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

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE INSTITUTE OF MATHEMATICAL STATISTICS

Articles

continued

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

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles—*continued*

The Annals of Applied Statistics
2014, Vol. 8, No. 1, 1–18
DOI: 10.1214/13-AOAS703

STOCHASTIC IDENTIFICATION OF MALWARE WITH DYNAMIC TRACES

BY CURTIS STORLIE*, BLAKE ANDERSON*, SCOTT VANDER WIEL*, DANIEL QUIST†, CURTIS HASH* AND NATHAN BROWN‡

Los Alamos National Laboratory, Bechtel Corporation*† *and Naval Postgraduate School*‡

A novel approach to malware classification is introduced based on analysis of instruction traces that are collected dynamically from the program in question. The method has been implemented online in a sandbox environment (i.e., a security mechanism for separating running programs) at Los Alamos National Laboratory, and is intended for eventual host-based use, provided the issue of sampling the instructions executed by a given process without disruption to the user can be satisfactorily addressed. The procedure represents an instruction trace with a Markov chain structure in which the transition matrix, **P**, has rows modeled as Dirichlet vectors. The malware class (malicious or benign) is modeled using a flexible spline logistic regression model with variable selection on the elements of **P**, which are observed with error. The utility of the method is illustrated on a sample of traces from malware and nonmalware programs, and the results are compared to other leading detection schemes (both signature and classification based). This article also has supplementary materials available online.

REFERENCES

ANDERSON, B., QUIST, D., NEIL, J., STORLIE, C. and LANE, T. (2011). Graph-based malware detection using dynamic analysis. *Journal in Computer Virology* **7** 247–258.

ANDERSON, B., QUIST, D., BROWN, N., STORLIE, C. and LANE, T. (2012). Improving malware classification: Bridging the static/dynamic gap. In *Proceedings of the 5th ACM Workshop on Security and Artificial Intelligence* 3–14. ACM, New York.

ANTIVIRUS COMPARATIVES (2011). Retrospective test (static detection of new/unknown malicious software). Available at http://www.av-comparatives.org/images/stories/test/ondret/avc_retro_nov2011.pdf.

BAYER, U., MOSER, A., KRUEGEL, C. and KIRDA, E. (2006). Dynamic analysis of malicious code. *Journal in Computer Virology* **2** 67–77.

BILAR, D. (2007). Opcodes as predictor for malware. *International Journal of Electronic Security and Digital Forensics* **1** 156–168.

CHRISTODORESCU, M. and JHA, S. (2003). Static analysis of executables to detect malicious patterns. In *Proceedings of the 12th USENIX Security Symposium* 169–186. USENIX Association, Berkeley, CA.

COVA, M., KRUEGEL, C. and VIGNA, G. (2010). Detection and analysis of drive-by-download attacks and malicious javascript code. In *Proceedings of the 19th International Conference on World Wide Web* 281–290. ACM, New York.

Key words and phrases. Malware detection, classification, elastic net, Relaxed Lasso, Adaptive Lasso, logistic regression, splines, empirical Bayes.

DAI, J., GUHA, R. and LEE, J. (2009). Efficient virus detection using dynamic instruction sequences. *Journal of Computers* **4** 405–414.

DINABURG, A., ROYAL, P., SHARIF, M. and LEE, W. (2008). Ether: Malware analysis via hardware virtualization extensions. In *Proceedings of the 15th ACM Conference on Computer and Communications Security* 51–62. ACM, New York.

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166

GOLDBERG, I., WAGNER, D., THOMAS, R. and BREWER, E. (1996). A secure environment for untrusted helper applications (confining the wily hacker). In *Proceedings of the Sixth USENIX UNIX Security Symposium* **6** 1. USENIX Association, Berkeley, CA.

GRAMACY, R. B. and POLSON, N. G. (2012). Simulation-based regularized logistic regression. *Bayesian Anal.* **7** 567–589. MR2981628

HASTIE, T. and TIBSHIRANI, R. (1996). Discriminant analysis by Gaussian mixtures. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58** 155–176. MR1379236

HOFMEYR, S. A., FORREST, S. and SOMAYAJI, A. (1998). Intrusion detection using sequences of system calls. *Journal of Computer Security* **6** 151–180.

KING, G. and ZENG, L. (2001). Logistic regression in rare events data. *Political Analysis* **9** 137–163.

KOLTER, J. Z. and MALOOF, M. A. (2006). Learning to detect and classify malicious executables in the wild. *J. Mach. Learn. Res.* **7** 2721–2744. MR2274458

LUK, C.-K., COHN, R., MUTH, R., PATIL, H., KLAUSER, A., LOWNEY, G., WALLACE, S., REDDI, V. J. and HAZELWOOD, K. (2005). Pin: Building customized program analysis tools with dynamic instrumentation. In *Proceedings of the ACM SIGPLAN Conference on Programming Language Design and Implementation* 190–200. ACM, New York.

MANSKI, C. F. and LERMAN, S. R. (1977). The estimation of choice probabilities from choice based samples. *Econometrica* **45** 1977–1988. MR0501708

MEINSHAUSEN, N. (2007). Relaxed Lasso. *Comput. Statist. Data Anal.* **52** 374–393. MR2409990

PANDALABS (2012). PandaLabs quarterly report. Available at http://press.pandasecurity.com/wp-content/uploads/2012/08/Quarterly-Report-PandaLabs-April-June-2012.pdf.

PERDISCI, R., DAGON, D., FOGLA, P. and SHARIF, M. (2006). Misleading worm signature generators using deliberate noise injection. In *Proceedings of the IEEE Symposium on Security and Privacy* 17–31. IEEE Computer Society Technical Committee on Security and Privacy.

PRENTICE, R. L. and PYKE, R. (1979). Logistic disease incidence models and case–control studies. *Biometrika* **66** 403–411. MR0556730

QUIST, D. (2012). Community malicious code research and analysis. Available at http://www.offensivecomputing.net/.

REDDY, D. K. S., DASH, S. and PUJARI, A. (2006). New malicious code detection using variable length n-grams. In *Information Systems Security. Lecture Notes in Computer Science* **4332** 276–288. Springer, Berlin.

REDDY, D. and PUJARI, A. (2006). N-gram analysis for computer virus detection. *Journal in Computer Virology* **2** 231–239.

RIECK, K., TRINIUS, P., WILLEMS, C. and HOLZ, T. (2011). Automatic analysis of malware behavior using machine learning. *Journal of Computer Security* **19** 639–668.

ROYAL, P., HALPIN, M., DAGON, D., EDMONDS, R. and LEE, W. (2006). Polyunpack: Automating the hidden-code extraction of unpackexecuting malware. In *Proceedings of the 22nd Annual Computer Security Applications Conference* 289–300.

SHAFIQ, M., KHAYAM, S. and FAROOQ, M. (2008). Embedded malware detection using Markov n-grams. In *Proceedings of the 5th International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment* 88–107. ACM, New York.

SHANKARAPANI, M., RAMAMOORTHY, S., MOVVA, R. and MUKKAMALA, S. (2010). Malware detection using assembly and API call sequences. *Journal in Computer Virology* **7** 1–13.

SKALETSKY, A., DEVOR, T., CHACHMON, N., COHN, R., HAZELWOOD, K., VLADIMIROV, V. and BACH, M. (2010). Dynamic program analysis of Microsoft Windows applications. In *2010 International Symposium on Performance Analysis of Software and Systems* (*ISPASS*) 2–12. IEEE Computer Society's Technical Committee on the Internet.

STOLFO, S., WANG, K. and LI, W.-J. (2007). Towards stealthy malware detection. In *Malware Detection*. *Advances in Information Security* **27** 231–249. Springer, New York.

STORLIE, C., ANDERSON, B., VANDER WIEL, S., QUIST, D., HASH, C. and BROWN, N. (2014). Supplement to "Stochastic identification of malware with dynamic traces." DOI:10.1214/13-AOAS703SUPP.

SYMANTEC (2008). Internet security threat report, trends for July–December 2007 (executive summary). White paper. Available at http://eval.symantec.com/mktginfo/enterprise/white_papers/b-whitepaper_exec_summary_internet_security_threat_report_xiii_04-2008.en-us.pdf.

SYMANTEC (2011). Internet security threat report, volume 16. White paper. Available at http://www.symantec.com/business/threatreport/index.jsp.

TADDY, M. (2013). Multinomial inverse regression for text analysis. *J. Amer. Statist. Assoc.* **108** 755–770.

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

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
2014, Vol. 8, No. 1, 19–47
DOI: 10.1214/13-AOAS685

HIERARCHICAL ARRAY PRIORS FOR ANOVA DECOMPOSITIONS OF CROSS-CLASSIFIED DATA

BY ALEXANDER VOLFOVSKY AND PETER D. HOFF

Harvard University and University of Washington

ANOVA decompositions are a standard method for describing and estimating heterogeneity among the means of a response variable across levels of multiple categorical factors. In such a decomposition, the complete set of main effects and interaction terms can be viewed as a collection of vectors, matrices and arrays that share various index sets defined by the factor levels. For many types of categorical factors, it is plausible that an ANOVA decomposition exhibits some consistency across orders of effects, in that the levels of a factor that have similar main-effect coefficients may also have similar coefficients in higher-order interaction terms. In such a case, estimation of the higher-order interactions should be improved by borrowing information from the main effects and lower-order interactions. To take advantage of such patterns, this article introduces a class of hierarchical prior distributions for collections of interaction arrays that can adapt to the presence of such interactions. These prior distributions are based on a type of array-variate normal distribution, for which a covariance matrix for each factor is estimated. This prior is able to adapt to potential similarities among the levels of a factor, and incorporate any such information into the estimation of the effects in which the factor appears. In the presence of such similarities, this prior is able to borrow information from well-estimated main effects and lower-order interactions to assist in the estimation of higher-order terms for which data information is limited.

REFERENCES

ALBRINK, M. J. and ULLRICH, I. H. (1986). Interaction of dietary sucrose and fiber on serum lipids in healthy young men fed high carbohydrate diets. *Am. J. Clin. Nutr.* **43** 419–428.

AUSTIN, G. L., OGDEN, L. G. and HILL, J. O. (2011). Trends in carbohydrate, fat, and protein intakes and association with energy intake in normal-weight, overweight, and obese individuals: 1971–2006. *Am. J. Clin. Nutr.* **93** 836–843.

BASIOTIS, P. P., THOMAS, R. G., KELSAY, J. L. and MERTZ, W. (1989). Sources of variation in energy intake by men and women as determined from one year's daily dietary records. *Am. J. Clin. Nutr.* **50** 448–453.

BERAN, R. (2005). ASP fits to multi-way layouts. *Ann. Inst. Statist. Math.* **57** 201–220. MR2160647

BERGER, J. O. and YANG, R.-Y. (1994). Noninformative priors and Bayesian testing for the AR(1) model. *Econometric Theory* **10** 461–482. MR1309107

CHANDALIA, M., GARG, A., LUTJOHANN, D., VON BERGMANN, K., GRUNDY, S. M. and BRINKLEY, L. J. (2000). Beneficial effects of high dietary fiber intake in patients with type 2 diabetes mellitus. *N. Engl. J. Med.* **342** 1392–1398.

Key words and phrases. Array-valued data, Bayesian estimation, cross-classified data, factorial design, MANOVA, penalized regression, tensor, Tucker product, sparse data.

CUI, Y., HODGES, J. S., KONG, X. and CARLIN, B. P. (2010). Partitioning degrees of freedom in hierarchical and other richly parameterized models. *Technometrics* **52** 124–136. MR2752111

DAWID, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika* **68** 265–274. MR0614963

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

GELMAN, A. (2005). Analysis of variance—Why it is more important than ever. *Ann. Statist.* **33** 1–53. MR2157795

GELMAN, A. and HILL, J. (2007). Data analysis using regression and multilevel hierarchical models. Unpublished manuscript.

GENKIN, A., LEWIS, D. D. and MADIGAN, D. (2007). Large-scale Bayesian logistic regression for text categorization. *Technometrics* **49** 291–304. MR2408634

HODGES, J. S., SARGENT, D. J., CUI, Y. and CARLIN, B. P. (2007). Smoothing balanced single-error-term analysis of variance. *Technometrics* **49** 12–25. MR2345448

HOFF, P. D. (2011). Separable covariance arrays via the Tucker product, with applications to multivariate relational data. *Bayesian Anal.* **6** 179–196. MR2806238

JOHANSSON, G., WIKMAN, A., AHREN, A. M., HALLMANS, G. and JOHANSSON, I. et al. (2001). Underreporting of energy intake in repeated 24-hour recalls related to gender, age, weight status, day of interview, educational level, reported food intake, smoking habits and area of living. *Public Health Nutrition* **4** 919–928.

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

KOLDA, T. G. and BADER, B. W. (2009). Tensor decompositions and applications. *SIAM Rev.* **51** 455–500. MR2535056

KRUSCHKE, J. (2011). *Doing Bayesian Data Analysis*: *A Tutorial Introduction with R and BUGS*. Academic Press, Boston, MA.

MILLER, R. and BROWN, B. (1997). *Beyond ANOVA*: *Basics of Applied Statistics*. Chapman & Hall/CRC, New York.

MOERMAN, C., DE MESQUITA, H. and RUNIA, S. (1993). Dietary sugar intake in the aetiology of biliary tract cancer. *International Journal of Epidemiology* **22** 207–214.

MONTONEN, J., KNEKT, P., JÄRVINEN, R., AROMAA, A. and REUNANEN, A. (2003). Whole-grain and fiber intake and the incidence of type 2 diabetes. *Am. J. Clin. Nutr.* **77** 622–629.

NIELSEN, S. J. and POPKIN, B. M. (2004). Changes in beverage intake between 1977 and 2001. *Am. J. Prev. Med.* **27** 205–210.

OLSON, C. (1976). On choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin* **83** 579.

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

PARK, D., GELMAN, A. and BAFUMI, J. (2006). State level opinions from national surveys: Post-stratification using multilevel logistic regression. In *Public Opinion in State Politics* 209–228. Stanford Univ. Press, Stanford, CA.

PARK, Y., SUBAR, A. F., HOLLENBECK, A. and SCHATZKIN, A. (2011). Dietary fiber intake and mortality in the NIH-AARP diet and health study. *Arch. Intern. Med.* **171** 1061–1068.

PITTAU, M., ZELLI, R. and GELMAN, A. (2010). Economic disparities and life satisfaction in European regions. *Social Indicators Research* **96** 339–361.

USDA. (2010). Food and Nutrient Database for Dietary Studies 4.1. U.S. Dept. Agriculture, Agricultural Research Service, Food Surveys Research Group, Beltsville, MD.

VERLY JUNIOR, E., FISBERG, R. M., CESAR, C. L. G. and MARCHIONI, D. M. L. (2010). Sources of variation of energy and nutrient intake among adolescents in São Paulo. *Brazil. Cadernos de Saúde Pública* **26** 2129–2137.

VOLFOVSKY, A. and HOFF, P. (2013). Supplement to "Hierarchical array priors for ANOVA decompositions of cross-classified data." DOI:10.1214/13-AOAS685SUPP.

YANG, E. J., CHUNG, H. K., KIM, W. Y., KERVER, J. M. and SONG, W. O. (2003). Carbohydrate intake is associated with diet quality and risk factors for cardiovascular disease in us adults: Nhanes iii. *Journal of the American College of Nutrition* **22** 71–79.

YUAN, M. and LIN, Y. (2005). Efficient empirical Bayes variable selection and estimation in linear models. *J. Amer. Statist. Assoc.* **100** 1215–1225. MR2236436

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

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

The Annals of Applied Statistics
2014, Vol. 8, No. 1, 48–73
DOI: 10.1214/13-AOAS699

ESTIMATION OF CAUSAL EFFECTS USING INSTRUMENTAL VARIABLES WITH NONIGNORABLE MISSING COVARIATES: APPLICATION TO EFFECT OF TYPE OF DELIVERY NICU ON PREMATURE INFANTS

BY FAN YANG*, SCOTT A. LORCH† AND DYLAN S. SMALL*

*University of Pennsylvania** *and The Children's Hospital of Philadelphia*†

Understanding how effective high-level NICUs (neonatal intensive care units that have the capacity for sustained mechanical assisted ventilation and high volume) are compared to low-level NICUs is important and valuable for both individual mothers and for public policy decisions. The goal of this paper is to estimate the effect on mortality of premature babies being delivered in a high-level NICU vs. a low-level NICU through an observational study where there are unmeasured confounders as well as nonignorable missing covariates. We consider the use of excess travel time as an instrumental variable (IV) to control for unmeasured confounders. In order for an IV to be valid, we must condition on confounders of the IV—outcome relationship, for example, month prenatal care started must be conditioned on for excess travel time to be a valid IV. However, sometimes month prenatal care started is missing, and the missingness may be nonignorable because it is related to the not fully measured mother's/infant's risk of complications. We develop a method to estimate the causal effect of a treatment using an IV when there are nonignorable missing covariates as in our data, where we allow the missingness to depend on the fully observed outcome as well as the partially observed compliance class, which is a proxy for the unmeasured risk of complications. A simulation study shows that under our nonignorable missingness assumption, the commonly used estimation methods, complete-case analysis and multiple imputation by chained equations assuming missingness at random, provide biased estimates, while our method provides approximately unbiased estimates. We apply our method to the NICU study and find evidence that high-level NICUs significantly reduce deaths for babies of small gestational age, whereas for almost mature babies like 37 weeks, the level of NICUs makes little difference. A sensitivity analysis is conducted to assess the sensitivity of our conclusions to key assumptions about the missing covariates. The method we develop in this paper may be useful for many observational studies facing similar issues of unmeasured confounders and nonignorable missing data as ours.

REFERENCES

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

Key words and phrases. Instrumental variable, causal inference, sensitivity analysis, nonignorable missing data.

ANGRIST, J. D. and KRUEGER, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics* **106** 979–1014.

BAIOCCHI, M., SMALL, D. S., LORCH, S. and ROSENBAUM, P. R. (2010). Building a stronger instrument in an observational study of perinatal care for premature infants. *J. Amer. Statist. Assoc.* **105** 1285–1296. MR2796550

BOYLE, M. H., TORRANCE, G. W., SINCLAIR, J. C. and HORWOOD, S. P. (1983). Economic evaluation of neonatal intensive care of very-low-birth-weight infants. *N. Engl. J. Med.* **308** 1330–13337.

BROOKHART, M. A. and SCHNEEWEISS, S. (2007). Preference-based instrumental variable methods for the estimation of treatment effects: Assessing validity and interpreting results. *Int. J. Biostat.* **3** Art. 14, 25. MR2383610

BROOKHART, M. A., WANG, P. S., SOLOMON, D. H. and SCHNEEWEISS, S. (2006). Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. *Epidemiology* **17** 268–275.

CHEN, H., GENG, Z. and ZHOU, X.-H. (2009). Identifiability and estimation of causal effects in randomized trials with noncompliance and completely nonignorable missing data. *Biometrics* **65** 675–682. MR2649840

CHUNG, J. H., PHIBBS, C. S., BOSCARDIN, W. J., KOMINSKI, G. F., ORTEGA, A. N. and NEEDLEMAN, J. (2010). The effect of neonatal intensive care level and hospital volume on mortality of very low birth weight infants. *Med. Care* **48** 635–644.

DOYLE, L. W. and VICTORIAN INFANT COLLABORATIVE STUDY GROUP (2004). Evaluation of neonatal intensive care for extremely low birth weight infants in Victoria over two decades: II. Efficiency. *Pediatrics* **113** 510–514.

FRANGAKIS, C. E. and RUBIN, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika* **86** 365–379. MR1705410

GUO, Z., CHENG, J., LORCH, S. A. and SMALL, D. S. (2014). Using an instrumental variable to test for unmeasured confounding. Preprint.

HOWELL, E. M., RICHARDSON, D., GINSBURG, P. and FOOT, B. (2002). Deregionalization of neonatal intensive care in urban areas. *Am. J. Publ. Health* **92** 119–124.

IMAI, K. and RATKOVIC, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *Ann. Appl. Stat.* **7** 443–470. MR3086426

KORN, E. L. and BAUMRIND, S. (1998). Clinician preferences and the estimation of causal treatment differences. *Statist. Sci.* **13** 209–235. MR1665709

LASSWELL, S. M., BARFIELD, W. D., ROCHAT, R. W. and BLACKMON, L. (2010). Perinatal regionalization for very low-birth-weight and very preterm infants: A meta-analysis. *J. Am. Med. Assoc.* **304** 992–1000.

LEVY, D. E., O'MALLEY, A. J. and NORMAND, S. T. (2004). Covariate adjustment in clinical trials with nonignorable missing data and noncompliance. *Stat. Med.* **23** 2319–2339.

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

LORCH, S. A., BAIOCCHI, M., AHLBERG, C. E. and SMALL, D. S. (2012). The differential impact of delivery NICU on the outcomes of premature infance. *Pediatrics* **130** 1–9.

MEALLI, F., IMBENS, G., FERRO, S. and BIGGERI, A. (2004). Analyzing a randomized trial on breast self examination with noncompliance and missing outcomes. *Biostatistics* **5** 207–222.

OLKIN, I. and TATE, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Ann. Inst. Statist. Math.* **32** 448–465. MR0152062

PENG, Y., LITTLE, R. J. A. and RAGHUNATHAN, T. E. (2004). An extended general location model for causal inferences from data subject to noncompliance and missing values. *Biometrics* **60** 598–607. MR2089434

PHIBBS, C. S., MARK, D. H., LUFT, H. S. et al. (1993). Choice of hospital for delivery: A comparison of high-risk and low-risk women. *Health Serv. Res.* **28** 201–222.

PHIBBS, C. S., BAKER, L. C., CAUGHEY, A. B., DANIELSEN, B., SCHMITT, S. K. and PHIBBS, R. H. (2007). Level and volume of neonatal intensive care and mortality in very-low-birth-weight infants. *N. Engl. J. Med.* **356** 2165–2175.

PROFIT, J., LEE, D., ZUPANCIC, J. A., PAPILE, L., GUTIERREZ, C., GOLDIE, S. J., GONZALEZ-PIER, E. and SALOMON, J. A. (2010). Clinical benefits, costs, and cost-effectiveness of neonatal intensive care in Mexico. *PLoS Medicine* **7** 1–10.

RICHARDSON, D. K., REED, K., CUTLER, J. C. et al. (1995). Perinatal regionalization vs hospital competition: The Hartford example. *Pediatrics* **96** 417–423.

ROGOWSKI, J. A., HORBAR, J. D., STAIGER, D. O., KENNY, M., CARPENTER, J. and GEPPERT, J. (2004). Indirect vs direct hospital quality indicators for very low-birth-weight infants. *J. Am. Med. Assoc.* **291** 202–209.

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

ROSENBAUM, P. R. and RUBIN, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *J. Amer. Statist. Assoc.* **79** 516–524.

ROY, J. and HENNESSY, S. (2011). Bayesian hierarchical pattern mixture models for comparative effectiveness of drugs and drug classes using healthcare data: A case study involving antihypertensive medications. *Statistics in Biosciences* **3** 79–93.

SCHAFER, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London. MR1692799

SMALL, D. S. and CHENG, J. (2009). Discussions of "Identifiability and estimation of causal effects in randomized trials with noncompliance and completely nonignorable missing data." *Biometrics* **65** 682–686. MR2766612

VAN BUUREN, S. and GROOTHUIS-OUDSHOORN, K. (2011). mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **45** 1–67.

WALKER, A. (1996). Confounding by indication. *Epidemiology* **7** 335–336.

YANG, F., LORCH, S. A. and SMALL, D. S. (2014). Supplement to "Estimation of causal effects using instrumental variables with nonignorable missing covariates: Application to effect of type of delivery NICU on premature infants." DOI:10.1214/13-AOAS699SUPP.

YEAST, J. D., POSKIN, M., STOCKBAUER, J. W. and SHAFFER, S. (1998). Changing patterns in regionalization of perinatal care and the impact on neonatal mortality. *Am. J. Obstet. Gynecol.* **178** 131–135.

