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

THE ANNALS *of* APPLIED STATISTICS

An Official Journal of the Institute of Mathematical Statistics

Articles

continued

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

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE INSTITUTE OF MATHEMATICAL STATISTICS

Articles—*Continued from front cover*

Vol. 12, No. 3—September 2018

THE ANNALS OF APPLIED STATISTICS Vol. 12, No. 3, pp. 1361–2021 September 2018

INSTITUTE OF MATHEMATICAL STATISTICS

(Organized September 12, 1935)

The purpose of the Institute is to foster the development and dissemination of the theory and applications of statistics and probability.

IMS OFFICERS

President: Xiao-Li Meng, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138-2901, USA

President-Elect: Susan Murphy, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138-2901, USA

Past President: Alison Etheridge, Department of Statistics, University of Oxford, Oxford, OX1 3LB, United Kingdom

Executive Secretary: Edsel Peña, Department of Statistics, University of South Carolina, Columbia, South Carolina 29208-001, USA

Treasurer: Zhengjun Zhang, Department of Statistics, University of Wisconsin, Madison, Wisconsin 53706-1510, USA

Program Secretary: Ming Yuan, Department of Statistics, Columbia University, New York, NY 10027-5927, USA

IMS PUBLICATIONS

The Annals of Statistics. *Editors*: Edward I. George, Department of Statistics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA; Tailen Hsing, Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109-1107, USA

The Annals of Applied Statistics. *Editor-In-Chief*: Tilmann Gneiting, Heidelberg Institute for Theoretical Studies, HITS gGmbH, Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany, and Institute for Stochastics, Karlsruhe Institute of Technology, Englerstr. 2, 76128 Karlsruhe, Germany

The Annals of Probability. *Editor*: Amir Dembo, Department of Statistics and Department of Mathematics, Stanford University, Stanford, California 94305, USA

The Annals of Applied Probability. *Editor*: Bálint Tóth, School of Mathematics, University of Bristol, University Walk, BS8 1TW, Bristol, U.K. and Institute of Mathematics, Technical University Budapest, Egry József u. 1, H-1111 Budapest, Hungary

Statistical Science. *Editor*: Cun-Hui Zhang, Department of Statistics, Rutgers University, Piscataway, New Jersey 08854, USA

The IMS Bulletin. *Editor*: Vlada Limic, UMR 7501 de l'Université de Strasbourg et du CNRS, 7 rue René Descartes, 67084 Strasbourg Cedex, France

The Annals of Applied Statistics **[ISSN 1932-6157 (print); ISSN 1941-7330 (online)],** Volume 12, Number 3, September 2018. Published quarterly by the Institute of Mathematical Statistics, 3163 Somerset Drive, Cleveland, Ohio 44122, USA. Periodicals postage pending at Cleveland, Ohio, and at additional mailing offices.

POSTMASTER: Send address changes to *The Annals of Applied Statistics*, Institute of Mathematical Statistics, Dues and Subscriptions Office, 9650 Rockville Pike, Suite L 2310, Bethesda, Maryland 20814-3998, USA.

The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1361–1384
https://doi.org/10.1214/16-AOAS915

REFINING CELLULAR PATHWAY MODELS USING AN ENSEMBLE OF HETEROGENEOUS DATA SOURCES[1]

BY ALEXANDER M. FRANKS, FLORIAN MARKOWETZ[2,3] AND EDOARDO M. AIROLDI[3]

University of California, Santa Barbara, University of Cambridge and Temple University

Improving current models and hypotheses of cellular pathways is one of the major challenges of systems biology and functional genomics. There is a need for methods to build on established expert knowledge and reconcile it with results of new high-throughput studies. Moreover, the available sources of data are heterogeneous, and the data need to be integrated in different ways depending on which part of the pathway they are most informative for. In this paper, we introduce a compartment specific strategy to integrate edge, node and path data for refining a given network hypothesis. To carry out inference, we use a local-move Gibbs sampler for updating the pathway hypothesis from a compendium of heterogeneous data sources, and a new network regression idea for integrating protein attributes. We demonstrate the utility of this approach in a case study of the pheromone response MAPK pathway in the yeast *S. cerevisiae*.

REFERENCES

ALBERTS, B., JOHNSON, A., LEWIS, J., RAFF, M., ROBERTS, K. and WALTER, P. (2002). *Molecular Biology of the Cell*, 4th ed. Garland Science, New York.

BALBIN, O. A., PRENSNER, J. R., SAHU, A., YOCUM, A., SHANKAR, S., MALIK, R., FERMIN, D., DHANASEKARAN, S. M., CHANDLER, B., THOMAS, D., BEER, D. G., CAO, X., NESVIZHSKII, A. I. and CHINNAIYAN, A. M. (2013). Reconstructing targetable pathways in lung cancer by integrating diverse omics data. *Nat. Commun.* **4** Article ID 2617. DOI:10.1038/ncomms3617.

BERNARD, A. and HARTEMINK, A. J. (2005). Informative structure priors: Joint learning of dynamic regulatory networks from multiple types of data. In *Pacific Symposium on Biocomputing* 459–470.

BREM, R. B. and KRUGLYAK, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natl. Acad. Sci. USA* **102** 1572–1577.

FRANKS, A., MARKOWETZ, F. and AIROLDI, E. (2018). Supplement to "Refining cellular pathway models using an ensemble of heterogeneous data sources." DOI:10.1214/16-AOAS915SUPP.

FRIEDMAN, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science* **303** 799–805.

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

Key words and phrases. Multi-level modeling, statistical network analysis, Bayesian inference, regulation and signaling dynamics.

FRÖHLICH, H., BEISSBARTH, T., TRESCH, A., KOSTKA, D., JACOB, J., SPANG, R. and MARKOWETZ, F. (2008a). Analyzing gene perturbation screens with nested effects models in R and Bioconductor. *Bioinformatics* **24** 2549–2550.

FRÖHLICH, H., FELLMANN, M., SÜLTMANN, H., POUSTKA, A. and BEISSBARTH, T. (2008b). Predicting pathway membership via domain signatures. *Bioinformatics* **24** 2137–2142.

GASCH, A. P., SPELLMAN, P. T., KAO, C. M., CARMEL-HAREL, O., EISEN, M. B., STORZ, G., BOTSTEIN, D. and BROWN, P. O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11** 4241–4257.

GAT-VIKS, I. and SHAMIR, R. (2007). Refinement and expansion of signaling pathways: The osmotic response network in yeast. *Genome Res.* **17** 358–367.

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

GITTER, A., CARMI, M., BARKAI, N. and BAR-JOSEPH, Z. (2013). Linking the signaling cascades and dynamic regulatory networks controlling stress responses. *Genome Res.* **23** 365–376.

GRUHLER, A., OLSEN, J. V., MOHAMMED, S., MORTENSEN, P., FAERGEMAN, N. J., MANN, M. and JENSEN, O. N. (2005). Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. *Mol. Cell. Proteomics* **4** 310–327.

GUAN, Y., MYERS, C. L., HESS, D. C., BARUTCUOGLU, Z., CAUDY, A. A. and TROYANSKAYA, O. G. (2008). Predicting gene function in a hierarchical context with an ensemble of classifiers. *Genome Biol.* **9**(Suppl. 1) S3.

GUAN, Y., GORENSHTEYN, D., BURMEISTER, M., WONG, A. K., SCHIMENTI, J. C., HANDEL, M. A., BULT, C. J., HIBBS, M. A. and TROYANSKAYA, O. G. (2012). Tissue-specific functional networks for prioritizing phenotype and disease genes. *PLoS Comput. Biol.* **8** Article ID e1002694.

HAHNE, F., MEHRLE, A., ARLT, D., POUSTKA, A., WIEMANN, S. and BEISSBARTH, T. (2008). Extending pathways based on gene lists using InterPro domain signatures. *BMC Bioinform.* **9** Article ID 3. DOI:10.1186/1471-2105-9-3.

HARA, K., ONO, T., KURODA, K. and UEDA, M. (2012). Membrane-displayed peptide ligand activates the pheromone response pathway in *Saccharomyces cerevisiae*. *J. Biochem.* **151** 551–557.

HARBISON, C. T., GORDON, D. B., LEE, T. I., RINALDI, N. J., MACISAAC, K. D., DANFORD, T. W., HANNETT, N. M., TAGNE, J.-B., REYNOLDS, D. B., YOO, J., JENNINGS, E. G., ZEITLINGER, J., POKHOLOK, D. K., KELLIS, M., ROLFE, P. A., TAKUSAGAWA, K. T., LANDER, E. S., GIFFORD, D. K., FRAENKEL, E. and YOUNG, R. A. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature* **431** 99–104.

HIBBS, M. A., HESS, D. C., MYERS, C. L., HUTTENHOWER, C., LI, K. and TROYANSKAYA, O. G. (2007). Exploring the functional landscape of gene expression: Directed search of large microarray compendia. *Bioinformatics* **23** 2692–2699.

HIBBS, M. A., MYERS, C. L., HUTTENHOWER, C., HESS, D. C., LI, K., CAUDY, A. A. et al. (2009). Directing experimental biology: A case study in mitochondrial biogenesis. *PLoS Comput. Biol.* **5**(3) Article ID e1000322.

HUGHES, T. R., MARTON, M. J., JONES, A. R., ROBERTS, C. J., STOUGHTON, R., ARMOUR, C. D., BENNETT, H. A., COFFEY, E., DAI, H., HE, Y. D., KIDD, M. J., KING, A. M., MEYER, M. R., SLADE, D., LUM, P. Y., STEPANIANTS, S. B., SHOEMAKER, D. D., GACHOTTE, D., CHAKRABURTTY, K., SIMON, J., BARD, M. and FRIEND, S. H. (2000). Functional discovery via a compendium of expression profiles. *Cell* **102** 109–126.

HYDUKE, D. R. and PALSSON, B. Ø. (2010). Towards genome-scale signalling network reconstructions. *Nat. Rev. Genet.* **11** 297–307.

ISCI, S., DOGAN, H., OZTURK, C. and OTU, H. H. (2014). Bayesian network prior: Network analysis of biological data using external knowledge. *Bioinformatics* **30** 860–867.

KANEHISA, M. and GOTO, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. **28** 27–30.

KIROUAC, D. C., SAEZ-RODRIGUEZ, J., SWANTEK, J., BURKE, J. M., LAUFFENBURGER, D. A. and SORGER, P. K. (2012). Creating and analyzing pathway and protein interaction compendia for modelling signal transduction networks. *BMC Syst. Biol*. **6** Article ID 29.

KNAPP, B. and KADERALI, L. (2013). Reconstruction of cellular signal transduction networks using perturbation assays and linear programming. *PLoS ONE* **8** Article ID e69220.

KOFAHL, B. and KLIPP, E. (2004). Modelling the dynamics of the yeast pheromone pathway. *Yeast* **21** 831–850.

LETUNIC, I., DOERKS, T. and BORK, P. (2012). SMART 7: Recent updates to the protein domain annotation resource. *Nucleic Acids Res*. **40** D302–D305.

LI, J., WEI, H., LIU, T. and ZHAO, P. X. (2013). GPLEXUS: Enabling genome-scale gene association network reconstruction and analysis for very large-scale expression data. *Nucleic Acids Res*. **42** Article ID e32.

LLEWELLYN, R. and EISENBERG, D. S. (2008). Annotating proteins with generalized functional linkages. *Proc. Natl. Acad. Sci. USA* **105** 17700–17705.

LO, K., RAFTERY, A. E., DOMBEK, K. M., ZHU, J., SCHADT, E. E., BUMGARNER, R. E. and YEUNG, K. Y. (2012). Integrating external biological knowledge in the construction of regulatory networks from time-series expression data. *BMC Syst. Biol*. **6** Article ID 101.

MARKOWETZ, F. and SPANG, R. (2007). Inferring cellular networks—A review. *BMC Bioinform*. **8**(Suppl. 6) S5.

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

MCCLEAN, M. N., MODY, A., BROACH, J. R. and RAMANATHAN, S. (2007). Cross-talk and decision making in MAP kinase pathways. *Nat. Genet*. **39** 409–414.

MUKHERJEE, S. and SPEED, T. P. (2008). Network inference using informative priors. *Proc. Natl. Acad. Sci. USA* **105** 14313–14318.

MULDER, K. W., WANG, X., ESCRIU, C., ITO, Y., SCHWARZ, R. F., GILLIS, J., SIROKMÁNY, G., DONATI, G., URIBE-LEWIS, S., PAVLIDIS, P., MURRELL, A., MARKOWETZ, F. and WATT, F. M. (2012a). Diverse epigenetic strategies interact to control epidermal differentiation. *Nat. Cell Biol*. **14** 753–763.

MÜLLER, P., KUTTENKEULER, D., GESELLCHEN, V., ZEIDLER, M. P. and BOUTROS, M. (2005). Identification of JAK/STAT signalling components by genome-wide RNA interference. *Nature* **436** 871–875.

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

NAGIEC, M. J. and DOHLMAN, H. G. (2012). Checkpoints in a yeast differentiation pathway coordinate signaling during hyperosmotic stress. *PLoS Genet*. **8** Article ID e1002437.

NARIAI, N., KIM, S., IMOTO, S. and MIYANO, S. (2004). Using protein-protein interactions for refining gene networks estimated from microarray data by Bayesian networks. In *Pacific Symposium on Biocomputing* 336–347.

OURFALI, O., SHLOMI, T., IDEKER, T., RUPPIN, E. and SHARAN, R. (2007). SPINE: A framework for signaling-regulatory pathway inference from cause-effect experiments. *Bioinformatics* **23** i359–i366.

PHAM, L., CHRISTADORE, L., SCHAUS, S. and KOLACZYK, E. D. (2011). Network-based prediction for sources of transcriptional dysregulation using latent pathway identification analysis. *Proc. Natl. Acad. Sci. USA* **108** 13347–13352.

PHAM, L. M., CARVALHO, L., SCHAUS, S. and KOLACZYK, E. D. (2016). Perturbation detection through modeling of gene expression on a latent biological pathway network: A Bayesian hierarchical approach. *J. Amer. Statist. Assoc*. **111** 73–92. MR3494639

POUNDS, S. and MORRIS, S. W. (2003). Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics* **19** 1236–1242.

PUNTA, M., COGGILL, P. C., EBERHARDT, R. Y., MISTRY, J., TATE, J., BOURSNELL, C., PANG, N., FORSLUND, K., CERIC, G., CLEMENTS, J., HEGER, A., HOLM, L., SONNHAMMER, E. L. L., EDDY, S. R., BATEMAN, A. and FINN, R. D. (2012). The Pfam protein families database. *Nucleic Acids Res*. **40** D290–D301.

REGULY, T., BREITKREUTZ, A., BOUCHER, L., BREITKREUTZ, B.-J., HON, G. C., MYERS, C. L., PARSONS, A., FRIESEN, H., OUGHTRED, R., TONG, A., STARK, C., HO, Y., BOTSTEIN, D., ANDREWS, B., BOONE, C., TROYANSKYA, O. G., IDEKER, T., DOLINSKI, K., BATADA, N. N. and TYERS, M. (2006). Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. *J. Biol*. **5** Article ID 11.

REN, B., ROBERT, F., WYRICK, J. J., APARICIO, O., JENNINGS, E. G., SIMON, I., ZEITLINGER, J., SCHREIBER, J., HANNETT, N., KANIN, E., VOLKERT, T. L., WILSON, C. J., BELL, S. P. and YOUNG, R. A. (2000). Genome-wide location and function of DNA binding proteins. *Science* **290** 2306–2309.

ROBERTS, C. J., NELSON, B., MARTON, M. J., STOUGHTON, R., MEYER, M. R., BENNETT, H. A., HE, Y. D., DAI, H., WALKER, W. L., HUGHES, T. R., TYERS, M., BOONE, C. and FRIEND, S. H. (2000). Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* **287** 873–880.

RYAN, C. J., CIMERMANI, P., SZPIECH, Z. A., SALI, A., HERNANDEZ, R. D. and KROGAN, N. J. (2013). High-resolution network biology: Connecting sequence with function. *Nat. Rev. Genet*. **14** 865–879.

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

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

SCHULTZ, J., MILPETZ, F., BORK, P. and PONTING, C. P. (1998). SMART, a simple modular architecture research tool: Identification of signaling domains. *Proc. Natl. Acad. Sci. USA* **95** 5857–5864.

SCOTT, J., IDEKER, T., KARP, R. M. and SHARAN, R. (2006). Efficient algorithms for detecting signaling pathways in protein interaction networks. *J. Comput. Biol*. **13** 133–144. MR2255250

SEGAL, E., WANG, H. and KOLLER, D. (2003). Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics* **19**(Suppl. 1) i264–i271.

SEGAL, E., SHAPIRA, M., REGEV, A., PE'ER, D., BOTSTEIN, D., KOLLER, D. and FRIEDMAN, N. (2003). Module networks: Identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet*. **34** 166–176.

SIMON, I., BARNETT, J., HANNETT, N., HARBISON, C. T., RINALDI, N. J., VOLKERT, T. L., WYRICK, J. J., ZEITLINGER, J., GIFFORD, D. K., JAAKKOLA, T. S. and YOUNG, R. A. (2001). Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell* **106** 697–708.

STARK, C., BREITKREUTZ, B.-J., REGULY, T., BOUCHER, L., BREITKREUTZ, A. and TYERS, M. (2006). BioGRID: A general repository for interaction datasets. *Nucleic Acids Res*. **34** D535–D539.

STELNIEC-KLOTZ, I., LEGEWIE, S., TCHERNITSA, O., WITZEL, F., KLINGER, B., SERS, C., HERZEL, H., BLÜTHGEN, N. and SCHÄFER, R. (2012b). Reverse engineering a hierarchical regulatory network downstream of oncogenic KRAS. *Mol. Syst. Biol*. **8** Article ID 601.

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

VAGA, S., BERNARDO-FAURA, M., COKELAER, T., MAIOLICA, A., BARNES, C. A., GILLET, L. C., HEGEMANN, B., VAN DROGEN, F., SHARIFIAN, H., KLIPP, E., PETER, M., SAEZ-RODRIGUEZ, J. and AEBERSOLD, R. (2014). Phosphoproteomic analyses reveal novel cross-modulation mechanisms between two signaling pathways in yeast. *Mol. Syst. Biol.* **10** Article ID 767.

WANG, X., CASTRO, M. A., MULDER, K. W. and MARKOWETZ, F. (2012). Posterior association networks and functional modules inferred from rich phenotypes of gene perturbations. *PLoS Comput. Biol.* **8** Article ID e1002566. MR2958374

WANG, X., YUAN, K., HELLMAYR, C., LIU, W. and MARKOWETZ, F. (2014). Reconstructing evolving signalling networks by hidden Markov nested effects models. *Ann. Appl. Stat.* **8** 448–480. MR3191998

WERHLI, A. V. and HUSMEIER, D. (2007). Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge. *Stat. Appl. Genet. Mol. Biol.* **6** Article ID 15. MR2349908

WORKMAN, C. T., MAK, H. C., MCCUINE, S., TAGNE, J.-B., AGARWAL, M., OZIER, O., BEGLEY, T. J., SAMSON, L. D. and IDEKER, T. (2006). A systems approach to mapping DNA damage response pathways. *Science* **312** 1054–1059.

YATES, P. D. and MUKHOPADHYAY, N. D. (2013). An inferential framework for biological network hypothesis tests. *BMC Bioinform.* **14** Article ID 94.

YEANG, C.-H., MAK, H. C., MCCUINE, S., WORKMAN, C., JAAKKOLA, T. and IDEKER, T. (2005). Validation and refinement of gene-regulatory pathways on a network of physical interactions. *Genome Biol.* **6** Article ID R62.

YIP, K. Y., ALEXANDER, R. P., YAN, K.-K. and GERSTEIN, M. (2010). Improved reconstruction of in silico gene regulatory networks by integrating knockout and perturbation data. *PLoS ONE* **5** Article ID e8121.

The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1385–1421
https://doi.org/10.1214/17-AOAS1107

STATISTICAL SHAPE ANALYSIS OF SIMPLIFIED NEURONAL TREES

BY ADAM DUNCAN, ERIC KLASSEN AND ANUJ SRIVASTAVA

Florida State University

Neuron morphology plays a central role in characterizing cognitive health and functionality of brain structures. The problem of quantifying neuron shapes and capturing statistical variability of shapes is difficult because neurons differ both in geometry and in topology. This paper develops a mathematical representation of neuronal trees, restricting to the trees that consist of: (1) a main branch viewed as a parameterized curve in $\mathbb{R}^3$, and (2) some number of secondary branches—also parameterized curves in $\mathbb{R}^3$—which emanate from the main branch at arbitrary points. It imposes a metric on the representation space, in order to compare neuronal shapes, and to obtain optimal deformations (geodesics) across arbitrary trees. The key idea is to impose certain equivalence relations that allow trees with different geometries and topologies to be compared efficiently. The combinatorial problem of matching side branches across trees is reduced to a linear assignment with well-known efficient solutions. This framework is then applied to comparing, clustering, and classifying neurons using fully automated algorithms. The framework is illustrated on three datasets of neuron reconstructions, specifically showing geodesics paths and cross-validated classification between experimental groups.

REFERENCES

ANDERSSON-ENGELS, S., AF KLINTEBERG, C., SVANBERG, K. and SVANBERG, S. (1997). In vivo fluorescence imaging for tissue diagnostics. *Phys. Med. Biol.* **42** 815–824.

ASCOLI, G. A., DONOHUE, D. E. and HALAVI, M. (2007). NeuroMorpho.Org: A central resource for neuronal morphologies. *J. Neurosci.* **27** 9247–9251.

AYDIN, B., PATAKI, G., WANG, H., BULLITT, E. and MARRON, J. S. (2009). A principal component analysis for trees. *Ann. Appl. Stat.* **3** 1597–1615. MR2752149

AYDIN, B., PATAKI, G., WANG, H., LADHA, A. and BULLITT, E. (2011). Visualizing the structure of large trees. *Electron. J. Stat.* **5** 405–420. MR2802049

BASSELL, G. J. and WARREN, S. T. (2008). Fragile X syndrome: Loss of local mRNA regulation alters synaptic development and function. *Neuron* **60** 201–214.

BILLERA, L. J., HOLMES, S. P. and VOGTMANN, K. (2001). Geometry of the space of phylogenetic trees. *Adv. in Appl. Math.* **27** 733–767. MR1867931

BRUVERIS, M. (2015). Optimal reparametrizations in the square-root velocity framework. *SIAM Journal on Mathematical Analysis* **48** 4335–4354.

CHAN-PALAY, V. and ASAN, E. (1989). Alterations in catecholamine neurons of the locus coeruleus in senile dementia of the Alzheimer type and in Parkinson's disease with and without dementia and depression. *The Journal of Comparative Neurology* **287** 373–392.

Key words and phrases. Neuron morphology, elastic shape analysis, tree registration, neuron deformation, tree geodesics.

CHEN, J.-R., WANG, B.-N., TSENG, G.-F., WANG, Y.-J., HUANG, Y.-S. and WANG, T.-J. (2014). Morphological changes of cortical pyramidal neurons in hepatic encephalopathy. *BMC Neuroscience* **15** 15.

COLEMAN, P. D. and FLOOD, D. G. (1987). Neuron numbers and dendritic extent in normal aging and Alzheimer's disease. *Neurobiol. Aging* **8** 521–545.

CUNTZ, H., FORSTNER, F., HAAG, J. and BORST, A. (2008). The morphological identity of insect dendrites. *PLoS Comput. Biol.* **4** e1000251.

DRYDEN, I. L. and MARDIA, K. V. (1998). *Statistical Shape Analysis*. Wiley, Chichester. MR1646114

DUNCAN, A., SRIVASTAVA, A., DESCOMBES, X. and KLASSEN, E. (2015). Geometric analysis of axonal tree structures. In *DIFF-CV: Differential Geometric Techniques in Computer Vision*.

ENGLE, E. C. (2008). Human genetic disorders of axon guidance. *Neuron* **60** 201–214.

FERAGEN, A. (2012). Complexity of computing distances between geometric trees. In *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshop, SSPR&SPR 2012, Hiroshima, Japan, November 7–9, 2012. Proceedings* 89–97. Springer, Berlin.

FERAGEN, A., LAUZE, F. and HAUBERG, S. (2015). Geodesic exponential kernels: When curvature and linearity conflict. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

FERAGEN, A., HAUBERG, S., NIELSEN, M. and LAUZE, F. (2011). Means in spaces of tree-like shapes. In *Computer Vision (ICCV), 2011 IEEE International Conference on* 736–746.

FERAGEN, A., OWEN, M., PETERSEN, J., WILLE, M. M., THOMSEN, L. H., DIRKSEN, A. and DE BRUIJNE, M. (2013a). Tree-space statistics and approximations for large-scale analysis of anatomical trees. In *Information Processing in Medical Imaging: 23rd International Conference, IPMI 2013, Asilomar, CA, USA, June 28–July 3, 2013. Proceedings* 74–85. Springer, Berlin.

FERAGEN, A., LO, P., DE BRUIJNE, M., NIELSEN, M. and LAUZE, F. (2013b). Toward a theory of statistical tree-shape analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **35** 2008–2021.

FERAGEN, A., PETERSEN, J., OWEN, M., LO, P., THOMSEN, L. H., WILLE, M. M., DIRKSEN, A. and DE BRUIJNE, M. (2015). Geodesic atlas-based labeling of anatomical trees: Application and evaluation on airways extracted from CT. *IEEE Trans. Med. Imag.* **34** 1212–1226.

GIBSON, D. A. and MA, L. (2011). Developmental regulation of axon branching in the vertebrate nervous system. *Development* **138** 183–195.

HALAVI, M., HAMILTON, K. A., PAREKH, R. and ASCOLI, G. A. (2012). Digital reconstructions of neuronal morphology: Three decades of research trends. *Frontiers in Neuroscience* **6**.

HEUMAN, H. and WITTUM, G. (2009). The tree-edit-distance, a measure for quantifying neuronal morphology. *Neuroinformatics* **7** 179–190.

HIROKAWA, N., NIWA, S. and TANAKA, Y. (2010). Molecular motors in neurons: Transport mechanisms and roles in brain function, development, and disease. *Neuron* **68** 610–638.

JONKER, R. and VOLGENANT, A. (1987). A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing* **38** 325–340. MR0902027

JOSHI, S. H., KLASSEN, E., SRIVASTAVA, A. and JERMYN, I. (2007). A novel representation for Riemannian analysis of elastic curves in Rn. In *2007 IEEE Conference on Computer Vision and Pattern Recognition* 1–7.

KABASO, D., COSKREN, P., HENRY, B., HOF, P. and WEARNE, S. (2009). The electrotonic structure of pyramidal neurons contributing to prefrontal cortical circuits in macaque monkeys is significantly altered in aging. *Cerebral Cortex* **19** 2248–2268.

KENDALL, D. G., BARDEN, D., CARNE, T. K. and LE, H. (1999). *Shape and Shape Theory*. Wiley, Chichester. MR1891212

KURTEK, S., SRIVASTAVA, A., KLASSEN, E. and DING, Z. (2012). Statistical modeling of curves using shapes and related features. *J. Amer. Statist. Assoc.* **107** 1152–1165. MR3010902

LAHIRI, S., ROBINSON, D. and KLASSEN, E. (2015). Precise matching of PL curves in $\mathbb{R}^N$ in the square root velocity framework. *Geom. Imaging Comput.* **2** 133–186. MR3501512

LEDDEROSE, J., SENCION, L., SALGADO, H., ARIAS-CARRION, O. and TREVINO, M. (2014). A software tool for the analysis of neuronal morphology data. *Int. Archive Medicine* **7** 1–9.

LIU, W., SRIVASTAVA, A. and KLASSEN, E. (2008). Joint shape and texture analysis of objects boundaries in images using a Riemannian approach. In *Asilomar Conference on Signals, Systems, and Computers*.

LIU, W., SRIVASTAVA, A. and ZHANG, J. (2011). A mathematical framework for protein structure comparison. *PLoS Comput. Biol.* **7** e1001075, 10. MR2788145

MEDIONI, C., RAMIALISON, M., EPHRUSSI, A. and BESSE, F. (2014). Imp promotes axonal remodeling by regulating profilin mRNA during brain development. *Curr. Biol.* **24** 793–800.

MIO, W. and SRIVASTAVA, A. (2004). Elastic-string models for representation and analysis of planar shapes. In *Proceedings of the* 2004 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004. *CVPR* 2004. **2** II–10–II–15.

NTZIACHRISTOS, V. (2006). Fluorescence molecular imaging. *Annu Rev Biomed Eng* **8** 1–33.

SCHOENBERG, I. J. (1938). Metric spaces and positive definite functions. *Trans. Amer. Math. Soc.* **44** 522–536. MR1501980

SELKOW, S. M. (1977). The tree-to-tree editing problem. *Inform. Process. Lett.* **6** 184–186. MR0458995

SRIVASTAVA, A. and KLASSEN, E. P. (2016). *Functional and Shape Data Analysis*. Springer, New York.

SRIVASTAVA, A., KLASSEN, E., JOSHI, S. H. and JERMYN, I. H. (2011). Shape analysis of elastic curves in Euclidean spaces. *IEEE Trans. Pattern Anal. Mach. Intell.* **33** 1415–1428.

SUO, L., LU, H., YING, G., CAPECCHI, M. R. and WU, Q. (2012). Protocadherin clusters and cell adhesion kinase regulate dendrite complexity through Rho GTPase. *Journal of Molecular Cell Biology* **4** 362.

TAI, K. C. (1979). The tree-to-tree correction problem. *J. Assoc. Comput. Mach.* **26** 422–433. MR0535263

WANG, H. and MARRON, J. S. (2007). Object oriented data analysis: Sets of trees. *Ann. Statist.* **35** 1849–1873. MR2363955

WEST, M. J., COLEMAN, P. D., FLOOD, D. G. and TRONCOSO, J. C. (1994). Differences in the pattern of hippocampal neuronal loss in normal ageing and Alzheimer's disease. *Lancet* **344** 769–772.

WHITEHOUSE, P. J., PRICE, D. L., STRUBLE, R. G., CLARK, A. W., COYLE, J. T. and DELON, M. R. (1982). Alzheimer's disease and senile dementia: Loss of neurons in the basal forebrain. *Science* **215** 1237–1239.

WU, C. H. et al. (2012). Mutations in the profilin 1 gene cause familial amyotrophic lateral sclerosis. *Nature* **488** 499–503.

ZHANG, K. (1996). A constrained edit distance between unordered labeled trees. *Algorithmica* **15** 205–222. MR1368250

The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1422–1450
https://doi.org/10.1214/17-AOAS1116

TPRM: TENSOR PARTITION REGRESSION MODELS WITH APPLICATIONS IN IMAGING BIOMARKER DETECTION

BY MICHELLE F. MIRANDA[*,†,1], HONGTU ZHU[*,‡,2] AND JOSEPH G. IBRAHIM[‡,3], FOR THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE[4]

University of Texas MD Anderson Cancer Center, Universidade de São Paulo*[†] *and University of North Carolina at Chapel Hill*[‡]

Medical imaging studies have collected high-dimensional imaging data to identify imaging biomarkers for diagnosis, screening, and prognosis, among many others. These imaging data are often represented in the form of a multi-dimensional array, called a tensor. The aim of this paper is to develop a tensor partition regression modeling (TPRM) framework to establish a relationship between low-dimensional clinical outcomes (e.g., diagnosis) and high-dimensional tensor covariates. Our TPRM is a hierarchical model and efficiently integrates four components: (i) a partition model, (ii) a canonical polyadic decomposition model, (iii) a principal components model, and (iv) a generalized linear model with a sparse inducing normal mixture prior. This framework not only reduces ultra-high dimensionality to a manageable level, resulting in efficient estimation, but also optimizes prediction accuracy in the search for informative sub-tensors. Posterior computation proceeds via an efficient Markov chain Monte Carlo algorithm. Simulation shows that TPRM outperforms several other competing methods. We apply TPRM to predict disease status (Alzheimer versus control) by using structural magnetic resonance imaging data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) study.

REFERENCES

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

BADER, B. W., KOLDA, T. G. et al. (2015). MATLAB Tensor Toolbox Version 2.6. Available online.

BAIR, E., HASTIE, T., PAUL, D. and TIBSHIRANI, R. (2006). Prediction by supervised principal components. *J. Amer. Statist. Assoc.* **101** 119–137. MR2252436

BECKMANN, C. F. and SMITH, S. M. (2005). Tensorial extensions of independent component analysis for multisubject FMRI analysis. *NeuroImage* **25** 294–311.

BICKEL, P. J. and LEVINA, E. (2004). Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli* **10** 989–1010.

