J. Public Econ.* **84** 305–339.

FRECH, H., HOPKINS, S. and MACDONALD, G. (2003). The Australian private health insurance boom: Was it subsidies or liberalized regulation? *Oxf. Econ. Pap.* **22** 58–64.

GIJBELS, I. and GODERNIAUX, A.-C. (2004). Bandwidth selection for changepoint estimation in nonparametric regression. *Technometrics* **46** 76–86. MR2043389

GRUBER, J. and POTERBA, D. (1994). Tax incentives and the decision to purchase health insurance: Evidence from the self-employed. *Q. J. Econ.* **123** 831–862.

HALL, P. and MEISTER, A. (2007). A ridge-parameter approach to deconvolution. *Ann. Statist.* **35** 1535–1558. MR2351096

JOO, J.-H. and QIU, P. (2009). Jump detection in a regression curve and its derivative. *Technometrics* **51** 289–305. MR2562274

KANG, Y., GONG, X., GAO, J. and QIU, P. (2015). Supplement to "Jump detection in generalized error-in-variables regression with an application to Australian health tax policies." DOI:10.1214/15-AOAS814SUPP.

LEE, D. and LEMIEUX, T. (2010). Regression discontinuity designs in economics. *J. Econ. Lit.* **48** 281–355.

LIANG, H. and WANG, N. (2005). Partially linear single-index measurement error models. *Statist. Sinica* **15** 99–116. MR2125722

MEISTER, A. (2009). *Deconvolution Problems in Nonparametric Statistics*. Springer, Berlin. MR2768576

MITRINOVIĆ, D. S., PEČARIĆ, J. E. and FINK, A. M. (1993). *Classical and New Inequalities in Analysis*. Kluwer Academic, Dordrecht. MR1220224

MÜLLER, H.-G. (1992). Change-points in nonparametric regression analysis. *Ann. Statist.* **20** 737–761. MR1165590

MÜLLER, C. H. (2002). Robust estimators for estimating discontinuous functions. *Metrika* **55** 99–109 (electronic). MR1903286

PALANGKARAYA, A. and YONG, J. (2005). Effects of recent carrot-and-stick policy initiatives on private health insurance coverage in Australia. *Econ. Rec.* **81** 262–272.

PALANGKARAYA, A., YONG, J., WEBSTER, E. and DAWKINS, P. (2009). The income distributive implications of recent private health insurance policy reforms in Australia. *Eur. J. Health Econ.* **10** 135–148.

QIU, P. H. (1991). Estimation of a kind of jump regression function. *Systems Sci. Math. Sci.* **4** 1–13. MR1182761

QIU, P. H. (1994). Estimation of the number of jumps of the jump regression functions. *Comm. Statist. Theory Methods* **23** 2141–2155. MR1293176

QIU, P. (2005). *Image Processing and Jump Regression Analysis*. Wiley, Hoboken, NJ. MR2111430

QIU, P. and YANDELL, B. (1998). A local polynomial jump-detection algorithm in nonparametric regression. *Technometrics* **40** 141–152. MR1626927

RODRÍGUEZ, M. and STOYANOVA, A. (2004). The effect of private insurance access on the choice of GP/specialist and public/private provider in Spain. *Health Econ.* **13** 689–703.

STAUDENMAYER, J. and RUPPERT, D. (2004). Local polynomial regression and simulation–extrapolation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 17–30. MR2035756

STEFANSKI, L. A. (2000). Measurement error models. *J. Amer. Statist. Assoc.* **95** 1353–1358. MR1825293

STEFANSKI, L. A. and COOK, J. R. (1995). Simulation–extrapolation: The measurement error jackknife. *J. Amer. Statist. Assoc.* **90** 1247–1256. MR1379467

TAUPIN, M.-L. (2001). Semi-parametric estimation in the nonlinear structural errors-in-variables model. *Ann. Statist.* **29** 66–93. MR1833959

WU, J. S. and CHU, C. K. (1993). Kernel-type estimators of jump points and values of a regression function. *Ann. Statist.* **21** 1545–1566. MR1241278

The Annals of Applied Statistics
2015, Vol. 9, No. 2, 901–925
DOI: 10.1214/15-AOAS815

HMMSEQ: A HIDDEN MARKOV MODEL FOR DETECTING DIFFERENTIALLY EXPRESSED GENES FROM RNA-SEQ DATA

BY SHIQI CUI[*], SUBHARUP GUHA[*,1], MARCO A. R. FERREIRA[†,2] AND ALLISON N. TEGGE[†]

University of Missouri[*] *and Virginia Tech*[†]

We introduce hmmSeq, a model-based hierarchical Bayesian technique for detecting differentially expressed genes from RNA-seq data. Our novel hmmSeq methodology uses hidden Markov models to account for potential co-expression of neighboring genes. In addition, hmmSeq employs an integrated approach to studies with technical or biological replicates, automatically adjusting for any extra-Poisson variability. Moreover, for cases when paired data are available, hmmSeq includes a paired structure between treatments that incoporates subject-specific effects. To perform parameter estimation for the hmmSeq model, we develop an efficient Markov chain Monte Carlo algorithm. Further, we develop a procedure for detection of differentially expressed genes that automatically controls false discovery rate. A simulation study shows that the hmmSeq methodology performs better than competitors in terms of receiver operating characteristic curves. Finally, the analyses of three publicly available RNA-seq data sets demonstrate the power and flexibility of the hmmSeq methodology. An R package implementing the hmmSeq framework will be submitted to CRAN upon publication of the manuscript.

REFERENCES

AGRAWAL, R. and GOMEZ-PINILLA, F. (2012). 'Metabolic syndrome' in the brain: Deficiency in omega-3 fatty acid exacerbates dysfunctions in insulin receptor signalling and cognition. *J. Gen. Physiol.* **590** 2485–2499.

ALBERTS, B., BRAY, D., LEWIS, J., RAFF, M., ROBERTS, K. and WATSON, J. D. (1994). *Molecular Biology of the Cell*. Garland Science, New York.

AUER, P. L. and DOERGE, R. W. (2011). A two-stage Poisson model for testing RNA-Seq data. *Stat. Appl. Genet. Mol. Biol.* **10** Art. 26, 28. MR2800690

AUER, P. L., SRIVASTAVA, S. and DOERGE, R. W. (2012). Differential expression—the next generation and beyond. *Brief. Funct. Genomics* **11** 57–62.

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

BLEKHMAN, R., MARIONI, J. C., ZUMBO, P., STEPHENS, M. and GILAD, Y. (2010). Sex-specific and lineage-specific alternative splicing in primates. *Genome Res.* **20** 180–189.

BULLARD, J. H., PURDOM, E., HANSEN, K. D. and DUDOIT, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. *BMC Bioinformatics* **11** 94.

Key words and phrases. Bayesian hierarchical model, first order dependence, next-generation sequencing, overdispersion, serial correlation.

CARON, H., VAN SCHAIK, B., VAN DER MEE, M., BAAS, F., RIGGINS, G., VAN SLUIS, P., HERMUS, M.-C., VAN ASPEREN, R., BOON, K., VOUTE, P. A. et al. (2001). The human transcriptome map: Clustering of highly expressed genes in chromosomal domains. *Science* **291** 1289–1292.

CHIB, S. and GREENBERG, E. (1994). Bayes inference in regression models with ARMA(p, q) errors. *J. Econometrics* **64** 183–206. MR1310523

CUI, S., GUHA, S., FERREIRA, M. and TEGGE, A. N. (2015). Supplement to "hmmSeq: A hidden Markov model for detecting differentially expressed genes from RNA-seq data." DOI:10.1214/15-AOAS815SUPP.

EDELMAN, L. B. and FRASER, P. (2012). Transcription factories: Genetic programming in three dimensions. *Curr. Opin. Genet. Dev.* **22** 110–114.

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

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

GOGOLLA, N., GALIMBERTI, I., DEGUCHI, Y. and CARONI, P. (2009). Wnt signaling mediates experience-related regulation of synapse numbers and mossy fiber connectivities in the adult hippocampus. *Neuron* **62** 510–525.

GUHA, S., LI, Y. and NEUBERG, D. (2008). Bayesian hidden Markov modeling of array CGH data. *J. Amer. Statist. Assoc.* **103** 485–497. MR2523987

HARDCASTLE, T. J. (2009). baySeq: Empirical Bayesian analysis of patterns of differential expression in count data. R package version 1.10.0.

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

HENN, A. D., WU, S., QIU, X., RUDA, M., STOVER, M., YANG, H., LIU, Z., WELLE, S. L., HOLDEN-WILTSE, J., WU, H. and ZAND, M. S. (2013). High-resolution temporal response patterns to influenza vaccine reveal a distinct human plasma cell gene signature. *Sci. Rep.* **3** 2327.

HUANG, D. W., SHERMAN, B. T. and LEMPICKI, R. A. (2009a). Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37** 1–13.

HUANG, D. W., SHERMAN, B. T. and LEMPICKI, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4** 44–57.

HURST, L. D., PÁL, C. and LERCHER, M. J. (2004). The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.* **5** 299–310.

KALITA, A., GUPTA, S., SINGH, P., SUROLIA, A. and BANERJEE, K. (2013). IGF-1 stimulated upregulation of cyclin D1 is mediated via STAT5 signaling pathway in neuronal cells. *IUBMB Life* **65** 462–471.

KARLEBACH, G. and SHAMIR, R. (2008). Modelling and analysis of gene regulatory networks. *Nat. Rev., Mol. Cell Biol.* **9** 770–780.

KVAM, V. M., LIU, P. and SI, Y. (2012). A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *Am. J. Bot.* **99** 248–256.

LANGMEAD, B., HANSEN, K. D., LEEK, J. T. et al. (2010). Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biol.* **11** R83.

LEE, J., JI, Y., LIANG, S., CAI, G. and MÜLLER, P. (2011). On differential gene expression using RNA-seq data. *Cancer Inform.* **10** 205–215.

LOUHIMO, R. and HAUTANIEMI, S. (2011). CNAmet: An R package for integrating copy number, methylation and expression data. *Bioinformatics* **27** 887–888.

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

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

MERCER, T. R. and MATTICK, J. S. (2013). Understanding the regulatory and transcriptional complexity of the genome through structure. *Genome Res.* **23** 1081–1088.

MICHALAK, P. (2008). Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. *Genomics* **91** 243–248.

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

NEWTON, M. A., NOUEIRY, A., SARKAR, D. and AHLQUIST, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5** 155–176.

PE'ER, D. and HACOHEN, N. (2011). Principles and strategies for developing network models in cancer. *Cell* **144** 864–873.

RABINER, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77** 257–286.

ROBERTSON, S. D., MATTHIES, H. J., OWENS, W. A., SATHANANTHAN, V., CHRISTIANSON, N. S. B., KENNEDY, J. P., LINDSLEY, C. W., DAWS, L. C. and GALLI, A. (2010). Insulin reveals akt signaling as a novel regulator of norepinephrine transporter trafficking and norepinephrine homeostasis. *J. Neurosci.* **30** 11305–11316.

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

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

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

SCOTT, S. L. (2002). Bayesian methods for hidden Markov models. *J. Amer. Statist. Assoc.* **97** 337–351.

SI, Y. and LIU, P. (2013). An optimal test with maximum average power while controlling FDR with application to RNA-Seq data. *Biometrics* **69** 594–605. MR3106587

SINGER, G. A., LLOYD, A. T., HUMINIECKI, L. B. and WOLFE, K. H. (2005). Clusters of co-expressed genes in mammalian genomes are conserved by natural selection. *Mol. Biol. Evol.* **22** 767–775.

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

STOREY, J. D. and TIBSHIRANI, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100** 9440–9445. MR1994856

TEGGE, A. N., CALDWELL, C. W. and XU, D. (2012). Pathway correlation profile of gene–gene co-expression for identifying pathway perturbation. *PLoS ONE* **7** e52127.

TITTERINGTON, D. M., SMITH, A. F. M. and MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, Chichester. MR0838090

VAN ARENSBERGEN, J., VAN STEENSEL, B. and BUSSEMAKER, H. J. (2014). In search of the determinants of enhancer–promoter interaction specificity. *Trends Cell Biol.* **24** 695–702.

WILHELM, S. and MANJUNATH, B. G. (2013). tmvtnorm: Truncated multivariate normal and student t distribution. R package version 1.4-8.

