STATISTICAL SCIENCE

Volume 35, Number 1 February 2020

Special Issue on Statistics and Science

Statistical Science **[ISSN 0883-4237 (print); ISSN 2168-8745 (online)]**, Volume 35, Number 1, February 2020. Published quarterly by the Institute of Mathematical Statistics, 3163 Somerset Drive, Cleveland, OH 44122, USA. Periodicals postage paid at Cleveland, Ohio and at additional mailing offices.

POSTMASTER: Send address changes to *Statistical Science*, Institute of Mathematical Statistics, Dues and Subscriptions Office, 9650 Rockville Pike—Suite L2310, Bethesda, MD 20814-3998, USA.

Statistical Science **Volume 35, Number 1 (1–157) February 2020**

Volume 35 Number 1 February 2020

Special Issue on Statistics and Science

Statistical Science
2020, Vol. 35, No. 1, 1
https://doi.org/10.1214/20-STS351IN

Introduction to the Special Issue

Statistical Science
2020, Vol. 35, No. 1, 2–13
https://doi.org/10.1214/19-STS714

Model-Based Approach to the Joint Analysis of Single-Cell Data on Chromatin Accessibility and Gene Expression

Zhixiang Lin, Mahdi Zamanighomi, Timothy Daley, Shining Ma and Wing Hung Wong

Abstract. Unsupervised methods, including clustering methods, are essential to the analysis of single-cell genomic data. Model-based clustering methods are under-explored in the area of single-cell genomics, and have the advantage of quantifying the uncertainty of the clustering result. Here we develop a model-based approach for the integrative analysis of single-cell chromatin accessibility and gene expression data. We show that combining these two types of data, we can achieve a better separation of the underlying cell types. An efficient Markov chain Monte Carlo algorithm is also developed.

Key words and phrases: Single-cell genomics, coupled clustering, Bayesian modeling, MCMC.

REFERENCES

BACHER, R. and KENDZIORSKI, C. (2016). Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol*. **17** 63. https://doi.org/10.1186/s13059-016-0927-y

BENAGLIA, T., CHAUVEAU, D., HUNTER, D. R. and YOUNG, D. (2009). mixtools: An R package for analyzing finite mixture models. *J. Stat. Softw.* **32** 1–29.

BUENROSTRO, J. D., GIRESI, P. G., ZABA, L. C., CHANG, H. Y. and GREENLEAF, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10** 1213.

BUENROSTRO, J. D., WU, B., CHANG, H. Y. and GREENLEAF, W. J. (2015a). ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol*. **109** 21–29.

BUENROSTRO, J. D., WU, B., LITZENBURGER, U. M., RUFF, D., GONZALES, M. L., SNYDER, M. P., CHANG, H. Y. and GREENLEAF, W. J. (2015b). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523** 486–490.

CORCES, M. R., BUENROSTRO, J. D., WU, B., GREENSIDE, P. G., CHAN, S. M., KOENIG, J. L., SNYDER, M. P., PRITCHARD, J. K., KUNDAJE, A. et al. (2016). Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet*. **48** 1193–1203.

CUSANOVICH, D. A., DAZA, R., ADEY, A., PLINER, H. A., CHRISTIANSEN, L., GUNDERSON, K. L., STEEMERS, F. J., TRAPNELL, C. and SHENDURE, J. (2015). Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348** 910–914.

DIEBOLT, J. and ROBERT, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *J. Roy. Statist. Soc. Ser. B* **56** 363–375. MR1281940

DUNHAM, I., KUNDAJE, A., ALDRED, S. et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489** 57–74. https://doi.org/10.1038/nature11247

DUREN, Z., CHEN, X., JIANG, R., WANG, Y. and WONG, W. H. (2017). Modeling gene regulation from paired expression and chromatin accessibility data. *Proc. Natl. Acad. Sci. USA* **114** E4914–E4923. https://doi.org/10.1073/pnas.1704553114

DUREN, Z., CHEN, X., ZAMANIGHOMI, M., ZENG, W., SATPATHY, A., CHANG, H., WANG, Y. and WONG, W. H. (2018). Integrative analysis of single cell genomics data by coupled nonnegative matrix factorizations. *Proc. Natl. Acad. Sci. USA* **115** 7723–7728. https://doi.org/10.1073/pnas.1805681115

GRÜN, D., MURARO, M. J., BOISSET, J.-C., WIEBRANDS, K., LYUBIMOVA, A., DHARMADHIKARI, G., VAN DEN BORN, M., VAN ES, J., JANSEN, E. et al. (2016). De novo prediction of stem cell identity using single-cell transcriptome data. *Cell Stem Cell* **19** 266–277. https://doi.org/10.1016/j.stem.2016.05.010

HICKS, S. C., TOWNES, F. W., TENG, M. and IRIZARRY, R. A. (2018). Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* **19** 562–578. MR3867412 https://doi.org/10.1093/biostatistics/kxx053