BRAAK, H. and BRAAK, E. (1998). Evolution of neuronal changes in the course of Alzheimer's disease. In *Ageing and Dementia* (K. Jellinger, F. Fazekas and M. Windisch, eds.). *Journal of Neural Transmission. Supplementa* **53** 127–140. Springer, Vienna.

Key words and phrases. Bayesian hierarchical model, big data, MCMC, tensor decomposition, tensor regression.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. Wadsworth Advanced Books and Software, Belmont, CA. MR0726392

CAFFO, B. S., CRAINICEANU, C. M., VERDUZCO, G., JOEL, S., MOSTOFSKY, S. H., BASSETT, S. S. and PEKAR, J. J. (2010). Two-stage decompositions for the analysis of functional connectivity for fMRI with application to Alzheimer's disease risk. *NeuroImage* **51** 1140–1149.

CAMPBELL, S. and MACQUEEN, G. (2004). The role of the hippocampus in the pathophysiology of major depression. *J. Psychiatry Neurosci.* **29** 417–426.

DAVATZIKOS, C., GENC, A., XU, D. and RESNICK, S. M. (2001). Voxel-based morphometry using the RAVENS maps: Methods and validation using simulated longitudinal atrophy. *NeuroImage* **14** 1361–1369.

DING, X., HE, L. and CARIN, L. (2011). Bayesian robust principal component analysis. *IEEE Trans. Image Process.* **20** 3419–3430.

EICKHOFF, S. B., STEPHAN, K. E., MOHLBERG, H., GREFKES, C., FINK, G. R., AMUNTS, K. and ZILLES, K. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage* **25** 1325–1335.

FAN, J. and FAN, Y. (2008). High-dimensional classification using features annealed independence rules. *Ann. Statist.* **36** 2605–2637. MR2485009

FOUNDAS, A. L., LEONARD, C. M., MAHONEY, S. M., AGEE, O. F. and HEILMAN, K. M. (1997). Atrophy of the hippocampus, parietal cortex, and insula in Alzheimer's disease: A volumetric magnetic resonance imaging study. *Neuropsychiatry Neuropsychol. Behav. Neurol.* **10** 81–89.

FRIEDMAN, J. H. (1991). Multivariate adaptive regression splines. *Ann. Statist.* **19** 1–141. MR1091842

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

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

GILLIES, R. J., KINAHAN, P. E. and HRICAK, H. (2016). Radiomics: Images are more than pictures, they are data. *Radiology* **278** 563–577.

GONÇALVES, F. B., GAMERMAN, D. and SOARES, T. M. (2013). Simultaneous multifactor DIF analysis and detection in item response theory. *Comput. Statist. Data Anal.* **59** 144–160.

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

HU, X., MEIBERTH, D., NEWPORT, B. and JESSEN, F. (2015). Anatomical correlates of the neuropsychiatric symptoms in Alzheimer's disease. *Current Alzheimer Research* **12** 266–277.

HUANG, M., NICHOLS, T., HUANG, C., YANG, Y., LU, Z., FENG, Q., KNICKMEYERE, R. C., ZHU, H. and THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2015). FVGWAS: Fast voxelwise genome wide association analysis of large-scale imaging genetic data. *NeuroImage* **118** 613–627.

JACK JR., C. R. and HOLTZMAN, D. M. (2013). Biomarker modeling of Alzheimer's disease. *Neuron* **80** 1347–1358.

JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693.

KARAS, G. B., SCHELTENS, P., ROMBOUTS, S. A. R. B., VISSER, P. J., VAN SCHIJNDEL, R. A., FOX, N. C. and BARKHOF, F. (2004). Global and local gray matter loss in mild cognitive impairment and Alzheimer's disease. *NeuroImage* **23** 708–716.

KOLDA, T. G. (2006). Multilinear operators for higher-order decompositions Technical report.

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

KRISHNAN, A., WILLIAMS, L. J., MCINTOSH, A. R. and ABDI, H. (2011). Partial least squares (PLS) methods for neuroimaging: A tutorial and review. *NeuroImage* **56** 455–475.

MARTINEZ, E., VALDES, P., MIWAKEICHI, F., GOLDMAN, R. I. and COHEN, M. S. (2004). Concurrent EEG/fMRI analysis by multiway partial least squares. *NeuroImage* **22** 1023–1034.

MAYRINK, V. D. and LUCAS, J. E. (2013). Sparse latent factor models with interactions: Analysis of gene expression data. *Ann. Appl. Stat.* **7** 799–822. MR3112918

MIRANDA, M. F., ZHU, H. and IBRAHIM, J. G. (2018). Supplement to "TPRM: Tensor partition regression models with applications in imaging biomarker detection." DOI:10.1214/17-AOAS1116SUPPA, DOI:10.1214/17-AOAS1116SUPPB.

MITCHELL, T. J. and BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83** 1023–1032.

MÜLLER, H.-G. and YAO, F. (2008). Functional additive models. *J. Amer. Statist. Assoc.* **103** 1534–1544. MR2504202

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

REISS, P. T. and OGDEN, R. T. (2010). Functional generalized linear models with images as predictors. *Biometrics* **66** 61–69. MR2756691

ROČKOVÁ, V. and GEORGE, E. I. (2014). EMVS: The EM approach to Bayesian variable selection. *J. Amer. Statist. Assoc.* **109** 828–846. MR3223753

SALMINEN, L. E., SCHOFIELD, P. R., LANE, E. M., HEAPS, J. M., PIERCE, K. D., CABEEN, R., LAIDLAW, D. H., AKBUDAK, E., CONTURO, T. E., CORREIA, S. and PAUL, R. H. (2013). Neuronal fiber bundle lengths in healthy adult carriers of the ApoE4 allele: A quantitative tractography DTI study. *Brain Imaging Behav.* **7** 274–281.

SCHUFF, N., WOERNER, N. BORETA, L., KORNFIELD, T., SHAW, L. M., TROJANOWSKI, J. Q., THOMPSON, P. M., JACK JR., C. R., WEINER, M. W. and ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2009). MRI of hippocampal volume loss in early Alzheimer's disease in relation to ApoE genotype and biomarkers. *Brain* **132** 1067–1077.

TIBSHIRANI, R., HASTIE, T., NARASIMHAN, B. and CHU, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA* **99** 6567–6572.

YASMIN, H., NAKATA, Y., AOKI, S., ABE, O., SATO, N., NEMOTO, K., ARIMA, K., FURUTA, N., UNO, M., HIRAI, S., MASUTANI, Y. and OHTOMO, K. (2008). Diffusion abnormalities of the uncinate fasciculus in Alzheimer's disease: Diffusion tensor tract-specific analysis using a new method to measure the core of the tract. *Neuroradiology* **50** 293–299.

ZHANG, H. P. and SINGER, B. H. (2010). *Recursive Partitioning and Applications*, 2nd ed. Springer, New York.

ZHOU, H., LI, L. and ZHU, H. (2013). Tensor regression with applications in neuroimaging data analysis. *J. Amer. Statist. Assoc.* **108** 540–552.

The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1451–1478
https://doi.org/10.1214/17-AOAS1117

COMPLEX-VALUED TIME SERIES MODELING FOR IMPROVED ACTIVATION DETECTION IN FMRI STUDIES

BY DANIEL W. ADRIAN*, RANJAN MAITRA†,[1] AND DANIEL B. ROWE‡,[2]

Grand Valley State University, Iowa State University† and Marquette University‡*

A complex-valued data-based model with pth order autoregressive errors and general real/imaginary error covariance structure is proposed as an alternative to the commonly used magnitude-only data-based autoregressive model for fMRI time series. Likelihood-ratio-test-based activation statistics are derived for both models and compared for experimental and simulated data. For a dataset from a right-hand finger-tapping experiment, the activation map obtained using complex-valued modeling more clearly identifies the primary activation region (left functional central sulcus) than the magnitude-only model. Such improved accuracy in mapping the left functional central sulcus has important implications in neurosurgical planning for tumor and epilepsy patients. Additionally, we develop magnitude and phase detrending procedures for complex-valued time series and examine the effect of spatial smoothing. These methods improve the power of complex-valued data-based activation statistics. Our results advocate for the use of the complex-valued data and the modeling of its dependence structures as a more efficient and reliable tool in fMRI experiments over the current practice of using only magnitude-valued datasets.

REFERENCES

ADRIAN, D. W, MAITRA, R. and ROWE, D. B. (2018). Supplement to "complex-valued time series modeling for improved activation detection in fMRI studies." DOI:10.1214/17-AOAS1117SUPP.

BANDETTINI, P. A., PETRIDOU, N. and BODURKA, J. (2005). Direct detection of neuronal activity with MRI: Fantasy, possibility, or reality? *Appl. Magn. Reson.* **29** 65–88.

BANDETTINI, P. A., JESMANOWICZ, A., WONG, E. C. and HYDE, J. S. (1993). Processing strategies for time-course data sets in functional MRI of the human brain. *Magn. Reson. Med.* **30** 161–173.

BELLIVEAU, J. W., KENNEDY, D. N., MCKINSTRY, R. C., BUCHBINDER, B. R., WEISSKOFF, R. M., COHEN, M. S., VEVEA, J. M., BRADY, T. J. and ROSEN, B. R. (1991). Functional mapping of the human visual cortex by magnetic resonance imaging. *Science* **254** 716–719.

BROWN, KINCAID B. M, T. R. and UGURBIL, K. (1982). NMR chemical shift imaging in three dimensions. *Proc. Natl. Acad. Sci. USA* **79** 3523–3526.

BRUCE, I. P., KARAMAN, M. M. and ROWE, D. B. (2011). A statistical examination of SENSE image reconstruction via an isomorphism representation. *Magn. Reson. Imag.* **29** 1267–1287.

BULLMORE, E., BRAMMER, M., WILLIAMS, S. C. R., RABE-HESKETH, S., JANOT, N., DAVID, A., MELLERS, J., HOWARD, R. and SHAM, P. (1996). Statistical methods of estimation and inference for function MR image analysis. *Magn. Reson. Med.* **35** 261–277.

Key words and phrases. Area under the ROC curve, contrast-to-noise ratio, finger-tapping motor experiment, hemodynamic response function, Kronecker product, neurosurgical planning guide, phase information, signal-to-noise ratio, structured covariance matrix.

COX, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed.l Res. Internat.l J.* **29** 162–173.

COX, R. W. (2012). AFNI: What a long strange trip it has been. *NeuroImage* **62** 743–747.

COX, R. W. and HYDE, J. S. (1997). Software tools for analysis and visualization of fMRI data. *NMR Biomed.* **10** 171–178.

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

FENG, Z., CAPRIHAN, A., BLAGOEV, K. B. and CALHOUN, V. D. (2009). Biophysical modeling of phase changes in BOLD fMRI. *NeuroImage* **47** 540–548.

FISHER, N. I. and LEE, A. J. (1992). Regression models for an angular response. *Biometrics* **48** 665–677. MR1187598

FORMAN, S. D., COHEN, J. D., FITZGERALD, M., EDDY, W. F., MINTUN, M. A. and NOLL, D. C. (1995). Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): Use of a cluster-size threshold. *Magn. Reson. Med.* **33** 636–647.

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

FRISTON, K. J., FRITH, C. D., LIDDLE, P. F., DOLAN, R. J., LAMMERTSMA, A. A. and FRACKOWIAK, R. S. J. (1990). The relationship between global and local changes in PET scans. *J. Cereb. Blood Flow Metab.* **10** 458–466.

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

FRISTON, K. J., JOSEPHS, O., ZARAHN, E., HOLMES, A. P., ROUQETTE, S. and POLINE, J.-B. (2000). To smooth or not to smooth? Bias and efficiency in fMRI time-series analysis. *NeuroImage* **12** 196–208.

GENOVESE, C. R., LAZAR, N. A. and NICHOLS, T. E. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate:. *NeuroImage* **15** 870–878.

GLISSON, T. H. (2011). *Introduction to Circuit Analysis and Design*. Springer, The Netherlands.

GLOVER, G. H. (1999). Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage* **9** 416–429.

GUDBJARTSSON, H. and PATZ, S. (1995). The Rician distribution of noisy data. *Magn. Reson. Med.* **34** 910–914.

HAHN, A. D., NENCKA, A. S. and ROWE, D. B. (2009). Improving robustness and reliability of phase-sensitive fMRI analysis using temporal off-resonance alignment of single-echo timeseries (TOAST). *NeuroImage* **44** 742–752.

HAHN, A. D., NENCKA, A. S. and ROWE, D. B. (2012). Enhancing the utility of complex-valued functional magnetic resonance imaging detection of neurobiological processes through postacquisition estimation and correction of dynamic B(0) errors and motion. *Hum. Brain Mapp.* **33** 288–306.

HAHN, A. D. and ROWE, D. B. (2012). Physiologic noise regression, motion regression, and TOAST dynamic field correction in complex-valued fMRI time series. *NeuroImage* **59** 2231–2240.

HARNSBERGER, H. R., OSBORN, A. G., ROSS, J. S., MOORE, K. R., SALZMAN, K. L., CARRASCO, C. R., HALMITON, B. E., DAVIDSON, H. C. and WIGGINS, R. H. (2007). *Diagnostic and Surgical Imaging Anatomy*: *Brain*, *Head and Neck*, *Spine*, 3rd ed. Amirsys, Salt Lake City, UT.

HOOGENRAD, F. G., REICHENBACH, J. R., HAACKE, E. M., LAI, S., KUPPUSAMY, K. and SPRENGER, M. (1998). In vivo measurement of changes in venous blood-oxygenation with high resolution functional MRI at 95 tesla by measuring changes in susceptibilty and velocity. *Magn. Res. Med.* **39** 97–107.

JACCARD, P. (1901). Ètude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull. Soc. Vaudoise Sci. Nat.* **37** 547–579.

JAIN, A. K. (1989). *Fundaments of Digital Image Processing*. Prentice Hall, New York.

JESMANOWICZ, A., NENCKA, A. and HYDE, J. S. (2014). Direct radiofrequency phase control in MRI by digital waveform playback at the larmor frequency. *Magn. Reson. Imag.* **71** 846–852.

JESMANOWICZ, A., WONG, E. C. and HYDE, J. S. (1993). Phase correction for EPI using internal reference lines. In *Proceedings from the International Society of Magnetic Resonance in Medicine* **12** 1239.

JEZZARD, P. and CLARE, S. (2001). Principles of nuclear magnetic resonance and MRI. In *Functional MRI: An Introduction to Methods* (P. Jezzard, P. M. Matthews and S. M. Smith, eds.) **3** 67–92. Oxford Univ. Press, New York.

KARAMAN, M., BRUCE, I. P. and ROWE, D. B. (2015). Incorporating relaxivities to more accurately reconstruct MR images. *Magn. Reson. Imag.* **33** 374–384.

KARAMAN, M., NENCKA, A. S., BRUCE, I. P. and ROWE, D. B. (2014). Quantification of the statistical effect of spatiatemporal processing of nontask fMRI data. *Brain Connectivity* **4** 649–661.

KRUGGEL, F. and VON CRAMON, D. Y. (1999). Modeling the hemodynamic response in single-trial functional MRI experiments. *Magn. Res. Med.* **42** 787–797.

KUMAR, A., WELTI, D. and ERNST, R. R. (1975). NMR Fourier zeugmatography. *J. Magn. Res.* **18** 69–83.

KWONG, K. K., BELLIVEAU, J. W., CHESLER, D. A., GOLDBERG, I. E., WEISSKOFF, R. M., PONCELET, B. P., KENNEDY, D. N., HOPPEL, B. E., COHEN, M. S., TURNER, R., CHENG, H.-M., BRADY, T. J. and ROSEN, B. R. (1992). Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. In *Proceedings from the National Academy of Sciences, USA* **89** 5675–5679.

LAI, S. and GLOVER, G. (1997). Detection of BOLD fMRI signals using complex data. In *Proceedings from the International Society of Magnetic Resonance in Medicine* **5** 1671.

LAZAR, N. A. (2008). *The Statistical Analysis of Functional MRI Data*. Springer, Berlin.

LEE, C. C., JACK, C. R. and RIEDERER, S. J. (1998). Mapping of the central sulcus with functional MR: Active versus passive activation tasks. *Amer. J. Neurorad.* **19** 847–852.

LEE, C. C., WARD, H. A., SHARBROUGH, F. W., MEYER, F. B., MARSH, W. R., RAFFEL, C., SO, E. L., CASCINO, G. D., SHIN, C., XU, Y., RIEDERER, S. J. and JACK, C. R. (1999). Assessment of functional MR imaging in neurosurgical planning. *Amer. J. Neurorad.* **20** 1511–1519.

LEE, J., SHAHRAM, M., SCHWARTZMAN, A. and PAULY, J. M. (2007). Complex data analysis in high-resolution SSFP fMRI. *Magn. Reson. Med.* **57** 905–917.

LJUNGGREN, S. (1983). A simple graphical representation of Fourier-based imaging methods. *J. Magn. Res.* **54** 338–343.

LOGAN, B. R., GELIAZKOVA, M. P. and ROWE, D. B. (2008). An evaluation of spatial thresholding techniques in fMRI analysis. *Hum. Brain Mapp.* **29** 1379–1389.

LOGAN, B. R. and ROWE, D. B. (2004). An evaluation of thresholding techniques in fMRI analysis. *NeuroImage* **22** 95–108.

MAITRA, R. (2010). A re-defined and generalized percent-overlap-of-activation measure for studies of fMRI reproducibility and its use in identifying outlier activation maps. *NeuroImage* **50** 124–135.

MARCHINI, J. L. and RIPLEY, B. D. (2000). A new statistical approach to detecting significant activation in functional MRI. *NeuroImage* **12** 366–380.

MENON, R. S. (2002). Postacquisition suppression of large-vessel BOLD signals in high-resolution fMRI. *Magn. Res. Med.* **47** 1–9.

MILLER, J. W. (1995). Exact maximum likelihood estimation in autoregressive processes. *J. Time Series Anal.* **16** 607–615.

NAN, F. Y. and NOWAK, R. D. (1999). Generalized likelihood ratio detection for fMRI using complex data. *IEEE Trans. Med. Imag.* **18** 320–329.

NENCKA, A. S., HAHN, A. D. and ROWE, D. B. (2008). The use of three navigator echoes in Cartesian EPI reconstruction reduces Nyquist ghosting. In *Proceedings from the International Society of Magnetic Resonance in Medicine* **16** 3032.

NENCKA, A. S., HAHN, A. D. and ROWE, D. B. (2009). A mathematical model for understanding the STatistical effects of k-space (AMMUST-k) preprocessing on observed voxel measurements in fcMRI and fMRI. *J. Neurosci. Met.* **181** 268–282.

OGAWA, S., LEE, T. M., NAYAK, A. S. and GLYNN, P. (1990). Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magn. Reson. Med.* **14** 68–78.

PETRIDOU, N., SCHAFER, A., GOWLAND, P. and BOWTELL, R. (2009). Phase vs. magnitude information in functional magnetic resonace imaging time series: Toward understanding the noise. *Magn. Reson. Imag.* **27** 1046–1057.

POURAHMADI, M. (2001). *Foundations of Time Series Analysis and Prediction Theory*. Wiley, New York. MR1849562

PURDON, P. L. and WEISSKOFF, R. M. (1998). Effect of temporal autocorrelation due to physiological noise and stimulus paradigm on voxel-level false-positive rates in fMRI. *Hum. Brain Mapp.* **6** 239–249.

ROSEN, B. R. and SAVOY, R. L. (2012). fMRI at 20: Has it changed the world? *NeuroImage* **62** 1316–1324.

ROWE, D. B. (2005). Modeling both the magnitude and phase of complex-valued fMRI data. *NeuroImage* **25** 1310–1324.

ROWE, D. B. (2016). Image Reconstruction in Functional MRI. In *Handbook of Neuroimaging Data Analysis* 205–232 Chapman & Hall/CRC, London.

ROWE, D. B. and LOGAN, B. R. (2004). A complex way to compute fMRI activation. *NeuroImage* **23** 1078–1092.

ROWE, D. B., MELLER, C. P. and HOFFMAN, R. G. (2007). Characterizing phase-only fMRI data with an angular regression model. *J. Neurosci. Met.* **161** 331–341.

RUMEAU, C., TZOURIO, N., MURAYAMA, N., PERETTI-VITON, P., LEVRIER, O., JOLIOT, M., MAZOYER, B. and SALAMON, G. (1994). Location of hand function in the sensorimotor cortex: MR and functional correlation. *Amer. J. Neurorad.* **15** 567–572.

SMITH, S. M. (2001). Preparing fMRI data for statistical analysis. In *Functional MRI*: *An Introduction to Methods* Chapter 12. Oxford Univ. Press, Oxford.

SØRENSEN, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biologiske Skrifter/Kongelige Danske Videnskabernes Selskab* **5** 1–34.

TWEIG, D. B. (1983). The k-trajectory formulation of the NMR imaging process with applications in analysis and synthesis of imaging methods. *Med. Phys.* **10** 610–21.

WANG, T. and WEI, T. (1994). Statistical analysis of MR imaging and its applications in image modeling. In *Proceedings of the IEEE International Conference on Image Proccessing and Neural Networks* **1** 866–870.

WOO, C.-W., KRISHNAN, A. and WAGER, T. D. (2014). Cluster-extent based thresholding in fMRI analyses. *NeuroImage* **91** 412–419.

WORSLEY, K. J., MARRETT, S., NEELIN, P., VANDAL, A. C., FRISTON, K. J. and EVANS, A. C. (1996). A unified statistical approach for determining significant voxels in images of cerebral activation. *Hum. Brain Mapp.* **4** 58–73.

ZARAHN, E., AGUIRRE, G. K. and D'ESPOSITO, M. (1997). Empirical analyses of BOLD fMRI statistics. I. Spatially unsmoothed data collected under null-hypothesis conditions. *NeuroImage* **5** 179–197.

ZHAO, F., JIN, T., WANG, P., HU, X. and KIM, S.-G. (2007). Sources of phase changes in BOLD and CBV-weighted fMRI. *Magn. Reson. Med.* **57** 520–527.

The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1479–1505
https://doi.org/10.1214/17-AOAS1118

OPTIMAL MULTILEVEL MATCHING USING NETWORK FLOWS: AN APPLICATION TO A SUMMER READING INTERVENTION

BY SAMUEL D. PIMENTEL*, LINDSAY C. PAGE†, MATTHEW LENARD‡ AND LUKE KEELE§

University of California, Berkeley, University of Pittsburgh†, Wake County Public Schools‡ and Georgetown University§*

Many observational studies of causal effects occur in settings with clustered treatment assignment. In studies of this type, treatment is applied to entire clusters of units. For example, an educational intervention might be administered to all the students in a school. We develop a matching algorithm for multilevel data based on a network flow algorithm. Earlier work on multilevel matching relied on integer programming, which allows for balance targeting on specific covariates but can be slow with larger data sets. Although we cannot directly specify minimal levels of balance for individual covariates, our algorithm is fast and scales easily to larger data sets. We apply this algorithm to assess a school-based intervention through which students in treated schools were exposed to a new reading program during summer school. In one variant of the algorithm, where we match both schools and students, we change the causal estimand through optimal subset matching to better maintain common support. In a second variant, we relax the common support assumption to preserve the causal estimand by only matching on schools. We find that the summer intervention does not appear to increase reading test scores. In a sensitivity analysis, however, we determine that an unobserved confounder could easily mask a larger treatment effect.

REFERENCES

ARPINO, B. and MEALLI, F. (2011). The specification of the propensity score in multilevel observational studies. *Comput. Statist. Data Anal.* **55** 1770–1780. MR2748678

BARNOW, B. S., CAIN, G. G. and GOLDBERGER, A. S. (1980). Issues in the analysis of selectivity bias. In *Evaluation Studies* (E. Stromsdorfer and G. Farkas, eds.) **5** 43–59. Sage, San Francisco, CA.

BORMAN, G. D., BENSON, J. and OVERMAN, L. T. (2005). Families, schools, and summer learning. *Elem. Sch. J.* **106** 131–150.

BORMAN, G. D. and DOWLING, N. M. (2006). Longitudinal achievement effects of multiyear summer school: Evidence from the teach Baltimore randomized field trial. *Educ. Eval. Policy Anal.* **28** 25–48.

COCHRAN, W. G. (1965). The planning of observational studies of human populations. *J. Roy. Statist. Soc. Ser. A* **128** 234–265.

COCHRAN, W. G. and RUBIN, D. B. (1973). Controlling bias in observational studies. *Sankhya Ser. A* **35** 417–446.

Key words and phrases. Causal inference, hierarchical/multilevel data, observational study, optimal matching.

COOPER, H., NYE, B., CHARLTON, K., LINDSAY, J. and GREATHOUSE, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Rev. Educ. Res.* **66** 227–268.

COOPER, H., CHARLTON, K., VALENTINE, J. C., MUHLENBRUCK, L. and BORMAN, G. D. (2000). Making the most of summer school: A meta-analytic and narrative review. *Monogr. Soc. Res. Child Dev.* **65** 1–127.

CORP, C. (2015). myON: A complete digital literacy program. Available at http://thefutureinreading.myon.com/overview/complete-literacy-program.

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

ENTWISLE, D. R. and ALEXANDER, K. L. (1992). Summer setback: Race, poverty, school composition, and mathematics achievement in the first two years of school. *Am. Sociol. Rev.* **57** 72–84.

HANSEN, B. B., ROSENBAUM, P. R. and SMALL, D. S. (2014). Clustered treatment assignments and sensitivity to unmeasured biases in observational studies. *J. Amer. Statist. Assoc.* **109** 133–144. MR3180552

HODGES, J. L. JR. and LEHMANN, E. L. (1963). Estimates of location based on rank tests. *Ann. Math. Stat.* **34** 598–611. MR0152070

HONG, G. and RAUDENBUSH, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *J. Amer. Statist. Assoc.* **101** 901–910. MR2324091

LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. Springer, New York. MR2135927

LI, F., ZASLAVSKY, A. M. and LANDRUM, M. B. (2013). Propensity score weighting with multilevel data. *Stat. Med.* **32** 3373–3387. MR3074363

PAGE, L. C. and SCOTT-CLAYTON, J. (2016). Improving college access in the United States: Barriers and policy responses. *Econ. Educ. Rev.* **51** 4–22.

PIMENTEL, S. D. and KELZ, R. (2017). Optimal tradeoffs in matching designs for observational studies. Unpublished manuscript.

PIMENTEL, S. D., KELZ, R. R., SILBER, J. H. and ROSENBAUM, P. R. (2015). Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. *J. Amer. Statist. Assoc.* **110** 515–527. MR3367244

QUINN, D. M. (2015). Black–white summer learning gaps interpreting the variability of estimates across representations. *Educ. Eval. Policy Anal.* **37** 50–69.

RAMBO-HERNANDEZ, K. E. and MCCOACH, D. B. (2015). High-achieving and average students' reading growth: Contrasting school and summer trajectories. *J. Educ. Res.* **108** 112–129.

ROSENBAUM, P. R. (1989). Optimal matching for observational studies. *J. Amer. Statist. Assoc.* **84** 1024–1032.

ROSENBAUM, P. R. (2002). *Observational Studies*, 2nd ed. Springer, New York. MR1899138

ROSENBAUM, P. R. (2003). Exact confidence intervals for nonconstant effects by inverting the signed rank test. *Amer. Statist.* **57** 132–138. MR1969770

ROSENBAUM, P. R. (2008). Testing hypotheses in order. *Biometrika* **95** 248–252. MR2409727

ROSENBAUM, P. R. (2010). *Design of Observational Studies*. Springer, New York. MR2561612

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

ROSENBAUM, P. R. (2012b). Testing one hypothesis twice in observational studies. *Biometrika* **99** 763–774. MR2999159

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

ROSENBAUM, P. R. and RUBIN, D. B. (1985). Constructing a control group using multivariate matched sampling methods. *Amer. Statist.* **39** 33–38.

ROSENBAUM, P. R. and SILBER, J. H. (2009). Sensitivity analysis for equivalence and difference in an observational study of neonatal intensive care units. *J. Amer. Statist. Assoc.* **104** 501–511. MR2751434

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **6** 688–701.

RUBIN, D. B. (2008). For objective causal inference, design trumps analysis. *Ann. Appl. Stat.* **2** 808–804. MR2516795

SILBER, J. H., ROSENBAUM, P. R., TRUDEAU, M. E., EVEN-SHOSHAN, O., CHEN, W., ZHANG, X. and MOSHER, R. E. (2001). Multivariate matching and bias reduction in the surgical outcomes study. *Med. Care* **39** 1048–1064.

SKIBBE, L. E., GRIMM, K. J., BOWLES, R. P. and MORRISON, F. J. (2012). Literacy growth in the academic year versus summer from preschool through second grade: Differential effects of schooling across four skills. *Sci. Stud. Read.* **16** 141–165.

SPLAWA-NEYMAN, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–472. Translated from the Polish and edited by D. M. Dabrowska and T. P. Speed. MR1092986

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

WIRT, J., CHOY, S., GRUNER, A., SABLE, J., TOBIN, R., BAE, Y., SEXTON, J., STENNETT, J., WATANABE, S., ZILL, N. et al. (2000). *The Condition of Education*, 2000. ERIC, Washington, DC.

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. MR2959630

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. MR3036400

ZUBIZARRETA, J. R. and KEELE, L. (2017a). Optimal multilevel matching in clustered observational studies: A case study of the effectiveness of private schools under a large-scale voucher system. *J. Amer. Statist. Assoc.* **112** 547–560. MR3671751

ZUBIZARRETA, J. R. and KEELE, L. (2017b). Optimal multilevel matching in clustered observational studies: A case study of the effectiveness of private schools under a large-scale voucher system. *J. Amer. Statist. Assoc.* **112** 547–560. MR3671751

ZUBIZARRETA, J. R., PAREDES, R. D. and ROSENBAUM, P. R. (2014). Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile. *Ann. Appl. Stat.* **8** 204–231. MR3191988

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

ZVOCH, K. and STEVENS, J. J. (2015). Identification of summer school effects by comparing the in-and out-of-school growth rates of struggling early readers. *Elem. Sch. J.* **115** 433–456.

The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1506–1534
https://doi.org/10.1214/17-AOAS1119

TOPOLOGICAL DATA ANALYSIS OF SINGLE-TRIAL ELECTROENCEPHALOGRAPHIC SIGNALS

BY YUAN WANG[*,1], HERNANDO OMBAO[†,‡,2] AND MOO K. CHUNG[*,1]

University of Wisconsin–Madison[*]*, University of California, Irvine*[†] *and King Abdullah University of Science and Technology*[‡]

Epilepsy is a neurological disorder marked by sudden recurrent episodes of sensory disturbance, loss of consciousness, or convulsions, associated with abnormal electrical activity in the brain. Statistical analysis of neurophysiological recordings, such as electroencephalography (EEG), facilitates the understanding of epileptic seizures. Standard statistical methods typically analyze amplitude and frequency information in EEG signals. In the current study, we propose a topological data analysis (TDA) framework to analyze single-trial EEG signals. The framework denoises signals with a weighted Fourier series (WFS), and tests for differences between the topological features—persistence landscapes (PLs) of denoised signals through resampling in the frequency domain. Simulation studies show that the test is robust for topologically similar signals while bearing sensitivity to topological tearing in signals. In an application to single-trial epileptic EEG signals, EEG signals in the diagnosed seizure origin and its symmetric site are found to have similar PLs before and during a seizure attack, in contrast to signals at other sites showing significant statistical difference in the PLs of the two phases.

REFERENCES

ABRAMOVICH, F. and BENJAMINI, Y. (1996). Adaptive thresholding of wavelet coefficients. *Comput. Statist. Data Anal.* **22** 351–361. MR1411575

ADLER, R. J., BOBROWSKI, O., BORMAN, M. S., SUBAG, E. and WEINBERGER, S. (2010). Persistent homology for random fields and complexes. In *Borrowing Strength: Theory Powering Applications—a Festschrift for Lawrence D. Brown. Inst. Math. Stat. (IMS) Collect.* **6** 124–143. IMS, Beachwood, OH. Available at arXiv:1003.1001. MR2798515

AHMED, M., FASY, B. T. and WENK, C. (2014). Local persistent homology based distance between maps. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* 43–52.

BANCAUD, J., BRUNET-BOURGIN, F., CHAUVEL, P. and HALGREN, E. (1994). Anatomical origin of déjà vu and vivid 'memories' in human temporal lobe epilepsy. *Brain* **117** 71–90.

BARTOLOMEI, F., CHAUVEL, P. and WENDLING, F. (2008). Epileptogenicity of brain structures in human temporal lobe epilepsy: A quantified study from intracerebral EEG. *Brain* **131** 1818–1830.

