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

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Special section in memory of Stephen E. Fienberg (1942–2016)
AOAS Editor-in-Chief 2013–2015

continued

Vol. 12, No. 2—June 2018

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

THE ANNALS *of* APPLIED STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Continued from front cover

Articles

Vol. 12, No. 2—June 2018

The Annals of Applied Statistics
2018, Vol. 12, No. 2, iii-x
https://doi.org/10.1214/17-AOAS122ED

SPECIAL SECTION IN MEMORY OF STEPHEN E. FIENBERG (1942–2016) AOAS EDITOR-IN-CHIEF 2013–2015

REFERENCES

EROSHEVA, E. and SLAVKOVIC, A. (2017). Obituary: Stephen E. Fienberg 1942–2016. *IMS Bulletin* **46** 4–5.

KADANE, J. B. (2017). Stephen Fienberg 1942–2016. *J. Roy. Statist. Soc. Ser. A* **180** 927–928.

MEJIA, R. (2017). Stephen E. Fienberg (1942–2016). Statistician who campaigned for better science in court. *Nature* **542** 415.

STRAF, M. L. and TANUR, J. M. (2013). A conversation with Stephen E. Fienberg. *Statist. Sci.* **28** 447–463. MR3135541

The Annals of Applied Statistics
2018, Vol. 12, No. 2, 683–684
https://doi.org/10.1214/17-AOAS1131SF

ON STEPHEN E. FIENBERG AS A DISCUSSANT AND A FRIEND

BY DONALD B. RUBIN

Harvard University (emeritus), Tsinghua University and Temple University

The Annals of Applied Statistics
2018, Vol. 12, No. 2, 685–726
https://doi.org/10.1214/18-AOAS1161SF

STATISTICAL PARADISES AND PARADOXES IN BIG DATA (I): LAW OF LARGE POPULATIONS, BIG DATA PARADOX, AND THE 2016 US PRESIDENTIAL ELECTION[1]

BY XIAO-LI MENG

Harvard University

Statisticians are increasingly posed with thought-provoking and even paradoxical questions, challenging our qualifications for entering the statistical paradises created by Big Data. By developing measures for data quality, this article suggests a framework to address such a question: "Which one should I trust more: a 1% survey with 60% response rate or a self-reported administrative dataset covering 80% of the population?" A 5-element Euler-formula-like identity shows that for any dataset of size n, probabilistic or not, the difference between the sample average $\overline{X}_n$ and the population average $\overline{X}_N$ is the product of three terms: (1) a *data quality* measure, $\rho_{R,X}$, the correlation between X_j and the response/recording indicator R_j; (2) a *data quantity* measure, $\sqrt{(N-n)/n}$, where N is the population size; and (3) a *problem difficulty* measure, σ_X, the standard deviation of X. This decomposition provides multiple insights: (I) Probabilistic sampling ensures high data quality by controlling $\rho_{R,X}$ at the level of $N^{-1/2}$; (II) When we lose this control, the impact of N is no longer canceled by $\rho_{R,X}$, leading to a *Law of Large Populations* (LLP), that is, our estimation error, relative to the benchmarking rate $1/\sqrt{n}$, increases with $\sqrt{N}$; and (III) the "bigness" of such Big Data (for population inferences) should be measured by the *relative size* $f = n/N$, not the *absolute size* n; (IV) When combining data sources for population inferences, those relatively tiny but higher quality ones should be given far more weights than suggested by their sizes.

Estimates obtained from the Cooperative Congressional Election Study (CCES) of the 2016 US presidential election suggest a $\rho_{R,X} \approx -0.005$ for self-reporting to vote for Donald Trump. Because of LLP, this seemingly minuscule data defect correlation implies that the simple sample proportion of the self-reported voting preference for Trump from 1% of the US eligible voters, that is, $n \approx 2{,}300{,}000$, has the same mean squared error as the corresponding sample proportion from a genuine simple random sample of size $n \approx 400$, a 99.98% reduction of sample size (and hence our confidence). The CCES data demonstrate LLP vividly: on average, the larger the state's voter populations, the further away the actual Trump vote shares from the usual 95% confidence intervals based on the sample proportions. This should remind us that, without taking data quality into account, population inferences with Big Data are subject to a *Big Data Paradox*: the more the data, the surer we fool ourselves.

Key words and phrases. Bias-variance tradeoff, data defect correlation, data defect index (d.d.i.), data confidentiality and privacy, data quality-quantity tradeoff, Euler identity, Monte Carlo and Quasi Monte Carlo (MCQMC), non-response bias.

REFERENCES

ANDERSON, M. and FIENBERG, S. E. (1999). *Who Counts? The Politics of Census-Taking in Contemporary America*. Russell Sage Foundation.

ANSOLABEHERE, S. and HERSH, E. (2012). Validation: What big data reveal about survey misreporting and the real electorate. *Polit. Anal.* **20** 437–459.

ANSOLABEHERE, S., SCHAFFNER, B. F. and LUKS, S. (2017). Guide to the 2016 Cooperative Congressional Election Survey. Available at http://dx.doi.org/10.7910/DVN/GDF6Z0.

ARGENTINI, G. (2007). A matrix generalization of Euler identity $e^{ix} = \cos(x) + i\sin(x)$. Preprint. Available at arXiv:math/0703448.

BAYARRI, M. J., BENJAMIN, D. J., BERGER, J. O. and SELLKE, T. M. (2016). Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *J. Math. Psych.* **72** 90–103. MR3506028

BETHLEHEM, J. (2009). The rise of survey sampling. *CBS Discussion Paper* **9015**.

BURDEN, B. C. (2000). Voter turnout and the national election studies. *Polit. Anal.* **8** 389–398.

CHEN, C., DUAN, N., MENG, X.-L. and ALEGRIA, M. (2006). Power-shrinkage and trimming: Two ways to mitigate excessive weights. In *Proceedings of the Survey Research Methods Section of the American Statistical Association* 2839–2846.

CHEN, Y., MENG, X.-L., WANG, X., VAN DYK, D. A., MARSHALL, H. L. and KASHYAP, V. L. (2018). Calibration concordance for astronomical instruments via multiplicative shrinkage. *J. Amer. Statist. Assoc.* To appear.

COHN, N. (2017). Election review: Why crucial state polls turned out to be wrong. *The New York Times*, *June 1st*.

DONOHO, D. (2017). 50 years of data science. *J. Comput. Graph. Statist.* **26** 745–766. MR3765335

DUNCAN, G. T. and FIENBERG, S. E. (1997). Obtaining information while preserving privacy: A Markov perturbation method for tabular data. In *Joint Statistical Meetings* 351–362.

EFRON, B. and HINKLEY, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika* **65** 457–487. MR0521817

FIENBERG, S. E. (1994). Conflicts between the needs for access to statistical information and demands for confidentiality. *J. Off. Stat.* **10** 115–132.

FIENBERG, S. E. (1996). Applying statistical concepts and approaches in academic administration. In *Education in a Research University* 65–82. Stanford Univ. Press, Stanford.

FIENBERG, S. E. (2007). *The Analysis of Cross-Classified Categorical Data*, Springer Science & Business Media.

FIENBERG, S. E. (2010). The relevance or irrelevance of weights for confidentiality and statistical analyses. *Journal of Privacy and Confidentiality* **1** 183–195.

FIENBERG, S. E., PETROVIĆ, S. and RINALDO, A. (2011). Algebraic statistics for p_1 random graph models: Markov bases and their uses. In *Looking Back*. *Lect. Notes Stat. Proc.* **202** 21–38. Springer, New York. MR2856692

FIENBERG, S. E., RINALDO, A. and YANG, X. (2010). Differential privacy and the risk-utility tradeoff for multi-dimensional contingency tables. In *International Conference on Privacy in Statistical Databases* 187–199. Springer, Berlin.

FIRTH, D. and BENNETT, K. E. (1998). Robust models in probability sampling (with discussions). *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 3–21. MR1625672

FRÉCHET, M. (1951). Sur les tableaux de corrélation dont les marges sont données. *Ann. Univ. Lyon. Sect. A.* (3) **14** 53–77. MR0049518

FULLER, W. A. (2011). *Sampling Statistics* Wiley, New York.

GELMAN, A. (2007). Struggles with survey weighting and regression modeling (with discussions). *Statist. Sci.* **22** 153–188.

GELMAN, A. and AZARI, J. (2017). 19 things we learned from the 2016 election (with discussions). *Statistics and Public Policy* **4** 1–10.

HARTLEY, H. O. and ROSS, A. (1954). Unbiased ratio estimators. *Nature* **174** 270–271.

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

HICKERNELL, F. J. (2006). *Koksma–Hlawka Inequality*. Wiley Online Library.

HICKERNELL, F. J. (2018). The trio identity for Quasi-Monte Carlo error analysis. In *Monte Carlo and Quasi Monte Carlo* (P. Glynn and A. Owen, eds.) 13–37. Springer.

HÖFFDING, W. (1940). Masstabinvariante Korrelationstheorie. *Schr. Math. Inst. u. Inst. Angew. Math. Univ. Berlin* **5** 181–233. MR0004426

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

KEIDING, N. and LOUIS, T. A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys (with discussions). *J. Roy. Statist. Soc. Ser. A* **179** 319–376. MR3461587

KIM, J. K. and KIM, J. J. (2007). Nonresponse weighting adjustment using estimated response probability. *Canad. J. Statist.* **35** 501–514. MR2381396

KIM, J. K. and RIDDLES, M. K. (2012). Some theory for propensity-score-adjustment estimators in survey sampling. *Surv. Methodol.* **38** 157.

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

KONG, A., MCCULLAGH, P., MENG, X.-L., NICOLAE, D. and TAN, Z. (2003). A theory of statistical models for Monte Carlo integration (with discussions). *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 585–618. MR1998624

KONG, A., MCCULLAGH, P., MENG, X.-L. and NICOLAE, D. L. (2007). Further explorations of likelihood theory for Monte Carlo integration. In *Advances in Statistical Modeling and Inference*. *Ser. Biostat.* **3** 563–592. World Sci. Publ., Hackensack, NJ. MR2416134

LIU, J. S. (1996). Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Stat. Comput.* **6** 113–119.

LIU, K. and MENG, X.-L. (2016). There is individualized treatment. Why not individualized inference? *The Annual Review of Statistics and Its Applications* **3** 79–111.

LIU, J., MENG, X.-L., CHEN, C. and ALEGRIA, M. (2013). Statistics can lie but can also correct for lies: Reducing response bias in NLAAS via Bayesian imputation. *Stat. Interface* **6** 387–398. MR3105229

LOHR, S. L. (2009). *Sampling: Design and Analysis*. Nelson Education.

MCDONALD, M. P. (2017). 2016 November general election turnout rates. Available at http://www.electproject.org/2016g.

MEHRHOFF, J. (2016). Executive summary: Meng, X.-L. (2014), "A trio of inference problems that could win you a Nobel prize in statistics (if you help fund it)". Conference handout.

MENG, X.-L. (1993). On the absolute bias ratio of ratio estimators. *Statist. Probab. Lett.* **18** 345–348. MR1247444

MENG, X.-L. (2005). Comment: Computation, survey and inference. *Statist. Sci.* **20** 21–28.

MENG, X.-L. (2014). A trio of inference problems that could win you a Nobel prize in statistics (if you help fund it). In *Past, Present, and Future of Statistical Science* (X. Lin et al., eds.) 537–562. CRC Press.

MENG, X.-L. (2018). Statistical paradises and paradoxes in big data (II): Multi-resolution inference, Simpson's paradox, and individualized treatments. Preprint.

OWEN, A. B. (2013). Monte Carlo Theory, Methods and Examples. Available at http://statweb.stanford.edu/~owen/mc/.

ROYALL, R. (1968). An old approach to finite population sampling theory. *J. Amer. Statist. Assoc.* **63** 1269–1279.

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

SENN, S. (2007). Trying to be precise about vagueness. *Stat. Med.* **26** 1417–1430. MR2359149

SHIRANI-MEHR, H., ROTHSCHILD, D., GOEL, S. and GELMAN, A. (2018). Disentangling bias and variance in election polls. Unpublished manuscript. Available at http://www.stat.columbia.edu/~gelman/research/unpublished/polling-errors.pdf.

SQUIRE, P. (1988). Why the 1936 literary digest poll failed. *Public Opin. Q.* **52** 125–133.

TROXEL, A. B., MA, G. and HEITJAN, D. F. (2004). An index of local sensitivity to nonignorability. *Statist. Sinica* **14** 1221–1237. MR2126350

The Annals of Applied Statistics
2018, Vol. 12, No. 2, 727–749
https://doi.org/10.1214/18-AOAS1155SF

HYPOTHESIS TESTING FOR HIGH-DIMENSIONAL MULTINOMIALS: A SELECTIVE REVIEW[1]

BY SIVARAMAN BALAKRISHNAN AND LARRY WASSERMAN

Carnegie Mellon University

In memory of Stephen E. Fienberg

The statistical analysis of discrete data has been the subject of extensive statistical research dating back to the work of Pearson. In this survey we review some recently developed methods for testing hypotheses about high-dimensional multinomials. Traditional tests like the χ^2-test and the likelihood ratio test can have poor power in the high-dimensional setting. Much of the research in this area has focused on finding tests with asymptotically normal limits and developing (stringent) conditions under which tests have normal limits. We argue that this perspective suffers from a significant deficiency: it can exclude many high-dimensional cases when—despite having non-normal null distributions—carefully designed tests can have high power. Finally, we illustrate that taking a minimax perspective and considering refinements of this perspective can lead naturally to powerful and practical tests.

REFERENCES

ACHARYA, J., DASKALAKIS, C. and KAMATH, G. (2015). Optimal testing for properties of distributions. In *Advances in Neural Information Processing Systems* 3591–3599.

ACHARYA, J., DAS, H., JAFARPOUR, A., ORLITSKY, A., PAN, S. and SURESH, A. (2012). Competitive classification and closeness testing. In *Proceedings of the* 25*th Annual Conference on Learning Theory* **23** 22.1–22.18.

ADDARIO-BERRY, L., BROUTIN, N., DEVROYE, L. and LUGOSI, G. (2010). On combinatorial testing problems. *Ann. Statist.* **38** 3063–3092. MR2722464

ARIAS-CASTRO, E., CANDÈS, E. J. and DURAND, A. (2011). Detection of an anomalous cluster in a network. *Ann. Statist.* **39** 278–304. MR2797847

ARIAS-CASTRO, E., PELLETIER, B. and SALIGRAMA, V. (2018). Remember the curse of dimensionality: The case of goodness-of-fit testing in arbitrary dimension. *J. Nonparametr. Stat.* **30** 448–471. MR3794401

BALAKRISHNAN, S. and WASSERMAN, L. (2017). Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates. Available at arXiv:1706.10003.

BALAKRISHNAN, S. and WASSERMAN, L. (2018). Hypothesis testing for densities and high-dimensional multinomials II: Sharp local minimax rates. Forthcoming.

BARRON, A. R. (1989). Uniformly powerful goodness of fit tests. *Ann. Statist.* **17** 107–124. MR0981439

BATU, T., KUMAR, R. and RUBINFELD, R. (2004). Sublinear algorithms for testing monotone and unimodal distributions. In *Proceedings of the* 36*th Annual ACM Symposium on Theory of Computing* 381–390. ACM, New York. MR2121623

Key words and phrases. Hypothesis testing, high-dimensional multinomials.

BATU, T., FORTNOW, L., RUBINFELD, R., SMITH, W. D. and WHITE, P. (2000). Testing that distributions are close. In 41*st Annual Symposium on Foundations of Computer Science* (*Redondo Beach*, *CA*, 2000) 259–269. IEEE Comput. Soc., Los Alamitos, CA. MR1931824

BERGER, R. L. and BOOS, D. D. (1994). *P* values maximized over a confidence set for the nuisance parameter. *J. Amer. Statist. Assoc.* **89** 1012–1016. MR1294746

BERTHET, Q. and RIGOLLET, P. (2013). Optimal detection of sparse principal components in high dimension. *Ann. Statist.* **41** 1780–1815. MR3127849

BHATTACHARYA, B. and VALIANT, G. (2015). Testing closeness with unequal sized samples. In *Advances in Neural Information Processing Systems* 2611–2619.

BICKEL, P. J., RITOV, Y. and STOKER, T. M. (2006). Tailor-made tests for goodness of fit to semiparametric hypotheses. *Ann. Statist.* **34** 721–741. MR2281882

BISHOP, Y. M. M., FIENBERG, S. E. and HOLLAND, P. W. (1977). *Discrete Multivariate Analysis*: *Theory and Practice*. MIT Press, Cambridge, MA. With the collaboration of Richard J. Light and Frederick Mosteller. Third printing. MR0431514

CAI, T. T. and LOW, M. G. (2015). A framework for estimation of convex functions. *Statist. Sinica* **25** 423–456. MR3379081

CANONNE, C. L. (2018). A survey on distribution testing: Your data is big. But is it blue? *Theory Comput.* To appear.

CANONNE, C. L., DIAKONIKOLAS, I., GOULEAKIS, T. and RUBINFELD, R. (2016). Testing shape restrictions of discrete distributions. In 33*rd Symposium on Theoretical Aspects of Computer Science*. *LIPIcs*. *Leibniz Int*. *Proc*. *Inform*. **47** Art. No. 25. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern. MR3539122

CHAN, S.-O., DIAKONIKOLAS, I., VALIANT, G. and VALIANT, P. (2014). Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the Twenty-Fifth Annual ACM–SIAM Symposium on Discrete Algorithms* 1193–1203. ACM, New York. MR3376448

CHATTERJEE, S., GUNTUBOYINA, A. and SEN, B. (2015). On risk bounds in isotonic and other shape restricted regression problems. *Ann. Statist.* **43** 1774–1800. MR3357878

DASKALAKIS, C., KAMATH, G. and WRIGHT, J. (2018). Which distribution distances are sublinearly testable? In *Proceedings of the Twenty-Ninth Annual ACM–SIAM Symposium on Discrete Algorithms* 2747–2764. SIAM, Philadelphia, PA.

DEVROYE, L. and GYÖRFI, L. (1985). *Nonparametric Density Estimation*: *The* L_1 *View*. Wiley, New York. MR0780746

DIAKONIKOLAS, I. and KANE, D. M. (2016). A new approach for testing properties of discrete distributions. In 57*th Annual IEEE Symposium on Foundations of Computer Science—FOCS* 2016 685–694. IEEE Comput. Soc., Los Alamitos, CA. MR3631031

DIAKONIKOLAS, I., KANE, D. M. and NIKISHKIN, V. (2015a). Optimal algorithms and lower bounds for testing closeness of structured distributions. In 2015 *IEEE* 56*th Annual Symposium on Foundations of Computer Science—FOCS* 2015 1183–1202. IEEE Comput. Soc., Los Alamitos, CA. MR3473364

DIAKONIKOLAS, I., KANE, D. M. and NIKISHKIN, V. (2015b). Testing identity of structured distributions. In *Proceedings of the Twenty-Sixth Annual ACM–SIAM Symposium on Discrete Algorithms* 1841–1854. SIAM, Philadelphia, PA. MR3451147

DIAKONIKOLAS, I., KANE, D. M. and NIKISHKIN, V. (2017). Near-optimal closeness testing of discrete histogram distributions. In 44*th International Colloquium on Automata*, *Languages*, *and Programming*. *LIPIcs*. *Leibniz Int*. *Proc*. *Inform*. **80** Art. No. 8. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern. MR3685748

DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. MR2065195

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

DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.* **24** 508–539. MR1394974

ERMAKOV, M. S. (1991). Minimax detection of a signal in Gaussian white noise. *Theory Probab. Appl.* **35** 667–679. MR1090496

FIENBERG, S. E. (1979). The use of chi-squared statistics for categorical data problems. *J. Roy. Statist. Soc. Ser. B* **41** 54–64. MR0535545

FIENBERG, S. E. (1980). *The Analysis of Cross-Classified Categorical Data*, 2nd ed. MIT Press, Cambridge, MA. MR0623082

FIENBERG, S. E. and HOLLAND, P. W. (1973). Simultaneous estimation of multinomial cell probabilities. *J. Amer. Statist. Assoc.* **68** 683–691. MR0359153

GINÉ, E. and NICKL, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models. Cambridge Series in Statistical and Probabilistic Mathematics* **40**. Cambridge Univ. Press, New York. MR3588285

GOLDENSHLUGER, A. and LEPSKI, O. (2011). Bandwidth selection in kernel density estimation: Oracle inequalities and adaptive minimax optimality. *Ann. Statist.* **39** 1608–1632. MR2850214

GOLDREICH, O. (2017). *Introduction to Property Testing*. Cambridge Univ. Press, Cambridge.

HABERMAN, S. J. (1977). Log-linear models and frequency tables with small expected cell counts. *Ann. Statist.* **5** 1148–1169. MR0448675

HOEFFDING, W. (1965). Asymptotically optimal tests for multinomial distributions. *Ann. Math. Stat.* **36** 369–408. MR0173322

HOLST, L. (1972). Asymptotic normality and efficiency for certain goodness-of-fit tests. *Biometrika* **59** 137–145. MR0314193

INDYK, P., LEVI, R. and RUBINFELD, R. (2012). Approximating and testing k-histogram distributions in sub-linear time. In *Proceedings of the* 31*st ACM SIGMOD–SIGACT–SIGART Symposium on Principles of Database Systems*, *PODS* 2012 15–22.

INGSTER, YU. I. (1997). Adaptive chi-square tests. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov.* (*POMI*) **244** 150–166, 333. MR1700386

INGSTER, YU. I. and SUSLINA, I. A. (2003). *Nonparametric Goodness-of-Fit Testing Under Gaussian Models. Lecture Notes in Statistics* **169**. Springer, New York. MR1991446

INGSTER, Y. I., TSYBAKOV, A. B. and VERZELEN, N. (2010). Detection boundary in sparse regression. *Electron. J. Stat.* **4** 1476–1526. MR2747131

IVČENKO, G. I. and MEDVEDEV, JU. I. (1980). Decomposable statistics and hypothesis testing. The case of small samples. *Theory Probab. Appl.* **23** 540–551. MR0516276

JIAO, J., HAN, Y. and WEISSMAN, T. (2017). Minimax estimation of the l_1 distance. Available at arXiv:1705.00807.

KOEHLER, K. J. (1986). Goodness-of-fit tests for log-linear models in sparse contingency tables. *J. Amer. Statist. Assoc.* **81** 483–493. MR0845887

KOEHLER, K. J. and LARNTZ, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. *J. Amer. Statist. Assoc.* **75** 336–344.

LECAM, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** 38–53. MR0334381

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

LEPSKI, O. V. and SPOKOINY, V. G. (1999). Minimax nonparametric hypothesis testing: The case of an inhomogeneous alternative. *Bernoulli* **5** 333–358. MR1681702

MORRIS, C. (1975). Central limit theorems for multinomial sums. *Ann. Statist.* **3** 165–188. MR0370871

PANINSKI, L. (2008). A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Trans. Inform. Theory* **54** 4750–4755. MR2591136

READ, T. R. C. and CRESSIE, N. A. C. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer, New York. MR0955054

RUBINFELD, R. (2012). Taming big probability distributions. *XRDS* **19** 24–28.

SPOKOINY, V. G. (1996). Adaptive hypothesis testing using wavelets. *Ann. Statist.* **24** 2477–2498. MR1425962

VALIANT, G. and VALIANT, P. (2011). Estimating the unseen: An $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In *STOC'11—Proceedings of the 43rd ACM Symposium on Theory of Computing* 685–694. ACM, New York. MR2932019

VALIANT, G. and VALIANT, P. (2017). An automatic inequality prover and instance optimal identity testing. *SIAM J. Comput.* **46** 429–455. MR3614697

VAN DE GEER, S. (2016). *Estimation and Testing Under Sparsity*. *Lecture Notes in Math.* **2159**. Springer, Cham. MR3526202

WEI, Y. and WAINWRIGHT, M. J. (2017). The local geometry of testing in ellipses: Tight control via localized Kolomogorov widths. Available at arXiv:1712.00711.

The Annals of Applied Statistics
2018, Vol. 12, No. 2, 750–770
https://doi.org/10.1214/18-AOAS1160SF

WHEN SHOULD MODES OF INFERENCE DISAGREE? SOME SIMPLE BUT CHALLENGING EXAMPLES[1]

BY D. A. S. FRASER[*], N. REID[*] AND WEI LIN[†]

University of Toronto[*] *and AidVoice Lab*[†]

At a recent conference on Bayes, fiducial and frequentist inference, David Cox presented eight illustrative examples, chosen to highlight potential difficulties for the theory of inference. We discuss these examples in light of the efforts of the conference, and related meetings, to study the similarities and differences between the approaches to inference. Emphasis is placed on the goal of finding a distribution for an unknown parameter.

REFERENCES

BARNDORFF-NIELSEN, O. E. (1991). Modified signed log likelihood ratio. *Biometrika* **78** 557–563. MR1130923

BAYES, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philos. Trans. R. Soc.* **53** 370–418.

BERGER, J. (2006). The case for objective Bayesian analysis. *Bayesian Anal.* **1** 385–402. MR2221271

BLISS, C. (1935a). The calculation of dosage-mortality curves. *Ann. Appl. Biol.* **22** 134–167.

BLISS, C. (1935b). The comparison of dosage-mortality data. *Ann. Appl. Biol.* **22** 307–333.

BOX, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *J. Roy. Statist. Soc. Ser. A* **143** 383–430. MR0603745

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations. *J. Roy. Statist. Soc. Ser. B* **26** 211–252. MR0192611

BRAZZALE, A. R., DAVISON, A. C. and REID, N. (2007). *Applied Asymptotics: Case Studies in Small-Sample Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **23**. Cambridge Univ. Press, Cambridge. MR2342742

COX, D. R. (1958). Some problems connected with statistical inference. *Ann. Math. Stat.* **29** 357–372. MR0094890

COX, D. R. (2006). *Principles of Statistical Inference*. Cambridge Univ. Press, Cambridge. MR2278763

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. Chapman & Hall, London. MR0370837

COX, D. R. and REID, N. (1987). Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. Ser. B* **49** 1–39. MR0893334

DATTA, G. S. and GHOSH, M. (1995). Some remarks on noninformative priors. *J. Amer. Statist. Assoc.* **90** 1357–1363. MR1379478

DATTA, G. S. and MUKERJEE, R. (2004). *Probability Matching Priors: Higher Order Asymptotics. Lecture Notes in Statistics* **178**. Springer, New York. MR2053794

DAVISON, A. C., FRASER, D. A. S., REID, N. and SARTORI, N. (2014). Accurate directional inference for vector parameters in linear exponential families. *J. Amer. Statist. Assoc.* **109** 302–314. MR3180565

Key words and phrases. Asymptotic theory, confidence distribution, fiducial density, marginalization paradox, noninformative priors.

EFRON, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika* **80** 3–26. MR1225211

EFRON, B. (2013). Bayes' theorem in the 21st century. *Science* **340** 1177–1178. MR3087705

EVANS, M. (2015). *Measuring Statistical Evidence Using Relative Belief. Monographs on Statistics and Applied Probability* **144**. CRC Press, Boca Raton, FL. MR3616661

FIELLER, E. C. (1954). Symposium on interval estimation: Some problems in interval estimation. *J. Roy. Statist. Soc. Ser. B* **16** 175–185. MR0093076

FIENBERG, S. E. (2006). Does it make sense to be an "objective Bayesian"? (comment on articles by Berger and by Goldstein). *Bayesian Anal.* **1** 429–432. MR2221275

FISHER, R. A. (1930). Inverse probability. *Proc. Camb. Philos. Soc.* **26** 528–535.

FISHER, R. A. (1956). *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh.

FRASER, D. A. S. (1961). The fiducial method and invariance. *Biometrika* **48** 261–280. MR0133910

FRASER, D. A. S. (1966). Structural probability and a generalization. *Biometrika* **53** 1–9. MR0196840

FRASER, D. A. S. (1990). Tail probabilities from observed likelihoods. *Biometrika* **77** 65–76. MR1049409

FRASER, D. A. S. (2003). Likelihood for component parameters. *Biometrika* **90** 327–339. MR1986650

FRASER, D. A. S. (2011). Is Bayes posterior just quick and dirty confidence? *Statist. Sci.* **26** 299–316. MR2918001

FRASER, D. A. S. (2016). The p-value function: The core concept of modern statistical inference. *Ann. Rev. Stat. Appl.* **4** 1–14.

FRASER, D. A. S. and REID, N. (1995). Ancillaries and third order significance. *Util. Math.* **47** 33–53. MR1330888

FRASER, D. A. S., REID, N. and SARTORI, N. (2016). Accurate directional inference for vector parameters. *Biometrika* **103** 625–639. MR3551788

FRASER, D. A. S., REID, N. and WU, J. (1999). A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika* **86** 249–264. MR1705367

FRASER, D. A. S., WONG, A. and SUN, Y. (2009). Three enigmatic examples and inference from likelihood. *Canad. J. Statist.* **37** 161–181. MR2531825

FRASER, D. A. S., REID, N., MARRAS, E. and YI, G. Y. (2010). Default priors for Bayesian and frequentist inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 631–654. MR2758239

FRASER, D. A. S., BÉDARD, M., WONG, A., LIN, W. and FRASER, A. M. (2016). Bayes, reproducibility and the quest for truth. *Statist. Sci.* **31** 578–590. MR3598740

GHOSH, M. (2011). Objective priors: An introduction for frequentists. *Statist. Sci.* **26** 187–202. MR2858380

GOLDSTEIN, M. (2006). Subjective Bayesian analysis: Principles and practice. *Bayesian Anal.* **1** 403–420. MR2221272

HANNIG, J. (2009). On generalized fiducial inference. *Statist. Sinica* **19** 491–544. MR2514173

HANNIG, J., IYER, H., LAI, R. C. S. and LEE, T. C. M. (2016). Generalized fiducial inference: A review and new results. *J. Amer. Statist. Assoc.* **111** 1346–1361. MR3561954

HARTMANN, M., HOSACK, G. R., HILLARY, R. M. and VANHATALO, J. (2017). Gaussian process framework for temporal dependence and discrepancy functions in Ricker-type population growth models. *Ann. Appl. Stat.* **11** 1375–1402. MR3709563

IZBICKI, R., LEE, A. B. and FREEMAN, P. E. (2017). Photo-z estimation: An example of nonparametric conditional density estimation under selection bias. *Ann. Appl. Stat.* **11** 698–724. MR3693543

KEELE, L. and QUINN, K. M. (2017). Bayesian sensitivity analysis for causal effects from 2 × 2 tables in the presence of unmeasured confounding with application to presidential campaign visits. *Ann. Appl. Stat.* **11** 1974–1997. MR3743285

LAPLACE, P. S. (1812). *Théorie Analytique des Probabilités*. Courcier, Paris.

MCCULLAGH, P. (2002). What is a statistical model? *Ann. Statist.* **30** 1225–1310. MR1936320

NEYMAN, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philos. Trans. R. Soc. A* **237** 333–380.

NEYMAN, J. and PEARSON, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. A* **231** 289–337.