ZEGER, S. L. and KARIM, M. R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *J. Amer. Statist. Assoc.* **86** 79–86. MR1137101

ZENG, J., KONOPKA, G., HUNT, B. G., PREUSS, T. M., GESCHWIND, D. and YI, S. V. (2012). Divergent whole-genome methylation maps of human and chimpanzee brains reveal epigenetic basis of human regulatory evolution. *Am. J. Hum. Genet.* **91** 455–465.

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K. and Liu, X. (2014). Comparison of RNA-seq and microarray in transcriptome profiling of activated t cells. *PLoS ONE* **9** e78644.

The Annals of Applied Statistics
2015, Vol. 9, No. 2, 926–949
DOI: 10.1214/15-AOAS819

TRACKING RAPID INTRACELLULAR MOVEMENTS: A BAYESIAN RANDOM SET APPROACH

BY VASILEIOS MAROULAS AND ANDREAS NEBENFÜHR[1]

University of Tennessee

We focus on the biological problem of tracking organelles as they move through cells. In the past, most intracellular movements were recorded manually, however, the results are too incomplete to capture the full complexity of organelle motions. An automated tracking algorithm promises to provide a complete analysis of noisy microscopy data. In this paper, we adopt statistical techniques from a Bayesian random set point of view. Instead of considering each individual organelle, we examine a random set whose members are the organelle states and we establish a Bayesian filtering algorithm involving such set states. The propagated multi-object densities are approximated using a Gaussian mixture scheme. Our algorithm is applied to synthetic and experimental data.

REFERENCES

AVISAR, D., PROKHNEVSKY, A. I., MAKAROVA, K. S., KOONIN, E. V. and DOLJA, V. V. (2008). Myosin XI-K is required for rapid trafficking of Golgi stacks, peroxisomes, and mitochondria in leaf cells of Nicotiana benthamiana. *Plant Physiol.* **146** 1098–1108.

BAR-SHALOM, Y. and BLAIR, W. D. (2000). *Multitarget-Multisensor Tracking*: *Applications and Advances*. Norwood, MA, Artech House.

BLACKMAN, S. and POPOLI, R. (1999). *Design and Analysis of Modern Tracking Systems*. Artech House, Norwood, MA.

BRÉMAUD, P. (1981). *Point Processes and Queues*: *Martingale Dynamics*. Springer, New York. MR0636252

CLARK, D. E., PANTA, K. and VO, B.-N. (2006). The GM-PHD filter multiple target tracker. In 9*th International Conference on Information Fusion* 1–8. IEEE.

COLLINGS, D. A., HARPER, J. D. I., MARC, J., OVERALL, R. and MULLEN, L. R. T. (2002). Life in the fast lane: Actin-based motility of plant peroxisomes. *Canadian Journal of Botany* **80** 430–441.

CORTI, B. (1774). *Osservazioni microscopiche sulla tremella e sulla circolazione del fluido in una pianta acquajuola*. Lucca.

DALEY, D. J. and VERE-JONES, D. (1988). *An Introduction to the Theory of Point Processes*. Springer, New York. MR0950166

DANUSER, G. (2011). Computer vision in cell biology. *Cell* **147** 973–978.

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

FORTMANN, T. E., BAR-SHALOM, Y. and SCHEFFE, M. (1983). Sonar tracking of multiple targets using joint probabilistic data association. *Oceanic Engineering*, *IEEE Journal of* **8** 173–184.

Key words and phrases. Multi-object Bayesian filtering, cardinalized probability hypothesis density, Gaussian mixture implementation, monitoring intracellular movements, random finite set theory, finite set statistics.

GILKS, W. R. and BERZUINI, C. (2001). Following a moving target—Monte Carlo inference for dynamic Bayesian models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 127–146. MR1811995

GOODMAN, I. R., MAHLER, R. P. S. and NGUYEN, H. T. (1997). *Mathematics of Data Fusion. Theory and Decision Library. Series B: Mathematical and Statistical Methods* **37**. Kluwer Academic, Dordrecht. MR1635258

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

GUTIERREZ, R., LINDEBOOM, J. J., PAREDEZ, A. R., EMONS, A. M. C. and EHRHARDT, D. W. (2009). Arabidopsis cortical microtubules position cellulose synthase delivery to the plasma membrane and interact with cellulose synthase trafficking compartments. *Nature Cell Biology* **11** 797–806.

HAMADA, T., TOMINAGA, M., FUKAYA, T., NAKAMURA, M., NAKANO, A., WATANABE, Y., HASHIMOTO, T. and BASKIN, T. I. (2012). RNA processing bodies, peroxisomes, golgi bodies, mitochondria, and endoplasmic reticulum tubule junctions frequently pause at cortical microtubules. *Plant and Cell Physiology* **53** 699–798.

HOFFMAN, J. R. and MAHLER, R. P. S. (2002). Multitarget miss distance and its applications. In *Proceedings of the Fifth International Conference on Information Fusion* 1 149–155. IEEE.

HUGHES, J. and FRICKS, J. (2011). A mixture model for quantum dot images of kinesin motor assays. *Biometrics* **67** 588–595. MR2829027

HUGHES, J., FRICKS, J. and HANCOCK, W. (2010). Likelihood inference for particle location in fluorescence microscopy. *Ann. Appl. Stat.* **4** 830–848. MR2758423

JAQAMAN, K., LOERKE, D., METTLEN, M., KUWATA, H., GRINSTEIN, S., SCHMIDT, S. L. and DANUSER, G. (2008). Robust single-particle tracking in live-cell time-lapse sequences. *Nat. Methods* **5** 695–702.

KANG, K. and MAROULAS, V. (2013). Drift homotopy methods for a nonGaussian filter. In 16*th International Conference on Information Fusion* (*FUSION*) 1088–1094. IEEE, Istanbul.

KANG, K., MAROULAS, V. and SCHIZAS, I. D. (2014). Drift homotopy particle filter for non-Gaussian multi-target tracking. In 17*th International Conference on Information Fusion* (*FUSION*) 1–7. IEEE, Salamanca.

KLAUS, A. and HERIBERT, H. (2004). Reactive oxygen species: Metabolism, oxidative stress, and signal transduction. *Annual Review of Plant Biology* **55** 373–399.

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

LIU, J. S. and CHEN, R. (1998). Sequential Monte Carlo methods for dynamic systems. *J. Amer. Statist. Assoc.* **93** 1032–1044. MR1649198

LOGAN, D. C. and LEAVER, C. J. (2000). Mitochondria-targeted GFP highlights the heterogeneity of mitochondrial shape, size and movement within living plant cells. *J. Experimental Botany* **51** 865–871.

MADISON, S. L. and NEBENFÜHR, A. (2013). Understanding myosin functions in plants: Are we there yet? Preprint.

MAHLER, R. P. S. (2003). Multitarget Bayes filtering via first-order multitarget moments. *Aerospace and Electronic Systems, IEEE Transactions on* **39** 1152–1178.

MAHLER, R. P. S. (2007). *Statistical Multisource-Multitarget Information Fusion*. Artech House, Norwood, MA.

MAHLER, R. P. S. and MAROULAS, V. (2013). Tracking spawning objects. *Radar, Sonar Navigation, IET* **7** 321–331.

MAROULAS, V. and STINIS, P. (2012). Improved particle filters for multi-target tracking. *J. Comput. Phys.* **231** 602–611. MR2872093

MØLLER, J. and WAAGEPETERSEN, R. P. (2004). *Statistical Inference and Simulation for Spatial Point Processes. Monographs on Statistics and Applied Probability* **100**. Chapman & Hall/CRC, Boca Raton, FL. MR2004226

NEBENFÜHR, A., GALLAGHER, L. A., DUNAHAY, T. G., FROHLICK, J. A., MAZURKIEWICZ, A. M., MEEHL, J. B. and STAEHELIN, L. A. (1999). Stop-and-go movements of plant Golgi stacks are mediated by the actomyosin system. *Plant Physiol.* **121** 1127–1141.

NELSON, B. K., CAI, X. and NEBENFÜHR, A. (2007). A multi-color set of in vivo organelle markers for colocalization studies in Arabidopsis and other plants. *Plant Journal* **51** 1126–1136.

OJANGU, E.-L., TANNER, K., PATA, P., JÄRVE, K., HOLWEG, C. L., TRUVE, E. and PAVES, H. (2012). Myosins XI-K, XI-1, and XI-2 are required for development of pavement cells, trichomes, and stigmatic papillae in Arabidopsis. *BMC Plant Biol.* **12** 81.

PAREDEZ, A. R., SOMERVILLE, C. R. and EHRHARDT, D. W. (2006). Visualization of cellulose synthase demonstrates functional association with microtubules. *Science* **311** 1491–1495.

PEREMYSLOV, V. V., PROKHNEVSKY, A. I., AVISAR, D. and DOLJA, V. V. (2008). Two class XI myosins function in organelle trafficking and root hair development in Arabidopsis. *Plant Physiol.* **146** 1009–1116.

SBALZARINI, I. F. and KOUMOUTSAKOS, P. (2005). Feature point tracking and trajectory analysis for video imaging in cell biology. *J. Struct. Biol.* **151** 182–195.

SCHNEIDER, C. A., RASBAND, W. S. and ELICEIRI, K. W. (2012). NIH image to ImageJ: 25 years of image analysis. *Nat. Methods* **9** 671–675.

SCHUHMACHER, D., VO, B.-T. and VO, B.-N. (2008). A consistent metric for performance evaluation of multi-object filters. *IEEE Trans. Signal Process.* **56** 3447–3457. MR2516955

SCHUHMACHER, D. and XIA, A. (2008). A new metric between distributions of point processes. *Adv. in Appl. Probab.* **40** 651–672. MR2454027

SHIMMEN, T. (2007). The sliding theory of cytoplasmic streaming: Fifty years of progress. *J. Plant Res.* **120** 31–43.

SHIMMEN, T. and YOKOTA, E. (1994). Physiological and biochemical aspects of cytoplasmic streaming. *International Review of Cytology* **155** 97–139.

SMAL, I. (2009). Particle filtering methods for subcellular motion analysis. Ph.D. thesis, Erasmus Univ. Rotterdam, Rotterdam, The Netherlands.

SMAL, I., NIESSEN, W. and MEIJERING, E. (2006). Particle filtering for multiple object tracking in molecular cell biology. In *IEEE Nonlinear Statistical Signal Processing Workshop* 129–132. IEEE.

SMAL, I., DRAEGESTEIN, K., GALJART, N., NIESSEN, W. and MEIJERING, E. (2008). Particle filtering for multiple object tracking in dynamic fluorescence microscopy images: Application to microtubule growth analysis. *Medical Imaging, IEEE Transactions on* **27** 789–804.

SNYDER, C., BENGTSSON, T., BICKEL, P. and ANDERSON, J. (2008). Obstacles to high-dimensional particle filtering. *Mon. Wea. Rev.* **136** 4629–4640.

TOMINAGA, M., KOJIMA, H., YOKOTA, E., ORII, H., NAKAMORI, R., KATAYAMA, E., NASON, M., SHIMMEN, T. and OIWA, K. (2003). Higher plant myosin XI moves processively on actin with 35 nm steps at high velocity. *EMBO Journal* **22** 1263–1272.

VAN GESTEL, K., KÖHLER, R. H. and VERBELEN, J.-P. (2002). Plant mitochondria move on F-actin, but their positioning in the cortical cytoplasm depends on both F-actin and microtubules. *J. Experimental Botany* **53** 659–667.

VICK, J. K. and NEBENFÜHR, A. (2012). Putting on the breaks: Regulation of organelle movements in plant cellst. *Journal of Integrative Plant Biology* **54** 868–874.