KHARCHENKO, P. V., SILBERSTEIN, L. and SCADDEN, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11** 740–742. https://doi.org/10.1038/nmeth.2967

KISELEV, V. Y., KIRSCHNER, K., SCHAUB, M. T., ANDREWS, T., YIU, A., CHANDRA, T., NATARAJAN, K. N., REIK, W., BARA-

Zhixiang Lin is Assistant Professor, Department of Statistics, The Chinese University of Hong Kong, Sha Tin, Hong Kong SAR, China (e-mail: zhixianglin@cuhk.edu.hk). Mahdi Zamanighomi is Computational Biologist, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. Timothy Daley is Postdoctoral Fellow, Department of Statistics and Department of Bioengineering, Stanford University, Stanford, California, USA. Shining Ma is Postdoctoral Fellow, Department of Statistics, Stanford University, Stanford, California, USA. Wing Hung Wong is Professor, Department of Statistics and Department of Biomedical Data Science, Stanford University, Stanford, California, USA. (e-mail: whwong@stanford.edu).

HONA, M. et al. (2017). SC3: Consensus clustering of single-cell RNA-seq data. *Nat. Methods* **14** 483.

KUNDAJE, A., MEULEMAN, W., ERNST, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature* **518** 317–330. https://doi.org/10.1038/nature14248

LAKE, B. B., CHEN, S., SOS, B. C., FAN, J., KAESER, G. E., YUNG, Y. C., DUONG, T. E., GAO, D., CHUN, J. et al. (2018). Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.* **36** 70–80.

LIN, P., TROUP, M. and HO, J. W. (2017). CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol.* **18** 59.

LIN, Z., ZAMANIGHOMI, M., DALEY, T., MA, S. and WONG, W. H. (2020). Supplement to "Model-Based Approach to the Joint Analysis of Single-Cell Data on Chromatin Accessibility and Gene Expression." https://doi.org/10.1214/19-STS714SUPP.

LIU, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *J. Amer. Statist. Assoc.* **89** 958–966. MR1294740

LIU, J. S., WONG, W. H. and KONG, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81** 27–40. MR1279653 https://doi.org/10.1093/biomet/81.1.27

LOVE, M. I., HUBER, W. and ANDERS, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15** 550. https://doi.org/10.1186/s13059-014-0550-8

OLKIN, I. and RUBIN, H. (1964). Multivariate beta distributions and independence properties of the Wishart distribution. *Ann. Math. Stat.* **35** 261–269. MR0160297 https://doi.org/10.1214/aoms/1177703748

PAPASTAMOULIS, P. (2016). Label.switching: An R package for dealing with the label switching problem in MCMC outputs. *J. Stat. Softw.* **69** 1–24. https://doi.org/10.18637/jss.v069.c01.

PAPASTAMOULIS, P. and ILIOPOULOS, G. (2010). An artificial allocations based solution to the label switching problem in Bayesian analysis of mixtures of distributions. *J. Comput. Graph. Statist.* **19** 313–331. MR2758306 https://doi.org/10.1198/jcgs.2010.09008

PIERSON, E. and YAU, C. (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* **16** 241. https://doi.org/10.1186/s13059-015-0805-z

POLLEN, A. A., NOWAKOWSKI, T. J., SHUGA, J., WANG, X., LEYRAT, A. A. et al. (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* **32** 1053–1058.

RICHARDSON, S. and GREEN, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc. Ser. B* **59** 731–792. MR1483213 https://doi.org/10.1111/1467-9868.00095

RODRÍGUEZ, C. E. and WALKER, S. G. (2014). Label switching in Bayesian mixture models: Deterministic relabeling strategies. *J. Comput. Graph. Statist.* **23** 25–45. MR3173759 https://doi.org/10.1080/10618600.2012.735624

ROTEM, A., RAM, O., SHORESH, N., SPERLING, R. A., GOREN, A., WEITZ, D. A. and BERNSTEIN, B. E. (2015). Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.* **33** 1165–1172. https://doi.org/10.1038/nbt.3383

ROZENBLATT-ROSEN, O., STUBBINGTON, M. J., REGEV, A. and TEICHMANN, S. A. (2017). The human cell atlas: From vision to reality. *Nat. News* **550** 451.

SLOAN, C. A., CHAN, E. T., DAVIDSON, J. M., MALLADI, V. S., STRATTAN, J. S., HITZ, B. C. and CHERRY, J. M. (2015). ENCODE data at the ENCODE portal. *Nucleic Acids Res.* **44** D726–D732. https://doi.org/10.1093/nar/gkv1160

SMALLWOOD, S. A., LEE, H. J., ANGERMUELLER, C., KRUEGER, F., SAADEH, H., PEAT, J., ANDREWS, S. R., STEGLE, O., REIK, W. et al. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11** 817.