BENDICH, P., MARRON, J. S., MILLER, E., PIELOCH, A. and SKWERER, S. (2016). Persistent homology analysis of brain artery trees. *Ann. Appl. Stat.* **10** 198–218. MR3480493

BOTEV, Z. I., GROTOWSKI, J. F. and KROESE, D. P. (2010). Kernel density estimation via diffusion. *Ann. Statist.* **38** 2916–2957. MR2722460

Key words and phrases. Persistence landscape, persistent homology, weighted Fourier series, electroencephalogram, epilepsy.

BRAZIER, M. A. B. (1972). Spread of seizure discharges in epilepsy: Anatomical and electrophysiological considerations. *Exp. Neurol.* **36** 263–272.

BUBENIK, P. (2015). Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.* **16** 77–102. MR3317230

BUBENIK, P., CARLSON, G., KIM, P. T. and LUO, Z.-M. (2010). Statistical topology via Morse theory persistence and nonparametric estimation. In *Algebraic Methods in Statistics and Probability II. Contemporary Mathematics*. **516** 75–92.

BURNS, S. P., SANTANIELLO, S., YAFFE, R. B., JOUNY, C. C., CRONE, N. E., BERGEY, G. K., ANDERSON, W. S. and SARMA, S. V. (2014). Network dynamics of the brain and influence of the epileptic seizure onset zone. *Proc. Natl. Acad. Sci. USA* **111** E5321–E5330.

CARLSSON, G. (2009). Topology and data. *Bull. Amer. Math. Soc.* (*N.S.*) **46** 255–308. MR2476414

CHAUDHURI, P. and MARRON, J. S. (2000). Scale space view of curve estimation. *Ann. Statist.* **28** 408–428. MR1790003

CHAZAL, F., FASY, B. T., LECCI, F., MICHEL, B., RINALDO, A. and WASSERMAN, L. (2014). Subsampling methods for persistent homology. Available at arXiv:1406.1901.

CHUNG, M. K. (2014). *Statistical and Computational Methods in Brain Image Analysis*. Chapman & Hall/CRC, London.

CHUNG, M. K., BUBENIK, P. and KIM, P. T. (2009). Persistence diagrams of cortical surface data. In *Proceedings of the* 21*st International Conference on Information Processing in Medical Imaging* (*IPMI*) 386–397.

CHUNG, M., DALTON, K., SHEN, L., EVANS, A. C. and DAVIDSON, R. J. (2007). Weighted Fourier series representation and its application to quantifying the amount of gray matter. *IEEE Trans. Med. Imag.* **26** 566–581.

CHUNG, M. K., SCHAEFER, S. M., VAN REEKUM, C. M., PESCHKE-SCHMITZ, L., SUTTERER, M. J. and DAVIDSON, R. J. (2014). A unified kernel regression for diffusion wavelets on manifolds detects aging-related changes in the amygdala and hippocampus. In *Proceedings of the* 17*th International Conference on Medical Image Computing and Computer Assisted Intervention* (*MICCAI*). *LNCS* **8674** 791–798.

CHUNG, M. K., HANSON, J. L., YE, J., DAVIDSON, R. J. and POLLAK, S. D. (2015). Persistent homology in sparse regression and its application to brain morphometry. *IEEE Trans. Med. Imag.* **34** 1928–1939.

COHEN-STEINER, D. and EDELSBRUNNER, H. (2009). Lipschitz functions have L_p-stable persistence. *Found. Comput. Math.* **10** 127–139.

COHEN-STEINER, D., EDELSBRUNNER, H. and HARER, J. (2007). Stability of persistence diagrams. *Discrete Comput. Geom.* **37** 103–120.

DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455. MR1311089

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

DONOHO, D. L., MALLAT, S. and VON SACHS, R. (1998). Estimating covariances of locally stationary processes: Rates of convergence of best basis methods. Technical report, Dept. Statistics, Stanford Univ. Stanford, CA.

EDELSBRUNNER, H. and HARER, J. L. (2010). *Computational Topology*: *An Introduction*. Amer. Math. Soc., Providence, RI. MR2572029

EDELSBRUNNER, H., LETSCHER, D. and ZOMORODIAN, A. (2002). Topological persistence and simplification. *Discrete Comput. Geom.* **28** 511–533. MR1949898

FRIED, I. (1997). Auras and experiential responses arising in the temporal lobe. *J. Neuropsychiatry Clin. Neurosci.* **9** 420–428.

GAMBLE, J. and HEO, G. (2010). Exploring uses of persistent homology for statistical analysis of landmark-based shape data. *J. Multivariate Anal.* **101** 2184–2199.

HATCHER, A. (2002). *Algebraic Topology*. Cambridge Univ. Press, Cambridge.

HEO, G., GAMBLE, J. and KIM, P. T. (2012). Topological analysis of variance and the maxillary complex. *J. Amer. Statist. Assoc.* **107** 477–492. MR2980059

KHALID, A., KIM, B. S., CHUNG, M. K., YE, J. C. and JEON, D. (2014). Tracing the evolution of multi-scale functional networks in a mouse model of depression using persistent brain network homology. *NeuroImage* **101** 351–363.

KOBAU, R., LUO, Y., ZACK, M., HELMERS, S. and THURMAN, D. (2012). Epilepsy in adults and access to care—United States, 2010. *Morb. Mort. Wkly. Rep.* **61** 910–913.

LANGE, H., LIEB, J., ENGEL, J. J. and CRANDALL, P. (1983). Temporo-spatial patterns of pre-ictal spike activity in human temporal lobe epilepsy. *Electroencephalogr. Clin. Neurophysiol.* **1978** 543–555.

LEE, H., CHUNG, M. K., KANG, H., KIM, B. N. and LEE, D. S. (2011). Computing the shape of brain networks using graph filtration and Gromov–Hausdorff metric. In *Proceedings of the* 14*th International Conference on Medical Image Computing and Computer-Assisted Intervention* (*MICCAI*). *LNCS* **6892** 302–309.

MARIS, E. (2012). Statistical testing in electrophysiological studies. *Psychophysiology* **49** 549–565.

MARTINERIE, J., ADAM, C., LE VAN QUYEN, M., BAULAC, M., CLÉMENCEAU, S., RENAULT, B. and VARELA, F. (1998). Can epileptic seizure be anticipated by nonlinear analysis? *Nat. Med.* **4** 1173–1176.

MCSHARRY, P. E., SMITH, L. A. and TARASSENKO, L. (2003). Prediction of epileptic seizures: Are nonlinear methods relevant? *Nat. Med.* **9** 241–242; author reply 242.

MILEYKO, Y., MUKHERJEE, S. and HARER, J. (2011). Probability measures on the space of persistence diagrams. *Inverse Probl.* **27** 1–21.

MILNOR, J. (1963). *Morse Theory*. Princeton Univ. Press, Princeton.

MITRA, P. P. and PESARAN, B. (1999). Analysis of dynamic brain imaging data. *Biophys. J.* **76** 691–708.

MOHSENI, H. R., MAGHSOUDI, A. and SHAMSOLLAHI, M. B. (2006). Seizure detection in EEG signals: A comparison of different approaches. In *Proceedings of the* 28*th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* 6724–6727.

OMBAO, H., VON SACHS, R. and GUO, W. (2005). SLEX analysis of multivariate nonstationary time series. *J. Amer. Statist. Assoc.* **100** 519–531.

OMBAO, H. C., RAZ, J. A., VON SACHS, R. and MALOW, B. A. (2001). Automatic statistical analysis of bivariate nonstationary time series. *J. Amer. Statist. Assoc.* **96** 543–560.

OPPENHEIM, A. V. and SCHAFER, R. W. (1989). *Discrete-Time Signal Processing*. Prentice Hall, New York.

REININGHAUS, J., HUBER, S., BAUER, U., TU, M. and KWITT, R. (2015). A stable multi-scale kernel for topological machine learning. Available at arXiv:1412.6821.

SOUSBIE, T., PICHON, C. and KAWAHARA, H. (2011). The persistent cosmic web and its filamentary structure. *Mon. Not. R. Astron. Soc.* **414** 384–403.

TURNER, K., MUKHERJEE, S. and BOYER, D. M. (2014). Persistent homology transform for modeling shapes and surfaces. *Inf. Inference* **3** 310–344.

TURNER, K., MILEYKO, Y., MUKHERJEE, S. and HARER, J. (2014). Frechet means for distributions of persistence diagrams. *Discrete Comput. Geom.* **52** 44–70.

VAN QUYEN, M. L., MARTINERIE, J., NAVARRO, V., BAULAC, M. and VARELA, F. J. (2001). Characterizing neurodynamic changes before seizures. *J. Clin. Neurophysiol.* **18** 191–208.

WHO (2005). Atlas: Epilepsy care in the world. Technical report.

WORSLEY, K. J. (1995). Estimating the number of peaks in a random field using the Hadwiger characteristic of excursion sets, with applications to medical images. *Ann. Statist.* **23** 640–669. MR1332586

ZHU, X., VARTANIAN, A., BANSAL, M., NGUYEN, D. and BRANDL, L. (2016). Stochastic multiresolution persistent homology kernel. In *Proceedings of the* 25*th International Joint Conference on Artificial Intelligence* (*IJCAI*) 2449–2455.

The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1535–1557
https://doi.org/10.1214/17-AOAS1120

JOINT SIGNIFICANCE TESTS FOR MEDIATION EFFECTS OF SOCIOECONOMIC ADVERSITY ON ADIPOSITY VIA EPIGENETICS[1]

BY YEN-TSUNG HUANG

Academia Sinica

Mediation analysis has become a popular practice in biomedical research. We conduct mediation analyses to investigate whether epigenetic variations mediate the effect of socioeconomic disadvantage on adiposity. Mediation effects can be expressed as a product of two parameters: one for the exposure-mediator association and the other for the mediator-outcome association conditional on the exposure. Under multi-mediator models, we study joint significance tests which examine the two parameters separately and compare with the widely used product significance tests which focus on the product of two parameters. Normal approximation of product significance tests depends on both effect size and sample size. We show that joint significance tests are intersection-union tests with size α and asymptotically more powerful than the normality-based product significance tests. Based on the theoretical results, we construct powerful testing procedures for gene-based mediation analyses and path-specific analyses. Advantage of joint significance tests is supported by simulation as well as the results of locus-based and gene-based mediation analyses of chromosome 17. Our analyses suggest that methylation of *FASN* gene mediates the effect of socioeconomic adversity on adiposity.

REFERENCES

AGHA, G., HOUSEMAN, E. A., KELSEY, K. T., EATON, C. B., BUKA, S. L. and LOUCKS, E. B. (2015). Adiposity is associated with DNA methylation profile in adipose tissue. *Int. J. Epidemiol.* **44** 1277–1287.

ALBERT, J. M. and NELSON, S. (2011). Generalized causal mediation analysis. *Biometrics* **67** 1028–1038. MR2829237

AROIAN, L. A. (1947). The probability function of the product of two normally distributed variables. *Ann. Math. Stat.* **18** 265–271. MR0021284

AVIN, C., SHPITSER, I. and PEARL, J. (2005). Identifiability of path-specific effects. In *Proceedings of the International Joint Conferences on Artificial Intelligence* 357–363.

BARON, R. M. and KENNY, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical consideration. *J. Pers. Soc. Psychol.* **51** 1173–1182.

BERGER, R. L. and HSU, J. C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statist. Sci.* **11** 283–319. MR1445984

BOLLEN, K. A. and STINE, R. (1990). Direct and indirect effects: Classical and bootstrap estimates of variability. *Sociol. Method.* **20** 115–140.

Key words and phrases. Intersection-union test, joint significance test, mediation analyses, multivariate analyses, normal product distribution, path-specific effect.

BORGHOL, N., SUDERMAN, M., MCARDLE, W., RACINE, A., HALLETT, M., PEMBREY, M., HERTZMAN, C., POWER, C. and SZYF, M. (2012). Associations with early-life socioeconomic position in adult DNA methylation. *Int. J. Epidemiol.* **41** 62–74.

DARMON, N. and DREWNOWSKI, A. (2008). Does social class predict diet quality? *Am. J. Clin. Nutr.* **87** 1107–1117.

DAVIS, S., DU, P., BILKE, S., TRICHE, T. J. and BOOTWALLA, M. (2015). methylumi: Handle Illumina methylation data. R package version 2.14.0.

GISKES, K., AVENDANO, M., BRUG, J. and KUNST, A. E. (2010). A systematic review of studies on socioeconomic inequalities in dietary intakes associated with weight gain and overweight/obesity conducted among European adults. *Obes. Rev.* **11** 413–429.

HARDY, J. B. (1971). The Johns Hopkins collaborative perinatal project. Descriptive background. *Johns Hopkins Med. J.* **128** 238–243.

HUANG, Y.-T. (2018). Supplement to "Joint significance tests for mediation effects of socioeconomic adversity on adiposity via epigenetics." DOI:10.1214/17-AOAS1120SUPP.

HUANG, Y.-T. and CAI, T. (2016). Mediation analysis for survival data using semiparametric probit models. *Biometrics* **72** 563–574. MR3515783

HUANG, Y.-T. and PAN, W.-C. (2016). Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics* **72** 401–413. MR3515767

HUANG, Y. T., CHU, S., LOUCKS, E. B., LIN, C. L., EATON, C. B., BUKA, S. L. and KELSEY, K. T. (2016). Epigenome-wide profiling of DNA methylation in paired samples of adipose tissue and blood. *Epigenetics* **11** 227–236.

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

KADOTA, Y., KAWAKAMI, T., TAKASAKI, S., SATO, M. and SUZUKI, S. (2016). Gene expression related to lipid and glucose metabolism in white adipose tissue. *Obes. Res. Clin. Pract.* **10** 85–93.

KOVACS, P., HARPER, I., HANSON, R. I., INFANTE, A. M., BOGARDUS, C., TATARANNI, P. A. and BAIER, L. J. (2004). A novel missense substitution (Val1483Ile) in the fatty acid synthase gene (FAS) is associated with percentage of body fat and substrate oxidation rates in nondiabetic Pima Indians. *Diabetes* **53** 1915–1919.

LANGE, T. and HANSEN, J. V. (2011). Direct and indirect effects in a survival context. *Epidemiology* **22** 575–581.

LOFTUS, T. M., JAWORSKY, D. E., FREHYWOT, G. L., TOWNSEND, C. A., RONNETT, G. V., LANE, M. D. and KUHAJDA, F. P. (2000). Reduced food intake and body weight in mice treated with fatty acid synthase inhibitors. *Science* **288** 2379–2381.

LOUCKS, E. B., HUANG, Y. T., AGHA, G., CHU, S., EATON, C. B., GILMAN, S. E., BUKA, S. L. and KELSEY, K. T. (2016). Epigenetic mediators between childhood socioeconomic disadvantage and mid-life body mass index: The New England Family Study. *Psychosom. Med.* **78** 1053–1065.

MACKINNON, D. (2008). *Introduction to Statistical Mediation Analysis*. Taylor & Francis, London.

MACKINNON, D. P., LOCKWOOD, C. M., HOFFMAN, J. M., WEST, S. G. and SHEETS, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychol. Methods* **7** 83–104.

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

PREACHER, K. J. and HAYES, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behav. Res. Methods* **40** 879–891.

ROBINS, J. M. (2003). *Semantics of Causal DAG Models and the Identification of Direct and Indirect Effects*. Oxford Univ. Press, New York.

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

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

SENESE, L. C., ALMEIDA, N. D., FATH, A. K., SMITH, B. T. and LOUCKS, E. B. (2009). Associations between childhood socioeconomic position and adulthood obesity. *Epidemiol. Rev.* **31** 21–51.

SOBEL, M. E. (1982). *Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models*. American Sociological Association, Washington, DC.

TAYLOR, A. B., MACKINNON, D. P. and TEIN, J.-Y. (2008). Tests of the three-path mediated effect. *Organ. Res. Methods* **11** 241–269.

TCHETGEN TCHETGEN, E. J. (2011). On causal mediation analysis with a survival outcome. *Int. J. Biostat.* **7** Article 33. MR2843528

TESCHENDORFF, A. E., MARABITA, F., LECHNER, M., BARTLETT, T., TEGNER, J., GOMEZ-CABRERO, D. and BECK, S. (2013). A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* **29** 189–196.

VANDERWEELE, T. J. (2011). Causal mediation analysis with survival data. *Epidemiology* **22** 582–585.

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.

VANDERWEELE, T. J. and VANSTEELANDT, S. (2013). Mediation analysis with multiple mediators. *Epidemiol. Methods* **2** 95–115.

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

The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1558–1582
https://doi.org/10.1214/17-AOAS1121

ADAPTIVE-WEIGHT BURDEN TEST FOR ASSOCIATIONS BETWEEN QUANTITATIVE TRAITS AND GENOTYPE DATA WITH COMPLEX CORRELATIONS

BY XIAOWEI WU[*], TING GUAN[*], DAJIANG J. LIU[†], LUIS G. LEÓN NOVELO[‡] AND DIPANKAR BANDYOPADHYAY[§,1]

Virginia Tech[*], Pennsylvania State University College of Medicine[†], University of Texas Health Science Center[‡] and Virginia Commonwealth University[§]

High throughput sequencing has often been used to screen samples from pedigrees or with population structure, producing genotype data with complex correlations caused by both familial relation and linkage disequilibrium. With such data it is critical to account for these genotypic correlations when assessing the contribution of multiple variants by gene or pathway. Recognizing the limitations of existing association testing methods, we propose *Adaptive-weight Burden Test* (ABT), a retrospective, mixed model test for genetic association of quantitative traits on genotype data with complex correlations. This method makes full use of genotypic correlations across both samples and variants and adopts "data driven" weights to improve power. We derive the ABT statistic and its explicit distribution under the null hypothesis and demonstrate through simulation studies that it is generally more powerful than the fixed-weight burden test and family-based SKAT in various scenarios, controlling for the type I error rate. Further investigation reveals the connection of ABT with kernel tests, as well as the adaptability of its weights to the direction of genetic effects. The application of ABT is illustrated by a gene-based association analysis of fasting glucose using data from the NHLBI "Grand Opportunity" Exome Sequencing Project.

REFERENCES

ANSORGE, W. J. (2009). Next-generation DNA sequencing techniques. *New Biotechnol*. **25** 195–203.

ASIMIT, J. and ZEGGINI, E. (2010). Rare variant association analysis methods for complex traits. *Annu. Rev. Genet*. **44** 293–308.

CHEN, H., MEIGS, J. B. and DUPUIS, J. (2013). Sequence kernel association test for quantitative traits in family samples. *Genet. Epidemiol*. **37** 196–204.

CHUN, H., BALLARD, D. H., CHO, J. and ZHAO, H. (2011). Identification of association between disease and multiple markers via sparse partial least-squares regression. *Genet. Epidemiol*. **35** 479–486.

FANG, S., ZHANG, S. and SHA, Q. (2014). Detecting association of rare variants by testing an optimally weighted combination of variants for quantitative traits in general families. *Ann. Hum. Genet*. **77** 524–534.

FUENTES, M. (2006). Testing for separability of spatial-temporal covariance functions. *J. Statist. Plann. Inference* **136** 447–466.

Key words and phrases. Genetic association test, burden test, kernel test, adaptive weight, complex genotypic correlation.

GAUDERMAN, W. J., MURCRAY, C., GILLILAND, F. and CONTI, D. V. (2007). Testing association between disease and multiple SNPs in a candidate gene. *Genet. Epidemiol.* **31** 383–395.

HAN, F. and PAN, W. (2010). A data-adaptive sum test for disease association with multiple common or rare variants. *Hum. Hered.* **70** 42–54.

HANSEN, J., RINNOV, A., KROGH-MADSEN, R., FISCHER, C. P., ANDREASEN, A. S., BERG, R. M., MØLLER, K., PEDERSEN, B. K. and PLOMGAARD, P. (2013). Plasma follistatin is elevated in patients with type 2 diabetes: Relationship to hyperglycemia, hyperinsulinemia, and systemic low-grade inflammation. *Diabetes/Metab. Res. Rev.* **29** 463–472.

INGELSSON, E., LANGENBERG, C., HIVERT, M. F., PROKOPENKO, I., LYSSENKO, V., DUPUIS, J., MÄGI, R., SHARP, S., JACKSON, A. U., ASSIMES, T. L. et al. (2010). Detailed physiologic characterization reveals diverse mechanisms for novel genetic loci regulating glucose and insulin metabolism in humans. *Diabetes* **59** 1266–1275.

JAKOBSDOTTIR, J. and MCPEEK, M. S. (2013). MASTOR: Mixed-model association mapping of quantitative traits in samples with related individuals. *Am. J. Hum. Genet.* **92** 652–666.

JIANG, D. and MCPEEK, M. S. (2013). Robust rare variant association testing for quantitative traits in samples with related individuals. *Genet. Epidemiol.* **38** 1–20.

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, S., WU, M. C. and LIN, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13** 762–775.

LEE, S., EMOND, M. J., BAMSHAD, M. J., BARNES, K. C., RIEDER, M. J., NICKERSON, D. A., NHLBI GO EXOME SEQUENCING PROJECT-ESP LUNG PROJECT TEAM, CHRISTIANI, D. C., WURFEL, M. M. and LIN, X. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* **91** 224–237.

LI, B. and LEAL, S. M. (2008). Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *Am. J. Hum. Genet.* **83** 311–321.

LI, M. X., GUI, H. S., KWAN, J. S. and SHAM, P. C. (2011). GATES: A rapid and powerful gene-based association test using extended simes procedure. *Am. J. Hum. Genet.* **88** 283–293.

LIN, D. Y. and TANG, Z. Z. (2011). A general framework for detecting disease associations with rare variants in sequencing studies. *Am. J. Hum. Genet.* **89** 354–367.

LIU, D. J. and LEAL, S. M. (2010). A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet.* **6** e1001156.

MA, L., CLARK, A. G. and KEINAN, A. (2013). Gene-based testing of interactions in association studies of quantitative traits. *PLoS Genet.* **9** e1003321.

MADSEN, B. E. and BROWNING, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* **5** e1000384.

MCPEEK, M. S., WU, X. and OBER, C. (2004). Best linear unbiased allele-frequency estimation in complex pedigrees. *Biometrics* **60** 359–367.

MELKERSSON, K. I., SCORDO, M. G., GUNES, A. and DAHL, M.-L. (2007). Impact of CYP1A2 and CYP2D6 polymorphisms on drug metabolism and on insulin and lipid elevations and insulin resistance in clozapine-treated patients. *J. Clin. Psychiatry* **68** 697–704.

MORGENTHALER, S. and THILLY, W. G. (2007). A strategy to discover genes that carrymultiallelic or mono-allelic risk for common diseases: A Cohort Allelic Sums Test (CAST). *Mutation Research* **615** 28–56.

NEALE, B. M., RIVAS, M. A., VOIGHT, B. F., ALTSHULER, D., DEVLIN, B., ORHO-MELANDER, M., KATHIRESAN, S., PURCELL, S. M., ROEDER, K. and DALY, M. J. (2011). Testing for an unusual distribution of rare variants. *PLoS Genet.* **7** e1001322.

PALATINI, P., BENETTI, E., MOS, L., GARAVELLI, G., MAZZER, A., COZZIO, S., FANIA, C. and CASIGLIA, E. (2015). Association of coffee consumption and CYP1A2 polymorphism with risk of impaired fasting glucose in hypertensive patients. *Eur. J. Epidemiol.* **30** 209–217.

PRICE, A. L., KRYUKOV, G. V., DE BAKKER, P. I., PURCELL, S. M., STAPLES, J., WEI, L. J. and SUNYAEV, S. R. (2010a). Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* **86** 832–838.

PRICE, A. L., ZAITLEN, N. A., REICH, D. and PATTERSON, N. (2010b). New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11** 459–463.

QI, Q., BRAY, G. A., HU, F. B., SACKS, F. M. and QI, L. (2012). Weight-loss diets modify glucose-dependent insulinotropic polypeptide receptor rs2287019 genotype effects on changes in body weight, fasting glucose, and insulin resistance: The preventing overweight using novel dietary strategies trial. *Am. J. Clin. Nutr.* **95** 506–513.

SAXENA, R., HIVERT, M. F., LANGENBERG, C., TANAKA, T., PANKOW, J. S., VOLLENWEIDER, P., LYSSENKO, V., BOUATIA-NAJI, N., DUPUIS, J., JACKSON, A. U. et al. (2010). Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. *Nat. Genet.* **42** 142–148.

SCHAID, D. J., MCDONNELL, S. K., SINNWELL, J. P. and THIBODEAU, S. M. (2013). Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. *Genet. Epidemiol.* **37** 409–418.

SCHIFANO, E. D., EPSTEIN, M. P., BIELAK, L. F., JHUN, M. A., KARDIA, S. L. R., PEYSER, P. A. and LIN, X. (2012). SNP set association analysis for familial data. *Genet. Epidemiol.* **36** 797–810.

SHA, Q. and ZHANG, S. (2014). A novel test for testing the optimally weighted combination of rare and common variants based on data of parents and affected children. *Genet. Epidemiol.* **38** 135–143.

SHA, Q., WANG, X., WANG, X. and ZHANG, S. (2012). Detecting association of rare and common variants by testing an optimally weighted combination of variants. *Genet. Epidemiol.* **36** 561–571.

SHENDURE, J. and JI, H. (2008). Next-generation DNA sequencing. *Nat. Biotechnol.* **26** 1135–1145.

SPLANSKY, G. L., COREY, D., YANG, Q., ATWOOD, L. D., CUPPLES, L. A., BENJAMIN, E. J., D'AGOSTINO SR., R. B., FOX, C. S., LARSON, M. G., MURABITO, J. M. et al. (2007). The third generation cohort of the national heart, lung, and blood institute's framingham heart study: Design, recruitment, and initial examination. *Am. J. Epidemiol.* **165** 1328–1335.

THE 1000 GENOMES PROJECT CONSORTIUM (2010). A map of human genome variation from population-scale sequencing. *Nature* **467** 1061–1073.

THORNTON, T. and MCPEEK, M. S. (2007). Case-control association testing with related individuals: A more powerful quasi-likelihood score test. *Am. J. Hum. Genet.* **81** 321–337.

THORNTON, T. and MCPEEK, M. S. (2010). ROADTRIPS: Case-control association testing with partially or completely unknown population and pedigree structure. *Am. J. Hum. Genet.* **86** 172–184.

WANG, K. and ABBOTT, D. (2008). A principal components regression approach to multilocus genetic association studies. *Genet. Epidemiol.* **32** 108–118.

WANG, Y., CHEN, Y. H. and YANG, Q. (2012). Joint rare variant association test of the average and individual effects for sequencing studies. *PLoS ONE* **7** e32485.

WANG, T. and ELSTON, R. C. (2007). Improved power by use of a weighted score test for linkage disequilibrium mapping. *Am. J. Hum. Genet.* **80** 353–360.

WANG, X., MORRIS, N. J., ZHU, X. and ELSTON, R. C. (2013a). A variance component based multi-marker association test using family and unrelated data. *BMC Genet.* **14** 17.

WANG, X., LEE, S., ZHU, X., REDLINE, S. and LIN, X. (2013b). GEE-based SNP set association test for continuous and discrete traits in family based association studies. *Genet. Epidemiol.* **37** 778–786.

WU, M. C., KRAFT, P., EPSTEIN, M. P., TAYLOR, D. M., CHANOCK, S. J., HUNTER, D. J. and LIN, X. (2010). Powerful SNP-set analysis for case-control genome-wide association studies. *Am. J. Hum. Genet.* **86** 929–942.

WU, M. C., LEE, S., CAI, T., LI, Y., BOEHNKE, M. and LIN, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89** 82–93.

WU, H., WU, M., CHEN, Y., ALLAN, C. A., PHILLIPS, D. J. and HEDGER, M. P. (2012). Correlation between blood activin levels and clinical parameters of type 2 diabetes. *Exp. Diabetes Res.* **2012** 410579.

WU, X., GUAN, T., LIU, D. J., NOVELO, L. G. and BANDYOPADHYAY, D. (2018). Supplement to "Adaptive-weight burden test for associations between quantitative traits and genotype data with complex correlations." DOI:10.1214/17-AOAS1121SUPP.

The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1583–1604
https://doi.org/10.1214/17-AOAS1122

BAYESIAN AGGREGATION OF AVERAGE DATA: AN APPLICATION IN DRUG DEVELOPMENT

BY SEBASTIAN WEBER*, ANDREW GELMAN[†,1], DANIEL LEE[‡], MICHAEL BETANCOURT[†,1], AKI VEHTARI[§,2] AND AMY RACINE-POON*

Novartis Pharma AG, Columbia University[†], Generable[‡] and Aalto University[§]*

Throughout the different phases of a drug development program, randomized trials are used to establish the tolerability, safety and efficacy of a candidate drug. At each stage one aims to optimize the design of future studies by extrapolation from the available evidence at the time. This includes collected trial data and relevant external data. However, relevant external data are typically available as averages only, for example, from trials on alternative treatments reported in the literature. Here we report on such an example from a drug development for wet age-related macular degeneration. This disease is the leading cause of severe vision loss in the elderly. While current treatment options are efficacious, they are also a substantial burden for the patient. Hence, new treatments are under development which need to be compared against existing treatments.

The general statistical problem this leads to is *meta-analysis*, which addresses the question of how we can combine data sets collected under different conditions. Bayesian methods have long been used to achieve partial pooling. Here we consider the challenge when the model of interest is complex (hierarchical and nonlinear) and one data set is given as raw data while the second data set is given as averages only. In such a situation, common meta-analytic methods can only be applied when the model is sufficiently simple for analytic approaches. When the model is too complex, for example, nonlinear, an analytic approach is not possible. We provide a Bayesian solution by using simulation to approximately reconstruct the likelihood of the external summary and allowing the parameters in the model to vary under the different conditions. We first evaluate our approach using fake data simulations and then report results for the drug development program that motivated this research.

REFERENCES

AMBATI, J. and FOWLER, B. J. (2012). Mechanisms of age-related macular degeneration. *Neuron* **75** 26–39.

AUGOOD, C. A., VINGERLING, J. R., DE JONG, P. T., CHAKRAVARTHY, U., SELAND, J., SOUBRANE, G., TOMAZZOLI, L., TOPOUZIS, F., BENTHAM, G., RAHU, M., VIOQUE, J., YOUNG, I. S. and FLETCHER, A. E. (2006). Prevalence of age-related maculopathy in older Europeans. *Arch. Ophthalmol.* **124** 529–535.

BETANCOURT, M. (2016). Diagnosing suboptimal cotangent disintegrations in Hamiltonian Monte Carlo. Preprint. Available at arXiv:1604.00695 [stat].

Key words and phrases. Meta-analysis, hierarchical modeling, Bayesian computation, pharmacometrics, Stan.

BROWN, D. M., KAISER, P. K., MICHELS, M., SOUBRANE, G., HEIER, J. S., KIM, R. Y., SY, J. P. and SCHNEIDER, S. (2006). Ranibizumab versus Verteporfin for Neovascular age-related macular degeneration. *N. Engl. J. Med.* **355** 1432–1444.

BUSCHINI, E., PIRAS, A., NUZZI, R. and VERCELLI, A. (2011). Age related macular degeneration and drusen: Neuroinflammation in the retina. *Prog. Neurobiol.* **95** 14–25.

CARO, J. J. and ISHAK, K. J. (2010). No head-to-head trial? Simulate the missing arms. *PharmacoEcon.* **28** 957–967.

DOMINICI, F., PARMIGIANI, G., WOLPERT, R. L. and HASSELBLAD, V. (1999). Meta-analysis of migraine headache teatments: Combining information from heterogeneous designs. *J. Amer. Statist. Assoc.* **94** 16–28.

GELMAN, A. (2004). Parameterization and Bayesian modeling. *J. Amer. Statist. Assoc.* **99** 537–545. MR2109315

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2014). *Bayesian Data Analysis*, 3rd ed. CRC Press, Boca Raton, FL. MR3235677

HARRIER. Efficacy and Safety of RTH258 Versus Aflibercept - Study 2 - ClinicalTrials.gov. Available at https://clinicaltrials.gov/ct2/show/NCT02434328.

HART, W. M., ed. (1992). *Adler's Physiology of the Eye*: *Clinical Application*, 9th ed. Mosby, St. Louis.

HAWK. Efficacy and Safety of RTH258 Versus Aflibercept - ClinicalTrials.gov. Available at https://clinicaltrials.gov/ct2/show/NCT02307682.

HEIER, J. S., BROWN, D. M., CHONG, V., KOROBELNIK, J.-F., KAISER, P. K., NGUYEN, Q. D., KIRCHHOF, B., HO, A., OGURA, Y., YANCOPOULOS, G. D., STAHL, N., VITTI, R., BERLINER, A. J., SOO, Y., ANDERESI, M., GROETZBACH, G., SOMMERAUER, B., SANDBRINK, R., SIMADER, C. and SCHMIDT-ERFURTH, U. (2012). Intravitreal Aflibercept (VEGF trap-eye) in wet age-related macular degeneration. *Ophthalmology* **119** 2537–2548.