OGDEN, H. E. (2017). On asymptotic validity of naive inference with an approximate likelihood. *Biometrika* **104** 153–164. MR3626485

PIERCE, D. A. and BELLIO, R. (2017). Modern likelihood-frequentist inference. *Int. Stat. Rev.* **85** 519–541. MR3723615

REID, N. and COX, D. R. (2015). On some principles of statistical inference. *Int. Stat. Rev.* **83** 293–308. MR3377082

REID, N. and SUN, Y. (2010). Assessing sensitivity to priors using higher order approximations. *Comm. Statist. Theory Methods* **39** 1373–1386. MR2753513

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.* **12** 1151–1172. MR0760681

SARTORI, N. (2003). Modified profile likelihoods in models with stratum nuisance parameters. *Biometrika* **90** 533–549. MR2006833

SCHWEDER, T. and HJORT, N. L. (2016). *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions. Cambridge Series in Statistical and Probabilistic Mathematics* **41**. Cambridge Univ. Press, New York. MR3558738

SIMOIU, C., CORBETT-DAVIES, S. and GOEL, S. (2017). The problem of infra-marginality in outcome tests for discrimination. *Ann. Appl. Stat.* **11** 1193–1216. MR3709557

STEIN, C. (1959). An example of wide discrepancy between fiducial and confidence intervals. *Ann. Math. Stat.* **30** 877–880. MR0125680

TAK, H., MANDEL, K., VAN DYK, D. A., KASHYAP, V. L., MENG, X.-L. and SIEMIGINOWSKA, A. (2017). Bayesian estimates of astronomical time delays between gravitationally lensed stochastic light curves. *Ann. Appl. Stat.* **11** 1309–1348. MR3709561

WASSERMAN, L. (2006). Frequentist Bayes is objective (comment on articles by Berger and by Goldstein). *Bayesian Anal.* **1** 451–456. MR2221280

XIE, M. and SINGH, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review. *Int. Stat. Rev.* **81** 3–39. MR3047496

The Annals of Applied Statistics
2018, Vol. 12, No. 2, 771–787
https://doi.org/10.1214/18-AOAS1140

FINGERPRINT SCIENCE

BY JOSEPH B. KADANE

Carnegie Mellon University

This paper examines the extent to which data support the source attributions made by fingerprint examiners. It challenges the assumption that each person's fingerprints are unique, but finds that evidence of persistence of an individual's fingerprints is better founded. The use of the AFIS (Automatic Fingerprint Identification System) is problematic, because the algorithms used are proprietary. Additionally, the databases used in conjunction with AFIS are incomplete and not public. Finally, and most crucially, the finding of similarities between the mark found at a crime scene and a fingerprint on file does not permit estimation of the number of persons in a given population who share those characteristics. Consequently, there is no scientific basis for a source attribution; whether phrased as a "match," as "individualization" or otherwise.

REFERENCES

ARKOWITZ, H. and LILLEFIELD, S. D. (2010). Why science tells us not to rely on eyewitness accounts: Eyewitness testimony is fickle. and, all too often, shockingly inaccurate. *Sci. Am.* Available at https://www.scientificamerican.com/article/do-the-eyes-have-it/.

CECIL, M. (2014). *Hoover's FBI and the Fourth Estate*: *The Campaign to Control the Press and the Bureau's Image*. Univ. Press of Kansas, Lawrence, KS.

COLE, S. A. (2005). More than zero: Accounting for error in latent fingerprint identification. *J. Crim. Law Criminol*. **95** 985–1078.

COLE, S. A. (2014). Individualization is dead, long live individualization! Reforms in reporting practices for fingerprint analysis in the United States. *Law Prob. Risk* **13** 117–150.

COMMITTEE ON ASSESSING THE NEEDS OF THE FORENSIC SCIENCE COMMUNITY (2009). *Strengthening the Forensic Sciences in the United States: A Path Forward*. The National Academies Press, Washington, DC.

COMMITTEE ON SCIENTIFIC ASSESSMANT OF BULLET LEAD ELEMENTAL COMPOSITION COMPARISON (2004). *Forensic Analysis Weighing Bullet Lead Evidence*. The National Academies Press, Washington, DC.

COMMITTEE TO REVIEW THE SCIENTIFIC EVIDENCE ON THE POLYGRAPH (2003). *The Polygraph and Lie Detection*. The National Academies Press, Washington, DC.

DEGROOT, M., FIENBERG, S. E. and KADANE, J. B. (1983). *Statistics and the Law*. Wiley, New York.

DEPARTMENT OF JUSTICE (2006). A review of the FBI's handling of the Brandon Mayfield case. Unclassified executive summary. Available at https://oig.justice.gov/special/s0601/final.pdf.

DEPARTMENT OF JUSTICE (2018). Approved uniform language for testimony and reports for the forensic latent print discipline. Available at https://www.justice.gov/file/1037171/download.

Key words and phrases. Fingerprint uniqueness, fingerprint persistence, AIS, source attribution, individualization, match.

DOYLE, C. (1903). The adventure of the Norwood builder. *Strand Magazine* and *Collier's Magazine*, later published as Chapter 2 in "The Return of Sherlock Holmes."

DROR, I. E. and CHARLTON, D. (2006). Why experts make errors. *J. Forensic Identif.* **56** 600–616.

DROR, I. E., CHARLTON, D. and PERON, A. E. (2006). Contextual information renders experts vulnerable to making erroneous identifications. *Forensic Sci. Int.* **156** 74–78.

DROR, I. E. and ROSENTHAL, R. (2008). Meta-analytically quantifying the reliability and biasability of forensic experts. *J. Forensic Sci.* **53** 900–903.

DROR, I. E., CHAMPOD, C., LANGENBURG, G., CHARLTON, D., HUNT, H. and ROSENTHAL, R. (2011). Cognitive issues in fingerprint analysis: Inter- and intra-expert consistency and the effect of a 'target' comparison. *Forensic Sci. Int.* **208** 10–17.

DROR, I. E., THOMPSON, W. C., MEISSNER, C. A., KORNFIELD, I., KRANE, D., SAKS, M. and RISINGER, M. (2015). Lettter to the editor—Context management toolbox: A linear sequential unmaksing approach for minimizing cognitive bias in forensic decision making. *J. Forensic Sci.* **60** 1111–1112.

ERICSSON, K. and SIMON, H. A. (1980). Verbal reports as data. *Psychol. Rev.* **87** 215–251.

ERICSSON, K. and SIMON, H. A. (1993). *Protocol Analysis, Verbal Reports as Data*. MIT Press, Cambridge, MA.

FAULD, H. (1880). On the skin-furrows of the hand. *Letter in Nature*, October 8, 1880. Available at http://www.clpex.com/Articles/History/Faulds1880.htm.

FEDERAL BUREAU OF INVESTIGATION (1985). *The Science of Fingerprints: Classification and Uses*. Available at www.gutenberg.org/ebooks/19022.

FIENBERG, S. E., ed. (1989). *The Evolving Role of Statistical Assessments as Evidence in Court*. Springer, Berlin.

FIENBERG, S. E. (2005). Forensic science: The nexus of science and the law. In *Arthur M. Sackler Colloquia of the National Academy of Science*, *November* 16–18. Available at http://www.nasonline.org/programs/sackler-colloquia/completed_colloquia/forensic-science-the-nexus-of-science-and-the-law.html.

FIENBERG, S. E. and STRAF, M. (1982). Statistical assessments as evidence. *J. Roy. Statist. Soc. Ser. A* **145** 410–421.

FLORIDA V. S. HAYES (2015). 15CJ109, August 24, 2015. Tampa, Florida.

GALTON, F. (1888). Letter in *Nature*: Personal identification and description. May 25, 1888.

GALTON, F. (1892). *Finger Prints*. MacMillan, London.

HABER, L. and HABER, R. N. (2008). Scientific valitation of fingerprint evidence under Daubert. *Law Prob. Risk* **7** 87–109.

HERSCHEL, W. (1880). Skin furrows of the hand. *Letter in Nature*, November 28, 1880. Available at http://galton.org/fingerprints/herschel-1880-nature-furrows.pdf.

INTERNATIONAL ASSOCIATION FOR IDENTIFICATION (1979). Resolution 1979-7.

INTERNATIONAL ASSOCIATION FOR IDENTIFICATION (1980). Resolution 1980-5.

INTERNATIONAL ASSOCIATION FOR IDENTIFICATION (2007). IAI position concerning latent print identification. November 29, 2007. Available at http://onin.com/fp/IAI_Position_Statement_11-29-07.pdf.

INTERNATIONAL ASSOCIATION FOR IDENTIFICATION (2010). Resolution 2010-18.

KADANE, J. B. and KOEHLER, J. J. (2018). Certainty and uncertainty in reporting fingerprint evidence. *Daedalus* (in press).

LANGENBURG, G. (2011). *The Fingerprint Sourcebook* 14: *Scientific Research Supporting the Foundations of Friction Ridge Examinations*. National Institute of Justice, Washington, DC.

LANGENBURG, G. M. (2012). A critical analysis and study of the ACE-V process. Ph.D. thesis, Univ. Lausanne.

LANGENBURG, G., CHAMPOD, C. and GENESSAY, T. (2012). Informing the judments of fingerprint analysts using quality etrics and statistical assessment tools. *Forensic Sci. Int.* **219** 183–198.

LARKIN, J., MCDERMOTT, J., SIMON, D. P. and SIMON, H. A. (1980). Expert and novice performance in solving physics problems. *Science* **208** 1335–1342.

LIU, Y. and SRIHARI, S. N. (2009). A computational discriminability analysis on twin fingerprints. In *Proceedings of the Third International Workshop on Computational Forensics*. Springer, The Hague.

MACEO, A. (2009). Friction ridge skin: Morphogenesis and overview. In *Wiley Encyclopedia of Forensic Science*. DOI:10.1002/97804700061589.fsa358.

MOSES, K. R. (2011). *The Fingerprint Sourcebook* 6: *Automated Fingerprint Identification*. National Institute of Justice, Washington, DC. Available at https://www.ncjrs.gov/pdffiles1/nij/225320.pdf.

MUSTONEN, V., HAKKARAINEN, K. and TUUAINEN, J. (2015). Discrepancies in expert decision-making in forensic fingerprint examination. *Forensic Sci. Int.* **254** 215–226.

NATIONAL INSTITUTE OF JUSTICE (2011). *The Fingerprint Sourcebook*. National Institute of Justice, Washington, DC. Available at https://www.ncjrs.gov/pdffiles1/nij/225320.pdf.

NATIONAL RESEARCH COUNCIL (2009). *Strengthening Forensic Science in the United States: A Path Forward*. The National Academies Press, Washington, DC.

NEUMANN, C., CHAMPOD, C., YOO, M., GENESSAY, T. and LANGENBERG, G. (2013). Improving the understanding and the reliability of the concept of "sufficiency" in friction ridge examination. Award Number 2010-DN-BX-K267, NIJ. Available at https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=271468.

PACHEO, I., CERCHAI, B. and STOILOFF, S. (2014). Miami–Dade study of the ACE-V process: Accuracy and precision in latent fingerprint examinations. National Institute of Justice Final Report Document #248534, Miami–Dade Police Department.

PRESIDENT'S COUNCIL OF ADVSORS ON SCIENCE AND TECHNOLOGY (2016). Report to the President, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-comparison Methods.

ROEDER, K., ESCOBAR, M., KADANE, J. B. and BALZAS, I. (1998). Measuring heterogeneity in forensic databases using hierarchical Bayes models. *Biometrika* **85** 269–287.

SCIENTIFIC WORKING GROUP ON FRICTION RIDGE ANALYSIS. Standards for examining friction ridge impressions and resulting conclusions. Version 1 (Latent/Tenprint). Available at http://clpex.com/swgfast/documents/examinations-conclusions/111026_Examinations-Conclusions_1.0.pdf.

SPECTER, M. (2002). Do fingerprints lie? The gold standard of forensic evidence is now being challenged. *New Yorker* May 27.

SRIHARI, S. N., SRINIVASAN, H. and FANG, G. (2008). Discriminability of fingerprints of twins. *J. Forensic Identif.* **58** 109–127.

SWOFFORD, H. J. (2015). Use of the term "Identification" in latent print. Information Paper.

TANGEN, J. M., THOMPSON, M. B. and MCCARTHY, D. J. (2011). Identifying fingerprint expertise. *Psychol. Sci.* **22** 995–997.

THE DETAIL (2006). Available at http://www.clpex.com/legacy/TheDetail/200-299/TheDetail254.htm. Accessed 28 July 2016.

THOMPSON, M. B., TANGEN, J. M. and MCCARTHY, D. J. (2013). Expertise in fingerprint identification. *J. Forensic Sci.* **58** 1519–1530.

THOMPSON, M. B., TANGEN, J. M. and MCCARTHY, D. J. (2014). Human matching performance of genuine crime scene latent fingerprints. *Law Hum. Behav.* **38** 84–93.

TWAIN, M. (1883). A thumb print and what became of it. Later published as Chapter 31 in "Life on the Mississippi."

TWAIN, M. (1894). *The Tragedy of Pudd'nhead Wilson and the Comedy of Those Extraordinary Twins*. American Publishing Company, Hartford, CT.

ULERY, B. T., HICKLIN, A. R., BUSCAGLIA, J. and ROBERTS, M. A. (2011). Accuracy and reliability of forensic latent fingerprint decisions. *Proc. Natl. Acad. Sci. USA* **108** 7733–7738.

YOON, S. and JAIN, A. K. (2015). Longitudinal study of fingerprint recognition. *Proc. Natl. Acad. Sci. USA* **112** 8555–8560.

ZABELL, S. L. (2005). Fingerprint evidence. *J. Law Policy* **13** 143–179.

UNITED STATES CENSUS BUREAU (2017). World population, more than 7.5 billion on May 25, 2017. Available at https://www.census.gov/popclock/world.

The Annals of Applied Statistics
2018, Vol. 12, No. 2, 788–814
https://doi.org/10.1214/18-AOAS1180

STATISTICAL MODELING AND ANALYSIS OF TRACE ELEMENT CONCENTRATIONS IN FORENSIC GLASS EVIDENCE

BY KAREN D. H. PAN[1] AND KAREN KAFADAR[2]

University of Virginia

The question of the validity of procedures used to analyze forensic evidence was raised many years ago by Stephen Fienberg, most notably when he chaired the National Academy of Sciences' Committee that issued the report *The Polygraph and Lie Detection* [National Research Council (2003) The National Academies Press]; his role in championing this cause and drawing other statisticians to these issues continued throughout his life. We investigate the validity of three standards related to different test methods for forensic comparison of glass (micro X-ray fluorescence (μ-XRF) spectrometry, ICP-MS, LA-ICP-MS], all of which include a series of recommended calculations from which "it may be concluded that [the samples] did not originate from the same source." Using publicly available data and data from other sources, we develop statistical models based on estimates of means and covariance matrices of the measured trace element concentrations recommended in these standards, leading to population-based estimates of error rates for the comparison procedures stated in the standards. Our results therefore do not depend on internal comparisons between pairs of glass samples, the representativeness of which cannot be guaranteed: our results apply to any collection of glass samples that have been or can be measured via these technologies. They suggest potentially higher false positive rates than have been reported, and we propose alternative methods that will ensure lower error rates.

REFERENCES

ASTM INTERNATIONAL (2012). *ASTM E2330-12 Standard Test Method for Determination of Concentrations of Elements in Glass Samples Using Inductively Coupled Plasma Mass Spectrometry (ICP-MS) for Forensic Comparisons*. Retrieved from https://www.astm.org/Standards/E2330.htm. DOI:10.1520/E2330-12.

ASTM INTERNATIONAL (2013). *ASTM E2926-13 Standard Test Method for Forensic Comparison of Glass Using Micro X-ray Fluorescence (μ-XRF) Spectrometry*. Retrieved from https://www.astm.org/Standards/E2926.htm. DOI:10.1520/E2926.

ASTM INTERNATIONAL (2016). *ASTM E2927-16e1 Standard Test Method for Determination of Trace Elements in Soda-Lime Glass Samples Using Laser Ablation Inductively Coupled Plasma Mass Spectrometry for Forensic Comparisons*. Retrieved from https://www.astm.org/Standards/E2927.htm. DOI:10.1520/E2927-16E01.

BRILLINGER, D. R. and TUKEY, J. W. (1985). Spectrum analysis in the presence of noise: Some issues and examples. In *The Collected Works of John W. Tukey II. Time Series*: 1965–1984 (D. R. Brillinger, ed.) 1001–1141. Wadsworth, Monterey, CA.

Key words and phrases. Robust methods, exploratory data analysis, multivariate lognormal distribution, covariance matrix, standard errors, error rates, ROC curve.

DETTMAN, J. R., CASSABAUM, A. A., SAUNDERS, C. P., SNYDER, D. L. and BUSCAGLIA, J. (2014). Forensic discrimination of copper wire using trace element concentrations. *Anal. Chem.* **86** 8176–8182.

DORN, H., RUDDELL, D. E., HEYDON, A. and BURTON, B. D. (2015). Discrimination of float glass by LA-ICP-MS: Assessment of exclusion criteria using casework samples. *Can. Soc. Forensic Sci. J.* **48** 85–96. DOI:10.1080/00085030.2015.1019224.

GABEL-CINO, J. (2017). Expert witnesses and lawyers: Can we all get along? Presentation to the Second Annual Conference of the National Center for Forensic Science, Orlando, Florida, October 17, 2017.

GENZ, A., BRETZ, F., MIWA, T., MI, X., LEISCH, F., SCHEIPL, F., BORNKAMP, B., MAECHLER, M. and HOTHORN, T. (2016). Multivariate normal and t distributions, R package mvtnorm. R package version 1.0-5.

GIANNELLI, P. C. (2010). Comparative bullet lead analysis: A retrospective (September 1, 2011). *Crim. Law Bull.* **47** 306. Case Legal Studies Research Paper No. 2011-21.

KAFADAR, K. and EBERHARDT, K. R. (1983). Statistical analysis of some gas chromatographic measurements. *NBS J. Res.* **88** 37–46.

KAFADAR, K. and EBERHARDT, K. R. (1984). Some basic statistical methods for chromatographic data. In *Advances in Chromatography*, *Chapter* 1 (J. C. Giddings, E. Grushka, J. Cazes and P. R. Brown, eds.) **24** 1–34. Dekker, New York.

KOONS, R. D. (2003). Personal communication to K. Kafadar.

KOONS, R. D. and BUSCAGLIA, J. A. (2001). Interpretation of glass composition measurements: The effects of match criteria on discrimination capability. *J. Forensic Sci.* **47** 505–512.

KOONS, R. D. and BUSCAGLIA, J. (2005). Forensic significance of bullet lead compositions. *J. Forensic Sci.* **50** 341–351.

NATIONAL RESEARCH COUNCIL (2003). *The Polygraph and Lie Detection* (*Committee to Review the Scientific Evidence on the Polygraph*, *Division of Behavioral and Social Sciences and Education*). The National Academies Press, Washington, DC. DOI:10.17226/10420.

NATIONAL RESEARCH COUNCIL (2004). *Forensic Analysis*: *Weighing Bullet Lead Evidence* (K. O. MacFadden, Chair). The National Academies Press, Washington, DC.

NATIONAL RESEARCH COUNCIL (2009). *Strengthening Forensic Science in the United States: A Path Forward* (The Honorable H. T. Edwards and C. Gatsonis, Co-Chairs). The National Academies Press, Washington, DC. Available at http://books.nap.edu/catalog.php?record_id=12589.

PAN, K. D. and KAFADAR, K. (2018). Supplement to "Statistical modeling and analysis of trace element concentrations in forensic glass evidence." DOI:10.1214/18-AOAS1180SUPP.

RIPLEY, B. (2015). MASS: Support functions and datasets for venables and Ripley's MASS. R package version 7.3-45.

ROUSSEEUW, P. and VAN DREISSEN, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **41** 212–223.

SPIEGELMAN, C. H. and KAFADAR, K. (2006). Data integrity and the scientific method: The case of bullet lead data as forensic evidence. *Chance* **19** 17–26 (with discussion). MR2247019

TREJOS, T., KOONS, R., WEIS, P., BECKER, S., BERMAN, T., DALPE, C., DUECKING, M., BUSCAGLIA, J., ECKERT-LUMSDON, T., ERNST, T., HANLON, C., HEYDON, A., MOONEY, K., NELSON, R., OLSSON, K., SCHENK, E., PALENIK, C., POLLOCK, E. C., RUDELL, D., RYLAND, S., TARIFA, A., VALADEZ, M., VAN ES, A., ZDANOWICZ, V. and ALMIRALL, J. (2013). Forensic analysis of glass by μ-XRF, SN-ICP-MS, LA-ICP-MS and LA-ICP-OES: Evaluation of the performance of different criteria for comparing elemental composition. *J. Anal. At. Spectrom.* **28** 1270–1282. DOI:10.1039/c3ja50128k.

WEIS, P., DÜCKLING, M., WATZKE, P., MENGES, S. and BECKER, S. (2011). Establishing a match criterion in forensic comparison analysis of float glass using laser ablation inductively coupled plasma mass spectrometry. *J. Anal. At. Spectrom.* **26** 1273–1284.

The Annals of Applied Statistics
2018, Vol. 12, No. 2, 815–845
https://doi.org/10.1214/18-AOAS1164

LOGLINEAR MODEL SELECTION AND HUMAN MOBILITY[1]

BY ADRIAN DOBRA AND REZA MOHAMMADI

University of Washington and University of Amsterdam

Methods for selecting loglinear models were among Steve Fienberg's research interests since the start of his long and fruitful career. After we dwell upon the string of papers focusing on loglinear models that can be partly attributed to Steve's contributions and influential ideas, we develop a new algorithm for selecting graphical loglinear models that is suitable for analyzing hyper-sparse contingency tables. We show how multi-way contingency tables can be used to represent patterns of human mobility. We analyze a dataset of geolocated tweets from South Africa that comprises 46 million latitude/longitude locations of 476,601 Twitter users that is summarized as a contingency table with 214 variables.

REFERENCES

AGRESTI, A. (1990). *Categorical Data Analysis*. Wiley, New York. MR1044993

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

BALDI, P., BRUNAK, S., CHAUVIN, Y., ANDERSEN, C. A. and NIELSEN, H. (2000). Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics* **16** 412–424.

BALTAZAR, C. S., HORTH, R., INGUANE, C., SATHANE, I., CÉSAR, F., RICARDO, H., BOTÃO, C., AUGUSTO, Â., COOLEY, L., CUMMINGS, B., RAYMOND, H. F. and YOUNG, P. W. (2015). HIV prevalence and risk behaviors among Mozambicans working in South African mines. *AIDS Behav*. **19** 59–67.

BECKER, R., CÁCERES, R., HANSON, K., ISAACMAN, S., LOH, J. M., MARTONOSI, M., ROWLAND, J., URBANEK, S., VARSHAVSKY, A. and VOLINSKY, C. (2013). Human mobility characterization from cellular network data. *Commun*. *ACM* **56** 74–82.

BESAG, J. (1975). Statistical analysis of non-lattice data. *J. R. Stat. Soc*., *Ser. D Stat*. **24** 179–195.

BESAG, J. (1977). Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika* **64** 616–618. MR0494640

BHATTACHARYA, A. and DUNSON, D. B. (2012). Simplex factor models for multivariate unordered categorical data. *J. Amer. Statist. Assoc*. **107** 362–377. MR2949366

BISHOP, Y. M. M., FIENBERG, S. E. and HOLLAND, P. W. (1975). *Discrete Multivariate Analysis*: *Theory and Practice*. MIT Press, Cambridge, MA. With the collaboration of Richard J. Light and Frederick Mosteller. MR0381130

BROCKMANN, D., HUFNAGEL, L. and GEISEL, T. (2006). The scaling laws of human travel. *Nature* **439** 462–465.

CALABRESE, F., DIAO, M., LORENZO, G. D., FERREIRA JR., J. and RATTI, C. (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transp*. *Res*., *Part C*, *Emerg*. *Technol*. **26** 301–313.

Key words and phrases. Contingency tables, model selection, human mobility, graphical models, Bayesian structural learning, birth–death processes, pseudo-likelihood.

CANALE, A. and DUNSON, D. B. (2011). Bayesian kernel mixtures for counts. *J. Amer. Statist. Assoc.* **106** 1528–1539. MR2896854

CAPPÉ, O., ROBERT, C. P. and RYDÉN, T. (2003). Reversible jump, birth-and-death and more general continuous time Markov chain Monte Carlo samplers. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 679–700. MR1998628

CHENG, Y. and LENKOSKI, A. (2012). Hierarchical Gaussian graphical models: Beyond reversible jump. *Electron. J. Stat.* **6** 2309–2331. MR3020264

CLYDE, M. and GEORGE, E. I. (2004). Model uncertainty. *Statist. Sci.* **19** 81–94. MR2082148

DELLAPORTAS, P. and FORSTER, J. J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika* **86** 615–633. MR1723782

DELLAPORTAS, P. and TARANTOLA, C. (2005). Model determination for categorical data with factor level merging. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 269–283. MR2137325

DESCOMBES, X., MINLOS, R. and ZHIZHINA, E. (2009). Object extraction using a stochastic birth-and-death dynamics in continuum. *J. Math. Imaging Vision* **33** 347–359. MR2480967

DOBRA, A. and LENKOSKI, A. (2011). Copula Gaussian graphical models and their application to modeling functional disability data. *Ann. Appl. Stat.* **5** 969–993. MR2840183

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

DOBRA, A. and MASSAM, H. (2010). The mode oriented stochastic search (MOSS) algorithm for log-linear models with conjugate priors. *Stat. Methodol.* **7** 240–253. MR2643600

DOBRA, A. and MOHAMMADI, R. (2018). Supplement to "Loglinear model selection and human mobility." DOI:10.1214/18-AOAS1164SUPP.

DOBRA, A., WILLIAMS, N. E. and EAGLE, N. (2015). Spatiotemporal detection of unusual human population behavior using mobile phone data. *PLoS ONE* **10** 1–20.

DOBRA, A., BÄRNIGHAUSEN, T., VANDORMAEL, A. and TANSER, F. (2017). Space-time migration patterns and risk of HIV acquisition in rural South Africa. *AIDS* **31** 37–145.

DONATO, K. M. (1993). Current trends and patterns of female migration: Evidence from Mexico. *Int. Migr. Rev.* **27** 748–771.

DRTON, M. and MAATHUIS, M. H. (2017). Structure learning in graphical modeling. *Annu. Rev. Statist. Appl.* **4** 365–393.

DUNSON, D. B. and XING, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *J. Amer. Statist. Assoc.* **104** 1042–1051. MR2562004

DURAND, J., KANDEL, W., PARRADO, E. A. and MASSEY, D. S. (1996). International migration and development in Mexican communities. *Demography* **33** 249–264.

EDWARDS, D. and HAVRÁNEK, T. (1985). A fast procedure for model search in multidimensional contingency tables. *Biometrika* **72** 339–351. MR0801773

FIENBERG, S. E. (1970). The analysis of multidimensional contingency tables. *Ecology* **51** 419–433.

FIENBERG, S. E. (1980). *The Analysis of Cross-Classified Categorical Data*, 2nd ed. MIT Press, Cambridge, MA. MR0623082

FIENBERG, S. E. and RINALDO, A. (2007). Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation. *J. Statist. Plann. Inference* **137** 3430–3445. MR2363267

FIENBERG, S. E. and RINALDO, A. (2012). Maximum likelihood estimation in log-linear models. *Ann. Statist.* **40** 996–1023. MR2985941

GAMAL-ELDIN, A., DESCOMBES, X. and ZERUBIA, J. (2010). Multiple birth and cut algorithm for point process optimization. In 2010 *Sixth International Conference on Signal-Image Technology and Internet-Based Systems* (*SITIS*) 35–42. IEEE, Los Alamitos, CA.

GAMAL-ELDIN, A., DESCOMBES, X., CHARPIAT, G. and ZERUBIA, J. (2011). A fast multiple birth and cut algorithm using belief propagation. In 2011 18*th IEEE International Conference on Image Processing* 2813–2816. IEEE, Los Alamitos, CA.

GONZALEZ, M. C., HIDALGO, C. A. and BARABASI, A.-L. (2008). Understanding individual human mobility patterns. *Nature* **453** 779–782.

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

GUERZHOY, M. and HERTZMANN, A. (2014). Learning latent factor models of travel data for travel prediction and analysis. In *Advances in Artificial Intelligence*. *Lecture Notes in Computer Science* **8436** 131–142. Springer, Cham. MR3218638

HARRIS, J. R. and TODARO, M. P. (1970). Migration, unemployment and development: A two-sector analysis. *Am*. *Econ*. *Rev*. **60** 126–142.

HOFF, P. D. (2008). Multiplicative latent factor models for description and prediction of social networks. *Comput*. *Math*. *Organ*. *Theory* **15** Art. ID 261.

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

HØJSGAARD, S., EDWARDS, D. and LAURITZEN, S. (2012). *Graphical Models with R*. Springer, New York. MR2905395

IMAI, K. (2017). *Quantitative Social Science*: *An Introduction*. Princeton Univ. Press, Princeton, NJ.

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

JURDAK, R., ZHAO, K., LIU, J., ABOUJAOUDE, M., CAMERON, M. and NEWTH, D. (2015). Understanding human mobility from Twitter. *PLoS ONE* **10** 1–16.

KUNIHAMA, T. and DUNSON, D. B. (2013). Bayesian modeling of temporal dependence in large sparse contingency tables. *J*. *Amer*. *Statist*. *Assoc*. **108** 1324–1338. MR3174711

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

LEETARU, K., WANG, S., CAO, G., PADMANABHAN, A. and SHOOK, E. (2013). Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday* **18**. Available at http://firstmonday.org/ojs/index.php/fm/article/view/4366/3654.

LENKOSKI, A. and DOBRA, A. (2011). Computational aspects related to inference in Gaussian graphical models with the G-Wishart prior. *J*. *Comput*. *Graph*. *Statist*. **20** 140–157. Supplementary material available online. MR2816542

LETAC, G. and MASSAM, H. (2012). Bayes factors and the geometry of discrete hierarchical log-linear models. *Ann*. *Statist*. **40** 861–890. MR2985936

MADIGAN, D. and RAFTERY, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J*. *Amer*. *Statist*. *Assoc*. **89** 1535–1546.

MADIGAN, D. and YORK, J. (1995). Bayesian graphical models for discrete data. *Int*. *Stat*. *Rev*. **63** 215–232.

MADIGAN, D. and YORK, J. C. (1997). Bayesian methods for estimation of the size of a closed population. *Biometrika* **84** 19–31. MR1450189

MADIGAN, D., RAFTERY, A. E., VOLINSKY, C. and HOETING, J. (1996). Bayesian model averaging. In *Proceedings of the AAAI Workshop on Integrating Multiple Learned Models* 77–83.

MASSAM, H., LIU, J. and DOBRA, A. (2009). A conjugate prior for discrete hierarchical log-linear models. *Ann*. *Statist*. **37** 3431–3467. MR2549565

MASSEY, D. S. (1990). Social structure, household strategies, and the cumulative causation of migration. *Popul*. *Index* **56** 3–26.