VO, B.-T., VO, B.-N. and CANTONI, A. (2007). Analytic implementations of the cardinalized probability hypothesis density filter. *IEEE Trans. Signal Process.* **55** 3553–3567. MR2517522

WEARE, J. (2009). Particle filtering with path sampling and an application to a bimodal ocean current model. *J. Comput. Phys.* **228** 4312–4331. MR2531900

The Annals of Applied Statistics
2015, Vol. 9, No. 2, 950–968
DOI: 10.1214/15-AOAS820

BAYESIAN DETECTION OF EMBRYONIC GENE EXPRESSION ONSET IN *C. ELEGANS*[1]

BY JIE HU[*], ZHONGYING ZHAO[†], HARI KRISHNA YALAMANCHILI[‡], JUNWEN WANG[‡], KENNY YE[§] AND XIAODAN FAN[*]

Chinese University of Hong Kong[*], *Hong Kong Baptist University*[†], *University of Hong Kong*[‡] *and Albert Einstein College of Medicine*[§]

To study how a zygote develops into an embryo with different tissues, large-scale 4D confocal movies of *C. elegans* embryos have been produced recently by experimental biologists. However, the lack of principled statistical methods for the highly noisy data has hindered the comprehensive analysis of these data sets. We introduced a probabilistic change point model on the cell lineage tree to estimate the embryonic gene expression onset time. A Bayesian approach is used to fit the 4D confocal movies data to the model. Subsequent classification methods are used to decide a model selection threshold and further refine the expression onset time from the branch level to the specific cell time level. Extensive simulations have shown the high accuracy of our method. Its application on real data yields both previously known results and new findings.

REFERENCES

ANDERSEN, E. C., LU, X. and HORVITZ, H. R. (2006). *C. elegans* ISWI and NURF301 antagonize an Rb-like pathway in the determination of multiple cell fates. *Development* **133** 2695–2704.

BAO, Z., MURRAY, J. I., BOYLE, T., OOI, S. L., SANDEL, M. J. and WATERSTON, R. H. (2006). Automated cell lineage tracing in *caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA* **103** 2707–2712.

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

GOOD, K., CIOSK, R., NANCE, J., NEVES, A., HILL, R. J. and PRIESS, J. R. (2004). The t-box transcription factors tbx-37 and tbx-38 link glp-1/notch signaling to mesoderm induction in *C. elegans* embryos. *Development* **131** 1967–1968.

GURALNIK, V. and SRIVASTAVA, J. (1999). Event detection from time series data. In *KDD'99 Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **17** 33–42. ACM, San Diego, CA.

HARRIS, D., BURGES, J. C. C., KAUFMAN, L., SMOLA, J. A. and VLADIMIR, N. V. (1997). Support vector regression machines. *Adv. Neural Inf. Process. Syst.* **9** 155–161.

HU, J., ZHAO, Z., YALAMANCHILI, H., WANG J., YE, K. and FAN, X. (2015). Supplement to "Bayesian detection of embryonic gene expression onset in *C. elegans*." DOI:10.1214/15-AOAS820SUPPA, DOI:10.1214/15-AOAS820SUPPB, DOI:10.1214/15-AOAS820SUPPC, DOI:10.1214/15-AOAS820SUPPD, DOI:10.1214/15-AOAS820SUPPE, DOI:10.1214/15-AOAS820SUPPF.

Key words and phrases. 4D confocal microscopy, embryonic onset, change point detection, Bayesian method.

KRAUSE, M. (1995). Myod and myogenesis in *C. elegans*. *BioEssays* **17** 228.

LIBEN-NOWELL, D. and KLEINBERG, J. (2008). Tracing information flow on a global scale using Internet chain-letter data. *Proc. Natl. Acad. Sci. USA* **105** 4633–4638.

LIU, X., LONG, F., PENG, H., AERNI, S. J., JIANG, M., BLANCO, A. S., MURRAY, J. I., PRESTON, E., MERICLE, B., BATZOGLOU, S., MYERS, E. W. and KIM, S. K. (2009). Analysis of cell fate from single-cell gene expression profiles in *C. elegans*. *Cell* **139** 623–633.

LONG, F., PENG, H., LIU, X., KIM, S. K. and MYERS, E. (2009). A 3D digital atlas of *C. elegans* and its application to single-cell analyses. *Nat. Methods* **6** 667–672.

MADUROA, M. F., HILLB, R. J., HEIDC, P. J., SMITHA, E. D. N., ZHU, J., PRIESS, J. R. and ROTHMAN, J. H. (2005). Genetic redundancy in endoderm specification within the genus caenorhabditis. *Dev. Biol.* **284** 522.

MURRAY, J. I., BAO, Z., BOYLE, T. J., BOECK, M. E., MERICLE, B. L., NICHOLAS, T. J., ZHAO, Z., SANDEL, M. J. and WATERSTON, R. H. (2008). Automated analysis of embryonic gene expression with cellular resolution in *C. elegans*. *Nature Methods* **5** 703–709.

MURRAY, J. I., BOYLE, T. J., PRESTON, E., VAFEADOS, D., MERICLE, B., WEISDEPP, P., ZHAO, Z., BAO, Z., BOECK, M. and WATERSTON, R. H. (2012). Multidimensional regulation of gene expression in the *C. elegans* embryo. *Genome Research* **22** 1282–1294.

PERREAULT, L., BERNIER, J., BOBEE, B. and PARENT, E. (2000). Bayesian change-point analysis in hydrometeorological time series. *Journal of Hydrology* **235** 221–241.

PICARD, D. (1985). Testing and estimating change-points in time series. *Adv. in Appl. Probab.* **17** 841–867. MR0809433

SPENCER, W. C., ZELLER, G., WATSON, J. D., HENZ, S. R., WATKINS, K. L., MCWHIRTER, R. D., PETERSEN, S., SREEDHARAN, V. T., WIDMER, C., JO, J., REINKE, V., PETRELLA, L., STROME, S., STETINA, S. E. V., KATZ, M., SHAHAM, S., RATSCH, G. and MILLER, D. M. (2011). A spatial and temporal map of *C. elegans* gene expression. *Genome Research* **21** 325–341.

SULSTON, J. E., SCHIERENBERG, E., WHITE, J. G. and THOMSON, J. N. (1983). The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol.* **100** 64–119.

YALAMANCHILI, H. K., YAN, B., LI, M. J., QIN, J., ZHAO, Z., CHIN, F. Y. and WANG, J. (2013). Dynamic delay gene network inference from high temporal data using gapped local alignment. *Bioinformatics* **30** 377–383.

The Annals of Applied Statistics
2015, Vol. 9, No. 2, 969–991
DOI: 10.1214/15-AOAS821

ASSESSING PHENOTYPIC CORRELATION THROUGH THE MULTIVARIATE PHYLOGENETIC LATENT LIABILITY MODEL[1]

BY GABRIELA B. CYBIS[*,2], JANET S. SINSHEIMER[†], TREVOR BEDFORD[‡], ALISON E. MATHER[§], PHILIPPE LEMEY[¶] AND MARC A. SUCHARD[†]

Federal University of Rio Grande do Sul[*], *University of California, Los Angeles*[†], *Fred Hutchinson Cancer Research Center*[‡], *Wellcome Trust Sanger Institute*[§] *and KU Leuven*[¶]

Understanding which phenotypic traits are consistently correlated throughout evolution is a highly pertinent problem in modern evolutionary biology. Here, we propose a multivariate phylogenetic latent liability model for assessing the correlation between multiple types of data, while simultaneously controlling for their unknown shared evolutionary history informed through molecular sequences. The latent formulation enables us to consider in a single model combinations of continuous traits, discrete binary traits and discrete traits with multiple ordered and unordered states. Previous approaches have entertained a single data type generally along a fixed history, precluding estimation of correlation between traits and ignoring uncertainty in the history. We implement our model in a Bayesian phylogenetic framework, and discuss inference techniques for hypothesis testing. Finally, we showcase the method through applications to columbine flower morphology, antibiotic resistance in *Salmonella* and epitope evolution in influenza.

REFERENCES

BAELE, G., LEMEY, P., BEDFORD, T., RAMBAUT, A., SUCHARD, M. A. and ALEKSEYENKO, A. V. (2012). Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Mol. Biol. Evol.* **29** 2157–2167.

BLUM, M. G., DAMERVAL, C., MANEL, S. and FRANÇOIS, O. (2004). Brownian models and coalescent structures. *Theor. Popul. Biol.* **65** 249–261.

BOYD, D., PETERS, G. A., CLOECKAERT, A., BOUMEDINE, K. S., CHASLUS-DANCLA, E., IMBERECHTS, H. and MULVEY, M. R. (2001). Complete nucleotide sequence of a 43-kilobase genomic island associated with the multidrug resistance region of *Salmonella* enterica serovar Typhimurium DT104 and its identification in phage type DT120 and serovar Agona. *J. Bacteriol.* **183** 5725–5732.

BRESLAW, J. A. (1994). Random sampling from a truncated multivariate normal distribution. *Appl. Math. Lett.* **7** 1–6. MR1349883

BUSH, R. M., BENDER, C. A., SUBBARAO, K., COX, N. J. and FITCH, W. M. (1999). Predicting the evolution of human influenza A. *Science* **286** 1921–1925.

COX, N. J. and BENDER, C. A. (1995). The molecular epidemiology of influenza viruses. In *Seminars in Virology* **6** 359–370. Elsevier, Amsterdam.

Key words and phrases. Bayesian phylogenetics, threshold model, evolution, genotype-phenotype correlation.

CYBIS, G. B., SINSHEIMER, J. S., BEDFORD, T., MATHER, A. E., LEMEY, P. and SUCHARD, M. A. (2015). Supplement to "Assessing phenotypic correlation through the multivariate phylogenetic latent liability model." DOI:10.1214/15-AOAS821SUPP.

DAMIEN, P. and WALKER, S. G. (2001). Sampling truncated normal, beta, and gamma densities. *J. Comput. Graph. Statist.* **10** 206–215. MR1939697

DRUMMOND, A. J., HO, S. Y. W., PHILLIPS, M. J. and RAMBAUT, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4** e88.

DRUMMOND, A. J., SUCHARD, M. A., XIE, D. and RAMBAUT, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29** 1969–1973.

FALCONER, D. S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann. Hum. Genet.* **29** 51–76.

FARIA, N. R., SUCHARD, M. A., RAMBAUT, A., STREICKER, D. G. and LEMEY, P. (2013). Simultaneously reconstructing viral cross-species transmission history and identifying the underlying constraints. *Philosophical Transactions of the Royal Society B*: *Biological Sciences* **368** 20120196.

FELSENSTEIN, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17** 368–376.

FELSENSTEIN, J. (1985). Phylogenies and the comparative method. *Amer. Nat.* **125** 1–15.

FELSENSTEIN, J. (2005). Using the quantitative genetic threshold model for inferences between and within species. *Philosophical Transactions of the Royal Society B*: *Biological Sciences* **360** 1427–1434.

FELSENSTEIN, J. (2012). A comparative method for both discrete and continuous characters using the threshold model. *Amer. Nat.* **179** 145–156.

FITCH, W. M., LEITER, J. M., LI, X. Q. and PALESE, P. (1991). Positive Darwinian evolution in human influenza a viruses. *Proc. Natl. Acad. Sci. USA* **88** 4270–4274.

FRECKLETON, R. P. (2012). Fast likelihood calculations for comparative analyses. *Methods in Ecology and Evolution* **3** 940–947.

GELFAND, A. E., SMITH, A. F. M. and LEE, T.-M. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *J. Amer. Statist. Assoc.* **87** 523–532. MR1173816

GRAFEN, A. (1989). The phylogenetic regression. *Philosophical Transactions of the Royal Society of London*. *Series B*, *Biological Sciences* **326** 119–157.

HADFIELD, J. D. and NAKAGAWA, S. (2010). General quantitative genetic methods for comparative biology: Phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *J. Evol. Biol.* **23** 494–508.

HO, L. S. T. and ANÉ, C. (2014). A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Systematic Biology* **3** 397–402.

HUELSENBECK, J. P. and RANNALA, B. (2003). Detecting correlation between characters in a comparative analysis with uncertain phylogeny. *Evolution* **57** 1237–1247.

IVES, A. R. and GARLAND, T. JR. (2010). Phylogenetic logistic regression for binary dependent variables. *Syst. Biol.* **59** 9–26.

JEFFREYS, H. (1935). Some tests of significance, treated by the theory of probability. *Math. Proc. Cambridge Philos. Soc.* **31** 203–222.

KOEL, B. F., BURKE, D. F., BESTEBROER, T. M., VAN DER VLIET, S., ZONDAG, G. C., VERVAET, G., SKEPNER, E., LEWIS, N. S., SPRONKEN, M. I., RUSSELL, C. A. et al. (2013). Substitutions near the receptor binding site determine major antigenic change during influenza virus evolution. *Science* **342** 976–979.