HIGGINS, J. P. T. and GREEN, S. (2011). *Cochrane Handbook for Systematic Reviews of Interventions*, Version 5.1.0 ed. The Cochrane Collaboration.

HIGGINS, J. P. T. and WHITEHEAD, A. (1996). Borrowing strength from external trials in a meta-analysis. *Stat. Med.* **15** 2733–2749.

HOFFMAN, M. D. and GELMAN, A. (2014). The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **15** 1593–1623. MR3214779

ISHAK, K. J., PROSKOROVSKY, I. and BENEDICT, A. (2015). Simulation and matching-based approaches for indirect comparison of treatments. *PharmacoEcon.* **33** 537–549.

JUSKO, W. J. and KO, H. C. (1994). Physiologic indirect response models characterize diverse types of pharmacodynamic effects. *Clin. Pharmacol. Ther.* **56** 406–419.

KHANDHADIA, S., CIPRIANI, V., YATES, J. R. W. and LOTERY, A. J. (2012). Age-related macular degeneration and the complement system. *Immunobiology* **217** 127–146.

KINNUNEN, K., PETROVSKI, G., MOE, M. C., BERTA, A. and KAARNIRANTA, K. (2012). Molecular mechanisms of retinal pigment epithelium damage and development of age-related macular degeneration. *Acta Ophthalmol.* **90** 299–309.

POCOCK, S. J. (1976). The combination of randomized and historical controls in clinical trials. *J. Chronic Dis.* **29** 175–188.

ROSENFELD, P. J., BROWN, D. M., HEIER, J. S., BOYER, D. S., KAISER, P. K., CHUNG, C. Y. and KIM, R. Y. (2006). Ranibizumab for neovascular age-related macular degeneration. *N. Engl. J. Med.* **355** 1419–1431.

SCHMIDT-ERFURTH, U., ELDEM, B., GUYMER, R., KOROBELNIK, J.-F., SCHLINGEMANN, R. O., AXER-SIEGEL, R., WIEDEMANN, P., SIMADER, C., GEKKIEVA, M. and WEICHSELBERGER, A. (2011). Efficacy and safety of monthly versus quarterly Ranibizumab treatment in neovascular age-related macular degeneration: The EXCITE study. *Ophthalmology* **118** 831–839.

SHEINER, L. B. (1997). Learning versus confirming in clinical drug development. *Clin. Pharmacol. Ther.* **61** 275–291.

SIGNOROVITCH, J. E., WU, E. Q., YU, A. P., GERRITS, C. M., KANTOR, E., BAO, Y., GUPTA, S. R. and MULANI, P. M. (2010). Comparative effectiveness without head-to-head trials. *PharmacoEcon.* **28** 935–945.

STAN DEVELOPMENT TEAM (2017). Stan: A C++ library for probability and sampling.

WEBER, S., CARPENTER, B., LEE, D., BOIS, F. Y., GELMAN, A. and RACINE, A. (2014). Bayesian drug disease model with Stan: Using published longitudinal data summaries in population models, Population Approach Group Europe Meeting 2014, Alicante, Spain. Available at http://page-meeting.org/?abstract=3200.

WEBER, S., GELMAN, A., LEE, D., BETANCOURT, M., VEHTARI, A. and RACINE-POON, A. (2018). Supplement to "Bayesian aggregation of average data: An application in drug development." DOI:10.1214/17-AOAS1122SUPP.

XU, L., LU, T., TUOMI, L., JUMBE, N., LU, J., EPPLER, S., KUEBLER, P., DAMICO-BEYER, L. A. and JOSHI, A. (2013). Pharmacokinetics of Ranibizumab in patients with neovascular age-related macular degeneration: A population approach. *Investig. Ophthalmol. Vis. Sci.* **54** 1616–1624.

The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1605–1627
https://doi.org/10.1214/17-AOAS1123

BAYCOUNT: A BAYESIAN DECOMPOSITION METHOD FOR INFERRING TUMOR HETEROGENEITY USING RNA-SEQ COUNTS

BY FANGZHENG XIE*, MINGYUAN ZHOU† AND YANXUN XU*,[1]

*Johns Hopkins University** *and University of Texas at Austin*†

Tumors are heterogeneous. A tumor sample usually consists of a set of subclones with distinct transcriptional profiles and potentially different degrees of aggressiveness and responses to drugs. Understanding tumor heterogeneity is therefore critical for precise cancer prognosis and treatment. In this paper we introduce BayCount—a Bayesian decomposition method to infer tumor heterogeneity with highly over-dispersed RNA sequencing count data. Using negative binomial factor analysis, BayCount takes into account both the between-sample and gene-specific random effects on raw counts of sequencing reads mapped to each gene. For the posterior inference, we develop an efficient compound Poisson-based blocked Gibbs sampler. Simulation studies show that BayCount is able to accurately estimate the subclonal inference, including the number of subclones, the proportions of these subclones in each tumor sample, and the gene expression profiles in each subclone. For real world data examples, we apply BayCount to The Cancer Genome Atlas lung cancer and kidney cancer RNA sequencing count data and obtain biologically interpretable results. Our method represents the first effort in characterizing tumor heterogeneity using RNA sequencing count data that simultaneously removes the need of normalizing the counts, achieves statistical robustness, and obtains biologically/clinically meaningful insights. The R package `BayCount` implementing our model and algorithm is available for download.

REFERENCES

ABBAS, A. R., WOLSLEGEL, K., SESHASAYEE, D., MODRUSAN, Z. and CLARK, H. F. (2009). Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS ONE* **4** e6098.

ANSCOMBE, F. J. (1948). The transformation of Poisson, binomial and negative-binomial data. *Biometrika* **35** 246–254.

CANCER GENOME ATLAS RESEARCH NETWORK (2012). Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489** 519–525.

CANCER GENOME ATLAS RESEARCH NETWORK (2013). Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499** 43–49.

CARTER, S. L., CIBULSKIS, K., HELMAN, E., MCKENNA, A., SHEN, H., ZACK, T., LAIRD, P. W., ONOFRIO, R. C., WINCKLER, W., WEIR, B. A. et al. (2012). Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30** 413–421.

Key words and phrases. Cancer genomics, compound Poisson, Markov chain Monte Carlo, negative binomial, overdispersion.

DEPIANTO, D., KERNS, M. L., DLUGOSZ, A. A. and COULOMBE, P. A. (2010). Keratin 17 promotes epithelial proliferation and tumor growth by polarizing the immune response in skin. *Nat. Genet.* **42** 910–914.

DESHWAR, A. G., VEMBU, S., YUNG, C. K., JANG, G. H., STEIN, L. and MORRIS, Q. (2015). PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.* **16** 1.

DILLIES, M.-A., RAU, A., AUBERT, J., HENNEQUET-ANTIER, C., JEANMOUGIN, M., SERVANT, N., KEIME, C., MAROT, G., CASTEL, D., ESTELLE, J. et al. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.* **14** 671–683.

DING, L., LEY, T. J., LARSON, D. E., MILLER, C. A., KOBOLDT, D. C., WELCH, J. S., RITCHEY, J. K., YOUNG, M. A., LAMPRECHT, T., MCLELLAN, M. D. et al. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481** 506–510.

FAN, J., SALATHIA, N., LIU, R., KAESER, G. E., YUNG, Y. C., HERMAN, J. L., KAPER, F., FAN, J.-B., ZHANG, K., CHUN, J. et al. (2016). Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat. Methods* **13** 241–244.

GHAHRAMANI, Z., MOHAMED, S. and HELLER, K. A. (2014). *Partial Membership and Factor Analysis*. Chapman & Hall/CRC, London.

GONG, T., HARTMANN, N., KOHANE, I. S., BRINKMANN, V., STAEDTLER, F., LETZKUS, M., BONGIOVANNI, S. and SZUSTAKOWSKI, J. D. (2011). Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples. *PLoS ONE* **6** e27156.

HORE, V., VIÑUELA, A., BUIL, A., KNIGHT, J., MCCARTHY, M. I., SMALL, K. and MARCHINI, J. (2016). Tensor decomposition for multiple-tissue gene expression experiments. *Nat. Genet.* **48** 1094–1100.

JOHNSON, N. L., KOTZ, S. and BALAKRISHNAN, N. (1997). *Discrete Multivariate Distributions* **165**. Wiley, New York.

KARANTZA, V. (2011). Keratins in health and cancer: More than mere epithelial cell markers. *Oncogene* **30** 127–138.

KHARCHENKO, P. V., SILBERSTEIN, L. and SCADDEN, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11** 740–742.

KIM, K.-T., LEE, H. W., LEE, H.-O., KIM, S. C., SEO, Y. J., CHUNG, W., EUM, H. H., NAM, D.-H., KIM, J., JOO, K. M. et al. (2015). Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. *Genome Biol.* **16** 127.

KUDRIAVTSEVA, A., ANEDCHENKO, E., OPARINA, N. Y., KRASNOV, G., KASHKIN, K., DMITRIEV, A., ZBOROVSKAYA, I., KONDRATJEVA, T., VINOGRADOVA, E., ZINOVYEVA, M. et al. (2009). Expression of FTL and FTH genes encoding ferritin subunits in lung and renal carcinomas. *Mol. Biol.* **43** 972–981.

LÄHDESMÄKI, H., DUNMIRE, V., YLI-HARJA, O., ZHANG, W. et al. (2005). In silico microdissection of microarray data from heterogeneous cell populations. *BMC Bioinform.* **6** 1.

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

LEE, S., CHUGH, P. E., SHEN, H., EBERLE, R. and DITTMER, D. P. (2013). Poisson factor models with applications to non-normalized microRNA profiling. *Bioinformatics* **29** 1105–1111.

LEE, J., MÜLLER, P., SENGUPTA, S., GULUKOTA, K. and JI, Y. (2016). Bayesian inference for intratumour heterogeneity in mutations and copy number variation. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **65** 547–563. MR3522952

LIAO, Y., SMYTH, G. K. and SHI, W. (2014). Featurecounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30** 923–930.

MACOSKO, E. Z., BASU, A., SATIJA, R., NEMESH, J., SHEKHAR, K., GOLDMAN, M., TIROSH, I., BIALAS, A. R., KAMITAKI, N., MARTERSTECK, E. M. et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161** 1202–1214.

MARUSYK, A., ALMENDRO, V. and POLYAK, K. (2012). Intra-tumour heterogeneity: A looking glass for cancer? *Nat. Rev. Cancer* **12** 323–334.

NIK-ZAINAL, S., VAN LOO, P., WEDGE, D. C., ALEXANDROV, L. B., GREENMAN, C. D., LAU, K. W., RAINE, K., JONES, D., MARSHALL, J., RAMAKRISHNA, M. et al. (2012). The life history of 21 breast cancers. *Cell* **149** 994–1007.

OESPER, L., MAHMOODY, A. and RAPHAEL, B. J. (2013). THetA: Inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol.* **14** R80.

OSHLACK, A. and WAKEFIELD, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct* **4** 1.

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

QUENOUILLE, M. H. (1949). A relation between the logarithmic, Poisson, and negative binomial series. *Biometrics* **5** 162–164.

RAHMAN, M., JACKSON, L. K., JOHNSON, W. E., LI, D. Y., BILD, A. H. and PICCOLO, S. R. (2015). Alternative preprocessing of RNA-sequencing data in the Cancer Genome Atlas leads to improved analysis results. *Bioinformatics* **31** 3666–3672.

REPSILBER, D., KERN, S., TELAAR, A., WALZL, G., BLACK, G. F., SELBIG, J., PARIDA, S. K., KAUFMANN, S. H. and JACOBSEN, M. (2010). Biomarker discovery in heterogeneous tissue samples-taking the in-silico deconfounding approach. *BMC Bioinform.* **11** 1.

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

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

SHEN, H. and HUANG, J. Z. (2008). Forecasting time series of inhomogeneous Poisson processes with application to call center workforce management. *Ann. Appl. Stat.* **2** 601–623. MR2524348

SHEN-ORR, S. S., TIBSHIRANI, R., KHATRI, P., BODIAN, D. L., STAEDTLER, F., PERRY, N. M., HASTIE, T., SARWAL, M. M., DAVIS, M. M. and BUTTE, A. J. (2010). Cell type–specific gene expression differences in complex tissues. *Nat. Methods* **7** 287–289.

SHINTANI, Y., MAEDA, M., CHAIKA, N., JOHNSON, K. R. and WHEELOCK, M. J. (2008). Collagen I promotes epithelial-to-mesenchymal transition in lung cancer cells via transforming growth factor-beta signaling. *Am. J. Respir. Cell Mol. Biol.* **38** 95–104.

VENET, D., PECASSE, F., MAENHAUT, C. and BERSINI, H. (2001). Separation of samples into their constituents using gene expression data. *Bioinformatics* **17** S279–S287.

WANG, M., MASTER, S. R. and CHODOSH, L. A. (2006). Computational expression deconvolution in a complex mammalian organ. *BMC Bioinform.* **7** 1.

WANG, N., HOFFMAN, E. P., CHEN, L., CHEN, L., ZHANG, Z., LIU, C., YU, G., HERRINGTON, D. M., CLARKE, R. and WANG, Y. (2016). Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues. *Sci. Rep.* **6**.

WILKERSON, M. D., YIN, X., HOADLEY, K. A., LIU, Y., HAYWARD, M. C., CABANSKI, C. R., MULDREW, K., MILLER, C. R., RANDELL, S. H., SOCINSKI, M. A. et al. (2010). Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types. *Clin. Cancer Res.* **16** 4864–4875.

WILKS, C., CLINE, M. S., WEILER, E., DIEHKANS, M., CRAFT, B., MARTIN, C., MURPHY, D., PIERCE, H., BLACK, J., NELSON, D. et al. (2014). The Cancer Genomics Hub (CGHub): Overcoming cancer through the power of torrential data. *Database* **2014**.

XIE, F., ZHOU, M. and XU, Y. (2018). Supplement to "BayCount: A Bayesian decomposition method for inferring tumor heterogeneity using RNA-Seq counts." DOI:10.1214/17-AOAS1123SUPP.

XU, Y., MÜLLER, P., YUAN, Y., GULUKOTA, K. and JI, Y. (2015). MAD Bayes for tumor heterogeneity—feature allocation with exponential family sampling. *J. Amer. Statist. Assoc.* **110** 503–514. MR3367243

ZHOU, M. (2016). Nonparametric Bayesian negative binomial factor analysis. Preprint. Available at arXiv:1604.07464.

ZHOU, M. and CARIN, L. (2012). Augment-and-conquer negative binomial processes. In *Advances in Neural Information Processing Systems* 2546–2554.

ZHOU, M., HANNAH, L., DUNSON, D. B. and CARIN, L. (2012). Beta-negative binomial process and Poisson factor analysis. In *AISTATS* **22** 1462–1471.

ZHU, M. and GHODSI, A. (2006). Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Comput. Statist. Data Anal.* **51** 918–930.

ZHU, J., CHEN, X., LIAO, Z., HE, C. and HU, X. (2015). TGFBI protein high expression predicts poor prognosis in colorectal cancer patients. *Int. J. Clin. Exp. Pathol.* **8** 702.

The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1628–1654
https://doi.org/10.1214/17-AOAS1124

EXPLORING THE CONFORMATIONAL SPACE FOR PROTEIN FOLDING WITH SEQUENTIAL MONTE CARLO

BY SAMUEL W. K. WONG*, JUN S. LIU[†,1] AND S. C. KOU[†,2]

*University of Florida** *and Harvard University*[†]

Computational methods for protein structure prediction from amino acid sequence are of vital importance in modern applications, for example protein design in biomedicine. Efficient sampling of conformations according to a given energy function remains a bottleneck, yet is a vital step for energy-based structure prediction methods. While the Protein Data Bank of experimentally determined 3-D protein structures has steadily increased in size, structure predictions for new proteins tend to be unreliable in the amino acid segments where there is low sequence similarity with known structures. In this paper we introduce a new method for building such segments of protein structures, inspired by sequential Monte Carlo methods. We apply our method to examples of real 3-D structure predictions and demonstrate its promise for improving low confidence segments. We also provide applications to the prediction of reconstructed segments in known structures, and to the assessment of energy function accuracy. We find that our method is able to produce conformations that have both low energies and good coverage of the conformational space and hence can be a useful tool for protein design and structure prediction.

REFERENCES

ANFINSEN, C. (1973). Principles that govern the folding of protein chains. *Science* **181** 223–230.

BERNSTEIN, F. C., KOETZLE, T. F., WILLIAMS, G. J., MEYER, E. F., BRICE, M. D., RODGERS, J. R., KENNARD, O., SHIMANOUCHI, T. and TASUMI, M. (1977). The protein data bank. *Eur. J. Biochem.* **80** 319–324.

BROOKS, C. L., ONUCHIC, J. N. and WALES, D. J. (2001). Taking a walk on a landscape. *Science* **293** 612–613.

CANUTESCU, A. and DUNBRACK, R. (2003). Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Sci.* **12** 963–972.

COOPER, S., KHATIB, F., TREUILLE, A., BARBERO, J., LEE, J., BEENEN, M., LEAVER-FAY, A., BAKER, D., POPOVIĆ, Z. et al. (2010). Predicting protein structures with a multiplayer online game. *Nature* **466** 756–760.

COUTSIAS, E., SEOK, C., JACOBSON, M. and DILL, K. (2004). A kinematic view of loop closure. *J. Comput. Chem.* **25** 510–528.

DILL, K. A. and MACCALLUM, J. L. (2012). The protein-folding problem, 50 years on. *Science* **338** 1042–1046.

DOUC, R. and CAPPÉ, O. (2005). Comparison of resampling schemes for particle filtering. In *Image and Signal Processing and Analysis*, 2005. *ISPA* 2005. *Proceedings of the* 4*th International Symposium on* 64–69. IEEE, New York.

Key words and phrases. Protein structure prediction, particle filter, structure refinement, energy optimization.

DOUCET, A., DE FREITAS, N. and GORDON, N. (2001). An introduction to sequential Monte Carlo methods. In *Sequential Monte Carlo Methods in Practice*. 3–14. Springer, New York. MR1847784

EDDY, S. R. (2004). Where did the BLOSUM62 alignment score matrix come from? *Nat. Biotechnol.* **22** 1035–1036.

ENGH, R. and HUBER, R. (1991). Accurate bond and angle parameters for X-ray protein-structure refinement. *Acta Crystallogr. Sect. A* **47** 392–400.

FEARNHEAD, P. and CLIFFORD, P. (2003). On-line inference for hidden Markov models via particle filters. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 887–899. MR2017876

FISER, A. and ŠALI, A. (2003). Modeller: Generation and refinement of homology-based protein structure models. *Methods Enzymol.* **374** 461–491.

FRIESNER, R. A., PRIGOGINE, I. and RICE, S. A. (2002). *Computational Methods for Protein Folding*. Wiley, New York.

JONES, J. E. (1924). On the determination of molecular fields. II. From the equation of state of a gas. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **106** 463–477.

KABSCH, W. and SANDER, C. (1983). Dictionary of protein secondary structure—pattern-recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22** 2577–2637.

KENDREW, J. C., BODO, G., DINTZIS, H. M., PARRISH, R., WYCKOFF, H. and PHILLIPS, D. C. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **181** 662–666.

KHOURY, G. A., SMADBECK, J., KIESLICH, C. A. and FLOUDAS, C. A. (2014). Protein folding and de novo protein design for biotechnological applications. *Trends Biotechnol.* **32** 99–109.

KRISSINEL, E. (2007). On the relationship between sequence and structure similarities in proteomics. *Bioinformatics* **23** 717–723.

LAZARIDIS, T. and KARPLUS, M. (2000). Effective energy functions for protein structure prediction. *Curr. Opin. Struck. Biol.* **10** 139–145.

LEE, D., REDFERN, O. and ORENGO, C. (2007). Predicting protein function from sequence and structure. *Nat. Rev., Mol. Cell Biol.* **8** 995–1005.

LI, J., ABEL, R., ZHU, K., CAO, Y., ZHAO, S. and FRIESNER, R. A. (2011). The VSGB 2.0 model: A next generation energy model for high resolution protein structure modeling. *Proteins* **79** 2794–2812.

LIANG, S., ZHANG, C. and STANDLEY, D. M. (2011). Protein loop selection using orientation-dependent force fields derived by parameter optimization. *Proteins* **79** 2260–2267.

LIANG, S., ZHANG, C. and ZHOU, Y. (2014). LEAP: Highly accurate prediction of protein loop conformations by integrating coarse-grained sampling and optimized energy scores with all-atom refinement of backbone and side chains. *J. Comput. Chem.* **35** 335–341.

LIN, M., CHEN, R. and LIU, J. S. (2013). Lookahead strategies for sequential Monte Carlo. *Statist. Sci.* **28** 69–94. MR3075339

LIU, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer, New York. MR1842342

LIU, J. S. and CHEN, R. (1998). Sequential Monte Carlo methods for dynamic systems. *J. Amer. Statist. Assoc.* **93** 1032–1044. MR1649198

LIU, J. S., CHEN, R. and WONG, W. H. (1998). Rejection control and sequential importance sampling. *J. Amer. Statist. Assoc.* **93** 1022–1031. MR1649197

LIU, J. S., LIANG, F. and WONG, W. H. (2000). The multiple-try method and local optimization in Metropolis sampling. *J. Amer. Statist. Assoc.* **95** 121–134. MR1803145

MANDELL, D. J., COUTSIAS, E. A. and KORTEMME, T. (2009). Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nat. Methods* **6** 551–552.

MODI, V. and DUNBRACK, R. L. (2016). Assessment of refinement of template-based models in CASP11. *Proteins* **84** 260–281.

MOULT, J., FIDELIS, K., KRYSHTAFOVYCH, A., SCHWEDE, T. and TRAMONTANO, A. (2016). Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. *Proteins* **84** 4–14.

ONUCHIC, J. N., LUTHEY-SCHULTEN, Z. and WOLYNES, P. G. (1997). Theory of protein folding: The energy landscape perspective. *Annu. Rev. Phys. Chem.* **48** 545–600.

RAMACHANDRAN, G., RAMAKRISHNAN, C. and SAISEKHARAN, V. (1963). Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7** 95–99.

ROHL, C. A., STRAUSS, C. E., CHIVIAN, D. and BAKER, D. (2004). Modeling structurally variable regions in homologous proteins with rosetta. *Proteins* **55** 656–677.

SHAPOVALOV, M. V. and DUNBRACK, R. L. (2011). A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure* **19** 844–858.

SÖDING, J., BIEGERT, A. and LUPAS, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33** W244–W248.

SOTO, C. S., FASNACHT, M., ZHU, J., FORREST, L. and HONIG, B. (2008). Loop modeling: Sampling, filtering, and scoring. *Proteins* **70** 834–843.

TAN, K., GU, M., CLANCY, S. and JOACHIMIAK, A. (2016). The crystal structure of the catalytic domain of peptidoglycan N-acetylglucosamine deacetylase from Eubacterium rectale ATCC 33656 (CASP target). PDB ID: 5JMU. DOI:10.2210/pdb5jmu/pdb.

TANG, K., ZHANG, J. and LIANG, J. (2014). Fast protein loop sampling and structure prediction using distance-guided sequential chain-growth Monte Carlo method. *PLoS Comput. Biol.* **10** e1003539.

VLUGT, T., MARTIN, M., SMIT, B., SIEPMANN, J. and KRISHNA, R. (1998). Improving the efficiency of the configurational-bias Monte Carlo algorithm. *Mol. Phys.* **94** 727–733.

WANG, G. and DUNBRACK, R. L. (2003). PISCES: A protein sequence culling server. *Bioinformatics* **19** 1589–1591.

WANG, F. and LANDAU, D. P. (2001). Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.* **86** 2050–2053.

WICK, C. and SIEPMANN, J. (2000). Self-adapting fixed-end-point configurational-bias Monte Carlo method for the regrowth of interior segments of chain molecules with strong intramolecular interactions. *Macromolecules* **33** 7207–7218.

WONG, W., CUI, Y. and CHEN, R. (1998). Torsional relaxation for biopolymers. *J. Comput. Biol.* **5** 655–665.

WONG, S. W. K., LIU, J. S. and KOU, S. C. (2017). Fast *de novo* discovery of low-energy protein loop conformations. *Proteins* **85** 1402–1412.

ZHANG, J., KOU, S. C. and LIU, J. S. (2007). Biopolymer structure simulation and optimization via fragment regrowth Monte Carlo. *J. Chem. Phys.* **126** 225101. DOI:10.1063/1.2736681.

ZHANG, J., LIN, M., CHEN, R., LIANG, J. and LIU, J. S. (2007). Monte Carlo sampling of near-native structures of proteins with applications. *Proteins* **66** 61–68.

ZHOU, H. and ZHOU, Y. (2002). Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.* **11** 2714–2726.

The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1655–1678
https://doi.org/10.1214/17-AOAS1125

SEQUENTIAL DOUBLE CROSS-VALIDATION FOR ASSESSMENT OF ADDED PREDICTIVE ABILITY IN HIGH-DIMENSIONAL OMIC APPLICATIONS

BY MAR RODRÍGUEZ-GIRONDO[*], PERTTU SALO[†], TOMASZ BURZYKOWSKI[‡], MARKUS PEROLA[†], JEANINE HOUWING-DUISTERMAAT[*,§] AND BART MERTENS[*]

Leiden University Medical Center[*], National Institute For Health and Welfare[†], Hasselt University[‡] and Leeds University[§]

Enriching existing predictive models with new biomolecular markers is an important task in the new multi-omic era. Clinical studies increasingly include new sets of omic measurements which may prove their added value in terms of predictive performance. We introduce a two-step approach for the assessment of the added predictive ability of omic predictors, based on sequential double cross-validation and regularized regression models. We propose several performance indices to summarize the two-stage prediction procedure and a permutation test to formally assess the added predictive value of a second omic set of predictors over a primary omic source. The performance of the test is investigated through simulations. We illustrate the new method through the systematic assessment and comparison of the performance of transcriptomics and metabolomics sources in the prediction of body mass index (BMI) using longitudinal data from the Dietary, Lifestyle, and Genetic determinants of Obesity and Metabolic syndrome (DILGOM) study, a population-based cohort from Finland.

REFERENCES

APALASAMY, Y. D. and MOHAMED, Z. (2015). Obesity and genomics: Role of technology in unraveling the complex genetic architecture of obesity. *Am. J. Hum. Genet.* **134** 361–374.

BOULESTEIX, A.-L. and HOTHORN, T. (2010). Testing the additional predictive value of high-dimensional molecular data. *BMC Bioinform.* **11** 78.

BREIMAN, L. (1996). Stacked regressions. *Mach. Learn.* **24** 49–64.

BÜHLMANN, P. and HOTHORN, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statist. Sci.* **22** 477–505. MR2420454

DELONG, E. R., DELONG, D. M. and CLARKE-PEARSON, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **44** 837–845.

DUDOIT, S., FRIDLYAND, J. and SPEED, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Statist. Assoc.* **97** 77–87. MR1963389

FRIEDMAN, J. HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

Key words and phrases. Added predictive ability, double cross-validation, regularized regression, multiple omics sets.

HARDIN, J., GARCIA, S. R. and GOLAN, D. (2013). A method for generating realistic correlation matrices. *Ann. Appl. Stat.* **7** 1733–1762. MR3127966

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

HERLIHY, M. and SHAVIT, N. (2012). *The Art of Multiprocessor Programming* (*Revised Edition*). Elsevier, New York.

HILDEN, J. and GERDS, T. A. (2014). A note on the evaluation of novel biomarkers: Do not rely on integrated discrimination improvement and net reclassification index. *Stat. Med.* **33** 3405–3414. MR3260635

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

HÖFLING, H. and TIBSHIRANI, R. (2008). A study of pre-validation. *Ann. Appl. Stat.* **2** 643–664. MR2524350

INOUYE, M., KETTUNEN, J., SOININEN, P., SILANDER, K., RIPATTI, S. et al. (2010). Metabonomic, transcriptomic, and genomic variation of a population cohort. *Mol. Syst. Biol.* **6** 441.

JENKINSON, C. P., GOERING, H. H. H., ARYA, R., BLANGERO, J., DUGGIRALA, R. and DEFRONZO, R. A. (2016). Transcriptomics in type 2 diabetes: Bridging the gap between genotype and phenotype. *Genomics Data* **8** 25–36.

JOLLIFFE, I. T. (2002). *Principal Component Analysis*, 2nd ed. Springer, New York. MR2036084

JONATHAN, P., KRZANOWSKI, W. J. and MCCARTHY, M. V. (2000). On the use of cross-validation to assess performance in multivariate prediction. *Stat. Comput.* **10** 209–229.

KERR, K. F., WANG, Z., JANES, H., MCCLELLAND, R. L., PSATY, B. M. and PEPE, M. S. (2014). Net reclassification indices for evaluating risk-prediction instruments: A. *Critical Review Epidemiology* **25** 114–121.

KNEIB, T., HOTHORN, T. and TUTZ, G. (2009). Variable selection and model choice in geoadditive regression models. *Biometrics* **65** 626–634. MR2751488

LIU, H., D'ANDRADE, P., FULMER-SMENTEK, S., LORENZI, P., KOHN, K. W., WEINSTEIN, J. N., POMMIER, Y. and REINHOLD, W. C. (2010). mRNA and microRNA expression profiles of the NCI-60 integrated with drug activities. *Mol. Cancer Ther.* **9** 1080–1091.

MARTENS, H. and NÆS, T. (1989). *Multivariate Calibration*. Wiley, Chichester. MR1029523

MERTENS, B. J. A., DE NOO, M. E., TOLLENAAR, R. A. E. M. and DEELDER, A. M. (2006). Mass spectrometry proteomic diagnosis: Enacting the double cross-validatory paradigm. *J. Comput. Biol.* **13** 1591–1605. MR2287728

MERTENS, B. J. A., VAN DE BURGT, Y. E. M., VELSTRA, B., MESKER, W. E., TOLLENAAR, R. A. E. M. and DEELDER, A. M. (2011). On the use of double cross-validation for the combination of proteomic mass spectral data for enhanced diagnosis and prediction. *Statist. Probab. Lett.* **81** 759–766. MR2793741

PENCINA, M. J., D'AGOSTINO, R. B. SR., D'AGOSTINO, R. B. JR. and VASAN, R. S. (2008). Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Stat. Med.* **27** 157–172. MR2412695

PENCINA, M. J., D'AGOSTINO, R. B., PENCINA, K. M., JANSSENS, C. J. W. and GREENLAND, P. (2012). Interpreting incremental value of markers added to risk prediction models. *Am. J. Epidemiol.* **176** 473–481.

PEPE, M. S., JANES, H. and LI, C. I. (2014). Net risk reclassification p values: Valid or misleading? *J. Natl. Cancer Inst.* **106** dju041.

RODRÍ GUEZ-GIRONDO, M., KNEIB, T., CADARSO-SUÁREZ, C. and ABU-ASSI, E. (2013). Model building in nonproportional hazard regression. *Stat. Med.* **32** 5301–5314. MR3141375

RODRÍGUEZ-GIRONDO, M., SALO, P., BURZYKOWSKI, T., PEROLA, M., HOUWING-DUISTERMAAT, J. and MERTENS, B. (2018). Supplement to "Sequential double cross-validation for assessment of added predictive ability in high-dimensional omic applications." DOI:10.1214/17-AOAS1125SUPPA, DOI:10.1214/17-AOAS1125SUPPB.

ROSENWALD, A., WRIGHT, G., CHAN, W. C., CONNORS, J. M., CAMPO, E. et al. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *N. Engl. J. Med.* **346** 1937–1947.

SCHEMPER, M. (2003). Predictive accuracy and explained variation. *Stat. Med.* **22** 2299–2308.

SCHWAMBORN, K. and CAPRIOLI, R. M. (2010). Molecular imaging by mass spectroscopy — looking beyond classical histology. *Nat. Rev.* **10** 639–646.

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

SOININEN, P., KANGAS, A. J., WURTZ, P., TUKIAINEN, T., TYNKKYNEN, T., LAATIKAINEN, R., JARVELIN, M. R., KAHONEN, M., LEHTIMAKI, T., VIIKARI, J., RAITAKARI, O. T., SAVOLAINEN, M. J. and ALA-KORPELA, M. (2009). High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism. *Anal.* **134** 1781–1785.

STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *J. Roy. Statist. Soc. Ser. B* **36** 111–147. MR0356377

STROEVE, J. H., SACCENTI, E., BOUWMAN, J., DANE, A., STRASSBURG, K. et al. (2016). Weight loss predictability by plasma metabolic signatures in adults with obesity and morbid obesity of the DiOGenes study. *J. Obesity* **24** 379–388.