The Annals of Applied Statistics
2014, Vol. 8, No. 1, 74–88
DOI: 10.1214/13-AOAS684

BETA REGRESSION FOR TIME SERIES ANALYSIS OF BOUNDED DATA, WITH APPLICATION TO CANADA GOOGLE® FLU TRENDS

BY ANNAMARIA GUOLO AND CRISTIANO VARIN

Università di Verona and Università Ca' Foscari Venezia

Bounded time series consisting of rates or proportions are often encountered in applications. This manuscript proposes a practical approach to analyze bounded time series, through a beta regression model. The method allows the direct interpretation of the regression parameters on the original response scale, while properly accounting for the heteroskedasticity typical of bounded variables. The serial dependence is modeled by a Gaussian copula, with a correlation matrix corresponding to a stationary autoregressive and moving average process. It is shown that inference, prediction, and control can be carried out straightforwardly, with minor modifications to standard analysis of autoregressive and moving average models. The methodology is motivated by an application to the influenza-like-illness incidence estimated by the Google® Flu Trends project.

REFERENCES

BUTLER, D. (2013). When Google got flu wrong. *Nature* **494** 155–156.

CASARIN, R., DALLA VALLE, L. and LEISEN, F. (2012). Bayesian model selection for beta autoregressive processes. *Bayesian Anal*. **7** 385–409. MR2934956

COX, D. R. (1981). Statistical analysis of time series: Some recent developments. *Scand. J. Stat*. **8** 93–115. MR0623586

CRIBARI-NETO, F. and ZEILEIS, A. (2010). Beta regression in R. *Journal of Statistical Software* **34** 1–24.

DA SILVA, C. Q., MIGON, H. S. and CORREIA, L. T. (2011). Dynamic Bayesian beta models. *Comput. Statist. Data Anal*. **55** 2074–2089. MR2785115

DA-SILVA, C. Q. and MIGON, H. S. (2012). Hierarchical dynamic beta model. Technical Report 253. Dept. Statistics, Federal Univ. Rio de Janeiro.

DUNN, P. K. and SMYTH, G. K. (1996). Randomized quantile residuals. *J. Comput. Graph. Statist*. **5** 236–244.

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

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

GRÜN, B., KOSMIDIS, I. and ZEILEIS, A. (2012). Extended beta regression in R: Shaken, stirred, mixed, and partitioned. *Journal of Statistical Software* **48** 1–25.

Key words and phrases. Beta regression, bounded time series, Gaussian copula, Google® Flu Trends, surveillance.

GUOLO, A. and VARIN, C. (2013). Supplement to "Beta regression for time series analysis of bounded data, with application to Canada Google® Flu Trends." DOI:10.1214/13-AOAS684SUPP.

HUTWAGNER, L., THOMPSON, W. W., SEEMAN, G. M. and TREADWELL, T. (2003). The bioterrorism preparedness and response early aberration reporting system (EARS). *Journal of Urban Health* **80** 89–96.

KIESCHNICK, R. and MCCULLOUGH, B. D. (2003). Regression analysis of variates observed on (0, 1): Percentages, proportions and fractions. *Stat. Model.* **3** 193–213. MR2005473

LOVE, T. M. T., THURSON, S. W., KEEFER, M. C., DEWHURST, S. and LEE, H. Y. (2010). Mathematical modeling of ultradeep sequencing data reveals that acute CD8+ T-lymphocyte responses exert strong selective pressure in simian immunodeficiency virus-infected macaques but still fail to clear founder epitope sequences. *Journal of Virology* **84** 5802–5814.

MASAROTTO, G. and VARIN, C. (2012). Gaussian copula marginal regression. *Electron. J. Stat.* **6** 1517–1549. MR2988457

MONTGOMERY, D. C. (2009). *Introduction to Statistical Quality Control*, 6th ed. Wiley, New York.

OSPINA, R. and FERRARI, S. L. P. (2012). A general class of zero-or-one inflated beta regression models. *Comput. Statist. Data Anal.* **56** 1609–1623. MR2892364

PAOLINO, P. (2001). Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis* **9** 325–346.

R CORE TEAM. (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available at http://www.R-project.org/.

ROCHA, A. V. and CRIBARI-NETO, F. (2009). Beta autoregressive moving average models. *TEST* **18** 529–545. MR2566415

ROGERS, J. A., POLHAMUS, D., GILLESPIE, W. R., ITO, K., ROMERO, K., QIU, R., STEPHENSON, D., GASTONGUAY, M. R. and CORRIGAN, B. (2012). Combining patient-level and summary-level data for Alzheimer's disease modeling and simulation: A beta regression meta-analysis. *J. Pharmacokinet. Pharmacodyn.* **39** 479–498.

SMITHSON, M. and VERKUILEN, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychol. Methods* **11** 54–71.

SONG, P. X. K. (2007). *Correlated Data Analysis*: *Modeling*, *Analytics*, *and Applications*. Springer, New York. MR2377853

STASINOPOULOS, D. M. and RIGBY, R. A. (2007). Generalized additive models for location scale and shape (gamlss) in R. *Journal of Statistical Software* **23** 1–46.

UNKEL, S., FARRINGTON, C. P., GARTHWAITE, P. H., ROBERTSON, C. and ANDREWS, N. (2012). Statistical methods for the prospective detection of infectious disease outbreaks: A review. *J. Roy. Statist. Soc. Ser. A* **175** 49–82. MR2873791

WANG, X.-F. (2012). Joint generalized models for multidimensional outcomes: A case study of neuroscience data from multimodalities. *Biom. J.* **54** 264–280. MR2915307

WANG, W., SCHARFSTEIN, D., WANG, C., DANIELS, M., NEEDHAM, D. and BROWER, R. (2011). Estimating the causal effect of low tidal volume ventilation on survival in patients with acute lung injury. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **60** 475–496. MR2829186

WOODALL, W. (2006). The use of control chart in health-care and public-health surveillance. *Journal of Quality Technology* **38** 89–104.

ZOU, K. H., CARLSSON, M. O. and QUINN, S. A. (2010). Beta-mapping and beta-regression for changes of ordinal-rating measurements on Likert scales: A comparison of the change scores among multiple treatment groups. *Stat. Med.* **29** 2486–2500. MR2897363

The Annals of Applied Statistics
2014, Vol. 8, No. 1, 89–119
DOI: 10.1214/13-AOAS710

POWER-LAW DISTRIBUTIONS IN BINNED EMPIRICAL DATA

BY YOGESH VIRKAR* AND AARON CLAUSET*,†

*University of Colorado, Boulder** *and Santa Fe Institute*†

Many man-made and natural phenomena, including the intensity of earthquakes, population of cities and size of international wars, are believed to follow power-law distributions. The accurate identification of power-law patterns has significant consequences for correctly understanding and modeling complex systems. However, statistical evidence for or against the power-law hypothesis is complicated by large fluctuations in the empirical distribution's tail, and these are worsened when information is lost from binning the data. We adapt the statistically principled framework for testing the power-law hypothesis, developed by Clauset, Shalizi and Newman, to the case of binned data. This approach includes maximum-likelihood fitting, a hypothesis test based on the Kolmogorov–Smirnov goodness-of-fit statistic and likelihood ratio tests for comparing against alternative explanations. We evaluate the effectiveness of these methods on synthetic binned data with known structure, quantify the loss of statistical power due to binning, and apply the methods to twelve real-world binned data sets with heavy-tailed patterns.

REFERENCES

ABAN, I. B. and MEERSCHAERT, M. M. (2004). Generalized least-squares estimators for the thickness of heavy tails. *J. Statist. Plann. Inference* **119** 341–352. MR2019645

ARNOLD, B. C. (1983). *Pareto Distributions*. *Statistical Distributions in Scientific Work* **5**. International Co-operative Publishing House, Burtonsville, MD. MR0751409

ASAL, V. and RETHEMEYER, R. K. (2008). The nature of the beast: Organizational structures and the lethality of terrorist attacks. *The Journal of Politics* **70** 437–449.

BARNDORFF-NIELSEN, O. E. and COX, D. R. (1995). *Inference and Asymptotics*. Chapman & Hall, London.

BEIRLANT, J. and TEUGELS., J. L. (1989). Asymptotic normality of Hill's estimator. *Extreme Value Theory* **51** 148–155.

BREIMAN, L., STONE, C. J. and KOOPERBERG, C. (1990). Robust confidence bounds for extreme upper quantiles. *J. Stat. Comput. Simul.* **37** 127–149. MR1082452

CADEZ, I. V., SMYTH, P., MCLACHLAN, G. J. and MCLAREN, C. E. (2002). Maximum likelihood estimation of mixture of densities for binned and truncated multivariate data. *Machine Learning* **47** 7–34.

CHAPUIS, A. and TETZLAFF, T. (2012). The variability of tidewater-glacier calving: Origin of event-size and interval distributions. Available at arXiv:1205.1640.

CLAUSET, A., SHALIZI, C. R. and NEWMAN, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Rev.* **51** 661–703. MR2563829

CLAUSET, A. and WOODARD, R. (2013). Estimating the historical and future probabilities of large terrorist events. *Ann. Appl. Stat.* **7** 1838–1865.

Key words and phrases. Power-law distribution, heavy-tailed distributions, model selection, binned data.

CLAUSET, A., YOUNG, M. and GLEDITSCH, K. S. (2007). On the frequency of severe terrorist events. *Journal of Conflict Resolution* **51** 58–87.

CRAMÉR, H. (1946). A contribution to the theory of statistical estimation. *Skand. Aktuarietidskr.* **29** 85–94. MR0017505

DANIELSSON, J., DE HAAN, L., PENG, L. and DE VRIES, C. G. (2001). Using a bootstrap method to choose the sample fraction in tail index estimation. *J. Multivariate Anal.* **76** 226–248. MR1821820

DEKKERS, A. L. M. and DE HAAN, L. (1993). Optimal choice of sample fraction in extreme-value estimation. *J. Multivariate Anal.* **47** 173–195. MR1247373

DREES, H. and KAUFMANN, E. (1998). Selecting the optimal sample fraction in univariate extreme value estimation. *Stochastic Process. Appl.* **75** 149–172. MR1632189

EFRON, B. and TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap*. *Monographs on Statistics and Applied Probability* **57**. Chapman & Hall, New York. MR1270903

GABAIX, X. (2009). Power laws in economics and finance. *Annual Review of Economics* **1** 255–293.

GOH, K.-I., CUSICK, M. E., VALLE, D., CHILDS, B., VIDAL, M. and BARABÁSI, A.-L. (2007). The human disease network. *Proc. Natl. Acad. Sci. USA* **104** 8685–8690.

GOLDSTEIN, M. L., MORRIS, S. A. and YEN, G. G. (2004). Problems with fitting to the power-law distribution. *Eur. Phys. J. B* **41** 255–258.

GRÜNWALD, P. D. (2007). *The Minimum Length Description Principle*. MIT Press, Cambridge, MA.

HALL, P. (1982). On some simple estimates of an exponent of regular variation. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **44** 37–42. MR0655370

HANDCOCK, M. S. and JONES, J. H. (2004). Likelihood-based inference for stochastic models of sexual network evolution. *Theoretical Population Biology* **65** 413–422.

HERITAGE PROVIDER NETWORK (2012). Health heritage prize data files, HHP_release3. Available at http://bit.ly/wG8Psl.

HILL, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist.* **3** 1163–1174. MR0378204

HORN, S. D. (1977). Goodness-of-fit tests for discrete data: A review and an application to a health development scale. *Biometrics* **33** 237–247.

IJIRI, Y. and SIMON, H. A. (1977). *Skew Distributions and the Sizes of Business Firms*. North-Holland, Amsterdam.

JARVINEN, B., NEUMANN, C. and DAVIS, M. A. S. (2012). NHC data archive. National Hurricane Center. Available at http://1.usa.gov/cCcwTg.

KASS, R. E. and RAFTERY, A. E. (1994). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795.

KRATZ, M. and RESNICK, S. I. (1996). The QQ-estimator and heavy tails. *Comm. Statist. Stochastic Models* **12** 699–724. MR1410853

MCLACHLAN, G. J. and JONES, P. N. (1988). Fitting mixture models to grouped and truncated data via the EM algorithm. *Biometrics* **44** 571–578.

MITZENMACHER, M. (2004). A brief history of generative models for power law and lognormal distributions. *Internet Math.* **1** 226–251. MR2077227

MITZENMACHER, M. (2006). The future of power law research. *Internet Math.* **2** 525–534.

NEWMAN, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics* **46** 323–351.

NOETHER, G. E. (1963). Note on the Kolmogorov statistic in the discrete case. *Metrika* **7** 115–116. MR0158462

ORPHANET REPORT SERIES, RARE DISEASES COLLECTION (2011). Prevalence of rare diseases: Bibliographic data. Available at http://bit.ly/MezSZ6.

PERSING, J. and MONTGOMERY, M. T. (2003). Hurricane superintensity. *J. Atmospheric Sci.* **60** 2349–2371.

PRESS, W. H., TEUKOLSKY, S. A., VETTERLING, W. T. and FLANNERY, B. P. (1992). *Numerical Recipes in C*: *The Art of Scientific Computing*, 2nd ed. Cambridge Univ. Press, Cambridge. MR1201159

RAO, C. R. (1947). Minimum variance and the estimation of several parameters. *Proc. Cambridge Philos. Soc.* **43** 280–283. MR0019904

RAO, C. R. (1957). Maximum likelihood estimation for the multinomial distribution. *Sankhyā* **18** 139–148. MR0105183

REED, W. J. and HUGHES, B. D. (2002). From gene families and genera to income and internet file sizes: Why power laws are so common in nature. *Phys. Rev. E* (3) **66** 067103.

REISS, R.-D. and THOMAS, M. (2007). *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields*, 3rd ed. Birkhäuser, Basel. MR2334035

RICHARDSON, L. F. (1960). *Statistics of Deadly Quarrels*. The Boxwood Press, Pittsburgh.

SCHULTZE, J. and STEINEBACH, J. (1996). On least squares estimates of an exponential tail coefficient. *Statist. Decisions* **14** 353–372. MR1437826

SHINOKAZI, K., YODA, K., HOZUMI, K. and KIRA, T. (1964). A quantitative analysis of plant form—The pipe model theory II: Further evidence of the theory and its application in forest ecology. *Japanese Journal of Ecology* **14** 133–139.

SORNETTE, D. (2006). *Critical Phenomena in Natural Sciences*: *Chaos, Fractals, Selforganization and Disorder*: *Concepts and Tools*, 2nd ed. Springer, Berlin. MR2220576

STOEV, S. A., MICHAILIDIS, G. and TAQQU, M. S. (2011). Estimating heavy-tail exponents through max self-similarity. *IEEE Trans. Inform. Theory* **57** 1615–1636. MR2815838

STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **36** 111–147. MR0356377

STORM PREDICTION CENTER (2011). Severe weather database files (1950–2011). Available at http://1.usa.gov/Lj7cC9.

STUMPF, M. P. H. and PORTER, M. A. (2012). Critical truths about power laws. *Science* **335** 665–666. MR2932329

TATE, M. W. and HYE, L. A. (1973). Inaccuracy of the χ^2 test of goodness of fit when expected frequencies are small. *J. Amer. Statist. Assoc.* **68** 836–841.

VIRKAR, Y. and CLAUSET, A. (2014). Supplement to "Power-law distributions in binned empirical data." DOI:10.1214/13-AOAS710SUPP.

VUONG, Q. H. (1989). Likelihood ratio tests for model selection and nonnested hypotheses. *Econometrica* **57** 307–333. MR0996939

WASSERMAN, L. (2004). *All of Statistics*: *A Concise Course in Statistical Inference*. Springer, New York. MR2055670

WEST, G. B., ENQUIST, B. J. and BROWN, J. H. (2009). A general quantitative theory of forest structure and dynamics. *Proc. Natl. Acad. Sci. USA* **106** 7040–7045.

WORLD GLACIER MONITORING SERVICE and NATIONAL SNOW AND ICE DATA CENTER (2012). World glacier inventory. Available at http://bit.ly/MhLdt6.

YAMAMOTO, K. and KOBAYASHI, S. (1993). Analysis of crown structure based on the pipe model theory. *Journal of the Japanese Forestry Society* **75** 445–448.

The Annals of Applied Statistics
2014, Vol. 8, No. 1, 120–147
DOI: 10.1214/13-AOAS694

SEPARABLE FACTOR ANALYSIS WITH APPLICATIONS TO MORTALITY DATA

BY BAILEY K. FOSDICK AND PETER D. HOFF

Statistical and Applied Mathematical Sciences Institute and University of Washington

Human mortality data sets can be expressed as multiway data arrays, the dimensions of which correspond to categories by which mortality rates are reported, such as age, sex, country and year. Regression models for such data typically assume an independent error distribution or an error model that allows for dependence along at most one or two dimensions of the data array. However, failing to account for other dependencies can lead to inefficient estimates of regression parameters, inaccurate standard errors and poor predictions. An alternative to assuming independent errors is to allow for dependence along each dimension of the array using a separable covariance model. However, the number of parameters in this model increases rapidly with the dimensions of the array and, for many arrays, maximum likelihood estimates of the covariance parameters do not exist. In this paper, we propose a submodel of the separable covariance model that estimates the covariance matrix for each dimension as having factor analytic structure. This model can be viewed as an extension of factor analysis to array-valued data, as it uses a factor model to estimate the covariance along each dimension of the array. We discuss properties of this model as they relate to ordinary factor analysis, describe maximum likelihood and Bayesian estimation methods, and provide a likelihood ratio testing procedure for selecting the factor model ranks. We apply this methodology to the analysis of data from the Human Mortality Database, and show in a cross-validation experiment how it outperforms simpler methods. Additionally, we use this model to impute mortality rates for countries that have no mortality data for several years. Unlike other approaches, our methodology is able to estimate similarities between the mortality rates of countries, time periods and sexes, and use this information to assist with the imputations.

REFERENCES

ALLEN, G. I. and TIBSHIRANI, R. (2010). Transposable regularized covariance models with an application to missing data imputation. *Ann. Appl. Stat.* **4** 764–790. MR2758420

ANDERSON, T. W. and RUBIN, H. (1956). Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* 1954–1955, *Vol. V* 111–150. Univ. California Press, Berkeley and Los Angeles. MR0084943

BHATTACHARYA, A. and DUNSON, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98** 291–306. MR2806429

Key words and phrases. Array normal, Kronecker product, multiway data, Bayesian estimation, imputation.

BRASS, W. (1971). On the scale of mortality. In *Biological Aspects of Demography* 69–110. Taylor and Francis, London.

BROWNE, M. W. (1984). The decomposition of multitrait-multimethod matrices. *British J. Math. Statist. Psych.* **37** 1–21. MR0783496

CARTER, L. R. and LEE, R. D. (1992). Modeling and forecasting US sex differentials in mortality. *International Journal of Forecasting* **8** 393–411.

CARVALHO, C. M., CHANG, J., LUCAS, J. E., NEVINS, J. R., WANG, Q. and WEST, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *J. Amer. Statist. Assoc.* **103** 1438–1456. MR2655722

CHIOU, J.-M. and MÜLLER, H.-G. (2009). Modeling hazard rates as functional data for the analysis of cohort lifetables and mortality forecasting. *J. Amer. Statist. Assoc.* **104** 572–585. MR2751439

COALE, A. J. and DEMENY, P. (1966). *Regional Model Life Tables and Stable Populations*. Princeton Univ. Press, Princeton.

CONGDON, P. (1993). Statistical graduation in local demographic analysis and projection. *J. Roy. Statist. Soc. Ser. A* **156** 237–270.

CURRIE, I. D., DURBAN, M. and EILERS, P. H. C. (2004). Smoothing and forecasting mortality rates. *Stat. Model.* **4** 279–298. MR2086492

DAWID, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika* **68** 265–274. MR0614963

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

DE LATHAUWER, L., DE MOOR, B. and VANDEWALLE, J. (2000). A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.* **21** 1253–1278 (electronic). MR1780272

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

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

FELIPE, A., GUILLEN, M. and NIELSEN, J. P. (2001). Longevity studies based on kernel hazard estimation. *Insurance Math. Econom.* **28** 191–204.

GENTON, M. G. (2007). Separable approximations of space–time covariance matrices. *Environmetrics* **18** 681–695. MR2408938

GEWEKE, J. and ZHOU, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *Rev. Financ. Stud.* **9** 557–587.

HARTMANN, M. (1987). Past and recent attempts to model mortality at all ages. *J. Off. Stat.* **3** 19–36.

HELIGMAN, L. and POLLARD, J. H. (1980). The age pattern of mortality. *Journal of the Institute of Actuaries* **107** 49–80.

HOFF, P. D. (2011). Separable covariance arrays via the Tucker product, with applications to multivariate relational data. *Bayesian Anal.* **6** 179–196. MR2806238

HUMAN MORTALITY DATABASE UNIVERSITY OF CALIFORNIA, BERKELEY (USA) AND MAX PLANCK INSTITUTE FOR DEMOGRAPHIC RESEARCH (GERMANY) (2011). Available at www.mortality.org or www.humanmortality.de (data downloaded in 2011).

JENNRICH, R. I. and ROBINSON, S. M. (1969). A Newton–Raphson algorithm for maximum likelihood factor analysis. *Psychometrika* **34** 111–123. MR0239703

JÖRESKOG, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika* **32** 443–482. MR0221659

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

KIERS, H. A. L. (2000). Towards a standardized notation and terminology in multiway analysis. *J. Chemom.* **14** 105–122.

KOLDA, T. G. and BADER, B. W. (2009). Tensor decompositions and applications. *SIAM Rev.* **51** 455–500. MR2535056

KROONENBERG, P. M. (2008). *Applied Multiway Data Analysis*. Wiley, Hoboken, NJ. MR2378349

LAWLEY, D. N. (1940). The estimation of factor loadings by the method of maximum likelihood. *Proc. Roy. Soc. Edinburgh Sect. A* **60** 64–82. MR0002754

LEE, R. D. and CARTER, L. R. (1992). Modeling and forecasting U.S. mortality. *J. Amer. Statist. Assoc.* **87** 659–671.

LEE, S.-Y. and SONG, X.-Y. (2002). Bayesian selection on the number of factors in a factor analysis model. *Behaviormetrika* **29** 23–39. MR1894459

LI, N. and LEE, R. (2005). Coherent mortality forecasts for a group of populations: An extension of the Lee–Carter method. *Demography* **42** 575–594.

LIU, C. and RUBIN, D. B. (1998). Maximum likelihood estimation of factor analysis using the ECME algorithm with complete and incomplete data. *Statist. Sinica* **8** 729–747. MR1651505

LOPES, H. F. and WEST, M. (2004). Bayesian model assessment in factor analysis. *Statist. Sinica* **14** 41–67. MR2036762

MANCEUR, A. M. and DUTILLEUL, P. (2013). Maximum likelihood estimation for the tensor normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion. *J. Comput. Appl. Math.* **239** 37–49. MR2991957

MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1979). *Multivariate Analysis*. Academic Press, San Diego, CA.

MARTÍNEZ-RUIZ, F., MATEU, J., MONTES, F. and PORCU, E. (2010). Mortality risk assessment through stationary space–time covariance functions. *Stoch. Environ. Res. Risk Assess.* **24** 519–526.

MCNOWN, R. and ROGERS, A. (1989). Forecasting mortality: A parameterized time series approach. *Demography* **26** 645–660.

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

MODE, C. and BUSBY, R. (1982). An eight-parameter model of human mortality—The single decrement case. *Bull. Math. Biol.* **44** 647–659.

MURRAY, C. J. L., FERGUSON, B. D., LOPEZ, A. D., GUILLOT, M., SALOMON, J. A. and AHMAD, O. (2003). Modified logit life table system: Principles, empirical validation, and application. *Population Studies* **57** 165–182.

UNITED NATIONS (1982). Model life tables for developing countries. In *Population Studies* **77**. United Nations, New York.

OORT, F. J. (1999). Stochastic three-mode models for mean and covariance structures. *British J. Math. Statist. Psych.* **52** 243–272. MR1740267

RENSHAW, A. E. and HABERMAN, S. (2003a). Lee–Carter mortality forecasting with age-specific enhancement. *Insurance Math. Econom.* **33** 255–272. Papers presented at the 6th IME Conference (Lisbon, 2002). MR2039286

RENSHAW, A. and HABERMAN, S. (2003b). Lee–Carter mortality forecasting: A parallel generalized linear modelling approach for England and Wales mortality projections. *J. R. Stat. Soc. Ser. C Appl. Stat.* **52** 119–137. MR1959085

RENSHAW, A. E. and HABERMAN, S. (2003c). On the forecasting of mortality reduction factors. *Insurance Math. Econom.* **32** 379–401.

RENSHAW, A. E., HABERMAN, S. and HATZOPOULOS, P. (1996). The modelling of recent mortality trends in United Kingdom male assured lives. *British Actuarial Journal* **2** 449–477.