STEPHENS, M. (2000). Dealing with label switching in mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 795–809. MR1796293 https://doi.org/10.1111/1467-9868.00265

SUN, Z., WANG, T., DENG, K., WANG, X.-F., LAFYATIS, R., DING, Y., HU, M. and CHEN, W. (2017). DIMM-SC: A Dirichlet mixture model for clustering droplet-based single cell transcriptomic data. *Bioinformatics* **34** 139–146.

WANG, B., ZHU, J., PIERSON, E., RAMAZZOTTI, D. and BATZOGLOU, S. (2017). Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods* **14** 414–416. https://doi.org/10.1038/nmeth.4207

XU, C. and SU, Z. (2015). Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* **31** 1974–1980.

YANG, Y., HUH, R., CULPEPPER, H. W., LIN, Y., LOVE, M. I. and LI, Y. (2018). SAFE-clustering: Single-cell Aggregated (From Ensemble) clustering for single-cell RNA-seq data. *Bioinformatics*.

YAU, C. et al. (2016). PcaReduce: Hierarchical clustering of single cell transcriptional profiles. *BMC Bioinform.* **17** 140.

ZAMANIGHOMI, M., LIN, Z., DALEY, T., CHEN, X., DUREN, Z., SCHEP, A., GREENLEAF, W. J. and WONG, W. H. (2018). Unsupervised clustering and epigenetic classification of single cells. *Nat. Commun.* **9** 2410.

ZANG, C., WANG, T., DENG, K., LI, B., QIN, Q., XIAO, T., ZHANG, S., MEYER, C. A., HE, H. H. et al. (2016). High-dimensional genomic data bias correction and data integration using MANCIE. *Nat. Commun.* **7** 11305.

ZHU, L., LEI, J., DEVLIN, B. and ROEDER, K. (2018). A unified statistical framework for single cell and bulk RNA sequencing data. *Ann. Appl. Stat.* **12** 609–632. MR3773407 https://doi.org/10.1214/17-AOAS1110

ZHU, L., LEI, J., KLEI, L., DEVLIN, B. and ROEDER, K. (2019). Semisoft clustering of single-cell data. *Proc. Natl. Acad. Sci. USA* **116** 466–471. MR3904692 https://doi.org/10.1073/pnas.1817715116

Statistical Science
2020, Vol. 35, No. 1, 14–30
https://doi.org/10.1214/19-STS729

Risk Models for Breast Cancer and Their Validation

Adam R. Brentnall and Jack Cuzick

Abstract. Strategies to prevent cancer and diagnose it early when it is most treatable are needed to reduce the public health burden from rising disease incidence. Risk assessment is playing an increasingly important role in targeting individuals in need of such interventions. For breast cancer many individual risk factors have been well understood for a long time, but the development of a fully comprehensive risk model has not been straightforward, in part because there have been limited data where joint effects of an extensive set of risk factors may be estimated with precision. In this article we first review the approach taken to develop the IBIS (Tyrer–Cuzick) model, and describe recent updates. We then review and develop methods to assess calibration of models such as this one, where the risk of disease allowing for competing mortality over a long follow-up time or lifetime is estimated. The breast cancer risk model model and calibration assessment methods are demonstrated using a cohort of 132,139 women attending mammography screening in the State of Washington, USA.

Key words and phrases: Breast cancer, calibration, risk assessment, breast density, Tyrer–Cuzick model, IBIS model.

REFERENCES

[1] ALTMAN, D. G., MCSHANE, L. M., SAUERBREI, W. and TAUBE, S. E. (2012). Reporting recommendations for tumor marker prognostic studies (REMARK): Explanation and elaboration. *PLoS Med.* **9** e1001216+. https://doi.org/10.1371/journal.pmed.1001216.

[2] AMIR, E., EVANS, D. G., SHENTON, A., LALLOO, F., MORAN, A., BOGGIS, C., WILSON, M. and HOWELL, A. (2003). Evaluation of breast cancer risk assessment packages in the family history evaluation and screening programme. *J. Med. Genet.* **40** 807–814. https://doi.org/10.1136/jmg.40.11.807.

[3] ANDERSON, H., BLADSTROM, A., OLSSON, H. and MOLLER, T. R. (2000). Familial breast and ovarian cancer: A Swedish population-based register study. *Am. J. Epidemiol.* **152** 1154–1163. https://doi.org/10.1093/aje/152.12.1154.

[4] ANTONIOU, A. C., CUNNINGHAM, A. P., PETO, J., EVANS, D. G., LALLOO, F., NAROD, S. A., RISCH, H. A., EYFJORD, J. E., HOPPER, J. L. et al. (2008). The BOADICEA model of genetic susceptibility to breast and ovarian cancers: Updates and extensions. *Br. J. Cancer* **98** 1457–1466. https://doi.org/10.1038/sj.bjc.6604305.