THEODORATOU, E., THAÇI, K., AGAKOV, F., TIMOFEEVA, M. N., STAMBUK, J. et al. (2016). Glycosylation of plasma IgG in colorectal cancer prognosis. *Sci. Rep.* **6** 28098.

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

TIBSHIRANI, R. J. and EFRON, B. (2002). Pre-validation and inference in microarrays. *Stat. Appl. Genet. Mol. Biol.* **1** 1. MR2011184

TUTZ, G. and BINDER, H. (2006). Generalized additive modeling with implicit variable selection by likelihood-based boosting. *Biometrics* **62** 961–971. MR2297666

VAN DE WIEL, M. A., LIEN, T. G., VERLAAT, W., VAN WIERINGEN, W. N. and WILTING, S. M. (2016). Better prediction by use of co-data: Adaptive group-regularized ridge regression. *Stat. Med.* **35** 368–381. MR3455507

VARMA, S. and SIMON, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinform.* **7** 91.

WESTERHUIS, J. A., HOEFSLOOT, H. C. J., SMIT, S., VIS, D. J., SMILDE, A. K., VAN VELZEN, E. J. J., VAN DUIJNHOVEN, J. P. M. and VAN DORSTEN, F. A. (2008). Assessment of PLSDA cross validation. *Metabolomics* **4** 81–89.

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

ZHANG, B. and HORVATH, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4** 17. MR2170433

ZOLDOS, V., HORVAT, T. and LAUC, G. (2013). Glycomics meets genomics, epigenomics and other high throughput omics for system biology studies. *Curr. Opin. Chem. Biol.* **17** 34–40.

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
2018, Vol. 12, No. 3, 1679–1699
https://doi.org/10.1214/17-AOAS1126

JOINING THE INCOMPATIBLE: EXPLOITING PURPOSIVE LISTS FOR THE SAMPLE-BASED ESTIMATION OF SPECIES RICHNESS

BY ALESSANDRO CHIARUCCI*, ROSA MARIA DI BIASE†, LORENZO FATTORINI‡, MARZIA MARCHESELLI‡ AND CATERINA PISANI‡

University of Bologna, University of Tuscia† and University of Siena‡*

The lists of species obtained by purposive sampling by field ecologists can be used to improve the sample-based estimation of species richness. A new estimator is here proposed as a modification of the difference estimator in which the species inclusion probabilities are estimated by means of the species frequencies from incidence data. If the species list used to support the estimation is complete the estimator guesses the true richness without error. In the case of incomplete lists, the estimator provides values invariably greater than the number of species detected by the combination of sample-based and purposive surveys. An asymptotically conservative estimator of the mean squared error is also provided. A simulation study based on two artificial communities is carried out in order to check the obvious increase in accuracy and precision with respect to the widely applied estimators based on the sole sample information. Finally, the proposed estimator is adopted to estimate species richness in the Maremma Regional Park, Italy.

REFERENCES

ARRIGONI, P. V. (2003). The flora of the Maremma Natural Park (Tuscany, central Italy). *Webbia Journal of Plant Taxonomy and Geography* **58** 151–240.

BARABESI, L. and FATTORINI, L. (1998). The use of replicated plot, line and point sampling for estimating species abundances and ecological diversity. *Environ. Ecol. Stat.* **5** 353–370.

BUNGE, J. and FITZPATRICK, M. (1993). Estimating the number of species: A review. *J. Am. Stat. Assoc.* **88** 364–373.

CAYUELA, L., GOTELLI, N. J. and COLWELL, R. K. (2015). Ecological and biogeographic null hypotheses for comparing rarefaction curves. *Ecol. Monogr.* **85** 437–455.

CHAO, A. (1984). Nonparametric estimation of the number of classes in a population. *Scand. J. Stat.* **11** 265–270. MR0793175

CHAO, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* **43** 783–791.

CHAO, A. and COLWELL, R. H. (2017). Thirty years of progeny from Chao's inequality: Estimating and comparing richness with incidence data and incomplete sampling. *SORT* **41** 3–54.

CHAO, A. and LEE, M. (1992). Estimating the number of classes via sample coverage. *J. Am. Stat. Assoc.* **87** 210–217.

CHAO, A., GOTELLI, N. J., HSIEH, T. C., SANDER, E. L., MA, K. H., COLWELL, R. K. and ELLISON, A. M. (2014). Rarefaction and extrapolation with Hill numbers: A framework for sampling and estimation in species diversity studies. *Ecol. Monogr.* **84** 45–67.

Key words and phrases. Difference estimator, probabilistic sampling, purposive survey, supporting list, simulation.

CHIARUCCI, A. (2012). Estimating species richness: Still a long way off! *J. Veg. Sci.* **23** 1003–1005.
CHIARUCCI, A., BACARO, G. and SCHEINER, S. M. (2011). Old and new challenges in using species diversity for assessing biodiversity. *Philos. T. Roy. Soc. B* **366** 2426–2437.
CHIARUCCI, A., ENRIGHT, N. J., PERRY, G. L. W., MILLER, B. P. and LAMONT, B. B. (2003). Performance of nonparametric species richness estimators in a high diversity plant community. *Divers. Distrib.* **9** 283–295.
CHIARUCCI, A., DI BIASE, R. M., FATTORINI, L., MARCHESELLI, M. and PISANI, C. (2018). Supplement to "Joining the incompatible: Exploiting purposive lists for the sample-based estimation of species richness." DOI:10.1214/17-AOAS1126SUPP.
COLWELL, R. K. (2013). EstimateS: Statistical estimation of species richness and shared species from samples. Version 9. User's Guide and application. Published at http://purl.oclc.org/estimates.
COLWELL, R. K. and CODDINGTON, J. A. (1994). Estimating terrestrial biodiversity through extrapolation. *Philos. T. Roy. Soc. B* **345** 101–118.
COLWELL, R. K., CHAO, A., GOTELLI, N. J., LIN, S. Y., MAO, C. X., CHAZDON, R. L. and LONGINO, J. T. (2012). Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblage. *J. Plant Ecol.* **5** 3–21.
CONTI, F., ABBATE, G., ALESSANDRINI, A. and BLASI, C., eds. (2005). *An Annotated Checklist of the Italian Vascular Flora*. Palombi, Roma.
CORMACK, R. M. (1989). Log-linear models for capture-recapture. *Biometrics* **45** 395–413.
D'ALESSANDRO, L. and FATTORINI, L. (2002). Resampling estimators of species richness from presence-absence data: Why they don't work. *Metron* **60** 5–19. MR1973845
DIEKMANN, M., KÜHNE, A. and ISERMANN, M. (2007). Random vs non-random sampling: Effects on patterns of species abundance, species richness and vegetation-environment relationships. *Folia Geobot.* **42** 179–190.
FATTORINI, L. (2006). Applying the Horvitz–Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabilities. *Biometrika* **93** 269–278. MR2278082
FATTORINI, L. (2007). Statistical inference on accumulation curves for inventorying forest diversity: A design-based critical look. *Plant Biosyst.* **141** 231–242.
FATTORINI, L. (2009). An adaptive algorithm for estimating inclusion probabilities and performing Horvitz–Thompson criterion in complex designs. *Comput. Statist.* **24** 623–639.
FATTORINI, S. (2013). Regional insect inventories require long time, extensive spatial sampling and good will. *PLoS ONE* **8** e62118.
GASTON, K. J. (1996). Species richness: Measure and measurement. In *Biodiversity. A Biology of Numbers and Difference* (K. J. Gaston, ed.) 77–113. Blackwell Science, Oxford.
GOTELLI, N. J. and CHAO, A. (2013). Measuring and estimating species richness, species diversity, and biotic similarity from sampling data. In *Encyclopedia of Biodiversity*, 2nd ed. (S. A. Levin, ed.) **5** 195–211. Elsevier Ltd, Waltham.
GOTELLI, N. J. and COLWELL, R. K. (2001). Quantifying biodiversity: Procedures and pitfalls in the measurement and comparison of species richness. *Ecol. Lett.* **4** 379–391.
GOTELLI, N. J., ANDERSON, M. J., ARITA, H. T., CHAO, A., COLWELL, R. K., CONNOLLY, S. R., CURRIE, D. J., DUNN, R. R., GRAVES, G. R., GREEN, J. L., GRYTNES, J. A., JIANG, Y. H., JETZ, W., KATHLEEN LYONS, S., MCCAIN, C. M., MAGURRAN, A. E., RAHBEK, C., RANGEL, T. F., SOBERÓN, J., WEBB, C. O. and WILLIG, M. R. (2009). Patterns and causes of species richness: A general simulation model for macroecology. *Ecol. Lett.* **12** 873–886.
GREGOIRE, T. G. and VALENTINE, H. T. (2008). *Sampling Strategies for Natural Resources and the Environment*. Chapman & Hall, Boca Raton, FL.
HÉDL, R. (2007). Is sampling subjectivity a distorting factor in surveys for vegetation diversity? *Folia Geobot.* **42** 191–198.
HELLMANN, J. J. and FOWLER, G. W. (1999). Bias, precision, and accuracy of four measures of species richness. *Ecol. Appl.* **9** 824–834.

HELTSHE, J. F. and FORRESTER, N. E. (1983). Estimating species richness using the jackknife procedure. *Biometrics* **39** 1–11.

HOLDRIDGE, L. R., GRENKE, W. C., HATHEWAY, W. H., LIANG, T. and TOSI, J. A. (1971). *Forest Environments in Tropical Life Zones*. Pergamon Press, Oxford.

HORTAL, J., BORGES, P. A. V. and GASPAR, C. (2006). Evaluating the performance of species richness estimators: Sensitivity to sample grain size. *J. Anim. Ecol.* **75** 274–287.

HOWARD, P. C., VISKANIC, P., DAVENPORT, T. R. B., KIGENYI, F. W., BALTZER, M., DICKINSON, C. J., LWANGA, J. S., MATTHEWS, R. A. and BALMFORD, A. (1998). Complementarity and the use of indicator groups for reserve selection in Uganda. *Nature* **394** 472–475.

LEE, S. M. and CHAO, A. (1994). Estimating population size via sample coverage for closed capture-recapture models. *Biometrics* **50** 88–97.

MELO, A. S. (2004). A critique of the use of jackknife and related non-parametric techniques to estimate species richness. *Community Ecol.* **5** 149–157.

NICHOLS, J. D. and CONROY, M. J. (1996). Estimation of species richness. In *Measuring and Monitoring Biological Diversity. Standard Methods for Mammals* (D. E. Wilson, F. R. Cole, J. D. Nichols, R. Rudran and M. Forster, eds.) 226–234. Smithsonian Institution Press, Washington, DC.

OKSANEN, J., GUILLAUME BLANCHET, F., FRIENDLY, M., KINDT, R., LEGENDRE, P., MCGLINN, D., MINCHIN, P. R., O'HARA, R. B., SIMPSON, G. L., SOLYMOS, P., STEVENS, M. H. H., SZOECS, E. and WAGNER, H. (2016). vegan: Community ecology package. R package version 2.4-1. https://CRAN.R-project.org/package=vegan.

PALMER, M. W. (1990). The estimation of species richness by extrapolation. *Ecology* **71** 1195–1198.

PALMER, M. W. (1991). Estimating species richness: The second-order jackknife reconsidered. *Ecology* **72** 1512–1513.

PALMER, M. W., EARLS, P. G., HOAGLAND, B. W., WHITE, P. S. and WOHLGEMUTH, T. (2002). Quantitative tools for perfecting species lists. *Environmetrics* **13** 121–138.

PIELOU, E. C. (1977). *Mathematical Ecology*. Wiley, New York.

PIGNATTI, S. (1982). *Flora d'Italia*, *Vol.* 3, Edagricole edizioni.

SÄRNDAL, C. E. and LUNDSTRÖM, S. (2005). *Estimation in Survey with Nonresponse*. Wiley, New York.

SÄRNDAL, C. E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. Springer, New York.

SEBER, G. A. F. (1982). *The Estimation of Animal Abundance*. Griffin, London.

SKOV, F. and LAWESSON, J. E. (2000). Estimation of plant species richness from systematically placed plots in a managed forest ecosystem. *Nord. J. Bot.* **20** 477–483.

SMITH, E. P. and VAN BELLE, G. (1984). Nonparametric estimation of species richness. *Biometrics* **40** 119–129.

THOMPSON, S. K. (2002). *Sampling*, 2nd ed. Wiley, New York.

WALTHER, B. A. and MOORE, J. L. (2005). The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography* **28** 815–829.

WALTHER, B. A. and MORAND, S. (1998). Comparative performance of species richness estimation methods. *Parasitology* **116** 395–405.

WILSON, J. B., PEET, R. K., DENGLER, J. and PÄRTEL, M. (2012). Plant species richness: The world records. *J. Veg. Sci.* **23** 796–802.

XU, H., LIU, S., LI, Y., ZANG, R. and HE, F. (2012). Assessing non-parametric and area-based methods for estimating regional species richness. *J. Veg. Sci.* **23** 1006–1012.

The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1700–1726
https://doi.org/10.1214/17-AOAS1127

A GENERAL FRAMEWORK FOR ASSOCIATION ANALYSIS OF HETEROGENEOUS DATA

BY GEN LI[1] AND IRINA GAYNANOVA

Columbia University and Texas A&M University

Multivariate association analysis is of primary interest in many applications. Despite the prevalence of high-dimensional and non-Gaussian data (such as count-valued or binary), most existing methods only apply to low-dimensional data with continuous measurements. Motivated by the Computer Audition Lab 500-song (CAL500) music annotation study, we develop a new framework for the association analysis of two sets of high-dimensional and heterogeneous (continuous/binary/count) data. We model heterogeneous random variables using exponential family distributions, and exploit a structured decomposition of the underlying natural parameter matrices to identify shared and individual patterns for two data sets. We also introduce a new measure of the strength of association, and a permutation-based procedure to test its significance. An alternating iteratively reweighted least squares algorithm is devised for model fitting, and several variants are developed to expedite computation and achieve variable selection. The application to the CAL500 data sheds light on the relationship between acoustic features and semantic annotations, and provides effective means for automatic music annotation and retrieval.

REFERENCES

BACH, F. R. and JORDAN, M. I. (2005). A probabilistic interpretation of canonical correlation analysis. Technical Report 688, Dept. Statistics, Univ. California, Berkeley, Berkeley, CA.

BARRINGTON, L., CHAN, A., TURNBULL, D. and LANCKRIET, G. (2007). Audio information retrieval using semantic similarity. In *International Conference on Acoustics, Speech and Signal Processing* **2** 725–728. IEEE, New York.

BERTIN-MAHIEUX, T., ECK, D., MAILLET, F. and LAMERE, P. (2008). Autotagger: A model for predicting social tags from acoustic features on large music databases. *J. New Music Res.* **37** 115–135.

BJÖRCK, K. and GOLUB, G. H. (1973). Numerical methods for computing angles between linear subspaces. *Math. Comp.* **27** 579–594.

BROWNE, M. W. (1979). The maximum-likelihood solution in inter-battery factor analysis. *Br. J. Math. Stat. Psychol.* **32** 75–86. MR0553146

CHAUDHURI, K., KAKADE, S. M., LIVESCU, K. and SRIDHARAN, K. (2009). Multi-view clustering via canonical correlation analysis. In *Proceedings of the* 26*th Annual International Conference on Machine Learning* 129–136. ACM, New York.

CHEN, X. and LIU, H. (2012). An efficient optimization algorithm for structured sparse cca, with applications to eqtl mapping. *Stat. Biosci.* **4** 3–26.

Key words and phrases. Exponential family, inter-battery factor analysis, joint and individual structure, matrix decomposition, generalized linear model, association coefficient.

CHEN, M., GAO, C., REN, Z. and ZHOU, H. H. (2013). Sparse cca via precision adjusted iterative thresholding. ArXiv preprint. Available at arXiv:1311.6186.

CHENG, J., LI, T., LEVINA, E. and ZHU, J. (2017). High-dimensional mixed graphical models. *J. Comput. Graph. Statist.* **26** 367–378. MR3640193

COLLINS, M., DASGUPTA, S. and SCHAPIRE, R. E. (2001). A generalization of principal components analysis to the exponential family. In *NIPS'*01: *Proceedings of the* 14*th International Conference on Neural Information Processing Systems*: *Natural and Synthetic* 617–624. MIT Press, Cambridge, MA.

ELLIS, D. P., WHITMAN, B., BERENZWEIG, A. and LAWRENCE, S. (2002). The quest for ground truth in musical artist similarity. In *ISMIR* 2002 *Conference Proceedings*: *Third International Conference on Music Information Retrieval*: *October* 13–17, 2002, *IRCAM-Centre Pompidou*, *Paris*, *France*.

GOLDSMITH, J., ZIPUNNIKOV, V. and SCHRACK, J. (2015). Generalized multilevel function-on-scalar regression and principal component analysis. *Biometrics* **71** 344–353.

GOLUB, G. H. and VAN LOAN, C. F. (2013). *Matrix Computations*, 4th ed. *Johns Hopkins Studies in the Mathematical Sciences*. Johns Hopkins Univ. Press, Baltimore, MD. MR3024913

GOTO, M. and HIRATA, K. (2004). Recent studies on music information processing. *Acoust. Sci. Technol.* **25** 419–425.

HASTIE, T., TIBSHIRANI, R. and WAINWRIGHT, M. (2015). *Statistical Learning with Sparsity*: *The Lasso and Generalizations*. CRC Press, Boca Raton, FL.

HERLOCKER, J. L., KONSTAN, J. A. and RIEDL, J. (2000). Explaining collaborative filtering recommendations. In *Proceedings of the* 2000 *ACM Conference on Computer Supported Cooperative Work* 241–250. ACM, New York.

HOTELLING, H. (1936). Relations between two sets of variates. *Biometrika* **28** 321–377.

JIA, Y., SALZMANN, M. and DARRELL, T. (2010). Factorized latent spaces with structured sparsity. *Adv. Neural Inf. Process. Syst.* 982–990.

JOHNSON, N. L., KOTZ, S. and BALAKRISHNAN, N. (1997). *Discrete Multivariate Distributions* **165**. Wiley, New York.

KLAMI, A., VIRTANEN, S. and KASKI, S. (2010). Bayesian exponential family projections for coupled data sources. In *The Twenty-Sixth Conference on Uncertainty in Artificial Intelligence* 286–293. AUAI Press.

KLAMI, A., VIRTANEN, S. and KASKI, S. (2013). Bayesian canonical correlation analysis. *J. Mach. Learn. Res.* **14** 965–1003.

LANDGRAF, A. J. and LEE, Y. (2015). Generalized principal component analysis: Projection of saturated model parameters. Technical Report 892, Department of Statistics, Ohio State Univ.

LI, G. and GAYNANOVA, I. (2018). Supplement to "A general framework for association analysis of heterogeneous data." DOI:10.1214/17-AOAS1127SUPP.

LI, Q., CHENG, G., FAN, J. and WANG, Y. (2018). Embracing the blessing of dimensionality in factor models. *J. Amer. Statist. Assoc.* **113** 380–389. MR3803472

LOCK, E. F., HOADLEY, K. A., MARRON, J. S. and NOBEL, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann. Appl. Stat.* **7** 523–542.

LOGAN, B. (2000). Mel frequency cepstral coefficients for music modeling. In *International Symposium on Music Information Retrieval* (*ISMIR*).

LUO, C., LIU, J., DEY, D. K. and CHEN, K. (2016). Canonical variate regression. *Biostatistics* **17** 468–483.

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman & Hall, London. [Second edition of MR0727836.] MR3223057

SHE, Y. (2013). Reduced rank vector generalized linear models for feature extraction. *Stat. Interface* **6** 197–209.

TRYGG, J. and WOLD, S. (2003). O2–PLS, a two-block (X–Y) latent variable regression (LVR) method with an integral OSC filter. *J. Chemom.* **17** 53–64.

TSOUMAKAS, G., SPYROMITROS-XIOUFIS, E., VILCEK, J. and VLAHAVAS, I. (2011). Mulan: A Java library for multi-label learning. *J. Mach. Learn. Res.* **12** 2411–2414.

TUCKER, L. R. (1958). An inter-battery method of factor analysis. *Psychometrika* **23** 111–136. MR0099737

TURNBULL, D., BARRINGTON, L., TORRES, D. and LANCKRIET, G. (2007). Towards musical query-by-semantic-description using the cal500 data set. In *Proceedings of the* 30*th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* 439–446. ACM, New York.

TURNBULL, D., BARRINGTON, L., TORRES, D. and LANCKRIET, G. (2008). Semantic annotation and retrieval of music and sound effects. *IEEE/ACM Trans. Audio Speech Lang. Process.* **16** 467–476.

VIRTANEN, S., KLAMI, A. and KASKI, S. (2011). Bayesian cca via group sparsity. In *Proceedings of the* 28*th International Conference on Machine Learning* (*ICML* 2011) 457–464. ACM, New York.

WESTERHUIS, J. A., KOURTI, T. and MACGREGOR, J. F. (1998). Analysis of multiblock and hierarchical PCA and PLS models. *J. Chemom.* **12** 301–321.

WITTEN, D. M., TIBSHIRANI, R. and HASTIE, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10** 513–534.

YANG, D., MA, Z. and BUJA, A. (2014). A sparse singular value decomposition method for high-dimensional data. *J. Comput. Graph. Statist.* **23** 923–942. MR3270704

YANG, Z., NING, Y. and LIU, H. (2014). On semiparametric exponential family graphical models. ArXiv preprint. Available at arXiv:1412.8697.

ZHOU, G., CICHOCKI, A., ZHANG, Y. and MANDIC, D. P. (2016a). Group component analysis for multiblock data: Common and individual feature extraction. *IEEE Trans. Neural Netw. Learn. Syst.* **27** 2426–2439.

ZHOU, G., ZHAO, Q., ZHANG, Y., ADALI, T., XIE, S. and CICHOCKI, A. (2016b). Linked component analysis from matrices to high-order tensors: Applications to biomedical data. *Proc. IEEE* **104** 310–331.

ZOH, R. S., MALLICK, B., IVANOV, I., BALADANDAYUTHAPANI, V., MANYAM, G., CHAPKIN, R. S., LAMPE, J. W. and CARROLL, R. J. (2016). PCAN: Probabilistic correlation analysis of two non-normal data sets. *Biometrics* **72** 1358–1368. MR3591620

The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1727–1748
https://doi.org/10.1214/17-AOAS1128

CONFIDENT INFERENCE FOR SNP EFFECTS ON TREATMENT EFFICACY

BY YING DING*, YING GRACE LI†, YUSHI LIU†, STEPHEN J. RUBERG† AND JASON C. HSU†,‡

University of Pittsburgh, Eli Lilly and Company† and Ohio State University‡*

Our research is for finding SNPs that are predictive of treatment efficacy, to decide which subgroup (with enhanced treatment efficacy) to target in drug development. Testing SNPs for lack of association with treatment outcome is inherently challenging, because any linkage disequilibrium between a noncausal SNP with a causal SNP, however small, makes the zero-null (no association) hypothesis technically false. Control of Type I error rate in testing such null hypotheses are therefore difficult to interpret. We propose a completely different formulation to address this problem. For each SNP, we provide simultaneous confidence intervals directed toward detecting possible dominant, recessive, or additive effects. Across the SNPs, we control the expected number of SNPs with at least one false confidence interval coverage. Since our confidence intervals are constructed based on pivotal statistics, the false coverage control is guaranteed to be exact and unaffected by the true values of test quantities (whether zero or nonzero). Our method is applicable to the therapeutic areas of Diabetes and Alzheimer's diseases, and perhaps more, as a step toward confidently targeting a patient subgroup in a tailored drug development process.

REFERENCES

BERGER, R. L. and HSU, J. C. (1996). Bioequivalence trials, intersection-union tests, and equivalence confidence sets. *Statist. Sci.* **11** 283–315. MR1445984

DE BAKKER, P., MCVEAN, G., SABETI, P., MIRETTI, M., GREEN, T., MARCHINI, J., KE, X., MONSUUR, A., WHITTAKER, P., DELGADO, M., MORRISON, J., RICHARDSON, A., WALSH, E., GAO, X., GALVER, L., HART, J., HAFLER, D., PERICAK-VANCE, M., TODD, J., DALY, M., TROWSDALE, J., WIJMENGA, C., VYSE, T., BECK, S., MURRAY, S., CARRINGTON, M., GREGORY, S., DELOUKAS, P. and RIOUX, J. (2006). A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat. Genet.* **38** 1166–1172.

EFRON, B. (2007). Correlation and large-scale simultaneous significance testing. *J. Amer. Statist. Assoc.* **102** 93–103.

FDA (2005). Pharmacogenomic data submission: guidance for Industry. Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER), Center for Devices and Radiological Health (CDRH), U.S. Food and Drug Administration, Rockville, MD.

FDA (2008). Guidance for Industry on Diabetes Mellitus: Developing Drugs and Therapeutic Biologics for Treatment and Prevention. Center for Drug Evaluation and Research (CDER), U.S. Food and Drug Administration, Rockville, MD.

Key words and phrases. Multiple testing, simultaneous confidence intervals, SNP, tailored drug development, treatment efficacy.

GENZ, A. and BRETZ, F. (1999). Numerical computation of multivariate t-probabilities with application to power calculation of multiple contrasts. *J. Stat. Comput. Simul.* **63** 361–378.

HOTHORN, L. and HOTHORN, T. (2009). Order-restricted scores test for the evaluation of population-based case-control studies when the genetic model is unknown. *Biom. J.* **51** 659–669.

HSU, J. C. (1996). *Multiple Comparisons: Theory and Methods*. Chapman & Hall, London.

LETTRE, G., LANGE, C. and HIRSCHHORN, J. (2007). Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genet. Epidemiol.* **31** 358–362.

LIPKOVICH, I., DMEITRIENKO, A. and D'AGOSTINO, R. B. (2017). Tutorial in biostatistics: Data-driven subgroup identification and analysis in clinical trials. *Stat. Med.* **36** 136–196.

LOH, W.-Y., HE, X. and MAN, M. (2015). A regression tree approach to identifying subgroups with differential treatment effects. *Stat. Med.* **34** 1818–1833.

MALLAL, S., NOLAN, D., WITT, C., MASEL, G., MARTIN, A. M., MOORE, C., SAYER, D., CASTLEY, A., MAMOTTE, C., MAXWELL, D., JAMES, I. and CHRISTIANSEN, F. T. (2002). Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. *Lancet* **359** 727–732.

MALLAL, S., PHILLIPS, E., CAROSI, G., MOLINA, J.-M., WORKMAN, C., TOMAŽIČ, J., JÄGEL-GUEDES, E., RUGINA, S., KOZYREV, O., CID, J. F., HAY, P., NOLAN, D., HUGHES, S., HUGHES, A., RYAN, S., FITCH, N., THORBORN, D. and BENBOW, A. (2008). HLA-B*5701 screening for hypersensitivity to abacavir. *N. Engl. J. Med.* **358** 568–579.

SO, H.-C. and SHAM, P. C. (2011). Robust association tests under different genetic models, allowing for binary or quantitative traits and covariates. *Behav. Genet.* **41** 768–775.

THE 1000 GENOMES PROJECT CONSORTIUM (2010). A map of human genome variation from population-scale sequencing. *Nature* **467** 1061–1073.

THE 1000 GENOMES PROJECT CONSORTIUM (2012). An integrated map of genetic variation from 1092 human genomes. *Nature* **491** 56–65.

THE 1000 GENOMES PROJECT CONSORTIUM (2015). A global reference for human genetic variation. *Nature* **526** 68–74.

TUKEY, J. W. (1992). Where should multiple comparisons go next? In *Multiple Comparisons, Selection, and Applications in Biometry: A Festschrift in Honor of Charles W. Dunnett* (F. M. Hoppe, ed.) Chapter 12 187–208. Dekker, New York.

The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1749–1772
https://doi.org/10.1214/17-AOAS1129

NONPARAMETRIC BAYESIAN LEARNING OF HETEROGENEOUS DYNAMIC TRANSCRIPTION FACTOR NETWORKS

BY XIANGYU LUO[1] AND YINGYING WEI[2]

The Chinese University of Hong Kong

Gene expression is largely controlled by transcription factors (TFs) in a collaborative manner. Therefore, an understanding of TF collaboration is crucial for the elucidation of gene regulation. The co-activation of TFs can be represented by networks. These networks are dynamic in diverse biological conditions and heterogeneous across the genome within each biological condition. Existing methods for construction of TF networks lack solid statistical models, analyze each biological condition separately, and enforce a single network for all genomic locations within one biological condition, resulting in low statistical power and misleading spurious associations. In this paper, we present a novel Bayesian nonparametric dynamic Poisson graphical model for inference on TF networks. Our approach automatically teases out genome heterogeneity and borrows information across conditions to improve signal detection from very few replicates, thus offering a valid and efficient measure of TF co-activations. We develop an efficient parallel Markov chain Monte Carlo algorithm for posterior computation. The proposed approach is applied to study TF associations in ENCODE cell lines and provides novel findings.

REFERENCES

ALDOUS, D. J. (1985). *Exchangeability and Related Topics*. Springer, New York.

BICKEL, P. J. and LEVINA, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969

BICKEL, P. J., BOLEY, N., BROWN, J. B., HUANG, H. and ZHANG, N. R. (2010). Subsampling methods for genomic inference. *Ann. Appl. Stat.* **4** 1660–1697.

CARTER, S. L., BRECHBÜHLER, C. M., GRIFFIN, M. and BOND, A. T. (2004). Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics* **20** 2242–2250.

CHENG, Y. and LENKOSKI, A. (2012). Hierarchical Gaussian graphical models: Beyond reversible jump. *Electron. J. Stat.* **6** 2309–2331.

CHENG, C., ALEXANDER, R., MIN, R., LENG, J., YIP, K. Y., ROZOWSKY, J., YAN, K.-K., DONG, X., DJEBALI, S., RUAN, Y. et al. (2012). Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res.* **22** 1658–1667.

CHUN, H., ZHANG, X. and ZHAO, H. (2015). Gene regulation network inference with joint sparse Gaussian graphical models. *J. Comput. Graph. Statist.* **24** 954–974.

DANAHER, P., WANG, P. and WITTEN, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 373–397.

Key words and phrases. Poisson graphical model, nonparametric Bayes, parallel Markov chain Monte Carlo, next generation sequencing.

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.

EBERT, P. and BOCK, C. (2015). Improving reference epigenome catalogs by computational prediction. *Nat. Biotechnol.* **33** 354–355.

ENCODE PROJECT CONSORTIUM (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489** 57–74.

ERNST, J. and KELLIS, M. (2012). ChromHMM: Automating chromatin-state discovery and characterization. *Nat. Methods* **9** 215–216.

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

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

GAO, C., ZHU, Y., SHEN, X. and PAN, W. (2016). Estimation of multiple networks in Gaussian mixture models. *Electron. J. Stat.* **10** 1133–1154.

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

GERSTEIN, M. B., KUNDAJE, A., HARIHARAN, M., LANDT, S. G., YAN, K.-K., CHENG, C., MU, X. J., KHURANA, E., ROZOWSKY, J., ALEXANDER, R. et al. (2012). Architecture of the human regulatory network derived from ENCODE data. *Nature* **489** 91–100.

GRANDORI, C., COWLEY, S. M., JAMES, L. P. and EISENMAN, R. N. (2000). The Myc/Max/Mad network and the transcriptional control of cell behavior. *Annu. Rev. Cell Dev. Biol.* **16** 653–699.

GROPP, W., LUSK, E. and SKJELLUM, A. (1999). *Using MPI: Portable Parallel Programming with the Message-Passing Interface*, *Vol.* 1. MIT Press, Cambridge, MA.

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

GUO, J., LEVINA, E., MICHAILIDIS, G. and ZHU, J. (2015). Estimating heterogeneous graphical models for discrete data with an application to roll call voting. *Ann. Appl. Stat.* **9** 821–848. MR3371337

HANLEY, J. A. and MCNEIL, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143** 29–36.

HOBERT, O. (2008). Gene regulation by transcription factors and microRNAs. *Science* **319** 1785–1786.

INOUYE, D. I., YANG, E., ALLEN, G. I. and RAVIKUMAR, P. (2017). A review of multivariate distributions for count data derived from the Poisson distribution. *Wiley Interdiscip. Rev.: Comput. Stat.* **9** e1398, 25. MR3648601

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

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

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

KARLIS, D. (2003). An EM algorithm for multivariate Poisson distribution and related models. *J. Appl. Stat.* **30** 63–77.

KARLIS, D. and MELIGKOTSIDOU, L. (2007). Finite mixtures of multivariate Poisson distributions with application. *J. Statist. Plann. Inference* **137** 1942–1960.

KAWAMURA, K. (1979). The structure of multivariate Poisson distribution. *Kodai Math. J.* **2** 337–345.