MASSEY, D. S. and ESPINOSA, K. E. (1997). What's driving Mexico–U.S. migration? A theoretical, empirical, and policy analysis. *Am*. *J*. *Sociol*. **102** 939–999.

MASSEY, D. S., ARANGO, J., HUGO, G., KOUAOUCI, A., PELLEGRINO, A. and TAYLOR, J. E. (1993). Theories of international migration: A review and appraisal. *Popul*. *Dev*. *Rev*. **19** 431–466.

MASSEY, D. S., WILLIAMS, N., AXINN, W. G. and GHIMIRE, D. (2010). Community services and out-migration. *Int*. *Migr*. **48** 1–41.

MOHAMMADI, A. and DOBRA, A. (2017). The R package BDgraph for Bayesian structure learning in graphical models. *ISBA Bull*. **4** 11–16.

MOHAMMADI, A., MASSAM, H. and LETAC, G. (2017). The ratio of normalizing constants for Bayesian graphical Gaussian model selection. Preprint. Available at arXiv:1706.04416.

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

MOHAMMADI, R. and WIT, E. C. (2017). BDgraph: An R package for Bayesian structure learning in graphical models. Preprint. Available at arXiv:1501.05108v4.

MOHAMMADI, R. and WIT, E. C. and DOBRA, A. (2018). BDgraph: Bayesian structure learning in graphical models using birth–death MCMC. R package version 2.49.

MOHAMMADI, A., ABEGAZ, F., VAN DEN HEUVEL, E. and WIT, E. C. (2017). Bayesian modelling of Dupuytren disease by using Gaussian copula graphical models. *J. R. Stat. Soc. Ser. C. Appl. Stat*. **66** 629–645. MR3632345

NARDI, Y. and RINALDO, A. (2012). The log-linear group-lasso estimator and its asymptotic properties. *Bernoulli* **18** 945–974. MR2948908

NEUBAUER, G., HUBER, H., VOGL, A., JAGER, B., PREINERSTORFER, A., SCHIRNHOFER, S., SCHIMAK, G. and HAVLIK, D. (2015). On the volume of geo-referenced tweets and their relationship to events relevant for migration tracking. In *Environmental Software Systems. Infrastructures, Services and Applications*: 11*th IFIP WG* 5.11 *International Symposium, ISESS* 2015, *Melbourne, VIC, Australia, March* 25–27, 2015. *Proceedings* (R. Denzer, R. M. Argent, G. Schimak and J. Hřebíček, eds.) 520–530. Springer, Cham.

OPENMP ARCHITECTURE REVIEW BOARD (2008). OpenMP application program interface version 3.0.

PENSAR, J., NYMAN, H., NIIRANEN, J. and CORANDER, J. (2017). Marginal pseudo-likelihood learning of discrete Markov network structures. *Bayesian Anal*. **12** 1195–1215. MR3724983

PRESTON, C. (1975). Spatial birth-and-death processes. *Bull. Inst. Int. Stat*. **46** 371–391, 405–408 (1975). With discussion. MR0474532

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

RAYMER, J., ABEL, G. and SMITH, P. W. F. (2007). Combining census and registration data to estimate detailed elderly migration flows in England and Wales. *J. Roy. Statist. Soc. Ser. A* **170** 891–908. MR2408983

RAYMER, J., WIŚNIOWSKI, A., FORSTER, J. J., SMITH, P. W. F. and BIJAK, J. (2013). Integrated modeling of European migration. *J. Amer. Statist. Assoc*. **108** 801–819. MR3174664

SCOTT, J. G. and CARVALHO, C. M. (2008). Feature-inclusion stochastic search for Gaussian graphical models. *J. Comput. Graph. Statist*. **17** 790–808. MR2649067

SMAPP (2017). smappR package: Tools for analysis of Twitter data, Social Media and Participation, New York University. Available at https://github.com/SMAPPNYU/smappR.

SMITH, P. W. F., RAYMER, J. and GIULIETTI, C. (2010). Combining available migration data in England to study economic activity flows over time. *J. Roy. Statist. Soc. Ser. A* **173** 733–753. MR2759963

STARK, O. and BLOOM, D. E. (1985). The new economics of labor migration. *Am. Econ. Rev.* **75** 173–178.

STARK, O. and TAYLOR, J. E. (1985). Migration incentives, migration types: The role of relative deprivation. *Econ. J.* **101** 1163–1178.

STOPHER, P. R. and GREAVES, S. P. (2007). Household travel surveys: Where are we going? *Transp. Res., Part A Policy Pract*. **41** 367–381.

TARANTOLA, C. (2004). MCMC model determination for discrete graphical models. *Stat. Model*. **4** 39–61. MR2037813

TATEM, A. J. (2014). Mapping population and pathogen movements. *Int. Health* **6** 5–11.

TAYLOR, J. E. (1987). Undocumented Mexico–U.S. migration and the returns to households in rural Mexico. *Am. J. Agric. Econ.* **69** 616–638.

TODARO, M. P. (1969). A model of labor migration and urban unemployment in less developed countries. *Am. Econ. Rev.* **59** 138–148.

TODARO, M. P. and MARUSZKO, L. (1987). Illegal immigration and U.S. immigration reform: A conceptual framework. *Popul. Dev. Rev.* **13** 101–114.

TSAMARDINOS, I., BROWN, L. E. and ALIFERIS, C. F. (2006). The max–min hill-climbing Bayesian network structure learning algorithm. *Mach. Learn.* **65** 31–78.

TWITTER, INC. (2017). Twitter REST APIs. Available at https://dev.twitter.com/rest/public.

VANWEY, L. K. (2005). Land ownership as a determinant of international and internal migration in Mexico and internal migration in Thailand. *Int. Migr. Rev.* **39** 141–172.

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

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

WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, Chichester. MR1112133

WILLIAMS, N. (2009). Education, gender, and migration in the context of social change. *Soc. Sci. Res.* **38** 883–896.

WILLIAMS, N. E., THOMAS, T. A., DUNBAR, M., EAGLE, N. and DOBRA, A. (2015). Measures of human mobility using mobile phone records enhanced with GIS data. *PLoS ONE* **10** 1–16.

WOLF, J., OLIVEIRA, M. and THOMPSON, M. (2003). Impact of underreporting on mileage and travel time estimates: Results from global positioning system-enhanced household travel survey. *Transp. Res. Rec.* **1854** 189–198.

The Annals of Applied Statistics
2018, Vol. 12, No. 2, 846–876
https://doi.org/10.1214/18-AOAS1169

ON THE USE OF BOOTSTRAP WITH VARIATIONAL INFERENCE: THEORY, INTERPRETATION, AND A TWO-SAMPLE TEST EXAMPLE

BY YEN-CHI CHEN, Y. SAMUEL WANG AND ELENA A. EROSHEVA

University of Washington

Variational inference is a general approach for approximating complex density functions, such as those arising in latent variable models, popular in machine learning. It has been applied to approximate the maximum likelihood estimator and to carry out Bayesian inference, however, quantification of uncertainty with variational inference remains challenging from both theoretical and practical perspectives. This paper is concerned with developing uncertainty measures for variational inference by using bootstrap procedures. We first develop two general bootstrap approaches for assessing the uncertainty of a variational estimate and the study the underlying bootstrap theory in both fixed- and increasing-dimension settings. We then use the bootstrap approach and our theoretical results in the context of mixed membership modeling with multivariate binary data on functional disability from the National Long Term Care Survey. We carry out a two-sample approach to test for changes in the repeated measures of functional disability for the subset of individuals present in 1989 and 1994 waves.

REFERENCES

AIROLDI, E., BLEI, D., XING, E. and FIENBERG, S. (2005). A latent mixed membership model for relational data. In *Proceedings of the* 3*rd International Workshop on Link Discovery* 82–89. ACM, New York.

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

AIROLDI, E. M., BLEI, D. M., EROSHEVA, E. A. and FIENBERG, S. E. (2015). Introduction to mixed membership models and methods. In *Handbook of Mixed Membership Models and Their Applications*. *Chapman & Hall/CRC Handb*. *Mod*. *Stat*. *Methods* 3–13. CRC Press, Boca Raton, FL. MR3380022

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

BABU, G. J. and SINGH, K. (1983). Inference on means using the bootstrap. *Ann*. *Statist*. **11** 999–1003. MR0707951

BERRY, A. C. (1941). The accuracy of the Gaussian approximation to the sum of independent variates. *Trans*. *Amer*. *Math*. *Soc*. **49** 122–136. MR0003498

BICKEL, P., CHOI, D., CHANG, X. and ZHANG, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Ann*. *Statist*. **41** 1922–1943. MR3127853

Key words and phrases. Variational inference, bootstrap, mixed membership model, increasing dimension, two-sample test.

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

BLEI, D. M., KUCUKELBIR, A. and MCAULIFFE, J. D. (2017). Variational inference: A review for statisticians. *J. Amer. Statist. Assoc*. **112** 859–877. MR3671776

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

BOX, G. E. P. (1976). Science and statistics. *J. Amer. Statist. Assoc*. **71** 791–799. MR0431440

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

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist*. **41** 2786–2819. MR3161448

DAMIANOU, A., TITSIAS, M. K. and LAWRENCE, N. D. (2011). Variational Gaussian process dynamical systems. In *Advances in Neural Information Processing Systems* 2510–2518.

DAMIANOU, A. C., TITSIAS, M. K. and LAWRENCE, N. D. (2016). Variational inference for latent variables and uncertain inputs in Gaussian processes. *J. Mach. Learn. Res*. **17** Paper No. 42. MR3491136

DOUGLAS, J. (1997). Joint consistency of nonparametric item characteristic curve and ability estimation. *Psychometrika* **62** 7–28. MR1439472

EFRON, B. (1982a). *The Jackknife, the Bootstrap and Other Resampling Plans*. *CBMS-NSF Regional Conference Series in Applied Mathematics* **38**. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. MR0659849

EFRON, B. (1982b). Bootstrap methods: Another look at the jackknife. In *Breakthroughs in Statistics* 569–593.

EROSHEVA, E. A., FIENBERG, S. E. and JOUTARD, C. (2007). Describing disability through individual-level mixture models for multivariate binary data. *Ann. Appl. Stat*. **1** 502–537. MR2415745

EROSHEVA, E. A. and WHITE, T. (2006). Operational definition of chronic disability in the national long term care survey: Problems and suggestions. Working Paper.

ESSEEN, C.-G. (1942). On the Liapounoff limit of error in the theory of probability. *Ark. Mat. Astron. Fys*. **28A** 19. MR0011909

FAN, J. and ZHOU, W.-X. (2016). Guarding against spurious discoveries in high dimensions. *J. Mach. Learn. Res*. **17** Paper No. 203. MR3580356

GHAHRAMANI, Z. and BEAL, M. J. (2000). Variational inference for Bayesian mixtures of factor analysers. In *Advances in Neural Information Processing Systems* 449–455.

HABERMAN, S. J. (1977). Maximum likelihood estimates in exponential response models. *Ann. Statist*. **5** 815–841. MR0501540

HALL, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York. MR1145237

HALL, P., ORMEROD, J. T. and WAND, M. P. (2011). Theory of Gaussian variational approximation for a Poisson mixed model. *Statist. Sinica* **21** 369–389. MR2796867

HALL, P., PHAM, T., WAND, M. P. and WANG, S. S. J. (2011). Asymptotic normality and valid inference for Gaussian variational approximation. *Ann. Statist*. **39** 2502–2532. MR2906876

HOROWITZ, J. L. (1997). Bootstrap methods in econometrics: Theory and numerical performance. *Econom. Soc. Monogr*. **28** 188–222.

JORDAN, M. I., GHAHRAMANI, Z., JAAKKOLA, T. S. and SAUL, L. K. (1999). An introduction to variational methods for graphical models. *Mach. Learn*. **37** 183–233.

KHAN, M. E., BOUCHARD, G., MURPHY, K. P. and MARLIN, B. M. (2010). Variational bounds for mixed-data factor analysis. In *Advances in Neural Information Processing Systems* 1108–1116.

KLAMI, A., VIRTANEN, S., LEPPÄAHO, E. and KASKI, S. (2015). Group factor analysis. *IEEE Trans. Neural Netw. Learn. Syst*. **26** 2136–2147. MR3453146

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

LEEB, H. and PÖTSCHER, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory* **21** 21–59. MR2153856

MAMMEN, E. (1989). Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *Ann. Statist.* **17** 382–400. MR0981457

MAMMEN, E. (1993). Bootstrap and wild bootstrap for high-dimensional linear models. *Ann. Statist.* **21** 255–285. MR1212176

MANTON, K. G., CORDER, L. S. and STALLARD, E. (1993). Estimates of change in chronic disability and institutional incidence and prevalence rates in the us elderly population from the 1982, 1984, and 1989 national long term care survey. *J. Gerontol.* **48** S153–S166.

NEYMAN, J. and SCOTT, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16** 1–32. MR0025113

O'HAGAN, A., MURPHY, T. B. and GORMLEY, I. C. (2015). On estimation of parameter uncertainty in model-based clustering. Preprint. Available at arXiv:1510.00551.

PORTNOY, S. (1984). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency. *Ann. Statist.* **12** 1298–1309. MR0760690

PORTNOY, S. (1985). Asymptotic behavior of M estimators of p regression parameters when p^2/n is large. II. Normal approximation. *Ann. Statist.* **13** 1403–1417. MR0811499

PORTNOY, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.* **16** 356–366. MR0924876

PRITCHARD, J. K., STEPHENS, M. and DONNELLY, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155** 945–959.

REDNER, R. A. and WALKER, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **26** 195–239. MR0738930

SINGH, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.* **9** 1187–1195. MR0630102

TROPP, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.* **12** 389–434. MR2946459

VAN DER VAART, A. W. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247

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

WANG, Y. and BLEI, D. M. (2017). Frequentist consistency of variational Bayes. Preprint. Available at arXiv:1705.03439.

WANG, Y. S. and EROSHEVA, E. A. (2015). Fitting mixed membership models using mixedmem.

WANG, Y. S., MATSUEDA, R. L. and EROSHEVA, E. A. (2017). A variational EM method for mixed membership models with multivariate rank data: An analysis of public policy preferences. *Ann. Appl. Stat.* **11** 1452–1480. MR3709566

WASSERMAN, L. (2006). *All of Nonparametric Statistics*. Springer, New York. MR2172729

WASSERMAN, L., KOLAR, M. and RINALDO, A. (2013). Estimating undirected graphs under weak assumptions. Preprint. Available at arXiv:1309.6933.

WESTLING, T. and MCCORMICK, T. H. (2015). Establishing consistency and improving uncertainty estimates of variational inference through m-estimation. Preprint. Available at arXiv:1510.08151.

WOODBURY, M. A., CLIVE, J. and GARSON, A. (1978). Mathematical typology: A grade of membership technique for obtaining disease definition. *Comput. Biomed. Res.* **11** 277–298.

The Annals of Applied Statistics
2018, Vol. 12, No. 2, 877–914
https://doi.org/10.1214/18-AOAS1171

PROVIDING ACCURATE MODELS ACROSS PRIVATE PARTITIONED DATA: SECURE MAXIMUM LIKELIHOOD ESTIMATION

BY JOSHUA SNOKE*,[1], TIMOTHY R. BRICK*, ALEKSANDRA SLAVKOVIĆ*,[1] AND MICHAEL D. HUNTER†,[2]

Pennsylvania State University and University of Oklahoma Health Sciences Center*†

This paper focuses on the privacy paradigm of providing access to researchers to remotely carry out analyses on sensitive data stored behind separate firewalls. We address the situation where the analysis demands data from multiple physically separate databases which cannot be combined. Motivating this work is a real model based on research data on kinship foster placement that came from multiple sources and could only be combined through a lengthy process with a trusted research network. We develop and demonstrate a method for accurate calculation of the multivariate normal likelihood, for a set of parameters given the partitioned data, which can then be maximized to obtain estimates. These estimates are achieved without sharing any data or any true intermediate statistics of the data across firewalls. We show that under a certain set of assumptions our method for estimation across these partitions achieves identical results as estimation with the full data. Privacy is maintained by adding noise at each partition. This ensures each party receives noisy statistics, such that the noise cannot be removed until the last step to obtain a single value, the true total log likelihood. Potential applications include all methods utilizing parameter estimation through maximizing the multivariate normal likelihood. We give detailed algorithms, along with available software, and present simulations and analyze the kinship foster placement data estimating structural equation models (SEMs) with partitioned data.

REFERENCES

ARBUCKLE, J. L., MARCOULIDES, G. A. and SCHUMACKER, R. E. (1996). Full information estimation in the presence of incomplete data. *Adv. Struct. Equ. Model. Issues Techn.* **243** 277.

BOKER, S. M., BRICK, T. R., PRITIKIN, J. N., WANG, Y., OERTZEN, T. V., BROWN, D., LACH, J., ESTABROOK, R., HUNTER, M. D., MAES, H. H. and NEALE, M. C. (2015). Maintained Individual Data Distributed Likelihood Estimation (MIDDLE). *Multivar. Behav. Res.* **50** 706–720.

CALANDRINO, J. A., KILZER, A., NARAYANAN, A., FELTEN, E. W. and SHMATIKOV, V. (2011). "You might also like:" privacy risks of collaborative filtering. In *Security and Privacy* (*SP*), 2011 *IEEE Symposium on* 231–246. IEEE.

DE MONTJOYE, Y.-A., SHMUELI, E., WANG, S. S. and PENTLAND, A. S. (2014). Openpds: Protecting the privacy of metadata through safeanswers. *PLoS ONE* **9** e98790.

Key words and phrases. Partitioned data, privacy, secure multiparty computation, structural equation models, distributed maximum likelihood estimation.

Di Crescenzo, G., Malkin, T. and Ostrovsky, R. (2000). Single database private information retrieval implies oblivious transfer. In *Advances in Cryptology—EUROCRYPT* 2000 (*Bruges*). *Lecture Notes in Computer Science* **1807** 122–138. Springer, Berlin. MR1772023

Dinur, I. and Nissim, K. (2003). Revealing information while preserving privacy. In *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* 202–210. ACM.

Dunst, C. J., Trivette, C. M. and Deal, A. G. (1988). *Enabling and Empowering Families*: *Principles and Guidelines for Practice*. Brookline Books, Cambridge, MA.

Dwork, C. (2008). Differential privacy: A survey of results. In *Theory and Applications of Models of Computation*. *Lecture Notes in Computer Science* **4978** 1–19. Springer, Berlin. MR2472670

Fienberg, S. E., Nardi, Y. and Slavković, A. B. (2009). Valid statistical analysis for logistic regression with multiple sources. In *Protecting Persons While Protecting the People* 82–94. Springer.

Fienberg, S. E. and Slavković, A. B. (2011). Data privacy and confidentiality. In *International Encyclopedia of Statistical Science* 342–345. Springer.

Fienberg, S. E., Fulp, W. J., Slavkovic, A. B. and Wrobel, T. A. (2006). "Secure" log-linear and logistic regression analysis of distributed databases. In *Privacy in Statistical Databases* 277–290. Springer.

Gaye, A., Marcon, Y., Isaeva, J., LaFlamme, P., Turner, A., Jones, E. M., Minion, J., Boyd, A. W., Newby, C. J., Nuotio, M.-L., Wilson, R., Butters, O., Murtagh, B., Demir, I., Doiron, D., Giepmans, L., Wallace, S. E., Budin-Ljøsne, I., Schmidt, C. O., Boffetta, P., Boniol, M., Bota, M., Carter, K. W., deKlerk, N., Dibben, C., Francis, R. W., Hiekkalinna, T., Hveem, K., Kvaløy, K., Millar, S., Perry, I. J., Peters, A., Phillips, C. M., Popham, F., Raab, G., Reischl, E., Sheehan, N., Waldenberger, M., Perola, M., van den Heuvel, E., Macleod, J., Knoppers, B. M., Stolk, R. P., Fortier, I., Harris, J. R., Woffenbuttel, B. H. R., Murtagh, M. J., Ferretti, V. and Burton, P. R. (2014). DataSHIELD: Taking the analysis to the data, not the data to the analysis. *Int*. *J*. *Epidemiol*. **43** 1929–1944.

Ghosh, J., Reiter, J. P. and Karr, A. F. (2007). Secure computation with horizontally partitioned data using adaptive regression splines. *Comput*. *Statist*. *Data Anal*. **51** 5813–5820. MR2407679

Goldwasser, S. (1997). Multi party computations: Past and present. In *Proceedings of the Sixteenth Annual ACM Symposium on Principles of Distributed Computing* 1–6. ACM.

Haynsworth, E. V. (1968). On the Schur complement. Technical Report, DTIC Document.

Hecht, D. B., Hunter, M. D. and Beasley, L. O. (2016). Family KINnections: A Kinship Navigation Program. Presented to the University of Oklahoma Health Sciences Center Department of Pediatrics Section of Developmental and Behavioral Pediatrics at the Section Research Meeting.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J. V., Stephan, D. A., Nelson, S. F. and Craig, D. W. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet*. **4** e1000167.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K. and de Wolf, P.-P. (2012). *Statistical Disclosure Control*. *Wiley Series in Survey Methodology*. Wiley, Chichester. MR3026260

Karr, A. F., Lin, X., Sanil, A. P. and Reiter, J. P. (2005). Secure regression on distributed databases. *J*. *Comput*. *Graph*. *Statist*. **14** 263–279. MR2160813

Karr, A. F., Fulp, W. J., Vera, F., Young, S. S., Lin, X. and Reiter, J. P. (2007). Secure, privacy-preserving analysis of distributed databases. *Technometrics* **49** 335–345. MR2408637

Karr, A. F., Lin, X., Sanil, A. P. and Reiter, J. P. (2009). Privacy-preserving analysis of vertically partitioned data using secure matrix products. *J*. *Off*. *Stat*. **25** 125.

Kissner, L. and Song, D. (2005). Privacy-preserving set operations. In *Advances in Cryptology—CRYPTO* 2005. *Lecture Notes in Computer Science* **3621** 241–257. Springer, Berlin. MR2237310

LIN, X. and KARR, A. F. (2010). Privacy-preserving maximum likelihood estimation for distributed data. *J. Priv. Confid.* **1** 6.

LINDELL, Y. and PINKAS, B. (2009). Secure multiparty computation for privacy-preserving data mining. *J. Priv. Confid.* **1** 5.

MEREDITH, W. and TISAK, J. (1990). Latent curve analysis. *Psychometrika* **55** 107–122.

NARDI, Y., FIENBERG, S. E. and HALL, R. J. (2012). Achieving both valid and secure logistic regression analysis on aggregated data from different private sources. *J. Priv. Confid.* **4** 9.

NASH, J. C. and VARADHAN, R. (2011). Unifying optimization algorithms to aid software system users: optimx for R. *J. Stat. Softw.* **43** 1–14.

NEALE, M. C., HUNTER, M. D., PRITIKIN, J. N., ZAHERY, M., BRICK, T. R., KIRKPATRICK, R. M., ESTABROOK, R., BATES, T. C., MAES, H. H. and BOKER, S. M. (2016). OpenMx 2.0: Extended structural equation and statistical modeling. *Psychometrika* **81** 535–549. MR3505378

R CORE TEAM (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

RAGHUNATHAN, T. E., REITER, J. P. and RUBIN, D. B. (2003). Multiple imputation for statistical disclosure limitation. *J. Off. Stat.* **19** 1–17.

REITER, J. P., KOHNEN, C. N., KARR, A. F., LIN, X. and SANIL, A. P. (2004). Partitioned, Vertically and Data, Partially Overlapping. Technical Report, NISS. Available at https://www.niss.org/sites/default/files/technicalreports/tr146.pdf.

SAMIZO, Y. (2016). Secure statistical analyses on vertically distributed databases. Master's thesis, The Pennsylvania State Univ.

SANIL, A. P., KARR, A. F., LIN, X. and REITER, J. P. (2004). Privacy preserving regression modelling via distributed computation. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 677–682. ACM.

SAVAGE, C. J. and VICKERS, A. J. (2009). Empirical study of data sharing by authors publishing in PLoS journals. *PLoS ONE* **4** e7078. DOI:10.1371/journal.pone.0007078.

SCHAFER, J. L. (1997). *Analysis of Incomplete Multivariate Data*. *Monographs on Statistics and Applied Probability* **72**. Chapman & Hall, London. MR1692799

SCHUR, I. (1905). Neue Begründung der Theorie der Gruppencharaktere. *Sitzungsberichte Königl. Preuss. Akad. Wiss.* 406–432.

SLAVKOVIC, A. B., NARDI, Y. and TIBBITS, M. M. (2007). "Secure" logistic regression of horizontally and vertically partitioned distributed databases. In *Data Mining Workshop*, 2007, *ICDM Workshops* 2007, *Seventh IEEE International Conference on Data Mining* 723–728.

SNOKE, J., BRICK, T. and SLAVKOVIĆ, A. (2016). Accurate estimation of structural equation models with remote partitioned data. In *International Conference on Privacy in Statistical Databases* 190–209. Springer.

SULLIVAN, C. M. (1992). *An Overview of Disclosure Principles*. Bureau of the Census.

VAIDYA, J. and CLIFTON, C. (2004). Privacy preserving naïve Bayes classifier for vertically partitioned data. In *Proceedings of the Fourth SIAM International Conference on Data Mining* 522–526. SIAM, Philadelphia, PA. MR2388481

VAIDYA, J., CLIFTON, C., KANTARCIOGLU, M. and PATTERSON, A. S. (2008). Privacy-preserving decision trees over vertically partitioned data. *ACM Trans. Knowl. Discov. Data* **2** 14.

WILLENBORG, L. and DE WAAL, T. (2001). *Elements of Statistical Disclosure Control*. *Lecture Notes in Statistics* **155**. Springer, New York. MR1866909

YAO, A. C. (1982). Protocols for secure computations. In 23*rd Annual Symposium on Foundations of Computer Science* (*Chicago*, *IL*, 1982) 160–164. IEEE, New York. MR0780394

The Annals of Applied Statistics
2018, Vol. 12, No. 2, 915–939
https://doi.org/10.1214/18-AOAS1173

CLUSTERING THE PREVALENCE OF PEDIATRIC CHRONIC CONDITIONS IN THE UNITED STATES USING DISTRIBUTED COMPUTING[1]

BY YUCHEN ZHENG AND NICOLETA SERBAN

Georgia Institute of Technology

This research paper presents an approach to clustering the prevalence of chronic conditions among children with public insurance in the United States. The data consist of prevalence estimates at the community level for 25 pediatric chronic conditions. We employ a spatial clustering algorithm to identify clusters of communities with similar chronic condition prevalences. The primary challenge is the computational effort needed to estimate the spatial clustering for all communities in the U.S. To address this challenge, we develop a distributed computing approach to spatial clustering. Overall, we found that the burden of chronic conditions in rural communities tends to be similar but with wide differences in urban communities. This finding suggests similar interventions for managing chronic conditions in rural communities but targeted interventions in urban areas.

REFERENCES

AMDAHL, G. M. (1967). Validity of the single processor approach to achieving large scale computing capabilities. In *Proceedings of the Spring Joint Computer Conference* 483–485. ACM, New York.

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

BESAG, J. and NEWELL, J. (1991). The detection of clusters in rare diseases. *J. Roy. Statist. Soc. Ser. A* **154** 143–155.

BEZANSON, J., EDELMAN, A., KARPINSKI, S. and SHAH, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Rev.* **59** 65–98. MR3605826

BIRANT, D. and KUT, A. (2007). ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data Knowl. Eng.* **60** 208–221.

CAMERON, E., BATTLE, K. E., BHATT, S., WEISS, D. J., BISANZIO, D., MAPPIN, B., DALRYMPLE, U., HAY, S. I., SMITH, D. L., GRIFFIN, J. T. et al. (2015). Defining the relationship between infection prevalence and clinical incidence of *Plasmodium falciparum* malaria. *Nat. Commun.* **6** Art. ID 8170.

CARSON, C., BELONGIE, S., GREENSPAN, H. and MALIK, J. (2002). Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Trans. Pattern Anal. Mach. Intell.* **24** 1026–1038.

CENTER FOR MEDICARE AND MEDICAID SERVICES (2017a). September 2017 Medicaid and CHIP enrollment data highlights. Available at https://www.medicaid.gov/medicaid/program-information/medicaid-and-chip-enrollment-data/report-highlights/index.html.

Key words and phrases. Distributed computing, Medicaid, pediatric chronic conditions, spatial clustering.

CENTER FOR MEDICARE AND MEDICAID SERVICES (2017b). Quality of care health disparities. Available at https://www.medicaid.gov/medicaid/quality-of-care/improvement-initiatives/health-disparities/index.html.

CHU, C.-T., KIM, S. K., LIN, Y.-A., YU, Y., BRADSKI, G., OLUKOTUN, K. and NG, A. Y. (2007). Map-reduce for machine learning on multicore. In *Advances in Neural Information Processing Systems* 281–288.

COCKERHAM, W. C., HAMBY, B. W. and OATES, G. R. (2017). The social determinants of chronic disease. *Am. J. Prev. Med.* **52** S5–S12.

CRESSIE, N. A. C. (2015). *Statistics for Spatial Data*, revised ed. Wiley, New York. Paperback edition of the 1993 edition [MR1239641]. MR3559472

DAVILA-PAYAN, C., DEGUZMAN, M., JOHNSON, K., SERBAN, N. and SWANN, J. (2015). Estimating prevalence of overweight or obese children and adolescents in small geographic areas using publicly available data. *Prev. Chronic Dis.* **12**. DOI:10.5888/pcd12.140229.

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. With discussion. MR0501537

DIGGLE, P. J. and GIORGI, E. (2016). Model-based geostatistics for prevalence mapping in low-resource settings. *J. Amer. Statist. Assoc.* **111** 1096–1120. MR3561931

DING, C. and HE, X. (2004). K-means clustering via principal component analysis. In *Proceedings of the* 21*st International Conference on Machine Learning* 29. ACM, New York.

ELLIOT, P., WAKEFIELD, J. C., BEST, N. G. and BRIGGS, D. J. (2000). *Spatial Epidemiology: Methods and Applications*. Oxford Univ. Press, Oxford.

ESTER, M., KRIEGEL, H.-P., SANDER, J., XU, X. et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the* 2*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD*'96) 226–231.

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

FRALEY, C. and RAFTERY, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *Comput. J.* **41** 578–588.

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

GOTWAY, C. A. and YOUNG, L. J. (2002). Combining incompatible spatial data. *J. Amer. Statist. Assoc.* **97** 632–648. MR1951636

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

JIANG, H. and SERBAN, N. (2012). Clustering random curves under spatial interdependence with application to service accessibility. *Technometrics* **54** 108–119. MR2929427

KOPEC, J. A., SAYRE, E. C., FLANAGAN, W. M., FINES, P., CIBERE, J., RAHMAN, M. M., BANSBACK, N. J., ANIS, A. H., JORDAN, J. M., SOBOLEV, B. et al. (2010). Development of a population-based microsimulation model of osteoarthritis in Canada. *Osteoarthr. Cartil.* **18** 303–311.

KRIEGEL, H.-P., KRÖGER, P., SANDER, J. and ZIMEK, A. (2011). Density-based clustering. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **1** 231–240.