[5] ANTONIOU, A. C., PHAROAH, P. D. P., MCMULLAN, G., DAY, N. E., STRATTON, M. R., PETO, J., PONDER, B. J. and EASTON, D. F. (2002). A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes. *Br. J. Cancer* **86** 76–83. https://doi.org/10.1038/sj.bjc.6600008

[6] ARJAS, E. (1988). A graphical method for assessing goodness of fit in Cox's proportional hazards model. *J. Amer. Statist. Assoc.* **83** 204–212.

[7] BANEGAS, M., GAIL, M., LACROIX, A., THOMPSON, B., MARTINEZ, M., WACTAWSKI-WENDE, J., JOHN, E., HUBBELL, YASMEEN, S. et al. (2012). Evaluating breast cancer risk projections for hispanic women. *Breast Cancer Res. Treat.* **132** 347–353. https://doi.org/10.1007/s10549-011-1900-9.

[8] BANEGAS, M. P., JOHN, E. M., SLATTERY, M. L., GOMEZ, S. L., YU, M., LACROIX, A. Z., PEE, D., CHLEBOWSKI, R. T., HINES, L. M. et al. (2017). Projecting individualized absolute invasive breast cancer risk in US hispanic women. *J. Natl. Cancer Inst.* **109** djw215. https://doi.org/10.1093/jnci/djw215. https://doi.org/10.1093/jnci/djw215

[9] BOUGHEY, J. C., HARTMANN, L. C., ANDERSON, S. S., DEGNIM, A. C., VIERKANT, R. A., REYNOLDS, C. A., FROST, M. H. and PANKRATZ, V. S. (2010). Evaluation of the Tyrer–Cuzick (international breast cancer intervention study) model for breast cancer risk prediction in women with atypical hyperplasia. *J. Clin. Oncol.* **28** 3591–3596. https://doi.org/10.1200/JCO.2010.28.0784.

[10] BOYD, N. F., GUO, H., MARTIN, L. J., SUN, L., STONE, J., FISHELL, E., JONG, R. A., HISLOP, G., CHIARELLI, A. et al. (2007). Mammographic density and the risk and detection of breast cancer. *N. Engl. J. Med.* **356** 227–236. https://doi.org/10.1056/nejmoa062790.

Adam R. Brentnall is Lecturer in Biostatistics, Centre for Cancer Prevention, Wolfson Institute of Preventive Medicine, Queen Mary University of London, Charterhouse square, London, EC1M 6BQ (e-mail: a.brentnall@qmul.ac.uk). Jack Cuzick is John Snow Professor of Epidemiology, Centre for Cancer Prevention, Wolfson Institute of Preventive Medicine, Queen Mary University of London, Charterhouse square, London, EC1M 6BQ (e-mail: j.cuzick@qmul.ac.uk).

[11] BOYD, N. F., O'SULLIVAN, B., CAMPBELL, J. E., FISHELL, E., SIMOR, I., COOKE, G. and GERMANSON, T. (1982). Mammographic signs as risk factors for breast cancer. *Br. J. Cancer* **45** 185–193.

[12] BRENTNALL, A. R., COHN, W. F., KNAUS, W. A., YAFFE, M. J., CUZICK, J. and HARVEY, J. A. (2019). A case-control study to add volumetric or clinical mammographic density into the Tyrer–Cuzick breast cancer risk model. *J. Breast Imaging* **1** 99–106. https://doi.org/10.1093/jbi/wbz006.

[13] BRENTNALL, A. R. and CUZICK, J. (2020). Supplement to "Risk models for breast cancer and their validation." https://doi.org/10.1214/19-STS729SUPP.

[14] BRENTNALL, A. R., CUZICK, J., BUIST, D. S. M. and BOWLES, E. J. A. (2018). Long-term Accuracy of Breast Cancer Risk Assessment Combining Classic Risk Factors and Breast Density. *JAMA Oncology* **4** e180174. https://doi.org/10.1001/jamaoncol.2018.0174.

[15] BRENTNALL, A. R., HARKNESS, E. F., ASTLEY, S. M., DONNELLY, L. S., STAVRINOS, P., SAMPSON, S., FOX, L., SERGEANT, J. C., HARVIE, M. N. et al. (2015). Mammographic density adds accuracy to both the Tyrer–Cuzick and Gail breast cancer risk models in a prospective UK screening cohort. *Breast Cancer Res.* **17** 147+. https://doi.org/10.1186/s13058-015-0653-5.

[16] CHEN, J., PEE, D., AYYAGARI, R., GRAUBARD, B., SCHAIRER, C., BYRNE, C., BENICHOU, J. and GAIL, M. H. (2006). Projecting absolute invasive breast cancer risk in white women with a model that includes mammographic density. *J. Natl. Cancer Inst.* **98** 1215–1226.

[17] CHLEBOWSKI, R. T. and ANDERSON, G. L. (2011). The influence of time from menopause and mammography on hormone therapy-related breast cancer risk assessment. *J. Natl. Cancer Inst.* **103** 284–285. https://doi.org/10.1093/jnci/djq561.