KITAMURA, Y., SHIMOHAMA, S., OTA, T., MATSUOKA, Y., NOMURA, Y. and TANIGUCHI, T. (1997). Alteration of transcription factors NF-κB and STAT1 in Alzheimer's disease brains. *Neurosci. Lett.* **237** 17–20.

KOCHERLAKOTA, S. and KOCHERLAKOTA, K. (1992). *Bivariate Discrete Distributions*. Wiley, New York.

LAN, K.-H., KANAI, F., SHIRATORI, Y., OHASHI, M., TANAKA, T., OKUDAIRA, T., YOSHIDA, Y., HAMADA, H. and OMATA, M. (1997). In vivo selective gene expression and therapy mediated by adenoviral vectors for human carcinoembryonic antigen-producing gastric carcinoma. *Cancer Res*. **57** 4279–4284.

LARA-MARQUEZ, M. L., O'DORISIO, M. S., O'DORISIO, T. M., SHAH, M. H. and KARACAY, B. (2001). Selective gene expression and activation-dependent regulation of vasoactive intestinal peptide receptor type 1 and type 2 in human T cells. *J. Immunol*. **166** 2522–2530.

LI, S.-H. and LI, X.-J. (2004). Huntingtin–protein interactions and the pathogenesis of Huntington's disease. *Trends Genet*. **20** 146–154.

LIN, Z., WANG, T., YANG, C. and ZHAO, H. (2017). On joint estimation of Gaussian graphical models for spatial and temporal data. *Biometrics* **73** 769–779.

LOCHAMY, J., ROGERS, E. M. and BOSS, J. M. (2007). CREB and phospho-CREB interact with RFX5 and CIITA to regulate MHC class II genes. *Mol. Immunol*. **44** 837–847.

LUO, X. and WEI, Y. (2018). Supplement to "Nonparametric Bayesian learning of heterogeneous dynamic transcription factor networks." DOI:10.1214/17-AOAS1129SUPP.

MACARTHUR, S., LI, X.-Y., LI, J., BROWN, J. B., CHU, H. C., ZENG, L., GRONDONA, B. P., HECHMER, A., SIMIRENKO, L., KERÄNEN, S. V. et al. (2009). Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biol*. **10** R80.

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

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

MITCHELL, P. J. and TJIAN, R. (1989). Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science* **245** 371–378.

MITRA, R., MÜLLER, P. and JI, Y. (2016). Bayesian graphical models for differential pathways. *Bayesian Anal*. **11** 99–124. MR3447093

MITRA, R., MÜLLER, P., LIANG, S., YUE, L. and JI, Y. (2013). A Bayesian graphical model for chip-seq data on histone modifications. *J. Amer. Statist. Assoc*. **108** 69–80.

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.

OGATA, H., GOTO, S., SATO, K., FUJIBUCHI, W., BONO, H. and KANEHISA, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. **27** 29–34.

PETERSON, C. B., STINGO, F. C. and VANNUCCI, M. (2015). Bayesian inference of multiple Gaussian graphical models. *J. Amer. Statist. Assoc*. **110** 159–174.

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

RODRIGUEZ, A., LENKOSKI, A., DOBRA, A. et al. (2011). Sparse covariance estimation in heterogeneous samples. *Electron. J. Stat*. **5** 981–1014.

SCHERZER, C. R., GRASS, J. A., LIAO, Z., PEPIVANI, I., ZHENG, B., EKLUND, A. C., NEY, P. A., NG, J., MCGOLDRICK, M., MOLLENHAUER, B. et al. (2008). GATA transcription factors directly regulate the Parkinson's disease-linked gene α-synuclein. *Proc. Natl. Acad. Sci. USA* **105** 10907–10912.

SHI, Q., LE, X., ABBRUZZESE, J. L., WANG, B., MUJAIDA, N., MATSUSHIMA, K., HUANG, S., XIONG, Q. and XIE, K. (1999). Cooperation between transcription factor AP-1 and NF-κB in the induction of interleukin-8 in human pancreatic adenocarcinoma cells by hypoxia. *J. Interferon Cytokine Res*. **19** 1363–1371.

SUBRAMANIAN, A., TAMAYO, P., MOOTHA, V. K., MUKHERJEE, S., EBERT, B. L., GILLETTE, M. A., PAULOVICH, A., POMEROY, S. L., GOLUB, T. R., LANDER, E. S. et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102** 15545–15550.

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82** 528–540.

WEI, Y. and WU, H. (2016). Measuring the spatial correlations of protein binding sites. *Bioinformatics* **32** 1766–1772.

XING, E. P., SOHN, K.-A. et al. (2007). Hidden Markov Dirichlet process: Modeling genetic inference in open ancestral space. *Bayesian Anal.* **2** 501–527.

XUE, W., KANG, J., BOWMAN, F. D., WAGER, T. D. and GUO, J. (2014). Identifying functional co-activation patterns in neuroimaging studies via Poisson graphical models. *Biometrics* **70** 812–822.

YANG, E., RAVIKUMAR, P. K., ALLEN, G. I. and LIU, Z. (2013). On Poisson graphical models. In *Advances in Neural Information Processing Systems* 1718–1726.

YANG, E., RAVIKUMAR, P., ALLEN, G. I. and LIU, Z. (2015). Graphical models via univariate exponential family distributions. *J. Mach. Learn. Res.* **16** 3813–3847.

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

ZERVOS, A. S., GYURIS, J. and BRENT, R. (1993). Mxi1, a protein that specifically interacts with Max to bind Myc-Max recognition sites. *Cell* **72** 223–232.

ZHANG, B. and HORVATH, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4** Article17.

ZHOU, H., CHERUVANKY, A., HU, X., MATSUMOTO, T., HIRAMATSU, N., CHO, M. E., BERGER, A., LEELAHAVANICHKUL, A., DOI, K., CHAWLA, L. S. et al. (2008). Urinary exosomal transcription factors, a new class of biomarkers for renal disease. *Kidney Int.* **74** 613–621.

The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1773–1795
https://doi.org/10.1214/17-AOAS1130

ESTIMATING AND COMPARING CANCER PROGRESSION RISKS UNDER VARYING SURVEILLANCE PROTOCOLS[1]

BY JANE M. LANGE[*,1,2], ROMAN GULATI[*,1,2], AMY S. LEONARDSON[*,1,3], DANIEL W. LIN[†,1,4], LISA F. NEWCOMB[*,1,4], BRUCE J. TROCK[‡,1], H. BALLENTINE CARTER[‡,1], PETER R. CARROLL[§,1,5], MATTHEW R. COOPERBERG[§,1,5], JANET E. COWAN[§,1,5], LAWRENCE H. KLOTZ[¶,6] AND RUTH ETZIONI[*,†,1,2]

Fred Hutchinson Cancer Research Center[*], University of Washington[†], Johns Hopkins University[‡], University of California, San Francisco[§] and University of Toronto[¶]

Outcomes after cancer diagnosis and treatment are often observed at discrete times via doctor-patient encounters or specialized diagnostic examinations. Despite their ubiquity as endpoints in cancer studies, such outcomes pose challenges for analysis. In particular, comparisons between studies or patient populations with different surveillance schema may be confounded by differences in visit frequencies. We present a statistical framework based on multistate and hidden Markov models that represents events on a continuous time scale given data with discrete observation times. To demonstrate this framework, we consider the problem of comparing risks of prostate cancer progression across multiple active surveillance cohorts with different surveillance frequencies. We show that the different surveillance schedules partially explain observed differences in the progression risks between cohorts. Our application permits the conclusion that differences in underlying cancer progression risks across cohorts persist after accounting for different surveillance frequencies.

REFERENCES

ANDERSEN, P. K. and KEIDING, N. (2002). Multi-state models for event history analysis. *Stat. Methods Med. Res.* **11** 91–115.

ARALIS, H. J. (2016). Modeling multistate processes with back transitions: Statistical challenges and applications. Ph.D. thesis, UCLA.

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

BLADT, M. and SORENSEN, M. (2005). Statistical inference for discretely observed Markov jump processes. *J. R. Stat. Soc., Ser. B Stat. Methodol.* **67** 395–410.

COLEY, R. Y., ZEGER, S. L., MAMAWALA, M., PIENTA, K. J. and CARTER, H. B. (2016). Prediction of the pathologic Gleason score to inform a personalized management program for prostate cancer. *Eur. Urol.* **72** 135–141.

Key words and phrases. Hidden Markov model, multistate model, panel data, prostate cancer, active surveillance.

CUMANI, A. (1982). On the canonical representation of homogeneous Markov processes modelling failure-time distributions. *Microelectron. Reliab.* **22** 583–602.

DALL'ERA, M. A. (2015). Patient and disease factors affecting the choice and adherence to active surveillance. *Curr. Opin. Neurol.* **25** 272–276.

DONNELLY, C., MCFETRIDGE, L. M., MARSHALL, A. H. and MITCHELL, H. J. (2017). A two-stage approach to the joint analysis of longitudinal and survival data utilising the Coxian phase-type distribution. *Stat. Methods Med. Res.* To appear. PMID: 28633604.

FOUCHER, Y., GIRAL, M., SOULILLOU, J.-P. and DAURES, J.-P. (2007). A semi-Markov model for multistate and interval-censored data with multiple terminal events. Application in renal transplantation. *Stat. Med.* **26** 5381–5393.

FRYDMAN, H. and SZAREK, M. (2009). Nonparametric estimation in a Markov "illness–death" process from interval censored observations with missing intermediate transition status. *Biometrics* **65** 143–151.

GIGNAC, G. A., MORRIS, M. J., HELLER, G., SCHWARTZ, L. H. and SCHER, H. I. (2008). Assessing outcomes in prostate cancer clinical trials: A twenty-first century tower of Babel. *Cancer* **113** 966–974.

GILBERT, P. and VARADHAN, R. (2012). numDeriv: Accurate numerical derivatives. R package version 2012.9-1.

GRÜGER, J., KAY, R. and SCHUMACHER, M. (1991). The validity of inferences based on incomplete observations in disease state models. *Biometrics* **47** 595–605.

HUANG, X. and WOLFE, R. A. (2002). A frailty model for informative censoring. *Biometrics* **58** 510–520.

HUDGENS, M. G., SATTEN, G. A. and LONGINI, I. M. (2001). Nonparametric maximum likelihood estimation for competing risks survival data subject to interval censoring and truncation. *Biometrics* **57** 74–80.

HUMPHREY, P. A. (2004). Gleason grading and prognostic factors in carcinoma of the prostate. *Mod. Pathol.* **17** 292–306.

INOUE, L. Y. T., TROCK, B. J., PARTIN, A. W., CARTER, H. B. and ETZIONI, R. (2014). Modeling grade progression in an active surveillance study. *Stat. Med.* **33** 930–939.

JACKSON, C. H., SHARPLES, L. D., THOMPSON, S. G. and DUFFY, S. W. (2003). Multistate Markov models for disease progression with classification error. *J. R. Stat. Soc., Ser. D Stat.* **52** 193–209.

KANG, M. and LAGAKOS, S. W. (2007). Statistical methods for panel data from a semi-Markov process, with application to HPV. *Biostatistics* **8** 252–264.

KLOTZ, L., VESPRINI, D., SETHUKAVALAN, P., JETHAVA, V., ZHANG, L., JAIN, S., YAMAMOTO, T., MAMEDOV, A. and LOBLAW, A. (2015). Long-term follow-up of a large active surveillance cohort of patients with prostate cancer. *J. Clin. Oncol.* **33** 272–277.

LANGE, J. M. and MININ, V. N. (2013). Fitting and interpreting continuous-time latent Markov models for panel data. *Stat. Med.* **32** 4581–4595.

LANGE, J. M., HUBBARD, R. A., INOUE, L. Y. T. and MININ, V. N. (2015). A joint model for multistate disease processes and random informative observation times, with applications to electronic medical records data. *Biometrics* **71** 90–101. MR3335353

LANGE, J. M., GULATI, R., LEONARDSON, A. S., LIN, D. W., NEWCOMB, L. F., TROCK, B. J., CARTER, H. B., COOPERBERG, M. R., COWAN, J. E., KLOTZ, L. H. and ETZIONI, R. (2018). Supplement to "Estimating and comparing cancer progression risks under varying surveillance protocols." DOI:10.1214/17-AOAS1130SUPPA, DOI:10.1214/17-AOAS1130SUPPB, DOI:10.1214/17-AOAS1130SUPPC, DOI:10.1214/17-AOAS1130SUPPD, DOI:10.1214/17-AOAS1130SUPPE.

MANDEL, M. (2010). Estimating disease progression using panel data. *Biostatistics* **11** 304–316.

MAO, L., LIN, D.-Y. and ZENG, D. (2017). Semiparametric regression analysis of interval-censored competing risks data. *Biometrics* **73** 857–865.

MARSHALL, G. and JONES, R. H. (1995). Multi-state models and diabetic retinopathy. *Stat. Med.* **14** 1975–1983.

MOLER, C. and LOAN, C. V. (2003). Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Rev.* **45** 801–836.

NAROD, S. A. and RAKOVITCH, E. (2014). A comparison of the risks of in-breast recurrence after a diagnosis of DCIS or early invasive breast cancer. *Curr. Oncol.* **21** 119–124.

NEWCOMB, L. F., THOMPSON JR., I. M., BOYER, H. D., BROOKS, J. D., CARROLL, P. R., COOPERBERG, M. R., DASH, A., ELLIS, W. J., FAZLI, L., FENG, Z., MARTIN, E., KUNJU, P., LANCE, R. S., MCKENNEY, J. K., MENG, M. V., MARLO, M., SANDA, M. G., SIMKO, J., SO, A., TRETIAKOVA, M. S., TROYER, D. A., TRUE, L. D., VAKAR-LOPEZ, F., VIRGIN, J., WAGNER, A. A., WEI, J. T., NELSON, P. S. and LIN, D. W. FOR THE CANARY PROSTATE ACTIVE SURVEILLANCE STUDY INVESTIGATORS (2016). Outcomes of active surveillance for the management of clinically localized prostate cancer in the prospective, multi-institutional Canary PASS cohort. *J. Urol.* **195** 206–221.

PALISAAR, J. R., NOLDUS, J., LÖPPENBERG, B., VON BODMAN, C., SOMMERER, F. and EGGERT, T. (2012). Comprehensive report on prostate cancer misclassification by 16 currently used low-risk and active surveillance criteria. *BJU Int.* **110**.

PENSON, D. F. (2012). Factors influencing patients' acceptance and adherence to active surveillance. *J. Natl. Cancer Inst. Monogr.* **45** 207–212.

PINSKY, P., PARNES, H. and FORD, L. (2008). Estimating rates of true high-grade disease in the Prostate Cancer Prevention Trial. *Cancer Prev. Res.* **1** 182–186.

POPIOLEK, M., RIDER, J. R., ANDRÉN, O., ANDERSSON, S.-O., HOLMBERG, L., ADAMI, H.-O. and JOHANSSON, J.-E. (2013). Natural history of early, localized prostate cancer: A final report from three decades of follow-up. *Eur. Urol.* **63** 428–435.

ROSS, A. E., LOEB, S., LANDIS, P., PARTIN, A. W., EPSTEIN, J. I., KETTERMANN, A., FENG, Z., CARTER, H. B. and WALSH, P. C. (2010). Prostate-specific antigen kinetics during follow-up are an unreliable trigger for intervention in a prostate cancer surveillance program. *J. Clin. Oncol.* **28** 2810–2816.

ROUANET, A., JOLY, P., DARTIGUES, J.-F., PROUST-LIMA, C. and JACQMIN-GADDA, H. (2016). Joint latent class model for longitudinal data and interval-censored semi-competing events: Application to dementia. *Biometrics* **72** 1123–1135.

SRIDHARA, R., MANDREKAR, S. J. and DODD, L. E. (2013). Missing data and measurement variability in assessing progression-free survival endpoint in randomized clinical trials. *Clin. Cancer Res.* **19** 2613–2620.

STEELE, R. J. and RAFTERY, A. E. (2010). Performance of Bayesian model selection criteria for Gaussian mixture models. In *Frontiers of Statistical Decision Making and Bayesian Analysis* (M.-H. Chen, P. Muller, D. Sun, K. Ye and D. K. Dey, eds.) 113–130. Springer, New York.

STEPHENSON, A. J., KATTAN, M. W., EASTHAM, J. A., DOTAN, Z. A., BIANCO, F. J., LILJA, H. and SCARDINO, P. T. (2006). Defining biochemical recurrence of prostate cancer after radical prostatectomy: A proposal for a standardized definition. *J. Clin. Oncol.* **24** 3973–3978.

TITMAN, A. C. and SHARPLES, L. D. (2010). Semi-Markov models with phase-type sojourn distributions. *Biometrics* **66** 742–752.

TOSOIAN, J. J., MAMAWALA, M., EPSTEIN, J. I., LANDIS, P., WOLF, S., TROCK, B. J. and CARTER, H. B. (2015). Intermediate and longer-term outcomes from a prospective active-surveillance program for favorable-risk prostate cancer. *J. Clin. Oncol.* **33** 3379–3385.

TOSOIAN, J. J., CARTER, H. B., LEPOR, A. and LOEB, S. (2016). Active surveillance for prostate cancer: Contemporary state of practice. *Nat. Rev. Urol.* **116** 1477–1490.

WELTY, C. J., COWAN, J. E., NGUYEN, H., SHINOHARA, K., PEREZ, N., GREENE, K. L., CHAN, J. M., MENG, M. V., SIMKO, J. P., COOPERBERG, M. R. and CARROLL, P. R. (2015). Extended followup and risk factors for disease reclassification in a large active surveillance cohort for localized prostate cancer. *J. Urol.* **193** 807–811.

ZENG, L., COOK, R. J., WEN, L. and BORUVKA, A. (2015). Bias in progression-free survival analysis due to intermittent assessment of progression. *Stat. Med.* **34** 3181–3193.

The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1796–1830
https://doi.org/10.1214/17-AOAS1132

ANALYSING PLANT CLOSURE EFFECTS USING TIME-VARYING MIXTURE-OF-EXPERTS MARKOV CHAIN CLUSTERING[1]

BY SYLVIA FRÜHWIRTH-SCHNATTER[*], STEFAN PITTNER[*], ANDREA WEBER[*,†,‡] AND RUDOLF WINTER-EBMER[§,¶]

Vienna University of Economics and Business[*], CEU Budapest[†], WIFO[‡], Johannes Kepler Universität Linz[§] and IHS[¶]

In this paper we study data on discrete labor market transitions from Austria. In particular, we follow the careers of workers who experience a job displacement due to plant closure and observe—over a period of 40 quarters—whether these workers manage to return to a steady career path. To analyse these discrete-valued panel data, we apply a new method of Bayesian Markov chain clustering analysis based on inhomogeneous first order Markov transition processes with time-varying transition matrices. In addition, a mixture-of-experts approach allows us to model the probability of belonging to a certain cluster as depending on a set of covariates via a multinomial logit model. Our cluster analysis identifies five career patterns after plant closure and reveals that some workers cope quite easily with a job loss whereas others suffer large losses over extended periods of time.

REFERENCES

AITKIN, M. and ALFÓ, M. (1998). Regression models for binary longitudinal responses. *Stat. Comput.* **8** 289–307.

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

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

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

CADEZ, I., HECKERMAN, D., MEEK, C., SMYTH, P. and WHITE, S. (2003). Model-based clustering and visualization of navigation patterns on a web site. *Data Min. Knowl. Discov.* **7**(4) 399–424. MR2011154

COUCH, K., JOLLEY, N. and PLACZEK, D. (2010). Earnings impact of job displacement revisited. *Am. Econ. Rev.* **100** 572–589.

DEL BONO, E., WEBER, A. and WINTER-EBMER, R. (2012). Clash of career and family: Fertility decisions after job displacement. *J. Eur. Econ. Assoc.* **10** 659–683.

DIAS, J. G. and VERMUNT, J. K. (2007). Latent class modeling of website users' search patterns: Implications for online market segmentation. *J. Retail. Consum. Serv.* **14** 359–368.

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

Key words and phrases. Transition data, Markov chain Monte Carlo, multinomial logit, panel data, inhomogeneous Markov chains.

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

DRAPER, N. R. and SMITH, H. (1998). *Applied Regression Analysis*, 3rd ed. *Wiley Series in Probability and Statistics*: *Texts and References Section*. Wiley, New York. With 1 IBM-PC floppy disk (3.5 inch; DD). MR1614335

ELIASON, M. and STORRIE, D. (2004). The echo of job displacement. ISER Working Paper 20.

FALLICK, B. (1996). A review of the recent empirical literature on displaced workers. *Ind. Labor Relat. Rev.* **50** 5–16.

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

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

FRÜHWIRTH-SCHNATTER, S. (2011). Panel data analysis: A survey on model-based clustering of time series. *Adv. Data Anal. Classif.* **5** 251–280. MR2860101

FRÜHWIRTH-SCHNATTER, S. and FRÜHWIRTH, R. (2010). Data augmentation and MCMC for binary and multinomial logit models. In *Statistical Modelling and Regression Structures* (T. Kneib and G. Tutz, eds.) 111–132. Physica-Verlag/Springer, Heidelberg. Also available at http://www.ifas.jku.at/ifas/content/e114480, IFAS Research Paper Series 2010-48. MR2664631

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

FRÜHWIRTH-SCHNATTER, S., PAMMINGER, C., WEBER, A. and WINTER-EBMER, R. (2012). Labor market entry and earnings dynamics: Bayesian inference using mixtures-of-experts Markov chain clustering. *J. Appl. Econometrics* **27** 1116–1137. MR3041877

FRÜHWIRTH-SCHNATTER, S., PAMMINGER, C., WEBER, A. and WINTER-EBMER, R. (2016). Mothers' long-run career patterns after first birth. *J. Roy. Statist. Soc. Ser. A* **179** 707–725. MR3501429

FRYDMAN, H. (2005). Estimation in the mixture of Markov chains moving with different speeds. *J. Amer. Statist. Assoc.* **100** 1046–1053. MR2201031

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

GOLLINI, I. and MURPHY, T. B. (2014). Mixture of latent trait analyzers for model-based clustering of categorical data. *Stat. Comput.* **24** 569–588. MR3223542

GOODMAN, L. A. (1961). Statistical methods for the mover-stayer model. *J. Amer. Statist. Assoc.* **56** 841–868. MR0136436

GORMLEY, I. C. and MURPHY, T. B. (2008). A mixture of experts model for rank data with applications in election studies. *Ann. Appl. Stat.* **2** 1452–1477. MR2655667

HECKMAN, J. (1981). The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process. In *Structural Analysis of Discrete Data with Econometric Applications* (C. F. Manski and D. McFadden, eds.) 179–195. MIT Press, Cambridge, MA.

HUTTUNEN, K., MOEN, J. and SALVANES, K. G. (2011). How destructive is creative destruction? Effects of job loss on job mobility, withdrawal and income. *J. Eur. Econ. Assoc.* **9** 840–870.

ICHINO, A., SCHWERDT, G., WINTER-EBMER, R. and ZWEIMÜLLER, J. (2017). Too old to work, too young to retire? *J. Econ. Ageing* **9** 14–29.

IMBENS, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev. Econ. Stat.* **86** 4–29.

JACOBSON, L. S., LALONDE, R. J. and SULLIVAN, D. G. (1993). Earnings losses of displaced workers. *Am. Econ. Rev.* **83** 685–709.

JACOBSON, L. S., LALONDE, R. J. and SULLIVAN, D. G. (2005). Estimating the returns to community college schooling for displaced workers. *J. Econometrics* **125** 271–304. MR2143378

JUÁREZ, M. A. and STEEL, M. F. J. (2010). Model-based clustering of non-Gaussian panel data based on skew-*t* distributions. *J. Bus. Econom. Statist.* **28** 52–66. MR2650600

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

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

PAMMINGER, C. and FRÜHWIRTH-SCHNATTER, S. (2010). Model-based clustering of categorical time series. *Bayesian Anal.* **5** 345–368. MR2719656

PAMMINGER, C. and TÜCHLER, R. (2011). A Bayesian analysis of female wage dynamics using Markov chain clustering. *Austr. J. Stat.* **40** 281–296.

PENG, F., JACOBS, R. A. and TANNER, M. A. (1996). Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. *J. Amer. Statist. Assoc.* **91** 953–960.

RAMONI, M., SEBASTIANI, P. and COHEN, P. (2002). Bayesian clustering by dynamics. *Mach. Learn.* **47** 91–121.

RUHM, C. J. (1991). Are workers permanently scarred by job displacements? *Am. Econ. Rev.* **81** 319–324.

SCHWERDT, G., ICHINO, A., RUF, O., WINTER-EBMER, R. and ZWEIMÜLLER, J. (2010). Does the color of the collar matter? Employment and earnings after plant closure. *Econom. Lett.* **108** 137–140.

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

SKRONDAL, A. and RABE-HESKETH, S. (2014). Handling initial conditions and endogenous covariates in dynamic/transition models for binary data with unobserved heterogeneity. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **63** 211–237. MR3234341

SULLIVAN, D. and VON WACHTER, T. (2009). Job displacement and mortality: An analysis using administrative data. *Q. J. Econ.* **124** 1265–1306.

WINTER-EBMER, R. (2016). Long-term effects of unemployment: What can we learn from plant-closure studies? In *Long-Term Unemployment After the Great Recession* (S. Bentolila and M. Jansen, eds.) 33–42. CEPR Press, London.

WOOLDRIDGE, J. M. (2005). Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *J. Appl. Econometrics* **20** 39–54. MR2138202

ZWEIMÜLLER, J., WINTER-EBMER, R., LALIVE, R., KUHN, A., WUELLRICH, J.-P., RUF, O. and BÜCHI, S. (2009). The Austrian Social Security Database (ASSD). Technical report.

The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1831–1852
https://doi.org/10.1214/17-AOAS1133

USING MISSING TYPES TO IMPROVE PARTIAL IDENTIFICATION WITH APPLICATION TO A STUDY OF HIV PREVALENCE IN MALAWI

BY ZHICHAO JIANG AND PENG DING[1]

Princeton University and University of California, Berkeley

Frequently, empirical studies are plagued with missing data. When the data are missing not at random, the parameter of interest is not identifiable in general. Without additional assumptions, we can derive bounds of the parameters of interest, which, unfortunately, are often too wide to be informative. Therefore, it is of great importance to sharpen these worst-case bounds by exploiting additional information. Traditional missing data analysis uses only the information of the binary missing data indicator, that is, a certain data point is either missing or not. Nevertheless, real data often provide more information than a binary missing data indicator, and they often record different types of missingness. In a motivating HIV status survey, missing data may be due to the units' unwillingness to respond to the survey items or their hospitalization during the visit, and may also be due to the units' temporarily absence or relocation. It is apparent that some missing types are more likely to be missing not at random, but other missing types are more likely to be missing at random. We show that making full use of the missing types results in narrower bounds of the parameters of interest. In a real-life example, we demonstrate substantial improvement of more than 50% reduction in bound widths for estimating the prevalence of HIV in rural Malawi. As we illustrate using the HIV study, our strategy is also useful for conducting sensitivity analysis by gradually increasing or decreasing the set of types that are missing at random. In addition, we propose an easy-to-implement method to construct confidence intervals for partially identified parameters with bounds expressed as the minimums and maximums of finite parameters, which is useful for not only our problem but also many other problems involving bounds.

REFERENCES

ANDREWS, D. W. K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica* **68** 399–405. MR1748009

ANGLEWICZ, P., ADAMS, J., OBARE, F., KOHLER, H.-P. and WATKINS, S. (2009). The Malawi Diffusion and Ideational Change Project 2004–06: Data collection, data quality, and analysis of attrition. *Demogr. Res*. **20** 503–540.

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.

ARPINO, B., DE CAO, E. and PERACCHI, F. (2014). Using panel data for partial identification of human immunodeficiency virus prevalence when infection status is missing not at random. *J. Roy. Statist. Soc. Ser. A* **177** 587–606. MR3256525

Key words and phrases. Longitudinal data, partial identification, sensitivity analysis, sharp bound, testable condition.

BALKE, A. and PEARL, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *J. Amer. Statist. Assoc.* **92** 1171–1176.
BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–972. MR2216189
CHENG, J. and SMALL, D. S. (2006). Bounds on causal effects in three-arm trials with noncompliance. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 815–836. MR2301296
CHERNOZHUKOV, V., LEE, S. and ROSEN, A. M. (2013). Intersection bounds: Estimation and inference. *Econometrica* **81** 667–737. MR3043345
COCHRAN, W. G. (1953). *Sampling Techniques*. Wiley, New York. MR0054199
COPAS, J. B. and LI, H. G. (1997). Inference for non-random samples. *J. Roy. Statist. Soc. Ser. B* **59** 55–95. MR1436555
DING, P. and GENG, Z. (2014). Identifiability of subgroup causal effects in randomized experiments with nonignorable missing covariates. *Stat. Med.* **33** 1121–1133. MR3247784
HAREL, O. and SCHAFER, J. L. (2009). Partial and latent ignorability in missing-data problems. *Biometrika* **96** 37–50. MR2482133
HOROWITZ, J. L. and MANSKI, C. F. (1998). Censoring of outcomes and regressors due to survey nonresponse: Identification and estimation using weights and imputations. *J. Econometrics* **84** 37–58. MR1621936
HOROWITZ, J. L. and MANSKI, C. F. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *J. Amer. Statist. Assoc.* **95** 77–88. MR1803142
IMBENS, G. W. and MANSKI, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica* **72** 1845–1857. MR2095534
JIANG, Z. and DING, P. (2018). Supplement to "Using missing types to improve partial identification with application to a study of HIV prevalence in Malawi." DOI:10.1214/17-AOAS1133SUPP.
JIANG, Z., DING, P. and GENG, Z. (2016). Principal causal effect identification and surrogate end point evaluation by multiple trials. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 829–848. MR3534352
JIN, H. and RUBIN, D. B. (2008). Principal stratification for causal inference with extended partial compliance. *J. Amer. Statist. Assoc.* **103** 101–111. MR2463484
KANG, J. D. Y. and SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22** 523–539. MR2420458
KITAGAWA, T. (2015). A test for instrument validity. *Econometrica* **83** 2043–2063. MR3414199
LEE, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Rev. Econ. Stud.* **76** 1071–1102.
LITTLE, R. J. (1993). Pattern-mixture models for multivariate incomplete data. *J. Amer. Statist. Assoc.* **88** 125–134.
LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley-Interscience, Hoboken, NJ. MR1925014
LITTLE, R. J., RUBIN, D. B. and ZANGENEH, S. Z. (2017). Conditions for ignoring the missing-data mechanism in likelihood inferences for parameter subsets. *J. Amer. Statist. Assoc.* **112** 314–320. MR3646573
LONG, D. M. and HUDGENS, M. G. (2013). Sharpening bounds on principal effects with covariates. *Biometrics* **69** 812–819. MR3146777
MA, W.-Q., GENG, Z. and HU, Y.-H. (2003). Identification of graphical models for nonignorable nonresponse of binary outcomes in longitudinal studies. *J. Multivariate Anal.* **87** 24–45. MR2007260
MANSKI, C. F. (2003). *Partial Identification of Probability Distributions*. Springer, New York. MR2151380
MANSKI, C. F. (2009). *Identification for Prediction and Decision*. Harvard Univ. Press, Cambridge.
MANSKI, C. F. and PEPPER, J. V. (2000). Monotone instrumental variables: With an application to the returns to schooling. *Econometrica* **68** 997–1010. MR1771587

MATTEI, A., MEALLI, F. and PACINI, B. (2014). Identification of causal effects in the presence of nonignorable missing outcome values. *Biometrics* **70** 278–288. MR3258033

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

MEALLI, F. and RUBIN, D. B. (2015). Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika* **102** 995–1000. MR3431570

MIAO, W., DING, P. and GENG, Z. (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *J. Amer. Statist. Assoc.* **111** 1673–1683. MR3601726

MOLENBERGHS, G., KENWARD, M. G. and GOETGHEBEUR, E. (2001). Sensitivity analysis for incomplete contingency tables: The Slovenian plebiscite case. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **50** 15–29.

ROMANO, J. P. and SHAIKH, A. M. (2010). Inference for the identified set in partially identified econometric models. *Econometrica* **78** 169–211. MR2642860

ROTNITZKY, A., SCHARFSTEIN, D., SU, T.-L. and ROBINS, J. (2001). Methods for conducting sensitivity analysis of trials with potentially nonignorable competing causes of censoring. *Biometrics* **57** 103–113. MR1833295

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. With comments by R. J. A. Little and a reply by the author. MR0455196

RUBIN, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys*. Wiley-Interscience, Hoboken, NJ. Reprint of the 1987 edition. MR2117498

RUBIN, D. B. (2005). Comment on "Multiple-bias modelling for analysis of observational data" by S. Greenland. *J. Roy. Statist. Soc. Ser. A* **168** 302.