LAWSON, A., BIGGERI, A., BOHNING, D., LESAFFRE, E., VIEL, J.-F. and BERTOLLINI, R. (1999). *Disease Mapping and Risk Assessment for Public Health*. Wiley, New York.

LIU, Q. and IHLER, A. (2012). Distributed parameter estimation via pseudo-likelihood. In *International Conference on Machine Learning* (*ICML*) 1487–1494.

MEYER, S. and HELD, L. (2014). Power-law models for infectious disease spread. *Ann. Appl. Stat.* **8** 1612–1639. MR3271346

NEFF, J. M., SHARP, V. L., MULDOON, J., GRAHAM, J., POPALISKY, J. and GAY, J. C. (2002). Identifying and classifying children with chronic conditions using administrative data with the clinical risk group classification system. *Ambul. Pediatr.* **2** 71–79.

OPENSHAW, S., CHARLTON, M., WYMER, C. and CRAFT, A. (1987). A mark 1 geographical analysis machine for the automated analysis of point data sets. *Int. J. Geogr. Inf. Syst.* **1** 335–358.

RAND, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. Amer. Statist. Assoc.* **66** 846–850.

RIPLEY, B. D. (2005). *Spatial Statistics*. Wiley, New York. MR0624436

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

RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. Roy. Statist. Soc. Ser. B* **71** 319–392. MR2649602

RUE, H. and TJELMELAND, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scand. J. Stat.* **29** 31–49. MR1894379

THE WORLD HEALTH ORGANIZATION (2005). Chronic diseases and their common risk factors. Available at http://www.who.int/chp/chronic_disease_report/media/Factsheet1.pdf.

UNITED STATES DEPARTMENT OF AGRICULTURE (2004). Measuring rurality: Rural-urban continuum codes. Available at https://www.ers.usda.gov/data-products/rural-urban-continuum-codes/.

WAKEFIELD, J. C. (2006). Disease mapping and spatial regression with count data. *Biostatistics* **8** 158–183.

WALLER, L. A. and GOTWAY, C. A. (2004). *Applied Spatial Statistics for Public Health Data*. Wiley, Hoboken, NJ. MR2075123

WANG, M., WANG, A. and LI, A. (2006). Mining spatial-temporal clusters from geo-databases. In *Advanced Data Mining and Applications*. *Lecture Notes in Artificial Intelligence* **4093** 263–270. Springer, Berlin.

WOLFE, J., HAGHIGHI, A. and KLEIN, D. (2008). Fully distributed EM for very large datasets. In *Proceedings of the* 25*th International Conference on Machine Learning* 1184–1191. ACM, New York.

ZHENG, Y. and SERBAN, N. (2018). Supplement to “Clustering the prevalence of pediatric chronic conditions in the United States using distributed computing.” DOI:10.1214/18-AOAS1173SUPP.

The Annals of Applied Statistics
2018, Vol. 12, No. 2, 940–970
https://doi.org/10.1214/18-AOAS1175

ESTIMATING LARGE CORRELATION MATRICES FOR INTERNATIONAL MIGRATION

BY JONATHAN J. AZOSE[*,†] AND ADRIAN E. RAFTERY[†]

Pacific Northwest National Laboratory[*] *and University of Washington*[†]

The United Nations is the major organization producing and regularly updating probabilistic population projections for all countries. International migration is a critical component of such projections, and between-country correlations are important for forecasts of regional aggregates. However, in the data we consider there are 200 countries and only 12 data points, each one corresponding to a five-year time period. Thus a 200 × 200 correlation matrix must be estimated on the basis of 12 data points. Using Pearson correlations produces many spurious correlations. We propose a maximum *a posteriori* estimator for the correlation matrix with an interpretable informative prior distribution. The prior serves to regularize the correlation matrix, shrinking *a priori* untrustworthy elements towards zero. Our estimated correlation structure improves projections of net migration for regional aggregates, producing narrower projections of migration for Africa as a whole and wider projections for Europe. A simulation study confirms that our estimator outperforms both the Pearson correlation matrix and a simple shrinkage estimator when estimating a sparse correlation matrix.

REFERENCES

ABEL, G. (2013). Estimating global migration flow tables using place of birth data. *Demogr. Res.* **28** 505–546.

ANTONIADIS, A. and FAN, J. (2001). Regularization of wavelet approximations. *J. Amer. Statist. Assoc.* **96** 939–967. With discussion and a rejoinder by the authors. MR1946364

AZOSE, J. J. and RAFTERY, A. E. (2015). Bayesian probabilistic projection of international migration. *Demography* **52** 1627–1650.

AZOSE, J. J., ŠEVČÍKOVÁ, H. and RAFTERY, A. E. (2016). Probabilistic population projections with migration uncertainty. *Proc. Natl. Acad. Sci. USA* **113** 6460–6465.

BARBÉ, E. and JOHANSSON-NOGUÉS, E. (2008). The EU as a modest 'force for good': The European Neighbourhood Policy. *Int. Aff.* **84** 81–96.

BARNARD, J., MCCULLOCH, R. and MENG, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statist. Sinica* **10** 1281–1311. MR1804544

BECK, A. and TEBOULLE, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2** 183–202. MR2486527

BICKEL, P. J. and LEVINA, E. (2008a). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

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

Key words and phrases. Correlation estimation, international migration, maximum a posteriori estimation, high-dimension.

BIEN, J. and TIBSHIRANI, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika* **98** 807–820. MR2860325

BIJAK, J. and WIŚNIOWSKI, A. (2010). Bayesian forecasting of immigration to selected European countries by using expert knowledge. *J. Roy. Statist. Soc. Ser. A* **173** 775–796. MR2759965

BIJAK, J., KUPISZEWSKA, D., KUPISZEWSKI, M., SACZUK, K. and KICINGER, A. (2007). Population and labour force projections for 27 European countries, 2002–2052: Impact of international migration on population ageing. *Eur. J. Popul.* **23** 1–31.

BROWN, S. K. and BEAN, F. D. (2012). Population growth. In *Debates on U.S. Immigration* (J. Gans, E. M. Replogle and D. J. Tichenor, eds.). SAGE, Thousand Oaks, CA.

CHAUDHURI, S., DRTON, M. and RICHARDSON, T. S. (2007). Estimation of a covariance matrix with zeros. *Biometrika* **94** 199–216. MR2307904

CHEN, X., XU, M. and WU, W. B. (2013). Covariance and precision matrix estimation for high-dimensional time series. *Ann. Statist.* **41** 2994–3021. MR3161455

CHI, E. C. and LANGE, K. (2014). Stable estimation of a covariance matrix guided by nuclear norm penalties. *Comput. Statist. Data Anal.* **80** 117–128. MR3240481

CRUSH, J. (1999). Fortress South Africa and the deconstruction of apartheid's migration regime. *Geoforum* **30** 1–11.

CUI, Y., LENG, C. and SUN, D. (2016). Sparse estimation of high-dimensional correlation matrices. *Comput. Statist. Data Anal.* **93** 390–403. MR3406221

DE BEER, J., RAYMER, J., VAN DER ERF, R. and VAN WISSEN, L. (2010). Overcoming the problems of inconsistent international migration data: A new method applied to flows in Europe. *Eur. J. Popul.* **26** 459–481.

DENG, X. and TSUI, K.-W. (2013). Penalized covariance matrix estimation using a matrix-logarithm transformation. *J. Comput. Graph. Statist.* **22** 494–512. MR3173726

EL KAROUI, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.* **36** 2717–2756. MR2485011

FAN, J., HAN, F. and LIU, H. (2014). Challenges of big data analysis. *Nat. Sci. Rev.* **1** 293–314.

FAN, J., HUANG, T. and LI, R. (2007). Analysis of longitudinal data with semiparametric estimation of convariance function. *J. Amer. Statist. Assoc.* **102** 632–641. MR2370857

FAN, J., LIAO, Y. and LIU, H. (2016). An overview of the estimation of large covariance and precision matrices. *Econom. J.* **19** C1–C32. MR3501529

FAN, J., LIAO, Y. and MINCHEVA, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 603–680. With 33 discussions by 57 authors and a reply by Fan, Liao and Mincheva. MR3091653

FASSMANN, H. and MUNZ, R. (1994). European East–West migration, 1945–1992. *Int. Migr. Rev.* **28** 520–538.

FOSDICK, B. K. and RAFTERY, A. E. (2014). Regional probabilistic fertility forecasting by modeling between-country correlations. *Demogr. Res.* **30** 1011–1034.

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

FURRER, R. and BENGTSSON, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *J. Multivariate Anal.* **98** 227–255. MR2301751

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

HARRIS, J. R. and TODARO, M. P. (1970). Migration, unemployment and development: A two-sector analysis. *Am. Econ. Rev.* **60** 126–142.

HERSBACH, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecast.* **15** 559–570.

HUANG, A. and WAND, M. P. (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Anal.* **8** 439–451. MR3066948

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

INTERNATIONAL ORGANIZATION FOR MIGRATION (2015). *Migration Governance Framework* (*C*/106/40). International Organization for Migration, Geneva. Available at https://governingbodies.iom.int/system/files/en/council/106/C-106-40-Migration-Governance-Framework.pdf.

INTERNATIONAL ORGANIZATION FOR MIGRATION and MCKINSEY & COMPANY (2018). *More than Numbers*: *How Migration Data Can Deliver Real-Life Benefits for Migrants and Governments*. International Organization for Migration, Geneva. Available at https://publications.iom.int/system/files/pdf/more_than_numbers.pdf.

JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. In *Proc*. 4*th Berkeley Sympos*. *Math*. *Statist*. *and Prob*., *Vol*. *I* 361–379. Univ. California Press, Berkeley, CA. MR0133191

LEDOIT, O. and WOLF, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *J*. *Empir*. *Finance* **10** 603–621.

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

LEDOIT, O. and WOLF, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Ann*. *Statist*. **40** 1024–1060. MR2985942

LEE, E. S. (1966). A theory of migration. *Demography* **3** 47–57.

LEONARD, T. and HSU, J. S. J. (1992). Bayesian inference for a covariance matrix. *Ann*. *Statist*. **20** 1669–1696. MR1193308

LEVINA, E., ROTHMAN, A. and ZHU, J. (2008). Sparse estimation of large covariance matrices via a nested Lasso penalty. *Ann*. *Appl*. *Stat*. **2** 245–263. MR2415602

LIECHTY, J. C., LIECHTY, M. W. and MÜLLER, P. (2004). Bayesian correlation estimation. *Biometrika* **91** 1–14. MR2050456

LIU, H., WANG, L. and ZHAO, T. (2014). Sparse covariance matrix estimation with eigenvalue constraints. *J*. *Comput*. *Graph*. *Statist*. **23** 439–459. MR3215819

MAYER, T. and ZIGNAGO, S. (2011). Notes on CEPII's distances measures: The GeoDist database.

NOCEDAL, J. and WRIGHT, S. J. (2006). *Numerical Optimization*, 2nd ed. Springer, New York. MR2244940

OKOLSKI, M. Regional dimension of international migration in Central and Eastern Europe. *Genus* **54** 11–36.

POURAHMADI, M. (2011). Covariance estimation: The GLM and regularization perspectives. *Statist*. *Sci*. **26** 369–387. MR2917961

RAYMER, J., WIŚNIOWSKI, A., FORSTER, J. J., SMITH, P. W. F. and BIJAK, J. (2013). Integrated modeling of European migration. *J*. *Amer*. *Statist*. *Assoc*. **108** 801–819. MR3174664

ROGERS, A. (1990). Requiem for the net migrant. *Geogr*. *Anal*. **22** 283–300.

SJAASTAD, L. A. (1962). The costs and returns of human migration. *J*. *Polit*. *Econ*. **70** 80–93.

STARK, O. and BLOOM, D. E. (1985). The new economics of labor migration. *Am*. *Econ*. *Rev*. **75** 173–178.

THIELEMANN, E. (2008). The future of the common European asylum system. *Eur*. *Policy Anal*. **1** 1–8.

TIERNEY, L. and KADANE, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J*. *Amer*. *Statist*. *Assoc*. **81** 82–86. MR0830567

U. S. SOCIAL SECURITY ADMINISTRATION (2013). The 2013 Annual Report of the Board of Trustees of the Federal Old-age and Survivors Insurance and Federal Disability Insurance Trust Funds. Board of Trustees, Federal Old-Age and Survivors Insurance and Federal Disability Insurance Trust Funds.

UNITED NATIONS (2012). *World Population Prospects*: *The* 2012 *Revision*. United Nations, New York.

UNITED NATIONS (2016). *Agreement Concerning the Relationship Between the United Nations and the International Organization for Migration* (*A/RES/*70*/*976). United Nations, New York. Available at https://digitallibrary.un.org/record/837208/files/A_RES_70_296-EN.pdf.

UNITED NATIONS (2017). *World Population Prospects*: *The* 2017 *Revision*. United Nations, New York.

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

WIŚNIOWSKI, A., SMITH, P. W., BIJAK, J., RAYMER, J. and FORSTER, J. J. (2015). Bayesian population forecasting: Extending the Lee–Carter method. *Demography* **52** 1035–1059.

WRIGHT, E. (2010). 2008-based national population projections for the United Kingdom and constituent countries. *Popul. Trends* **139** 91–114.

ZHANG, T. and ZOU, H. (2014). Sparse precision matrix estimation via lasso penalized D-trace loss. *Biometrika* **101** 103–120. MR3180660

The Annals of Applied Statistics
2018, Vol. 12, No. 2, 971–1012
https://doi.org/10.1214/18-AOAS1176

TRACKING NETWORK DYNAMICS: A SURVEY USING GRAPH DISTANCES[1]

BY CLAIRE DONNAT AND SUSAN HOLMES[2]

Stanford University

From longitudinal biomedical studies to social networks, graphs have emerged as essential objects for describing evolving interactions between agents in complex systems. In such studies, after pre-processing, the data are encoded by a set of graphs, each representing a system's state at a different point in time or space. The analysis of the system's dynamics depends on the selection of the appropriate analytical tools. In particular, after specifying properties characterizing similarities between states, a critical step lies in the choice of a distance between graphs capable of reflecting such similarities.

While the literature offers a number of distances to choose from, their properties have been little investigated and no guidelines regarding the choice of such a distance have yet been provided. In particular, most graph distances consider that the nodes are exchangeable—ignoring node "identities." Alignment of the graphs according to identified nodes enables us to enhance these distances' sensitivity to perturbations in the network and detect important changes in graph dynamics. Thus the selection of an adequate metric is a decisive—yet delicate—practical matter.

In the spirit of Goldenberg et al.'s seminal 2009 review [*Found. Trends Mach. Learn.* **2** (2010) 129–233], this article provides an overview of commonly-used graph distances and an explicit characterization of the structural changes that they are best able to capture. We show how these choices affect real-life situations, and we use these distances to analyze both a longitudinal microbiome dataset and a brain fMRI study. One contribution of the present study is a coordinated suite of data analytic techniques, displays and statistical tests using "metagraphs": a graph of graphs based on a chosen metric. Permutation tests can uncover the effects of covariates on the graphs' variability. Furthermore, synthetic examples provide intuition as to the qualities and drawbacks of the different distances. Above all, we provide some guidance on choosing one distance over another in different contexts. Finally, we extend the scope of our analyses from temporal to spatial dynamics and apply these different distances to a network created from worldwide recipes.

REFERENCES

AHN, Y.-Y., AHNERT, S. E., BAGROW, J. P. and BARABÁSI, A.-L. (2011). Flavor network and the principles of food pairing. *Sci. Rep.* **1** 196.

BANERJEE, A. (2008). The spectrum of the graph Laplacian as a tool for analyzing structure and evolution of networks. Ph.D. thesis, Univ. Leipzig.

Key words and phrases. Temporal networks, longitudinal analysis, graph distances, graph signal processing, wavelets, microbiome, longitudinal analysis.

BANERJEE, A. and JOST, J. (2008). Spectral plot properties: Towards a qualitative classification of networks. *Netw. Heterog. Media* **3** 395–411. MR2395239

BARBERÁN, A., BATES, S. T., CASAMAYOR, E. O. and FIERER, N. (2012). Using network analysis to explore co-occurrence patterns in soil microbial communities. *ISME J.* **6** 343–351.

BOLLOBÁS, B. (1998). *Modern Graph Theory*. *Graduate Texts in Mathematics* **184**. Springer, New York. MR1633290

BONATO, A., GLEICH, D. F., KIM, M., MITSCHE, D., PRALAT, P., TIAN, Y. and YOUNG, S. J. (2014). Dimensionality of social networks using motifs and eigenvalues. *PLoS ONE* **9** e106052.

CHAKERIAN, J. and HOLMES, S. (2012). Computational tools for evaluating phylogenetic and hierarchical clustering trees. *J. Comput. Graph. Statist.* **21** 581–599. MR2970909

CHAMPIN, P.-A. and SOLNON, C. (2003). Measuring the similarity of labeled graphs. In *International Conference on Case-Based Reasoning* 80–95. Springer, Berlin.

CHUNG, F. (2007). The heat kernel as the PageRank of a graph. *Proc. Natl. Acad. Sci. USA* **104** 19735–19740.

CVETKOVIĆ, D. (2012). Spectral recognition of graphs. *Yugosl. J. Oper. Res.* **22** 145–161. MR3007483

DETHLEFSEN, L. and RELMAN, D. A. (2011). Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proc. Natl. Acad. Sci. USA* **108** 4554–4561.

DIGIULIO, D. B., CALLAHAN, B. J., MCMURDIE, P. J., COSTELLO, E. K., LYELL, D. J., ROBACZEWSKA, A., SUN, C. L., GOLTSMAN, D. S., WONG, R. J., SHAW, G. et al. (2015). Temporal and spatial variation of the human microbiota during pregnancy. *Proc. Natl. Acad. Sci. USA* **112** 11060–11065.

DONNAT, C. and HOLMES, S. (2018). Supplement to "Tracking network dynamics: A survey using graph distances." DOI:10.1214/18-AOAS1176SUPP.

DONNAT, C., ZITNIK, M., HALLAC, D. and LESKOVEC, J. (2017). Spectral graph wavelets for structural role similarity in networks. Preprint. Available at ArXiv:1710.10321.

FERRER, M., BARDAJÍ, I., VALVENY, E., KARATZAS, D. and BUNKE, H. (2013). Median graph computation by means of graph embedding into vector spaces. In *Graph Embedding for Pattern Analysis* 45–71. Springer, Berlin.

FUKUYAMA, J., MCMURDIE, P. J., DETHLEFSEN, L., RELMAN, D. A. and HOLMES, S. (2012). Comparisons of distance methods for combining covariates and abundances in microbiome studies. In *Biocomputing* 2012 213–224. World Scientific, Singapore.

GERBER, G. K. (2014). The dynamic microbiome. *FEBS Lett.* **588** 4131–4139.

GOLDENBERG, A., ZHENG, A. X., FIENBERG, S. E. and AIROLDI, E. M. (2010). A survey of statistical network models. *Found. Trends Mach. Learn.* **2** 129–233.

GU, J., JOST, J., LIU, S. and STADLER, P. F. (2016). Spectral classes of regular, random, and empirical graphs. *Linear Algebra Appl.* **489** 30–49. MR3421836

HAMMOND, D. K., VANDERGHEYNST, P. and GRIBONVAL, R. (2011). Wavelets on graphs via spectral graph theory. *Appl. Comput. Harmon. Anal.* **30** 129–150. MR2754772

HULLAR, M. A. and LAMPE, J. W. (2012). The gut microbiome and obesity. In *Obesity Treatment and Prevention: New Directions* **73** 67–79. Karger, Basel.

IPSEN, M. and MIKHAILOV, A. S. (2002). Evolutionary reconstruction of networks. *Phys. Rev. E* (3) **66** 6–9. DOI:10.1103/PhysRevE.66.046109.

JOST, J. and JOY, M. P. (2002). Evolving networks with distance preferences. *Phys. Rev. E* (3) **66** 036126.

JURMAN, G., VISINTAINER, R. and FURLANELLO, C. (2011). An introduction to spectral distances in networks. *Frontiers Artificial Intelligence Appl.* **226** 227–234. DOI:10.3233/978-1-60750-692-8-227.

JURMAN, G., VISINTAINER, R., RICCADONNA, S., FILOSI, M. and FURLANELLO, C. (2012). A glocal distance for network comparison. Preprint. Available at ArXiv:1201.2931.

JURMAN, G., VISINTAINER, R., FILOSI, M., RICCADONNA, S. and FURLANELLO, C. (2015). The HIM glocal metric and kernel for network comparison and classification. In *Proceedings of the* 2015 *IEEE International Conference on Data Science and Advanced Analytics*, *DSAA* 2015 **7** 46109. DOI:10.1109/DSAA.2015.7344816.

JURMAN, G., FILOSI, M., RICCADONNA, S., VISINTAINER, R. and FURLANELLO, C. (2016). Differential network analysis and graph classification: A glocal approach. 1–13.

KELLY, C., ZUO, X.-N., GOTIMER, K., COX, C. L., LYNCH, L., BROCK, D., IMPERATI, D., GARAVAN, H., ROTROSEN, J., CASTELLANOS, F. X. et al. (2011). Reduced interhemispheric resting state functional connectivity in cocaine addiction. *Biological Psychiatry* **69** 684–692.

KELMANS, A. K. (1976). Comparison of graphs by their number of spanning trees. *Discrete Math*. **16** 241–261. MR0463000

KELMANS, A. K. (1997). Transformations of a graph increasing its Laplacian polynomial and number of spanning trees. *European J. Combin*. **18** 35–48. MR1427603

KOUTRA, D., PARIKH, A., RAMDAS, A. and XIANG, J. (2011). Algorithms for graph similarity and subgraph matching. In *Proc. Ecol. Inference Conf.*

KOUTRA, D., SHAH, N., VOGELSTEIN, J. T., GALLAGHER, B. and FALOUTSOS, C. (2016). DeltaCon: Principled massive-graph similarity function with attribution. *ACM Transactions on Knowledge Discovery from Data* (*TKDD*) **10** 28.

LAYEGHIFARD, M., HWANG, D. M. and GUTTMAN, D. S. (2017). Disentangling interactions in the microbiome: A network perspective. *Trends Microbiol*. **25** 217–228.

LEVANDOWSKY, M. and WINTER, D. (1971). Distance between sets. *Nature* **234** 34–35.

LUQMAN, M. M., RAMEL, J.-Y. and LLADÓS, J. (2013). Multilevel analysis of attributed graphs for explicit graph embedding in vector spaces. In *Graph Embedding for Pattern Analysis* 1–26. Springer, Berlin.

MONNIG, N. D. and MEYER, F. G. (2018). The resistance perturbation distance: A metric for the analysis of dynamic networks. *Discrete Appl. Math*. **236** 347–386. MR3739796

PAPADIMITRIOU, P., DASDAN, A. and GARCIA-MOLINA, H. (2010). Web graph similarity for anomaly detection. *Journal of Internet Services and Applications* **1** 19–30.

PROULX, S. R., PROMISLOW, D. E. and PHILLIPS, P. C. (2005). Network thinking in ecology and evolution. *Trends in Ecology & Evolution* **20** 345–353.

SHIMADA, Y., HIRATA, Y., IKEGUCHI, T. and AIHARA, K. (2016). Graph distance for complex networks. *Sci Rep* **6** 34944.

SHUMAN, D. I., RICAUD, B. and VANDERGHEYNST, P. (2016). Vertex-frequency analysis on graphs. *Appl. Comput. Harmon. Anal*. **40** 260–291. MR3440174

SHUMAN, D., NARANG, S., FROSSARD, P., ORTEGA, A. and VANDERGHEYNST, P. (2013). The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Process. Mag*. **30** 83–98.

SPIELMAN, D. A. (2007). Spectral graph theory and its applications. In *Foundations of Computer Science*, 2007. *FOCS'*07. 48*th Annual IEEE Symposium on* 29–38. IEEE Press, New York.

TÉTREAULT, P., MANSOUR, A., VACHON-PRESSEAU, E., SCHNITZER, T. J., APKARIAN, A. V. and BALIKI, M. N. (2016). Brain connectivity predicts placebo response across chronic pain clinical trials. *PLoS Biology* **14** e1002570.

THÜNE, M. (2012). Eigenvalues of matrices and graphs. Ph.D. thesis, Univ. Leipzig.

TREMBLAY, N. and BORGNAT, P. (2014). Graph wavelets for multiscale community mining. *IEEE Trans. Signal Process*. **62** 5227–5239. MR3268107

TURNBAUGH, P. J., LEY, R. E., MAHOWALD, M. A., MAGRINI, V., MARDIS, E. R. and GORDON, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444** 1027–131.

TURNBAUGH, P. J., BÄCKHED, F., FULTON, L. and GORDON, J. I. (2008). Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. *Cell Host & Microbe* **3** 213–223.

VISHWANATHAN, S. V. N., SCHRAUDOLPH, N. N., KONDOR, R. and BORGWARDT, K. M. (2010). Graph kernels. *J. Mach. Learn. Res.* **11** 1201–1242. MR2645450

WEISS, S., VAN TREUREN, W., LOZUPONE, C., FAUST, K., FRIEDMAN, J., DENG, Y., XIA, L. C., XU, Z. Z., URSELL, L., ALM, E. J. et al. (2016). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *The ISME Journal* **10** 1669–1681.

ZAGER, L. A. and VERGHESE, G. C. (2008). Graph similarity scoring and matching. *Appl. Math. Lett.* **21** 86–94. MR2435232

The Annals of Applied Statistics
2018, Vol. 12, No. 2, 1013–1038
https://doi.org/10.1214/18-AOAS1178

BAYESIAN PROPAGATION OF RECORD LINKAGE UNCERTAINTY INTO POPULATION SIZE ESTIMATION OF HUMAN RIGHTS VIOLATIONS[1]

BY MAURICIO SADINLE

University of Washington

Multiple-systems or capture–recapture estimation are common techniques for population size estimation, particularly in the quantitative study of human rights violations. These methods rely on multiple samples from the population, along with the information of which individuals appear in which samples. The goal of record linkage techniques is to identify unique individuals across samples based on the information collected on them. Linkage decisions are subject to uncertainty when such information contains errors and missingness, and when different individuals have very similar characteristics. Uncertainty in the linkage should be propagated into the stage of population size estimation. We propose an approach called *linkage-averaging* to propagate linkage uncertainty, as quantified by some Bayesian record linkage methodologies, into a subsequent stage of population size estimation. Linkage-averaging is a two-stage approach in which the results from the record linkage stage are fed into the population size estimation stage. We show that under some conditions the results of this approach correspond to those of a proper Bayesian joint model for both record linkage and population size estimation. The two-stage nature of linkage-averaging allows us to combine different record linkage models with different capture–recapture models, which facilitates model exploration. We present a case study from the Salvadoran civil war, where we are interested in estimating the total number of civilian killings using lists of witnesses' reports collected by different organizations. These lists contain duplicates, typographical and spelling errors, missingness, and other inaccuracies that lead to uncertainty in the linkage. We show how linkage-averaging can be used for transferring the uncertainty in the linkage of these lists into different models for population size estimation.

REFERENCES

ANDERSON, M. J. and FIENBERG, S. E. (1999). *Who Counts?*: *The Politics of Census-Taking in Contemporary America*, Revised paperback (2001) ed. Russell Sage Foundation, New York.

BALL, P. (2000). The Salvadoran human rights commission: Data processing, data representation, and generating analytical reports. In *Making the Case*: *Investigating Large Scale Human Rights Violations Using Information Systems and Data Analysis* (P. Ball, H. F. Spirer and L. Spirer, eds.) American Association for the Advancement of Science, Washington, DC.

BILENKO, M., MOONEY, R. J., COHEN, W. W., RAVIKUMAR, P. and FIENBERG, S. E. (2003). Adaptive name matching in information integration. *IEEE Intell*. *Syst*. **18** 16–23.

Key words and phrases. Capture–recapture, counting casualties, data linkage, decomposable graphical model, duplicate detection, entity resolution, multiple-systems estimation, multiple record linkage.

BIRD, S. M. and KING, R. (2018). Multiple systems estimation (or capture–recapture estimation) to inform public policy. *Ann. Rev. Statist. Appl.* **5** 95–118.

BISHOP, Y. M. M., FIENBERG, S. E. and HOLLAND, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA. With the collaboration of Richard J. Light and Frederick Mosteller. MR0381130

CASTLEDINE, B. J. (1981). A Bayesian analysis of multiple-recapture sampling for a closed population. *Biometrika* **68** 197–210. MR0614956

CHRISTEN, P. (2012). A survey of indexing techniques for scalable record linkage and deduplication. *IEEE Trans. Knowl. Data Eng.* **24** 1537–1555.

COMMISSION ON THE TRUTH FOR EL SALVADOR (1993). From madness to hope: The 12-year war in El Salvador: Report of the Commission on the Truth for El Salvador. Available at http://www.usip.org/files/file/ElSalvador-Report.pdf [Accessed May 21, 2018]. UN Security Council.

DAWID, A. P. and LAURITZEN, S. L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21** 1272–1317. MR1241267

EDWARDS, D. (2000). *Introduction to Graphical Modelling*, 2nd ed. Springer, New York. MR1880319

ELMAGARMID, A. K., IPEIROTIS, P. G. and VERYKIOS, V. S. (2007). Duplicate record detection: A survey. *IEEE Trans. Knowl. Data Eng.* **19** 1–16.

ERICKSEN, E. P., KADANE, J. B. and TUKEY, J. W. (1989). Adjusting the 1980 census of population and housing. *J. Amer. Statist. Assoc.* **84** 927–944.

FELLEGI, I. P. and SUNTER, A. B. (1969). A theory for record linkage. *J. Amer. Statist. Assoc.* **64** 1183–1210.

FIENBERG, S. E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika* **59** 591–603. MR0383619

FIENBERG, S. E., JOHNSON, M. S. and JUNKER, B. W. (1999). Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *J. Roy. Statist. Soc. Ser. A* **162** 383–405.

FORTINI, M., NUCCITELLI, A., LISEO, B. and SCANU, M. (2002). Modeling issues in record linkage: A Bayesian perspective. In *Proceedings of the Section on Survey Research Methods* 1008–1013. American Statistical Association, Alexandria, VA.

GEORGE, E. I. and ROBERT, C. P. (1992). Capture-recapture estimation via Gibbs sampling. *Biometrika* **79** 677–683. MR1209469

GUTMAN, R., AFENDULIS, C. C. and ZASLAVSKY, A. M. (2013). A Bayesian procedure for file linking to analyze end-of-life medical costs. *J. Amer. Statist. Assoc.* **108** 34–47. MR3174601

HERZOG, T. N., SCHEUREN, F. J. and WINKLER, W. E. (2007). *Data Quality and Record Linkage Techniques*. Springer, New York.

HOGAN, H. (1992). The 1990 post-enumeration survey: An overview. *Amer. Statist.* **46** 261–269.

HOGAN, H. (1993). The 1990 post-enumeration survey: Operations and results. *J. Amer. Statist. Assoc.* **88** 1047–1060.

HOWLAND, T. (2008). How El Rescate, a small nongovernmental organization, contributed to the transformation of the human rights situation in El Salvador. *Hum. Rights Q.* **30** 703–757.