ROBERTSON, D. and SYMONS, J. (2007). Maximum likelihood factor analysis with rank-deficient sample covariance matrices. *J. Multivariate Anal.* **98** 813–828. MR2322130

RUBIN, D. B. and THAYER, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika* **47** 69–76. MR0668505

SILER, W. (1983). Parameters of mortality in human populations with widely varying life spans. *Stat. Med.* **2** 373–380.

SPALL, J. C. (2005). Monte Carlo computation of the Fisher information matrix in nonstandard settings. *J. Comput. Graph. Statist.* **14** 889–909. MR2211372

SPEARMAN, C. (1904). "General intelligence," objectively determined and measured. *Am. J. Psychol.* **15** 201–292.

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

WANG, H. (2012). Bayesian graphical lasso models and efficient posterior computation. *Bayesian Anal.* **7** 867–886. MR3000017

WANG, H. and WEST, M. (2009). Bayesian analysis of matrix normal graphical models. *Biometrika* **96** 821–834. MR2564493

WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50** 1–25. MR0640163

ZHAO, J.-H., YU, P. L. H. and JIANG, Q. (2008). ML estimation for factor analysis: EM or non-EM? *Stat. Comput.* **18** 109–123. MR2390813

The Annals of Applied Statistics
2014, Vol. 8, No. 1, 148–175
DOI: 10.1214/13-AOAS705

A HIERARCHICAL BAYESIAN MODEL FOR INFERENCE OF COPY NUMBER VARIANTS AND THEIR ASSOCIATION TO GENE EXPRESSION

BY ALBERTO CASSESE[*], MICHELE GUINDANI[†], MAHLET G. TADESSE[‡], FRANCESCO FALCIANI[§] AND MARINA VANNUCCI[*]

Rice University[*], MD Anderson Cancer Center[†], Georgetown University[‡] and University of Liverpool[§]

A number of statistical models have been successfully developed for the analysis of high-throughput data from a single source, but few methods are available for integrating data from different sources. Here we focus on integrating gene expression levels with comparative genomic hybridization (CGH) array measurements collected on the same subjects. We specify a measurement error model that relates the gene expression levels to latent copy number states which, in turn, are related to the observed surrogate CGH measurements via a hidden Markov model. We employ selection priors that exploit the dependencies across adjacent copy number states and investigate MCMC stochastic search techniques for posterior inference. Our approach results in a unified modeling framework for simultaneously inferring copy number variants (CNV) and identifying their significant associations with mRNA transcripts abundance. We show performance on simulated data and illustrate an application to data from a genomic study on human cancer cell lines.

REFERENCES

BARNES, C., PLAGNOL, V., FITZGERALD, T., REDON, R., MARCHINI, J., CLAYTON, D. and HURLES, M. E. (2008). A robust statistical method for case–control association testing with copy number variation. *Nature Genetics* **40** 1245–1252.

BELFIORE, A., GENUA, M. and MALAGUARNERA, R. (2009). PPAR-gamma agonists and their effects on IGF-I receptor signaling: Implications for cancer. *PPAR Research* **2009** Article ID 830501.

BREHENY, P., CHALISE, P., BATZLER, A., WANG, L. and FRIDLEY, B. L. (2012). Genetic association studies of copy-number variation: Should assignment of copy number states precede testing? *PLoS ONE* **7** e34262.

BROET, P., LEWIN, A., RICHARDSON, S., DALMASSO, C. and MAGDELENAT, H. (2004). A mixture model-based strategy for selecting sets of genes in multiclass response microarray experiments. *Bioinformatics* **20** 2562–2571.

BROWN, P. J., VANNUCCI, M. and FEARN, T. (1998). Multivariate Bayesian variable selection and prediction. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **60** 627–641. MR1626005

BUSSEY, K. J., CHIN, K., LABABIDI, S., REIMERS, M., REINHOLD, W. C., KU, W.-L., GWADRY, F., KOUROS-MEHR, A. H., FRIDLYAND, J., JAIN, A., COLLINS, C., NISHIZUKA, S.,

Key words and phrases. Bayesian hierarchical models, comparative genomic hybridization arrays, gene expression, hidden Markov models, measurement error, variable selection.

TONON, G., ROSCHKE, A., GEHLHAUS, K., KIRSCH, I., SCUDIERO, D. A., GRAY, J. W. and WEINSTEIN, J. N. (2006). Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel. *Molecular Cancer Therapeutics* **5** 853–867.

CARDIN, N., HOLMES, C., DONNELLY, P., WELLCOME TRUST CASE CONTROL CONSORTIUM and MARCHINI, J. (2011). Bayesian hierarchical mixture modeling to assign copy number from a targeted CNV array. *Genetic Epidemiology* **35** 536–548.

CASSESE, A., GUINDANI, M., TADESSE, M. G., FALCIANI, F. and VANNUCCI, M. (2014). Supplement to "A hierarchical Bayesian model for inference of copy number variants and their association to gene expression." DOI:10.1214/13-AOAS705SUPP.

CHEN, X., WANG, L. and ISHWARAN, H. (2010). An integrative pathway-based clinical-genomic model for cancer survival prediction. *Statist. Probab. Lett.* **80** 1313–1319. MR2669767

CHIN, K., DEVRIES, S., FRIDLYAND, J., SPELLMAN, P. T., ROYDASGUPTA, R., KUO, W. L., LAPUK, A., NEVE, R. M., QIAN, Z., RYDER, T., CHEN, F., FEILER, H., TOKUYASU, T., KINGSLEY, C., DAIRKEE, S., MENG, Z., CHEW, K., PINKEL, D., JAIN, A., LJUNG, B. M., ESSERMAN, L., ALBERTSON, D. G., WALDMAN, F. M. and GRAY, J. W. (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell.* **10** 529–541.

CHOI, H., QIN, Z. S. and GHOSH, D. (2010). A double-layered mixture model for the joint analysis of DNA copy number and gene expression data. *J. Comput. Biol.* **17** 121–137. MR2593954

COLELLA, S., YAU, C., TAYLOR, J. M., MIRZA, G., BUTLER, H., CLOUSTON, P., BASSETT, A. S., SELLER, A., HOLMES, C. C. and RAGOUSSIS, J. (2007). QuantiSNP: An objective Bayes hidden-Markov model to detect and accurately map copy number variation using SNP genotyping data. *Nucleid Acids Research* **35** 2013–2025.

CORDELL, H. J. (2002). Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics* **11** 2463–2468.

COSTA, T., GUINDANI, M., BASSETTI, F., LEISEN, F. and AIROLDI, E. M. (2013). Generalized species sampling priors with latent beta reinforcements. Available at arXiv:1012.0866.

DALENC, F., DROUET, J., ADER, I., DELMAS, C., ROCHAIX, P., FAVRE, G., COHEN-JONATHAN, E. and TOULAS, C. (2012). Increased expression of a COOH-truncated nucleophosmin resulting from alternative splicing is associated with cellular resistance to ionizing radiation in HeLa cells. *Int. J. Cancer* **100** 662–668.

DRIER, Y., SHEFFER, M. and DOMANY, E. (2013). Pathway-based personalized analysis of cancer. *Proc. Natl. Acad. Sci. USA* **110** 6388–6393.

DU, L., CHEN, M., LUCAS, J. and CARIN, L. (2010). Sticky hidden Markov modeling of comparative genomic hybridization. *IEEE Trans. Signal Process.* **58** 5353–5368. MR2722675

FOX, E. B., SUDDERTH, E. B., JORDAN, M. I. and WILLSKY, A. S. (2011). A sticky HDP–HMM with application to speaker diarization. *Ann. Appl. Stat.* **5** 1020–1056. MR2840185

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

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

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

HEIDELBERGER, P. and WELCH, P. D. (1981). A spectral method for confidence interval generation and run length control in simulations. *Comm. ACM* **24** 233–245. MR0611745

JONES, B., CARVALHO, C., DOBRA, A., HANS, C., CARTER, C. and WEST, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statist. Sci.* **20** 388–400. MR2210226

KACZYNSKI, J., HANSSON, G. and WALLERSTEDT, S. (2009). Wallerstedtincreased porphyrins in primary liver cancer mainly reflect a parallel liver disease. *Gastroenterology Research and Practice* **2009** Article ID 402394.

MARIONI, J. C., THORNE, N. P. and TAVARE, S. (2006). BioHMM: A heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics* **22** 1144–1146.

MONNI, S. and TADESSE, M. G. (2009). A stochastic partitioning method to associate high-dimensional responses and covariates. *Bayesian Anal.* **4** 413–436. MR2545166

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, 667. MR2432418

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.

NOOR, R., MITTAL, S. and IQBAL, J. (2002). Superoxide dismutase-applications and relevance to human diseases. *Med. Sci. Monit.* **8** 9.

OVACIK, M. A., SUKUMARAN, S., ALMON, R. R., DUBOIS, D. C., JUSKO, W. J. and ANDROULAKIS, I. P. (2010). Circadian signatures in rat liver: From gene expression to pathways. *BMC Bioinformatics* **11** 540.

PICARD, F., ROBIN, S., LEBARBIER, E. and DAUDIN, J.-J. (2007). A segmentation/clustering model for the analysis of array CGH data. *Biometrics* **63** 758–766. MR2395713

RABER, P., OCHOA, A. C. and RODRÍGUEZ, P. C. (2012). Metabolism of L-arginine by myeloid-derived suppressor cells in cancer: mechanisms of T cell suppression and therapeutic perspectives. *Immunol. Invest.* **41** 614–634.

REDON, R., ISHIKAWA, S., FITCH, K. R., FEUK, L., PERRY, G. H., ANDREWS, T. D., FIEGLER, H., SHAPERO, M. H., CARSON, A. R., CHEN, W. ET AL. (2006). Global variation in copy number in the human genome. *Nature* **444** 444–454.

RICHARDSON, S., BOTTOLO, L. and ROSENTHAL, J. S. (2010). Bayesian models for sparse regression analysis of high dimensional data. *Bayesian Statistics* **9** 539–569.

RICHARDSON, S. and GILKS, W. R. (1993). Conditional independence models for epidemiological studies with covariate measurement error. *Stat. Med.* **12** 1703–1722.

RODRIGUEZ, R. R. R., DURAN, R. C. D., FALCIANI, F., PEÑA, J. G. T. and TREVINO, V. (2012). COMPADRE: An R and web resource for pathway activity analysis by component decompositions. *Bioinformatics* **28** 2701–2702.

SCOTT-BOYER, M. P., IMHOLTE, G. C., TAYEB, A., LABBE, A., DESCHEPPER, C. F. and GOTTARDO, R. (2012). An integrated hierarchical Bayesian model for multivariate eQTL mapping. *Stat. Appl. Genet. Mol. Biol.* **11** 1515–1544. MR2958606

SEBAT, J., LAKSHMI, B., TROGE, J., ALEXANDER, J., YOUNG, J., LUNDIN, P., MANER, S., MASSA, H., WALKER, M., CHI, M. ET AL. (2004). Large-scale copy number polymorphism in the human genome. *Science* **305** 525–528.

SHA, N., VANNUCCI, M., TADESSE, M. G., BROWN, P. J., DRAGONI, I., DAVIES, N., ROBERTS, T. C., CONTESTABILE, A., SALMON, M., BUCKLEY, C. and FALCIANI, F. (2004). Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics* **60** 812–828. MR2089459

SOMWAR, H., ERDJUMENT-BROMAGE, R., LARSSON, E., SHUM, D., LOCKWOOD, W. W., YANG, G., SANDER, C., OUERFELLI, O., TEMPST, P. J., DJABALLAH, H. and VARMUS, H. E. (2011). Superoxide dismutase 1 (SOD1) is a target for a small molecule identified in a screen for inhibitors of the growth of lung adenocarcinoma cell lines. *PNAS* **108** 39.

STINGO, F. C., CHEN, Y. A., VANNUCCI, M., BARRIER, M. and MIRKES, P. E. (2010). A Bayesian graphical modeling approach to microRNA regulatory network inference. *Ann. Appl. Stat.* **4** 2024–2048. MR2829945

STRANGER, B. E., FORREST, M. S., DUNNING, M., INGLE, C. E., BEAZLEY, C., THORNE, N., REDON, R., BIRD, C. P., DE GRASSI, A., LEE, C., TYLER-SMITH, C., CARTER, N., SCHERER, S. W., TAVARÉ, S., DELOUKAS, P., HURLES, M. E. and DERMITZAKIS, E. T. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315** 848–853.

SU, J., YOON, B.-J. and DOUGHERTY, E. R. (2009). Accurate and reliable cancer classification based on probabilistic inference of pathway activity. *PLoS ONE* **4** e8161.

SUBIRANA, I., DIAZ-URIARTE, R., LUCAS, G. and GONZALEZ, J. R. (2011). CNVassoc: Association analysis of CNV data using R. *BMC Med. Genomics* **4** 47.

VENKATRAMAN, E. S. and OLSHEN, A. B. (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23** 657–663.

WANG, K., LI, M., HADLEY, D., LIU, R., GLESSNER, J., GRANT, S. F. A., HAKONARSON, H. and BUCAN, M. (2007). PennCNV: An integrated hidden Markov model deisigned for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research* **17** 1665–1674.

WANG, K., CHEN, Z., TADESSE, M. G., GLESSNER, J., GRANT, S. F. A., HAKONARSON, H., BUCAN, M. and LI, M. (2008). Modeling genetic inheritance of copy number variations. *Nucleid Acids Research* **36** 21.

WU, G., GUO, Z., CHATTERJEE, A., HUANG, X., RUBIN, E., WU, F., MAMBO, E., CHANG, X., OSADA, M., KIM, M. S., MOON, C., CALIFANO, J. A., RATOVITSKI, E. A., GOLLIN, S. M., SUKUMAR, S., SIDRANSKY, D. and TRINK, B. (2006). Overexpression of glycosylphosphatidylinositol (GPI) transamidase subunits phosphatidylinositol glycan class T and/or GPI anchor attachment 1 induces tumorigenesis and contributes to invasion in human breast cancer. *Cancer Res.* **66** 9829–9836.

YANG, Y. and BEDFORD, M. T. (2013). Protein arginine methyltransferases and cancer. *Nat. Rev. Cancer* **13** 37–50.

The Annals of Applied Statistics
2014, Vol. 8, No. 1, 176–203
DOI: 10.1214/13-AOAS695

BAYESIAN METHODS FOR GENETIC ASSOCIATION ANALYSIS WITH HETEROGENEOUS SUBGROUPS: FROM META-ANALYSES TO GENE–ENVIRONMENT INTERACTIONS

BY XIAOQUAN WEN AND MATTHEW STEPHENS

University of Michigan and University of Chicago

Genetic association analyses often involve data from multiple potentially-heterogeneous subgroups. The expected amount of heterogeneity can vary from modest (e.g., a typical meta-analysis) to large (e.g., a strong gene–environment interaction). However, existing statistical tools are limited in their ability to address such heterogeneity. Indeed, most genetic association meta-analyses use a "fixed effects" analysis, which assumes no heterogeneity. Here we develop and apply Bayesian association methods to address this problem. These methods are easy to apply (in the simplest case, requiring only a point estimate for the genetic effect and its standard error, from each subgroup) and effectively include standard frequentist meta-analysis methods, including the usual "fixed effects" analysis, as special cases. We apply these tools to two large genetic association studies: one a meta-analysis of genome-wide association studies from the Global Lipids consortium, and the second a cross-population analysis for expression quantitative trait loci (eQTLs). In the Global Lipids data we find, perhaps surprisingly, that effects are generally quite homogeneous across studies. In the eQTL study we find that eQTLs are generally shared among different continental groups, and discuss consequences of this for study design.

REFERENCES

BRAVATA, D. and OLKIN, I. (2001). Simple pooling versus combining in meta-analysis. *Eval. Health Prof.* **24** 218–230.

BROWN, C., MANGRAVITE, L. M. and ENGELHARDT, B. E. (2012). Integrative modeling of eQTLs and cis-regulatory elements suggest mechanisms underlying cell type specifcity of eQTLs. Preprint. Available at arXiv:1210.3294.

BURGESS, S., THOMPSON, S. G. and ANDREWS, G. et al. (2010). Bayesian methods for meta-analysis of causal relationships estimated using genetic instrumental variables. *Stat. Med.* **29** 1298–1311.

BUTLER, R. W. and WOOD, A. T. A. (2002). Laplace approximations for hypergeometric functions with matrix argument. *Ann. Statist.* **30** 1155–1177. MR1926172

DE IORIO, M., NEWCOMBE, P. J., TACHMAZIDOU, I., VERZILLI, C. J. and WHITTAKER, J. C. (2011). Bayesian semiparametric meta-analysis for genetic association studies. *Genet. Epidemiol.* **35** 333–340.

DIMAS, A. S., DEUTSCH, S., STRANGER, B. E., MONTGOMERY, S. B., BOREL, C. et al. (2009). Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* **325** 1246–1250.

Key words and phrases. Meta-analysis, gene–environment interaction, Bayes factor, Bayesian hypothesis testing, heterogeneity.

DuMouchel, W. H. and Harris, J. E. (1983). Bayes methods for combining the results of cancer studies in humans and other species. *J. Amer. Statist. Assoc.* **78** 293–315. MR0711105

Durbin, R. M., Altshuler, D. L., Abecasis, G. R., Bentley, D. R., Chakravarti, A. et al. (2010). A map of human genome variation from population-scale sequencing. *Nature* **467** 1061–1073.

Eddy, D. M., Hasselblad, V. and Schachter, R. (1990). A Bayesian method for synthesizing evidence. *International Journal of Technical Assistance in Health Care* **6** 31–55.

Fledel-Alon, A., Leffler, E. M., Guan, Y., Stephens, M., Coop, G. et al. (2011). Variation in human recombination rates and its genetic determinants. *PloS One* **6** e20321.

Flutre, T., Wen, X., Pritchard, J. K. and Stephens, M. (2013). A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genetics* **9** e1003486.

Gilad, Y., Rifkin, S. A. and Pritchard, J. K. (2008). Revealing the architecture of gene regulation: The promise of eQTL studies. *Trends Genet.* **24** 408–415.

Givens, G. H., Smith, D. D. and Tweedie, R. L. (1997). Publication bias in meta-analysis: A Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. *Statist. Sci.* **12** 221–250.

Guan, Y. and Stephens, M. (2008). Practical issues in imputation-based association mapping. *PLoS Genetics* **4** e1000279.

Han, B. and Eskin, E. (2011). Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am. J. Hum. Genet.* **88** 586–598.

Johnson, V. E. (2005). Bayes factors based on test statistics. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 689–701. MR2210687

Johnson, V. E. (2008). Properties of Bayes factors based on test statistics. *Scand. J. Stat.* **35** 354–368. MR2418746

Kong, A., Thorleifsson, G., Stefansson, H., Masson, G. et al. (2008). Sequence variants in the RNF212 gene associate with genome-wide recombination rate. *Science* **319** 1398–1401.

Lebrec, J. J., Stijnen, T. and van Houwelingen, H. C. (2010). Dealing with heterogeneity between cohorts in genomewide SNP association studies dealing with heterogeneity between cohorts in genomewide SNP association studies. *Stat. Appl. Genet. Mol. Biol.* **9** Art. 8, 22 pp. MR2594947

Li, Z. and Begg, C. B. (1994). Random effects models for combining results from controlled and uncontrolled studies in a meta-analysis. *J. Amer. Statist. Assoc.* **89** 1523–1527. MR1310241

Mila, A. L. and Ngugi, H. K. (2011). A Bayesian approach to meta-analysis of plant pathology studies. *Phytopathology* **101** 42–51.

Owen, A. B. (2009). Karl Pearson's meta-analysis revisited. *Ann. Statist.* **37** 3867–3892. MR2572446

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F. et al. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing *Nature* **464** 768–772.

Servin, B. and Stephens, M. (2008). Imputation-based analysis of association studies: Candidate regions and quantitative traits. *PLoS Genetics* **3** e114.

Stangl, D. K. and Berry, D. A. (2000). *Meta-Analysis in Medicine and Health Policy*. Dekker, New York.

Stephens, M. (2013). A unified framework for association analysis with multiple related phenotypes. *PLoS One* **8** e65245.

Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C. P. et al. (2007). Population genomics of human gene expression. *Nat. Genet.* **39** 1217–1224.

Sutton, A. J. and Abrams, K. R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Stat. Methods Med. Res.* **10** 277–303.

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M. et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466** 707–713.

VERZILLI, C. J., SHAH, T., CASAS, J. P., CHAPMAN, J., SANDHU, M. et al. (2008). Bayesian meta-analysis of genetic association studies with different sets of markers. *Am. J. Hum. Genet.* **82** 859–872.

WAKEFIELD, J. (2009). Bayes factors for genome-wide association studies: Comparison with P-values. *Genet. Epidemiol.* **33** 79–86.

WEN, X. (2011). Bayesian analysis of genetic association data, accounting for heterogeneity. Ph.D. thesis, Dept. Statistics, Univ. Chicago.

WEN, X. and STEPHENS, M. (2014). Supplement to "Bayesian methods for genetic association analysis with heterogeneous subgroups: From meta-analyses to gene–environment interactions." DOI:10.1214/13-AOAS695SUPP.

WHITEHEAD, A. and WHITEHEAD, J. (1991). A general parametric approach to the meta-analysis of randomized clinical trials. *Stat. Med.* **10** 1665–1677.

WILLER, C. J., LI, Y. and ABECASIS, G. R. (2010). METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26** 2190–2191.

The Annals of Applied Statistics
2014, Vol. 8, No. 1, 204–231
DOI: 10.1214/13-AOAS713

MATCHING FOR BALANCE, PAIRING FOR HETEROGENEITY IN AN OBSERVATIONAL STUDY OF THE EFFECTIVENESS OF FOR-PROFIT AND NOT-FOR-PROFIT HIGH SCHOOLS IN CHILE

BY JOSÉ R. ZUBIZARRETA, RICARDO D. PAREDES AND PAUL R. ROSENBAUM

Columbia University, Pontificia Universidad Católica de Chile and University of Pennsylvania

Conventionally, the construction of a pair-matched sample selects treated and control units and pairs them in a single step with a view to balancing observed covariates **x** and reducing the heterogeneity or dispersion of treated-minus-control response differences, Y. In contrast, the method of cardinality matching developed here first selects the maximum number of units subject to covariate balance constraints and, with a balanced sample for **x** in hand, then separately pairs the units to minimize heterogeneity in Y. Reduced heterogeneity of pair differences in responses Y is known to reduce sensitivity to unmeasured biases, so one might hope that cardinality matching would succeed at both tasks, balancing **x**, stabilizing Y. We use cardinality matching in an observational study of the effectiveness of for-profit and not-for-profit private high schools in Chile—a controversial subject in Chile—focusing on students who were in government run primary schools in 2004 but then switched to private high schools. By pairing to minimize heterogeneity in a cardinality match that has balanced covariates, a meaningful reduction in sensitivity to unmeasured biases is obtained.

REFERENCES

ANGRIST, J. D., PATHAK, P. A. and WALTERS, C. R. (2013). Explaining charter school effectiveness. *Am. Econ. J.* **5** 1–27.

BAIOCCHI, M. (2011). Designing robust studies using propensity score and prognostic score matching. Chapter 3 in *Methodologies for Observational Studies of Health Care Policy*. Ph.D. thesis, Dept. Statistics, The Wharton School, Univ. Pennsylvania, Philadelphia, PA.

BELLEI, C. (2009). Does lengthening the school day increase students academic achievement? Results from a natural experiment in Chile. *Econ. Educ. Rev.* **28** 629–640.

BROWN, B. M. (1981). Symmetric quantile averages and related estimators. *Biometrika* **68** 235–242. MR0614960

CORNFIELD, J., HAENSZEL, W., HAMMOND, E., LILIENFELD, A., SHIMKIN, M. and WYNDER, E. (1959). Smoking and lung cancer. *J. Natl. Cancer Inst.* **22** 173–203.

COX, D. R. (1958). *Planning of Experiments. A Wiley Publication in Applied Statistics*. Wiley, New York. MR0095561

CRUMP, R. K., HOTZ, V. J., IMBENS, G. W. and MITNIK, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96** 187–199. MR2482144

DEYO, R. A., CHERKIN, D. C. and CIOL, M. A. (1992). Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. *J. Clin. Epidemiol.* **45** 613–619.

Key words and phrases. Design sensitivity, integer programming, testing twice.