[18] CLAUS, E. B., RISCH, N., THOMPSON, W. D., CLAUS, E. B., RISCH, N. and THOMPSON, W. D. (1993). The calculation of breast cancer risk for women with a first degree family history of ovarian cancer. *Breast Cancer Res. Treat.* **28** 115–120. https://doi.org/10.1007/bf00666424.

[19] COSTANTINO, J. P., GAIL, M. H., PEE, D., ANDERSON, S., REDMOND, C. K., BENICHOU, J. and WIEAND, H. S. (1999). Validation studies for models projecting the risk of invasive and total breast cancer incidence. *J. Natl. Cancer Inst.* **91** 1541–1548.

[20] CROWDER, M. J. (2001). *Classical Competing Risks*, 1st ed. CRC Press/CRC, Boca Raton, FL.

[21] CROWSON, C. S., ATKINSON, E. J. and THERNEAU, T. M. (2016). Assessing calibration of prognostic risk scores. *Stat. Methods Med. Res.* **25** 1692–1706. MR3541121 https://doi.org/10.1177/0962280213497434

[22] CUZICK, J., BRENTNALL, A. R., SEGAL, C., BYERS, H., REUTER, C., DETRE, S., LOPEZ-KNOWLES, E., SESTAK, I., HOWELL, A. et al. (2017). Impact of a panel of 88 single nucleotide polymorphisms on the risk of breast cancer in high-risk women: Results from two randomized tamoxifen prevention trials. *J. Clin. Oncol.* **35** 743–750. https://doi.org/10.1200/JCO.2016.69.8944.

[23] CUZICK, J., SESTAK, I., CAWTHORN, S., HAMED, H., HOLLI, K., HOWELL, A. and FORBES, J. F. (2015). Tamoxifen for prevention of breast cancer: Extended long-term follow-up of the IBIS-I breast cancer prevention trial. *Lancet Oncol.* **16** 67–75. https://doi.org/10.1016/s1470-2045(14)71171-4.

[24] DECARLI, A., CALZA, S., MASALA, G., SPECCHIA, C., PALLI, D. and GAIL, M. H. (2006). Gail model for prediction of absolute risk of invasive breast cancer: Independent evaluation in the Florence-European prospective investigation into cancer and nutrition cohort. *J. Natl. Cancer Inst.* **98** 1686–1693. https://doi.org/10.1093/jnci/djj463.

[25] DEGNIM, A. C., VISSCHER, D. W., BERMAN, H. K., FROST, M. H., SELLERS, T. A., VIERKANT, R. A., MALONEY, S. D., PANKRATZ, V. S., DE GROEN, P. C. et al. (2007). Stratification of breast cancer risk in women with atypia: A mayo cohort study. *J. Clin. Oncol.* **25** 2671–2677. https://doi.org/10.1200/JCO.2006.09.0217.

[26] DE STAVOLA, B. (1987). Statistical facts about cancers on which doctor Rigoni–Stern based his contribution to the surgeons' subgroup of the IV Congress of the Italian scientists on 23 September 1842. *Stat. Med.* **6** 881–884. https://doi.org/10.1002/sim.4780060803.

[27] DUPONT, W. D., DEGNIM, A. C., SANDERS, M. E., SIMPSON, J. F. and HARTMANN, L. C. (2018). Risk Factors for Breast Carcinoma in Women With Proliferative Breast Disease. In *The Breast* 264–271.e2. Elsevier, Amsterdam. https://doi.org/10.1016/b978-0-323-35955-9.00020-9.

[28] EASTON, D. F., PHAROAH, P. D., ANTONIOU, A. C., TISCHKOWITZ, M., TAVTIGIAN, S. V., NATHANSON, K. L., DEVILEE, P., MEINDL, A., COUCH, F. J. et al. (2015). Gene-panel sequencing and the prediction of breast-cancer risk. *N. Engl. J. Med.* **372** 2243–2257.

[29] EASTON, D. F., PHAROAH, P. D. P., ANTONIOU, A. C., TISCHKOWITZ, M., TAVTIGIAN, S. V., NATHANSON, K. L., DEVILEE, P., MEINDL, A., COUCH, F. J. et al. (2015). Gene-panel sequencing and the prediction of breast-cancer risk. *N. Engl. J. Med.* **372** 2243–2257. https://doi.org/10.1056/NEJMsr1501341.

[30] EVANS, D. G., BRENTNALL, A., BYERS, H., HARKNESS, E., STAVRINOS, P., HOWELL, A., NEWMAN, W. G. and CUZICK, J. (2017). The impact of a panel of 18 SNPs on breast cancer risk in women attending a UK familial screening clinic: A case-control study. *J. Med. Genet.* **54** 111–113. https://doi.org/10.1136/jmedgenet-2016-104125.