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

JARO, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *J. Amer. Statist. Assoc.* **84** 414–420.

LAPORTE, R. E., MCCARTY, D., BRUNO, G., TAJIMA, N. and BABA, S. (1993). Counting diabetes in the next millennium: Application of capture–recapture technology. *Diabetes Care* **16** 528–534.

LARSEN, M. D. and RUBIN, D. B. (2001). Iterative automated record linkage using mixture models. *J. Amer. Statist. Assoc.* **96** 32–41. MR1973781

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

LISEO, B. and TANCREDI, A. (2011). Bayesian estimation of population size via linkage of multivariate normal data sets. *J. Off. Stat.* **27** 491–505.

LUM, K., PRICE, M. E. and BANKS, D. (2013). Applications of multiple systems estimation in human rights research. *Amer. Statist.* **67** 191–200. MR3303809

MADIGAN, D. and YORK, J. C. (1997). Bayesian methods for estimation of the size of a closed population. *Biometrika* **84** 19–31. MR1450189

MANRIQUE-VALLIER, D. (2016). Bayesian population size estimation using Dirichlet process mixtures. *Biometrics* **72** 1246–1254. MR3591609

MATSAKIS, N. E. (2010). Active duplicate detection with Bayesian nonparametric models. Ph.D. thesis, Massachusetts Institute of Technology.

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

POLLOCK, K. H. (2000). Capture–recapture models. *J. Amer. Statist. Assoc.* **95** 293–296.

PRICE, M. and BALL, P. (2015). Selection bias and the statistical patterns of mortality in conflict. *Statist. J. IAOS* **31** 263–272.

PRICE, M., GOHDES, A. and BALL, P. (2015). Documents of war: Understanding the Syrian conflict. *Significance* **12** 14–19.

SADINLE, M. (2014). Detecting duplicates in a homicide registry using a Bayesian partitioning approach. *Ann. Appl. Stat.* **8** 2404–2434. MR3292503

SADINLE, M. (2017). Bayesian estimation of bipartite matchings for record linkage. *J. Amer. Statist. Assoc.* **112** 600–612. MR3671755

STEORTS, R. C. (2015). Entity resolution with empirically motivated priors. *Bayesian Anal.* **10** 849–875. MR3432242

STEORTS, R. C., HALL, R. and FIENBERG, S. E. (2016). A Bayesian approach to graphical record linkage and deduplication. *J. Amer. Statist. Assoc.* **111** 1660–1672. MR3601725

TANCREDI, A. and LISEO, B. (2011). A hierarchical Bayesian approach to record linkage and population size problems. *Ann. Appl. Stat.* **5** 1553–1585. MR2849786

WINKLER, W. E. (1988). Using the EM algorithm for weight computation in the Fellegi–Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods* 667–671. American Statistical Association, Alexandria, VA.

WINKLER, W. E. (1990). String comparator metrics and enhanced decision rules in the Fellegi–Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods* 354–359. American Statistical Association, Alexandria, VA.

The Annals of Applied Statistics
2018, Vol. 12, No. 2, 1039–1067
https://doi.org/10.1214/18-AOAS1163

UNIQUE ENTITY ESTIMATION WITH APPLICATION TO THE SYRIAN CONFLICT

BY BEIDI CHEN[*,1], ANSHUMALI SHRIVASTAVA[*,2] AND REBECCA C. STEORTS[†,3]

Rice University[*] *and Duke University*[†]

Entity resolution identifies and removes duplicate entities in large, noisy databases and has grown in both usage and new developments as a result of increased data availability. Nevertheless, entity resolution has tradeoffs regarding assumptions of the data generation process, error rates, and computational scalability that make it a difficult task for real applications. In this paper, we focus on a related problem of unique entity estimation, which is the task of estimating the unique number of entities and associated standard errors in a data set with duplicate entities. Unique entity estimation shares many fundamental challenges of entity resolution, namely, that the computational cost of all-to-all entity comparisons is intractable for large databases. To circumvent this computational barrier, we propose an efficient (near-linear time) estimation algorithm based on locality sensitive hashing. Our estimator, under realistic assumptions, is unbiased and has provably low variance compared to existing random sampling based approaches. In addition, we empirically show its superiority over the state-of-the-art estimators on three real applications. The motivation for our work is to derive an accurate estimate of the documented, identifiable deaths in the ongoing Syrian conflict. Our methodology, when applied to the Syrian data set, provides an estimate of 191,874 ± 1,772 documented, identifiable deaths, which is very close to the Human Rights Data Analysis Group (HRDAG) estimate of 191,369. Our work provides an example of challenges and efforts involved in solving a real, noisy challenging problem where modeling assumptions may not hold.

REFERENCES

ALEKSANDROV, P. S. (1947). *Combinatorial Topology* 1. Courier Corporation. MR0025723

ANDONI, A. and INDYK, P. (2004). E2lsh: Exact Euclidean locality sensitive hashing. Technical report.

BAXTER, R., CHRISTEN, P., CHURCHES, T. et al. (2003). A comparison of fast blocking methods for record linkage. In *ACM SIGKDD* **3** 25–27.

BHATTACHARYA, I. and GETOOR, L. (2006). A latent Dirichlet model for unsupervised entity resolution. In *Proceedings of the Sixth SIAM International Conference on Data Mining* 47–58. SIAM, Philadelphia, PA. MR2337922

BRODER, A. Z. (1997a). On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of Sequences* 1997 (*SEQUENCES'*97) 21–29. IEEE Computer Society, Washington, DC.

BRODER, A. Z. (1997b). On the resemblance and containment of documents. In *The Compression and Complexity of Sequences* 21–29.

Key words and phrases. Syrian conflict, entity resolution, clustering, hashing.

CHAZELLE, B., RUBINFELD, R. and TREVISAN, L. (2005). Approximating the minimum spanning tree weight in sublinear time. *SIAM J. Comput.* **34** 1370–1379. MR2165745

CHEN, B., SHRIVASTAVA, A. and STEORTS, R. C. (2018). Supplement to "Unique entity estimation with application to the Syrian conflict." DOI:10.1214/18-AOAS1163SUPP.

CHEN, B., XU, Y. and SHRIVASTAVA, A. (2018). LSH sampling breaks the computational chicken-and-egg loop in adaptive stochastic gradient estimation.

CHRISTEN, P. (2012). A survey of indexing techniques for scalable record linkage and deduplication. *IEEE Trans. Knowl. Data Eng.* **24** 1537–1555.

CHRISTEN, P. (2014). Preparation of a real voter data set for record linkage and duplicate detection research. Tech. report.

DEMING, W. E. and GLASSER, G. J. (1959). On the problem of matching lists by samples. *J. Amer. Statist. Assoc.* **54** 403–415. MR0105768

ERDŐS, P. and RÉNYI, A. (1960). On the evolution of random graphs. *Magy. Tud. Akad. Mat. Kut. Intéz. Közl.* **5** 17–61. MR0125031

FELLEGI, I. and SUNTER, A. (1969). A theory for record linkage. *J. Amer. Statist. Assoc.* **64** 1183–1210.

FRANK, O. (1978). Estimation of the number of connected components in a graph by using a sampled subgraph. *Scand. J. Stat.* **5** 177–188. MR0515656

GIONIS, A., INDYK, P., MOTWANI, R. et al. (1999). Similarity search in high dimensions via hashing. In *Very Large Data Bases* (*VLDB*) **99** 518–529.

GRILLO, C. (2016). Judges in Habre trial cite HRDAG analysis.

GUTMAN, R., AFENDULIS, C. C. and ZASLAVSKY, A. M. (2013). A Bayesian procedure for file linking to analyze end-of-life medical costs. *J. Amer. Statist. Assoc.* **108** 34–47. MR3174601

INDYK, P. and MOTWANI, R. (1999). Approximate nearest neighbors: Towards removing the curse of dimensionality. In *STOC'*98 (*Dallas*, *TX*) 604–613. ACM, New York. MR1715608

LIANG, H., WANG, Y., CHRISTEN, P. and GAYLER, R. (2014). Noise-tolerant approximate blocking for dynamic real-time entity resolution. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* 449–460. Springer, Berlin.

LISEO, B. and TANCREDI, A. (2013). Some advances on Bayesian record linkage and inference for linked data. Available at https://pdfs.semanticscholar.org/8926/9690219564cddf7d0b91ec5f692fef13b9a9.pdf.

LUO, C. and SHRIVASTAVA, A. (2017). Arrays of (locality-sensitive) count estimators (ACE): High-speed anomaly detection via cache lookups. Preprint. Available at arXiv:1706.06664.

LUO, C. and SHRIVASTAVA, A. (2018). Scaling-up split-merge MCMC with Locality Sensitive Sampling (LSS). Preprint. Available at arXiv:1802.07444.

MCCALLUM, A., NIGAM, K. and UNGAR, L. H. (2000). Efficient clustering of high-dimensional data sets with application to reference matching. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 169–178. ACM, New York.

MCCALLUM, A. and WELLNER, B. (2004). Conditional models of identity uncertainty with application to noun coreference. In *Advances in Neural Information Processing Systems* (*NIPS'*04) 905–912. MIT Press, Cambridge, MA.

PAULEVÉ, L., JÉGOU, H. and AMSALEG, L. (2010). Locality sensitive hashing: A comparison of hash function types and querying mechanisms. *Pattern Recogn. Lett.* **31** 1348–1358.

PRICE, M., KLINGNER, J., QTIESH, A. and BALL, P. (2014). Updated statistical analysis of documentation of killings in the Syrian Arab Republic. *United Nations Office of the UN High Commissioner for Human Rights*.

PROVAN, J. S. and BALL, M. O. (1983). The complexity of counting cuts and of computing the probability that a graph is connected. *SIAM J. Comput.* **12** 777–788. MR0721012

RAJARAMAN, A. and ULLMAN, J. D. (2012). *Mining of Massive Datasets*. Cambridge Univ. Press, Cambridge, MA. MR3155538

SADINLE, M. (2014). Detecting duplicates in a homicide registry using a Bayesian partitioning approach. *Ann. Appl. Stat.* **8** 2404–2434. MR3292503

SADOSKY, P., SHRIVASTAVA, A., PRICE, M. and STEORTS, R. C. (2015). Blocking methods applied to casualty records from the Syrian conflict. ArXiv preprint. Available ar arXiv:1510.07714.

SHRIVASTAVA, A. and LI, P. (2014a). Densifying one permutation hashing via rotation for fast near neighbor search. In *Proceedings of the* 31*st International Conference on Machine Learning* 557–565.

SHRIVASTAVA, A. and LI, P. (2014b). Improved densification of one permutation hashing. In *Proceedings of the* 30*th Conference on Uncertainty in Artificial Intelligence*.

SHRIVASTAVA, A. and LI, P. (2014c). In defense of Minhash over Simhash. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics* 886–894.

SPRING, R. and SHRIVASTAVA, A. (2017a). A new unbiased and efficient class of LSH-based samplers and estimators for partition function computation in log-linear models. Preprint. Available at arXiv:1703.05160.

SPRING, R. and SHRIVASTAVA, A. (2017b). Scalable and sustainable deep learning via randomized hashing. In *Proceedings of the* 23*rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 445–454. ACM, New York.

STEORTS, R. C. (2015). Entity resolution with empirically motivated priors. *Bayesian Anal.* **10** 849–875. MR3432242

STEORTS, R. C., HALL, R. and FIENBERG, S. E. (2014). SMERED: A Bayesian approach to graphical record linkage and de-duplication. *J. Mach. Learn. Res.* **33** 922–930.

STEORTS, R. C., HALL, R. and FIENBERG, S. E. (2016). A Bayesian approach to graphical record linkage and deduplication. *J. Amer. Statist. Assoc.* **111** 1660–1672. MR3601725

STEORTS, R. C., VENTURA, S. L., SADINLE, M. and FIENBERG, S. E. (2014). A comparison of blocking methods for record linkage. In *International Conference on Privacy in Statistical Databases* 253–268.

TANCREDI, A. and LISEO, B. (2011). A hierarchical Bayesian approach to record linkage and population size problems. *Ann. Appl. Stat.* **5** 1553–1585. MR2849786

VATSALAN, D., CHRISTEN, P., O'KEEFE, C. M. and VERYKIOS, V. S. (2014). An evaluation framework for privacy-preserving record linkage. *J. Priv. Confident.* **6** 3.

WANG, Y., SHRIVASTAVA, A. and RYU, J. (2017). FLASH: Randomized algorithms accelerated over CPU-GPU for ultra-high dimensional similarity search. ArXiv preprint. Available at arXiv:1709.01190.

WINKLER, W. E. (2005). Approximate string comparator search strategies for very large administrative lists. *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*.

WINKLER, W. E. (2006). Overview of record linkage and current research directions. In *U.S. Bureau of the Census*. Washington, DC. Available at https://www.census.gov/srd/papers/pdf/rrs2006-02.pdf.

ZANELLA, G., BETANCOURT, B., MILLER, J. W., WALLACH, H., ZAIDI, A. and STEORTS, R. (2016). Flexible models for microclustering with application to entity resolution. In *Advances in Neural Information Processing Systems* 1417–1425.

The Annals of Applied Statistics
2018, Vol. 12, No. 2, 1068–1095
https://doi.org/10.1214/18-AOAS1190

ADJUSTED REGULARIZATION IN LATENT GRAPHICAL MODELS: APPLICATION TO MULTIPLE-NEURON SPIKE COUNT DATA

BY GIUSEPPE VINCI[*,1], VALÉRIE VENTURA[†,§,2]
MATTHEW A. SMITH[‡,§,3] AND ROBERT E. KASS[†,§,2]

Rice University[*], *Carnegie Mellon University*[†],
University of Pittsburgh[‡] *and Center for the Neural Basis of Cognition*[§]

A major challenge in contemporary neuroscience is to analyze data from large numbers of neurons recorded simultaneously across many experimental replications (trials), where the data are counts of neural firing events, and one of the basic problems is to characterize the dependence structure among such multivariate counts. Methods of estimating high-dimensional covariation based on ℓ_1-regularization are most appropriate when there are a small number of relatively large partial correlations, but in neural data there are often large numbers of relatively small partial correlations. Furthermore, the variation across trials is often confounded by Poisson-like variation within trials. To overcome these problems we introduce a comprehensive methodology that imbeds a Gaussian graphical model into a hierarchical structure: the counts are assumed Poisson, conditionally on latent variables that follow a Gaussian graphical model, and the graphical model parameters, in turn, are assumed to depend on physiologically-motivated covariates, which can greatly improve correct detection of interactions (nonzero partial correlations). We develop a Bayesian approach to fitting this covariate-adjusted generalized graphical model and we demonstrate its success in simulation studies. We then apply it to data from an experiment on visual attention, where we assess functional interactions between neurons recorded from two brain areas.

REFERENCES

ANDREWS, D. F. and MALLOWS, C. L. (1974). Scale mixtures of normal distributions. *J. Roy. Statist. Soc. Ser. B* **36** 99–102. MR0359122

BANERJEE, S. and GHOSAL, S. (2015). Bayesian structure learning in graphical models. *J. Multivariate Anal.* **136** 147–162. MR3321485

BEHSETA, S., BERDYYEVA, T., OLSON, C. R. and KASS, R. E. (2009). Bayesian correction for attenuation of correlation in multi-trial spike count data. *J. Neurophysiol.* **101** 2186–2193.

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

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

Key words and phrases. Bayesian inference, Gaussian graphical models, Gaussian scale mixture, high dimensionality, lasso, latent variable models, macaque prefrontal cortex, macaque visual cortex, Poisson-lognormal, sparsity, spike-counts.

BRILLINGER, D. R. (1988). Maximum likelihood analysis of spike trains of interacting nerve cells. *Biol. Cybernet.* **59** 189–200.

CHANDRASEKARAN, V., PARRILO, P. A. and WILLSKY, A. S. (2012). Latent variable graphical model selection via convex optimization. *Ann. Statist.* **40** 1935–1967. MR3059067

CHURCHLAND, A. K., KIANI, R., CHAUDHURI, R., WANG, X. J., POUGET, A. and SHADLEN, M. N. (2011). Variance as a signature of neural computations during decision making. *Neuron* **69** 818–831.

COHEN, M. R. and MAUNSELL, J. H. (2009). Attention improves performance primarily by reducing interneuronal correlations. *Nat. Neurosci.* **12** 1594–1600.

COX, D. R. and LEWIS, P. A. W. (1972). Multivariate point processes. 401–448. MR0413254

ECKER, A. S., BERENS, P., COTTON, R. J., SUBRAMANIYAN, M., DENFIELD, G. H., CADWELL, C. R., SMIRNAKIS, S. M., BETHGE, M. and TOLIAS, A. S. (2014). State dependence of noise correlations in macaque primary visual cortex. *Neuron* **82** 235–248.

EFRON, B., TIBSHIRANI, R., STOREY, J. D. and TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151–1160. MR1946571

FAN, J., FENG, Y. and WU, Y. (2009). Network exploration via the adaptive lasso and SCAD penalties. *Ann. Appl. Stat.* **3** 521–541. MR2750671

FIENBERG, S. E. (1974). Stochastic models for single neuron firing trains: A survey. *Biometrics* **30** 399–427. MR0359082

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

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

GEORGOPOULOS, A. P. and ASHE, J. (2000). One motor cortex, two different views. *Nat. Neurosci.* **3** 964–963; author reply 965.

GIRAUD, C. and TSYBAKOV, A. (2012). Discussion: Latent variable graphical model selection via convex optimization [MR3059067]. *Ann. Statist.* **40** 1984–1988. MR3059071

GORIS, R. L., MOVSHON, J. A. and SIMONCELLI, E. P. (2014). Partitioning neuronal variability. *Nat. Neurosci.* **17** 858–865.

HINNE, M., AMBROGIONI, L., JANSSEN, R. J., HESKES, T. and VAN GERVEN, M. A. J. (2014). Structurally-informed Bayesian functional connectivity analysis. *NeuroImage* **86** 294–305.

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

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

KASS, R. E., EDEN, U. T. and BROWN, E. N. (2014). *Analysis of Neural Data*. Springer, New York. MR3244261

KASS, R. E., VENTURA, V. and BROWN, E. N. (2005). Statistical issues in the analysis of neuronal data. *J. Neurophysiol.* **94** 8–25.

KASS, R. E., AMARI, S. I., ARAI, K., BROWN, E. N., DIEKMAN, C. O., DIESMANN, M. . . . and FUKAI, T. (2018). Computational neuroscience: Mathematical and statistical perspectives. *Ann. Rev. Statist. Appl.* **5** 183–214.

MAUNSELL, J. H. (2015). Neuronal mechanisms of visual attention. *Ann. Rev. Vision Sci.* **1** 373–391.

MAZUMDER, R. and HASTIE, T. (2012). The graphical lasso: New insights and alternatives. *Electron. J. Stat.* **6** 2125–2149. MR3020259

MITCHELL, J. F., SUNDBERG, K. A. and REYNOLDS, J. H. (2009). Spatial attention decorrelates intrinsic activity fluctuations in macaque area V4. *Neuron* **63** 879–888.

NG, B., VAROQUAUX, G., POLINE, J. B. and THIRION, B. (2012). A novel sparse graphical approach for multimodal brain connectivity inference. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 707–714. Springer, Berlin, Heidelberg.

PERKEL, D. H., GERSTEIN, G. L. and MOORE, G. P. (1967a). Neuronal spike trains and stochastic point processes: I. The single spike train. *Biophys. J.* **7** 391–418.

PERKEL, D. H., GERSTEIN, G. L. and MOORE, G. P. (1967b). Neuronal spike trains and stochastic point processes: II. Simultaneous spike trains. *Biophys. J.* **7** 419–440.

PINEDA-PARDO, J. A., BRUÑA, R., WOOLRICH, M., MARCOS, A., NOBRE, A. C., MAESTÚ, F. and VIDAURRE, D. (2014). Guiding functional connectivity estimation by structural connectivity in MEG: An application to discrimination of conditions of mild cognitive impairment. *NeuroImage* **101** 765–777.

POLSON, N. G., SCOTT, J. G. and WINDLE, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *J. Amer. Statist. Assoc.* **108** 1339–1349. MR3174712

RABINOWITZ, N. C., GORIS, R. L., COHEN, M. and SIMONCELLI, E. P. (2015). Attention stabilizes the shared gain of V4 populations. *eLife* **4** e08998.

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

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

RUFF, D. A. and COHEN, M. R. (2016). Stimulus dependence of correlated variability across cortical areas. *J. Neurosci.* **36** 7546–7556.

SCOTT, J. G., KELLY, R. C., SMITH, M. A., ZHOU, P. and KASS, R. E. (2015). False discovery rate regression: An application to neural synchrony detection in primary visual cortex. *J. Amer. Statist. Assoc.* **110** 459–471. MR3367240

SMITH, M. A. and KOHN, A. (2008). Spatial and temporal scales of neuronal correlation in primary visual cortex. *J. Neurosci.* **28** 12591–12603.

SMITH, M. A. and SOMMER, M. A. (2013). Spatial and temporal scales of neuronal correlation in visual area V4. *J. Neurosci.* **33** 5422–5432.

SNYDER, A. C., MORAIS, M. J. and SMITH, M. A. (2016). Dynamics of excitatory and inhibitory networks are differentially altered by selective attention. *J. Neurophysiol.* **116** 1807–1820.

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

VINCI, G., VENTURA, V., SMITH, M. A. and KASS, R. E. (2016). Separating spike count correlation from firing rate correlation. *Neural Comput.* **28** 849–881.

VINCI, G., VENTURA, V., SMITH, M. A. and KASS, R. E. (2018a). Adjusted regularization of cortical covariance. *J. Comput. Neurosci.* To appear.

VINCI, G., VENTURA, V., SMITH, M. A. and KASS, R. E. (2018b). Supplement to "Adjusted regularization in latent graphical models: Application to multiple-neuron spike count data." DOI:10.1214/18-AOAS1190SUPP.

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

WANG, H. (2015). Scaling it up: Stochastic search structure learning in graphical models. *Bayesian Anal.* **10** 351–377. MR3420886

WILSON, H. R. and COWAN, J. D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophys. J.* **12** 1–24.

YATSENKO, D., JOSIĆ, K., ECKER, A. S., FROUDARAKIS, E., COTTON, R. J. and TOLIAS, A. S. (2015). Improved estimation and interpretation of correlations in neural circuits. *PLoS Comput. Biol.* **11** e1004083.

YU, B. M., CUNNINGHAM, J. P., SANTHANAM, G., RYU, S. I., SHENOY, K. V. and SAHANI, M. (2009). Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. In *Advances in Neural Information Processing Systems* 1881–1888.

YUAN, M. (2012). Discussion: Latent variable graphical model selection via convex optimization [MR3059067]. *Ann. Statist.* **40** 1968–1972. MR3059068

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

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. 2, 1096–1123
https://doi.org/10.1214/18-AOAS1191

DISCOVERING POLITICAL TOPICS IN FACEBOOK DISCUSSION THREADS WITH GRAPH CONTEXTUALIZATION

BY YILIN ZHANG*,[1], MARIE POUX-BERTHE†,[2], CHRIS WELLS‡,
KAROLINA KOC-MICHALSKA†,[2] AND KARL ROHE*,[1]

University of Wisconsin-Madison, Audencia Business School† and Boston University‡*

We propose a graph contextualization method, `pairGraphText`, to study political engagement on Facebook during the 2012 French presidential election. It is a spectral algorithm that contextualizes graph data with text data for online discussion thread. In particular, we examine the Facebook posts of the eight leading candidates and the comments beneath these posts. We find evidence of both (i) candidate-centered structure, where citizens primarily comment on the wall of one candidate and (ii) issue-centered structure (i.e., on political topics), where citizens' attention and expression is primarily directed toward a specific set of issues (e.g., economics, immigration, etc). To identify issue-centered structure, we develop `pairGraphText`, to analyze a network with high-dimensional features on the interactions (i.e., text). This technique scales to hundreds of thousands of nodes and thousands of unique words. In the Facebook data, spectral clustering without the contextualizing text information finds a mixture of (i) candidate and (ii) issue clusters. The contextualized information with text data helps to separate these two structures. We conclude by showing that the novel methodology is consistent under a statistical model.

REFERENCES

ADAMIC, L. A. and GLANCE, N. (2005). The political blogosphere and the 2004 us election: Divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery* 36–43. ACM.

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

BAKSHY, E., MESSING, S. and ADAMIC, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science* **348** 1130–1132. MR3380630

BINKIEWICZ, N., VOGELSTEIN, J. T. and ROHE, K. (2017). Covariate-assisted spectral clustering. *Biometrika* **104** 361–377. MR3698259

BLEI, D. M. (2012). Probabilistic topic models. *Commun. ACM* **55** 77–84.

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

BOYD, D. N. and ELLISON, N. B. (2007). Social network sites: Definition, history, and scholarship. *J. Comput.-Mediat. Commun.* **13** 210–230.

CHANG, J. and BLEI, D. (2009). Relational topic models for document networks. In *Artificial Intelligence and Statistics* 81–88.

Key words and phrases. Network, Facebook, topic, spectral clustering, node covariate, stochastic co-Blockmodel.

CHANG, J. and BLEI, D. M. (2010). Hierarchical relational models for document networks. *Ann. Appl. Stat.* **4** 124–150. MR2758167

CHOY, M., CHEONG, M. L., LAIK, M. N. and SHUNG, K. P. (2011). A sentiment analysis of Singapore presidential election 2011 using Twitter data with census correction. Preprint. Available at arXiv:1108.5520.

COLLEONI, E., ROZZA, A. and ARVIDSSON, A. (2014). Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *J. Commun.* **64** 317–332.

GONZALEZ-BAILON, S., KALTENBRUNNER, A. and BANCHS, R. E. (2010). The structure of political discussion networks: A model for the analysis of online deliberation. *J. Inf. Technol.* **25** 230–243.

GRIMMER, J. and STEWART, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Polit. Anal.* **21** 267–297.

HEBSHI, S. and O'GARA (2011). The rohe of online social networking in the 2008 democratic presidential primary campains. Preprint. Available at http://www.shoshanahebshi.com/wp-content/uploads/2011/08/Social-Medias-role-in-primary-campaigns.pdf.

HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. MR0718088

JOACHIMS, T. (1996). A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. Technical report, Carnegie-Mellon Univ., Pittsburgh, PA, Dept. of Computer Science.

KAPLAN, A. M. and HAENLEIN, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Bus. Horiz.* **53** 59–68.

KARRER, B. and NEWMAN, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **83** 016107. MR2788206

KIM, Y. M. (2009). Issue publics in the new information environment: Selectivity, domain specificity, and extremity. *Communic. Res.* **36** 254–284.

KREISS, D. and MCGREGOR, S. C. (2018). Technology firms shape political communication: The work of Microsoft, Facebook, Twitter, and Google with campaigns during the 2016 US presidential cycle. *Polit. Commun.* **35** 155–177.

KUSHIN, M. J. and KITCHENER, K. (2009). Getting political on social network sites: Exploring online political discourse on Facebook. *First Monday* **14** 11-2.

PANG, B. and LEE, L. (2008). Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.* **2** 1–135.

PAPACHARISSI, Z. (2002). The virtual sphere: The Internet as a public sphere. *New Media Soc.* **4** 9–27.

QIN, T. and ROHE, K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems* 3120–3128.

RAMAGE, H. R., CONNOLLY, L. E. and COX, J. S. (2009). Comprehensive functional analysis of Mycobacterium tuberculosis toxin-antitoxin systems: Implications for pathogenesis, stress responses, and evolution. *PLoS Genet.* **5** e1000767.

RAMOS, J. (2003). Using TF-IDF to determine word relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning*.

ROBERTSON, S. P., VATRAPU, R. K. and MEDINA, R. (2010). Off the wall political discourse: Facebook use in the 2008 US presidential election. *Information Polity* **15** 11–31.

ROHE, K., QIN, T. and YU, B. (2016). Co-clustering directed graphs to discover asymmetries and directional communities. *Proc. Natl. Acad. Sci. USA* **113** 12679–12684. MR3576189

SALTON, G., WONG, A. and YANG, C.-S. (1975). A vector space model for automatic indexing. *Commun. ACM* **18** 613–620.

SIVIC, J. and ZISSERMAN, A. (2003). Video Google: A text retrieval approach to object matching in videos. In *Proceedings Ninth IEEE International Conference on Computer Vision*. IEEE.

STIEGLITZ, S. and DANG-XUAN, L. (2012). Political communication and influence through microblogging—an empirical analysis of sentiment in Twitter messages and retweet behavior. In 45*th Hawaii International Conference on System Science*. IEEE.

STIEGLITZ, S. and DANG-XUAN, L. (2013). Social media and political communication: A social media analytics framework. *Soc. Netw. Anal. Min.* **3** 1277–1291.

TUMASJAN, A., SPRENGER, T. O., SANDNER, P. G. and WELPE, I. M. (2011). Election forecasts with Twitter: How 140 characters reflect the political landscape. *Soc. Sci. Comput. Rev.* **29** 402–418.

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

WANG, H., CAN, D., KAZEMZADEH, A., BAR, F. and NARAYANAN, S. (2012). A system for real-time Twitter sentiment analysis of 2012 US presidential election cycle. In *Proceedings of the ACL* 2012 *System Demonstrations* 115–120. Association for Computational Linguistics.

WATTAL, S., SCHUFF, D., MANDVIWALLA, M. and WILLIAMS, C. B. (2010). Web 2.0 and politics: The 2008 US presidential election and an e-politics research agenda. *MIS Q.* 669–688.

WEBSTER, J. G. (2014). *The Marketplace of Attention: How Audiences Take Shape in a Digital Age*. MIT Press.

WELLMAN, B., HAASE, A. Q., WITTE, J. and HAMPTON, K. (2001). Does the Internet increase, decrease, or supplement social capital? Social networks, participation, and community commitment. *Am. Behav. Sci.* **45** 436–455.

WILLIAMS, C. B. and GULATI, G. J. (2009). Explaining Facebook support in the 2008 congressional election cycle. OpenSIUC Working Papers, 26.

WILLIAMS, C. B. and GULATI, G. J. (2013). Social networks in political campaigns: Facebook and the congressional elections of 2006 and 2008. *New Media Soc.* **15** 52–71.

WITTEN, D. M. (2011). Classification and clustering of sequencing data using a Poisson model. *Ann. Appl. Stat.* **5** 2493–2518. MR2907124

ZHANG, Y., POUX-BERTHE, M., WELLS, C., KOC-MICHALSKA, K. and ROHE, K. (2018). Supplement to "Discovering political topics in Facebook discussion threads with graph contextualization." DOI:10.1214/18-AOAS1191SUPP.