ELACQUA, G. (2009). *The Impact of School Choice and Public Policy on Segregation*: *Evidence from Chile*. Centro de Políticas Comparadas de Educación, Univ. Diego Portales, Santiago, Chile.
FISHER, R. A. (1935). *The Design of Experiments*. Oliver & Boyd, Edinburgh.
GASTWIRTH, J. L. (1992). Methods for assessing the sensitivity of statistical comparisons used in Title VII cases to omitted variables. *Jurimetrics* **33** 19–34.
HANSEN, B. B. (2007). Optmatch: Flexible, optimal matching for observational studies. *R News* **7** 18–24. (Package `optmatch` in `R`).
HANSEN, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika* **95** 481–488. MR2521594
HILL, J. and SU, Y.-S. (2013). Assessing lack of common support in causal inference using Bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *Ann. Appl. Stat.* **7** 1386–1420. MR3127952
HODGES, J. L. JR. and LEHMANN, E. L. (1963). Estimates of location based on rank tests. *Ann. Math. Statist.* **34** 598–611. MR0152070
HOSMAN, C. A., HANSEN, B. B. and HOLLAND, P. W. (2010). The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. *Ann. Appl. Stat.* **4** 849–870. MR2758424
IACUS, S. M., KING, G. and PORRO, G. (2009). Software for coarsened exact matching. *J. Stat. Softw.* **30** 1–27.
KNAUS, W. A., DRAPER, E. A., WAGNER, D. P. and ZIMMERMAN, J. E. (1985). APACHE II: A severity of disease classification system. *Crit. Care Med.* **13** 818–829.
LEHMANN, E. L. (1975). *Nonparametrics*. Holden-Day, San Francisco, CA.
LU, B., GREEVY, R., XU, X. and BECK, C. (2011). Optimal nonbipartite matching and its statistical applications. *Amer. Statist.* **65** 21–30. (Package `nbpmatching` in `R`). MR2899649
MARCUS, S. M. (1997). Using omitted variable bias to assess uncertainty in the estimation of an AIDS education treatment effect. *J. Educ. Statist.* **22** 193–201.
MARITZ, J. S. (1979). A note on exact robust confidence intervals for location. *Biometrika* **66** 163–166. MR0529161
NEYMAN, J. (1923, 1990). On the application of probability theory to agricultural experiments. *Statist. Sci.* **5** 463–480.
ROSENBAUM, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika* **74** 13–26. MR0885915
ROSENBAUM, P. R. (1993). Hodges–Lehmann point estimates of treatment effect in observational studies. *J. Amer. Statist. Assoc.* **88** 1250–1253. MR1245357
ROSENBAUM, P. R. (2002). *Observational Studies*, 2nd ed. Springer, New York. MR1899138
ROSENBAUM, P. R. (2004). Design sensitivity in observational studies. *Biometrika* **91** 153–164. MR2050466
ROSENBAUM, P. R. (2005). Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *Amer. Statist.* **59** 147–152. MR2133562
ROSENBAUM, P. R. (2007). Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies. *Biometrics* **63** 456–464. (`R` package `sensitivitymv`). MR2370804
ROSENBAUM, P. R. (2010a). *Design of Observational Studies*. *Springer Series in Statistics*. Springer, New York. MR2561612
ROSENBAUM, P. R. (2010b). Design sensitivity and efficiency in observational studies. *J. Amer. Statist. Assoc.* **105** 692–702. MR2724853
ROSENBAUM, P. R. (2011). A new U-statistic with superior design sensitivity in matched observational studies. *Biometrics* **67** 1017–1027. MR2829236
ROSENBAUM, P. R. (2012a). Testing one hypothesis twice in observational studies. *Biometrika* **99** 763–774. MR2999159

ROSENBAUM, P. R. (2012b). Optimal matching of an optimally chosen subset in observational studies. *J. Comput. Graph. Statist.* **21** 57–71. MR2913356

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

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

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

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

RUBIN, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *J. Amer. Statist. Assoc.* **74** 318–328.

SMALL, D. S. (2007). Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *J. Amer. Statist. Assoc.* **102** 1049–1058. MR2411664

STEPHENSON, W. R. (1981). A general class of one-sample nonparametric test statistics based on subsamples. *J. Amer. Statist. Assoc.* **76** 960–966. MR0650912

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

TRASKIN, M. and SMALL, D. S. (2011). Defining the study population for an observational study to ensure suffient overlap: A tree approach. *Statist. Biosci.* **3** 94–118.

WANG, L. and KRIEGER, A. M. (2006). Causal conclusions are most sensitive to unobserved binary covariates. *Stat. Med.* **25** 2257–2271. MR2240099

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

WOLFE, D. A. (1974). A characterization of population weighted-symmetry and related results. *J. Amer. Statist. Assoc.* **69** 819–822. MR0426239

YANAGAWA, T. (1984). Case–control studies: Assessing the effect of a confounding factor. *Biometrika* **71** 191–194. MR0738341

YANG, D., SMALL, D. S., SILBER, J. H. and ROSENBAUM, P. R. (2012). Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes. *Biometrics* **68** 628–636. (R package `finebalance`). MR2959630

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

ZUBIZARRETA, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *J. Amer. Statist. Assoc.* **107** 1360–1371. (R software `mip-match` at http://www-stat.wharton.upenn.edu/~josezubi/). MR3036400

ZUBIZARRETA, J. R., PAREDES, R. D. and ROSENBAUM, P. R. (2014). Supplement to: "Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile." DOI:10.1214/13-AOAS713SUPP.

ZUBIZARRETA, J. R., REINKE, C. E., KELZ, R. R., SILBER, J. H. and ROSENBAUM, P. R. (2011). Matching for several sparse nominal variables in a case–control study of readmission following surgery. *Amer. Statist.* **65** 229–238. MR2867507

The Annals of Applied Statistics
2014, Vol. 8, No. 1, 232–258
DOI: 10.1214/13-AOAS678

USING INFORMATIVE PRIORS IN THE ESTIMATION OF MIXTURES OVER TIME WITH APPLICATION TO AEROSOL PARTICLE SIZE DISTRIBUTIONS

BY DARREN WRAITH*, KERRIE MENGERSEN†, CLAIR ALSTON†, JUDITH ROUSSEAU‡ AND TAREQ HUSSEIN§

Institut National de Recherche en Informatique et en Automatique (INRIA), Queensland University of Technology†, CREST, INSEE and Université de Paris Dauphine‡, and University of Helsinki and University of Jordan§*

The issue of using informative priors for estimation of mixtures at multiple time points is examined. Several different informative priors and an independent prior are compared using samples of actual and simulated aerosol particle size distribution (PSD) data. Measurements of aerosol PSDs refer to the concentration of aerosol particles in terms of their size, which is typically multimodal in nature and collected at frequent time intervals. The use of informative priors is found to better identify component parameters at each time point and more clearly establish patterns in the parameters over time. Some caveats to this finding are discussed.

REFERENCES

AALTO, P., HÄMERI, K., BECKER, E., WEBER, R., SALM, J., MAKELA, J. M., HOELL, C., O'DOWD, C., KARLSSON, H., HANSSON, H. C., VAKEVA, M., KOPONEN, I. K., BUZORIS, G. and KULMALA, M. (2001). Physical characterization of aerosol particles during nucleation events. *Tellus* **53** 344–358.

ALSTON, C. L. and MENGERSEN, K. L. (2010). Allowing for the effect of data binning in a Bayesian normal mixture model. *Comput. Statist. Data Anal.* **54** 916–923. MR2580926

ALSTON, C. L., MENGERSEN, K. L., ROBERT, C. P., THOMPSON, J. M., LITTLEFIELD, P. J., PERRY, D. and BALL, A. J. (2007). Bayesian mixture models in a longitudinal setting for analysing sheep CAT scan images. *Comput. Statist. Data Anal.* **51** 4282–4296. MR2364445

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

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

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

DAL MASO, M., KULMALA, M., RIIPINEN, I., WAGNER, R., HUSSEIN, T., AALTO, P. P. and LEHTINEN, K. E. J. (2005). Formation and growth of fresh atmospheric aerosols: Eight years of aerosol size distribution data from SMEAR II, Hyytiala, Finland. *Boreal Env. Res.* **10** 323–336.

DUNSON, D. B. (2006). Bayesian dynamic modelling of latent trait distributions. *Biostatistics* **7** 551–568.

Key words and phrases. Bayesian statistics, mixture models, time series, aerosol particle size distribution.

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

FERNÁNDEZ, C. and GREEN, P. J. (2002). Modelling spatially correlated data via mixtures: A Bayesian approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 805–826. MR1979388

FRÜHWIRTH-SCHNATTER, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *J. Amer. Statist. Assoc.* **96** 194–209. MR1952732

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

GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Anal.* **1** 515–533 (electronic). MR2221284

GODSILL, S., DOUCET, A. and WEST, M. (2001). Maximum a posteriori sequence estimation using Monte Carlo particle filters. *Ann. Inst. Statist. Math.* **53** 82–96. MR1777255

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

GUSTAFSON, P. and WALKER, L. J. (2003). An extension of the Dirichlet prior for the analysis of longitudinal multinomial data. *J. Appl. Stat.* **30** 293–310. MR1960779

HOFF, P. D. (2003). Nonparametric modelling of hierarchically exchangeable data. Technical report, Dept. Statistics, Univ. Washington.

HUSSEIN, T., HÄMERI, K., AALTO, P., PAATERO, P. and KULMALA, M. (2005). Modal structure and spatial–temporal variations of urban and suburban aerosols in Helsinki–Finland. *Atmos. Environ.* **39** 1655–1668.

IBRAHIM, J. G., CHEN, M.-H. and SINHA, D. (2003). On optimality properties of the power prior. *J. Amer. Statist. Assoc.* **98** 204–213. MR1965686

JASRA, A., HOLMES, C. C. and STEPHENS, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statist. Sci.* **20** 50–67. MR2182987

JI, C. (2009). Advances in Bayesian modelling and computation: Spatio-temporal processes, model assessment and advances in Bayesian modelling and computation: Spatio-temporal processes, model assessment and adaptive MCMC. Ph.D. thesis, Dept. Statistical Science, Univ. Duke.

KULMALA, M., VEHKAMÄKIA, H., PETÄJÄA, T., DAL MASO, M., LAURIA, A., BIRMILI, W. and MCMURRY, P. H. (2004). Formation and growth rates of ultrafine atmospheric particles: A review of observations. *J. Aerosol Sci.* **35** 143–176.

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

MARIN, J. M., MENGERSEN, K. and ROBERT, C. P. (2005). Bayesian modelling and inference on mixtures of distributions. In *Handbook of Statistics* **25** (D. Dey and C. R. Rao, eds.). Elsevier, Amsterdam.

MCMURRY, P. H. (2000). A review of atmospheric aerosol measurements. *Atmos. Environ.* **34** 1959–1999.

RICHARDSON, S. and GREEN, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **59** 731–792. MR1483213

ROBERT, C. P. and CASELLA, G. (2004). *Monte Carlo Statistical Methods*, 2nd ed. Springer, New York. MR2080278

ROUSSEAU, J. and MENGERSEN, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73** 689–710. MR2867454

SPERRIN, M., JAKI, T. and WIT, E. (2010). Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models. *Stat. Comput.* **20** 357–366. MR2725393

STEPHENS, M. (1997). Bayesian methods for mixtures of normal distributions. Ph.D. thesis, Dept. Statistics, Univ. Oxford.

STEPHENS, M. (2000). Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods. *Ann. Statist.* **28** 40–74. MR1762903

STRICKLAND, C. M., SIMPSON, D. P., TURNER, I. W., DENHAM, R. and MENGERSEN, K. L. (2011). Fast Bayesian analysis of spatial dynamic factor models for multitemporal remotely sensed imagery. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **60** 109–124. MR2758572

VESALA, T., HATAJA, J., AALTO, P., ALTIMIR, N., BUZORIUS, G., GARAM, E., HÄMERI, K., ILVESNIEMI, H., JOKINEN, V., KERONEN, P., LAHTI, T., MARKKANEN, T., MÄKELÄ, J. M., NIKINMAA, E., PALMROTH, S., PALVA, L., POHJA, T., PUMPANEN, J., RANNIK, U., SIIVOLA, E., YLITALO, H., HARI, P. and KULMALA, M. (1998). Long-term field measurements of atmospheric–surface interactions in boreal forest ecology, micrometerology, aerosol physics, and atmospheric chemistry. *Trends Heat Mass Momentum Transf.* **4** 17–35.

WEST, M. and HARRISON, J. (1997). *Bayesian Forecasting and Dynamic Models*, 2nd ed. Springer, New York. MR1482232

WHITBY, E., MCMURRY, P. H. and SHANKER, U. (1991). Modal aerosol dynamics modelling. Technical report, US Environment Protection Agency, Atmospheric Research and Exposure Assessment Laboratory.

WHITBY, E. and MCMURRY, P. H. (1997). Modal aerosol dynamics modeling. *Aerosol Sci. Technol.* **27** 673–688.

WRAITH, D., MENGERSEN, K., ALSTON, C., ROUSSEAU, J. and HUSSEIN, T. (2014). Supplement to "Using informative priors in the estimation of mixtures over time with application to aerosol particle size distributions." DOI:10.1214/13-AOAS678SUPP.

YAO, W. (2012). Model based labeling for mixture models. *Stat. Comput.* **22** 337–347. MR2865020

WORLD HEALTH ORGANIZATION (2006). WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide Global update 2005, World Health Organisation.

The Annals of Applied Statistics
2014, Vol. 8, No. 1, 259–285
DOI: 10.1214/13-AOAS680

γ-SUP: A CLUSTERING ALGORITHM FOR CRYO-ELECTRON MICROSCOPY IMAGES OF ASYMMETRIC PARTICLES

BY TING-LI CHEN*, DAI-NI HSIEH*, HUNG HUNG†, I-PING TU*, PEI-SHIEN WU‡, YI-MING WU*, WEI-HAU CHANG* AND SU-YUN HUANG*

Academia Sinica, National Taiwan University† and Duke University‡*

Cryo-electron microscopy (cryo-EM) has recently emerged as a powerful tool for obtaining three-dimensional (3D) structures of biological macromolecules in native states. A minimum cryo-EM image data set for deriving a meaningful reconstruction is comprised of thousands of randomly orientated projections of identical particles photographed with a small number of electrons. The computation of 3D structure from 2D projections requires clustering, which aims to enhance the signal to noise ratio in each view by grouping similarly oriented images. Nevertheless, the prevailing clustering techniques are often compromised by three characteristics of cryo-EM data: high noise content, high dimensionality and large number of clusters. Moreover, since clustering requires registering images of similar orientation into the same pixel coordinates by 2D alignment, it is desired that the clustering algorithm can label misaligned images as outliers. Herein, we introduce a clustering algorithm γ-SUP to model the data with a q-Gaussian mixture and adopt the minimum γ-divergence for estimation, and then use a self-updating procedure to obtain the numerical solution. We apply γ-SUP to the cryo-EM images of two benchmark macromolecules, RNA polymerase II and ribosome. In the former case, simulated images were chosen to decouple clustering from alignment to demonstrate γ-SUP is more robust to misalignment outliers than the existing clustering methods used in the cryo-EM community. In the latter case, the clustering of real cryo-EM data by our γ-SUP method eliminates noise in many views to reveal true structure features of ribosome at the projection level.

REFERENCES

ADRIAN, M., DUBOCHET, J., LEPAULT, J. and MCDOWALL, A. W. (1984). Cryo-electron microscopy of viruses. *Nature* **308** 32–36.

AMARI, S.-I. and OHARA, A. (2011). Geometry of q-exponential family of probability distributions. *Entropy* **13** 1170–1185. MR2811982

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

BASU, A., HARRIS, I. R., HJORT, N. L. and JONES, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika* **85** 549–559. MR1665873

BRACEWELL, R. N. (1956). Strip integration in radio astronomy. *Austral. J. Phys.* **9** 198–217. MR0080017

Key words and phrases. Clustering algorithm, cryo-EM images, γ-divergence, k-means, mean-shift algorithm, multilinear principal component analysis, q-Gaussian distribution, robust statistics, self-updating process.

CARREIRA-PERPIÑÁN, M. A. (2006). Fast nonparametric clustering with Gaussian blurring mean-shift. In *Proceeding of the 23rd International Conference on Machine Learning* 153–160. ACM, Pittsburgh, PA.

CHANG, W.-H., CHIU, M.-K., CHEN, C.-Y., YEN, C.-F., LIN, Y.-C., WENG, Y.-P., CHANG, J.-C., WU, Y.-M., CHENG, H., FU, J. and TU, I.-P. (2010). Zernike phase plate cryo-electron microscopy facilitates single particle analysis of unstained asymmetric protein complexes. *Structure* **18** 17–27.

CHEN, T.-L. (2013). On the convergence and consistency of the blurring mean-shift process. Available at arXiv:1305.1040.

CHEN, T.-L. and SHIU, S.-Y. (2007). A clustering algorithm by self-updating process. In *JSM Proceedings, Statistical Computing Section* 2034–2038. American Statistical Association, Salt Lake City, UT.

CHENG, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **17** 790–799.

CICHOCKI, A. and AMARI, S.-I. (2010). Families of alpha- beta- and gamma-divergences: Flexible and robust measures of similarities. *Entropy* **12** 1532–1568. MR2659408

CROWTHER, R. A., AMOS, L. A., FINCH, J. T., DE ROSIER, D. J. and KLUG, A. (1970). Three dimensional reconstructions of spherical viruses by Fourier synthesis from electron micrographs. *Nature* **226** 421–425.

DUBOCHET, J. (2012). Cryo-EM–the first thirty years. *J. Microsc.* **245** 221–224.

EGUCHI, S., KOMORI, O. and KATO, S. (2011). Projective power entropy and maximum Tsallis entropy distributions. *Entropy* **13** 1746–1764. MR2851127

FIELD, C. and SMITH, B. (1994). Robust estimation: A weighted maximum likelihood approach. *International Statistical Review* **62** 405–424.

FRANK, J. (2002). Single-particle imaging of macromolecules by cryo-electron microscopy. *Annu. Rev. Biophys. Biomol. Struct.* **31** 303–319.

FRANK, J. (2009). Single-particle reconstruction of biological macromolecules in electron microscopy—30 years. *Q. Rev. Biophys.* **42** 139–158.

FRANK, J. (2012). Intermediate states during mRNA-tRNA translocation. *Curr. Opin. Struct. Biol.* **22** 778–785.

FRANK, J., RADERMACHERA, M., PENCZEKA, P., ZHUA, J., LI, Y., LADJADJ, M. and LEITHA, A. (1996). SPIDER and WEB: Processing and visualization of images in 3D electron microscopy and related fields. *Journal of Structural Biology* **116** 190–199.

FRIGYIK, B. A., SRIVASTAVA, S. and GUPTA, M. R. (2008). Functional Bregman divergence and Bayesian estimation of distributions. *IEEE Trans. Inform. Theory* **54** 5130–5139. MR2589887

FUJISAWA, H. and EGUCHI, S. (2008). Robust parameter estimation with a small bias against heavy contamination. *J. Multivariate Anal.* **99** 2053–2081. MR2466551

FUKUNAGA, K. and HOSTETLER, L. D. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans. Inform. Theory* **IT-21** 32–40. MR0388638

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

GOOD, I. J. (1971). Comment on "Measuring information and uncertainty." In *Foundation of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.) 265–273. Holt, Rinehart and Winston, Toronto.

GRASSUCCI, R., TAYLOR, D. and FRANK, J. (2011). Preparation of macromolecular complexes for cryo-electron microscopy. *Nature Protocols* **2** 3239–3246.

HALL, R. J., NOGALES, E. and GLAESER, R. M. (2011). Accurate modeling of single-particle cryo-EM images quantitates the benefits expected from using Zernike phase contrast. *J. Struct. Biol.* **174** 468–475.

HARTIGAN, J. A. (1975). *Clustering Algorithms*. Wiley, New York. MR0405726

HENDERSON, R. (1995). The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Q. Rev. Biophys.* **28** 171–193.

HUNG, H., WU, P., TU, I. and HUANG, S. (2012). On multilinear principal component analysis of order-two tensors. *Biometrika* **99** 569–583. MR2966770

JIANG, W., BAKER, M. L., JAKANA, J., WEIGELE, P. R., KING, J. and CHIU, W. (2008). Backbone structure of the infectious epsilon15 virus capsid revealed by electron cryomicroscopy. *Nature* **451** 1130–1134.

LEPAULT, J., BOOY, F. P. and DUBOCHET, J. (1983). Electron microscopy of frozen biological suspensions. *J. Microsc.* **129** 89–102.

LIU, H., JIN, L., KOH, S. B. S., ATANASOV, I., SCHEIN, S., WU, L. and ZHOU, Z. H. (2010). Atomic structure of human adenovirus by cryo-EM reveals interactions among protein networks. *Science* **329** 1038–1043.

LLOYD, S. P. (1982). Least squares quantization in PCM. *IEEE Trans. Inform. Theory* **28** 129–137. MR0651807

LU, H., PLATANIOTIS, K. N. K. and VENETSANOPOULOS, A. N. (2008). MPCA: Multilinear principal component analysis of tensor objects. *IEEE Trans. Neural. Netw.* **19** 18–39.

MANNING, C., RAGHAVAN, P. and SCHTZE, H. (2008). *Introduction to Information Retrieval*. Cambridge Univ. Press, New York.

MCQUEEN, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 291–297. Univ. California Press, Berkeley, CA.

MOLLAH, M. N. H., SULTANA, N., MINAMI, M. and EGUCHI, S. (2010). Robust extraction of local structures by the minimum beta-divergence method. *Neural Networks* **23** 226–238.

SAIBIL, H. R. (2000). Macromolecular structure determination by cryo-electron microscopy. *Acta Crystallographica Section D-Biological Crystallography* **56** 1215–1222.

SHIU, S.-Y. and CHEN, T.-L. (2012). Clustering by self-updating process. Available at arXiv:1201.1979.

SINGER, A., COIFMAN, R. R., SIGWORTH, F. J., CHESTER, D. W. and SHKOLNISKY, Y. (2010). Detecting consistent common lines in cryo-EM by voting. *J. Struct. Biol.* **169** 312–322.

SORZANO, C. O. S., MARABINI, R., VELÁZQUEZ-MURIEL, J., BILBAO-CASTRO, J. R., SCHERES, S. H. W., CARAZO, J. M. and PASCUAL-MONTANO, A. (2004). XMIPP: A new generation of an open-source image processing package for electron microscopy. *J. Struct. Biol.* **148** 194–204.

SORZANO, C. O. S., BILBAO-CASTRO, J. R., SHKOLNISKY, Y., ALCORLO, M., MELERO, R., CAFFARENA-FERNANDEZ, G., LI, M., XU, G., MARABINI, R. and CARAZO, J. M. (2010). A clustering approach to multireference alignment of single-particle projections in electron microscopy. *Journal of Structural Biology* **171** 197–206.

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

VAN HEEL, M. (1987). Angular reconstitution: A posteriori assignment of projection directions for 3D reconstruction. *Ultramicroscopy* **21** 111–124.

VAN HEEL, M., GOWEN, B., MATADEEN, R., ORLOVA, E. V., FINN, R., PAPE, T., COHEN, D., STARK, H., SCHMIDT, R., SCHATZ, M. and PATWARDHAN, A. (2000). Single-particle electron cryo-microscopy: Towards atomic resolution. *Q. Rev. Biophys.* **33** 307–369.

WILSOME, D. and CATE, J. (2012). The structure and function of the eukaryotic ribosome. *Cold Spring Harbor Perspectives in Biology* **4** a011536.

WINDHAM, M. P. (1995). Robustifying model fitting. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57** 599–609. MR1341326

YANG, Z., FANG, J., CHITTULURU, J., ASTURIAS, F. J. and PENCZEK, P. A. (2012). Iterative stable alignment and clustering of 2D transmission electron microscope images. *Structure* **20** 237–247.