LANDIS, M. J., SCHRAIBER, J. G. and LIANG, M. (2013). Phylogenetic analysis using Lévy processes: Finding jumps in the evolution of continuous traits. *Syst. Biol.* **62** 193–204.

LARTILLOT, N. and POUJOL, R. (2011). A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Mol. Biol. Evol.* **28** 729–744.

LEMEY, P., RAMBAUT, A., WELCH, J. J. and SUCHARD, M. A. (2010). Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.* **27** 1877–1885.

LEWIS, P. O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology* **50** 913–925.

LIU, J. S., LIANG, F. and WONG, W. H. (2000). The multiple-try method and local optimization in Metropolis sampling. *J. Amer. Statist. Assoc.* **95** 121–134. MR1803145

MATHER, A. E., MATTHEWS, L., MELLOR, D. J., REEVE, R., DENWOOD, M. J., BOERLIN, P., REID-SMITH, R. J., BROWN, D. J., COIA, J. E., BROWNING, L. M. et al. (2012). An ecological approach to assessing the epidemiology of antimicrobial resistance in animal and human populations. *Proceedings of the Royal Society B*: *Biological Sciences* **279** 1630–1639.

MATHER, A. E., REID, S. W. J., MASKELL, D. J., PARKHILL, J., FOOKES, M. C., HARRIS, S. R., BROWN, D. J., COIA, J. E., MULVEY, M. R., GILMOUR, M. W. et al. (2013). Distinguishable epidemics of multidrug-resistant *Salmonella* Typhimurium DT104 in different hosts. *Science* **341** 1514–1517.

MININ, V. N., BLOOMQUIST, E. W. and SUCHARD, M. A. (2008). Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol. Biol. Evol.* **25** 1459–1471.

NOVEMBRE, J. and SLATKIN, M. (2009). Likelihood-based inference in isolation-by-distance models using the spatial distributions of low frequency alleles. *Evolution* **63** 2914–2925.

PAGEL, M. (1994). Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society of London*. *Series B*: *Biological Sciences* **255** 37–45.

PLOTKIN, J. B. and DUSHOFF, J. (2003). Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza a virus. *Proc. Natl. Acad. Sci. USA* **100** 7152–7157.

PYBUS, O. G., SUCHARD, M. A., LEMEY, P., BERNARDIN, F. J., RAMBAUT, A., CRAWFORD, F. W., GRAY, R. R., ARINAMINPATHY, N., STRAMER, S. L., BUSCH, M. P. and DELWART, E. L. (2012). Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proc. Natl. Acad. Sci. USA* **109** 15066–15071.

REVELL, L. J. (2012). phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* **3** 217–223.

REVELL, L. J. (2014). Ancestral character estimation under the threshold model from quantitative genetics. *Evolution* **68** 743–759.

ROBERT, C. P. (1995). Simulation of truncated normal variables. *Stat. Comput.* **5** 121–125.

SUCHARD, M. A., WEISS, R. E. and SINSHEIMER, J. S. (2001). Bayesian selection of continuous-time Markov chain evolutionary models. *Molecular Biology and Evolution* **18** 1001–1013.

VAN DER NIET, T. and JOHNSON, S. D. (2012). Phylogenetic evidence for pollinator-driven diversification of angiosperms. *Trends in Ecology & Evolution* **27** 353–361.

WHITTALL, J. B. and HODGES, S. A. (2007). Pollinator shifts drive increasingly long nectar spurs in columbine flowers. *Nature* **447** 706–709.

WHITTALL, J. B., VOELCKEL, C., KLIEBENSTEIN, D. J. and HODGES, S. A. (2006). Convergence, constraint and the role of gene expression during adaptive radiation: Floral anthocyanins in Aquilegia. *Mol. Ecol.* **15** 4645–4657.

WRIGHT, S. (1934). An analysis of variability in number of digits in an inbred strain of guinea pigs. *Genetics* **19** 506.

The Annals of Applied Statistics
2015, Vol. 9, No. 2, 992–1023
DOI: 10.1214/15-AOAS822
In the Public Domain

EXAMINING SOCIOECONOMIC HEALTH DISPARITIES USING A RANK-DEPENDENT RÉNYI INDEX

BY MAKRAM TALIH

National Center for Health Statistics

The Rényi index (RI) is a one-parameter class of indices that summarize health disparities among population groups by measuring divergence between the distributions of disease burden and population shares of these groups. The *rank-dependent* RI introduced in this paper is a two-parameter class of health disparity indices that also accounts for the association between socioeconomic rank and health; it may be derived from a rank-dependent social welfare function. Two competing classes are discussed and the rank-dependent RI is shown to be more robust to changes in the distribution of either socioeconomic rank or health. The standard error and sampling distribution of the rank-dependent RI are evaluated using linearization and resampling techniques, and the methodology is illustrated using health survey data from the U.S. National Health and Nutrition Examination Survey and registry data from the U.S. Surveillance, Epidemiology and End Results Program. Such data underlie many population-based objectives within the U.S. Healthy People 2020 initiative. The rank-dependent RI provides a unified mathematical framework for eliciting various societal positions with regards to the policies that are tied to such wide-reaching public health initiatives. For example, if population groups with lower socioeconomic position were ascertained to be more likely to utilize costly public programs, then the parameters of the RI could be selected to reflect prioritizing those population groups for intervention or treatment.

REFERENCES

AABERGE, R. (2005). Asymptotic distribution theory of empirical rank-dependent measures of inequality. Discussion Papers No. 402, Statistics Norway, Research Department. Available at http://www.ssb.no/a/publikasjoner/pdf/DP/dp402.pdf.

AKERS, A. Y., NEWMAN, S. J. and SMITH, J. S. (2007). Factors underlying disparities in cervical cancer incidence, screening, and treatment in the United States. *Curr. Probl. Cancer* **31** 157–181.

ASADA, Y., YOSHIDA, Y. and WHIPP, A. M. (2013). Summarizing social disparities in health. *Milbank Q.* **91** 5–36.

ATKINSON, A. B. (1970). On the measurement of inequality. *J. Econom. Theory* **2** 244–263. MR0449508

BARON, S. (2012). Measuring occupation as an element of socioeconomic status/position. Presented at the March 2012 Hearing of the National Committee on Vital and Health Statistics on Minimum Data Standards for the Measurement of Socioeconomic Status in Federal Health Surveys. Available at http://www.ncvhs.hhs.gov/120308p2.pdf.

Key words and phrases. Social welfare function, concentration curve, Rényi divergence, statistical inference, complex survey data, health surveillance data.

BENNETT, C. J. and MITRA, S. (2013). Multidimensional poverty: Measurement, estimation, and inference. *Econometric Rev.* **32** 57–83. MR2988920

BERREBI, Z. M. and SILBER, J. (1981). Weighting income ranks and levels: A multiple-parameter generalization for absolute and relative inequality indices. *Econom. Lett.* **7** 391–397.

BIEWEN, M. and JENKINS, S. P. (2006). Variance estimation for generalized entropy and Atkinson inequality indices: The complex survey data case. *Oxford Bulletin of Economics and Statistics* **68** 371–383.

BLEICHRODT, H., ROHDE, K. I. M. and VAN OURTI, T. (2012). An experimental test of the concentration index. *J. Health Econ.* **31** 86–98.

BLEICHRODT, H. and VAN DOORSLAER, E. (2006). A welfare economics foundation for health inequality measurement. *J. Health Econ.* **25** 945–957.

BOMMIER, A. and STECKLOV, G. (2002). Defining health inequality: Why Rawls succeeds where social welfare theory fails. *J. Health Econ.* **21** 497–513.

BORRELL, L. N. and TALIH, M. (2012). Examining periodontal disease disparities among U.S. adults 20 years of age and older: NHANES III (1988–1994) and NHANES 1999–2004. *Public Health Rep.* **127** 497–506.

BORRELL, L. N. and TALIH, M. (2011). A symmetrized Theil index measure of health disparities: An example using dental caries in U.S. children and adolescents. *Stat. Med.* **30** 277–290. MR2758878

BOYD, M. and NAM, C. B. (2004). Occupational status in 2000: Over a century of census-based measurement. *Popul. Res. Policy Rev.* **23** 327–358.

BRAVEMAN, P. (2006). Health disparities and health equity: Concepts and measurement. *Annu. Rev. Public Health* **27** 167–194.

BRAVEMAN, P. A., CUBBIN, C., EGERTER, S., WILLIAMS, D. R. and PAMUK, E. (2010). Socioeconomic disparities in health in the United States: What the patterns tell us. *Am. J. Public Health* **100** S186–S196.

CHEN, Z., ROY, K. and CRAWFORD, C. A. G. (2012). Evaluation of variance estimators for the concentration and health achievement indices: A Monte Carlo simulation. *Health Econ.* **21** 1375–1381.

CHEN, J. T., BECKFIELD, J., WATERMAN, P. D. and KRIEGER, N. (2013). Can changes in the distributions of and associations between education and income bias temporal comparisons of health disparities? An exploration with causal graphs and simulations. *Am. J. Epidemiol.* **177** 870–881.

CHENG, N. F., HAN, P. Z. and GANSKY, S. A. (2008). Methods and software for estimating health disparities: The case of children's oral health. *Am. J. Epidemiol.* **168** 906–914.

CHERNOFF, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statistics* **23** 493–507. MR0057518

COSTA-FONT, J. and HERNÁNDEZ-QUEVEDO, C. (2012). Measuring inequalities in health: What do we know? What do we need to know? *Health Policy* **106** 195–206.

COWELL, F. A., DAVIDSON, R. and FLACHAIRE, E. (2011). Goodness of fit: An axiomatic approach. Discussion Paper DT 2011-50, Groupement de Recherche en Economie Quantitative d'Aix-Marseille (GREQAM). Available at http://halshs.archives-ouvertes.fr/docs/00/63/90/75/PDF/DTGREQAM2011_50.pdf.

COWELL, F. A. and GARDINER, K. (1999). Welfare weights. OFT Research Paper 202, STICERD—London School of Economics. Available at http://darp.lse.ac.uk/papersDB/Cowell-Gardiner_(OFT).pdf.

COWELL, F. A. and KUGA, K. (1981). Additivity and the entropy concept: An axiomatic approach to inequality measurement. *J. Econom. Theory* **25** 131–143. MR0636017

DEAN, H. D., WILLIAMS, K. M. and FENTON, K. A. (2013). From theory to action: Applying social determinants of health to public health practice. *Public Health Rep.* **128** S31–S34.

DECANCQ, K. and LUGO, M. A. (2009). Measuring inequality of well-being with a correlation-sensitive multidimensional Gini index. Working paper ECINEQ 2009-124, Society for the Study of Economic Inequality. Available at http://www.ecineq.org/milano/WP/ECINEQ2009-124.pdf.

DHHS (2014). HealthyPeople.gov. U.S. Department of Health and Human Services (DHHS), Washington, DC. Available at http://healthypeople.gov.

ERREYGERS, G. (2009a). Can a single indicator measure both attainment and shortfall inequality? *J. Health Econ.* **28** 885–893.

ERREYGERS, G. (2009b). Correcting the concentration index. *J. Health Econ.* **28** 504–515.

ERREYGERS, G. and VAN OURTI, T. (2011). Measuring socioeconomic inequality in health, health care and health financing by mean of rank-dependent indices: A recipe for good practice. *J. Health Econ.* **30** 685–694.

FOSTER, J. E., MCGILLIVRAY, M. and SETH, S. (2013). Composite indices: Rank robustness, statistical association, and redundancy. *Econometric Rev.* **32** 35–56. MR2988919

FROHLICH, K. L. and POTVIN, L. (2008). The inequality paradox: The population approach and vulnerable populations. *Am. J. Public Health* **98** 216–221.

GRAVELLE, H. (2003). Measuring income related inequality in health: Standardisation and the partial concentration index. *Health Econ.* **12** 803–819.

HARPER, S., LYNCH, J., MEERSMAN, S. C., BREEN, N., DAVIS, W. W. and REICHMAN, M. E. (2008). An overview of methods for monitoring social disparities in cancer with an example using trends in lung cancer incidence by area-socioeconomic position and race-ethnicity, 1992–2004. *Am. J. Epidemiol.* **167** 889–899.

HARPER, S., KING, N. B., MEERSMAN, S. C., REICHMAN, M. E., BREEN, N. and LYNCH, J. (2010). Implicit value judgments in the measurement of health inequalities. *Milbank Q.* **88** 4–29.