The Annals of Applied Statistics
2018, Vol. 12, No. 2, 1124–1156
https://doi.org/10.1214/18-AOAS1194

PROVIDING ACCESS TO CONFIDENTIAL RESEARCH DATA THROUGH SYNTHESIS AND VERIFICATION: AN APPLICATION TO DATA ON EMPLOYEES OF THE U.S. FEDERAL GOVERNMENT[1]

BY ANDRÉS F. BARRIENTOS*, ALEXANDER BOLTON†, TOM BALMAT*, JEROME P. REITER*, JOHN M. DE FIGUEIREDO*, ASHWIN MACHANAVAJJHALA*, YAN CHEN*, CHARLEY KNEIFEL* AND MARK DELONG*

Duke University and Emory University*†*

Data stewards seeking to provide access to large-scale social science data face a difficult challenge. They have to share data in ways that protect privacy and confidentiality, are informative for many analyses and purposes, and are relatively straightforward to use by data analysts. One approach suggested in the literature is that data stewards generate and release synthetic data, that is, data simulated from statistical models, while also providing users access to a verification server that allows them to assess the quality of inferences from the synthetic data. We present an application of the synthetic data plus verification server approach to longitudinal data on employees of the U.S. federal government. As part of the application, we present a novel model for generating synthetic career trajectories, as well as strategies for generating high dimensional, longitudinal synthetic datasets. We also present novel verification algorithms for regression coefficients that satisfy differential privacy. We illustrate the integrated use of synthetic data plus verification via analysis of differentials in pay by race. The integrated system performs as intended, allowing users to explore the synthetic data for potential pay differentials and learn through verifications which findings in the synthetic data hold up and which do not. The analysis on the confidential data reveals pay differentials across races not documented in published studies.

REFERENCES

ABOWD, J. M. and SCHMUTTE, I. M. (2017). Revisting the economics of privacy: Population statistics and confidentiality protection as public goods. Technical report, Working paper 17-37, Center for Economic Studies, U.S. Census Bureau.

ABOWD, J., STINSON, M. and BENEDETTO, G. (2006). Final report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project. Technical report, U.S. Census Bureau Longitudinal Employer–Household Dynamics Program. Available at http://www.census.gov/sipp/synth_data.html.

ABOWD, J. and VILHUBER, L. (2008). How protective are synthetic data? In *Privacy in Statistical Databases* (J. Domingo-Ferrer and Y. Saygun, eds.) 239–246. Springer, New York.

ALTONJI, J. G. and BLANK, R. M. (1999). Race and gender in the labor market. In *Handbook of Labor Economics* **3** 3143–3259. Elsevier, Amsterdam.

Key words and phrases. Disclosure, privacy, public, remote, synthetic.

BARAK, B., CHAUDHURI, K., DWORK, C., KALE, S., MCSHERRY, F. and TALWAR, K. (2007). Privacy, accuracy, and consistency too: A holistic solution to contingency table release. In *Proceedings of the* 26*th ACM SIGMOD International Conference on Management of Data / Principles of Database Systems* 273–282.

BARRIENTOS, A. F., BOLTON, A., BALMAT, T., REITER, J. P., DE FIGUEIREDO, J. M., MACHANAVAJJHALA, A., CHEN, Y., KNEIFEL, C. and DELONG, M. (2018). Supplement to "Providing access to confidential research data through synthesis and verification: An application to data on employees of the U.S. federal government." DOI:10.1214/18-AOAS1194SUPPA, DOI:10.1214/18-AOAS1194SUPPB, DOI:10.1214/18-AOAS1194SUPPC.

BLACK, D. A., KOLESNIKOVA, N., SANDERS, S. G. and TAYLOR, L. J. (2013). The role of location in evaluating racial wage disparity. *IZA J. Labor Econ.* **2** 2.

BLAU, F. D. and KAHN, L. M. (2017). The gender wage gap: Extent, trends, and expectations. *J. Econ. Lit.* **55** 789–865.

BLUM, A., LIGETT, K. and ROTH, A. (2008). A learning theory approach to non-interactive database privacy. In *Proceedings of the* 40*th ACM SIGACT Symposium on Theory of Computing* 609–618. ACM, New York. MR2582916

BOLTON, A. and DE FIGUEIREDO, J. M. (2016). Why have federal wages risen so rapidly? Technical report, Duke Univ. Law School.

BOLTON, A. and DE FIGUEIREDO, J. M. (2017). Measuring and explaining the gender wage gap in the U.S. federal government. Technical report, Duke Univ. Law School.

BOLTON, A., DE FIGUEIREDO, J. M. and LEWIS, D. E. (2016). Elections, ideology, and turnover in the U.S. federal government. Technical report, National Bureau of Economics Research Working Paper 22932.

BORJAS, G. J. (1980). Wage determination in the federal government: The role of constituents and bureaucrats. *J. Polit. Econ.* **88** 1110–1147.

BORJAS, G. J. (1982). The politics of employment discrimination in the federal bureaucracy. *J. Law Econ.* **25** 271–299.

BORJAS, G. J. (1983). The measurement of race and gender wage differentials: Evidence from the federal sector. *Ind. Labor Relat. Rev.* **37** 79–91.

CALLIER, V. (2015). How fake data could protect real people's privacy. *The Atlantic*, July 30, 2015.

CAMERON, A. C. and MILLER, D. L. (2015). A practitioner's guide to cluster-robust inference. *J. Hum. Resour.* **50** 317–373.

CANCIO, A. S., EVANS, T. D. and MAUME, D. J. J. (1996). Reconsidering the declining significance of race: Racial differences in early career wages. *Am. Sociol. Rev.* **61** 541–556.

CARD, D. and LEMIEUX, T. (1994). Changing wage structure and black–white wage differentials. *Am. Econ. Rev.* **84** 29–33.

CHAREST, A. S. (2010). How can we analyze differentially private synthetic datasets. *J. Priv. Confid.* **2** Article 3.

CHARLES, J. (2003). Diversity management: An exploratory assessment of minority group representation in state government. *Public Pers. Manag.* **32** 561–577.

CHEN, Y., MACHANAVAJJHALA, A., REITER, J. P. and BARRIENTOS, A. F. (2016). Differentially private regression diagnostics. In *Proceedings of the IEEE* 16*th International Conference on Data Mining* 81–90.

COMMISSION ON EVIDENCE-BASED POLICYMAKING (2017). The promise of evidence-based policymaking.

DRECHSLER, J. (2011). *Synthetic Datasets for Statistical Disclosure Control*: *Theory and Implementation*. *Lecture Notes in Statistics* **201**. Springer, New York. MR2809912

DRECHSLER, J., DUNDLER, A., BENDER, S., RÄSSLER, S. and ZWICK, T. (2008). A new approach for disclosure control in the IAB establishment panel—Multiple imputation for a better data access. *AStA Adv. Stat. Anal.* **92** 439–458. MR2461314

DWORK, C. (2006). Differential privacy. In *Automata*, *Languages and Programming*. *Part II*. *Lecture Notes in Computer Science* **4052** 1–12. Springer, Berlin. MR2307219

DWORK, C. and ROTH, A. (2013). The algorithmic foundations of differential privacy. *Found*. *Trends Theor*. *Comput*. *Sci*. **9** 211–487. MR3254020

ELRINGSSON, U., PIHUR, V. and KOROLOVA, A. (2014). Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the* 2014 *ACM SIGSAC Conference on Computer and Communications Security* 1054–1067.

FIENBERG, S. E. (1994). A radical proposal for the provision of micro-data samples and the preservation of confidentiality. Technical report, Dept. of Statistics, Carnegie-Mellon Univ.

GOVERNMENT ACCOUNTABILITY OFFICE (2009). Gender pay gap in the federal workforce narrows as differences in occupation, education, and experience diminish. Technical report, Government Accountability Office, Washington, DC.

HARDT, M., LIGETT, K. and MCSHERRY, F. (2012). A simple and practical algorithm for differentially private data release. *Adv*. *Neural Inf*. *Process*. *Syst*. **25** 2348–2356.

HARDT, M. and ROTHBLUM, G. N. (2010). A multiplicative weights mechanism for privacy-preserving data analysis. In 2010 *IEEE* 51*st Annual Symposium on Foundations of Computer Science—FOCS* 2010 61–70. IEEE Computer Soc., Los Alamitos, CA. MR3024776

HAWALA, S. (2008). Producing partially synthetic data to avoid disclosure. In *Proceedings of the Joint Statistical Meetings*. American Statistical Association, Alexandria, VA.

HOLAN, S. H., TOTH, D., FERREIRA, M. A. R. and KARR, A. F. (2010). Bayesian multiscale multiple imputation with implications for data confidentiality. *J*. *Amer*. *Statist*. *Assoc*. **105** 564–577. MR2759932

HOOVER, G. A., COMPTON, R. A. and GIEDEMAN, D. C. (2015). The impact of economic freedom on the black/white income gap. *Am*. *Econ*. *Rev*. *Pap*. *Proc*. **105** 587–592.

KARR, A. F. and REITER, J. P. (2014). Using statistics to protect privacy. In *Privacy*, *Big Data*, *and the Public Good*: *Frameworks for Engagement* (J. Lane, V. Stodden, S. Bender and H. Nissenbaum, eds.) 276–295. Cambridge Univ. Press, Cambridge.

KARR, A. F., FULP, W. J., VERA, F., YOUNG, S. S., LIN, X. and REITER, J. P. (2007). Secure, privacy-preserving analysis of distributed databases. *Technometrics* **49** 335–345. MR2408637

KARWA, V. and SLAVKOVIĆ, A. (2016). Inference using noisy degrees: Differentially private β-model and synthetic graphs. *Ann*. *Statist*. **44** 87–112. MR3449763

KIM, C.-K. (2004). Women and minorities in state government agencies. *Public Pers*. *Manag*. **33** 165–180.

KINNEY, S. K., REITER, J. P., REZNEK, A. P., MIRANDA, J., JARMIN, R. S. and ABOWD, J. M. (2011). Towards unrestricted public use business microdata: The synthetic Longitudinal Business Database. *Int*. *Stat*. *Rev*. **79** 363–384.

LEWIS, G. B. (1998). Continuing progress toward racial and gender pay equality in the federal service: An update. *Rev*. *Public Pers*. *Adm*. **18** 23–40.

LEWIS, G. B. and DURST, S. L. (1995). Will locality pay solve recruitment and retention problems in the federal civil service? *Public Adm*. *Rev*. **55** 371–380.

LEWIS, G. B. and NICE, D. (1994). Race, sex, and occupational segregation in state and local governments. *Am*. *Rev*. *Public Adm*. **24** 393–410.

LITTLE, R. J. A. (1993). Statistical analysis of masked data. *J*. *Off*. *Stat*. **9** 407–426.

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

LLORENS, J. J., WENGER, J. B. and KELLOUGH, J. E. (2007). Choosing public sector employment: The impact of wages on the representation of women and minorities in state bureaucracies. *J*. *Public Adm*. *Res*. *Theory* **18** 397–413.

MACHANAVAJJHALA, A., KIFER, D., ABOWD, J., GEHRKE, J. and VILHUBER, L. (2008). Privacy: Theory meets practice on the map. In *IEEE* 24*th International Conference on Data Engineering* 277–286.

MAXWELL, N. L. (1994). The effect of black–white wage differences of differences in the quantity and quality of education. *Ind. Labor Relat. Rev.* **47** 249–264.

MCCABE, B. C. and STREAM, C. (2000). Diversity by the numbers: Changes in state and local government workforces, 1980–1995. *Public Pers. Manag.* **29** 93–106.

MCCALL, L. (2001). Sources of racial wage inequality in metropolitan labor markets: Racial, ethnic, and gender differences. *Am. Sociol. Rev.* **66** 520–541.

MCCLURE, D. and REITER, J. P. (2012). Towards providing automated feedback on the quality of inferences from synthetic datasets. *J. Priv. Confid.* **4** Article 8.

MCCLURE, D. and REITER, J. P. (2016). Assessing disclosure risks for synthetic data with arbitrary intruder knowledge. *Stat. J. Int. Assoc. Off. Stat.* **32** 109–126.

MIR, D., ISAACMAN, S., CACERES, R., MARTONOSI, M. and WRIGHT, R. N. (2013). DP–WHERE: Differentially private modeling of human mobility. In *Proceedings of the IEEE Conference on Big Data* 580–588.

NARAYANAN, A. and SHMATIKOV, V. (2008). Robust de-anonymization of large sparse datasets. In *Proceedings of the IEEE Symposium on Security and Privacy* 111–125.

NEAL, D. and JOHNSON, W. R. (2003). The role of pre-market factors in black–white wage differences. *J. Polit. Econ.* **87** 567–594.

NISSIM, K., RASKHODNIKOVA, S. and SMITH, A. (2007). Smooth sensitivity and sampling in private data analysis. In *STOC'07—Proceedings of the* 39*th Annual ACM Symposium on Theory of Computing* 75–84. ACM, New York. MR2402430

O'NEILL, J. (1990). The role of human capital in earnings differences between black and white men. *J. Econ. Perspect.* **108** 937–975.

PARRY, M. (2011). Harvard researchers accused of breaching students' privacy. *The Chronicle of Higher Education*, July 11, 2011.

RAGHUNATHAN, T. E., REITER, J. P. and RUBIN, D. B. (2003). Multiple imputation for statistical disclosure limitation. *J. Off. Stat.* **19** 1–16.

REITER, J. P. (2005a). Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study. *J. Roy. Statist. Soc. Ser. A* **168** 185–205. MR2113234

REITER, J. P. (2005b). Using CART to generate partially synthetic, public use microdata. *J. Off. Stat.* **21** 441–462.

REITER, J. P., OGANIAN, A. and KARR, A. F. (2009). Verification servers: Enabling analysts to assess the quality of inferences from public use data. *Comput. Statist. Data Anal.* **53** 1475–1482. MR2657106

REITER, J. P. and RAGHUNATHAN, T. E. (2007). The multiple adaptations of multiple imputation. *J. Amer. Statist. Assoc.* **102** 1462–1471. MR2372542

REITER, J. P., WANG, Q. and ZHANG, B. (2014). Bayesian estimation of disclosure risks in multiply imputed, synthetic data. *J. Priv. Confid.* **6** Article 2.

RUBIN, D. B. (1993). Discussion: Statistical disclosure limitation. *J. Off. Stat.* **9** 462–468.

SAKANO, R. (2002). Are black and white income distributions converging? Time series analysis. *Rev. Black Polit. Econ.* **30** 91–106.

SPRINGER, L. M. (2005). Memorandum for chief human capital officers. Office of Personnel Management, November 9, 2005.

SWEENEY, L. (1997). Weaving technology and policy together to maintain confidentiality. *J. Law Med. Ethics* **25** 98–110.

SWEENEY, L. (2015). Only you, your doctor, and many others may know. *Technology Science*, September 29, 2015.

TANG, J., KOROLOVA, A., BAI, X., WANG, X. and WANG, X. (2017). Privacy loss in Apple's implementation of differential privacy on MacOS 10.12. Preprint. Available at 2709.03753.

ULLMAN, J. (2015). Private multiplicative weights beyond linear queries. In *Proceedings of the* 34*th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems* 303–312.

VILHUBER, L., ABOWD, J. A. and REITER, J. P. (2016). Synthetic establishment microdata around the world. *Stat. J. Int. Assoc. Off. Stat.* **32** 65–68.

WANG, H. and REITER, J. P. (2012). Multiple imputation for sharing precise geographies in public use data. *Ann. Appl. Stat.* **6** 229–252. MR2951536

The Annals of Applied Statistics
2018, Vol. 12, No. 2, 1157–1179
https://doi.org/10.1214/18-AOAS1184

THE INTERLOCKING WORLD OF SURVEYS AND EXPERIMENTS

BY STEPHEN E. FIENBERG AND JUDITH M. TANUR*

*Stony Brook University**

Random sampling and randomized experimentation are inextricably linked. Beginning with their common origins in the work of Fisher and Neyman from the 1920s and the 1930s, one can trace the development of parallel concepts and structures in the two areas (see Fienberg and Tanur [*Bull. Int. Stat. Inst.* **51** (1985) Art. ID 10.1; *Int. Stat. Rev.* **55** (1987) 75–96]). One of the more important lessons to be learned from the parallel concepts and structures is that they can profitably be linked and intertwined, with sampling embedded in experiments and formal experimental structures embedded in sampling designs.

In this paper, we trace some of parallels between sampling theory and theory of experimental design. We then explore some of the ways that experimental and sampling structures have been combined in statistical practice and the principles that underlie their combination; we also make some suggestions toward the improvement of practice.

REFERENCES

ARONSON, E., BREWER, M. and CARLSMITH, J. M. (1985). Experimentation in social psychology. In *Handbook of Social Psychology*, *Vol. I*, 3rd ed. (G. Lindzey and E. Aronson, eds.). Random House, New York.

BAILAR, B. A. (1983). Interpenetrating subsamples. In *Encyclopedia of Statistical Sciences*, *Vol.* 4 (S. Kotz and N. Johnson, eds.) 197–201. Wiley, New York.

BAILAR, B. and BIEMER, P. (1984). Some methods for evaluating nonsampling error in household censuses and surveys. In *W. G. Cochran's Impact on Statistics* (P. S. R. S. Rao and J. Sedransk, eds.) 253–275. Wiley, New York.

BARTLETT, M. S. (1978). Fisher, R. A. In *International Encyclopedia of Statisitcs* (W. H. Kruskal and J. M. Tanur, eds.) 352–358. Free Press, New York.

BOX, J. F. (1978). *R. A. Fisher*: *The Life of a Scientist*. Wiley, New York. MR0500579

CAMPBELL, D. P. (1957). Factors relevant to the validity of experiments in social settings. *Psychol. Bull.* **54** 297–312.

CAMPBELL, D. P. and STANLEY, J. C. (1963). Experimental and quasi-experimental designs for research. In *Handbook of Research on Teaching* (N. L. Gage, ed.) 171–246. Rand McNally, Chicago, IL.

COCHRAN, W. G. (1968). Errors of measurement in statistics. *Technometrics* **10** 637–666.

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

COCHRAN, W. G. and COX, G. M. (1957). *Experimental Designs*, 2nd ed. Wiley, New York; Chapman & Hall, London. MR0085682

COOK, T. D. and CAMPBELL, D. P. (1979). *Quasi-Experiments*: *Design and Analysis Issues for Field Settings*. Rand McNally, Skokie, IL.

Key words and phrases. External validity, internal validity, interviewer effects, randomized experiments, sample surveys, control, experimental design, embedding, randomization, sampling design.

WOOD (A. J.) RESEARCH CORPORATION (1959). *Woodchips*, 4, No. 1.

COX, D. R. (1958). *Planning of Experiments*. Wiley, New York; Chapman & Hall, London. MR0095561

DURBIN, J. and STUART, A. (1951). Differences in response rates of experienced and inexperienced interviewers. *J. Roy. Statist. Soc. Ser. A* **114** 163–206.

FATHI, D. C., SCHOOLER, J. and LOFTUS, E. E. (1984). Moving survey problems into the cognitive psychology laboratory. In *Proceedings of the American Statistical Association Section on Survey Research Methods* 19–21. Amer. Statist. Assoc., Washington, DC.

FEDERER, W. T. (1977). Sampling, blocking, and model considerations for split plot and split block designs. *Biom. J.* **19** 181–200.

FELLEGI, I. P. (1964). Response variance and its estimation. *J. Amer. Statist. Assoc.* **59** 1016–1041.

FIENBERG, S. E. (1971). Randomization and social affairs: The 1970 draft lottery. *Science* **171** 255–261.

FIENBERG, S. E., LOFTUS, E. E. and TANUR, J. M. (1985). Cognitive aspects of health survey methodology: An overview. *Milbank Mem. Fund Q.* **63** 547–564.

FIENBERG, S. E., SINGER, B. and TANUR, J. M., (1985). Large scale social experimentation in the U.S.A. In *A Celebration of Statistics* (A. C. Atkinson and S. E. Fienberg, eds.) 287–326. Springer, New York.

FIENBERG, S. E. and TANUR, J. M. (1985). A long and honorable tradition: Intertwining concepts and constructs in experimental design and sample surveys. *Bull. Int. Stat. Inst.* **51** Art. ID 10.1. MR0886247

FIENBERG, S. E. and TANUR, J. M. (1987). Experimental and sampling structures: Parallels diverging and meeting. *Int. Stat. Rev.* **55** 75–96. MR0962943

FIENBERG, S. E. and TANUR, J. M. (1989). Combining cognitive and statistical approaches to survey design. *Science* **243** 1017–1022. MR0986238

FISHER, R. A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh.

FISHER, R. A. (1926). The arrangement of field experiments. *J. Minist. Agric.* **33** 503–513.

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

FRANKEL, L. R. and STOCK, J. S. (1942). On the sample survey of unemployment. *J. Amer. Statist. Assoc.* **37** 77–80.

GILBERT, J. P., LIGHT, R. J. and MOSTELLER, F. (1975). Assessing social innovations: An empirical base for policy. In *Evaluation and Experiment: Some Critical Issues in Assessing Social Programs* (C. A. Bennett and A. A. Lumsdaine, eds.) 39–193. Academic Press, New York.

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

GROVES, R. M. and MAGILAVY, L. J. (1986). Measuring and explaining interviewer effects in centralized telephone surveys. *Public Opin. Q.* **50** 251–266.

HANSEN, M. H., HURWITZ, W. N. and BERSHAD, M. A. (1961). Measurement errors in censuses and surveys. *Bull. Int. Stat. Inst.* **38** 359–374.

HANSEN, M. H., HURWITZ, W. N. and MADOW, W. G. (1953a). *Sample Survey Methods and Theory, Vol. I: Methods and Applications*. Wiley, New York; Chapman & Hall, London. MR0058171

HANSEN, M. H., HURWITZ, W. N. and MADOW, W. G. (1953b). *Sample Survey Methods and Theory, Vol. II: Theory*. Wiley, New York; Chapman & Hall, London. MR0058172

HANSEN, M. H., HURWITZ, W. N., MARKS, E. S. and MAULDIN, W. P. (1951). Response errors in surveys. *J. Amer. Statist. Assoc.* **46** 147–190.

HARTLEY, H. O. and RAO, J. N. K. (1978). Estimation of nonsampling variance components in sample surveys. In *Survey Sampling and Measurement* (N. K. Namboodiri, ed.) 35–43. Academic Press, New York.

JABINE, T. B. and ROTHWELL, N. D. (1970). Split-panel tests of census and survey questionnaires. In *Proceedings of the American Statistical Association Social Statistics Section* 4–13. Amer. Statist. Assoc., Washington, DC.

JABINE, T. B., STRAF, M., TANUR, J. M. and TORANGEAU, R., eds. (1984). *Cognitive Aspects of Survey Methodology*: *Building a Bridge Between Disciplines*. National Academy Press, Washington, DC.

KAIZAR, E. E. (2011). Estimating treatment effect via simple cross design synthesis. *Stat. Med.* **30** 2986–3009. MR2851395

KEMPTHORNE, O. (1952). *The Design and Analysis of Experiments*. Wiley, New York; Chapman & Hall, London. MR0045368

KEMPTHORNE, O. (1955). The randomization theory of experimental inference. *J. Amer. Statist. Assoc.* **50** 946–967. MR0071696

KISH, L. (1962). Studies of interviewer variance for attitudinal variables. *J. Amer. Statist. Assoc.* **57** 92–115.

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

KISH, L. and FRANKEL, M. R. (1974). Inference from complex samples. *J. Roy. Statist. Soc. Ser. B* **36** 1–37. MR0365812

KRUSKAL, W. and MOSTELLER, F. (1980). Representative sampling. IV. The history of the concept in statistics, 1895–1939. *Int. Stat. Rev.* **48** 169–195. MR0586104

LAVRAKAS, P. J., TRAUGOTT, M., KENNEDY, C., DE LEEUW, E., HOLBROOK, A. and WEST, B., eds. (2018). *Experimental Methods in Survey Research*: *Techniques That Combine Random Assignment with Random Probability Sampling*. Wiley, New York. In press.

LOFTUS, E. E. and FATHI, D. (1985). Retrieving multiple autobiographical memories. *Social Cogn.* **3** 280–295.

MAHALANOBIS, P. C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *J. Roy. Statist. Soc.* **109** 325–378.

MOSTELLER, F. (1978). Nonsampling errors. In *International Encyclopedia of Statisitcs*, *Vol. I* (Kruskal, W. H. and Tanur, J. M., eds.) 208–229. Free Press, New York.

NEYMAN, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *J. Roy. Statist. Soc. Ser. A* **109** 558–606.

PANEL ON PRIVACY AND CONFIDENTIALITY AS FACTORS IN SURVEY RESPONSE, COMMITTEE ON NATIONAL STATISTICS (1979). *Privacy and Confidentiality as Factors in Survey Response*. National Academy of Sciences, Washington, DC.

PLATEK, R., RAO, J. N. K., SMRNDAL, C. E. and SINGH, M. B. (1986). *Small Area Statistics*: *An International Symposium*. Wiley, New York.

REID, C. (1982). *Neyman—From Life*. Springer, New York. MR0680939

ROSENTHAL, R. and RUBIN, D. B. (1979). Issues in summarizing the first 345 studies of interpersonal expectancy effects. *Behav. Brain Sci.* **3** 410–415.

SCHEFFÉ, H. (1959). *The Analysis of Variance*. Wiley, New York; Chapman & Hall, London. MR0116429

SCHUMAN, H., STEEH, C. and BOBO, L. (1985). *Racial Attitudes in America*: *Trends and Interpretations*. Harvard Univ. Press, Cambridge, MA.

SCOTT, C. (1961). Research on mail surveys. *J. Roy. Statist. Soc. Ser. A* **124** 143–205.

SCOTT, C. (1973). Experiments on recall error in African household budget surveys. Unpublished paper presented at meeting of International Association of Survey Statisticians, Vienna, Austria (August 1973).

SENG, Y. P. (1951). Historical survey of the development of sampling theories and practice. *J. Roy. Statist. Soc. Ser. A* **114** 214–231.

SMITH, T. M. F. and SUGDEN, R. A. (1985). Inference and the ignorability of selection for experiments and surveys. *Bull. Int. Stat. Inst.* **51** 10.2-1–10.2-12. MR0886248

STEPHAN, F. F. (1948). History of the uses of modern sampling procedures. *J. Amer. Statist. Assoc.* **43** 12–39.

STOKES, S. L. (1986). Estimation of interviewer effects in complex surveys with application to random digit dialing. In *Proceedings of Second Annual Research Conference* 21–31. U.S. Bureau of the Census, Washington, DC.

TCHUPROV, A. A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. *Metron* **2** 646–680.

TEPPING, B. J., HURWITZ, W. N. and DEMING, W. E. (1943). On the efficiency of deep stratification in block sampling. *J. Amer. Statist. Assoc.* **38** 93–100.

TOURANGEAU, R. (1986). Personal communication.

TOURANGEAU, R. and RASINSKI, K. A. (1986). Context effects in attitude surveys. Unpublished manuscript.

WAKSBERG, J. and PEARL, R. B. (1965). New methodological research on labor force measurement. In *Proceedings of the Social Statistics Section* 227–237. Amer. Statist. Assoc., Washington, DC.

WILK, M. B. and KEMPTHORNE, O. (1955). Fixed, mixed, and random models. *J. Amer. Statist. Assoc.* **50** 1144–1167.

WILK, M. B. and KEMPTHORNE, O. (1956). Some aspects of the analysis of factorial experiments in a completely randomized design. *Ann. Math. Stat.* **27** 950–985. MR0087283

WOLTMAN, H. I., TURNER, A. G. and BUSHERY, J. M. (1980). Comparison of three mixed-mode interviewing procedures in the National Crime Survey. *J. Amer. Statist. Assoc.* **75** 534–543.

YATES, E. (1981). *Sampling Methods for Censuses and Surveys*, 4th ed. Macmillan, New York.

YATES, F. (1985). Book review of "W.G. Cochran's Impact on Statistics", Ed. P.S.R.S. Rao and J. Sedransk. *Biometrics* **41** 591–592.

YATES, E. and COCHRAN, W. G. (1938). The analysis of groups of experiments. *J. Agric. Sci.* **28** 556–580.

ZARKOVICH, S. S. (1956). Note on the history of sampling methods in Russia. *J. Roy. Statist. Soc. Ser. A* **119** 336–338.

ZARKOVICH, S. S. (1962). A supplement to "Note on the history of sampling methods in Russia". *J. Roy. Statist. Soc. Ser. A* **125** 580–582.

The Annals of Applied Statistics
2018, Vol. 12, No. 2, 1180–1203
https://doi.org/10.1214/17-AOAS1083

A TESTING BASED APPROACH TO THE DISCOVERY OF DIFFERENTIALLY CORRELATED VARIABLE SETS

BY KELLY BODWIN[1], KAI ZHANG[2] AND ANDREW NOBEL[3]

University of North Carolina at Chapel Hill

Given data obtained under two sampling conditions, it is often of interest to identify variables that behave differently in one condition than in the other. We introduce a method for differential analysis of second-order behavior called Differential Correlation Mining (DCM). The DCM method identifies differentially correlated sets of variables, with the property that the average pairwise correlation between variables in a set is higher under one sample condition than the other. DCM is based on an iterative search procedure that adaptively updates the size and elements of a candidate variable set. Updates are performed via hypothesis testing of individual variables, based on the asymptotic distribution of their average differential correlation. We investigate the performance of DCM by applying it to simulated data as well as to recent experimental datasets in genomics and brain imaging.

REFERENCES

ANDERSON, T. W. (1959). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

BASSI, F. and HERO, A. (2012). Large scale correlation detection. In *Proc. of the IEEE International Symposium on Information Theory* 2591–2595.

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

BICKEL, P. J. and LEVINA, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

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

BOCKMAYR, M., KLAUSCHEN, F., GYÖRFFY, B., DENKERT, C. and BUDCZIES, J. (2013). New network topology approaches reveal differential correlation patterns in breast cancer. *BMC Syst. Biol.* **7** 78.

BODWIN, K., ZHANG, K. and NOBEL, A. (2018). Supplement to "A testing based approach to the discovery of differentially correlated variable sets." DOI:10.1214/17-AOAS1083SUPP.

BROWNE, M. W. and SHAPIRO, A. (1986). The asymptotic covariance matrix of sample correlation coefficients under general conditions. *Linear Algebra Appl.* **82** 169–176. MR0858970

CAI, T. T. and JIANG, T. (2011). Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *Ann. Statist.* **39** 1496–1525. MR2850210

CAI, T., LIU, W. and XIA, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *J. Amer. Statist. Assoc.* **108** 265–277. MR3174618

CAI, T. T. and ZHANG, A. (2014). Inference on high-dimensional differential correlation matrix. Technical report.

Key words and phrases. Differential correlation mining, association mining, biostatistics, genomics, high-dimensional data.

CAI, T. T., ZHANG, C.-H. and ZHOU, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38** 2118–2144. MR2676885

CHOI, Y. and KENDZIORSKI, C. (2009). Statistical methods for gene set co-expression analysis. *Bioinformatics* **25** 2780–2786.

CUI, X. and CHURCHILL, G. A. (2003). Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.* **4** 210.

DATTA, S. and DATTA, S. (2002). Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics* **19**.

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

DONNER, A. and ZOU, G. (2014). Testing the equality of dependent intraclass correlation coefficients. *J. R. Stat. Soc., D* **51** 367–379.

FRIGUET, C., KLOAREG, M. and CAUSEUR, D. (2009). A factor model approach to multiple testing under dependence. *J. Amer. Statist. Assoc.* **104** 1406–1415. MR2750571

FUKUSHIMA, A. (2013). DiffCorr: An R package to analyze and visualize differential correlations in biological networks. *Gene* **518** 209–214.