SCHARFSTEIN, D. O., MANSKI, C. F. and ANTHONY, J. C. (2004). On the construction of bounds in prospective studies with missing ordinal outcomes: Application to the good behavior game trial. *Biometrics* **60** 154–164. MR2044111

SCHARFSTEIN, D. O., ROTNITZKY, A. and ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Amer. Statist. Assoc.* **94** 1096–1146. MR1731478

SEAMAN, S., GALATI, J., JACKSON, D. and CARLIN, J. (2013). What is meant by "missing at random"? *Statist. Sci.* **28** 257–268. MR3112409

SHAO, J. and WANG, L. (2016). Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika* **103** 175–187. MR3465829

TANG, G., LITTLE, R. J. A. and RAGHUNATHAN, T. E. (2003). Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika* **90** 747–764. MR2024755

VANSTEELANDT, S., GOETGHEBEUR, E., KENWARD, M. G. and MOLENBERGHS, G. (2006). Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statist. Sinica* **16** 953–979. MR2281311

YANG, S. and KIM, J. K. (2016). A note on multiple imputation for method of moments estimation. *Biometrika* **103** 244–251. MR3465836

YANG, F. and SMALL, D. S. (2016). Using post-outcome measurement information in censoring-by-death problems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 299–318. MR3453657

The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1853–1870
https://doi.org/10.1214/18-AOAS1134

A COUPLED ETAS-I^{2}GMM POINT PROCESS WITH APPLICATIONS TO SEISMIC FAULT DETECTION[1]

BY YICHENG CHENG, MURAT DUNDAR AND GEORGE MOHLER

Indiana University–Purdue University Indianapolis

Epidemic-type aftershock sequence (ETAS) point process is a common model for the occurrence of earthquake events. The ETAS model consists of a stationary background Poisson process modeling spontaneous earthquakes and a triggering kernel representing the space–time-magnitude distribution of aftershocks. Popular nonparametric methods for estimation of the background intensity include histograms and kernel density estimators. While these methods are able to capture local spatial heterogeneity in the intensity of spontaneous events, they do not capture well patterns resulting from fault line structure over larger spatial scales. Here we propose a two-layer infinite Gaussian mixture model for clustering of earthquake events into fault-like groups over intermediate spatial scales. We introduce a Monte Carlo expectation-maximization (EM) algorithm for joint inference of the ETAS-I^{2}GMM model and then apply the model to the Southern California Earthquake Catalog. We illustrate the advantages of the ETAS-I^{2}GMM model in terms of both goodness of fit of the intensity and recovery of fault line clusters in the Community Fault Model 3.0 from earthquake occurrence data.

REFERENCES

ADELFIO, G. and CHIODI, M. (2015). Alternated estimation in semi-parametric space–time branching-type point processes with application to seismic catalogs. *Stoch. Environ. Res. Risk Assess.* **29** 443–450.

ANDREWS, J. L. and MCNICHOLAS, P. D. (2012). Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions: The tEIGEN family. *Stat. Comput.* **22** 1021–1029. MR2950082

ARCHAMBEAU, C. and VERLEYSEN, M. (2007). Robust Bayesian clustering. *Neural Netw.* **20** 129–138.

BAUDRY, J.-P., RAFTERY, A. E., CELEUX, G., LO, K. and GOTTARDO, R. (2010). Combining mixture components for clustering. *J. Comput. Graph. Statist.* **19** 332–353. MR2758307

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

FIGUEIREDO, M. A. and JAIN, A. K. (2002). Unsupervised learning of finite mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.* **24** 381–396.

FORBES, F. and WRAITH, D. (2014). A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweight: Application to robust clustering. *Stat. Comput.* **24** 971–984. MR3253848

GARDNER, J. K. and KNOPOFF, L. (1974). Is the sequence of earthquakes in Southern California, with aftershocks removed, Poissonian? *Bull. Seismol. Soc. Am.* **64** 1363–1367.

Key words and phrases. Infinite Gaussian mixture model, epidemic-type aftershock sequence, point process.

GE, Y. and SEALFON, S. C. (2012). FlowPeaks: A fast unsupervised clustering for flow cytometry data via K-means and density peak finding. *Bioinformatics* **28** 2052–2058.
HENNIG, C. (2010). Methods for merging Gaussian mixture components. *Adv. Data Anal. Classif.* **4** 3–34. MR2639661
KUHN, H. W. (1955). The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **2** 83–97. MR0075510
LAI, E., MOYER, D., YUAN, B., FOX, E., HUNTER, B., BERTOZZI, A. L. and BRANTINGHAM, J. (2014). Topic time series analysis of microblogs. Technical report, DTIC Document.
LEE, S. and MCLACHLAN, G. J. (2014). Finite mixtures of multivariate skew t-distributions: Some recent and new results. *Stat. Comput.* **24** 181–202. MR3165547
LEWIS, E. and MOHLER, G. (2011). A nonparametric EM algorithm for multiscale Hawkes processes. Preprint available at http://paleo.sscnet.ucla.edu/Lewis-Molher-EM_Preprint.pdf.
MARSAN, D. and LENGLINE, O. (2008). Extending earthquakes' reach through cascading. *Science* **319** 1076–1079.
MOHLER, G. (2013). Modeling and estimation of multi-source clustering in crime and security data. *Ann. Appl. Stat.* **7** 1525–1539. MR3127957
MOHLER, G. (2014). Marked point process hotspot maps for homicide and gun crime prediction in Chicago. *Int. J. Forecast.* **30** 491–497.
MOHLER, G. O., SHORT, M. B., BRANTINGHAM, P. J., SCHOENBERG, F. P. and TITA, G. E. (2011). Self-exciting point process modeling of crime. *J. Amer. Statist. Assoc.* **106** 100–108. MR2816705
MOHLER, G. O., SHORT, M. B., MALINOWSKI, S., JOHNSON, M., TITA, G. E., BERTOZZI, A. L. and BRANTINGHAM, P. J. (2015). Randomized controlled field trials of predictive policing. *J. Amer. Statist. Assoc.* **110** 1399–1411. MR3449035
OGATA, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *J. Amer. Statist. Assoc.* **83** 9–27.
OGATA, Y. (1998). Space–time point-process models for earthquake occurrences. *Ann. Inst. Statist. Math.* **50** 379–402.
PEEL, D. and MCLACHLAN, G. J. (2000). Robust mixture modelling using the t distribution. *Stat. Comput.* **10** 339–348.
PLESCH, A., SHAW, J. H., BENSON, C., BRYANT, W. A., CARENA, S., COOKE, M., DOLAN, J., FUIS, G., GATH, E. and GRANT, L. (2007). Community fault model (CFM) for southern California. *Bull. Seismol. Soc. Am.* **97** 1793–1802.
PORTER, M. D. and WHITE, G. (2012). Self-exciting hurdle models for terrorist activity. *Ann. Appl. Stat.* **6** 106–124. MR2951531
SCEDC (2013). Southern California earthquake center. https://service.scedc.caltech.edu/eq-catalogs/. Caltech.Dataset. DOI:10.7909/C3WD3xH1.
SIMMA, A. and JORDAN, M. I. (2012). Modeling events with cascades of Poisson processes. Preprint ArXiv:1203.3516.
SPENCE, W., SIPKIN, S. A. and CHOY, G. L. (1989). Measuring the size of an earthquake. *Earthquake Information Bulletin* (*USGS*) **21** 58–63.
STEPHENS, M. (2000). Dealing with label switching in mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 795–809. MR1796293
SUN, J., KABAN, A. and GARIBALDI, J. M. (2010). Robust mixture modeling using the Pearson type VII distribution. In *Neural Networks* (*IJCNN*), *the* 2010 *International Joint Conference on* 1–7. IEEE.
SVENSÉN, M. and BISHOP, C. M. (2005). Robust Bayesian mixture modelling. *Neurocomputing* **64** 235–252.
UTSU, T. (1961). A statistical study on the occurrence of aftershocks. *Geophys. Mag.* **30** 521–605.
VEEN, A. and SCHOENBERG, F. P. (2008). Estimation of space–time branching process models in seismology using an EM-type algorithm. *J. Amer. Statist. Assoc.* **103** 614–624. MR2523998

WHITE, G. and PORTER, M. D. (2014). GPU accelerated MCMC for modeling terrorist activity. *Comput. Statist. Data Anal.* **71** 643–651. MR3131995

WHITE, G., PORTER, M. D. and MAZEROLLE, L. (2013). Terrorism risk, resilience, and volatility: A comparison of terrorism in three Southeast Asian countries. *J. Quant. Criminol.* **29** 295–320.

YEREBAKAN, H. Z., RAJWA, B. and DUNDAR, M. (2014). The infinite mixture of infinite Gaussian mixtures. In *Advances in Neural Information Processing Systems* 28–36.

ZALIAPIN, I., GABRIELOV, A., KEILIS-BOROK, V. and WONG, H. (2008). Clustering analysis of seismicity and aftershock identification. *Phys. Rev. Lett.* **101** 018501.

ZHAO, Q., ERDOGDU, M. A., HE, H. Y., RAJARAMAN, A. and LESKOVEC, J. (2015). Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the* 21*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1513–1522. ACM.

ZHUANG, J. (2011). Next-day earthquake forecasts for the Japan region generated by the ETAS model. *Earth Planets Space* **63** 5.

ZHUANG, J., OGATA, Y. and VERE-JONES, D. (2002). Stochastic declustering of space–time earthquake occurrences. *J. Amer. Statist. Assoc.* **97** 369–380. MR1941459

The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1871–1893
https://doi.org/10.1214/18-AOAS1135

FUNCTIONAL PRINCIPAL VARIANCE COMPONENT TESTING FOR A GENETIC ASSOCIATION STUDY OF HIV PROGRESSION

BY DENIS AGNIEL[*,†], WEN XIE[‡], MYRON ESSEX[‡] AND TIANXI CAI[‡]

Harvard Medical School[*], RAND Corporation[†] and Harvard T. H. Chan School of Public Health[‡]

HIV-1C is the most prevalent subtype of HIV-1 and accounts for over half of HIV-1 infections worldwide. Host genetic influence of HIV infection has been previously studied in HIV-1B, but little attention has been paid to the more prevalent subtype C. To understand the role of host genetics in HIV-1C disease progression, we perform a study to assess the association between longitudinally collected measures of disease and more than 100,000 genetic markers located on chromosome 6. The most common approach to analyzing longitudinal data in this context is linear mixed effects models, which may be overly simplistic in this case. On the other hand, existing flexible and nonparametric methods either require densely sampled points, restrict attention to a single SNP, lack testing procedures, or are cumbersome to fit on the genome-wide scale. We propose a functional principal variance component (FPVC) testing framework which captures the nonlinearity in the CD4 and viral load with low degrees of freedom and is fast enough to carry out thousands or millions of times. The FPVC testing unfolds in two stages. In the first stage, we summarize the markers of disease progression according to their major patterns of variation via functional principal components analysis (FPCA). In the second stage, we employ a simple working model and variance component testing to examine the association between the summaries of disease progression and a set of single nucleotide polymorphisms. We supplement this analysis with simulation results which indicate that FPVC testing can offer large power gains over the standard linear mixed effects model.

REFERENCES

AGNIEL, D., XIE, W., ESSEX, M. and CAI, T. (2018). Supplement to "Functional principal variance component testing for a genetic association study of HIV progression." DOI:10.1214/18-AOAS1135SUPP.

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

BAUM, M. K., CAMPA, A., LAI, S., MARTINEZ, S. S., TSALAILE, L., BURNS, P., FARAHANI, M., LI, Y., VAN WIDENFELT, E., PAGE, J. B. et al. (2013). Effect of micronutrient supplementation on disease progression in asymptomatic, antiretroviral-naive, HIV-infected adults in Botswana: A randomized clinical trial. *JAMA* **310** 2154–2163.

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

Key words and phrases. Genomic association studies, HIV disease progression, functional principal component analysis, longitudinal data, mixed effects models, variance component testing.

CASTRO, P. E., LAWTON, W. H. and SYLVESTRE, E. A. (1986). Principal modes of variation for processes with continuous sample curves. *Technometrics* **28** 329–337.

CHIOU, J.-M., MÜLLER, H.-G. and WANG, J.-L. (2003). Functional quasi-likelihood regression models with smooth random effects. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 405–423. MR1983755

CHUBB, D., WEINHOLD, N., BRODERICK, P., CHEN, B., JOHNSON, D. C., FÖRSTI, A., VIJAYAKRISHNAN, J., MIGLIORINI, G., DOBBINS, S. E., HOLROYD, A. et al. (2013). Common variation at 3q26. 2, 6p21. 33, 17p11. 2 and 22q13. 1 influences multiple myeloma risk. *Nat. Genet.* **45** 1221–1225.

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

CRAINICEANU, C. M. and RUPPERT, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 165–185. MR2035765

FELLAY, J., SHIANNA, K. V., GE, D., COLOMBO, S., LEDERGERBER, B., WEALE, M., ZHANG, K., GUMBS, C., CASTAGNA, A., COSSARIZZA, A. et al. (2007). A whole-genome association study of major determinants for host control of HIV-1. *Science* **317** 944–947.

GERETTI, A. M. (2006). HIV-1 subtypes: Epidemiology and significance for HIV management. *Curr. Opin. Infect. Dis.* **19** 1–7.

GUO, W. (2002). Functional mixed effects models. *Biometrics* **58** 121–128. MR1891050

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

JIANG, J. (1998). Asymptotic properties of the empirical BLUP and BLUE in mixed linear models. *Statist. Sinica* **8** 861–885. MR1651513

JOINT UNITED NATIONS PROGRAMME ON HIV/AIDS (UNAIDS) (2012). Global Report: UNAIDS Report on the Global AIDS Epidemic: 2012. UNAIDS.

KRAFTY, R. T., GIMOTTY, P. A., HOLTZ, D., COUKOS, G. and GUO, W. (2008). Varying coefficient model with unknown within-subject covariance for analysis of tumor growth curves. *Biometrics* **64** 1023–1031. MR2522249

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

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

LINDSTROM, M. J. and BATES, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics* **46** 673–687. MR1085815

MIGUELES, S. A., SABBAGHIAN, M. S., SHUPERT, W. L., BETTINOTTI, M. P., MARINCOLA, F. M., MARTINO, L., HALLAHAN, C. W., SELIG, S. M., SCHWARTZ, D., SULLIVAN, J. et al. (2000). HLA B* 5701 is highly associated with restriction of virus replication in a subgroup of HIV-infected long term nonprogressors. *Proc. Natl. Acad. Sci. USA* **97** 2709–2714.

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

NALLS, M. A., COUPER, D. J., TANAKA, T., VAN ROOIJ, F. J., CHEN, M.-H., SMITH, A. V., TONIOLO, D., ZAKAI, N. A., YANG, Q., GREINACHER, A. et al. (2011). Multiple loci are associated with white blood cell phenotypes. *PLoS Genet.* **7** e1002113–e1002113.

O'BRIEN, S. J. and HENDRICKSON, S. L. (2013). Host genomic influences on HIV/AIDS. *Genome Biol.* **14** 201.

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

RICE, J. A. and SILVERMAN, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc. Ser. B* **53** 233–243. MR1094283

RICE, J. A. and WU, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57** 253–259. MR1833314

ROBINSON, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statist. Sci.* **6** 15–51. MR1108815

SKIBOLA, C. F., BRACCI, P. M., HALPERIN, E., CONDE, L., CRAIG, D. W., AGANA, L., IYADURAI, K., BECKER, N., BROOKS-WILSON, A., CURRY, J. D. et al. (2009). Genetic variants at 6p21. 33 are associated with susceptibility to follicular lymphoma. *Nat. Genet.* **41** 873–875.

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

VAN MANEN, D., KOOTSTRA, N. A., BOESER-NUNNINK, B., HANDULLE, M. A., VAN'T WOUT, A. B. and SCHUITEMAKER, H. (2009). Association of HLA-C and HCP5 gene regions with the clinical course of HIV-1 infection. *AIDS* **23** 19–28.

WU, M. C., LEE, S., CAI, T., LI, Y., BOEHNKE, M. and LIN, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89** 82–93.

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

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

The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1894–1913
https://doi.org/10.1214/18-AOAS1136

ESTIMATING A COMMON COVARIANCE MATRIX FOR NETWORK META-ANALYSIS OF GENE EXPRESSION DATASETS IN DIFFUSE LARGE B-CELL LYMPHOMA[1]

BY ANDERS ELLERN BILGRAU[*,†,2], RASMUS FROBERG BRØNDUM[†,2], POUL SVANTE ERIKSEN[*], KAREN DYBKÆR[†] AND MARTIN BØGSTED[*,†]

Aalborg University and Aalborg University Hospital†*

The estimation of covariance matrices of gene expressions has many applications in cancer systems biology. Many gene expression studies, however, are hampered by low sample size and it has therefore become popular to increase sample size by collecting gene expression data across studies. Motivated by the traditional meta-analysis using random effects models, we present a hierarchical random covariance model and use it for the meta-analysis of gene correlation networks across 11 large-scale gene expression studies of diffuse large B-cell lymphoma (DLBCL). We suggest to use a maximum likelihood estimator for the underlying common covariance matrix and introduce an EM algorithm for estimation. By simulation experiments comparing the estimated covariance matrices by cophenetic correlation and Kullback–Leibler divergence the suggested estimator showed to perform better or not worse than a simple pooled estimator. In a posthoc analysis of the estimated common covariance matrix for the DLBCL data we were able to identify novel biologically meaningful gene correlation networks with eigengenes of prognostic value. In conclusion, the method seems to provide a generally applicable framework for meta-analysis, when multiple features are measured and believed to share a common covariance matrix obscured by study dependent noise.

REFERENCES

AGNELLI, L., FORCATO, M., FERRARI, F., TUANA, G., TODOERTI, K., WALKER, B. A., MORGAN, G. J., LOMBARDI, L., BICCIATO, S. and NERI, A. (2011). The reconstruction of transcriptional networks reveals critical genes with implications for clinical outcome of multiple myeloma. *Clin. Cancer Res.* **17** 7402–7412.

BILGRAU, A. E. (2014). correlateR: Fast, efficient, and robust partial correlations. R package version 0.1. Available at http://github.com/AEBilgrau/correlateR.

BILGRAU, A. E., BRØNDUM, R. F., ERIKSEN, P. S., DYBKÆR, K. and BØGSTED, M. (2018). Supplement to "Estimating a common covariance matrix for network meta-analysis of gene expression datasets in diffuse large B-cell lymphoma." DOI:10.1214/18-AOAS1136SUPPA.

BORENSTEIN, M., HEDGES, L. V., HIGGINS, J. P. and ROTHSTEIN, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res. Synth. Methods* **1** 97–111.

Key words and phrases. Covariance estimation, precision estimation, integrative analysis, meta-analysis, network analysis.

CHENG, P., CORZO, C. A., LUETTEKE, N., YU, B., NAGARAJ, S., BUI, M. M., ORTIZ, M., NACKEN, W., SORG, C., VOGL, T. et al. (2008). Inhibition of dendritic cell differentiation and accumulation of myeloid-derived suppressor cells in cancer is regulated by S100A9 protein. *J. Exp. Med.* **205** 2235–2249.

CHOI, J. K., YU, U., KIM, S. and YOO, O. J. (2003). Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* **19** i84–i90.

CLARKE, C., MADDEN, S. F., DOOLAN, P., AHERNE, S. T., JOYCE, H., O'DRISCOLL, L., GALLAGHER, W. M., HENNESSY, B. T., MORIARTY, M., CROWN, J., KENNEDY, S. and CLYNES, M. (2013). Correlating transcriptional networks to breast cancer survival: A large-scale coexpression analysis. *Carcinogenesis* **34** 2300–2308.

COMPAGNO, M., LIM, W. K., GRUNN, A., NANDULA, S. V., BRAHMACHARY, M., SHEN, Q., BERTONI, F., PONZONI, M., SCANDURRA, M., CALIFANO, A. et al. (2009). Mutations of multiple genes cause deregulation of NF-κB in diffuse large B-cell lymphoma. *Nature* **459** 717–721.

DAI, M., WANG, P., BOYD, A. D., KOSTOV, G., ATHEY, B., JONES, E. G., BUNNEY, W. E., MYERS, R. M., SPEED, T. P., AKIL, H., WATSON, S. J. and MENG, F. (2005). Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.* **33** e175.

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

DERSIMONIAN, R. and LAIRD, N. (1986). Meta-analysis in clinical trials. *Control. Clin. Trials* **7** 177–188.

DYBKÆR, K., BØGSTED, M., FALGREEN, S., BØDKER, J. S., KJELDSEN, M. K., SCHMITZ, A., BILGRAU, A. E., XU-MONETTE, Z. Y., LI, L., BERGKVIST, K. S., LAURSEN, M. B., RODRIGO-DOMINGO, M., MARQUES, S. C., RASMUSSEN, S. B., NYEGAARD, M., GAIHEDE, M., MØLLER, M. B., SAMWORTH, R. J., SHAH, R. D., JOHANSEN, P., EL-GALALY, T. C., YOUNG, K. H. and JOHNSEN, H. E. (2015). A diffuse large B-cell lymphoma classification system that associates normal B-cell subset phenotypes with prognosis. *J. Clin. Oncol.* **33** 1379–1388.

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

FRANÇOIS, R., EDDELBUETTEL, D. and BATES, D. (2012). RcppArmadillo: Rcpp integration for Armadillo templated linear algebra library. R package version 0.3.6.1.

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

FULMER, T. (2008). Suppressing the suppressors. *SciBX* **1**(38). DOI:10.1038/scibx.2008.914.

GALILI, T. (2015). dendextend: An R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* **31** 3718–3720.

GAUTIER, L., COPE, L., BOLSTAD, B. M. and IRIZARRY, R. A. (2004). affy—Analysis of affymetrix GeneChip data at the probe level. *Bioinformatics* **20** 307–315.

HORVATH, S. (2011). *Weighted Network Analysis*: *Applications in Genomics and Systems Biology*. Springer, Berlin.

HUMMEL, M., BENTINK, S., BERGER, H., KLAPPER, W., WESSENDORF, S., BARTH, T. F., BERND, H.-W., COGLIATTI, S. B., DIERLAMM, J., FELLER, A. C. et al. (2006). A biologic definition of Burkitt's lymphoma from transcriptional and genomic profiling. *N. Engl. J. Med.* **354** 2419–2430.

INTERNATIONAL LYMPHOMA STUDY GROUP (1997). A clinical evaluation of the international lymphoma study group classification of non-Hodgkin's lymphoma. *Blood* **89** 3909–3918.

IRIZARRY, R. A., HOBBS, B., COLLIN, F., BEAZER-BARCLAY, Y. D., ANTONELLIS, K. J., SCHERF, U. and SPEED, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4** 249–264.

JIMA, D. D., ZHANG, J., JACOBS, C., RICHARDS, K. L., DUNPHY, C. H., CHOI, W. W., AU, W. Y., SRIVASTAVA, G., CZADER, M. B., RIZZIERI, D. A. et al. (2010). Deep sequencing of the small RNA transcriptome of normal and malignant human B cells identifies hundreds of novel microRNAs. *Blood* **116** e118–e127.

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

LEE, J. A., DOBBIN, K. K. and AHN, J. (2014). Covariance adjustment for batch effect in gene expression data. *Stat. Med.* **33** 2681–2695. MR3256670

LENZ, G., WRIGHT, G. W., EMRE, N. T., KOHLHAMMER, H., DAVE, S. S., DAVIS, R. E., CARTY, S., LAM, L. T., SHAFFER, A., XIAO, W. et al. (2008). Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways. *Proc. Natl. Acad. Sci. USA* **105** 13520–13525.

MATTIUSSI, V., TUMMINELLO, M., IORI, G. and MANTEGNA, R. N. (2011). Comparing correlation matrix estimators via Kullback–Leibler divergence. Preprint, DOI:10.2139/ssrn.1966714.

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

MONTI, S., CHAPUY, B., TAKEYAMA, K., RODIG, S. J., HAO, Y., YEDA, K. T., INGUILIZIAN, H., MERMEL, C., CURRIE, T., DOGAN, A. et al. (2012). Integrative analysis reveals an outcome-associated and targetable pattern of p53 and cell cycle deregulation in diffuse large B cell lymphoma. *Cancer Cell* **22** 359–372.

PHIPSON, B. and SMYTH, G. K. (2010). Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. *Stat. Appl. Genet. Mol. Biol.* **9** Art. 39, 14. MR2746025

R CORE TEAM (2012). *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

REIMAND, J., KOLDE, R. and ARAK, T. (2016). gProfileR: Interface to the 'g:Profiler' toolkit. R package version 0.6.1.

REIMAND, J., ARAK, T., ADLER, P., KOLBERG, L., REISBERG, S., PETERSON, H. and VILO, J. (2016). g:Profiler—A web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.* **44** W83–W89.

SALAVERRIA, I., PHILIPP, C., OSCHLIES, I., KOHLER, C. W., KREUZ, M., SZCZEPANOWSKI, M., BURKHARDT, B., TRAUTMANN, H., GESK, S., ANDRUSIEWICZ, M. et al. (2011). Translocations activating IRF4 identify a subtype of germinal center-derived B-cell lymphoma affecting predominantly children and young adults. *Blood* **118** 139–147.

SHROUT, P. E. and FLEISS, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychol. Bull.* **86** 420–428.

SOKAL, R. R. and ROHLF, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon* **11** 33–40.

STRONCEK, D. F., BUTTERFIELD, L. H., CANNARILE, M. A., DHODAPKAR, M. V., GRETEN, T. F., GRIVEL, J. C., KAUFMAN, D. R., KONG, H. H., KORANGY, F., LEE, P. P., MARINCOLA, F., RUTELLA, S., SIEBERT, J. C., TRINCHIERI, G. and SELIGER, B. (2017). Systematic evaluation of immune regulation and modulation. *J. Immunother. Cancer* **5** 21.

VAN WIERINGEN, W. N. and PEETERS, C. F. W. (2016). Ridge estimation of inverse covariance matrices from high-dimensional data. *Comput. Statist. Data Anal.* **103** 284–303. MR3522633

VISCO, C., LI, Y., XU-MONETTE, Z. Y., MIRANDA, R. N., GREEN, T. M., TZANKOV, A., WEN, W., LIU, W., KAHL, B., D'AMORE, E. et al. (2012). Comprehensive gene expression profiling and immunohistochemical studies support application of immunophenotypic algorithm for molecular subtype classification in diffuse large B-cell lymphoma: A report from the international DLBCL Rituximab-CHOP consortium program study. *Leukemia* **26** 2103–2113.

WILLIAMS, P. M., LI, R., JOHNSON, N. A., WRIGHT, G., HEATH, J.-D. and GASCOYNE, R. D. (2010). A novel method of amplification of FFPET-derived RNA enables accurate disease classification with microarrays. *J. Mol. Diagnostics* **12** 680–686.
XIE, Y. (2013). *Dynamic Documents with R and Knitr*. CRC Press, Boca Raton, FL.

The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1914–1938
https://doi.org/10.1214/18-AOAS1137

TREE-BASED REINFORCEMENT LEARNING FOR ESTIMATING OPTIMAL DYNAMIC TREATMENT REGIMES[1]

BY YEBIN TAO, LU WANG AND DANIEL ALMIRALL

University of Michigan

Dynamic treatment regimes (DTRs) are sequences of treatment decision rules, in which treatment may be adapted over time in response to the changing course of an individual. Motivated by the substance use disorder (SUD) study, we propose a tree-based reinforcement learning (T-RL) method to directly estimate optimal DTRs in a multi-stage multi-treatment setting. At each stage, T-RL builds an unsupervised decision tree that directly handles the problem of optimization with multiple treatment comparisons, through a purity measure constructed with augmented inverse probability weighted estimators. For the multiple stages, the algorithm is implemented recursively using backward induction. By combining semiparametric regression with flexible tree-based learning, T-RL is robust, efficient and easy to interpret for the identification of optimal DTRs, as shown in the simulation studies. With the proposed method, we identify dynamic SUD treatment regimes for adolescents.

REFERENCES

ALMIRALL, D., MCCAFFREY, D. F., GRIFFIN, B. A., RAMCHAND, R., YUEN, R. A. and MURPHY, S. A. (2012). Examining moderated effects of additional adolescent substance use treatment: Structural nested mean model estimation using inverse-weighted regression-with-residuals. Technical Report No. 12-121, Penn State Univ., Univiversity Park, PA.

BATHER, J. (2000). *Decision Theory*: *An Introduction to Dynamic Programming and Sequential Decisions*. Wiley, Chichester. MR1884596

BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. Wadsworth Advanced Books and Software, Belmont, CA. MR0726392

CHAKRABORTY, B. and MOODIE, E. E. M. (2013). *Statistical Methods for Dynamic Treatment Regimes*: *Reinforcement Learning*, *Causal Inference*, *and Personalized Medicine*. Springer, New York. MR3112454

CHAKRABORTY, B. and MURPHY, S. (2014). Dynamic treatment regimes. *Annual Review of Statistics and Its Application* **1** 447–464.

CORTES, C. and VAPNIK, V. (1995). Support-vector networks. *Mach. Learn.* **20** 273–297.

ELOMAA, T. and MALINEN, T. (2003). On lookahead heuristics in decision tree learning. In *International Symposium on Methodologies for Intelligent Systems*. *Lecture Notes in Artificial Intelligence* **2871** 445–453. Springer, Heidelberg.

ESMEIR, S. and MARKOVITCH, S. (2004). Lookahead-based algorithms for anytime induction of decision trees. In *Proceedings of the Twenty-First International Conference on Machine Learning* 257–264. ACM, New York.

Key words and phrases. Multi-stage decision-making, personalized medicine, classification, backward induction, decision tree.

GIFFORD, S. (2015). Difference between outpatient and inpatient treatment programs. Psych Central. Retrieved on July 6, 2016, from http://psychcentral.com/lib/differences-between-outpatient-and-inpatient-treatment-programs.

HERNÁN, M. A., BRUMBACK, B. and ROBINS, J. M. (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *J. Amer. Statist. Assoc.* **96** 440–448. MR1939347

HSER, Y.-I., ANGLIN, M. D., GRELLA, C., LONGSHORE, D. and PRENDERGAST, M. L. (1997). Drug treatment careers A conceptual framework and existing research findings. *J. Subst. Abuse Treat.* **14** 543–558.

HUANG, X., CHOI, S., WANG, L. and THALL, P. F. (2015). Optimization of multi-stage dynamic treatment regimes utilizing accumulated data. *Stat. Med.* **34** 3423–3443. MR3412642

LABER, E. B. and ZHAO, Y. Q. (2015). Tree-based methods for individualized treatment regimes. *Biometrika* **102** 501–514. MR3394271

LAKKARAJU, H. and RUDIN, C. (2017). Learning cost-effective and interpretable treatment regimes. *Proceedings of Machine Learning Research* **54** 166–175.

MARLATT, G. A. and DONOVAN, D. M. (2005). *Relapse Prevention*: *Maintenance Strategies in the Treatment of Addictive Behaviors*. Guilford Press, New York, NY.

MCLELLAN, A. T., LEWIS, D. C., O'BRIEN, C. P. and KLEBER, H. D. (2000). Drug dependence, a chronic medical illness: Implications for treatment, insurance, and outcomes evaluation. *J. Am. Med. Dir. Assoc.* **284** 1689–1695.

MENARD, S. (2002). *Applied Logistic Regression Analysis*, 2nd ed. Sage, Thousand Oaks, CA.

MOODIE, E. E. M., CHAKRABORTY, B. and KRAMER, M. S. (2012). Q-learning for estimating optimal dynamic treatment rules from observational data. *Canad. J. Statist.* **40** 629–645. MR2998853

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

MURPHY, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Stat. Med.* **24** 1455–1481. MR2137651

MURPHY, S. A., VAN DER LAAN, M. J. and ROBINS, J. M. (2001). Marginal mean models for dynamic regimes. *J. Amer. Statist. Assoc.* **96** 1410–1423. MR1946586

MURPHY, S. A., LYNCH, K. G., OSLIN, D., MCKAY, J. R. and TENHAVE, T. (2007). Developing adaptive treatment strategies in substance abuse research. *Drug Alcohol Depend.* **88** S24–S30.

MURTHY, S. and SALZBERG, S. (1995). Lookahead and pathology in decision tree induction. In *Proceedings of Fourteenth International Joint Conference on Artificial Intelligence* 1025–1031. Morgan Kaufmann, San Francisco, CA.

ORELLANA, L., ROTNITZKY, A. and ROBINS, J. M. (2010). Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, Part I: Main content. *Int. J. Biostat.* **6** Art. 8, 49. MR2602551

RAGHUNATHAN, T. E., SOLENBERGER, P. and VAN HOEWYK, J. (2002). IVEware: Imputation and variance estimation software user guide. Survey Methodology Program, Univ. Michigan, Ann Arbor, MI.