JOHNSON, C. L., PAULOSE-RAM, R., OGDEN, C. L., CARROLL, M. D., KRUSZAN-MORAN, D., DOHRMANN, S. M. and CURTIN, L. R. (2013). National health and nutrition examination survey: Analytic guidelines, 1999–2010. *Vital and Health Statistics*, *Series* 2 **161**.

KAKWANI, N., WAGSTAFF, A. and VAN DOORSLAER, E. (1997). Socioeconomic inequalities in health: Measurement, computation, and statistical inference. *J. Econom.* **77** 87–103.

KEPPEL, K., PAMUK, E., LYNCH, J., CARTER-POKRAS, O., KIM, I., MAYS, V., PEARCY, J., SCHOENBACH, V. and WEISSMAN, J. S. (2005). Methodological issues in measuring health disparities. *Vital and Health Statistics*, *Series* 2 **141**.

KJELLSSON, G. and GERDTHAM, U.-G. (2013). On correcting the concentration index for binary variables. *J. Health Econ.* **32** 659–670.

KOOLMAN, X. and VAN DOORSLAER, E. (2004). On the interpretation of a concentration index of inequality. *Health Econ.* **13** 649–656.

KRIEGER, N., WILLIAMS, D. R. and MOSS, N. E. (1997). Measuring social class in US public health research: Concepts, methodologies, and guidelines. *Annu. Rev. Public Health* **18** 341–378.

KUCZMARSKI, R. J., OGDEN, C. L., GUO, S. S., GRUMMER-STRAWN, L. M., FLEGAL, K. M., MEI, Z., WEI, R., CURTIN, L. R., ROCHE, A. F. and JOHNSON, C. L. (2002). 2000 CDC growth charts for the United States: Methods and development. *Vital and Health Statistics*, *Series* 11 **246**.

LAMBERT, P. and ZHENG, B. (2011). On the consistent measurement of attainment and shortfall inequality. *J. Health Econ.* **30** 214–219.

LANGEL, M. and TILLÉ, Y. (2013). Variance estimation of the Gini index: Revisiting a result several times published. *J. Roy. Statist. Soc. Ser. A* **176** 521–540. MR3045858

LUMLEY, T. (2004). Analysis of complex survey samples. *J. Stat. Softw.* **9** 1–19.

LUMLEY, T. (2011). "Survey": Analysis of complex survey samples. R package version 3.26.

LYNCH, J., SMITH, G. D., HARPER, S., HILLEMEIER, M., ROSS, N., KAPLAN, G. A. and WOLFSON, M. (2004). Is income inequality a determinant of population health? Part 1. A systematic review. *Milbank Q.* **82** 5–99.

MAASOUMI, E. (1986). The measurement and decomposition of multi-dimensional inequality. *Econometrica* **54** 991–998.

MACKENBACH, J. P. and KUNST, A. E. (1997). Measuring the magnitude of socio-economic inequalities in health: An overview of available measures illustrated with two examples from Europe. *Soc. Sci. Med.* **44** 757–771.

MAKDISSI, P. and YAZBECK, M. (2012). Avoiding blindness to health status: A new class of health achievement and inequality indices. Working Paper No. 1207E, Univ. Ottawa, Dept. Economics. Available at http://www.sciencessociales.uottawa.ca/sites/default/files/public/eco/fra/documents/1207E.pdf.

MECHANIC, D. (2002). Disadvantage, inequality, and social policy: Major initiatives intended to improve population health may also increase health disparities. *Health Aff.* **21** 48–59.

NAKAO, K. and TREAS, J. (1990). *Computing* 1989 *Occupational Prestige Scores*. *GSS Methodology Reports* **70**. NORC, Chicago, IL.

NCHS (2011). *Healthy People* 2010 *Final Review*. National Center for Health Statistics (NCHS), Hyattsville, MD.

O*NET (2014). Onetcenter.org. National Center for O*NET Development, Washington, DC. Available at http://www.onetcenter.org.

OGDEN, C. L., LAMB, M. M., CARROLL, M. D. and FLEGAL, K. M. (2010). Obesity and socioeconomic status in children and adolescents: United States, 2005–2008. *NCHS Data Brief* **51** 1–8.

OGDEN, C. L., CARROLL, M. D., KIT, B. K. and FLEGAL, K. M. (2012). Prevalence of obesity in the United States, 2009–2010. *NCHS Data Brief* **82** 1–8.

PAMUK, E. R. (1985). Social class inequality in mortality from 1921 to 1972 in England and Wales. *Popul. Stud.* **39** 17–31.

PAMUK, E. R. (1988). Social class inequality in infant mortality in England and Wales from 1921 to 1980. *European Journal of Population—Revue Européenne de Démographie* **4** 1–22.

PARUOLO, P., SAISANA, M. and SALTELLI, A. (2013). Ratings and rankings: Voodoo or science? *J. Roy. Statist. Soc. Ser. A* **176** 609–634. MR3067416

PETER, F. (2001). Health equity and social justice. *J. Appl. Philos.* **18** 159–170.

PRATT, J. W. (1964). Risk aversion in the small and the large. *Econometrica* **32** 122–136.

R DEVELOPMENT CORE TEAM (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RAO, J. N. K. and WU, C.-F. J. (1988). Resampling inference with complex survey data. *J. Amer. Statist. Assoc.* **83** 231–241. MR0941020

RAO, J. N. K., WU, C. F. J. and YUE, K. (1992). Some recent work in resampling methods. *Surv. Methodol.* **18** 209–217.

RAWLS, J. B. (1999). *A Theory of Justice*, Revised ed. Harvard Univ. Press, Cambridge, MA.

RÉNYI, A. (1961). On measures of entropy and information. In *Proc.* 4*th Berkeley Sympos. Math. Statist. and Prob.*, *Vol. I* 547–561. Univ. California Press, Berkeley, CA. MR0132570

ROSE, G. (1985). Sick individuals and sick populations. *Int. J. Epidemiol.* **14** 32–38.

ROSSEN, L. M. and TALIH, M. (2014). Social determinants of disparities in weight among US children and adolescents. *Ann. Epidemiol.* **24** 705–713.

RTI (2012). SUDAAN: Software for the Statistical Analysis of Correlated Data, Release 11. Research Triangle Institute (RTI), Research Triangle Park, NC.

SARAIYA, M., KING, J., THOMPSON, T., WATSON, M., AJANI, U., LI, J. and HOUSTON, K. A. (2013). Cervical cancer screening among women aged 18–30 years—United States, 2000–2010. *MMWR Morb. Mortal. Wkly. Rep.* **61** 1038–1042.

SAS INSTITUTE (2010). SAS Proprietary Software 9.3. SAS Institute Inc., Cary, NC.

SEER PROGRAM (2013). SEER*Stat Database: Incidence—SEER 18 Regs Research Data + Hurricane Katrina Impacted Louisiana Cases, Nov 2012 Sub (2000–2010) <Katrina/Rita Population Adjustment>—Linked to County Attributes—Total U.S., 1969–2011 Counties. National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, Bethesda, MD. Released April 2013, based on the November 2012 submission. Available at http://www.seer.cancer.gov.

SURVEILLANCE RESEARCH PROGRAM (2013). SEER*Stat Software 8.1.2. National Cancer Institute, Bethesda, MD.

TALIH, M. (2013a). Invited commentary: Can changes in the distributions of and associations between education and income bias estimates of temporal trends in health disparities? *Am. J. Epidemiol.* **177** 882–884.

TALIH, M. (2013b). A reference-invariant health disparity index based on Rényi divergence. *Ann. Appl. Stat.* **7** 1217–1243. MR3113507

TROIANO, R. P. and FLEGAL, K. M. (1998). Overweight children and adolescents: Description, epidemiology, and demographics. *Pediatrics* **101** 497–504.

TSUI, K.-Y. (1999). Multidimensional inequality and multidimensional generalized entropy measures: An axiomatic derivation. *Soc. Choice Welf.* **16** 145–157. MR1656440

WAGSTAFF, A. (2002). Inequality aversion, health inequalities and health achievement. *J. Health Econ.* **21** 627–641.

WAGSTAFF, A. (2011). The concentration index of a binary outcome revisited. *Health Econ.* **20** 1155–1160.

WAGSTAFF, A., PACI, P. and VAN DOORSLAER, E. (1991). On the measurement of inequalities in health. *Soc. Sci. Med.* **33** 545–557.

WHO-CSDH (2008). *Closing the Gap in a Generation*: *Health Equity Through Action on the Social Determinants of Health. Final Report of the Commission on Social Determinants of Health* (*CSDH*). World Health Organization, Geneva.

WILSON, J. (2009). Justice and the social determinants of health: An overview. *Public Health Ethics* **2** 210–213.

YIN, D., MORRIS, C., ALLEN, M., CRESS, R., BATES, J. and LIU, L. (2010). Does socioeconomic disparity in cancer incidence vary across racial/ethnic groups? *Cancer Causes Control* **21** 1721–1730.

YOUNG, L. JR., ROFFERS, S. D., RIES, L. A. G., FRITZ, A. G. and HURLBUT, A. A., eds. (2001). *SEER Summary Staging Manual*—2000: *Codes and Coding Instructions*. National Cancer Institute, Bethesda, MD.

The Annals of Applied Statistics
2015, Vol. 9, No. 2, 1024–1052
DOI: 10.1214/15-AOAS823

BAYESIAN STRUCTURED ADDITIVE DISTRIBUTIONAL REGRESSION WITH AN APPLICATION TO REGIONAL INCOME INEQUALITY IN GERMANY

BY NADJA KLEIN*,[1], THOMAS KNEIB*,[1],
STEFAN LANG†,[2] AND ALEXANDER SOHN*

Georg-August-University Göttingen and University of Innsbruck*†

We propose a generic Bayesian framework for inference in distributional regression models in which each parameter of a potentially complex response distribution and not only the mean is related to a structured additive predictor. The latter is composed additively of a variety of different functional effect types such as nonlinear effects, spatial effects, random coefficients, interaction surfaces or other (possibly nonstandard) basis function representations. To enforce specific properties of the functional effects such as smoothness, informative multivariate Gaussian priors are assigned to the basis function coefficients. Inference can then be based on computationally efficient Markov chain Monte Carlo simulation techniques where a generic procedure makes use of distribution-specific iteratively weighted least squares approximations to the full conditionals. The framework of distributional regression encompasses many special cases relevant for treating nonstandard response structures such as highly skewed nonnegative responses, overdispersed and zero-inflated counts or shares including the possibility for zero- and one-inflation. We discuss distributional regression along a study on determinants of labour incomes for full-time working males in Germany with a particular focus on regional differences after the German reunification. Controlling for age, education, work experience and local disparities, we estimate full conditional income distributions allowing us to study various distributional quantities such as moments, quantiles or inequality measures in a consistent manner in one joint model. Detailed guidance on practical aspects of model choice including the selection of several competing distributions for labour incomes and the consideration of different covariate effects on the income distribution complete the distributional regression analysis. We find that next to a lower expected income, full-time working men in East Germany also face a more unequal income distribution than men in the West, ceteris paribus.

REFERENCES

ARNOLD, B. C. (2008). Pareto and generalized Pareto distributions. In *Modeling Income Distributions and Lorenz Curves* (D. Chotikapanich, ed.) 119–145. Springer, New York.

ATKINSON, A. B. (1975). *The Economics of Inequality*. Clarendon Press, Oxford.

Key words and phrases. Generalised additive models for location, scale and shape, income distribution, iteratively weighted least squares proposal, Markov chain Monte Carlo, semiparametric regression, wage inequality.

AUTOR, D. H., KATZ, L. F. and KEARNEY, M. S. (2008). Trends in U.S. wage inequality: Revising the revisionists. *Rev. Econ. Stat.* **28** 300–323.

BACH, S., CORNEO, G. and STEINER, V. (2009). From bottom to top: The entire income distribution in Germany, 1992–2003. *Rev. Income Wealth* **55** 303–330.

BELITZ, C. and LANG, S. (2008). Simultaneous selection of variables and smoothing parameters in structured additive regression models. *Comput. Statist. Data Anal.* **53** 61–81. MR2528592

BELITZ, C., BREZGER, A., KLEIN, N., KNEIB, T., LANG, S. and UMLAUF, N. (2015). BayesX-software for Bayesian inference in structured additive regression models. Version 3.0. Available at http://www.bayesx.org.