GILL, R., DATTA, S. and DATTA, S. (2010). A statistical framework for differential network analysis from microarray data. *BMC Bioinform.* **11** 95.

GREICIUS, M. D., KRASNOW, B., REISS, A. L. and MENON, V. (2002). Functional connectivity in the resting brain: A network analysis of the default mode hypothesis. *Proc. Natl. Acad. Sci. USA* **100** 253–258.

HARMAN, H. H. (1960). *Modern Factor Analysis*. Univ. Chicago Press, Chicago, Ill. MR0159393

HU, R., QIU, X. and GLAZKO, G. (2010). A new gene selection procedure based on the covariance distance. *Bioinformatics* **26** 348–354.

IGLESIA, M. D., VINCENT, B. G., PARKER, J. S., HOADLEY, K. A., CAREY, L. A., PEROU, C. M. and SERODY, J. S. (2014). Prognostic B-cell signatures using mRNA-Seq in patients with subtype-specific breast and ovarian cancer. *Clin. Cancer Res.* **20** 3818–3829.

JIANG, D., TANG, C. and ZHANG, A. (2004). Cluster analysis for gene expression data: A survey. *IEEE Trans. Knowl. Data Eng.* **16** 1370–1386.

KRIEGEL, H.-P., KRÖGER, P. and ZIMEK, A. (2009). Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Trans. Knowl. Discov. Data* **3**.

LANGFELDER, P. and HORVATH, S. (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinform.* **9** 559.

LEWIS, K., KAUFMAN, J., GONZALEZ, M., WIMMER, A. and CHRISTAKIS, N. (2008). Tastes, ties, and time: A new social network dataset using Facebook.com. *Soc. Netw.* **30** 330–342.

LIU, B.-H., YU, H., TU, K., LI, C., LI, Y.-X. and LI, Y.-Y. (2010). DCGL: An R package for identifying differentially coexpressed genes and links from gene expression microarray data. *Bioinformatics* **26** 2637–2638.

MACMAHON, M. and GARLASCHELLI, D. (2015). Community detection for correlation matrices. *Phys. Rev. X* **5** 021006.

MUIRHEAD, R. J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley, New York. MR0652932

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

PEROU, C. M., SØRLIE, T., EISEN, M. B., VAN DE RIJN, M., JEFFREY, S. S., REES, C. A., POLLACK, J. R., ROSS, D. T., JOHNSEN, H., AKSLEN, L. A., FLUGE, O., PERGAMENSCHIKOV, A., WILLIAMS, C., ZHU, S. X., LONNING, P. E., BORRESEN-DALE, A.-L., BROWN, P. O. and BOTSTEIN, D. (2000). Molecular portraits of human breast tumours. *Nature* **406** 747–752.

PHAN, K. L., WAGER, T., TAYLOR, S. F. and LIBERZON, I. (2002). Functional neuroanatomy of emotion: A meta-analysis of emotion activation studies in PET and fMRI. *NeuroImage* **16** 331–348.

RAJARATNAM, B., MASSAM, H. and CARVALHO, C. M. (2008). Flexible covariance estimation in graphical Gaussian models. *Ann. Statist.* **36** 2818–2849. MR2485014

SHENG, E., WITTEN, D. and ZHOU, X.-H. (2016). Hypothesis testing for differentially correlated features. *Biostatistics* **17** 677–691. MR3604273

SONESON, C. and DELORENZI, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinform.* **14** 91.

STAMATATOS, E. (2009). A comparison of methods for differential expression analysis of RNA-seq data. *J. Am. Soc. Inf. Sci. Technol.* **60** 538–556.

STEIGER, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychol. Bull.* **87** 245–251.

STEIGER, J. H. and HAKSTIAN, A. R. (1982). The asymptotic distribution of elements of a correlation matrix: Theory and application. *Br. J. Math. Stat. Psychol.* **35** 208–215. MR0683508

TESSON, B. M., BREITLING, R. and JANSEN, R. C. (2010). DiffCoEx: A simple and sensitive method to find differentially coexpressed gene modules. *BMC Bioinform.* **11** 497.

VOETS, N. L., ADCOCK, J. E., FLITNEY, D. E., BEHRENS, T. E., HART, Y., STACEY, R., CARPENTER, K. and MATTHEWS, P. M. (2006). Distinct right frontal lobe activation in language processing following left hemisphere injury. *Brain* **129** 754–766.

WAINER, H. and BRAUN, H. I. (2013). *Test Validity*. Routledge, London.

WANG, J., FAN, L., WANG, Y., XU, W., JIANG, T., FOX, P. T., EICKHOFF, S. B., YU, C. and JIANG, T. (2015). Determination of the posterior boundary of Wernicke's area based on multimodal connectivity profiles. *Hum. Brain Mapp.* **36** 1908–1924.

WILSON, J. D., WANG, S., MUCHA, P. J., BHAMIDI, S. and NOBEL, A. B. (2014). A testing based extraction algorithm for identifying significant communities in networks. *Ann. Appl. Stat.* **8** 1853–1891. MR3271356

XIA, Y., CAI, T. and CAI, T. T. (2015). Testing differential networks with applications to the detection of gene–gene interactions. *Biometrika* **102** 247–266. MR3371002

ZHOU, C., HAN, F., ZHANG, X. and LIU, H. (2015). An extreme-value approach for testing the equality of large U-statistic based correlation matrices. Available at arXiv:1502.03211.

The Annals of Applied Statistics
2018, Vol. 12, No. 2, 1204–1227
https://doi.org/10.1214/17-AOAS1085

BIOMARKER ASSESSMENT AND COMBINATION WITH DIFFERENTIAL COVARIATE EFFECTS AND AN UNKNOWN GOLD STANDARD, WITH AN APPLICATION TO ALZHEIMER'S DISEASE

BY ZHEYU WANG AND XIAO-HUA ZHOU[1]

Johns Hopkins University and University of Washington

The continued efforts to evaluate biomarkers' predictive abilities and identify optimal biomarker combinations are often challenged by the absence of a gold standard, that is, the true disease status. Current methods that address this issue are mostly developed for binary or ordinal diagnostic tests, which do not fully utilize information provided by continuous biomarkers, or require strong parametric assumptions. Moreover, limited methods exist to allow for the inclusion of covariates—despite their crucial role in facilitating the accurate evaluation of biomarkers. In this paper, we proposed a latent profile approach to evaluating diagnostic accuracy of biomarkers without a gold standard. The method allows for flexible biomarker distributions and incorporation of previous knowledge about risk factors while simultaneously permitting researchers to model paticipants' characteristics that putatively affect biomarker levels, and therefore provides information needed to develop more personalized diagnostic procedures. Additionally, the proposed method presents a potential strategy for biomarker combination when gold standard information is unavailable, as it derives a composite risk score for the underlying disease status. The method is applied to evaluate different cerebral spinal fluid (CSF) biomarkers for Alzheimer's disease (AD) detection. The results show that CSF biomarkers hold significant potential for facilitating early AD detection and for continuous disease monitoring. Furthermore, they call attention to biomarker variability in subgroups and reexamination of CSF biomarker distributions. Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database.

REFERENCES

ALBERT, P. S. and DODD, L. E. (2004). A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics* **60** 427–435. MR2066277

ALBERT, P. S., MCSHANE, L. M. and SHIH, J. H. (2001). Latent class modelling approaches for assessing diagnostic error without a gold standard: With applications to p53 immunohistochemical assays in bladder tumors. *Biometrics* **57** 610–619. MR1855699

BANDEEN-ROCHE, K., MIGLIORETTI, D. L., ZEGER, S. L. and RATHOUZ, P. J. (1997). Latent variable regression for multiple discrete outcomes. *J. Amer. Statist. Assoc.* **92** 1375–1386. MR1615248

Key words and phrases. Diagnostic accuracy, latent profile model, finite mixture models, differential covariate effect, identifiability, Alzheimer's disease.

BATEMAN, R. J., XIONG, C., BENZINGER, T. L. S., FAGAN, A. M., GOATE, A., FOX, N. C., MARCUS, D. S., CAIRNS, N. J., XIE, X., BLAZEY, T. M., HOLTZMAN, D. M., SANTACRUZ, A., BUCKLES, V., OLIVER, A., MOULDER, K., AISEN, P. S., GHETTI, B., KLUNK, W. E., MCDADE, E., MARTINS, R. N., MASTERS, C. L., MAYEUX, R., RINGMAN, J. M., ROSSOR, M. N., SCHOFIELD, P. R., SPERLING, R. A., SALLOWAY, S., MORRIS, J. C. and DOMINANTLY INHERITED ALZHEIMER NETWORK (2012). Clinical and biomarker changes in dominantly inherited Alzheimer's disease. *N. Engl. J. Med.* **367** 795–804.

BENAGLIA, T., CHAUVEAU, D. and HUNTER, D. R. (2009). An EM-like algorithm for semi- and nonparametric estimation in multivariate mixtures. *J. Comput. Graph. Statist.* **18** 505–526. MR2749842

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations. *J. Roy. Statist. Soc. Ser. B* **26** 211–252. MR0192611

BRANSCUM, A. J., JOHNSON, W. O., HANSON, T. E. and GARDNER, I. A. (2008). Bayesian semiparametric ROC curve estimation and disease diagnosis. *Stat. Med.* **27** 2474–2496. MR2432500

BRANSCUM, A. J., JOHNSON, W. O., HANSON, T. E. and BARON, A. T. (2015). Flexible regression models for ROC and risk analysis, with or without a gold standard. *Stat. Med.* **34** 3997–4015.

CHENG, R. C. H. and TRAYLOR, L. (1995). Non-regular maximum likelihood problems. *J. Roy. Statist. Soc. Ser. B* **57** 3–44. MR1325377

COLLINS, J. and HUYNH, M. (2014). Estimation of diagnostic test accuracy without full verification: A review of latent class methods. *Stat. Med.* **33** 4141–4169. MR3267401

COOK, R. J., NG, E. T. M. and MEADE, M. O. (2000). Estimation of operating characteristics for dependent diagnostic tests based on latent Markov models. *Biometrics* **56** 1109–1117. MR1815590

EFRON, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika* **68** 589–599. MR0637776

EFRON, B. (1987). Better bootstrap confidence intervals. *J. Amer. Statist. Assoc.* **82** 171–200. MR0883345

GOODMAN, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61** 215–231. MR0370936

HEBERT, L. E., WEUVE, J., SCHERR, P. A. and EVANS, D. A. (2013). Alzheimer disease in the United States (2010–2050) estimated using the 2010 census. *Neurology* **80** 1778–1783.

HUANG, G.-H. and BANDEEN-ROCHE, K. (2004). Building an identifiable latent class model with covariate effects on underlying and measured variables. *Psychometrika* **69** 5–32. MR2272437

HUI, S. L. and WALTER, S. D. (1980). Estimating the error rates of diagnostic tests. *Biometrics* **36** 167–171.

JACK, C. R. JR., KNOPMAN, D. S., JAGUST, W. J., SHAW, L. M., AISEN, P. S., WEINER, M. W., PETERSEN, R. C. and TROJANOWSKI, J. Q. (2010). Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurol.* **9** 119–128.

JANES, H. and PEPE, M. S. (2009). Adjusting for covariate effects on classification accuracy using the covariate-adjusted receiver operating characteristic curve. *Biometrika* **96** 371–382. MR2507149

JONES, G., JOHNSON, W. O., VINK, W. D. and FRENCH, N. (2012). A framework for the joint modeling of longitudinal diagnostic outcome data and latent infection status: Application to investigating the temporal relationship between infection and disease. *Biometrics* **68** 371–379. MR2959603

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

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

MCHUGH, R. B. (1956). Efficient estimation and local identification in latent class analysis. *Psychometrika* **21** 331–347. MR0082427

McLachlan, G. and Peel, D. (2004). *Finite Mixture Models*. Wiley-Interscience, New York. MR1789474

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford Statistical Science Series* **28**. Oxford Univ. Press, Oxford. MR2260483

Pfeiffer, R. M., Carroll, R. J., Wheeler, W., Whitby, D. and Mbulaiteye, S. (2008). Combining assays for estimating prevalence of human herpesvirus 8 infection using multivariate mixture models. *Biostatistics* **9** 137–151.

Qu, Y., Tan, M. and Kutner, M. H. (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics* **52** 797–810. MR1411731

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **26** 195–239. MR0738930

Selkoe, D. J. (1991). The molecular pathology of Alzheimer's disease. *Neuron* **6** 487–498.

Storandt, M., Head, D., Fagan, A. M., Holtzman, D. M. and Morris, J. C. (2012). Toward a multifactorial model of Alzheimer disease. *Neurobiol. Aging* **33** 2262–2271.

van Smeden, M., Naaktgeboren, C. A., Reitsma, J. B., Moons, K. G. and de Groot, J. A. (2013). Latent class models in diagnostic studies when there is no reference standard—A systematic review. *Am. J. Epidemiol.* **179** 423–431.

Wang, Z. (2013). Latent Class and Latent Profile Analysis in Medical Diagnosis and Prognosis. Ph.D. thesis, University of Washington.

Wang, Z. and Zhou, X.-H. (2012). Random effects models for assessing diagnostic accuracy of traditional Chinese doctors in absence of a gold standard. *Stat. Med.* **31** 661–671. MR2900868

Wang, Z. and Zhou, X.-H. (2014). Nonparametric identifiability of finite mixture models with covariates for estimating error rate without a gold standard. UW Biostatistics Working Paper Series. Working Paper 403.

Wang, Z., Zhou, X.-H. and Wang, M. (2011). Evaluation of diagnostic accuracy in detecting ordered symptom statuses without a gold standard. *Biostatistics* **12** 567–581.

Wu, Z., Deloria-Knoll, M., Hammitt, L. L. and Zeger, S. L. (2016). Partially latent class models for case-control studies of childhood pneumonia aetiology. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **65** 97–114. MR3438240

Zhou, X.-H., Castelluccio, P. and Zhou, C. (2005). Nonparametric estimation of ROC curves in the absence of a gold standard. *Biometrics* **61** 600–609. MR2140934

The Annals of Applied Statistics
2018, Vol. 12, No. 2, 1228–1249
https://doi.org/10.1214/17-AOAS1087

ROBUST DEPENDENCE MODELING FOR HIGH-DIMENSIONAL COVARIANCE MATRICES WITH FINANCIAL APPLICATIONS

BY ZHE ZHU AND ROY E. WELSCH

MIT Sloan School of Management

A very important problem in finance is the construction of portfolios of assets that balance risk and reward in an optimal way. A critical issue in portfolio development is how to address data outliers that reflect very unusual, generally non-recurring, market conditions. Should we allow these to have a significant impact on our estimation and portfolio construction process or should they be considered separately as evidence of a regime shift and/or be used to adjust baseline results? In financial asset allocation, a fundamental step is often a mean-variance optimization problem that makes use of the location vector and dispersion matrix of the financial assets. In this paper, we introduce a new high- dimensional covariance estimator that is much less sensitive to outliers compared to its classical counterparts. We then apply this estimator to the active asset allocation application, and show that our proposed new estimator delivers better results compared to many existing asset allocation methods. An important bonus is that on our examples, the method has a smaller proportion of stock weights greater than 10% and, in many cases, a higher alpha. Covariance estimation is more challenging than mean estimation and only locally and not globally optimal solutions are available. Our proposed new robust covariance estimator uses a regular vine dependence structure and only pairwise robust partial correlation estimators. The resulting robust covariance estimator delivers high performance for identifying outliers for large high dimensional datasets, has a high breakdown point, and is positive definite. When the full vine structure is not available, we propose using a minimal spanning tree algorithm to replace missing vine structure.

REFERENCES

ANDERSON, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York; Chapman & Hall, London. MR0091588

BEDFORD, T. and COOKE, R. M. (2002). Vines—A new graphical model for dependent random variables. *Ann. Statist.* **30** 1031–1068. MR1926167

BILLOR, N., HADI, A. S. and VELLEMAN, P. F. (2000). BACON: Blocked adaptive computationally efficient outlier nominators. *Comput. Statist. Data Anal.* **34** 279–298.

DEMIGUEL, V., GARLAPPI, L. and UPPAL, R. (2009). Optimal versus naïve diversification: How inefficient is the $1/N$ portfolio strategy? *Rev. Financ. Stud.* **22** 1915–53.

DIESTEL, R. (2005). *Graph Theory*, 3rd ed. *Graduate Texts in Mathematics* **173**. Springer, Berlin. MR2159259

DISSMANN, J., BRECHMANN, E. C., CZADO, C. and KUROWICKA, D. (2012). Selecting and estimating regular vine copulae and application to financial returns. *Comput. Statist. Data Anal.* **59** 52–69.

Key words and phrases. Active asset allocation, portfolio selection, robust estimation, high-dimensional dependence modeling, covariance/correlation estimation, regular vine.

DONOHO, D. and HUBER, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann* (P. J. Bickel, K. A. Doksum and J. L. Hodges, eds.) 157–184. Wadsworth, Belmont, CA. MR0689745

FAN, J., FAN, Y. and LV, J. (2008). High dimensional covariance matrix estimation using a factor model. *J. Econometrics* **147** 186–197. MR2472991

GARCIA-ALVAREZ, L. and LUGER, R. (2011). Dynamic correlations, estimation risk, and portfolio management during the financial crisis. Working Paper, CEMFI, Madrid.

KUROWICKA, D. and COOKE, R. M. (2006). *Uncertainty Analysis with High Dimensional Dependence Modelling*. Wiley, Chichester.

KUROWICKA, D., COOKE, R. M. and CALLIES, U. (2006). Vines inference. *Braz. J. Probab. Stat.* **20** 103–120.

LOPUHAÄ, H. P. (1989). On the relation between S-estimators and M-estimators of multivariate location and covariance. *Ann. Statist.* **17** 1662–1683. MR1026304

LOPUHAÄ, H. P. and ROUSSEEUW, P. J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *Ann. Statist.* **19** 229–248. MR1091847

MAECHLER, M. and STAHEL, W. (2009). Robust scatter estimators—The barrow wheel benchmark. ICORS 2009, Parma.

MARKOWITZ, H. M. (1952). Portfolio selection. *J. Finance* **7** 77–91.

MARONNA, R. A. and ZAMAR, R. H. (2002). Robust estimates of location and dispersion for high-dimensional data sets. *Technometrics* **44** 307–317.

ROUSSEEUW, P. J. and DRIESSEN, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **41** 212–223.

ROUSSEEUW, P. J. and LEROY, A. M. (1987). *Robust Regression and Outlier Detection*. Wiley.

ROUSSEEUW, P. and YOHAI, V. (1984). Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series Analysis* (*Heidelberg*, 1983). *Lect. Notes Stat.* **26** 256–272. Springer, New York. MR0786313

RUDIN, W. (1976). *Principles of Mathematical Analysis*, 3rd ed. McGraw-Hill Book Co., New York–Auckland–Düsseldorf. MR0385023

TYLER, D. E. (1987). A distribution-free M-estimator of multivariate scatter. *Ann. Statist.* **15** 234–251. MR0885734

YOHAI, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Ann. Statist.* **15** 642–656. MR0888431

YULE, G. U. and KENDALL, M. G. (1965). *An Introduction to the Theory of Statistics*, 14th ed. Hafner Publishing Co., New York. MR0035938

The Annals of Applied Statistics
2018, Vol. 12, No. 2, 1250–1270
https://doi.org/10.1214/17-AOAS1097

NETWORK-BASED FEATURE SCREENING WITH APPLICATIONS TO GENOME DATA[1]

BY MENGYUN WU*, LIPING ZHU† AND XINGDONG FENG*

Shanghai University of Finance and Economics and*
Renmin University of China†

Modern biological techniques have led to various types of data, which are often used to identify important biomarkers for certain diseases with appropriate statistical methods, such as feature screening. Model-free feature screening has been extensively studied in the literature, and it is effective to select useful predictors for ultra-high dimensional data. These existing screening procedures are conducted based on certain marginal correlations between predictors and a response variable, therefore network structures connecting the predictors are usually ignored. Google's PageRank algorithm has achieved remarkable success. We adopt its spirit to adjust original screening approaches by incorporating the network information. We can then significantly improve the performance of those screening methods in choosing useful biomarkers, which is demonstrated in an intensive simulation study. A couple of real genome datasets along with a biological network are further analyzed by comparing results on both accuracy of predicting responses and stability of identifying biomarkers.

REFERENCES

BARABÁSI, A.-L., GULBAHCE, N. and LOSCALZO, J. (2011). Network medicine: A network-based approach to human disease. *Nat. Rev. Genet.* **12** 56–68.

BARABASI, A. L. and OLTVAI, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet.* **5** 101–113.

BARUT, E., FAN, J. and VERHASSELT, A. (2016). Conditional sure independence screening. *J. Amer. Statist. Assoc.* **111** 1266–1277. MR3561948

BRUNE, K., HONG, S.-M., LI, A. et al. (2008). Genetic and epigenetic alterations of familial pancreatic cancers. *Cancer Epidemiol. Biomark. Prev.* **17** 3536–3542.

CAMPAGNA, D., COPE, L., LAKKUR, S. S., HENDERSON, C., LAHERU, D., IACOBUZIO-DONAHUE, C. A. et al. (2008). Gene expression profiles associated with advanced pancreatic cancer. *Int. J. Clin. Exp. Pathol.* **1** 32–43.

CHEN, G., CHAKRAVARTI, N., AARDALEN, K. et al. (2014). Molecular profiling of patient-matched brain and extracranial melanoma metastases implicates the PI3K pathway as a therapeutic target. *Clin. Cancer Res.* **20** 5537–5546.

CHUANG, H., LEE, E., LIU, Y. T. et al. (2006). Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* **3** 140.

CUN, Y. and FRÖHLICH, H. (2012). Biomarker gene signature discovery integrating network knowledge. *Biol.* **1** 5–17.

Key words and phrases. Correlation, feature screening, model-free, network, ultra-high dimension, variable selection.

DAS, J. and YU, H. (2012). HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC Syst. Biol.* **6** 92.

FAN, J., HAN, F. and LIU, H. (2014). Challenges of Big Data analysis. *Nat. Sci. Rev.* **1** 293–314.

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

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

GUSTAFSSON, M., NESTOR, C. E., ZHANG, H. et al. (2014). Modules, networks and systems medicine for understanding disease and aiding diagnosis. *Gen. Med.* **6** 1–11.

HAWRYLYCZ, M., MILLER, J. A., MENON, V., FENG, D., DOLBEARE, T., GUILLOZET-BONGAARTS, A. L., JEGGA, A. G., ARONOW, B. J., LEE, C.-K., BERNARD, A., GLASSER, M. F., DIERKER, D. L., MENCHE, J., SZAFER, A., COLLMAN, F., GRANGE, P., BERMAN, K. A., MIHALAS, S., YAO, Z., STEWART, L., BARABÁSI, A.-L., SCHULKIN, J., PHILLIPS, J., NG, L., DANG, C., HAYNOR, D. R., JONES, A., ESSEN, D. C. V., KOCH, C. and LEIN, E. (2015). Canonical genetic signatures of the adult human brain. *Nat. Neurosci.* **18** 1832–1844.

HE, Z. and YU, W. (2010). Stable feature selection for biomarker discovery. *Comput. Biol. Chem.* **34** 215–225.

HONG, H. G., WANG, L. and HE, X. (2016). A data-driven approach to conditional screening of high-dimensional variables. *Statistics* **5** 200–212.

HRUBAN, R. H., GOGGINS, M., PARSONS, J. and KERN, S. E. (2000). Progression model for pancreatic cancer. *Clin. Cancer Res.* **6** 2969–2972.

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

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

HUANG, Z.-Q., BUCHSBAUM, D. J., RAISCH, K. P., BONNER, J. A., BLAND, K. I. and VICKERS, S. M. (2003). Differential responses by pancreatic carcinoma cell lines to prolonged exposure to Erbitux (IMC-C225) anti-EGFR antibody. *J. Surg. Res.* **111** 274–283.

JAGIRDAR, R., SOLENOV, E. I., HATZOGLOU, C., MOLYVDAS, P.-A., GOURGOULIANIS, K. I. and ZAROGIANNIS, S. G. (2013). Gene expression profile of aquaporin 1 and associated interactors in malignant pleural mesothelioma. *Genetics* **517** 99–105.

JAVLE, M., LI, Y., TAN, D., DONG, X., CHANG, P., KAR, S. and LI, D. (2014). Biomarkers of TGF-β signaling pathway and prognosis of pancreatic cancer. *PLoS ONE* **9** e85942.

LANGVILLE, A. N. and MEYER, C. D. (2012). *Google's PageRank and Beyond*: *The Science of Search Engine Rankings*. Princeton Univ. Press, Princeton, NJ. MR3052718

LEISERSON, M. D., VANDIN, F., WU, H. T. et al. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **47** 106–114.

LI, C. and LI, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* **24** 1175–1182.

LI, R., ZHONG, W. and ZHU, L. (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc.* **107** 1129–1139.

MARTINEZLEDESMA, E., VERHAAK, R. G. and TREVINO, V. (2015). Identification of a multi-cancer gene expression biomarker for cancer clinical outcomes using a network-based algorithm. *Sci. Rep.* **5**.

MOFFITT, R. A., MARAYATI, R., FLATE, E. L. et al. (2015). Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nat. Genet.* **47** 1168–1178.

PAN, W., XIE, B. and SHEN, X. (2010). Incorporating predictor network in penalized regression with application to microarray data. *Biometrics* **66** 474–484. MR2758827

ROBERT, C. P. and CASELLA, G. (1999). *Monte Carlo Statistical Methods*. Springer, New York. MR1707311

SHI, X., YI, H. and MA, S. (2015). Measures for the degree of overlap of gene signatures and applications to TCGA. *Brief. Bioinform.* **16** 266–272.

TASCILAR, M., SKINNER, H. G., ROSTY, C. et al. (2001). The SMAD4 protein and prognosis of pancreatic ductal adenocarcinoma. *Clin. Cancer Res.* **7** 4115–4121.

TAYLOR, I. W., LINDING, R., WARDEFARLEY, D. et al. (2009). Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat. Biotechnol.* **27** 199–204.

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

VIDAL, M., CUSICK, M. E. and BARABASI, A. L. (2011). Interactome networks and human disease: Cell. *Cell* **144** 986–998.

WANG, X. and LENG, C. (2016). High dimensional ordinary least squares projection for screening variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 589–611.

WONG, H. H. and LEMOINE, N. R. (2009). Pancreatic cancer: Molecular pathogenesis and new therapeutic targets. *Nat. Rev. Gastroenterol. Hepatol.* **6** 412–422.

WU, M., ZHU, L. and FENG, X. (2018). Supplement to "Network-based feature screening with applications to genome data." DOI:10.1214/17-AOAS1097SUPP.

YU, G. and LIU, Y. (2016). Sparse regression incorporating graphical structure among predictors. *J. Amer. Statist. Assoc.* **111** 707–720. MR3538699

ZHU, L.-P., LI, L., LI, R. and ZHU, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc.* **106** 1464–1475. MR2896849

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

The Annals of Applied Statistics
2018, Vol. 12, No. 2, 1271–1292
https://doi.org/10.1214/17-AOAS1109

COVARIATE MATCHING METHODS FOR TESTING AND QUANTIFYING WIND TURBINE UPGRADES

BY YEI EUN SHIN, YU DING[1] AND JIANHUA Z. HUANG[2]

Texas A&M University

In the wind industry, engineers perform retrofitting upgrades on in-service wind turbines for the purpose of improving power production capabilities. Considering how costly an upgrade can be, people often wonder about the upgrade effect: whether it indeed improves turbine performances, and if so, how much. One cannot simply compare power outputs for the purpose of assessing a turbine's improvement, as wind power generation is affected by an array of environmental covariates, including wind speed, wind direction, temperature, pressure as well as other atmosphere dynamics. For a fair comparison to discern the upgrade effect, it is critical to have these environmental effects controlled for while comparing power output differences. Most existing approaches rely on establishing a power curve model and let the model account for the environmental effects. In this paper, we propose a different approach, which is to devise a covariate matching method to ensure the environmental covariates to have comparable distribution profiles before and after an action of upgrade. Once the covariates are matched, paired t-tests can be applied to the power outputs for testing the significance of the upgrade effect. The relative increase in power production can also be quantified. The proposed approach is simple to use and relies on fewer assumptions than the power curve modeling approach.

REFERENCES

ACKERMANN, T. and SÖDER, L. (2005). Wind power in power systems: An introduction. In *Wind Power in Power Systems* 25–51.

ALBERS, A. (2012). Relative and integral wind turbine power performance evaluation. In *Proceedings of the* 2012 *European Wind Energy Conference & Exhibition* 22–25.

BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.

BYON, E., NTAIMO, L., SINGH, C. and DING, Y. (2013). Wind energy facility reliability and maintenance. In *Handbook of Wind Power Systems*: *Optimization*, *Modeling*, *Simulation and Economic Aspects* (P. M. Pardalos, S. Rebennack, M. V. F. Pereira, N. A. Iliadis and V. Pappu, eds.) 639–672. Springer, Berlin.

DELLE MONACHE, L., ECKEL, F. A., RIFE, D. L., NAGARAJAN, B. and SEARIGHT, K. (2013). Probabilistic weather prediction with an analog ensemble. *Mon. Weather Rev.* **141** 3498–3516.

DOE (2015). Windexchange: US installed wind capacity 2015. Technical report, U.S. Department of Energy's Energy Efficiency & Renewable Energy Website. Available at http://apps2.eere.energy.gov/wind/windexchange/wind_installed_capacity.asp.

IACUS, S. M., KING, G. and PORRO, G. (2012). Causal inference without balance checking: Coarsened exact matching. *Polit. Anal.* **20** 1–24.

Key words and phrases. Causal inference, Mahalanobis distance, matching methods, nearest neighbor matching, observational study, wind power curve.

IEC (2005). Wind turbines—part 12-1: Power performance measurements of electricity producing wind turbines; iec tc/sc 88. Technical Report 61400-12-1:2005, International Electrotechnical Commission.

JAMMALAMADAKA, S. R. and SENGUPTA, A. (2001). *Topics in Circular Statistics* **5**. World Scientific, Singapore.

JEON, J. and TAYLOR, J. W. (2012). Using conditional kernel density estimation for wind power density forecasting. *J. Amer. Statist. Assoc.* **107** 66–79.

KHALFALLAH, M. G. and KOLIUB, A. M. (2007). Suggestions for improving wind turbines power curves. *Desalination* **209** 221–229.

KUSIAK, A., ZHENG, H. and SONG, Z. (2009). Wind farm power prediction: A data-mining approach. *Wind Energy* **12** 275–293.

LEE, G., DING, Y., GENTON, M. G. and XIE, L. (2015a). Power curve estimation with multivariate environmental factors for inland and offshore wind farms. *J. Amer. Statist. Assoc.* **110** 56–67. MR3338486

LEE, G., DING, Y., XIE, L. and GENTON, M. G. (2015b). A kernel plus method for quantifying wind turbine performance upgrades. *Wind Energy* **18** 1207–1219.