[31] FORD, D., EASTON, D. F., STRATTON, M., NAROD, S., GOLDGAR, D., DEVILEE, P., BISHOP, D. T., WEBER, B., LENOIR, G. et al. (1998). Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. *Am. J. Hum. Genet.* **62** 676–689.

[32] GAIL, M. H., BRINTON, L. A., BYAR, D. P., CORLE, D. K., GREEN, S. B., SCHAIRER, C. and MULVIHILL, J. J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J. Natl. Cancer Inst.* **81** 1879–1886. https://doi.org/10.1093/jnci/81.24.1879.

[33] GAIL, M. H. and PFEIFFER, R. M. (2005). On criteria for evaluating models of absolute risk. *Biostatistics* **6** 227–239. https://doi.org/10.1093/biostatistics/kxi005.

[34] GRØNNESBY, J. K. and BORGAN, Ø. (1996). A method for checking regression models in survival analysis based on the risk score. *Lifetime Data Anal.* **2** 315–328. https://doi.org/10.1007/BF00127305.

[35] COLLABORATIVE GROUP ON HORMONAL FACTORS IN BREAST CANCER (2002). Alcohol, tobacco and breast cancer—collaborative reanalysis of individual data from 53 epidemiological studies, including 58 515 women with breast cancer and 95 067 women without the disease. *Br. J. Cancer* **87** 1234–1245. https://doi.org/10.1038/sj.bjc.6600596.

[36] HARRELL, F. E., LEE, K. L. and MARK, D. B. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **15** 361–387. https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4%3C361::AID-SIM168%3E3.0.CO;2-4.

[37] HIDAYAT, K., YANG, C. M. and SHI, B. M. (2018). Body fatness at a young age, body fatness gain and risk of breast cancer: Systematic review and meta-analysis of cohort studies. *Obes. Rev.* **19** 254–268. https://doi.org/10.1111/obr.12627.

[38] HOSMER, D. W., LEMESHOW, S. and STURDIVANT, R. X. (2013). *Applied Logistic Regression*. Wiley, Hoboken, NJ, USA. https://doi.org/10.1002/9781118548387.

[39] JOHN, E. M. (2005). Migration history, acculturation, and breast cancer risk in hispanic women. *Cancer Epidemiol. Biomark. Prev.* **14** 2905–2913. https://doi.org/10.1158/1055-9965.EPI-05-0483.

[40] KEIDING, N. (1987). The method of expected number of deaths, 1786–1886–1986. *Int. Stat. Rev.* **55** 1–20. MR0962938 https://doi.org/10.2307/1403267

[41] LAWLESS, J. F. (2003). *Statistical Models and Methods for Lifetime Data*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley-Interscience, Hoboken, NJ. MR1940115

[42] LEE, A. J., CUNNINGHAM, A. P., TISCHKOWITZ, M., SIMARD, J., PHAROAH, P. D., EASTON, D. F. and ANTONIOU, A. C. (2016). Incorporating truncating variants in PALB2, CHEK2, and ATM into the BOADICEA breast cancer risk model. *Genet. Med.* **18** 1190–1198. https://doi.org/10.1038/gim.2016.31.

[43] MACMAHON, B. and COLE, P. (1969). Endocrinology and epidemiology of breast cancer. *Cancer* **24** 1146–1150. https://doi.org/10.1002/1097-0142(196912)24:6%3C1146::aid-cncr2820240612%3E3.0.co;2-0.

[44] MAVADDAT, N., MICHAILIDOU, K., DENNIS, J., LUSH, M., FACHAL, L., LEE, A., TYRER, J. P., CHEN, T.-H., WANG, Q. et al. (2019). Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *Am. J. Hum. Genet.* **104** 21–34. https://doi.org/10.1016/j.ajhg.2018.11.002.

[45] MCCORMACK, V. A. and DOS SANTOS SILVA, I. (2006). Breast density and parenchymal patterns as markers of breast cancer risk: A meta-analysis. *Cancer Epidemiol. Biomark. Prev.* **15** 1159–1169. https://doi.org/10.1158/1055-9965.epi-06-0034.

[46] MEALIFFE, M. E., STOKOWSKI, R. P., RHEES, B. K., PRENTICE, R. L., PETTINGER, M. and HINDS, D. A. (2010). Assessment of clinical validity of a breast cancer risk model combining genetic and clinical information. *J. Natl. Cancer Inst.* **102** 1618–1627. https://doi.org/10.1093/jnci/djq388.

[47] COLLABORATIVE GROUP ON HORMONAL FACTORS IN BREAST CANCER (2012). Menarche, menopause, and breast cancer risk: Individual participant meta-analysis, including 118–964 women with breast cancer from 117 epidemiological studies. *Lancet Oncol.* **13** 1141–1151. https://doi.org/10.1016/s1470-2045(12)70425-4.

[48] PARMIGIANI, G., BERRY, D. A. and AGUILAR, O. (1998). Determining carrier probabilities for breast cancer-susceptibility genes BRCA1 and BRCA2. *Am. J. Hum. Genet.* **62** 145–158. https://doi.org/10.1086/301670.