REIF, S., GEORGE, P., BRAUDE, L., DOUGHERTY, R. H., DANIELS, A. S., GHOSE, S. S. and DELPHIN-RITTMON, M. E. (2014). Residential treatment for individuals with substance use disorders: Assessing the evidence. *Psychiatr. Serv.* (*Wash. D.C.*) **65** 301–312.

RIVEST, R. L. (1987). Learning decision lists. *Mach. Learn.* **2** 229–246.

ROBINS, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—Application to control of the healthy worker survivor effect. *Math. Model.* **7** 1393–1512. MR0877758

ROBINS, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Comm. Statist. Theory Methods* **23** 2379–2412. MR1293185

ROBINS, J. M. (1997). Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality*, 69–117. Springer, New York. MR1601279

ROBINS, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics*, 189–326. Springer, New York. MR2129402

ROBINS, J. M. and HERNÁN, M. A. (2009). Estimation of the causal effects of time-varying exposures. In *Longitudinal Data Analysis*, 553–599. CRC Press, Boca Raton, FL. MR1500133

ROTNITZKY, A., ROBINS, J. M. and SCHARFSTEIN, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *J. Amer. Statist. Assoc.* **93** 1321–1339. MR1666631

SCHULTE, P. J., TSIATIS, A. A., LABER, E. B. and DAVIDIAN, M. (2014). Q- and A-learning methods for estimating optimal dynamic treatment regimes. *Statist. Sci.* **29** 640–661. MR3300363

SUTTON, R. and BARTO, A. (1998). *Reinforcement Learning*: *An Introduction*. MIT Press, Cambridge.

TAO, Y. and WANG, L. (2017). Adaptive contrast weighted learning for multi-stage multi-treatment decision-making. *Biometrics* **73** 145–155. MR3632360

TAO, Y., WANG, L. and ALMIRALL, D. (2018a). Supplement to "Tree-based reinforcement learning for estimating optimal dynamic treatment regimes." DOI:10.1214/18-AOAS1137SUPPA.

TAO, Y., WANG, L. and ALMIRALL, D. (2018b). Supplement to "Tree-based reinforcement learning for estimating optimal dynamic treatment regimes." DOI:10.1214/18-AOAS1137SUPPB.

THALL, P. F., WOOTEN, L. H., LOGOTHETIS, C. J., MILLIKAN, R. E. and TANNIR, N. M. (2007). Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring. *Stat. Med.* **26** 4687–4702. MR2413392

VAN DER LAAN, M. J. and RUBIN, D. (2006). Targeted maximum likelihood learning. *Int. J. Biostat.* **2** Art. 11, 40. MR2306500

WAGNER, E. H., AUSTIN, B. T., DAVIS, C., HINDMARSH, M., SCHAEFER, J. and BONOMI, A. (2001). Improving chronic illness care: Translating evidence into action. *Health Aff.* (*Millwood, Va.*) **20** 64–78.

WANG, L., ROTNITZKY, A., LIN, X., MILLIKAN, R. E. and THALL, P. F. (2012). Evaluation of viable dynamic treatment regimes in a sequentially randomized trial of advanced prostate cancer. *J. Amer. Statist. Assoc.* **107** 493–508. MR2980060

WATKINS, C. J. and DAYAN, P. (1992). Q-learning. *Mach. Learn.* **8** 279–292.

ZHANG, B., TSIATIS, A. A., DAVIDIAN, M., ZHANG, M. and LABER, E. B. (2012). Estimating optimal treatment regimes from a classification perspective. *Stat* **1** 103–114.

ZHANG, Y., LABER, E. B., TSIATIS, A. and DAVIDIAN, M. (2015). Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics* **71** 895–904. MR3436715

ZHANG, Y., LABER, E. B., TSIATIS, A. and DAVIDIAN, M. (2016). Interpretable dynamic treatment regimes. arXiv preprint arXiv:1606.01472.

ZHAO, Y., ZENG, D., RUSH, A. J. and KOSOROK, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *J. Amer. Statist. Assoc.* **107** 1106–1118. MR3010898

ZHAO, Y.-Q., ZENG, D., LABER, E. B. and KOSOROK, M. R. (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. *J. Amer. Statist. Assoc.* **110** 583–598. MR3367249

ZHOU, X., MAYER-HAMBLETT, N., KHAN, U. and KOSOROK, M. R. (2017). Residual weighted learning for estimating individualized treatment rules. *J. Amer. Statist. Assoc.* **112** 169–187. MR3646564

ZHU, R., ZENG, D. and KOSOROK, M. R. (2015). Reinforcement learning trees. *J. Amer. Statist. Assoc.* **110** 1770–1784. MR3449072

The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1939–1968
https://doi.org/10.1214/18-AOAS1138

A FREQUENCY-CALIBRATED BAYESIAN SEARCH FOR NEW PARTICLES

BY SHIRIN GOLCHI[1] AND RICHARD LOCKHART[1]

Simon Fraser University

The statistical procedure used in the search for new particles is investigated in this paper. The discovery of the Higgs particles is used to lay out the problem and the existing procedures. A Bayesian hierarchical model is proposed to address inference about the parameters of interest while incorporating uncertainty about the nuisance parameters into the model. In addition to inference, a decision making procedure is proposed. A loss function is introduced that mimics the important features of a discovery problem. Given the importance of controlling the "false discovery" and "missed detection" error rates in discovering new phenomena, the proposed procedure is calibrated to control for these error rates.

REFERENCES

ATLAS COLLABORATION (2012). Observation of a new particle in the search for the standard model higgs boson with the ATLAS detector at the LHC. *Phys. Lett. B* **716** 1–29.

BERGER, J. O. (1980). *Statistical Decision Theory*: *Foundations*, *Concepts*, *and Methods*. Springer, New York. MR0580664

CMS COLLABORATION (2012). Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys. Lett. B* **716** 30–61.

CMS COLLABORATION (2013). Updated measurements of the Higgs boson at 125 GeV in the two photon decay channel. Technical Report CMS-PAS-HIG-13-001, CERN, Geneva.

CMS COLLABORATION (2014). Observation of the diphoton decay of the Higgs boson and measurement of its properties. *Eur. Phys. J. C* **74** 3076.

COWAN, G., CRANMER, K., GROSS, E. and VITELLS, O. (2011). Asymptotic formulae for likelihood-based tests of new physics. *Eur. Phys. J. C* **71** 1554–1573.

DAVIES, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **74** 33–43. MR0885917

DEL MORAL, P., DOUCET, A. and JASRA, A. (2006). Sequential Monte Carlo samplers. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 411–436. MR2278333

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

ENGLERT, F. and BROUT, R. (1964). Broken symmetry and the mass of gauge vector mesons. *Phys. Rev. Lett.* **13** 321–323. MR0174314

FELDMAN, G. J. and COUSINS, R. D. (1998). Unified approach to the classical statistical analysis of small signals. *Phys. Rev. D* **57** 3873–3889.

GOLCHI, S. and LOCKHART, R. (2018). Supplement to "A Frequency-calibrated Bayesian search for new particles." DOI:10.1214/18-AOAS1138SUPP.

Key words and phrases. Bayes rule, decision set, Higgs boson, linear loss function, sequential Monte Carlo.

GROSS, E. and VITELLS, O. (2010). Trial factors for the look elsewhere effect in high energy physics. *Eur. Phys. J. C* **70** 525–562.

GURALNIK, G. S. (2009). The history of the Guralnik, Hagen and Kibble development of the theory of spontaneous symmetry breaking and gauge particles. *Internat. J. Modern Phys. A* **24** 2601–2627.

GURALNIK, G. S., HAGEN, C. R. and KIBBLE, T. W. B. (1964). Global conservation laws and massless particles. *Phys. Rev. Lett.* **13** 585–587.

HIGGS, P. W. (1964). Broken symmetries and the masses of gauge bosons. *Phys. Rev. Lett.* **13** 508–509. MR0175554

JASRA, A., STEPHENS, D. A., DOUCET, A. and TSAGARIS, T. (2011). Inference for Lévy-driven stochastic volatility models via adaptive sequential Monte Carlo. *Scand. J. Stat.* **38** 1–22. MR2760137

JOHNSON, V. E. (2013). Uniformly most powerful Bayesian tests. *Ann. Statist.* **41** 1716–1741. MR3127847

PAULO, R. (2005). Default priors for Gaussian processes. *Ann. Statist.* **33** 556–582. MR2163152

RUBINO, G. and TUFFIN, B., eds. (2009). *Rare Event Simulation Using Monte Carlo Methods*. Wiley, Chichester. MR2742378

RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 319–392. MR2649602

The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1969–1992
https://doi.org/10.1214/18-AOAS1139

BAYESIAN RANDOMIZED RESPONSE TECHNIQUE WITH MULTIPLE SENSITIVE ATTRIBUTES: THE CASE OF INFORMATION SYSTEMS RESOURCE MISUSE[1]

BY RAY S. W. CHUNG*, AMANDA M. Y. CHU† AND MIKE K. P. SO*

*The Hong Kong University of Science and Technology** *and Hang Seng Management College*†

The randomized response technique (RRT) is a classical and effective method used to mitigate the distortion arising from dishonest answers. The traditional RRT usually focuses on the case of a single sensitive attribute, and discussion of the case of multiple sensitive attributes is limited. Here, we study a business case to identify some individual and organizational determinants driving information systems (IS) resource misuse in the workplace. People who actually engage in IS resource misuse are probably not willing to provide honest answers, given the sensitivity of the topic. Yet, to develop the causal relationship between IS resource misuse and its determinants, a version of the RRT for multivariate analysis is required. To implement the RRT with multiple sensitive attributes, we propose a Bayesian approach for estimating covariance matrices with incomplete information (resulting from the randomization procedure in the RRT case). The proposed approach (i) accommodates the positive definite condition and other intrinsic parameter constraints in the posterior to improve statistical precision, (ii) incorporates Bayesian shrinkage estimation for covariance matrices despite incomplete information, and (iii) adopts a quasi-likelihood method to achieve Bayesian semiparametric inference for enhancing flexibility. We show the effectiveness of the proposed method in a simulation study. We also apply the Bayesian RRT method and structural equation modeling to identify the causal relationship between IS resource misuse and its determinants.

REFERENCES

ANDO, T. (2011). Predictive Bayesian model selection. *Amer. J. Math. Management Sci.* **31** 13–38. MR2976700

BLAIR, G. and IMAI, K. (2012). Statistical analysis of list experiments. *Polit. Anal.* **20** 47–77.

BLAIR, G., IMAI, K. and ZHOU, Y.-Y. (2015). Design and analysis of the randomized response technique. *J. Amer. Statist. Assoc.* **110** 1304–1319. MR3420703

BLAIR, G. and ZHOU, Y.-Y. (2016). Bayesian randomized response regression. The rr R package [Online]. Available at https://github.com/SensitiveQuestions/rr/blob/master/R/rrBayes.R.

CHAUDHURI, A. (2011). *Randomized Response and Indirect Questioning Techniques in Surveys*. CRC Press, Boca Raton, FL. MR2759226

CHEN, C. C. and SINGH, S. (2011). Pseudo-Bayes and pseudo-empirical Bayes estimators in randomized response sampling. *J. Stat. Comput. Simul.* **81** 779–793. MR2821424

Key words and phrases. Causal modeling, Markov chain Monte Carlo, quasi-likelihood, sensitive responses, shrinkage estimation of covariance matrix, unrelated question design.

CHRISTOFIDES, T. C. (2005). Randomized response technique for two sensitive characteristics at the same time. *Metrika* **62** 53–63. MR2236296

CHU, A. M. Y. and CHAU, P. Y. K. (2014). Development and validation of instruments of information security deviant behavior. *Decis. Support Syst.* **66** 93–101.

CHU, A. M. Y., CHAU, P. Y. K. and SO, M. K. P. (2015). Developing a typological theory using a quantitative approach: A case of information security deviant behavior. *Commun. Assoc. Inf. Syst.* **37** 510–535.

CHUNG, R. S., CHU, A. M. and SO, M. K. (2018). Supplement to "Bayesian Randomized Response Technique with Multiple Sensitive Attributes: The Case of Information Systems Resource Misuse." DOI:10.1214/18-AOAS1139SUPP.

COHEN, A. (1996). On the discriminant validity of the Meyer and Allen measure of organizational commitment: How does it fit with the work commitment construct? *Educ. Psychol. Meas.* **56** 494–503.

COUTTS, E. and JANN, B. (2011). Sensitive questions in online surveys: Experimental results for the randomized response technique (RRT) and the unmatched count technique (UCT). *Sociol. Methods Res.* **40** 169–193. MR2758303

CRUYFF, M. J. L. F., VAN DEN HOUT, A. and VAN DER HEIJDEN, P. G. M. (2008). The analysis of randomized response sum score variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 21–30. MR2412629

D'ARCY, J. and DEVARAJ, S. (2012). Employee misuse of information technology resources: Testing a contemporary deterrence model. *Decis. Sci.* **43** 1091–1124.

D'ARCY, J., HOVAV, A. and GALLETTA, D. (2009). User awareness of security countermeasures and its impact on information systems misuse: A deterrence approach. *Inf. Syst. Res.* **20** 79–98.

FOX, J. A. and TRACY, P. E. (1984). Measuring associations with randomized response. *Soc. Sci. Res.* **13** 188–197.

FOX, J. A. and TRACY, P. E. (1986). *Randomized Response*: *A Method for Sensitive Surveys*. SAGE Publications, Thousand Oaks, CA.

GJESTVANG, C. R. and SINGH, S. (2006). A new randomized response model. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 523–530. MR2278339

GREENBERG, B. G., ABUL-ELA, A.-L. A., SIMMONS, W. R. and HORVITZ, D. G. (1969). The unrelated question randomized response model: Theoretical framework. *J. Amer. Statist. Assoc.* **64** 520–539. MR0247719

GREENBERG, B. G., KUEBLER, R. R., ABERNATHY, J. R. and HORVITZ, D. G. (1971). Application of the randomized response technique in obtaining quantitative data. *J. Amer. Statist. Assoc.* **66** 243–250.

HANSEN, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50** 1029–1054. MR0666123

HÖGLINGER, M., JANN, B. and DIEKMANN, A. (2016). Sensitive questions in online surveys: An experimental evaluation of different implementations of the randomized response technique and the crosswise model. *Surv. Res. Methods* **10** 171–187.

HORVITZ, D. G., SHAH, B. V. and SIMMONS, W. R. (1967). The unrelated question randomized response model. In *Proceedings of Social Statistics Section* 65–72. Amer. Statist. Assoc., Alexandria, VA.

HSIEH, J. J. P.-A., RAI, A. and KEIL, M. (2008). Understanding digital inequality: Comparing continued use behavioral models of the socio-economically advantaged and disadvantaged. *MIS Q.* **32** 97–126.

HUANG, J. Z., LIU, N., POURAHMADI, M. and LIU, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93** 85–98. MR2277742

IMAI, K. (2011). Multivariate regression analysis for the item count technique. *J. Amer. Statist. Assoc.* **106** 407–416. MR2866971

IMAI, K., PARK, B. and GREENE, K. F. (2015). Using the predicted responses from list experiments as explanatory variables in regression models. *Polit. Anal.* **23** 180–196.
JAROS, S. J. (1997). An assessment of Meyer and Allen's (1991) three-component model of organizational commitment and turnover intentions. *J. Vocat. Behav.* **51** 319–337.
JAYARAJ, A., ODUMADE, O. and SINGH, S. (2014). A new quasi-empirical Bayes estimate in randomized response technique. *JSM* 2014.
KAPLAN, D. (2009). *Structural Equation Modeling*, 2nd ed. SAGE Publications, Inc., Thousand Oaks, CA.
KUK, A. Y. C. (1990). Asking sensitive questions indirectly. *Biometrika* **77** 436–438. MR1064822
KWAN, S. S. K., SO, M. K. P. and TAM, K. Y. (2010). Research note—Applying the randomized response technique to elicit truthful responses to sensitive questions in IS research: The case of software piracy behavior. *Inf. Syst. Res.* **21** 941–959.
LEE, C. S., SEDORY, S. A. and SINGH, S. (2016). Cramer–Rao lower bounds of variance for estimating two proportions and their overlap by using two-decks of cards. In *Handbook of Statistics* **34** 353–385.
LIN, C.-P. and DING, C. G. (2003). Modeling information ethics: The joint moderating role of locus of control and job insecurity. *J. Bus. Ethics* **48** 335–346.
LIU, J. S., LIANG, F. and WONG, W. H. (2000). The multiple-try method and local optimization in Metropolis sampling. *J. Amer. Statist. Assoc.* **95** 121–134. MR1803145
LOCANDER, W., SUDMAN, S. and BLACKBURN, N. (1976). An investigation of interview method, threat and response distortion. *J. Amer. Statist. Assoc.* **71** 269–275.
MANGAT, N. S. (1994). An improved randomized response strategy. *J. Roy. Statist. Soc. Ser. B* **56** 93–95. MR1257798
MANGAT, N. S. and SINGH, R. (1990). An alternative randomized response procedure. *Biometrika* **77** 439–442. MR1064823
MEYER, J. P., ALLEN, N. J. and SMITH, C. A. (1993). Commitment to organizations and occupations: Extension and test of a three-component conceptualization. *J. Appl. Psychol.* **78** 538–551.
MINSKY-KELLY, D., HAMBERGER, L. K., PAPE, D. A. and WOLFF, M. (2005). We've had training, now what? Qualitative analysis of barriers to domestic violence screening and referral in a health care setting. *J. Interpers. Violence* **20** 1288–1309.
MUIRHEAD, R. J. (2005). *Aspects of Multivariate Statistical Theory*. Wiley-Interscience, New York. MR0652932
NUNO, A. and ST. JOHN, F. A. V. (2015). How to ask sensitive questions in conservation: A review of specialized questioning techniques. *Biol. Conserv.* **189** 5–15.
PANACCIO, A. and VANDENBERGHE, C. (2009). Perceived organizational support, organizational commitment and psychological well-being: A longitudinal study. *J. Vocat. Behav.* **75** 224–236.
PARK, T. and CASELLA, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.* **103** 681–686. MR2524001
PEACE, A. G., GALLETTA, D. F. and THONG, J. Y. L. (2003). Software piracy in the workplace: A model and empirical test. *J. Manage Inf. Syst.* **20** 153–177.
POLLOCK, K. H. and BEK, Y. (1976). A comparison of three randomized response models for quantitative data. *J. Amer. Statist. Assoc.* **71** 884–886.
RAGHAVARAO, D. and FEDERER, W. T. (1979). Block total response as an alternative to the randomized response method in surveys. *J. Roy. Statist. Soc. Ser. B* **41** 40–45. MR0535543
ROSENFELD, B., IMAI, K. and SHAPIRO, J. N. (2016). An empirical validation study of popular survey methodologies for sensitive questions. *Amer. J. Polit. Sci.* **60** 783–802.
SHAO, J. (2003). *Mathematical Statistics*, 2nd ed. Springer, New York. MR2002723
SHEPHARD, N. and PITT, M. K. (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika* **84** 653–667. MR1603940
SINGH, S. (2003). *Advanced Sampling Theory with Applications*: *How Michael 'Selected' Amy*, *Vol. I*. Kluwer Academic, Dordrecht. MR2032219

SINGH, S. and SEDORY, S. A. (2011). Cramer–Rao lower bound of variance in randomized response sampling. *Sociol. Methods Res.* **40** 536–546. MR2829152

SO, M. K. P., CHEN, C. W. S. and CHEN, M.-T. (2005). A Bayesian threshold nonlinearity test for financial time series. *J. Forecast.* **24** 61–75. MR2143086

TAMHANE, A. C. (1981). Randomized response techniques for multiple sensitive attributes. *J. Amer. Statist. Assoc.* **76** 916–923. MR0650904

TAN, M. T., TIAN, G.-L. and TANG, M.-L. (2009). Sample surveys with sensitive questions: A non-randomized response approach. *Amer. Statist.* **63** 9–16. MR2655697

WARNER, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *J. Amer. Statist. Assoc.* **60** 63–69.

YIN, G. (2009). Bayesian generalized method of moments. *Bayesian Anal.* **4** 191–207. MR2507358

The Annals of Applied Statistics
2018, Vol. 12, No. 3, 1993–2021
https://doi.org/10.1214/18-AOAS1141

DIRECT LIKELIHOOD-BASED INFERENCE FOR DISCRETELY OBSERVED STOCHASTIC COMPARTMENTAL MODELS OF INFECTIOUS DISEASE[1]

BY LAM SI TUNG HO*, FORREST W. CRAWFORD† AND MARC A. SUCHARD*

*University of California, Los Angeles** *and Yale University*†

Stochastic compartmental models are important tools for understanding the course of infectious diseases epidemics in populations and in prospective evaluation of intervention policies. However, calculating the likelihood for discretely observed data from even simple models—such as the ubiquitous susceptible-infectious-removed (SIR) model—has been considered computationally intractable, since its formulation almost a century ago. Recently researchers have proposed methods to circumvent this limitation through data augmentation or approximation, but these approaches often suffer from high computational cost or loss of accuracy. We develop the mathematical foundation and an efficient algorithm to compute the likelihood for discretely observed data from a broad class of stochastic compartmental models. We also give expressions for the derivatives of the transition probabilities using the same technique, making possible inference via Hamiltonian Monte Carlo (HMC). We use the 17th century plague in Eyam, a classic example of the SIR model, to compare our recursion method to sequential Monte Carlo, analyze using HMC, and assess the model assumptions. We also apply our direct likelihood evaluation to perform Bayesian inference for the 2014–2015 Ebola outbreak in Guinea. The results suggest that the epidemic infectious rates have decreased since October 2014 in the Southeast region of Guinea, while rates remain the same in other regions, facilitating understanding of the outbreak and the effectiveness of Ebola control interventions.

REFERENCES

ABATE, J. and WHITT, W. (1992). The Fourier-series method for inverting transforms of probability distributions. *Queueing Syst.* **10** 5–87. MR1149995

ALTHAUS, C. L. (2014). Estimating the reproduction number of Ebola virus (EBOV) during the 2014 outbreak in West Africa. *PLOS Currents Outbreaks* **6**.

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

ARULAMPALAM, M. S., MASKELL, S., GORDON, N. and CLAPP, T. (2002). A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Trans. Signal Process.* **50** 174–188.

BECKER, N. G. and BRITTON, T. (1999). Statistical studies of infectious disease incidence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **61** 287–307. MR1680342

BLUM, M. G. and TRAN, V. C. (2010). HIV with contact tracing: A case study in approximate Bayesian computation. *Biostatistics* **11** 644–660.

Key words and phrases. Epidemic model, multivariate birth process, infectious disease, transition probabilities, Ebola.

BRAUER, F. (2008). Compartmental models in epidemiology. In *Mathematical Epidemiology*. *Lecture Notes in Math*. **1945** 19–79. Springer, Berlin. MR2428372

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

COX, J. C., INGERSOLL, J. E. JR. and ROSS, S. A. (1985). A theory of the term structure of interest rates. *Econometrica* **53** 385–407. MR0785475

CRAWFORD, F. W., STUTZ, T. C. and LANGE, K. (2016). Coupling bounds for approximating birth-death processes by truncation. *Statist. Probab. Lett.* **109** 30–38. MR3434957

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

CSILLÉRY, K., BLUM, M. G., GAGGIOTTI, O. E. and FRANÇOIS, O. (2010). Approximate Bayesian computation (ABC) in practice. *Trends Ecol. Evol.* **25** 410–418.

DE DONDER, T., VAN DEN DUNGEN, F. and VAN LERBERGHE, G. (1920). *Leçons de Thermodynamique et de Chimie Physique*. *Number V.* 1 *in Leçons de Thermodynamique et de Chimie Physique*. Gauthier-Villars, Paris.

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

DUONG, T. (2007). ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R. *J. Stat. Softw.* **21 (7)** 1–16.

FADDY, M. J. (1977). Stochastic compartmental models as approximations to more general stochastic systems with the general stochastic epidemic as an example. *Adv. in Appl. Probab.* **9** 448–461. MR0464445

FELLER, W. (1968). *An Introduction to Probability Theory and Its Applications*. *Vol. I*, 3rd ed. Wiley, New York. MR0228020

GIBSON, G. J. and RENSHAW, E. (1998). Estimating parameters in stochastic compartmental models using Markov chain methods. *Math. Med. Biol.* **15** 19–40.

GOLIGHTLY, A. and WILKINSON, D. J. (2005). Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics* **61** 781–788. MR2196166

HO, L. S. T., XU, J., CRAWFORD, F. W., MININ, V. N. and SUCHARD, M. A. (2018). Birth/birth-death processes and their computable transition probabilities with biological applications. *J. Math. Biol.* **76** 911–944. MR3764582

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

KAREV, G. P., BEREZOVSKAYA, F. S. and KOONIN, E. V. (2005). Modeling genome evolution with a diffusion approximation of a birth-and-death process. *Bioinformatics* **21** iii12–iii19.

KERMACK, W. and MCKENDRICK, A. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A* **115** 700–721.

KING, A. A., NGUYEN, D. and IONIDES, E. L. (2016). Statistical inference for partially observed Markov processes via the R package pomp. *J. Stat. Softw.* **69** 1–43.

LEVIN, D. (1973). Development of non-linear transformations of improving convergence of sequences. *Int. J. Comput. Math.* **3** 371–388. MR0359261

MCKENDRICK, A. (1926). Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematics Society* **44** 98–130.

NEAL, R. M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*. 113–162. CRC Press, Boca Raton, FL. MR2858447

O'NEILL, P. D. (2002). A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. *Math. Biosci.* **180** 103–114. MR1950750

O'NEILL, P. D. and ROBERTS, G. O. (1999). Bayesian inference for partially observed stochastic epidemics. *J. Roy. Statist. Soc. Ser. A* **162** 121–129.

O'NEILL, P. D. and WEN, C. H. (2012). Modelling and inference for epidemic models featuring non-linear infection pressure. *Math. Biosci.* **238** 38–48. MR2947082

OWEN, J., WILKINSON, D. J. and GILLESPIE, C. S. (2015). Scalable inference for Markov processes with intractable likelihoods. *Stat. Comput.* **25** 145–156. MR3304916

RAGGETT, G. (1982). A stochastic model of the Eyam plague. *J. Appl. Stat.* **9** 212–225.

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

REUTER, G. E. H. (1957). Denumerable Markov processes and the associated contraction semigroups on l. *Acta Math.* **97** 1–46. MR0102123

ROBERT, C. P., CORNUET, J.-M., MARIN, J.-M. and PILLAI, N. S. (2011). Lack of confidence in approximate Bayesian computation model choice. *Proc. Natl. Acad. Sci. USA* **108** 15112–15117.

ROBERTS, M., ANDREASEN, V., LLOYD, A. and PELLIS, L. (2015). Nine challenges for deterministic epidemic models. *Epidemics* **10** 49–53.

SCHRANZ, H. W., YAP, V. B., EASTEAL, S., KNIGHT, R. and HUTTLEY, G. A. (2008). Pathological rate matrices: From primates to pathogens. *BMC Bioinform.* **9** 550.

SEVERO, N. C. (1969). Generalizations of some stochastic epidemic models. *Math. Biosci.* **4** 395–402. MR0245166

SIDJE, R. B. (1998). Expokit: A software package for computing matrix exponentials. *ACM Trans. Math. Software* **24** 130–156.

SUNNÅKER, M., BUSETTO, A. G., NUMMINEN, E., CORANDER, J., FOLL, M. and DESSIMOZ, C. (2013). Approximate Bayesian computation. *PLoS Comput. Biol.* **9** e1002803, 10. MR3032718

VERDINELLI, I. and WASSERMAN, L. (1995). Computing Bayes factors using a generalization of the Savage–Dickey density ratio. *J. Amer. Statist. Assoc.* **90** 614–618. MR1340514

WHO EBOLA RESPONSE TEAM (2014). Ebola virus disease in West Africa—The first 9 months of the epidemic and forward projections. *N. Engl. J. Med.* **371** 1481–1495.

WHO EBOLA RESPONSE TEAM (2015). West African Ebola epidemic after one year-slowing but not yet under control. *N. Engl. J. Med.* **372** 584–587.

WORLD HEALTH ORGANIZATION (2015). Statement on the 4th meeting of the IHR Emergency Committee on the 2014 Ebola outbreak in West Africa. World Health Organization, IHR Emergency Committee regarding Ebola.

The Annals of Applied Statistics

Next Issues

Continued

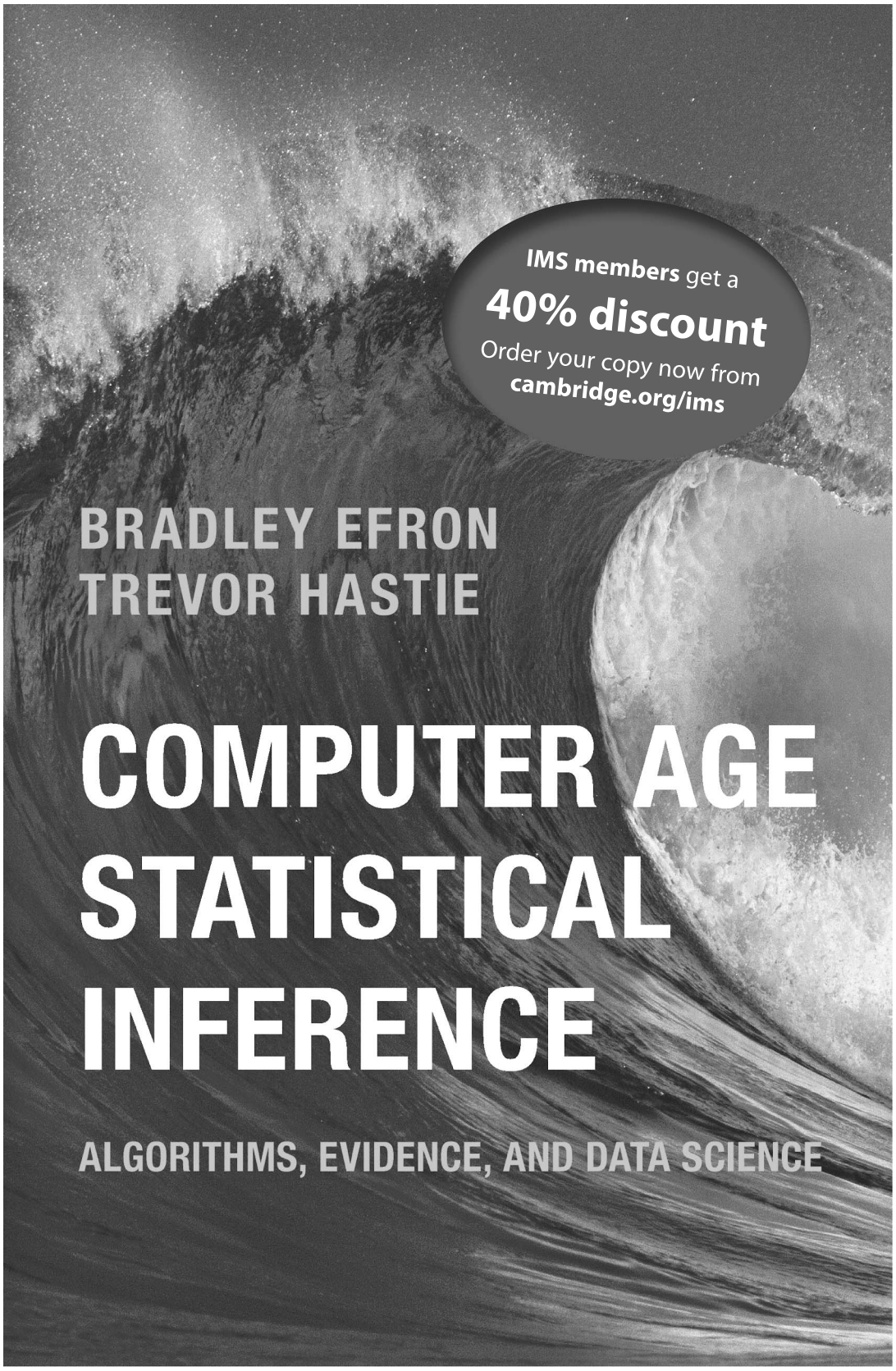
The Annals of Applied Statistics

Next Issues–*Continued*

Continued

The Annals of Applied Statistics

Next Issues–*Continued*


IMS members get a
40% discount
Order your copy now from
cambridge.org/ims
BRADLEY EFRON
TREVOR HASTIE
COMPUTER AGE
STATISTICAL
INFERENCE
ALGORITHMS, EVIDENCE, AND DATA SCIENCE