MAHALANOBIS, P. C. (1936). On the generalized distance in statistics. *Proc. Natl. Inst. Sci.* (*Calcutta*) **2** 49–55.

OSADCIW, L. A., YAN, Y., YE, X., BENSON, G. and WHITE, E. (2010). Wind turbine diagnostics based on power curve using particle swarm optimization. In *Wind Power Systems* (*Green Energy and Technology*) (L. Wang, C. Singh and A. Kusiak, eds.) 151–165. Springer, Berlin.

ØYE, S. (1995). The effect of vortex generators on the performance of the ELKRAFT 1000 kW turbine. In *Aerodynamics of Wind Turbines*: 9*th IEA Symposium* 9–14, Stockholm, Sweden.

PINSON, P., NIELSEN, H. A., MADSEN, H. and NIELSEN, T. S. (2008). Local linear regression with adaptive orthogonal fitting for the wind power application. *Stat. Comput.* **18** 59–71. MR2416439

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.

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

RUBIN, D. B. (1973). Matching to remove bias in observational studies. *Biometrics* **29** 159–183.

RUBIN, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Serv. Outcomes Res. Methodol.* **2** 169–188.

SANCHEZ, I. (2006). Short-term prediction of wind energy production. *Int. J. Forecast.* **22** 43–56.

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

TORBEN, NIELSEN, S., NIELSEN, H. A. and MADSEN, H. (2002). Prediction of wind power using time-varying coefficient functions. In *Proceedings of the XV IFAC World Congress on Automatic Control*, Barcelona, Spain.

ULUYOL, O., PARTHASARATHY, G., FOSLIEN, W. and KIM, K. (2011). Power curve analytic for wind turbine performance monitoring and prognostics. In *Annual Conference of the Prognostics and Health Management Society* **2**. Publication Control Number 049, Montreal, Canada.

WAN, Y.-H., ELA, E. and ORWIG, K. (2010). Development of an equivalent wind plant power curve. Technical Report NREL/CP-550-48146, National Renewable Energy Laboratory. Available at http://www.nrel.gov/docs/fy10osti/48146.pdf.

WANG, L., TANG, X. and LIU, X. (2012). Blade design optimisation for fixed-pitch fixed-speed wind turbines. *ISRN Renew. Energy* **2012** Article ID 682859.

YAN, Y., OSADCIW, L. A., BENSON, G. and WHITE, E. (2009). Inverse data transformation for change detection in wind turbine diagnostics. In *Proceedings of the* 22*nd IEEE Canadian Conference on Electrical and Computer Engineering* 944–949. St. John's, Newfoundland, Canada.

ZHAO, Z. (2004). Using matching to estimate treatment effects: Data requirements, matching metrics, and Monte Carlo evidence. *Rev. Econ. Stat.* **86** 91–107.

The Annals of Applied Statistics
2018, Vol. 12, No. 2, 1293–1311
https://doi.org/10.1214/17-AOAS1111

NONSTATIONARY MODELLING OF TAIL DEPENDENCE OF TWO SUBJECTS' CONCENTRATION

BY KSHITIJ SHARMA[*,†], VALÉRIE CHAVEZ-DEMOULIN[*] AND PIERRE DILLENBOURG[†]

University of Lausanne[*] *and École Polytechnique Férdérale de Lausanne*[†]

We analyse eye-tracking data to understand how people collaborate. Our dataset consists of time series of measurements for eye movements, such as spatial entropy, calculated for each subject during an experiment when several pairs of participants collaborate to accomplish a task. We observe that pairs with high collaboration quality obtain their highest values of concentration (or equivalently lowest values of spatial entropy) occurring simultaneously. In this paper, we propose a flexible model that describes the tail dependence structure between two subjects' entropy when the pair is collaborating. More generally, we develop a generalized additive model (GAM) framework for tail dependence coefficients in the presence of covariates. As for any GAM-type model, the methodology can be used to predict collaboration quality or to explore how joint concentration depends on other cognitive operations and varies over time.

REFERENCES

AAS, K., CZADO, C., FRIGESSI, A. and BAKKEN, H. (2009). Pair-copula constructions of multiple dependence. *Insurance Math. Econom.* **44** 182–198. MR2517884

ALLOPENNA, P. D., MAGNUSON, J. S. and TANENHAUS, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *J. Mem. Lang.* **38** 419–439.

BALLARD, D. H., HAYHOE, M. M., LI, F., WHITEHEAD, S. D., FRISBY, J. P., TAYLOR, J. G. and FISHER, R. B. (1992). Hand-eye coordination during sequential tasks [and discussion]. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **337** 331–339.

CHASE, W. G. and SIMON, H. A. (1973). Perception in chess. *Cogn. Psychol.* **4** 55–81.

COLES, S. G. and TAWN, J. A. (1996). Modelling extremes of the areal rainfall process. *J. Roy. Statist. Soc. Ser. B* **58** 329–347. MR1377836

EMBRECHTS, P., LINDSKOG, F. and MCNEIL, A. (2003). Modelling dependence with copulas and applications to risk management. In *Handbook of Heavy Tailed Distributions in Finance* (S. Rachev, ed.) 329–384. Elsevier, Amsterdam.

FERREIRA, M. (2013). Nonparametric estimation of the tail-dependence coefficient. *REVSTAT* **11** 1–16. MR3048720

GARDES, L. and GIRARD, S. (2015). Nonparametric estimation of the conditional tail copula. *J. Multivariate Anal.* **137** 1–16. MR3332795

GRANT, E. R. and SPIVEY, M. J. (2003). Eye movements and problem solving guiding attention guides thought. *Psychol. Sci.* **14** 462–466.

Key words and phrases. Collaborative learning, copulas, entropy, generalized additive models, tail dependence.

GREEN, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *Int. Stat. Rev.* **55** 245–259. MR0963142

GREEN, P. J. and SILVERMAN, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach. Monographs on Statistics and Applied Probability* **58**. Chapman & Hall, London. MR1270012

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models. Monographs on Statistics and Applied Probability* **43**. Chapman & Hall, London. MR1082147

HMELO-SILVER, C. E. (2006). Analyzing collaborative learning: Multiple approaches to understanding processes and outcomes. In *Proceedings of the* 7*th International Conference on Learning Sciences* 1059–1065. International Society of the Learning Sciences.

JACOB, R. J. and KARN, K. S. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. *Mind* **2** 4.

JOE, H. (1997). *Multivariate Models and Dependence Concepts. Monographs on Statistics and Applied Probability* **73**. Chapman & Hall, London. MR1462613

JONES, G. (2003). Testing two cognitive theories of insight. *J. Exper. Psychol., Learn., Mem., Cogn.* **29** 1017.

JUST, M. A. and CARPENTER, P. A. (1976). Eye fixations and cognitive processes. *Cogn. Psychol.* **8** 441–480.

KNOBLICH, G., OHLSSON, S. and RANEY, G. E. (2001). An eye movement study of insight problem solving. *Mem. Cogn.* **29** 1000–1009.

LI, F. (2016). Modeling covariate-contingent correlation and tail-dependence with copulas. Available at arXiv:1401.0100.

MCNEIL, A. J., FREY, R. and EMBRECHTS, P. (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton Univ. Press, Princeton, NJ. MR2175089

NELSEN, R. B. (1999). *An Introduction to Copulas. Lecture Notes in Statistics* **139**. Springer, New York. MR1653203

NÜSSLI, M.-A. (2011). Dual eye-tracking methods for the study of remote collaborative problem solving.

PIETINEN, S., BEDNARIK, R. and TUKIAINEN, M. (2010). Shared visual attention in collaborative programming: A descriptive analysis. In *Proceedings of the* 2010 *ICSE Workshop on Cooperative and Human Aspects of Software Engineering* 21–24. ACM, New York.

RICHARDSON, D. C., DALE, R. and KIRKHAM, N. Z. (2007). The art of conversation is coordination common ground and the coupling of eye movements during dialogue. *Psychol. Sci.* **18** 407–413.

SANGIN, M., MOLINARI, G., NÜSSLI, M.-A. and DILLENBOURG, P. (2008). How learners use awareness cues about their peer's knowledge?: Insights from synchronized eye-tracking data. In *Proceedings of the* 8*th International Conference on International Conference for the Learning Sciences* **2** 287–294. International Society of the Learning Sciences.

SCHMIDT, R. and STADTMÜLLER, U. (2006). Non-parametric estimation of tail dependence. *Scand. J. Stat.* **33** 307–335. MR2279645

SCHNEIDER, B., SHARMA, K., CUENDET, S., ZUFFEREY, G., DILLENBOURG, P. and PEA, R. D. (2015). 3D tangibles facilitate joint visual attention in dyads. In *Proceedings of* 11*th International Conference of Computer Supported Collaborative Learning* **1** 156–165. EPFL-CONF-223609.

SHARMA, K., CHAVEZ-DEMOULIN, V. and DILLENBOURG, P. (2017). An application of extreme value theory to learning analytics: Predicting collaboration quality from eye-tracking data. *J. Learn. Anal.* **4** (3) 140–164.

SHARMA, K., JERMANN, P., NÜSSLI, M.-A. and DILLENBOURG, P. (2012). Gaze evidence for different activities in program understanding. In 24*th Annual Conference of Psychology of Programming Interest Group*. EPFL-CONF-184006.

SHARMA, K., JERMANN, P., NÜSSLI, M.-A. and DILLENBOURG, P. (2013). Understanding collaborative program comprehension: Interlacing gaze and dialogues. In *Computer Supported Collaborative Learning* (*CSCL* 2013).

SIBUYA, M. (1960). Bivariate extreme statistics. I. *Ann. Inst. Statist. Math. Tokyo* **11** 195–210. MR0115241

SKLAR, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Stat. Univ. Paris* **8** 229–231. MR0125600

SMITH, R. L. (1990). Max-stable processes and spatial extremes. Unpublished manuscript.

VATTER, T. and CHAVEZ-DEMOULIN, V. (2015). Generalized additive models for conditional dependence structures. *J. Multivariate Anal.* **141** 147–167. MR3390064

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

ZELINSKY, G. J. and MURPHY, G. L. (2000). Synchronizing visual and language processing: An effect of object name length on eye movements. *Psychol. Sci.* **11** 125–131.

The Annals of Applied Statistics
2018, Vol. 12, No. 2, 1312–1331
https://doi.org/10.1214/17-AOAS1113

A SPATIALLY VARYING STOCHASTIC DIFFERENTIAL EQUATION MODEL FOR ANIMAL MOVEMENT[1]

BY JAMES C. RUSSELL*, EPHRAIM M. HANKS†, MURALI HARAN† AND DAVID HUGHES†

Muhlenberg College and The Pennsylvania State University†*

Animal movement exhibits complex behavior which can be influenced by unobserved environmental conditions. We propose a model which allows for a spatially varying movement rate and spatially varying drift through a semiparametric potential surface and a separate motility surface. These surfaces are embedded in a stochastic differential equation framework which allows for complex animal movement patterns in space. The resulting model is used to analyze the spatially varying behavior of ants to provide insight into the spatial structure of ant movement in the nest.

REFERENCES

ALBERTSEN, C. M., WHORISKEY, K., YURKOWSKI, D., NIELSEN, A. and MILLS, J. (2015). Fast fitting of non-Gaussian state-space models to animal movement data via Template Model Builder. *Ecology* **96** 2598–2604.

ALTIZER, S., BARTEL, R. and HAN, B. A. (2011). Animal migration and infectious disease risk. *Science* **331** 296–302.

AVGAR, T., BAKER, J. A., BROWN, G. S., HAGENS, J. S., KITTLE, A. M., MALLON, E. E., MCGREER, M. T., MOSSER, A., NEWMASTER, S. G., PATTERSON, B. R. et al. (2015). Space-use behaviour of woodland caribou based on a cognitive movement model. *J. Anim. Ecol.* **84** 1059–1070.

BANERJEE, S., CARLIN, B. P. and GELFAND, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press, Boca Raton.

BAYLIS, A. M., ORBEN, R. A., ARNOULD, J. P., CHRISTIANSEN, F., HAYS, G. C. and STANILAND, I. J. (2015). Disentangling the cause of a catastrophic population decline in a large marine mammal. *Ecology* **96** 2834–2847.

BESTLEY, S., JONSEN, I. D., HINDELL, M. A., HARCOURT, R. G. and GALES, N. J. (2015). Taking animal tracking to new depths: Synthesizing horizontal–vertical movement relationships for four marine predators. *Ecology* **96** 417–427.

BEYER, H. L., MORALES, J. M., MURRAY, D. and FORTIN, M.-J. (2013). The effectiveness of Bayesian state-space models for estimating behavioural states from movement paths. *Methods Ecol. Evol.* **4** 433–441.

BRILLINGER, D. R. (2003). Simulating constrained animal motion using stochastic differential equations. In *Probability, Statistics and Their Applications: Papers in Honor of Rabi Bhattacharya. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **41** 35–48. IMS, Beachwood, OH. MR1999413

BRILLINGER, D. R. (2007). A potential function approach to the flow of play in soccer. *J. Quant. Anal. Sports* **3** 3. MR2304568

Key words and phrases. Animal movement, stochastic differential equations, potential surface, *Camponotus pennsylvanicus*.

BRILLINGER, D. (2012). A particle migrating randomly on a sphere. In *Selected Works of David Brillinger* 73–87. Springer, New York.

BRILLINGER, D. R. and STEWART, B. S. (1998). Elephant-seal movements: Modelling migration. *Canad. J. Statist.* **26** 431–443.

BRILLINGER, D., PREISLER, H. and WISDOM, M. (2011). Modelling particles moving in a potential field with pairwise interactions and an application. *Braz. J. Probab. Stat.* **25** 421–436.

BRILLINGER, D. R., STEWART, B. S. and LITTNAN, C. L. (2008). Three months journeying of a Hawaiian monk seal. In *Probability and Statistics*: *Essays in Honor of David A. Freedman* 246–264. IMS, Beachwood, OH.

BRILLINGER, D. R., PREISLER, H. K., AGER, A. A., KIE, J. and STEWART, B. S. (2001). Modelling movements of free-ranging animals. Univ. Calif. Berkeley Statistics, Technical Report 610.

BRILLINGER, D. R., PREISLER, H. K., AGER, A. A., KIE, J. G. and STEWART, B. S. (2002). Employing stochastic differential equations to model wildlife motion. *Bull. Braz. Math. Soc.* (*N.S.*) **33** 385–408.

BRILLINGER, D. R., PREISLER, H. K., AGER, A. A. and KIE, J. (2012). The use of potential functions in modelling animal movement. In *Selected Works of David Brillinger* 385–409. Springer, Berlin.

BROST, B. M., HOOTEN, M. B., HANKS, E. M. and SMALL, R. J. (2015). Animal movement constraints improve resource selection inference in the presence of telemetry error. *Ecology* **96** 2590–2597.

CREMER, S., ARMITAGE, S. A. O. and SCHMID-HEMPEL, P. (2007). Social immunity. *Curr. Biol.* **17** R693–R702.

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

DODGE, S., BOHRER, G., WEINZIERL, R., DAVIDSON, S. C., KAYS, R., DOUGLAS, D., CRUZ, S., HAN, J., BRANDES, D. and WIKELSKI, M. (2013). The environmental-data automated track annotation (Env-DATA) system: Linking animal tracks with environmental data. *Mov. Ecol.* **1** 3.

EILERS, P. H. and MARX, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statist. Sci.* 89–102.

FLEGAL, J., HARAN, M. and JONES, G. (2008). Markov chain Monte Carlo: Can we trust the third significant figure? *Statist. Sci.* **23** 250–260.

GARDINER, C. (1986). Handbook of stochastic methods for physics, chemistry and the natural sciences. *Springer Ser. Synergetics* **13** 149–168.

GELFAND, A. E., DIGGLE, P., GUTTORP, P. and FUENTES, M. (2010). *Handbook of Spatial Statistics*. CRC Press, Boca Raton.

GIBERT, J. P., CHELINI, M.-C., ROSENTHAL, M. F. and DELONG, J. P. (2016). Crossing regimes of temperature dependence in animal movement. *Glob. Change Biol.* **22** 1722–1736.

HANKS, E. M., HOOTEN, M. B. and ALLDREDGE, M. W. (2015). Continuous-time discrete-space models for animal movement. *Ann. Appl. Stat.* **9** 145–165.

IHAKA, R. and GENTLEMAN, R. (1996). R: A language for data analysis and graphics. *J. Comput. Graph. Statist.* **5** 299–314.

JOHNSON, D. (2013). crawl: Fit continuous-time correlated random walk models to animal movement data. R package version 1.4-1.

JOHNSON, D., LONDON, J., LEA, M. and DURBAN, J. (2008). Continuous-time correlated random walk model for animal telemetry data. *Ecology* **89** 1208–1215.

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

JONSEN, I. (2015). bsam: Bayesian state-space models for animal movement. R package version 0.43.1.

KILLEEN, J., THURFJELL, H., CIUTI, S., PATON, D., MUSIANI, M. and BOYCE, M. S. (2014). Habitat selection during ungulate dispersal and exploratory movement at broad and fine scale with implications for conservation management. *Mov. Ecol.* **2** 15.

KLEBANER, F. C. (2005). *Introduction to Stochastic Calculus with Applications*. Imperial College Press, London.

KLOEDEN, P. E. and PLATEN, E. (1992). *Numerical Solution of Stochastic Differential Equations*. Springer, Berlin.

MERSCH, D. P., CRESPI, A. and KELLER, L. (2013). Tracking individuals shows spatial fidelity is a key regulator of ant social organization. *Science* **340** 1090–1093.

NORTHRUP, J. M., ANDERSON, C. R. and WITTEMYER, G. (2015). Quantifying spatial habitat loss from hydrocarbon development through assessing habitat selection patterns of mule deer. *Glob. Change Biol.* **21** 3961–3970.

PREISLER, H. K., AGER, A. A. and WISDOM, M. J. (2013). Analyzing animal movement patterns using potential functions. *Ecosphere* **4** art32.

PREISLER, H. K. and AKERS, R. P. (1995). Autoregressive-type models for the analysis of bark beetle tracks. *Biometrics* 259–267.

PREISLER, H. K., AGER, A. A., JOHNSON, B. K. and KIE, J. G. (2004). Modeling animal movements using stochastic differential equations. *Environmetrics* **15** 643–657.

QUEVILLON, L. E., HANKS, E. M., BANSAL, S. and HUGHES, D. P. (2015). Social, spatial, and temporal organization in a complex insect society. Scientific Reports.

RODE, K. D., WILSON, R. R., REGEHR, E. V., MARTIN, M. S., DOUGLAS, D. C. and OLSON, J. (2015). Increased land use by Chukchi Sea polar bears in relation to changing sea ice conditions. *PLoS ONE* **10** e0142213.

RUSSELL, J. C., HANKS, E. M., HARAN, M. and HUGHES, D. (2018). Supplement to "A spatially varying stochastic differential equation model for animal movement." DOI:10.1214/17-AOAS1113SUPP.

SCHARF, H. R., HOOTEN, M. B., FOSDICK, B. K., JOHNSON, D. S., LONDON, J. M. and DURBAN, J. W. (2016). Dynamic social networks based on movement. *Ann. Appl. Stat.* **10** 2182–2202. MR3592053

THIEBAULT, A. and TREMBLAY, Y. (2013). Splitting animal trajectories into fine-scale behaviorally consistent movement units: Breaking points relate to external stimuli in a foraging seabird. *Behav. Ecol. Sociobiol.* **67** 1013–1026.

TOLEDO, S., KISHON, O., ORCHAN, Y., BARTAN, Y., SAPIR, N., VORTMAN, Y. and NATHAN, R. (2014). Lightweight low-cost wildlife tracking tags using integrated transceivers. In *Education and Research Conference* (*EDERC*), 2014 6*th European Embedded Design* 287–291.

WATKINS, K. S. and ROSE, K. A. (2013). Evaluating the performance of individual-based animal movement models in novel environments. *Ecological Modelling* **250** 214–234.

WIKLE, C. K. and HOOTEN, M. B. (2010). A general science-based framework for dynamical spatio-temporal models. *Test* **19** 417–451.

The Annals of Applied Statistics
2018, Vol. 12, No. 2, 1332–1359
https://doi.org/10.1214/17-AOAS1115

TORUS PRINCIPAL COMPONENT ANALYSIS WITH APPLICATIONS TO RNA STRUCTURE

BY BENJAMIN ELTZNER*,[1], STEPHAN HUCKEMANN*,[1] AND
KANTI V. MARDIA†,‡

Georg-August-University Göttingen, University of Oxford† and
University of Leeds‡*

There are several cutting edge applications needing PCA methods for data on tori, and we propose a novel torus-PCA method that adaptively favors low-dimensional representations while preventing overfitting by a new test—both of which can be generally applied and address shortcomings in two previously proposed PCA methods. Unlike tangent space PCA, our torus-PCA features structure fidelity by honoring the cyclic topology of the data space and, unlike geodesic PCA, produces nonwinding, nondense descriptors. These features are achieved by deforming tori into spheres with self-gluing and then using a variant of the recently developed principal nested spheres analysis. This PCA analysis involves a step of subsphere fitting, and we provide a new test to avoid overfitting. We validate our torus-PCA by application to an RNA benchmark data set. Further, using a larger RNA data set, torus-PCA recovers previously found structure, now globally at the one-dimensional representation, which is not accessible via tangent space PCA.

REFERENCES

ALTIS, A., OTTEN, M., NGUYEN, P. H., RAINER, H. and STOCK, G. (2008). Construction of the free energy landscape of biomolecules via dihedral angle principal component analysis. *J. Chem. Phys.* **128** 245102.

ARSIGNY, V., COMMOWICK, O., PENNEC, X. and AYACHE, N. (2006). A log-Euclidean framework for statistics on diffeomorphisms. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI* 2006 924–931. Springer, Berlin.

BOISVERT, J., PENNEC, X., LABELLE, H., CHERIET, F. and AYACHE, N. (2006). Principal spine shape deformation modes using Riemannian geometry and articulated models. In *Articulated Motion and Deformable Objects* 346–355. Springer, Berlin.

BREWER, J. W. (2013). Regulatory crosstalk within the mammalian unfolded protein response. *Cell. Mol. Life Sci.* **71** 1067–1079.

ČECH, P., KUKAL, J., ČERNỲ, J., SCHNEIDER, B. and SVOZIL, D. (2013). Automatic workflow for the classification of local DNA conformations. *BMC Bioinform.* **14** 205.

CHAKRABARTI, A., CHEN, A. W. and VARNER, J. D. (2011). A review of the mammalian unfolded protein response. *Biotechnol. Bioeng.* **108** 2777–2793.

CHAPMAN, R., SIDRAUSKI, C. and WALTER, P. (1998). Intracellular signaling from the endoplasmic reticulum to the nucleus. *Annu. Rev. Cell Dev. Biol.* **14** 459–485.

CHEN, A. A. and GARCÍA, A. E. (2013). High-resolution reversible folding of hyperstable RNA tetraloops using molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA* **110** 16820–16825.

Key words and phrases. Statistics on manifolds, tori deformation, directional statistics, dimension reduction, dihedral angles, fitting small spheres, principal nested spheres analysis.

DAVIS, I. W., LEAVER-FAY, A., CHEN, V. B., BLOCK, J. N., KAPRAL, G. J., WANG, X., MURRAY, L. W., ARENDALL, W. B., SNOEYINK, J., RICHARDSON, J. S. et al. (2007). MolProbity: All-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res*. **35** W375–W383.

DRYDEN, I. L. and MARDIA, K. V. (2016). *Statistical Shape Analysis*: *With Applications in R*. Wiley, New York.

DUARTE, C. M. and PYLE, A. M. (1998). Stepping through an RNA structure: A novel approach to conformational analysis. *J. Mol. Biol*. **284** 1465–1478.

DÜMBGEN, L. and WALTHER, G. (2008). Multiscale inference about a density. *Ann. Statist*. **36** 1758–1785.

DUNBRACK, R. L. and KARPLUS, M. (1994). Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. *Nat. Struct. Mol. Biol*. **1** 334–340.

EGLI, M., PORTMANN, S. and USMAN, N. (1996). RNA hydration: A detailed look. *Biochemistry* **35** 8489–8494.

ELTZNER, B., HUCKEMANN, S. and MARDIA, K. V. (2018a). Supplement to "Torus principal component analysis with applications to RNA structure." DOI:10.1214/17-AOAS1115SUPPA.

ELTZNER, B., HUCKEMANN, S. and MARDIA, K. V. (2018b). Supplement to "Torus principal component analysis with applications to RNA structure." DOI:10.1214/17-AOAS1115SUPPB.

ELTZNER, B., HUCKEMANN, S. and MARDIA, K. V. (2018c). Supplement to "Torus principal component analysis with applications to RNA structure." DOI:10.1214/17-AOAS1115SUPPC.

ESTARELLAS, C., OTYEPKA, M., KOČA, J., BANÁŠ, P., KREPL, M. and ŠPONER, J. (2015). Molecular dynamic simulations of protein/RNA complexes: CRISPR/Csy4 endoribonuclease. *Biochimica et Biophysica Acta* (*BBA*)—*General Subjects* **1850** 1072–1090.

FLETCHER, P. T., LU, C., PIZER, S. M. and JOSHI, S. C. (2004). Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Trans. Med. Im*. **23** 995–1005.

FRELLSEN, J., MOLTKE, I., THIIM, M., MARDIA, K. V., FERKINGHOFF-BORG, J. and HAMELRYCK, T. (2009). A probabilistic model of RNA conformational space. *PLoS Comput. Biol*. **5** e1000406.

GOWER, J. C. (1975). Generalized Procrustes analysis. *Psychometrika* **40** 33–51. MR0405725

GREEN, P. J. and MARDIA, K. V. (2006). Bayesian alignment using hierarchical models, with applications in protein bioinformatics. *Biometrika* **93** 235–254.

HOTZ, T. and HUCKEMANN, S. (2014). Intrinsic means on the circle: Uniqueness, locus and asymptotics. *Ann. Inst. Statist. Math*. **67** 177–193.

HUCKEMANN, S. F. and ELTZNER, B. (2015). Polysphere PCA with applications. In *Proceedings of the Leeds Annual Statistical Research* (*LASR*) *Workshop* 2015.

HUCKEMANN, S., HOTZ, T. and MUNK, A. (2010). Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. *Statist. Sinica* **20** 1–58. MR2640651

HUCKEMANN, S. and ZIEZOLD, H. (2006). Principal component analysis for Riemannian manifolds, with an application to triangular shape spaces. *Adv. in Appl. Probab*. **2** 299–319.

HUCKEMANN, S., KIM, K.-R., MUNK, A., REHFELDT, F., SOMMERFELD, M., WEICKERT, J. and WOLLNIK, C. (2016). The circular SiZer, inferred persistence of shape parameters and application to early stem cell differentiation. *Bernoulli* **22** 2113–2142.

JAIN, S., RICHARDSON, D. C. and RICHARDSON, J. S. (2015). Computational methods for RNA structure validation and improvement. In *Structures of Large RNA Molecules and Their Complexes* (S. A. Woodson and F. H. Allain, eds.) **558** 181–212. Academic Press, Cambridge, MA.

JUNG, S., DRYDEN, I. L. and MARRON, J. S. (2012). Analysis of principal nested spheres. *Biometrika* **99** 551–568.

JUNG, S., FOSKEY, M. and MARRON, J. S. (2011). Principal arc analysis on direct product manifolds. *Ann. Appl. Stat*. **5** 578–603.

JUNG, S., LIU, X., MARRON, J. S. and PIZER, S. M. (2010). Generalized PCA via the backward stepwise approach in image analysis. In *Brain*, *Body and Machine*: *Proceedings of an International Symposium on the* 25*th Anniversary of McGill University Centre for Intelligent Machines*, *Advances in Intelligent and Soft Computing*. *Body and Machine* **83** 111–123. Springer, Berlin.

KENT, J. T. and MARDIA, K. V. (2009). Principal component analysis for the wrapped normal torus model. In *Proceedings of the Leeds Annual Statistical Research* (*LASR*) *Workshop* 2009.

KENT, J. T. and MARDIA, K. V. (2015). The winding number for circular data. In *Proceedings of the Leeds Annual Statistical Research* (*LASR*) *Workshop* 2015.

LABORDE, J., ROBINSON, D., SRIVASTAVA, A., KLASSEN, E. and ZHANG, J. (2013). RNA global alignment in the joint sequence–structure space using elastic shape analysis. *Nucleic Acids Res*. **41** e114–e114.

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

MARDIA, K. V. (2013). Statistical approaches to three key challenges in protein structural bioinformatics. *J*. *R*. *Stat*. *Soc*. *Ser*. *C*. *Appl*. *Stat*. **62** 487–514. MR3060628

MARDIA, K. V. and JUPP, P. E. (2000). *Directional Statistics*. **49**. Wiley, Chichester. Revised reprint of *Statistics of Directional Data* by Mardia [MR 0336854]. MR1828667

MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1979). *Multivariate Analysis*. Academic Press, London. MR0560319

MURRAY, L. J. W., ARENDALL, W. B. I., RICHARDSON, D. C. and RICHARDSON, J. S. (2003). RNA backbone is rotameric. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **100** 13904–13909.

RICHARDSON, J. S., SCHNEIDER, B., MURRAY, L. W., KAPRAL, G. J., IMMORMINO, R. M., HEADD, J. J., RICHARDSON, D. C., HAM, D., HERSHKOVITS, E., WILLIAMS, L. D., KEATING, K. S., PYLE, A. M., MICALLEF, D., WESTBROOK, J. and BERMAN, H. M. (2008). RNA backbone: Consensus all-angle conformers and modular string nomenclature (an RNA ontology consortium contribution). *RNA* **14** 465–481.

SARGSYAN, K., WRIGHT, J. and LIM, C. (2012). GeoPCA: A new tool for multivariate analysis of dihedral angles based on principal component geodesics. *Nucleic Acids Res*. **40** e25.

SCHMIDT-HIEBER, J., MUNK, A. and DÜMBGEN, L. (2013). Multiscale methods for shape constraints in deconvolution: Confidence statements for qualitative features. *Ann*. *Statist*. **41** 1299–1328.

SCHNEIDER, B., MORÁVEK, Z. and BERMAN, H. M. (2004). RNA conformational classes. *Nucleic Acids Res*. **32** 1666–1677.

SEETIN, M. G. and MATHEWS, D. H. (2012). RNA structure prediction: An overview of methods. In *Bacterial Regulatory RNA*: *Methods and Protocols* 99–122. Springer, New York.

SOMMER, S. (2013). Horizontal dimensionality reduction and iterated frame bundle and development. In *Geometric Science of Information*. *Lecture Notes in Computer Science* **8085** 76–83.

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

WADLEY, L. M., KEATING, K. S., DUARTE, C. M. and PYLE, A. M. (2007). Evaluating and learning from RNA pseudotorsional space: Quantitative validation of a reduced representation for RNA structure. *Journal of Molecular Biology* **372** 942–957.

YANG, H., JOSSINET, F., LEONTIS, N., CHEN, L., WESTBROOK, J., BERMAN, H. and WESTHOF, E. (2003). Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res*. **31** 3450–3460.