[49] PIKE, M. C., KRAILO, M. D., HENDERSON, B. E., CASAGRANDE, J. T. and HOEL, D. G. (1983). 'Hormonal' risk factors, 'breast tissue age' and the age-incidence of breast cancer. *Nature* **303** 767–770.

[50] PRENTICE, R. L., CAAN, B., CHLEBOWSKI, R. T., PATTERSON, R., KULLER, L. H., OCKENE, J. K., MARGOLIS, K. L., LIMACHER, M. C., MANSON, J. E. et al. (2006). Low-fat dietary pattern and risk of invasive breast cancer. *JAMA* **295** 629. https://doi.org/10.1001/jama.295.6.629.

[51] REEVES, G. K., BERAL, V., GREEN, J., GATHANI, T. and BULL, D. (2006). Hormonal therapy for menopause and breast-cancer risk by histological type: A cohort study and meta-analysis. *Lancet Oncol.* **7** 910–918. https://doi.org/10.1016/S1470-2045(06)70911-1.

[52] RICE, M. S., TWOROGER, S. S., HANKINSON, S. E., TAMIMI, R. M., ELIASSEN, A. H., WILLETT, W. C., COLDITZ, G. and ROSNER, B. (2017). Breast cancer risk prediction: An update to the Rosner–Colditz breast cancer incidence model. *Breast Cancer Res. Treat.* **166** 227–240. https://doi.org/10.1007/s10549-017-4391-5.

[53] RIGONI-STERN (1842). Fatti statistici relativi alle malattie cancerose. *Giorn. Prog. Patol. Terap.* **2** 507–517.

[54] ROCKHILL, B., SPIEGELMAN, D., BYRNE, C., HUNTER, D. J. and COLDITZ, G. A. (2001). Validation of the Gail et al. Model of breast cancer risk prediction and implications for chemoprevention. *J. Natl. Cancer Inst.* **93** 358–366. https://doi.org/10.1093/jnci/93.5.358.

[55] ROSNER, B. A., COLDITZ, G. A., HANKINSON, S. E., SULLIVAN-HALLEY, J., LACEY, J. V. and BERNSTEIN, L. (2013). Validation of Rosner–Colditz breast cancer incidence model using an independent data set, the California Teachers Study. *Breast Cancer Res. Treat.* **142** 187–202. https://doi.org/10.1007/s10549-013-2719-3.

[56] SASIENI, P. and BRENTNALL, A. R. (2017). On standardized relative survival. *Biometrics* **73** 473–482. MR3665964 https://doi.org/10.1111/biom.12578

[57] SASLOW, D., BOETES, C., BURKE, W., HARMS, S., LEACH, M. O., LEHMAN, C. D., MORRIS, E., PISANO, E., SCHNALL, M. et al. (2007). American cancer society guidelines for breast screening with MRI as an adjunct to mammography. *CA Cancer J. Clin.* **57** 75–89. https://doi.org/10.3322/canjclin.57.2.75.

[58] SCHOEMAKER, M. J., NICHOLS, H. B., WRIGHT, L. B., BROOK, M. N., JONES, M. E., O'BRIEN, K. M., ADAMI, H.-O., BAGLIETTO, L., BERNSTEIN, L. et al. (2018). Association of body mass index and age with subsequent breast cancer risk in premenopausal women. *JAMA Oncol.* **4** e181771. https://doi.org/10.1001/jamaoncol.2018.1771.

[59] STEEL, M., THOMPSON, A. and CLAYTON, J. (1991). Genetic aspects of breast cancer. *Br. Med. Bull.* **47** 504–518.

[60] TEAMS, F. C. (2010). Mammographic surveillance in women younger than 50 years who have a family history of breast cancer: Tumour characteristics and projected effect on mortality in the prospective, single-arm, FH01 study. *Lancet Oncol.* **11** 1127–1134. https://doi.org/10.1016/s1470-2045(10)70263-1.

[61] TICE, J. A., CUMMINGS, S. R., SMITH-BINDMAN, R., ICHIKAWA, L., BARLOW, W. E. and KERLIKOWSKE, K. (2008). Using clinical factors and mammographic breast density to estimate breast cancer risk: Development and validation of a new predictive model. *Ann. Intern. Med.* **148** 337–347.

[62] TICE, J. A., MIGLIORETTI, D. L., LI, C.-S., VACHON, C. M., GARD, C. C. and KERLIKOWSKE, K. (2015). Breast density and benign breast disease: Risk assessment to identify women at high risk of breast cancer. *J. Clin. Oncol.* **33** JCO.2015.60.8869–3143. https://doi.org/10.1200/jco.2015.60.8869.

[63] TYRER, J., DUFFY, S. W. and CUZICK, J. (2004). A breast cancer prediction model incorporating familial and personal risk factors. *Stat. Med.* **23** 1111–1130. https://doi.org/10.1002/sim.1668.