The Annals of Applied Statistics
2014, Vol. 8, No. 1, 286–308
DOI: 10.1214/13-AOAS686

APPLYING MULTIPLE TESTING PROCEDURES TO DETECT CHANGE IN EAST AFRICAN VEGETATION

BY NICOLLE CLEMENTS[*], SANAT K. SARKAR[†], ZHIGEN ZHAO[†] AND DONG-YUN KIM[‡,§]

Saint Joseph's University[*], Temple University[†], National Institute of Health[‡] and Virginia Tech[§]

The study of vegetation fluctuations gives valuable information toward effective land use and development. We consider this problem for the East African region based on the Normalized Difference Vegetation Index (NDVI) series from satellite remote sensing data collected between 1982 and 2006 over 8-kilometer grid points. We detect areas with significant increasing or decreasing monotonic vegetation changes using a multiple testing procedure controlling the mixed directional false discovery rate (mdFDR). Specifically, we use a three-stage directional Benjamini–Hochberg (BH) procedure with proven mdFDR control under independence and a suitable adaptive version of it. The performance of these procedures is studied through simulations before applying them to the vegetation data. Our analysis shows increasing vegetation in the Northern hemisphere as well as coastal Tanzania and generally decreasing Southern hemisphere vegetation trends, which are consistent with historical evidence.

REFERENCES

ABELSON, R. P. and TUKEY, J. W. (1963). Efficient utilization of non-numerical information in quantitative analysis: General theory and the case of simple order. *Ann. Math. Statist.* **34** 1347–1369. MR0156411

BENJAMINI, Y. and HELLER, R. (2007). False discovery rates for spatial signals. *J. Amer. Statist. Assoc.* **102** 1272–1281. MR2412549

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

BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245

BENJAMINI, Y. and YEKUTIELI, D. (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters. *J. Amer. Statist. Assoc.* **100** 71–93. MR2156820

BRILLINGER, D. R. (1989). Consistent detection of a monotonic trend superposed on a stationary time series. *Biometrika* **76** 23–30. MR0991419

CHEN, J., JONSSON, P. and TAMURA, M. (2004). A simple method for reconstructing a high-quality NDVI time-series data set based on the Savitzky–Golay filter. *Remote Sensing of Environment* **91** 332–344.

CLEMENTS, N., SARKAR, S. K. and GUO, W. (2011). Astronomical transient detection controlling the false discovery rate. In *Statistical Challenges in Modern Astronomy V* (E. D. Feigelson and G. J. Babu, eds.) 383–396. Springer, New York.

Key words and phrases. False discovery rate, directional false discovery rate, NDVI, East Africa vegetation.

COLE, J. E., DUNBAR, R. B., MCCLANAHAN, T. R. and MUTHIGA, N. A. (2000). Tropical pacific forcing of decadal SST variability in the western Indian Ocean over the past two centuries. *Science* **287** 617–619.

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

CURRAN, P. J. (1980). Multispectral remote sensing of vegetation amount. *Progress in Physical Geography* **4** 315–341.

DUVEILLER, G., DEFOURNY, P., DESCLEE, B. and MAYAUX, P. (2007). Deforestation in Central Africa: Estimates at regional, national and landscape levels by advanced processing of systematically-disturbed Landsat extracts. *Remote Sensing of Environment* **112** 1969–1981.

FOODY, G. M. (2003). Geographical weighting as a further refinement to regression modeling: An example focused on the NDVI–rainfall relationship. *Remote Sensing of Environment* **88** 283–293.

GUO, W. and SARKAR, S. (2012). Adaptive controls of the FWER and FDR under block dependence. Unpublished manuscript. Available at http://web.njit.edu/~wguo/research.html.

HAYES, D. J. and SADER, S. A. (2001). Comparison of change-detection techniques for monitoring tropical forest clearing and vegetation regrowth in a time series. *Photogrammetric Engineering and Remote Sensing* **67** 1067–1075.

JACKSON, R. D., SLATER, P. N. and PINTER, P. J. (1983). Discrimination of growth and water stress in wheat by various vegetation indices through clear and turbid atmospheres. *Remote Sensing of Environment* **13** 187–208.

OCHA (2011). Eastern Africa drought humanitarian report No. 3. *OCHA, UN Office for the Coordination of Humanitarian Affairs* reliefweb.int.

PACIFICO, M. P., GENOVESE, C., VERDINELLI, I. and WASSERMAN, L. (2004). False discovery control for random fields. *J. Amer. Statist. Assoc.* **99** 1002–1014. MR2109490

SARKAR, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Ann. Statist.* **30** 239–257. MR1892663

STOREY, J. D., TAYLOR, J. E. and SIEGMUND, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66** 187–205. MR2035766

TOBLER, W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography* **46** 234–240.

TUCKER, C., PINZON, J., BROWN, M., SLAYBACK, D., PAK, E., MAHONEY, R., VERMOTE, E. and SALEOUS, N. (2005). An extended AVHRR 8-km NDVI data set compatible with MODIS and SPOT vegetation NDVI data. *International Journal of Remote Sensing* **26** 4485–4498.

USONGO, L. and NAGAHUEDI, J. (2008). Participatory land-use planning for priority landscapes of the Congo Basin. *Unasylva* **230** 17–24.

VRIELING, A., DE BEURS, K. M. and BROWN, M. E. (2008). Recent trends in agricultural production of Africa based on AVHRR NDVI time series. *Proceedings of the SPIE Conference: Remote Sensing for Agriculture, Ecosystems and Hydrology X*.

The Annals of Applied Statistics
2014, Vol. 8, No. 1, 309–330
DOI: 10.1214/13-AOAS687

QUANTIFYING ALTERNATIVE SPLICING FROM PAIRED-END RNA-SEQUENCING DATA

BY DAVID ROSSELL* CAMILLE STEPHAN-OTTO ATTOLINI† MANUEL KROISS‡,§ AND ALMOND STÖCKER‡

University of Warwick, Institute for Research in Biomedicine of Barcelona†, LMU Munich‡ and TU Munich§*

RNA-sequencing has revolutionized biomedical research and, in particular, our ability to study gene alternative splicing. The problem has important implications for human health, as alternative splicing may be involved in malfunctions at the cellular level and multiple diseases. However, the high-dimensional nature of the data and the existence of experimental biases pose serious data analysis challenges. We find that the standard data summaries used to study alternative splicing are severely limited, as they ignore a substantial amount of valuable information. Current data analysis methods are based on such summaries and are hence suboptimal. Further, they have limited flexibility in accounting for technical biases. We propose novel data summaries and a Bayesian modeling framework that overcome these limitations and determine biases in a nonparametric, highly flexible manner. These summaries adapt naturally to the rapid improvements in sequencing technology. We provide efficient point estimates and uncertainty assessments. The approach allows to study alternative splicing patterns for individual samples and can also be the basis for downstream analyses. We found a severalfold improvement in estimation mean square error compared popular approaches in simulations, and substantially higher consistency between replicates in experimental data. Our findings indicate the need for adjusting the routine summarization and analysis of alternative splicing RNA-seq studies. We provide a software implementation in the R package `casper`.[4]

REFERENCES

AMEUR, A., WETTERBOM, A., FEUK, L. and GYLLENSTEN, U. (2010). Global and unbiased detection of splice junctions from RNA-seq data. *Genome Biol*. **11** R34.

BLENCOWE, B. J. (2006). Alternative splicing: New insights from global analyses. *Cell* **126** 37–47.

CASELLA, G. and BERGER, R. L. (2001). *Statistical Inference*, 2nd ed. Duxbury, N. Scituate.

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

ENCODE PROJECT CONSORTIUM (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306** 636–640.

GLAUS, P., HONKELA, A. and RATTRAY, M. (2012). Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics* **28** 1721–1728.

Key words and phrases. Alternative splicing, RNA-Seq, Bayesian modeling, estimation.

[4] http://www.bioconductor.org/packages/release/bioc/html/casper.html.

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

HOLT, R. A. and JONES, S. J. M. (2008). The new paradigm of flow cell sequencing. *Genome Research* **18** 839–846.

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

KAPLAN, E. L. and MEIER, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53** 457–481. MR0093867

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

LACROIX, V., SAMMETH, M., GUIGO, R. and BERGERON, A. (2008). Exact Transcriptome Reconstruction from Short Sequence Reads. In *Proceedings of the* 8*th International Workshop on Algorithms in Bioinformatics*. 50–63. Springer, Berlin.

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

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

LI, R., YU, C., LI, Y., LAM, T. W., YIU, S. M., KRISTIANSEN, K. and WANG, J. (2009). SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics* **25** 1966–1967.

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.

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

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

ROBERTS, A., TRAPNELL, C., DONAGHEY, J., RINN, J. L. and PACHTER, L. (2011a). Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol.* **12** R22.

ROBERTS, A., PIMENTEL, H., TRAPNELL, C. and PACHTER, L. (2011b). Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27** 2325–2329.

ROGERS, M. F., THOMAS, J., REDDY, A. S. and BEN-HUR, A. (2012). SpliceGrapher: Detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data. *Genome Biol.* **13** R4.

ROSSELL, D., STEPHAN-OTTO ATTOLINI, C., KROISS, M. and STÖCKER, A. (2014). Supplement to "Quantifying alternative splicing from paired-end RNA-sequencing data." DOI:10.1214/13-AOAS687SUPP.

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

THERNEAU, T. and LUMLEY, T. (2011). Survival: Survival analysis, including penalised likelihood. R package version 2.36-10.

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

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

TRAPNELL, C., ROBERTS, A., GOFF, L., PERTEA, G., KIM, D., KELLEY, D. R., PIMENTEL, H., SALZBERG, S. L., RINN, J. L. and PACHTER, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* **7** 562–578.

WU, Z., WANG, X. and ZHANG, X. (2011). Using non-uniform read distribution models to improve isoform expression inference in RNA-Seq. *Bioinformatics* **27** 502–508.

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

XING, Y., YU, T., WU, Y. N., ROY, M., KIM, J. and LEE, C. (2006). An expectation–maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. *Nucleic. Acids Res*. **34** 3150–3160.

The Annals of Applied Statistics
2014, Vol. 8, No. 1, 331–351
DOI: 10.1214/13-AOAS688

A SEMI-PARAMETRIC BAYESIAN MODEL OF INTER- AND INTRA-EXAMINER AGREEMENT FOR PERIODONTAL PROBING DEPTH

BY E. G. HILL AND E. H. SLATE

Medical University of South Carolina and Florida State University

Periodontal probing depth is a measure of periodontitis severity. We develop a Bayesian hierarchical model linking true pocket depth to both observed and recorded values of periodontal probing depth, while permitting correlation among measures obtained from the same mouth and between duplicate examiners' measures obtained at the same periodontal site. Periodontal site-specific examiner effects are modeled as arising from a Dirichlet process mixture, facilitating identification of classes of sites that are measured with similar bias. Using simulated data, we demonstrate the model's ability to recover examiner site-specific bias and variance heterogeneity and to provide cluster-adjusted point and interval agreement estimates. We conclude with an analysis of data from a probing depth calibration training exercise.

REFERENCES

ALBANDAR, J. M., BRUNELLE, J. A. and KINGMAN, A. (1999). Destructive periodontal disease in adults 30 years of age and older in the United States, 1988–1994. *J. Periodontol*. **70** 13–29.

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

CELEUX, G., FORBES, F., ROBERT, C. P. and TITTERINGTON, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Anal*. **1** 651–673 (electronic). MR2282197

COCHRAN, W. G. (1977). *Sampling Techniques*, 3rd ed. Wiley, New York. MR0474575

COHEN, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull*. **70** 213–220.

CONGDON, P. (2001). *Bayesian Statistical Modelling*. Wiley, Chichester. MR1852012

CONGDON, P. (2007). Bayesian modelling strategies for spatially varying regression coefficients: A multivariate perspective for multiple outcomes. *Comput. Statist. Data Anal*. **51** 2586–2601. MR2338990

DAHL, D. B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. In *Bayesian Inference for Gene Expression and Proteomics* 201–218. Cambridge Univ. Press, New York.

DEROUEN, T. A., MANCL, L. and HUJOEL, P. (1991). Measurement of associations in periodontal diseases using statistical methods for dependent data. *J. Periodont. Res*. **26** 218–229.

EISENHAUER, E. A., THERASSE, P., BOGAERTS, J., SCHWARTZ, L. H., SARGENT, D., FORD, R., DANCEY, J., ARBUCK, S., GWYTHER, S., MOONEY, M., RUBINSTEIN, L., SHANKAR, L., DODD, L., KAPLAN, R., LACOMBE, D. and VERWEIJ, J. (2009). New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *Eur. J. Cancer* **45** 228–247.

Key words and phrases. Agreement, cluster-correlated data, clustering, Dirichlet process mixture model, measurement error, periodontal disease, weighted kappa.

ESCOBAR, M. D. and WEST, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588. MR1340510

ESCOBAR, M. D. and WEST, M. (1998). Computing nonparametric hierarchical models. In *Practical Nonparametric and Semiparametric Bayesian Statistics* 1–22. Springer, New York. MR1630073

FLEISS, J. L. (2001). *Statistical Methods for Rates and Proportions*. Wiley, New York.

FLEISS, J. L., MANN, J., PAIK, M., GOULTCHIN, J. and CHILTON, N. W. (1991). A study of inter- and intra-examiner reliability of pocket depth and attachment level. *J. Periodont. Res.* **26** 122–128.

GUGGENMOOS-HOLZMANN, I. and VONK, R. (1998). Kappa-like indices of observer agreement viewed from a latent class perspective. *Stat. Med.* **17** 797–812.

HILL, E. G. and SLATE, E. H. (2014). Supplement to "A semi-parametric Bayesian model of inter- and intra-examiner agreement for periodontal probing depth." DOI:10.1214/13-AOAS688SUPP.

HILL, E. G., SLATE, E. H., WIEGAND, R. E., GROSSI, S. G. and SALINAS, C. F. (2006). Study design for calibration of clinical examiners measuring periodontal parameters. *J. Periodontol.* **77** 1129–1141.

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

JONES, G. L., HARAN, M., CAFFO, B. S. and NEATH, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *J. Amer. Statist. Assoc.* **101** 1537–1547. MR2279478

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

ODEN, N. L. (1991). Estimating kappa from binocular data. *Stat. Med.* **10** 1303–1311.

OSBORN, J. B., STOLTENBERG, J. L., HUSO, B. A., AEPPLI, D. M. and PHILSTROM, B. L. (1992). Comparison of measurement variability in subjects with moderate periodontitis using a conventional and constant force periodontal probe. *J. Periodontol.* **63** 283–289.

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

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

SLATE, E. H. and HILL, E. G. (2012). Discovering factors influencing examiner agreement for periodontal measures. *Community Dentistry and Oral Epidemiology* **40** 21–27. Suppl. 1.

WILLIAMSON, J. M., LIPSITZ, S. R. and MANATUNGA, A. K. (2000). Modeling kappa for measuring dependent categorical agreement data. *Biostatistics* **1** 191–202.

WILLIAMSON, J. M. and MANATUNGA, A. K. (1997). Assessing interrater agreement from dependent data. *Biometrics* **53** 707–714.

The Annals of Applied Statistics
2014, Vol. 8, No. 1, 352–376
DOI: 10.1214/13-AOAS690

JOINT ANALYSIS OF SNP AND GENE EXPRESSION DATA IN GENETIC ASSOCIATION STUDIES OF COMPLEX DISEASES

BY YEN-TSUNG HUANG*, TYLER J. VANDERWEELE† AND XIHONG LIN†

Brown University and Harvard University†*

Genetic association studies have been a popular approach for assessing the association between common Single Nucleotide Polymorphisms (SNPs) and complex diseases. However, other genomic data involved in the mechanism from SNPs to disease, for example, gene expressions, are usually neglected in these association studies. In this paper, we propose to exploit gene expression information to more powerfully test the association between SNPs and diseases by jointly modeling the relations among SNPs, gene expressions and diseases. We propose a variance component test for the total effect of SNPs and a gene expression on disease risk. We cast the test within the causal mediation analysis framework with the gene expression as a potential mediator. For eQTL SNPs, the use of gene expression information can enhance power to test for the total effect of a SNP-set, which is the combined direct and indirect effects of the SNPs mediated through the gene expression, on disease risk. We show that the test statistic under the null hypothesis follows a mixture of χ^2 distributions, which can be evaluated analytically or empirically using the resampling-based perturbation method. We construct tests for each of three disease models that are determined by SNPs only, SNPs and gene expression, or include also their interactions. As the true disease model is unknown in practice, we further propose an omnibus test to accommodate different underlying disease models. We evaluate the finite sample performance of the proposed methods using simulation studies, and show that our proposed test performs well and the omnibus test can almost reach the optimal power where the disease model is known and correctly specified. We apply our method to reanalyze the overall effect of the SNP-set and expression of the *ORMDL3* gene on the risk of asthma.

REFERENCES

CAI, T., LIN, X. and CARROLL, R. J. (2012). Identifying genetic marker sets associated with phenotypes via an efficient adaptive score test. *Biostatistics* **13** 776–790.

CHEUNG, V. G., SPIELMAN, R. S., EWENS, K. G., WEBER, T. M., MORLEY, M. and BURDICK, J. T. (2005). Mapping determinants of human gene expression by regional and genome-wide association. *Nature* **437** 1365–1369.

CUSANOVICH, D. A., BILLSTRAND, C., ZHOU, X., CHAVARRIA, C., LEON, S. D., MICHELINI, K. ET AL. (2012). The combination of a genome-wide association study of lymphocyte count and analysis of gene expression data reveals novel asthma candidate genes. *Hum. Mol. Genet.* **21** 2111–2123.

Key words and phrases. Causal inference, data integration, mediation analysis, mixed models, score test, SNP set analysis, variance component test.

DAVIES, R. (1980). The distribution of a linear combination of chi-square random variables. *Appl. Stat*. **29** 323–333.

DERMITZAKIS, E. T. (2008). From gene expression to disease risk. *Nat. Genet*. **40** 492–493.

DICKSON, S. P., WANG, K., KRANTZ, I., HAKONARSON, H. and GOLDSTEIN, D. B. (2010). Rare variants create synthetic genome-wide associations. *PLoS Biol*. **8** e1000294.

DIXON, A. L., LIANG, L., MOFFATT, M. F., CHEN, W., HEATH, S., WONG, K. C. C. ET AL. (2007). A genome-wide association study of global gene expression. *Nat. Genet*. **39** 1202–1207.

FU, J., KEURENTJES, J. J. B., BOUWMEESTER, H., AMERICA, T., VERSTAPPEN, F. W. A., WARD, J. L., BEALE, M. H., DE VOS, R. C. H., DIJKSTRA, M., SCHELTEMA, R. A., JOHANNES, F., KOORNNEEF, M., VREUGDENHIL, D., BREITLING, R. and JANSEN, R. C. (2009). System-wide molecular evidence for phenotypic buffering in Arabidopsis. *Nat. Genet*. **41** 166–167.

HAGEMAN, R. S., LEDUC, M. S., KORSTANJE, R., PAIGEN, B. and CHURCHILL, G. A. (2011). A Bayesian framework for inference of the genotype–phenotype map for segregating populations. *Genetics* **187** 1163–1170.

HSU, Y. H., ZILLILKENS, M., WILSON, S., FARBER, C., DEMISSIE, S., SORANZO, N. ET AL. (2010). An integration of genome-wide association study and expression profiling to prioritize the discovery of susceptibility loci for osteoporosis-related traits. *PLoS Genet*. **6** e1000977.

HUANG, Y. T., VANDERWEELE, T. J. and LIN, X. (2013). Supplement to "Joint analysis of SNP and gene expression data in genetic association studies of complex diseases." DOI:10.1214/13-AOAS690SUPP.

HUNTER, D. and CHANOCK, S. (2010). Genome-wide association studies and "the art of the soluble". *J. Natl. Cancer Inst*. **102** 1–2.

IMAI, K., KEELE, L. and YAMAMOTO, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statist. Sci*. **25** 51–71. MR2741814

INNOCENTI, F., COOPER, G. M., STANAWAY, I. B., GAMAZON, E. R., SMITH, J. D., MIRKOV, S. ET AL. (2011). Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS Genet*. **7** e1002078.

JOHANNES, F., COLOT, V. and JANSEN, R. C. (2008). Epigenome dynamics: A quantitative genetics perspective. *Nat. Rev. Genet*. **9** 883–890.

KLINE, P. and SANTOS, A. (2012). A score based approach to wild bootstrap inference. *Journal of Econometric Methods* **1** 23–41.

KWEE, L. C., LIU, D., LIN, X., GHOSH, D. and EPSTEIN, M. P. (2008). A powerful and flexible multilocus association test for quantitative traits. *Am. J. Hum. Genet*. **82** 386–397.

LEE, P. H. and SHATKAY, H. (2008). F-SNP: Computationally predicted functional SNPs for disease association studies. *Nucleic Acids Res*. **36** D820–D824.

LI, Y., ALVAREZ, O. A., GUTTELING, E. W., TIJSTERMAN, M., FU, J., RIKSEN, J. A., HAZENDONK, E., PRINS, P., PLASTERK, R. H., JANSEN, R. C., BREITLING, R. and KAMMENGA, J. E. (2006). Mapping determinants of gene expression plasticity by genetical genomics in *C. elegans*. *PLoS Genet*. **2** e222.

LI, Y., TESSON, B. M., CHURCHILL, G. A. and JANSEN, R. C. (2010). Critical reasoning on causal inference in genome-wide linkage and association studies. *Trends Genet*. **26** 493–498.

LIN, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika* **84** 309–326. MR1467049

MARCHINI, J., HOWIE, B., MYERS, S., MCVEAN, G. and DONNELLY, P. (2007). A new multipoint method for genome-wide association studies via imputation of genotypes. *Nat. Genet*. **39** 906–913.

MOFFATT, M. F., KABESCH, M., LIANG, L., DIXON, A. L., STRACHAN, D., HEATH, S. ET AL. (2007). Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* **448** 470–473.

MORLEY, M., MOLONY, C. M., WEBER, T. M., DEVLIN, J. L., EWENS, K. G., SPIELMAN, R. S. ET AL. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature* **430** 743–747.

NETO, E. C., BROMAN, A. T., KELLER, M. P., ATTIE, A. D., ZHANG, B., ZHU, J. and YANDELL, B. S. (2013). Modeling causality for pairs of phenotypes in system genetics. *Genetics* **193** 1003–1013.

PARZEN, M. I., WEI, L. J. and YING, Z. (1994). A resampling method based on pivotal estimating functions. *Biometrika* **81** 341–350. MR1294895

PEARL, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence* 411–420. Morgan Kaufmann, San Francisco.

ROBINS, J. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. In *Highly Structured Stochastic Systems* (P. Green, N. L. Hjort and S. Richardson, eds.) 70–81. Oxford Univ. Press, Oxford. MR2082403

ROBINS, J. M. and GREENLAND, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3** 143–155.

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

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

SATTERTHWAITE, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics* **2** 110–114.

SCHADT, E. E., MONKS, S. A., DRAKE, T. A., LUSIS, A. J., CHE, N., COLINAYO, V. ET AL. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422** 297–302.

SCHADT, E. E., LAMB, J., YANG, X., ZHU, J., EDWARDS, S., GUHATHAKURTA, D. ET AL. (2005). An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.* **37** 710–717.

SMITH, D. G. and EBRAHIM, S. (2003). Mendelian randomization: Can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32** 1–22.

SMITH, D. G. and EBRAHIM, S. (2005). What can Mendelian randomisation tell us about modifiable behavioural and environmental exposures? *British Medical Journal* **330** 1076–1079.

STOREY, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 479–498. MR1924302

VANDERWEELE, T. J. and VANSTEELANDT, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Stat. Interface* **2** 457–468. MR2576399

VANDERWEELE, T. J. and VANSTEELANDT, S. (2010). Odds ratios for mediation analysis for a dichotomous outcome. *Am. J. Epidemiol.* **172** 1339–1348.

WU, M., KRAFT, P., EPSTEIN, M., TAYLOR, D., CHANOCK, S., HUNTER, D. ET AL. (2010). Powerful SNP set analysis for case–control genomewide association studies. *Am. J. Hum. Genet.* **86** 929–942.

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

ZHANG, M., LIANG, L., MORAR, N., DIXON, A. L., LATHROP, G. M., DING, J. ET AL. (2012). Integrating pathway analysis and genetics of gene expression for genome-wide association study of basal cell carcinoma. *Hum. Genet.* **131** 615–623.

ZHONG, H., BEAULAURIER, J., LUM, P. Y., MOLONY, C., YANG, X., MACNEIL, D. J. ET AL. (2010). Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. *PLoS Genet.* **6** e1000932.

ZHU, J., ZHANG, B., SMITH, E. N., DREES, B., BREM, R. B., KRUGLYAK, L., BUMGARNER, R. E. and SCHADT, E. E. (2008). Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nat. Genet.* **40** 854–861.