BIEWEN, M. (2000). Income inequality in Germany during the 1980s and 1990s. *Rev. Income Wealth* **46** 1–19.

BIEWEN, M. and JENKINS, S. P. (2005). A framework for the decomposition of poverty differences with an application to poverty differences between countries. *Empir. Econ.* **30** 331–358.

BREZGER, A. and LANG, S. (2006). Generalized structured additive regression based on Bayesian P-splines. *Comput. Statist. Data Anal.* **50** 967–991. MR2210741

CARD, D. E., HEINING, J. and KLINE, P. (2013). Workplace heterogeneity and the rise of German wage inequality. *Q. J. Bus. Econ.* **128** 967–1015.

DAGUM, C. (1977). A new model of personal income distribution: Specification and estimation. *Economie Applicée* **30** 413–437.

DAGUM, C. (2008). A new model of personal income distribution: Specification and estimation. In *Modeling Income Distributions and Lorenz Curves* (D. Chotikapanich, ed.) 3–25. Springer, New York.

DINARDO, J., FORTIN, N. M. and LEMIEUX, T. (1996). Labor market institutions and the distribution of wages, 1973–1992: A semiparametric approach. *Econometrica* **64** 1001–1044.

DONALD, S. G., GREEN, D. A. and PAARSCH, H. J. (2000). Differences in wage distributions between Canada and the United States: An application of a flexible estimator of distribution functions in the presence of covariates. *Rev. Econ. Stud.* **67** 609–633.

DUCLOS, J.-Y., ESTEBAN, J. and RAY, D. (2004). Polarization: Concepts, measurement, estimation. *Econometrica* **72** 1737–1772. MR2095531

DUNN, P. K. and SMYTH, G. K. (1996). Randomized quantile residuals. *J. Comput. Graph. Statist.* **5** 236–245.

DUSTMANN, C., LUDSTECK, J. and SCHÖNBERG, U. (2009). Revisiting the German wage structure. *Q. J. Econ.* **124** 843–881.

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

FAHRMEIR, L., KNEIB, T. and LANG, S. (2004). Penalized structured additive regression for space-time data: A Bayesian perspective. *Statist. Sinica* **14** 731–761. MR2087971

FAHRMEIR, L., KNEIB, T., LANG, S. and MARX, B. (2013). *Regression: Models, Methods and Applications*. Springer, Heidelberg. MR3075546

FORTIN, N. M., LEMIEUX, T. and FIRPO, S. (2011). Decomposition methods in economics. In *Handbook of Labor Economics* (O. Ashenfelter and D. E. Card, eds.) **4A** 1–102. North-Holland, Amsterdam.

FUCHS-SCHÜNDELN, N., KRUEGER, D. and SOMMER, M. (2010). Inequality trends for Germany in the last two decades: A tale of two countries. *Rev. Econ. Dyn.* **13** 103–132.

GALVAO, A. F., LAMARCHE, C. and LIMA, L. R. (2013). Estimation of censored quantile regression for panel data with fixed effects. *J. Amer. Statist. Assoc.* **108** 1075–1089. MR3174685

GAMERMAN, D. (1997). Sampling from the posterior distribution in generalized linear mixed models. *Stat. Comput.* **7** 57–68.

GNEITING, T. (2011a). Making and evaluating point forecasts. *J. Amer. Statist. Assoc.* **106** 746–762. MR2847988

GNEITING, T. (2011b). Quantiles as optimal point forecasts. *Int. J. Forecast.* **27** 197–207.

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

GNEITING, T. and RANJAN, R. (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *J. Bus. Econom. Statist.* **29** 411–422. MR2848512

GRADÍN, C. (2000). Polarization by sub-populations in Spain, 1973–1991. *Rev. Income Wealth* **46** 457–474.

GREENE, W. H. (2008). *Econometric Analysis*, 6th ed. Pearson Prentice Hall, Upper Saddle River.

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

HELLER, G., STASINOPOULOS, D. and RIGBY, R. (2006). The zero-adjusted inverse Gaussian distribution as a model for insurance data. In *Proceedings of the* 21*th International Workshop on Statistical Modelling* (J. Hinde, J. Einbeck and J. Newell, eds.) 226–233.

KLASEN, S. (2008). The efficiency of equity. *Rev. Polit. Econ.* **20** 257–274.

KLEIBER, C. (1996). Dagum vs. Singh–Maddala income distributions. *Econom. Lett.* **57** 39–44.

KLEIBER, C. and KOTZ, S. (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*. Wiley, Hoboken, NJ. MR1994050

KLEIN, N., DENUIT, M., LANG, S. and KNEIB, T. (2014). Nonlife ratemaking and risk management with Bayesian generalized additive models for location, scale, and shape. *Insurance Math. Econom.* **55** 225–249. MR3179814

KLEIN, N., KNEIB, T. and LANG, S. (2015). Bayesian generalized additive models for location, scale and shape for zero-inflated and overdispersed count data. *J. Amer. Statist. Assoc.* **110** 405–419. MR3338512

KLEIN, N., KNEIB, T., KLASEN, S. and LANG, S. (2015a). Bayesian structured additive distributional regression for multivariate responses. *J. R. Stat. Soc. Ser. C. Appl. Stat.* To appear.

KLEIN, N., KNEIB, T., LANG, S. and SOHN, A. (2015b). Supplement to "Bayesian structured additive distributional regression with an application to regional income inequality in Germany." DOI:10.1214/15-AOAS823SUPPA.

KLEIN, N., KNEIB, T., LANG, S. and SOHN, A. (2015c). Supplement to "Bayesian structured additive distributional regression with an application to regional income inequality in Germany." DOI:10.1214/15-AOAS823SUPPB.

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

KOENKER, R. and BASSETT, G. JR. (1978). Regression quantiles. *Econometrica* **46** 33–50. MR0474644

KOHN, K. and ANTONCZYK, D. (2011). The aftermath of reunification: Sectoral transition, gender, and rising wage inequality in East Germany. IZA Discussion paper series No. 5708. Available at http://hdl.handle.net/10419/51717.

LAIO, F. and TAMEA, S. (2007). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrol. Earth Syst. Sci.* **11** 1267–1277.

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

LANG, S., UMLAUF, N., WECHSELBERGER, P., HARTTGEN, K. and KNEIB, T. (2014). Multilevel structured additive regression. *Stat. Comput.* **24** 223–238. MR3165550

LEMIEUX, T. (2006). The "Mincer equation." Thirty years after Schooling, Experience, and Earnings. In *Jacob Mincer*: *A Pioneer of Modern Labor Economics* (S. Grossbard, ed.) 127–145. Kluwer Academic, Boston.

MINCER, J. (1974). *Schooling*, *Experience*, *and Earnings*. Columbia Univ. Press, New York.

MISZTAL, B. (2013). *Trust in Modern Societies*: *The Search for the Bases of Social Order*. Wiley, Hoboken.

MORDUCH, J. and SICULAR, T. (2002). Rethinking inequality decomposition, with evidence from rural China. *Econ. J.* **112** 93–106.

NEWEY, W. K. and POWELL, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica* **55** 819–847. MR0906565

OSBAND, K. and REICHELSTEIN, S. (1985). Information-eliciting compensation schemes. *J. Public Econ.* **27** 107–115.

PIKETTY, T. and SAEZ, E. (2007). Income and wage inequality in the United States, 1913–2002. In *Top Incomes Over the Twentieth Century* (A. B. Atkinson and T. Piketty, eds.) 141–225. Oxford Univ. Press, Oxford.

PUDNEY, S. (1999). On some statistical methods for modelling the incidence of poverty. *Oxf. Bull. Econ. Stat.* **61** 385–408.

RIGBY, R. A. and STASINOPOULOS, D. M. (2005). Generalized additive models for location, scale and shape. *J. Roy. Statist. Soc. Ser. C* **54** 507–554. MR2137253

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

RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003). *Semiparametric Regression. Cambridge Series in Statistical and Probabilistic Mathematics* **12**. Cambridge Univ. Press, Cambridge. MR1998720

SALEM, A. B. Z. and MOUNT, T. D. (1974). A convenient descriptive model of income distribution. *Econometrica* **42** 1115–1127.

SARABIA, J. M. (2008). Parametric Lorenz curves: Models and applications. In *Modeling Iincome Distributions and Lorenz Curves* (D. Chotikapanich, ed.) 167–190. Springer, New York.

SCHEIPL, F., FAHRMEIR, L. and KNEIB, T. (2012). Spike-and-slab priors for function selection in structured additive regression models. *J. Amer. Statist. Assoc.* **107** 1518–1532. MR3036413

SCHNABEL, S. K. and EILERS, P. H. C. (2009). Optimal expectile smoothing. *Comput. Statist. Data Anal.* **53** 4168–4177. MR2744314

SILBER, J. (1999). Introduction—thirty years of intensive research on income inequality measurement. In *Handbook of Income Inequality Measurement* (J. Silber, ed.) 1–18. Kluwer Academic, Boston.

SKIDELSKY, R. (2010). *Keynes: The Return of the Master*, 1st ed. Public Affairs, New York.

SOBOTKA, F. and KNEIB, T. (2012). Geoadditive expectile regression. *Comput. Statist. Data Anal.* **56** 755–767. MR2888723

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

STATISTISCHES BUNDESAMT (2012). Verbraucherpreisindizes für Deutschland—Lange Reihen ab 1948, Preise.

UMLAUF, N., KLEIN, N., LANG, S. and ZEILEIS, A. (2014). bamlss: Bayesian additive models for location scale and shape (and beyond). R package Version 0.1-1. Available at http://bayesr.r-forge.r-project.org.

WAGNER, G. G., FRICK, J. R. and SCHUPP, J. (2007). The German socio-economic panel study (SOEP)—scope, evolution and enhancements. *Schmollers Jahrbuch* **127** 139–169.

WOLFSON, M. C. (1994). When inequalities diverge. *Am. Econ. Rev.* **84** 353–358.

WOOD, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *J. Amer. Statist. Assoc.* **99** 673–686. MR2090902

WOOD, S. N. (2008). Fast stable direct fitting and smoothness selection for generalized additive models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 495–518. MR2420412

YU, K. and MOYEED, R. A. (2001). Bayesian quantile regression. *Statist. Probab. Lett.* **54** 437–447. MR1861390

The Annals of Applied Statistics
2015, Vol. 9, No. 2, 1053–1075
DOI: 10.1214/15-AOAS825

SAMPLE SIZE DETERMINATION FOR TRAINING CANCER CLASSIFIERS FROM MICROARRAY AND RNA-seq DATA

BY SANDRA SAFO[1,2], XIAO SONG[3] AND KEVIN K. DOBBIN[1,2]

University of Georgia

The objective of many high-dimensional microarray and RNA-seq studies is to develop a classifier of cancer patients based on characteristics of their disease. The germinal center B-cell (GCB) classifier study in lymphoma and the National Cancer Institute's Director's Challenge lung (DC-lung) study are two examples. In recent years, such classifiers are often developed using regularized regression, such as the lasso. A critical question is whether a better classifier can be developed from a larger training set size and, if so, how large the training set should be. This paper examines these two questions using an existing sample size method and a novel sample size method developed here specifically for lasso logistic regression. Both methods are based on pilot data. We reexamine the lymphoma and lung cancer data sets to evaluate the sample sizes, and use resampling to assess the estimation methods. We also study application to an RNA-seq data set. We find that it is feasible to estimate sample size for regularized logistic regression if an adequate pilot data set exists. The GCB and the DC-lung data sets appear adequate, under specific assumptions. Existing human RNA-seq data sets are by and large inadequate, and cannot be used as pilot data. Pilot RNA-seq data can be simulated, and the methods in this paper can be used for sample size estimation. A MATLAB program is made available.

REFERENCES

AMBROISE, C. and MCLACHLAN, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci. USA* **14** 6562–6566.

BI, X., REXER, B., ARTEAGA, C. L., GUO, M. and MAHADEVAN-JANSEN, A. (2014). Evaluating HER2 amplification status and acquired drug resistance in breast cancer cells using Raman spectroscopy. *J. Biomed. Opt.* **19** 25001.

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

CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A. and CRAINICEANU, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd ed. *Monographs on Statistics and Applied Probability* **105**. Chapman & Hall/CRC, Boca Raton, FL. MR2243417

COOK, J. R. and STEFANSKI, L. A. (1994). Simulation-extrapolation estimation in parametric measurement errror models. *J. Amer. Statist. Assoc.* **89** 1314–1328.

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

Key words and phrases. Sample size, lasso, classification, regularized logistic regression, conditional score, high-dimensional data, measurement error.

DETTLING, M. and BÜHLMANN, P. (2003). Boosting for tumor classification with gene expression. *Bioinformatics* **19** 1061–1069.

DOBBIN, K. K. and SIMON, R. M. (2007). Sample size planning for developing classifiers using high-dimensional DNA microarray data. *Biostatistics* **8** 101–117.

DOBBIN, K. K. and SONG, X. (2013). Sample size requirements for training high-dimensional risk predictors. *Biostatistics* **14** 639–652.

DYRSKJØT, L. (2003). Classification of bladder cancer by microarray expression profiling: Towards a general clinical use of microarrays in cancer diagnostics. *Expert Rev. Mol. Diagn.* **3** 635–647.

EFRON, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *J. Amer. Statist. Assoc.* **70** 892–898. MR0391403

EFRON, B. and TIBSHIRANI, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *J. Amer. Statist. Assoc.* **92** 548–560. MR1467848

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

FRAZEE, A. C., LANGMEAD, B. and LEEK, J. T. (2011). ReCount: A multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics* **12** 449.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

GEISSER, S. (1993). *Predictive Inference: An Introduction*. Chapman & Hall, New York. MR1252174

GRAVELEY, B. R., BROOKS, A. N., CARLSON, J. W., DUFF, M. O., LANDOLIN, J. M., YANG, L., ARTIERI, C. G., VAN BAREN, M. J., BOLEY, N., BOOTH, B. W., BROWN, J. B., CHERBAS, L., DAVIS, C. A., DOBIN, A., LI, R., LIN, W., MALONE, J. H., MATTIUZZO, N. R., MILLER, D., STURGILL, D., TUCH, B. B., ZALESKI, C., ZHANG, D., BLANCHETTE, M., DUDOIT, S., EADS, B., GREEN, R. E., HAMMONDS, A., JIANG, L., KAPRANOV, P., LANGTON, L., PERRIMON, N., SANDLER, J. E., WAN, K. H., WILLINGHAM, A., ZHANG, Y., ZOU, Y., ANDREWS, J., BICKEL, P. J., BRENNER, S. E., BRENT, M. R., CHERBAS, P., GINGERAS, T. R., HOSKINS, R. A., KAUFMAN, T. C., OLIVER, B. and CELNIKER, S. E. (2011). The developmental transcriptome of *Drosophila melanogaster*. *Nature* **471** 473–479.

HANASH, S. M., BAIK, C. L. and KALLIONIEMI, O. (2011). Emerging molecular biomarkers—blood-based strategies to detect and monitor cancer. *Nat. Rev. Clin. Oncol.* **8** 142–150.

HANFELT, J. J. and LIANG, K.-Y. (1995). Approximate likelihood ratios for general estimating functions. *Biometrika* **82** 461–477. MR1366274

HANFELT, J. J. and LIANG, K.-Y. (1997). Approximate likelihoods for generalized linear errors-in-variables models. *J. Roy. Statist. Soc. Ser. B* **59** 627–637. MR1452030

HUANG, Y. and WANG, C. Y. (2000). Cox regression with accurate covariates unascertainable: A nonparametric-correction approach. *J. Amer. Statist. Assoc.* **95** 1209–1219. MR1804244

HUANG, Y. and WANG, C. Y. (2001). Consistent functional methods for logistic regression with errors in covariates. *J. Amer. Statist. Assoc.* **96** 1469–1482. MR1946591

MCSHANE, L. M. and HAYES, D. F. (2012). Publication of tumor marker research results: The necessity for complete and transparent reporting. *J. Clin. Oncol.* **30** 4223–4232.

MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2008). The group Lasso for logistic regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 53–71. MR2412631

MOEHLER, T. M., SECKINGER, A., HOSE, D., ANDRULIS, M., MOREAUX, J., HIELSCHER, T., WILLLHAUCK-FLECKENSTEIN, M., MERLING, A., BERTSCH, U., JAUCH, A., GOLDSCHMIDT, H., KLEIN, B. and SCHWARTZ-ALBIEZ, R. (2013). The glycome of normal and malignant plasma cells. *PLoS ONE* **8** e83719.

MUKHERJEE, S., TAMAYO, P., ROGERS, S., RIFKIN, R., ENGLE, A., CAMPBELL, C., GOLUB, T. R. and MESIROV, J. P. (2003). Estimating dataset size requirements for classifying DNA microarray data. *J. Comput. Biol.* **10** 119–142.

NOVICK, S. J. and STEFANSKI, L. A. (2002). Corrected score estimation via complex variable simulation extrapolation. *J. Amer. Statist. Assoc.* **97** 472–481. MR1941464

PFEFFER, U., ROMEO, F., NOONAN, D. M. and ALBINI, A. (2009). Predictin of breast cancer metastasis by genomic profiling: Where do we stand? *Clin. Exp. Metastasis* **26** 547–558.

ROSENWALD, A., WRIGHT, G., CHAN, W. C., CONNORS, J. M., CAMPO, E., FISHER, R. I., GASCOYNE, R. D., MULLER-HERMELINK, H. K., SMELAND, E. B., GILTNANE, J. M., HURT, E. M., ZHAO, H., AVERETT, L., YANG, L., WILSON, W. H., JAFFE, E. S., SIMON, R., KLAUSNER, R. D., POWELL, J., DUFFEY, P. L., LONGO, D. L., GREINER, T. C., WEISENBURGER, D. D., SANGER, W. G., DAVE, B. J., LYNCH, J. C., VOSE, J., ARMITAGE, J. O., MONTSERRAT, E., LÓPEZ-GUILLERMO, A., GROGAN, T. M., MILLER, T. P., LEBLANC, M., OTT, G., KVALOY, S., DELABIE, J., HOLTE, H., KRAJCI, P., STOKKE, T. and STAUDT, L. M. (LYMPHOMA/LEUKEMIA MOLECULAR PROFILING PROJECT) (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N. Engl. J. Med.* **346** 1937–1947.

SAFO, S., SONG, X. and DOBBIN, K. K. (2015). Supplement to "Sample size determination for training cancer classifiers from microarray and RNA-seq data." DOI:10.1214/15-AOAS825SUPP.

SHEDDEN, K., TAYLOR, J. M., ENKEMANN, S. A., TSAO, M. S., YEATMAN, T. J., GERALD, W. L., ESCHRICH, S., JURISICA, I., GIORDANO, T. J., MISEK, D. E., CHANG, A. C., ZHU, C. Q., STRUMPF, D., HANASH, S., SHEPHERD, F. A., DING, K., SEYMOUR, L., NAOKI, K., PENELL, N., WEIR, B., VERHAAK, R., LADD-ACOSTA, C., GOLUB, T., GRUIDL, M., SHARMA, A., SZOKE, J., ZAKOWSKI, M., RUSCH, V., KRIS, M., VIALE, A., MOTOI, N., TRAVIS, W., CONLEY, B., SESHAN, V. E., MEYERSON, M., KUICK, R., DOBBIN, K. K., LIVELY, T., JACOBSON, J. W. and BEER, D. G. (2008). Gene expression-based survival prediction in lung adenocarcinoma: A multisite, blinded validation study. *Nat. Med.* **14** 822–827.

SIMON, R. (2010). Clinical trials for predictive medicine: New challenges and paradigms. *Clin. Trials* **7** 516–524.

SIMON, R. M., RADMACHER, M. D., DOBBIN, K. K. and MCSHANE, L. M. (2003). Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J. Natl. Cancer Inst.* **95** 14–18.

STEFANSKI, L. A. and CARROLL, R. J. (1987). Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika* **74** 703–716. MR0919838

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

VARMA, S. and SIMON, R. M. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* **7** 91.

ZHANG, J. X., SONG, W., CHEN, Z. H., WEI, J. H., LIAO, Y. J., LEI, J., HU, M., CHEN, G. Z., LIAO, B., LU, J., ZHAO, H. W., CHEN, W., HE, Y. L., WANG, H. Y., XIE, D. and LUO, J. H. (2013). Prognostic and predictive value of a microRNA signature in stage II colon cancer: A microRNA expression analysis. *Lancet Oncol.* **14** 1295–1306.

ZHU, J. and HASTIE, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics* **5** 427–443.

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

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

ZWIENER, I., FRISCH, B. and BINDER, H. (2014). Transforming RNA-seq data to improve the performance of prognostic gene signatures. *PLoS ONE* **8** e85150.

The Annals of Applied Statistics
2015, Vol. 9, No. 2, 1076–1101
DOI: 10.1214/15-AOAS829

WAVELET-DOMAIN REGRESSION AND PREDICTIVE INFERENCE IN PSYCHIATRIC NEUROIMAGING

BY PHILIP T. REISS[1,*,†], LAN HUO[2,*], YIHONG ZHAO[*], CLARE KELLY[2,*] AND R. TODD OGDEN[3,‡]

New York University[*], *Nathan S. Kline Institute for Psychiatric Research*[†] *and Columbia University*[‡]

An increasingly important goal of psychiatry is the use of brain imaging data to develop predictive models. Here we present two contributions to statistical methodology for this purpose. First, we propose and compare a set of wavelet-domain procedures for fitting generalized linear models with scalar responses and image predictors: sparse variants of principal component regression and of partial least squares, and the elastic net. Second, we consider assessing the contribution of image predictors over and above available scalar predictors, in particular, via permutation tests and an extension of the idea of confounding to the case of functional or image predictors. Using the proposed methods, we assess whether maps of a spontaneous brain activity measure, derived from functional magnetic resonance imaging, can meaningfully predict presence or absence of attention deficit/hyperactivity disorder (ADHD). Our results shed light on the role of confounding in the surprising outcome of the recent ADHD-200 Global Competition, which challenged researchers to develop algorithms for automated image-based diagnosis of the disorder.

REFERENCES

ADHD-200 CONSORTIUM (2012). The ADHD-200 consortium: A model to advance the translational potential of neuroimaging in clinical neuroscience. *Front. Syst. Neurosci.* **6** 62.

BERLINET, A., BIAU, G. and ROUVIÈRE, L. (2008). Functional supervised classification with wavelets. *Ann. I.S.U.P.* **52** 61–80. MR2435041

BROWN, P. J., FEARN, T. and VANNUCCI, M. (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *J. Amer. Statist. Assoc.* **96** 398–408. MR1939343

BROWN, M. R. G., SIDHU, G. S., GREINER, R., ASGARIAN, N., BASTANI, M., SILVERSTONE, P. H., GREENSHAW, A. J. and DURSUN, S. M. (2012). ADHD-200 global competition: Diagnosing ADHD using personal characteristic data can outperform resting state fMRI measurements. *Front. Syst. Neurosci.* **6** 69.

CAFFO, B., ELOYAN, A., HAN, F., LIU, H., MUSCHELLI, J., NEBEL, M. B., ZHAO, T. and CRAINICEANU, C. (2012). SMART thoughts on the ADHD 200 Competition. Available at http://www.smart-stats.org/?q=content/repost-our-document-adhd-competition.

CAI, T. T. and HALL, P. (2006). Prediction in functional linear regression. *Ann. Statist.* **34** 2159–2179. MR2291496

Key words and phrases. ADHD-200, elastic net, functional confounding, functional magnetic resonance imaging, functional regression, sparse principal component regression, sparse partial least squares.

CARDOT, H., FERRATY, F. and SARDA, P. (1999). Functional linear model. *Statist. Probab. Lett.* **45** 11–22. MR1718346

CARDOT, H., FERRATY, F. and SARDA, P. (2003). Spline estimators for the functional linear model. *Statist. Sinica* **13** 571–591. MR1997162

CHANG, C., CHEN, Y. and OGDEN, R. T. (2014). Functional data classification: A wavelet approach. *Comput. Statist.* **29** 1497–1513.

CHI, E. C. and SCOTT, D. W. (2014). Robust parametric classification and variable selection by a minimum distance criterion. *J. Comput. Graph. Statist.* **23** 111–128.