The Annals of Applied Statistics
2014, Vol. 8, No. 1, 377–405
DOI: 10.1214/13-AOAS691

THE RANDOM SUBGRAPH MODEL FOR THE ANALYSIS OF AN ECCLESIASTICAL NETWORK IN MEROVINGIAN GAUL

BY YACINE JERNITE[*,§], PIERRE LATOUCHE[*], CHARLES BOUVEYRON[†], PATRICK RIVERA[‡], LAURENT JEGOU[‡] AND STÉPHANE LAMASSÉ[‡]

Laboratoire SAMM, EA 4543, Université Paris 1 Panthéon–Sorbonne[*], Laboratoire MAP5, UMR CNRS 8145, Université Paris Descartes and Sorbonne Paris Cité[†] Laboratoire LAMOP, UMR 8589, Université Paris 1 Panthéon–Sorbonne[‡] and Ecole Polytechnique[§]

In the last two decades many random graph models have been proposed to extract knowledge from networks. Most of them look for communities or, more generally, clusters of vertices with homogeneous connection profiles. While the first models focused on networks with binary edges only, extensions now allow to deal with valued networks. Recently, new models were also introduced in order to characterize connection patterns in networks through mixed memberships. This work was motivated by the need of analyzing a historical network where a partition of the vertices is given and where edges are typed. A known partition is seen as a decomposition of a network into subgraphs that we propose to model using a stochastic model with unknown latent clusters. Each subgraph has its own mixing vector and sees its vertices associated to the clusters. The vertices then connect with a probability depending on the subgraphs only, while the types of edges are assumed to be sampled from the latent clusters. A variational Bayes expectation-maximization algorithm is proposed for inference as well as a model selection criterion for the estimation of the cluster number. Experiments are carried out on simulated data to assess the approach. The proposed methodology is then applied to an ecclesiastical network in Merovingian Gaul. An R code, called *Rambo*, implementing the inference algorithm is available from the authors upon request.

REFERENCES

AIROLDI, E. M., BLEI, D. M., FIENBERG, S. E. and XING, E. P. (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9** 1981–2014.

ALBERT, R. and BARABÁSI, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Modern Phys.* **74** 47–97. MR1895096

AMBROISE, C., GRASSEAU, G., HOEBEKE, M., LATOUCHE, P., MIELE, V. and PICARD, F. (2010). The mixer R package. Available at http://cran.r-project.org/web/packages/mixer/.

BICKEL, P. J. and CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106** 21068–21073.

BIERNACKI, C., CELEUX, G. and GOVAERT, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Comput. Statist. Data Anal.* **41** 561–575. MR1968069

Key words and phrases. Ecclesiastical network, subgraphs, stochastic bloc models, random subgraph model.

BILMES, J. A. (1998). A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute* **4** 126.

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

CELISSE, A., DAUDIN, J.-J. and PIERRE, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electron. J. Stat.* **6** 1847–1899. MR2988467

DAUDIN, J. J., PICARD, F. and ROBIN, S. (2008). A mixture model for random graphs. *Stat. Comput.* **18** 173–183. MR2390817

FIENBERG, S. E. and WASSERMAN, S. S. (1981). Categorical data analysis of single sociometric relations. *Sociol. Method.* **12** 156–192.

FRANK, O. and HARARY, F. (1982). Cluster inference by using transitivity indices in empirical graphs. *J. Amer. Statist. Assoc.* **77** 835–840. MR0686407

GIRVAN, M. and NEWMAN, M. E. J. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **99** 7821–7826 (electronic). MR1908073

GOLDENBERG, A., ZHENG, A. X. and FIENBERG, S. E. (2010). *A Survey of Statistical Network Models*. Now Publishers, Hanover, MA.

HANDCOCK, M. S., RAFTERY, A. E. and TANTRUM, J. M. (2007). Model-based clustering for social networks. *J. Roy. Statist. Soc. Ser. A* **170** 301–354. MR2364300

HOFMAN, J. M. and WIGGINS, C. H. (2008). Bayesian approach to network modularity. *Phys. Rev. Lett.* **100** 258701.

HOLLAND, P. W. and LEINHARDT, S. (1981). An exponential family of probability distributions for directed graphs. *J. Amer. Statist. Assoc.* **76** 33–65. MR0608176

JEFFREYS, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. London. Ser. A.* **186** 453–461. MR0017504

JERNITE, Y., LATOUCHE, P., BOUVEYRON, C., RIVERA, P., JEGOU, L. and LAMASSÉ, S. (2013). Supplement to "The random subgraph model for the analysis of an ecclesiastical network in Merovingian Gaul." DOI:10.1214/13-AOAS691SUPP.

KEMP, C., TENENBAUM, J. B., GRIFFITHS, T. L., YAMADA, T. and UEDA, N. (2006). Learning systems of concepts with an infinite relational model. In *Proceedings of the National Conference on Artificial Intelligence* **21** 381. AAAI Press, Boston, MA.

LATOUCHE, P., BIRMELÉ, E. and AMBROISE, C. (2009). Advances in Data Analysis Data Handling and Business Intelligence Bayesian methods for graph clustering, 229–239. Springer, Berlin.

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

LATOUCHE, P., BIRMELÉ, E. and AMBROISE, C. (2012). Variational Bayesian inference and complexity control for stochastic block models. *Stat. Model.* **12** 93–115. MR2953099

MARIADASSOU, M. and MATIAS, C. (2014). Convergence of the groups posterior distribution in latent or stochastic block models. *Bernoulli*. To appear.

MILO, R., SHEN-ORR, S., ITZKOVITZ, S., KASHTAN, D., CHKLOVSKII, D. and ALON, U. (2002). Network motifs: Simple building blocks of complex networks. *Science* **298** 824–827.

MORENO, J. L. (1934). *Who Shall Survive?*: *A New Approach to the Problem of Human Interrelations*. Nervous and Mental Disease Publishing Co, Washington, DC.

NOWICKI, K. and SNIJDERS, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *J. Amer. Statist. Assoc.* **96** 1077–1087. MR1947255

PALLA, G., DERÉNYI, I., FARKAS, I. and VICSEK, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435** 814–818.

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

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

SNIJDERS, T. A. B. and NOWICKI, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *J. Classification* **14** 75–100. MR1449742

SOUFIANI, H. A. and AIROLDI, E. M. (2012). Graphlet decomposition of a weighted network. *J. Mach. Learn. Res.* **22** 54–63.

VILLA, N., ROSSI, F. and TRUONG, Q. D. (2008). Mining a medieval social network by kernel SOM and related methods. In *Proceedings of MASHS 2008* (*Modèles et Apprentissage en Sciences Humaines et Sociales*), Créteil, France, June 2008.

KEMP, C., TENENBAUM, J. B., GRIFFITHS, T. L., YAMADA, T. and UEDA, N. (2006). Learning systems of concepts with an infinite relational model. In *Proceedings of the National Conference on Artificial Intelligence* **21** 381. AAAI Press, Boston, MA.

WANG, Y. J. and WONG, G. Y. (1987). Stochastic blockmodels for directed graphs. *J. Amer. Statist. Assoc.* **82** 8–19. MR0883333

WHITE, H. C., BOORMAN, S. A. and BREIGER, R. L. (1976). Social structure from multiple networks. I. Blockmodels of roles and positions. *Am. J. Sociol.* 730–780.

XING, E. P., FU, W. and SONG, L. (2010). A state-space mixed membership blockmodel for dynamic network tomography. *Ann. Appl. Stat.* **4** 535–566. MR2758639

The Annals of Applied Statistics
2014, Vol. 8, No. 1, 406–429
DOI: 10.1214/13-AOAS692

A FUNCTIONAL DATA ANALYSIS APPROACH FOR GENETIC ASSOCIATION STUDIES

BY MATTHEW REIMHERR AND DAN NICOLAE

Pennsylvania State University and University of Chicago

We present a new method based on Functional Data Analysis (FDA) for detecting associations between one or more scalar covariates and a longitudinal response, while correcting for other variables. Our methods exploit the temporal structure of longitudinal data in ways that are otherwise difficult with a multivariate approach. Our procedure, from an FDA perspective, is a departure from more established methods in two key aspects. First, the raw longitudinal phenotypes are assembled into functional trajectories prior to analysis. Second, we explore an association test that is not directly based on principal components. We instead focus on quantifying the reduction in L^2 variability as a means of detecting associations. Our procedure is motivated by longitudinal genome wide association studies and, in particular, the childhood asthma management program (CAMP) which explores the long term effects of daily asthma treatments. We conduct a simulation study to better understand the advantages (and/or disadvantages) of an FDA approach compared to a traditional multivariate one. We then apply our methodology to data coming from CAMP. We find a potentially new association with a SNP negatively affecting lung function. Furthermore, this SNP seems to have an interaction effect with one of the treatments.

REFERENCES

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

BOSQ, D. (2000). *Linear Processes in Function Spaces*. Springer, New York. MR1783138

CARDOT, H., FERRATY, F., MAS, A. and SARDA, P. (2003). Testing hypotheses in the functional linear model. *Scand. J. Stat.* **30** 241–255. MR1965105

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

DUCHESNE, P. and LAFAYE DE MICHEAUX, P. (2010). Computing the distribution of quadratic forms: Further comparisons between the Liu–Tang–Zhang approximation and exact methods. *Comput. Statist. Data Anal.* **54** 858–862. MR2580921

FAN, J. and ZHANG, J.-T. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **62** 303–322. MR1749541

GOHBERG, I., GOLDBERG, S. and KAASHOEK, M. A. (2003). *Basic Classes of Linear Operators*. Birkhäuser, Basel. MR2015498

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

Key words and phrases. Functional data analysis, longitudinal data analysis, genome wide association study, functional linear model, functional analysis of variance, hypothesis testing.

HALL, P., MÜLLER, H.-G. and WANG, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.* **34** 1493–1517. MR2278365

IMHOF, J. P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika* **48** 419–426. MR0137199

KOKOSZKA, P. and REIMHERR, M. (2013). Determining the order of the functional autoregressive model. *J. Time Series Anal.* **34** 116–129. MR3008019

KOKOSZKA, P., MASLOVA, I., SOJKA, J. and ZHU, L. (2008). Testing for lack of dependence in the functional linear model. *Canad. J. Statist.* **36** 207–222. MR2431682

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

MATLAB (2013). *Version* 8.1 (*R*2013*a*). The MathWorks Inc., Natick, MA.

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

REIMHERR, M. (2013). Functional data methods for genome-wide association studies. Ph.D. thesis, Chicago, IL.

REISS, P. T., HUANG, L. and MENNES, M. (2010). Fast function-on-scalar regression with penalized basis expansions. *Int. J. Biostat.* **6** Art. 28, 30. MR2683940

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.

TANG, R. and MÜLLER, H.-G. (2009). Time-synchronized clustering of gene expression trajectories. *Biostatistics* **10** 32–45.

TANTISIRA, K. G., LASKY-SU, J., HARADA, M., MURPHY, A., LITONJUA, A. A., HIMES, B. E., LANGE, C., LAZARUS, R., SYLVIA, J., KLANDERMAN, B., DUAN, Q. L., QIU, W., HIROTA, T., MARTINEZ, F. D., MAUGER, D., SORKNESS, C., SZEFLER, S., LAZARUS, S. C., LEMANSKE, R. F., PETERS, S. P., LIMA, J. J., NAKAMURA, Y., TAMARI, M. and WEISS, S. T. (2011). Genomewide association between GLCCI1 and response to glucocorticoid therapy in asthma. *N. Engl. J. Med.* **365** 1173–1183.

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

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

VERZELEN, N., TAO, W. and MÜLLER, H.-G. (2012). Inferring stochastic dynamics from functional data. *Biometrika* **99** 533–550. MR2966768

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.

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

ZHANG, J.-T. and CHEN, J. (2007). Statistical inferences for functional data. *Ann. Statist.* **35** 1052–1079. MR2341698

ZIPUNNIKOV, V., CAFFO, B., YOUSEM, D. M., DAVATZIKOS, C., SCHWARTZ, B. S. and CRAINICEANU, C. (2011). Functional principal component model for high-dimensional brain imaging. *NeuroImage* **58** 772–784.

The Annals of Applied Statistics
2014, Vol. 8, No. 1, 430–447
DOI: 10.1214/13-AOAS693

A TIME-VARYING SHARED FRAILTY MODEL WITH APPLICATION TO INFECTIOUS DISEASES

BY DOYO G. ENKI, ANGELA NOUFAILY AND C. PADDY FARRINGTON

Open University

We propose a new parametric time-varying shared frailty model to represent changes over time in population heterogeneity, for use with bivariate current status data. The model uses a power transformation of a time-invariant frailty U, and is particularly convenient when U is a member of the generalized gamma family. This model avoids some shortcomings of a previously suggested time-varying frailty model, notably time-dependent support. We describe some key properties of the model, including its relative frailty variance function in different settings and how the model can be fitted to data. We describe several applications to shared frailty modeling of bivariate current status data on infectious diseases, in which the frailty represents age-dependent heterogeneity in contact rates or susceptibility to infection.

REFERENCES

AALEN, O. O., BORGAN, Ø. and GJESSING, H. K. (2008). *Survival and Event History Analysis*: *A Process Point of View*. Springer, New York. MR2449233

BALAKRISHNAN, N. and PENG, Y. (2006). Generalized gamma frailty model. *Stat. Med.* **25** 2797–2816. MR2242204

COX, C., CHU, H., SCHNEIDER, M. F. and MUÑOZ, A. (2007). Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Stat. Med.* **26** 4352–4374. MR2405358

DUCHATEAU, L. and JANSSEN, P. (2008). *The Frailty Model*. Springer, New York. MR2723929

FARRINGTON, C. P., UNKEL, S. and ANAYA-IZQUIERDO, K. (2012). The relative frailty variance and shared frailty models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **74** 673–696. MR2965955

FARRINGTON, C. P., WHITAKER, H. J., UNKEL, S. and PEBODY, R. (2013). Correlated infections: Quantifying heterogeneity in the spread of infectious diseases. *American Journal of Epidemiology* **177** 474–486.

NOUFAILY, A. and JONES, C. (2013). Parametric quantile regression based on the generalised gamma distribution. *J. R. Stat. Soc. Ser. C Appl. Stat.* **62** 723–740.

OAKES, D. (1989). Bivariate survival models induced by frailties. *J. Amer. Statist. Assoc.* **84** 487–493. MR1010337

PAIK, M. C., TSAI, W. Y. and OTTMAN, R. (1994). Multivariate survival analysis using piecewise gamma frailty. *Biometrics* **50** 975–988.

R DEVELOPMENT CORE TEAM (2012). R: A language an environment for statistical compuring. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Available at http://www.R-project.org.

UNKEL, S. and FARRINGTON, C. P. (2012). A new measure of time-varying association for shared frailty models with bivariate current status data. *Biostatistics* **13** 665–679.

Key words and phrases. Current status data, frailty, gamma, generalized gamma, heterogeneity, infectious disease, shared frailty model, time-varying frailty.

UNKEL, S., FARRINGTON, C. P., WHITAKER, H. J. and PEBODY, R. (2014). Time varying frailty models and the estimation of heterogeneities in transmission of infectious diseases. *J. R. Stat. Soc. Ser. C Appl. Stat.* **63** 141–158.

WIENKE, A. (2011). *Frailty Models in Survival Analysis*. *Chapman & Hall/CRC Biostatistics Series* **37**. CRC Press, Boca Raton, FL. MR2682965

The Annals of Applied Statistics
2014, Vol. 8, No. 1, 448–480
DOI: 10.1214/13-AOAS696

RECONSTRUCTING EVOLVING SIGNALLING NETWORKS BY HIDDEN MARKOV NESTED EFFECTS MODELS

BY XIN WANG, KE YUAN, CHRISTOPH HELLMAYR, WEI LIU AND FLORIAN MARKOWETZ

University of Cambridge

Inferring time-varying networks is important to understand the development and evolution of interactions over time. However, the vast majority of currently used models assume direct measurements of node states, which are often difficult to obtain, especially in fields like cell biology, where perturbation experiments often only provide indirect information of network structure. Here we propose hidden Markov nested effects models (HM-NEMs) to model the evolving network by a Markov chain on a state space of signalling networks, which are derived from nested effects models (NEMs) of indirect perturbation data. To infer the hidden network evolution and unknown parameter, a Gibbs sampler is developed, in which sampling network structure is facilitated by a novel structural Metropolis–Hastings algorithm. We demonstrate the potential of HM-NEMs by simulations on synthetic time-series perturbation data. We also show the applicability of HM-NEMs in two real biological case studies, in one capturing dynamic crosstalk during the progression of neutrophil polarisation, and in the other inferring an evolving network underlying early differentiation of mouse embryonic stem cells.

REFERENCES

AHMED, A. and XING, E. P. (2009). Recovering time-varying networks of dependencies in social and biological studies. *Proc. Natl. Acad. Sci. USA* **106** 11878–11883.

ANCHANG, B., SADEH, M. J., JACOB, J., TRESCH, A., VLAD, M. O., OEFNER, P. J. and SPANG, R. (2009). Modeling the temporal interplay of molecular signaling and gene expression by using dynamic nested effects models. *Proc. Natl. Acad. Sci. USA* **106** 6447–6452.

BOUTROS, M. and AHRINGER, J. (2008). The art and design of genetic screens: RNA interference. *Nat. Rev. Genet.* **9** 554–566.

BOYER, L. A., LEE, T. I., COLE, M. F., JOHNSTONE, S. E., LEVINE, S. S., ZUCKER, J. P., GUENTHER, M. G., KUMAR, R. M., MURRAY, H. L., JENNER, R. G. et al. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122** 947–956.

CASTRO, M. A., WANG, X., FLETCHER, M. N., MEYER, K. B. and MARKOWETZ, F. (2012). RedeR: R/Bioconductor package for representing modular structures, nested networks and multiple levels of hierarchical associations. *Genome Biol.* **13** R29.

FAILMEZGER, H., PRAVEEN, P., TRESCH, A. and FRÖHLICH, H. (2013). Learning gene network structure from time laps cell imaging in RNAi Knock downs. *Bioinformatics* **29** 1534–1540.

FRIEDMAN, N. and KOLLER, D. (2003). Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning* **50** 95–125.

Key words and phrases. Dynamic, signalling networks, gene perturbation, hidden Markov, nested effects models, MCMC.

FRÖHLICH, H., PRAVEEN, P. and TRESCH, A. (2011). Fast and efficient dynamic nested effects models. *Bioinformatics* **27** 238–244.

FRÖHLICH, H., FELLMANN, M., SUELTMANN, H., POUSTKA, A. and BEISSBARTH, T. (2007). Large scale statistical inference of signaling pathways from RNAi and microarray data. *BMC Bioinformatics* **8** 386.

FRÖHLICH, H., FELLMANN, M., SÜLTMANN, H., POUSTKA, A. and BEISSBARTH, T. (2008). Estimating large-scale signaling networks through nested effect models with intervention effects from microarray data. *Bioinformatics* **24** 2650–2656.

GELMAN, A., ROBERTS, G. O. and GILKS, W. R. (1996). Efficient Metropolis jumping rules. *Bayesian Stat*. **5** 599–607.

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

GEYER, C. (2011). Introduction to Markov chain Monte Carlo. In *Handbook of Markov Chain Monte Carlo* (S. Brooks, A. Gelman, G. Jones and X. L. Meng, eds.). CRC Press, Boca Raton, FL. MR2742422

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

GRZEGORCZYK, M. and HUSMEIER, D. (2008). Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move. *Machine Learning* **71** 265–305.

GRZEGORCZYK, M. and HUSMEIER, D. (2009). Nonstationary continuous dynamic Bayesian networks. *Advances in Neural Information Processing Systems* (*NIPS*) **22** 682–690.

GUO, F., HANNEKE, S., FU, W. and XING, E. P. (2007). Recovering temporally rewiring networks: A model-based approach. In *Proceedings of the* 24*th International Conference on Machine Learning* 321–328. ACM, Corvallis, OR.

HANNEKE, S. and XING, E. P. (2006). Discrete temporal models of social networks. In *Proceedings of the* 2006 *Conference on Statistical Network Analysis* 115–125. Springer, Berlin.

HOUSE, C. D., VASKE, C. J., SCHWARTZ, A. M., OBIAS, V., FRANK, B., LUU, T., SARVAZYAN, N., IRBY, R., STRAUSBERG, R. L., HALES, T. G., STUART, J. M. and LEE, N. H. (2010). Voltage-gated Na+ channel SCN5A is a key regulator of a gene transcriptional network that controls colon cancer invasion. *Cancer Res*. **70** 6957–6967.

HUSMEIER, D., DONDELINGER, F. and LEBRE, S. (2010). Inter-time segment information sharing for nonhomogeneous dynamic Bayesian networks. *Adv. Neural Inf. Process. Syst*. **23** 901–909.

IVANOVA, N., DOBRIN, R., LU, R. et al. (2006). Dissecting self-renewal in stem cells with RNA interference. *Nature* **442** 533–538.

KALMAR, T., LIM, C., HAYWARD, P., MUÑOZ-DESCALZO, S., NICHOLS, J., GARCIA-OJALVO, J. and ARIAS, A. M. (2009). Regulated fluctuations in Nanog expression mediate cell fate decisions in embryonic stem cells. *PLoS Biology* **7** e1000149.

KU, C.-J., WANG, Y., WEINER, O. D., ALTSCHULER, S. J. and WU, L. F. (2012). Network crosstalk dynamically changes during neutrophil polarization. *Cell* **149** 1073–1083.

LÈBRE, S. (2007). Stochastic process analysis for Genomics and Dynamic Bayesian Networks inference. Ph.D. thesis, Univ. d'Évry Val-d'Essonne, France.

LOH, Y.-H., WU, Q., CHEW, J.-L., VEGA, V. B., ZHANG, W., CHEN, X., BOURQUE, G., GEORGE, J., LEONG, B., LIU, J. et al. (2006). The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet*. **38** 431–440.

MADIGAN, D., YORK, J. and ALLARD, D. (1995). Bayesian graphical models for discrete data. *International Statistical Review/Revue Internationale de Statistique* **63** 215–232.

MARKOWETZ, F. (2006). Probabilistic models for gene silencing data. Ph.D. thesis, Free Univ. Berlin, Germany.

MARKOWETZ, F. (2010). How to understand the cell by breaking it: Network analysis of gene perturbation screens. *PLoS Comput. Biol*. **6** e1000655.

MARKOWETZ, F., BLOCH, J. and SPANG, R. (2005). Nontranscriptional pathway features reconstructed from secondary effects of RNA interference. *Bioinformatics* **21** 4026–4032.

MARKOWETZ, F., KOSTKA, D., TROYANSKAYA, O. G. and SPANG, R. (2007). Nested effects models for high-dimensional phenotyping screens. *Bioinformatics* **23** i305–i312.

MASUI, S., NAKATAKE, Y., TOYOOKA, Y. et al. (2007). Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells. *Nat. Cell Biol.* **9** 625–635.

MATOBA, R., NIWA, H., MASUI, S., OHTSUKA, S., CARTER, M. G., SHAROV, A. A. and KO, M. S. (2006). Dissecting Oct3/4-regulated gene networks in embryonic stem cells by expression profiling. *PLoS One* **1** e26.

MURPHY, K. P. (2002). Dynamic Bayesian networks: Representation, inference and learning. Ph.D. thesis, Univ. California.

NAVARRO, P., FESTUCCIA, N., COLBY, D., GAGLIARDI, A., MULLIN, N. P., ZHANG, W., KARWACKI-NEISIUS, V., OSORNO, R., KELLY, D., ROBERTSON, M. et al. (2012). OCT4/SOX2-independent Nanog autorepression modulates heterogeneous Nanog gene expression in mouse ES cells. *The EMBO Journal* **31** 4547–4562.

NEUMANN, B., WALTER, T., JEAN-KARIM, H. et al. (2010). Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature* **464** 721–727.

NIEDERBERGER, T., ETZOLD, S., LIDSCHREIBER, M., MAIER, K. C., MARTIN, D. E., FRÖHLICH, H., CRAMER, P. and TRESCH, A. (2012). MC EMiNEM maps the interaction landscape of the Mediator. *PLoS Comput. Biol.* **8** e1002568.

NIWA, H., OGAWA, K., SHIMOSATO, D. and ADACHI, K. (2009). A parallel circuit of LIF signalling pathways maintains pluripotency of mouse ES cells. *Nature* **460** 118–122.

ROBINSON, J. W. and HARTEMINK, A. J. (2009). Nonstationary dynamic Bayesian networks. *Adv. Neural Inf. Process. Syst.* **21** 1369–1376.