CHUN, H. and KELEŞ, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 3–25. MR2751241

COOK, R. D. (2007). Fisher lecture: Dimension reduction in regression. *Statist. Sci.* **22** 1–26. MR2408655

CRADDOCK, R. C., HOLTZHEIMER III, P. E., HU, X. P. and MAYBERG, H. S. (2009). Disease state prediction from resting state functional connectivity. *Magn. Reson. Med.* **62** 1619–1628.

CRAINICEANU, C. M., REISS, P. T., GOLDSMITH, J., HUANG, L., HUO, L. and SCHEIPL, F. (2014). refund: Regression with functional data. R package version 0.1-10.

DAUBECHIES, I. (1988). Orthonormal bases of compactly supported wavelets. *Comm. Pure Appl. Math.* **41** 909–996. MR0951745

DELAIGLE, A. and HALL, P. (2012a). Achieving near perfect classification for functional data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 267–286. MR2899863

DELAIGLE, A. and HALL, P. (2012b). Methodology and theory for partial least squares applied to functional data. *Ann. Statist.* **40** 322–352. MR3014309

DING, B. and GENTLEMAN, R. (2005). Classification using generalized partial least squares. *J. Comput. Graph. Statist.* **14** 280–298. MR2160814

DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455. MR1311089

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

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

FAY, M. P., KIM, H.-J. and HACHEY, M. (2007). On using truncated sequential probability ratio test boundaries for Monte Carlo implementation of hypothesis tests. *J. Comput. Graph. Statist.* **16** 946–967. MR2412490

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

GOLDSMITH, J., HUANG, L. and CRAINICEANU, C. M. (2014). Smooth scalar-on-image regression via spatial Bayesian variable selection. *J. Comput. Graph. Statist.* **8** 1045–1064.

GOLDSMITH, J., BOBB, J., CRAINICEANU, C. M., CAFFO, B. and REICH, D. (2011). Penalized functional regression. *J. Comput. Graph. Statist.* **20** 830–851. MR2878950

GOLLAND, P. and FISCHL, B. (2003). Permutation tests for classification: Towards statistical significance in image-based studies. In *Information Processing in Medical Imaging: Proceedings of the* 18*th International Conference* (C. J. Taylor and J. A. Noble, eds.) 330–341. Springer, Berlin.

GROSENICK, L., KLINGENBERG, B., KATOVICH, K., KNUTSON, B. and TAYLOR, J. E. (2013). Interpretable whole-brain prediction analysis with GraphNet. *NeuroImage* **72** 304–321.

GUILLAS, S. and LAI, M.-J. (2010). Bivariate splines for spatial functional regression models. *J. Nonparametr. Stat.* **22** 477–497. MR2662608

HALL, P. and HOROWITZ, J. L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.* **35** 70–91. MR2332269

HONORIO, J., TOMASI, D., GOLDSTEIN, R. Z., LEUNG, H.-C. and SAMARAS, D. (2012). Can a single brain region predict a disorder? *IEEE Trans. Med. Imaging* **31** 2062–2072.

HUANG, L., GOLDSMITH, J., REISS, P. T., REICH, D. S. and CRAINICEANU, C. M. (2013). Bayesian scalar-on-image regression with application to association between intracranial DTI and cognitive outcomes. *NeuroImage* **83** 210–223.

HUO, L., REISS, P. and ZHAO, Y. (2014). refund.wave: Wavelet–domain regression with functional data. R package version 0.1.

JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693. MR2751448

KAPUR, S., PHILLIPS, A. G. and INSEL, T. R. (2012). Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Mol. Psychiatry* **17** 1174–1179.

MALLAT, S. G. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **11** 674–693.

MALLAT, S. (2009). *A Wavelet Tour of Signal Processing: The Sparse Way*, 3rd ed. Academic Press, Burlington, MA. MR2479996

MALLOY, E. J., MORRIS, J. S., ADAR, S. D., SUH, H., GOLD, D. R. and COULL, B. A. (2010). Wavelet-based functional linear mixed models: An application to measurement error-corrected distributed lag models. *Biostatistics* **11** 432–452.

MARX, B. D. (1996). Iteratively reweighted partial least squares estimation for generalized linear regression. *Technometrics* **38** 374–381.

MARX, B. D. and EILERS, P. H. C. (1999). Generalized linear regression on sampled signals and curves: A P-spline approach. *Technometrics* **41** 1–13.

MARX, B. D. and EILERS, P. H. C. (2005). Multidimensional penalized signal regression. *Technometrics* **47** 13–22. MR2135789

MASSY, W. F. (1965). Principal components regression in exploratory statistical research. *J. Amer. Statist. Assoc.* **60** 234–256.

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

MEINSHAUSEN, N., MEIER, L. and BÜHLMANN, P. (2009). p-values for high-dimensional regression. *J. Amer. Statist. Assoc.* **104** 1671–1681. MR2750584

MILHAM, M. P. (2012). Open neuroscience solutions for the connectome-wide association era. *Neuron* **73** 214–218.

MORRIS, J. S., BALADANDAYUTHAPANI, V., HERRICK, R. C., SANNA, P. and GUTSTEIN, H. (2011). Automated analysis of quantitative image data using isomorphic functional mixed models, with application to proteomics data. *Ann. Appl. Stat.* **5** 894–923. MR2840180

MÜLLER, H.-G. and STADTMÜLLER, U. (2005). Generalized functional linear models. *Ann. Statist.* **33** 774–805. MR2163159

NADLER, B. and COIFMAN, R. R. (2005). The prediction error in CLS and PLS: The importance of feature selection prior to multivariate calibration. *J. Chemom.* **19** 107–118.

NASON, G. P. (2008). *Wavelet Methods in Statistics with R*. Springer, New York. MR2445580

NASON, G. (2013). wavethresh: Wavelets statistics and transforms. R package version 4.6.2.

NGUYEN, D. V. and ROCKE, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* **18** 39–50.

NICHOLS, T. E. (2012). Multiple testing corrections, nonparametric methods, and random field theory. *NeuroImage* **62** 811–815.

OGDEN, R. T. (1997). *Essential Wavelets for Statistical Applications and Data Analysis*. Birkhäuser, Boston, MA. MR1420193

OJALA, M. and GARRIGA, G. C. (2010). Permutation tests for studying classifier performance. *J. Mach. Learn. Res.* **11** 1833–1863. MR2660654

ÖZKAYA, S. G. and VAN DE VILLE, D. (2011). Anatomically adapted wavelets for integrated statistical analysis of fMRI data. In 2011 *IEEE International Symposium on Biomedical Imaging*: *From Nano to Macro* 469–472. IEEE, New York.

POTTER, D. M. (2005). A permutation test for inference in logistic regression with small- and moderate-sized data sets. *Stat*. *Med*. **24** 693–708. MR2134534

PREDA, C. and SAPORTA, G. (2005). PLS regression on a stochastic process. *Comput*. *Statist*. *Data Anal*. **48** 149–158. MR2134488

R DEVELOPMENT CORE TEAM (2012). *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

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

RASMUSSEN, P. M., HANSEN, L. K., MADSEN, K. H., CHURCHILL, N. W. and STROTHER, S. C. (2012). Model sparsity and brain pattern interpretation of classification models in neuroimaging. *Pattern Recognition* **45** 2085–2100.

REISS, P. T. (2006). Regression with signals and images as predictors. Ph.D. Thesis, Dept. of Biostatistics, Columbia Univ., New York.

REISS, P. T. (2015). Cross-validation and hypothesis testing in neuroimaging: An irenic comment on the exchange between Friston and Lindquist et al. *NeuroImage*. To appear.

REISS, P. T. and OGDEN, R. T. (2007). Functional principal component regression and functional partial least squares. *J*. *Amer. Statist*. *Assoc*. **102** 984–996. MR2411660

REISS, P. T. and OGDEN, R. T. (2010). Functional generalized linear models with images as predictors. *Biometrics* **66** 61–69. MR2756691

REISS, P. T., HUO, L., ZHAO, Y., KELLY, C. and OGDEN, R. T. (2015). Supplement to "Wavelet-domain regression and predictive inference in psychiatric neuroimaging." DOI:10.1214/15-AOAS829SUPP.

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

ROTHMAN, K. J. (2012). *Epidemiology*: *An Introduction*, 2nd ed. Oxford Univ. Press, New York.

RUTTIMANN, U. E., UNSER, M., RAWLINGS, R. R., RIO, D., RAMSEY, N. F., MATTAY, V. S., HOMMER, D. W., FRANK, J. A. and WEINBERGER, D. R. (1998). Statistical analysis of functional MRI data in the wavelet domain. *IEEE Trans*. *Med*. *Imaging* **17** 142–154.

SABUNCU, M. R., VAN LEEMPUT, K. and ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2012). The relevance voxel machine (RVoxM): A self-tuning Bayesian model for informative image-based prediction. *IEEE Trans*. *Med*. *Imaging* **31** 2290–2306.

SHEN, H. and HUANG, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *J*. *Multivariate Anal*. **99** 1015–1034. MR2419336

STONE, M. and BROOKS, R. J. (1990). Continuum regression: Cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *J*. *Roy. Statist*. *Soc*. *Ser. B* **52** 237–269. MR1064418

SUN, D., VAN ERP, T. G. M., THOMPSON, P. M., BEARDEN, C. E., DALEY, M., KUSHAN, L., HARDT, M. E., NUECHTERLEIN, K. H., TOGA, A. W. and CANNON, T. D. (2009). Elucidating a magnetic resonance imaging-based neuroanatomic biomarker for psychosis: Classification analysis using probabilistic brain atlas and machine learning algorithms. *Biological Psychiatry* **66** 1055–1060.

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

VAN DE VILLE, D., SEGHIER, M. L., LAZEYRAS, F., BLU, T. and UNSER, M. (2007). WSPM: Wavelet-based statistical parametric mapping. *NeuroImage* **37** 1205–1217.

VIDAKOVIC, B. (1999). *Statistical Modeling by Wavelets*. Wiley, New York. MR1681904

WAND, M. P. and ORMEROD, J. T. (2011). Penalized wavelets: Embedding wavelets into semiparametric regression. *Electron*. *J*. *Stat*. **5** 1654–1717. MR2870147

WANG, X., RAY, S. and MALLICK, B. K. (2007). Bayesian curve classification using wavelets. *J. Amer. Statist. Assoc.* **102** 962–973. MR2354408

WANG, X., NAN, B., ZHU, J., KOEPPE, R. and THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2014). Regularized 3D functional regression for brain image data via Haar wavelets. *Ann. Appl. Stat.* **8** 1045–1064.

WITTEN, D. M., TIBSHIRANI, R. and HASTIE, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10** 515–534.

WOLD, H. (1966). Nonlinear estimation by iterative least square procedures. In *Research Papers in Statistics* (*Festschrift J. Neyman*) (F. N. David, ed.) 411–444. Wiley, London. MR0210250

YANG, H., WU, Q.-Z., GUO, L.-T., LI, Q.-Q., LONG, X.-Y., HUANG, X.-Q., CHAN, R. C. K. and GONG, Q.-Y. (2011). Abnormal spontaneous brain activity in medication-naïve ADHD children: A resting state fMRI study. *Neurosci. Lett.* **502** 89–93.

ZHAO, Y., CHEN, H. and OGDEN, R. T. (2015). Wavelet-based weighted LASSO and screening approaches in functional linear regression. *J. Comput. Graph. Statist.* To appear.

ZHAO, Y., OGDEN, R. T. and REISS, P. T. (2012). Wavelet-based LASSO in functional linear regression. *J. Comput. Graph. Statist.* **21** 600–617. MR2970910

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

ZHU, H., BROWN, P. J. and MORRIS, J. S. (2012). Robust classification of functional and quantitative image data using functional mixed models. *Biometrics* **68** 1260–1268. MR3040032

ZHU, J. and HASTIE, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics* **5** 427–443.

ZHU, H., VANNUCCI, M. and COX, D. D. (2010). A Bayesian hierarchical model for classification with selection of functional predictors. *Biometrics* **66** 463–473. MR2758826

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

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

ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.* **15** 265–286. MR2252527

ZOU, Q.-H., ZHU, C.-Z., YANG, Y., ZUO, X.-N., LONG, X.-Y., CAO, Q.-J., WANG, Y.-F. and ZANG, Y.-F. (2008). An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: Fractional ALFF. *J. Neurosci. Methods* **172** 137–141.