ROUDER, J. N., SPECKMAN, P. L., SUN, D., MOREY, R. D. and IVERSON, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon. Bull. Rev.* **16** 225–237.

SADEH, M. J., MOFFA, G. and SPANG, R. (2013). Considering unknown unknowns-reconstruction of nonconfoundable causal relations in biological networks. In *Research in Computational Molecular Biology* 234–248. Springer, Berlin.

SMALL, J. V., GEIGER, B., KAVERINA, I. and BERSHADSKY, A. (2002). How do microtubules guide migrating cells? *Nat. Rev. Mol. Cell Biol.* **3** 957–964.

SONG, L., KOLAR, M. and XING, E. P. (2009). Time-varying dynamic Bayesian networks. *Adv. Neural Inf. Process. Syst.* **22** 1732–1740.

TRESCH, A. and MARKOWETZ, F. (2008). Structure learning in nested effects models. *Stat. Appl. Genet. Mol. Biol.* **7** Art. 9, 28. MR2386326

VAN DEN BERG, D. L., ZHANG, W., YATES, A., ENGELEN, E., TAKACS, K., BEZSTAROSTI, K., DEMMERS, J., CHAMBERS, I. and POOT, R. A. (2008). Estrogen-related receptor beta interacts with Oct4 to positively regulate Nanog gene expression. *Mol. Cell. Biol.* **28** 5986–5995.

VASKE, C. J., HOUSE, C., LUU, T., FRANK, B., YEANG, C.-H., LEE, N. H. and STUART, J. M. (2009). A factor graph nested effects model to identify networks from genetic perturbations. *PLoS Comput. Biol.* **5** e1000274, 16. MR2486699

The Annals of Applied Statistics
2014, Vol. 8, No. 1, 481–498
DOI: 10.1214/13-AOAS697

REPLICABILITY ANALYSIS FOR GENOME-WIDE ASSOCIATION STUDIES

BY RUTH HELLER AND DANIEL YEKUTIELI

Tel-Aviv University

The paramount importance of replicating associations is well recognized in the genome-wide associaton (GWA) research community, yet methods for assessing replicability of associations are scarce. Published GWA studies often combine separately the results of primary studies and of the follow-up studies. Informally, reporting the two separate meta-analyses, that of the primary studies and follow-up studies, gives a sense of the replicability of the results. We suggest a formal empirical Bayes approach for discovering whether results have been replicated across studies, in which we estimate the optimal rejection region for discovering replicated results. We demonstrate, using realistic simulations, that the average false discovery proportion of our method remains small. We apply our method to six type two diabetes (T2D) GWA studies. Out of 803 SNPs discovered to be associated with T2D using a typical meta-analysis, we discovered 219 SNPs with replicated associations with T2D. We recommend complementing a meta-analysis with a replicability analysis for GWA studies.

REFERENCES

BENJAMINI, Y. and HELLER, R. (2008). Screening for partial conjunction hypotheses. *Biometrics* **64** 1215–1222. MR2522270

BENJAMINI, Y., HELLER, R. and YEKUTIELI, D. (2009). Selective inference in complex research. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **267** 1–17.

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

BENJAMINI, Y., KRIEGER, A. M. and YEKUTIELI, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* **93** 491–507. MR2261438

BENJAMINI, Y. and YEKUTIELI, D. (2005). Quantitative trait Loci analysis using the false discovery rate. *Genetics* **171** 783–790.

BOGOMOLOV, M. and HELLER, R. (2013). Discovering findings that replicate from a primary study of high dimension to a follow-up study. *J. Amer. Statist. Assoc.* **108** 1480–1492.

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

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

EFRON, B. (2010). *Large-Scale Inference*. Cambridge Univ. Press, Cambridge. MR2724758

EFRON, B. and TIBSHIRANI, R. (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genet. Epidemiol.* **23** 70–86.

Key words and phrases. Combined analysis, empirical Bayes, false discovery rate, meta-analysis, replication, reproducibility, type 2 diabetes.

HEDGES, L. V. and OLKIN, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press, Orlando, FL. MR0798597

HELLER, R. and YEKUTIELI, D. (2014). Supplement to "Replicability analysis for genome-wide association studies." DOI:10.1214/13-AOAS697SUPP.

IOANNIDIS, J. P. A. and KHOURY, M. J. (2011). Improving validation practices in "omics" research. *Science* **334** 1230–1232.

JIN, J. and CAI, T. T. (2007). Estimating the null and the proportional of nonnull effects in large-scale multiple comparisons. *J. Amer. Statist. Assoc.* **102** 495–506. MR2325113

KRAFT, P., ZEGGINI, E. and IOANNIDIS, J. P. A. (2009). Replication in genome-wide association studies. *Statist. Sci.* **24** 561–573. MR2779344

MCCARTHY, M. I., ABECASIS, G. R., CARDON, L. R., GOLDSTEIN, D. B., LITTLE, J., IOANNIDIS, J. P. A. and HIRSCHHORN, J. N. (2008). Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9** 356–369.

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

MURALIDHARAN, O. (2010). An empirical Bayes mixture method for effect size and false discovery rate estimation. *Ann. Appl. Stat.* **4** 422–438. MR2758178

NATARAJAN, L., PU, M. and MESSER, K. (2012). Statistical tests for the intersection of independent lists of genes: Sensitivity, FDR, and type I error control. *Ann. Appl. Stat.* **6** 521–541. MR2976481

NCI-NHGRI (2007). Replicating genotype-phenotype associations. *Nature* **447** 655–660.

OWEN, A. B. (2009). Karl Pearson's meta-analysis revisited. *Ann. Statist.* **37** 3867–3892. MR2572446

SKOL, A. D., SCOTT, L. J., ABECASIS, G. R. and BOEHNKE, M. (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet.* **38** 209–213.

STOREY, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann. Statist.* **31** 2013–2035. MR2036398

STOREY, J. D. (2007). The optimal discovery procedure: A new approach to simultaneous significance testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 347–368. MR2323757

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

STRIMMER, K. (2008). A unified approach to false discovery rate estimation. *BMC Bioinformatics* **9** 303.

SU, Z., MARCHINI, J. and DONNELLY, P. (2011). HAPGEN2: Simulation of multiple disease SNPs. *Bioinformatics* **27** 2304–2305.

SUN, W. and CAI, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.* **102** 901–912. MR2411657

SUN, W. and WEI, Z. (2011). Multiple testing for pattern identification, with applications to microarray time-course experiments. *J. Amer. Statist. Assoc.* **106** 73–88. MR2816703

THE INTERNATIONAL HAPMAP CONSORTIUM (2003). The International Hapmap project. *Nature* **426** 789–796.

VOIGHT, B. F., SCOTT, L. J., STEINTHORSDOTTIR, V., MORRIS, A. P., DINA, C., WELCH, R. P., ZEGGINI, E., HUTH, C., AULCHENKO, Y. S., THORLEIFSSON, G., MCCULLOCH, L. J., FERREIRA, T., GRALLERT, H., AMIN, N., WU, G., WILLER, C. J., RAYCHAUDHURI, S., MCCARROLL, S. A., LANGENBERG, C., HOFMANN, O. M., DUPUIS, J., QI, L., SEGRÈ, A. V., VAN HOEK, M., NAVARRO, P., ARDLIE, K., BALKAU, B., BENEDIKTSSON, R., BENNETT, A. J., BLAGIEVA, R., BOERWINKLE, E., BONNYCASTLE, L. L., BENGTSSON BOSTRÖM, K., BRAVENBOER, B., BUMPSTEAD, S., BURTT, N. P., CHARPENTIER, G., CHINES, P. S., CORNELIS, M., COUPER, D. J., CRAWFORD, G., DONEY, A. S., ELLIOTT, K. S., ELLIOTT, A. L., ERDOS, M. R., FOX, C. S., FRANKLIN, C. S., GANSER, M., GIEGER, C., GRARUP, N., GREEN, T., GRIFFIN, S., GROVES, C. J., GUIDUCCI, C., HADJADJ, S., HASSANALI, N.,

HERDER, C., ISOMAA, B., JACKSON, A. U., JOHNSON, P. R., JORGENSEN, T., KAO, W. H., KLOPP, N., KONG, A., KRAFT, P., KUUSISTO, J., LAURITZEN, T., LI, M., LIEVERSE, A., LINDGREN, C. M., LYSSENKO, V., MARRE, M., MEITINGER, T., MIDTHJELL, K., MORKEN, M. A., NARISU, N., NILSSON, P., OWEN, K. R., PAYNE, F., PERRY, J. R., PETERSEN, A. K., PLATOU, C., PROENÇA, C., PROKOPENKO, I., RATHMANN, W., RAYNER, N. W., ROBERTSON, N. R., ROCHELEAU, G., RODEN, M., SAMPSON, M. J., SAXENA, R., SHIELDS, B. M., SHRADER, P., SIGURDSSON, G., SPARSØ, T., STRASSBURGER, K., STRINGHAM, H. M., SUN, Q., SWIFT, A. J., THORAND, B., TICHET, J., TUOMI, T., VAN DAM, R. M., VAN HAEFTEN, T. W., VAN HERPT, T., VAN VLIET-OSTAPTCHOUK, J. V., WALTERS, G. B., WEEDON, M. N., WIJMENGA, C., WITTEMAN, J., BERGMAN, R. N., CAUCHI, S., COLLINS, F. S., GLOYN, A. L., GYLLENSTEN, U., HANSEN, T., HIDE, W. A., HITMAN, G. A., HOFMAN A., HUNTER, D. J., HVEEM, K., LAAKSO, M., MOHLKE, K. L., MORRIS, A. D., PALMER, C. N., PRAMSTALLER, P. P., RUDAN, I., SIJBRANDS, E., STEIN, L. D., TUOMILEHTO, J., UITTERLINDEN, A., WALKER, M., WAREHAM, N. J., WATANABE, R. M., ABECASIS, G. R., BOEHM, B. O., CAMPBELL, H., DALY, M. J., HATTERSLEY, A. T., HU, F. B., MEIGS, J. B., PANKOW, J. S., PEDERSEN, O., WICHMANN, H. E., BARROSO, I., FLOREZ, J. C., FRAYLING, T. M., GROOP, L., SLADEK, R., THORSTEINSDOTTIR, U., WILSON, J. F., ILLIG, T., FROGUEL, P., VAN DUIJN, C. M., STEFANSSON, K., ALTSHULER, D., BOEHNKE, M., MCCARTHY, M. I., MAGIC INVESTIGATORS AND GIANT CONSORTIUM (2010). Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature Genetics* **42** 579–589.

WAKEFIELD, J. (2007). A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am. J. Hum. Genet.* **81** 208–227.

ZEGGINI, E., WEEDON, M. N., LINDGREN, C. M., FRAYLING, T. M., ELLIOTT, K. S., LANGO, H., TIMPSON, N. J., PERRY, J. R. B., RAYNER, N. W. et al.(2007). Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316** 1336–1341.

The Annals of Applied Statistics
2014, Vol. 8, No. 1, 499–529
DOI: 10.1214/13-AOAS698

CONCISE COMPARATIVE SUMMARIES (CCS) OF LARGE TEXT CORPORA WITH A HUMAN EXPERIMENT

BY JINZHU JIA*, LUKE MIRATRIX†, BIN YU‡, BRIAN GAWALT‡, LAURENT EL GHAOUI‡, LUKE BARNESMOORE§ AND SOPHIE CLAVIER§

Peking University, Harvard University†, University of California, Berkeley‡ and San Francisco State University§*

In this paper we propose a general framework for topic-specific summarization of large text corpora and illustrate how it can be used for the analysis of news databases. Our framework, concise comparative summarization (CCS), is built on sparse classification methods. CCS is a lightweight and flexible tool that offers a compromise between simple word frequency based methods currently in wide use and more heavyweight, model-intensive methods such as latent Dirichlet allocation (LDA). We argue that sparse methods have much to offer for text analysis and hope CCS opens the door for a new branch of research in this important field.

For a particular topic of interest (e.g., China or energy), CSS automatically labels documents as being either on- or off-topic (usually via keyword search), and then uses sparse classification methods to predict these labels with the high-dimensional counts of all the other words and phrases in the documents. The resulting small set of phrases found as predictive are then harvested as the summary.

To validate our tool, we, using news articles from the New York Times international section, designed and conducted a human survey to compare the different summarizers with human understanding. We demonstrate our approach with two case studies, a media analysis of the framing of "Egypt" in the New York Times throughout the Arab Spring and an informal comparison of the New York Times' and Wall Street Journal's coverage of "energy." Overall, we find that the Lasso with L^2 normalization can be effectively and usefully used to summarize large corpora, regardless of document size.

REFERENCES

BISCHOF, J. M. and AIROLDI, E. M. (2012). Summarizing topical content with word frequency and exclusivity. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)* 201–208. Edinburgh, Scotland.

BLEI, D. and MCAULIFFE, J. (2008). Supervised topic models. In *Advances in Neural Information Processing Systems* 20 (J. C. Platt, D. Koller, Y. Singer and S. Roweis, eds.) 121–128. MIT Press, Cambridge, MA.

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

Key words and phrases. Text summarization, high-dimensional analysis, sparse modeling, Lasso, L1 regularized logistic regression, co-occurrence, tf-idf, L2 normalization.

CHANG, J., BOYD-GRABER, J., GERRISH, S., WANG, C. and BLEI, D. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems* 22 (Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams and A. Culotta, eds.) 288–296. Vancouver, BC, Canada.

CLAVIER, S., EL GHAOUI, L., BARNESMOORE, L. and LI, G.-C. (2010). All the news that's fit to compare: Comparing Chinese representations in the American Press and US representations in the Chinese press.

DAI, X., JIA, J., EL GHAOUI, L. and YU., B. (2011). SBA-term: Sparse bilingual association for terms. In *Fifth IEEE International Conference on Semantic Computing* (*ICSC*) 189–192. Stanford Univ., Palo Alto, CA.

EISENSTEIN, J., SMITH, N. A. and XING, E. P. (2011). Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the* 49*th Annual Meeting of the Association for Computational Linguistics*: *Human Language Technologies* **1** 1365–1374. Association for Computational Linguistics, Portland, OR.

EL GHAOUI, L., VIALLON, V. and RABBANI, T. (2010). Safe feature elimination in sparse supervised learning. Technical Report No. UC/EECS-2010-126. EECS Dept., Univ. California, Berkeley.

EL GHAOUI, L., LI, G.-C., DUONG, V.-A., PHAM, V., SRIVASTAVA, A. and BHADURI, K. (2011). Sparse machine learning methods for understanding large text corpora: Application to flight reports. In *Conference on Intelligent Data Understanding* 159–173. Mountain View, CA.

ENTMAN, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication* **43** 52–57.

ENTMAN, R. M. (2004). *Projections of power framing news*, *public opinion*, *and U.S. foreign policy*. Univ. Chicago, Chicago, IL.

FORMAN, G. (2003). An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.* **3** 1289–1305.

FRANK, E., PAYNTER, G. W., WITTEN, I. H., GUTWIN, C. and NEVILL-MANNING, C. G. (1999). Domain-specific keyphrase extraction. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence* (*IJCAI*-99) 668–673. Morgan Kaufmann, San Francisco, CA.

GAWALT, B., JIA, J., MIRATRIX, L. W., GHAOUI, L., YU, B. and CLAVIER, S. (2010). Discovering word associations in news media via feature selection and sparse classification. In *Proceedings of the International Conference on Multimedia Information Retrieval* (*MIR'*10) 211–220. Philadelphia, PA.

GENKIN, A., LEWIS, D. D. and MADIGAN, D. (2007). Large-scale Bayesian logistic regression for text categorization. *Technometrics* **49** 291–304. MR2408634

GOFFMAN, E. (1974). *Frame Analysis*: *An Essay on the Organization of Experience*. Harvard Univ. Press, Cambridge, MA.

GOLDSTEIN, J., MITTAL, V., CARBONELL, J. and KANTROWITZ, M. (2000). Multi-document summarization by sentence extraction. In *NAACL-ANLP* 2000 *Workshop on Automatic Summarization* 40–48. Seattle, WA.

GRIMMER, J., SHOREY, R., WALLACH, H. and ZLOTNICK, F. (2011). A class of Bayesian semiparametric cluster-topic models for political texts.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. H. (2011). *The Elements of Statistical Learning*, *Vol.* 1. Springer, New York.

HENNIG, L. (2009). Topic-based multi-document summarization with probabilistic latent semantic analysis. In *Recent Advances in Natural Language Processing* (*RANLP*) 144–149. Association for Computational Linguistics, Borovets, Bulgaria.

HOPKINS, D. and KING, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science* **54** 229–247.

IFRIM, G., BAKIR, G. and WEIKUM, G. (2008). Fast logistic regression for text categorization with variable-length N-grams. In 14*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 354–362. ACM, New York.

JIA, J., MIRATRIX, L. W., GAWALT, B., YU, B. and EL GHAOUI, L. (2011). What is in the news on a subject: Automatic and sparse summarization of large document corpora. Technical Report #801, Dept. Statistics, Univ. California, Berkeley.

KIOUSIS, S. and WU, X. (2008). International agenda-building and agenda-setting: Exploring the influence of public relations counsel on US news media and public perceptions of foreign nations. *The International Communications Gazette* **70** 58–75.

KUNCZIK, M. (2000). Globalisation: News media, images of nations and the flow of international capital with special reference to the role of rating agencies. *J. International Communication* **8** 39–79.

LAZER, D., PENTLAND, A., ADAMIC, L., ARAL, S., BARABASI, A.-L., BREWER, D., CHRISTAKIS, N., CONTRACTOR, N., FOWLER, J., GUTMANN, M., JEBARA, T., KING, G., MACY, M., ROY, D. and VAN ALSTYNE, M. (2009). Computational social science. *Science* **323** 721–723.

LEE, L. and CHEN, S. (2006). New methods for text categorization based on a new feature selection method and a new similarity measure between documents. *Lecture Notes in Comput. Sci.* **4031** 1280.

MCLEOD, M., KOSICKI, G. M. and PAN, Z. (1991). *On Understanding and Misunderstanding Media Effects*. Edward Arnold, London.

MONROE, B. L., COLARESI, M. P. and QUINN, K. M. (2008). Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis* **16** 372–403.

MOSTELLER, F. and WALLACE, D. L. (1984). *Applied Bayesian and Classical Inference*: *The Case of The Federalist Papers*, 2nd ed. Springer, New York. MR0766742

NETO, J. L., FREITAS, A. A. and KAESTNER, C. A. A. (2002). Automatic text summarization using a machine learning approach. In *Advances in Artificial Intelligence*. *Lecture Notes in Computer Science* **2507** 205–215. Springer, Berlin. MR2048852

PAUL, M. J., ZHAI, C. and GIRJU, R. (2010). Summarizing contrastive viewpoints in opinionated text. In *Proceedings of the* 2010 *Conference on Empirical Methods in Natural Language Processing* 66–76. Association for Computational Linguistics, Stroudsburg, PA.

POTTKER, H. (2003). News and its communicative quality: The inverted pyramid—when and why did it appear? *Journalism Studies* **4** 501–511.

ROSE, S., ENGEL, D., CRAMER, N. and COWLEY, W. (2010). Automatic keyword extraction from individual documents. In *Text Mining*: *Applications and Theory* (M. W. Berry and J. Kogan, eds.). Wiley, Chichester.

SALTON, G. (1991). Developments in automatic text retrieval. *Science* **253** 974–980.

SALTON, G. and BUCKLEY, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management* **24** 513–523.

SENELLART, P. and BLONDEL, V. D. (2008). Automatic discovery of similar words. In *Survey of Text Mining II*. Springer, Berlin.

SHAHAF, D., GUESTRIN, C. and HORVITZ, E. (2012). Trains of thought: Generating information maps. In *Proceedings of the* 21*st International Conference on World Wide Web* 899–908. ACM, Lyon, France.

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

WAGSTAFF, K. L. (2012). Machine learning that matters. In 29*th International Conference on Machine Learning* 1–6. Edinburgh, Scotland.

YANG, Y. and PENDERSEN, I. O. (1997). A comparative study on feature selection in text categorization. In *ICML*-97, 14*th International Conference on Machine Learning* 412–420. Nashville, TN.

ZHANG, T. and OLES, F. J. (2001). Text categorization based on regularized linear classfiication methods. *Information Retrieval* **4** 5–31.

ZHAO, P. and YU, B. (2007). Stagewise lasso. *J. Mach. Learn. Res.* **8** 2701–2726. MR2383572

ZUBIAGA, A., SPINA, D., FRESNO, V. and MARTÍNEZ, R. (2011). Classifying trending topics: A typology of conversation triggers on Twitter. In *Proceedings of the* 20*th ACM International Conference on Information and Knowledge Management* (*CIKM*'11) 2461–2464. ACM, New York.

The Annals of Applied Statistics
2014, Vol. 8, No. 1, 530–552
DOI: 10.1214/13-AOAS701

STATISTICAL ANALYSIS OF TRAJECTORIES ON RIEMANNIAN MANIFOLDS: BIRD MIGRATION, HURRICANE TRACKING AND VIDEO SURVEILLANCE

BY JINGYONG SU*, SEBASTIAN KURTEK†, ERIC KLASSEN‡ AND ANUJ SRIVASTAVA‡

Texas Tech University, Ohio State University† and Florida State University‡*

We consider the statistical analysis of trajectories on Riemannian manifolds that are observed under arbitrary temporal evolutions. Past methods rely on cross-sectional analysis, with the given temporal registration, and consequently may lose the mean structure and artificially inflate observed variances. We introduce a quantity that provides both a cost function for temporal registration and a proper distance for comparison of trajectories. This distance is used to define statistical summaries, such as sample means and covariances, of synchronized trajectories and "Gaussian-type" models to capture their variability at discrete times. It is invariant to identical time-warpings (or temporal reparameterizations) of trajectories. This is based on a novel mathematical representation of trajectories, termed transported square-root vector field (TSRVF), and the $\mathbb{L}^2$ norm on the space of TSRVFs. We illustrate this framework using three representative manifolds—$\mathbb{S}^2$, SE(2) and shape space of planar contours—involving both simulated and real data. In particular, we demonstrate: (1) improvements in mean structures and significant reductions in cross-sectional variances using real data sets, (2) statistical modeling for capturing variability in aligned trajectories, and (3) evaluating random trajectories under these models. Experimental results concern bird migration, hurricane tracking and video surveillance.

REFERENCES

AGGARWAL, J. K. and CAI, Q. (1999). Human motion analysis: A review. *Comput. Vis. Image Underst.* **73** 428–440.

BEG, M. F., MILLER, M. I., TROUVE, A. and YOUNES, L. (2005). Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *Int. J. Comput. Vis.* **61** 139–157.

BERTSEKAS, D. P. (2007). *Dynamic Programming and Optimal Control*, 3rd ed. Athena Scientific, Belmont, MA.

CHRISTENSEN, G. E. and JOHNSON, H. J. (2001). Consistent image registration. *IEEE Trans. Med. Imag.* **20** 568–582.

DRYDEN, I. L. and MARDIA, K. V. (1998). *Statistical Shape Analysis*. Wiley, Chichester. MR1646114

GAVRILA, D. M. (1999). The visual analysis of human movement: A survey. *Comput. Vis. Image Underst.* **73** 82–98.

JUPP, P. E. and KENT, J. T. (1987). Fitting smooth paths to spherical data. *J. R. Stat. Soc. Ser. C Appl. Stat.* **36** 34–46. MR0887825

Key words and phrases. Riemannian manifold, time warping, variance reduction, temporal trajectory, rate invariant, parallel transport.

KENOBI, K., DRYDEN, I. L. and LE, H. (2010). Shape curves and geodesic modelling. *Biometrika* **97** 567–584. MR2672484

KNEIP, A. and RAMSAY, J. O. (2008). Combining registration and fitting for functional models. *J. Amer. Statist. Assoc.* **103** 1155–1165. MR2528838

KUME, A., DRYDEN, I. L. and LE, H. (2007). Shape-space smoothing splines for planar landmark data. *Biometrika* **94** 513–528. MR2410005

KURTEK, S., WU, W. and SRIVASTAVA, A. (2011). Signal estimation under random time warpings and its applications in nonlinear signal alignments. *Adv. Neural Inf. Process. Syst.* **24** 676–683.

LE, H. (2003). Unrolling shape curves. *J. Lond. Math. Soc.* (2) **68** 511–526. MR1994697

LE, H. and KUME, A. (2000). The Fréchet mean shape and the shape of the means. *Adv. in Appl. Probab.* **32** 101–113. MR1765168

LIU, X. and MÜLLER, H.-G. (2004). Functional convex averaging and synchronization for time-warped random curves. *J. Amer. Statist. Assoc.* **99** 687–699. MR2090903

MARDIA, K. V. and JUPP, P. E. (2000). *Directional Statistics*, 3rd ed. Wiley, Chichester. MR1828667

MICHOR, P. W. and MUMFORD, D. (2007). An overview of the Riemannian metrics on spaces of curves using the Hamiltonian approach. *Appl. Comput. Harmon. Anal.* **23** 74–113. MR2333829

OWEN, J. C. and MOORE, F. R. (2008). Swainson's thrushes in migratory disposition exhibit reduced immune function. *Journal of Ethology* **26** 383–388.

ROBINSON, D. (2012). Functional analysis and partial matching in the square root velocity framework. Ph.D. thesis, Florida State Univ., Tallahassee, FL.

SHAH, J. (2008). H^0-type Riemannian metrics on the space of planar curves. *Quart. Appl. Math.* **66** 123–137. MR2396654

SRIVASTAVA, A., WU, W., KURTEK, S., KLASSEN, E. and MARRON, J. S. (2011a). Registration of functional data using Fisher–Rao metric. Preprint. Available at arXiv:1103.3817v2.

SRIVASTAVA, A., KLASSEN, E., JOSHI, S. H. and JERMYN, I. H. (2011b). Shape analysis of elastic curves in Euclidean spaces. *IEEE Trans. Pattern Anal. Mach. Intell.* **33** 1415–1428.

SUNDARAMOORTHI, G., MENNUCCI, A., SOATTO, S. and YEZZI, A. (2011). A new geometric metric in the space of curves, and applications to tracking deforming objects by prediction and filtering. *SIAM J. Imaging Sci.* **4** 109–145. MR2792407

TROUVÉ, A. and YOUNES, L. (2000). On a class of diffeomorphic matching problems in one dimension. *SIAM J. Control Optim.* **39** 1112–1135. MR1814269

TUCKER, J. D., WU, W. and SRIVASTAVA, A. (2013). Generative models for functional data using phase and amplitude separation. *Comput. Statist. Data Anal.* **61** 50–66. MR3063000

VEERARAGHAVAN, A., SRIVASTAVA, A., ROY-CHOWDHURY, A. K. and CHELLAPPA, R. (2009). Rate-invariant recognition of humans and their activities. *IEEE Trans. Image Process.* **18** 1326–1339. MR2742162

YOUNES, L., MICHOR, P. W., SHAH, J. and MUMFORD, D. (2008). A metric on shape space with explicit geodesics. *Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei* (9) *Mat. Appl.* **19** 25–57. MR2383560

The Annals of Applied Statistics
2014, Vol. 8, No. 1, 553–576
DOI: 10.1214/13-AOAS704

TESTING FOR SHIELDING OF SPECIAL NUCLEAR WEAPON MATERIALS

BY KUNG-SIK CHAN, JINZHENG LI, WILLIAM EICHINGER AND ERWEI BAI

University of Iowa

Nuclear-weapon-material detection via gamma-ray sensing is routinely applied, for example, in monitoring cross-border traffic. Natural or deliberate shielding both attenuates and distorts the shape of the gamma-ray spectra of specific radionuclides, thereby making such routine applications challenging. We develop a Lagrange multiplier (LM) test for shielding. A strong advantage of the LM test is that it only requires fitting a much simpler model that assumes no shielding. We show that, under the null hypothesis and some mild regularity conditions and as the detection time increases, LM test statistic for (composite) shielding is asymptotically Chi-square with the degree of freedom equal to the presumed number of shielding materials. We also derive the local power of the LM test. Extensive simulation studies suggest that the test is robust to the number and nature of the intervening materials, which owes to the fact that common intervening materials have broadly similar attenuation functions.

REFERENCES

ALLISON, J., AMAKO, K., APOSTOLAKIS, J., ARAUJO, H., DUBOIS, P. A., ASAI, M., BARRAND, G. A. B. G., CAPRA, R. A. C. R., CHAUVIE, S. A. C. S., CHYTRACEK, R. A. C. R. et al. (2006). Geant4 developments and applications. *IEEE Transactions on Nuclear Science* **53** 270–278.

AUGUST, R. and WHITLOCK, R. (2005). HELGA II: Autonomous passive detection of nuclear weapons materials. 2005 NRL Review, Naval Research Laboratory, Washington, DC.

BAI, E., CHAN, K.-S., EICHINGER, W. and KUMP, P. (2011). Detection of radionuclides from weak and poorly resolved spectra using Lasso and subsampling techniques. *Radiation Measurements* **46** 1138–1146.

BENJAMINI, Y. (2010). Simultaneous and selective inference: Current successes and future challenges. *Biom. J.* **52** 708–721. MR2758547

BRYAN, J. C. (2008). *Introduction to Nuclear Science*. CRC Press, Boca Raton, FL.

BURR, T. and HAMADA, M. S. (2009). Radio-isotope identification algorithms for NaI gamma spectra. *Algorithms* **2** 339–360.

CANTONE, M. C. and HOESCHEN, C., eds. (2011). *Radiation Physics for Nuclear Medicine*. Springer, Berlin.

CASHWELL, E. D., EVERETT, C. J. and RECHARD, O. W. (1957). A practical manual on the Monte Carlo method for random walk problems. Los Alamos Scientific Lab., Los Alamos, New Mexico. Available at http://catalog.hathitrust.org/Record/012213450.

Key words and phrases. Gamma ray detection, homeland security, Lagrange multiplier test, local power, multicollinearity, Poisson regression.

CHAN, K.-S., LI, J., EICHINGER, W. and BAI, E. (2012). A new physics-based method for detecting weak nuclear signals via spectral decomposition. *Nuclear Instruments and Methods in Physics Research Section A* **667** 16–25.

CHAN, K.-S., LI, J., EICHINGER, W. and BAI, E. (2013). Supplement to "Testing for shielding of special nuclear weapon materials." DOI:10.1214/13-AOAS704SUPP.

FETTER, S., COCHRAN, T. B., GRODZINS, L., LYNCH, H. L. and ZUCKER, M. S. (1990). Gamma-ray measurements of a Soviet cruise-missile warhead. *Science* **248** 828–834.

GARDNER, R. P. and XU, L. (2009). Status of the Monte Carlo library least-squares (MCLLS) approach for nonlinear radiation analyzer problems. *Radiation Physics and Chemistry* **78** 843–851.

JARMAN, K. H., DALY, D. S., ANDERSON, K. K. and WAHL, K. L. (2003). A new approach to automated peak detection. *Chemometrics and Intelligent Laboratory Systems* **69** 61–76.

KUMP, P., BAI, E.-W., CHAN, K.-S., EICHINGER, B. and LI, K. (2012). Variable selection via RIVAL (removing irrelevant variables amidst Lasso iterations) and its application to nuclear material detection. *Automatica J. IFAC* **48** 2107–2115. MR2956886

LO PRESTI, C. A., WEIER, D., KOUZES, R. and SCHWEPPE, J. (2006). Baseline suppression of vehicle portal monitor gamma count profiles: A characterization study. *Nuclear Instruments and Methods in Physics Research Section A* **562** 281–297.

MAHER, K. and WIKIBOOKS CONTRIBUTORS (2008). Basic physics of nuclear medicine. Libronomia company. Available at http://en.wikibooks.org/wiki/Basic_Physics_of_Nuclear_Medicine.

MARSHALL, J. H. and ZUMBERGE, J. F. (1989). On-line measurements of bulk coal using prompt gamma neutron activation analysis. *Nuclear Geophysics* **3** 445–459.

MEDALIA, J. (2009). Detection of nuclear weapons and materials: Science, technologies, observations. Technical Report No. R40154, CRS report for Congress.

MITCHELL, A. L., BORGARDT, J. D. and KOUZES, R. T. (2009). Skyshine contribution to gamma ray background between 0 and 4 MeV. Pacific Northwest National Laboratory. Available at http://www.pnl.gov/main/publications/external/technical_reports/PNNL-18666.pdf.

MITCHELL, D. J., SANGER, H. M. and MARLOW, K. W. (1989). Gamma-ray response functions for scintillation and semiconductor detectors. *Nuclear Instruments and Methods In Physics Research Section A* **276** 574–556.

MOSS, J., GEESAMAN, D., SCHROEDER, L., SIMON-GILLO, J. and KEISTER, B. (2002). Report on the workshop on the role of the nuclear physics research community in combating terrorism. Report No. DOE/SC-0062, U.S. Department of Energy, Washington, DC.

SMITH, H. A. JR. and LUCAS, M. (2002). Gamma-ray detectors. Los Alamos Technical reports, Federation of American Scientists. Available at http://www.lanl.gov/orgs/n/n1/panda/00326398.pdf.

The Annals of Applied Statistics
2014, Vol. 8, No. 1, 577–594
DOI: 10.1214/13-AOAS708

PREDICTIVE REGRESSIONS FOR MACROECONOMIC DATA

BY FUKANG ZHU*, ZONGWU CAI† AND LIANG PENG‡

Jilin University, University of Kansas and Xiamen University†, and Georgia Institute of Technology‡*

Researchers have constantly asked whether stock returns can be predicted by some macroeconomic data. However, it is known that macroeconomic data may exhibit nonstationarity and/or heavy tails, which complicates existing testing procedures for predictability. In this paper we propose novel empirical likelihood methods based on some weighted score equations to test whether the monthly CRSP value-weighted index can be predicted by the log dividend-price ratio or the log earnings-price ratio. The new methods work well both theoretically and empirically regardless of the predicting variables being stationary or nonstationary or having an infinite variance.

REFERENCES

AMIHUD, Y. and HURVICH, C. M. (2004). Predictive regressions: A reduced-bias estimation method. *J. Financ. Quant. Anal.* **39** 813–841.

AMIHUD, Y., HURVICH, C. M. and WANG, Y. (2009). Multiple-predictor regressions: Hypothesis testing. *Rev. Financ. Stud.* **22** 413–434.

CAI, Z. and WANG, Y. (2014). Testing predictive regression models with nonstationary regressors. *J. Econometrics* **178** 4–14. MR3137888

CAMPBELL, J. Y. and YOGO, M. (2005). Implementing the econometric methods. In *"Efficient Tests of Stock Return Predictability."* Univ. Pennsylvania. Unpublished manuscript.

CAMPBELL, J. Y. and YOGO, M. (2006). Efficient tests of stock return predictability. *Journal of Financial Economics* **81** 27–60.

CAVANAGH, C. L., ELLIOTT, G. and STOCK, J. H. (1995). Inference in models with nearly integrated regressors. *Econometric Theory* **11** 1131–1147. MR1458951

CHAN, N. H., LI, D. and PENG, L. (2012). Toward a unified interval estimation of autoregressions. *Econometric Theory* **28** 705–717. MR2927926

CHEN, W. W. and DEO, R. S. (2009). Bias reduction and likelihood-based almost exactly sized hypothesis testing in predictive regressions using the restricted likelihood. *Econometric Theory* **25** 1143–1179. MR2540496

CHUANG, C.-S. and CHAN, N. H. (2002). Empirical likelihood for autoregressive models, with applications to unstable time series. *Statist. Sinica* **12** 387–407. MR1902716

DATTA, S. (1996). On asymptotic properties of bootstrap for AR(1) processes. *J. Statist. Plann. Inference* **53** 361–374. MR1407648

FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications. Vol. II*. 2nd ed. Wiley, New York. MR0270403

HALL, P. and HEYDE, C. C. (1980). *Martingale Limit Theory and Its Application*. Academic Press, New York. MR0624435

Key words and phrases. Autoregressive process, empirical likelihood, long memory process, nearly integrated, predictive regressions, unit root, weighted estimation.

HALL, P. and JING, B.-Y. (1998). Comparison of bootstrap and asymptotic approximations to the distribution of a heavy-tailed mean. *Statist. Sinica* **8** 887–906. MR1651514

JANSSON, M. and MOREIRA, M. J. (2006). Optimal inference in regression models with nearly integrated regressors. *Econometrica* **74** 681–714. MR2217613

LEWELLEN, J. (2004). Predicting returns with financial ratios. *Journal of Financial Economics* **74** 209–235.

LING, S. (2005). Self-weighted least absolute deviation estimation for infinite variance autoregressive models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 381–393. MR2155344

OWEN, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75** 237–249. MR0946049

OWEN, A. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.* **18** 90–120. MR1041387

OWEN, A. B. (2001). *Empirical Likelihood*. Chapman & Hall, New York.

SAMORODNITSKY, G., RACHEV, S. T., KURZ-KIM, J.-R. and STOYANOV, S. V. (2007). Asymptotic distribution of unbiased linear estimators in the presence of heavy-tailed stochastic regressors and residuals. *Probab. Math. Statist.* **27** 275–302. MR2445998

STAMBAUGH, R. F. (1999). Predictive regressions. *Journal of Financial Economics* **54** 375–421.

ZHOU, M. (2012). Empirical likelihood ratio for censored/truncated data. *R package version 0.9-8-2.* http://CRAN.R-project.org/package=emplik.

The Annals of Applied Statistics
2014, Vol. 8, No. 1, 595–621
DOI: 10.1214/13-AOAS709

THE ROLE OF THE INFORMATION SET FOR FORECASTING—WITH APPLICATIONS TO RISK MANAGEMENT

BY HAJO HOLZMANN AND MATTHIAS EULERT

Philipps-Universität Marburg

Predictions are issued on the basis of certain information. If the forecasting mechanisms are correctly specified, a larger amount of available information should lead to better forecasts. For point forecasts, we show how the effect of increasing the information set can be quantified by using strictly consistent scoring functions, where it results in smaller average scores. Further, we show that the classical Diebold–Mariano test, based on strictly consistent scoring functions and asymptotically ideal forecasts, is a consistent test for the effect of an increase in a sequence of information sets on h-step point forecasts. For the value at risk (VaR), we show that the average score, which corresponds to the average quantile risk, directly relates to the expected shortfall. Thus, increasing the information set will result in VaR forecasts which lead on average to smaller expected shortfalls. We illustrate our results in simulations and applications to stock returns for unconditional versus conditional risk management as well as univariate modeling of portfolio returns versus multivariate modeling of individual risk factors. The role of the information set for evaluating probabilistic forecasts by using strictly proper scoring rules is also discussed.

REFERENCES

ACERBI, C. and TASCHE, D. (2002). On the coherence of expected shortfall. *J. Banking Finance* **26** 1487–1503.

BAO, Y., LEE, T.-H. and SALTOĞLU, B. (2006). Evaluating predictive performance of value-at-risk models in emerging markets: A reality check. *J. Forecast*. **25** 101–128. MR2226780

BERKOWITZ, J., CHRISTOFFERSEN, P. F. and PELLETIER, D. (2011). Evaluating value-at-risk models with desk-level data. *Management Science* **57** 2213–2227.

BRÖCKER, J. (2009). Reliability, sufficiency, and the decomposition of proper scores. *Q. J. Roy. Meteor. Soc*. **135** 1512–1519.

CHRISTOFFERSEN, P. F. (1998). Evaluating interval forecasts. *Internat. Econom. Rev.* **39** 841–862. MR1661906

CHRISTOFFERSEN, P. F. (2009). Value-at-risk models. In *Handbook of Financial Time Series* (T. Mikosch, J. P. Kreiß, R. A. Davis and T. G. Andersen, eds.) 753–766. Springer, Berlin.

DEGROOT, M. H. and FIENBERG, S. E. (1983). The comparison and evaluation of forecasters. *J. Roy. Stat. Soc. Ser. D* (*The Statistician*) **32** 12–22.

DIEBOLD, F. X. (2012). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold–Mariano tests. Working Paper No. 18391, NBER.

DIEBOLD, F. X. and MARIANO, R. S. (1995). Comparing predictive accuracy. *J. Bus. Econom. Statist*. **13** 253–263.

Key words and phrases. Forecast, information set, scoring function, scoring rule, value at risk.

DURRETT, R. (2005). *Probability*: *Theory and Examples*, 3rd ed. Thomson Brooks/Cole, Belmont, CA.

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

ESCANCIANO, J. C. and OLMO, J. (2011). Robust backtesting tests for value-at-risk models. *J. Financ. Economet.* **9** 132–161.

GIACOMINI, R. and WHITE, H. (2006). Tests of conditional predictive ability. *Econometrica* **74** 1545–1578. MR2268409

GNEITING, T. (2011). Making and evaluating point forecasts. *J. Amer. Statist. Assoc.* **106** 746–762. MR2847988

GNEITING, T., BALABDAOUI, F. and RAFTERY, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69** 243–268. MR2325275

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

HEINRICH, C. (2014). The mode functional is not elicitable. *Biometrika*. To appear.

JORION, P. (2006). *Value-at-Risk*: *The New Benchmark for Managing Financial Risk*. McGraw Hill, New York.

KLENKE, A. (2008). *Probability Theory*: *A Comprehensive Course*. Springer London, London. MR2372119

MCNEIL, A. J., FREY, R. and EMBRECHTS, P. (2005). *Quantitative Risk Management*: *Concepts, Techniques and Tools*. Princeton Univ. Press, Princeton, NJ. MR2175089

MITCHELL, J. and WALLIS, K. F. (2011). Evaluating density forecasts: Forecast combinations, model mixtures, calibration and sharpness. *J. Appl. Econometrics* **26** 1023–1040. MR2843116

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

PATTON, A. J. and TIMMERMANN, A. (2012). Forecast rationality tests based on multi-horizon bounds. *J. Bus. Econom. Statist.* **30** 1–17. MR2899176

ROCKAFELLAR, R. T. and URYASEV, S. (2000). Optimization of conditional value-at-risk. *J. Risk* **2** 21–41.

TSYPLAKOV, A. (2011). Evaluating density forecasts: A comment. Paper No. 31233, MPRA. Available at http://mpra.ub.uni-muenchen.de/31233.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*: *With Applications to Statistics*. Springer, New York. MR1385671

The Annals of Applied Statistics
2014, Vol. 8, No. 1, 622–647
DOI: 10.1214/13-AOAS711

MODELING EXTREME VALUES OF PROCESSES OBSERVED AT IRREGULAR TIME STEPS: APPLICATION TO SIGNIFICANT WAVE HEIGHT

BY NICOLAS RAILLARD[*,†,‡], PIERRE AILLIOT[*] AND JIANFENG YAO[§]

Université de Brest[*], IFREMER[†], Université de Rennes 1[‡] and University of Hong Kong[§]

This work is motivated by the analysis of the extremal behavior of buoy and satellite data describing wave conditions in the North Atlantic Ocean. The available data sets consist of time series of significant wave height (Hs) with irregular time sampling. In such a situation, the usual statistical methods for analyzing extreme values cannot be used directly. The method proposed in this paper is an extension of the peaks over threshold (POT) method, where the distribution of a process above a high threshold is approximated by a max-stable process whose parameters are estimated by maximizing a composite likelihood function. The efficiency of the proposed method is assessed on an extensive set of simulated data. It is shown, in particular, that the method is able to describe the extremal behavior of several common time series models with regular or irregular time sampling. The method is then used to analyze Hs data in the North Atlantic Ocean. The results indicate that it is possible to derive realistic estimates of the extremal properties of Hs from satellite data, despite its complex space–time sampling.

REFERENCES

AILLIOT, P., THOMPSON, C. and THOMSON, P. (2011). Mixed methods for fitting the GEV distribution. *Water Resour. Res*. **47** W05551.

AILLIOT, P., BAXEVANI, A., CUZOL, A., MONBET, V. and RAILLARD, N. (2011). Space–time models for moving fields with an application to significant wave height fields. *Environmetrics* **22** 354–369. MR2843390

BEIRLANT, J., GOEGEBEUR, Y., TEUGELS, J. and SEGERS, J. (2004). *Statistics of Extremes*: *Theory and Applications*. Wiley, Chichester. MR2108013

BENTON, D. and KRISHNAMOORTHY, K. (2002). Performance of the parametric bootstrap method in small sample interval estimates. *Adv. Appl. Stat*. **2** 269–285. MR1984757

BORTOT, P. and GAETAN, C. (2014). A latent process model for temporal extremes. *Scand. J. Stat.* To appear.

CAIRES, S. and STERL, A. (2005). 100-year return value estimates for ocean wind speed and significant wave height from the ERA-40 data. *J. Climate* **18** 1032–1048.

CHALLENOR, P. G., FOALE, S. and WEBB, D. J. (1990). Seasonal changes in the global wave climate measured by the Geosat altimeter. *Int. J. Remote Sens*. **11** 2205–2213.

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

Key words and phrases. Extreme values, time series, max-stable process, composite likelihood, irregular time sampling, significant wave height, satellite data.

COOLEY, D., NYCHKA, D. and NAVEAU, P. (2007). Bayesian spatial modeling of extreme precipitation return levels. *J. Amer. Statist. Assoc.* **102** 824–840. MR2411647

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

DAVISON, A. C. and SMITH, R. L. (1990). Models for exceedances over high thresholds. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **52** 393–442. MR1086795

DE HAAN, L. (1984). A spectral representation for max-stable processes. *Ann. Probab.* **12** 1194–1204. MR0757776

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

DREES, H., DE HAAN, L. and LI, D. (2006). Approximations to the tail empirical distribution function with application to testing extreme value conditions. *J. Statist. Plann. Inference* **136** 3498–3538. MR2284668

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

FAWCETT, L. and WALSHAW, D. (2007). Improved estimation for temporally clustered extremes. *Environmetrics* **18** 173–188. MR2345653

FAWCETT, L. and WALSHAW, D. (2012). Estimating return levels from serially dependent extremes. *Environmetrics* **23** 272–283. MR2914208

FISHER, R. A. and TIPPETT, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Math. Proc. Cambridge Philos. Soc.* **24** 180–190.

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

JEON, S. and SMITH, R. L. (2012). Dependence structure of spatial extremes using threshold approach. Preprint. Available at arXiv:1209.6344.

LEADBETTER, M. R., LINDGREN, G. and ROOTZÉN, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer, New York. MR0691492

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

MENÉNDEZ, M., MÉNDEZ, F. J., LOSADA, I. J. and GRAHAM, N. E. (2008). Variability of extreme wave heights in the northeast Pacific Ocean based on buoy measurements. *Geophys. Res. Lett.* **35** L22607.

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

QUEFFEULOU, P. (2004). Long-term validation of wave height measurements from altimeters. *Mar. Geod.* **27** 495–510.

RAILLARD, N., AILLIOT, P. and YAO, J. (2014). Supplement to "Modeling extreme values of processes observed at irregular time steps: Application to significant wave height." DOI:10.1214/13-AOAS711SUPP.

REICH, B. J. and SHABY, B. A. (2012). A hierarchical max-stable spatial model for extreme precipitation. *Ann. Appl. Stat.* **6** 1430–1451. MR3058670

REICH, B. J., SHABY, B. A. and COOLEY, D. (2013). A hierarchical model for serially-dependent extremes: A study of heat waves in the western US. *J. Agric. Biol. Environ. Stat.* 1–17.

RIBATET, M., OUARDA, T. B. M. J., SAUQUET, E. and GRESILLON, J. M. (2009). Modeling all exceedances above a threshold using an extremal dependence structure: Inferences on several flood characteristics. *Water Resour. Res.* **45** W03407.

RIBEREAU, P., NAVEAU, P. and GUILLOU, A. (2011). A note of caution when interpreting parameters of the distribution of excesses. *Adv. Water Resour.* **34** 1215–1221.

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

SILVA, R. D. S. and LOPES, H. F. (2008). Copula, marginal distributions and model selection: A Bayesian note. *Stat. Comput.* **18** 313–320. MR2413387

SMITH, R. L. (1990). Max-stable processes and spatial extremes. Unpublished manuscript.

SMITH, R. L., TAWN, J. A. and COLES, S. G. (1997). Markov chain models for threshold exceedances. *Biometrika* **84** 249–268. MR1467045

TOURNADRE, J. and EZRATY, R. (1990). Local climatology of wind and sea state by means of satellite radar altimeter measurements. *J. Geophys. Res.* **95** 18255–18268.

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

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

VINOTH, J. and YOUNG, I. R. (2011). Global estimates of extreme wind speed and wave height. *J. Climate* **24** 1647–1665.

WIMMER, W., CHALLENOR, P. and RETZLER, C. (2006). Extreme wave heights in the North Atlantic from altimeter data. *Renewable Energy* **31** 241–248.