[64] VACHON, C. M., PANKRATZ, V. S., SCOTT, C. G., HAEBERLE, L., ZIV, E., JENSEN, M. R., BRANDT, K. R., WHALEY, D. H., OLSON, J. E. et al. (2015). The Contributions of Breast Density and Common Genetic Variation to Breast Cancer Risk. *J Natl Cancer Inst* **107** dju397+. https://doi.org/10.1093/jnci/dju397.

[65] VAN VEEN, E. M., BRENTNALL, A. R., BYERS, H., HARKNESS, E. F., ASTLEY, S. M., SAMPSON, S., HOWELL, A., NEWMAN, W. G., CUZICK, J. et al. (2018). Use of single-nucleotide polymorphisms and mammographic density plus clas-

sic risk factors for breast cancer risk prediction. *JAMA Oncol*. **4** 476. https://doi.org/10.1001/jamaoncol.2017.4881.

[66] WACHOLDER, S., HARTGE, P., PRENTICE, R., GARCIA-CLOSAS, M., FEIGELSON, H. S., DIVER, W. R., THUN, M. J., COX, D. G., HANKINSON, S. E. et al. (2010). Performance of common genetic variants in breast-cancer risk models. *N. Engl. J. Med*. **362** 986–993. https://doi.org/10.1056/nejmoa0907727.

[67] WANG, C., BRENTNALL, A. R., CUZICK, J., HARKNESS, E. F., EVANS, D. G. and ASTLEY, S. (2017). A novel and fully automated mammographic texture analysis for risk prediction: Results from two case-control studies. *Breast Cancer Res*. **19** 114. https://doi.org/10.1186/s13058-017-0906-6.

[68] WARWICK, J., BIRKE, H., STONE, J., WARREN, R. M. L., PINNEY, E., BRENTNALL, A. R., DUFFY, S. W., HOWELL, A. and CUZICK, J. (2014). Mammographic breast density refines Tyrer–Cuzick estimates of breast cancer risk in high-risk women: findings from the placebo arm of the International Breast Cancer Intervention Study I. *Breast Cancer Research* **16** 451+. https://doi.org/10.1186/s13058-014-0451-5.

[69] WOLFE, J. N. (1976). Breast patterns as an index of risk for developing breast cancer. *Am. J. Roentgenol*. **126** 1130–1137.

[70] WOOD, S. N. (2006). *Generalized Additive Models*: *An Introduction with R*. *Texts in Statistical Science Series*. CRC Press/CRC, Boca Raton, FL. MR2206355

[71] WU, Y., ZHANG, D. and KANG, S. (2013). Physical activity and risk of breast cancer: A meta-analysis of prospective studies. *Breast Cancer Res. Treat*. **137** 869–882. https://doi.org/10.1007/s10549-012-2396-7.

[72] ZHANG, X., RICE, M., TWOROGER, S. S., ROSNER, B. A., ELIASSEN, A. H., TAMIMI, R. M., JOSHI, A. D., LINDSTROM, S., QIAN, J. et al. (2018). Addition of a polygenic risk score, mammographic density, and endogenous hormones to existing breast cancer risk prediction models: A nested case-control study. *PLoS Med*. **15** e1002644. https://doi.org/10.1371/journal.pmed.1002644.

Statistical Science
2020, Vol. 35, No. 1, 31–41
https://doi.org/10.1214/19-STS730

Some Statistical Issues in Climate Science

Michael L. Stein

Abstract. Climate science is a field that is arguably both data-rich and data-poor. Data rich in that huge and quickly increasing amounts of data about the state of the climate are collected every day. Data poor in that important aspects of the climate are still undersampled, such as the deep oceans and some characteristics of the upper atmosphere. Data rich in that modern climate models can produce climatological quantities over long time periods with global coverage, including quantities that are difficult to measure and under conditions for which there is no data presently. Data poor in that the correspondence between climate model output to the actual climate, especially for future climate change due to human activities, is difficult to assess. The scope for fruitful interactions between climate scientists and statisticians is great, but requires serious commitments from researchers in both disciplines to understand the scientific and statistical nuances arising from the complex relationships between the data and the real-world problems. This paper describes a small fraction of some of the intellectual challenges that occur at the interface between climate science and statistics, including inferences for extremes for processes with seasonality and long-term trends, the use of climate model ensembles for studying extremes, the scope for using new data sources for studying space-time characteristics of environmental processes and a discussion of non-Gaussian space-time process models for climate variables. The paper concludes with a call to the statistical community to become more engaged in one of the great scientific and policy issues of our time, anthropogenic climate change and its impacts.

Key words and phrases: Statistical climatology, climate extremes, Argo network, non-Gaussian processes.

REFERENCES

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

ANDERSON, B. and BELL, M. (2009). Weather-related mortality: How heat, cold, and heat waves affect mortality in the United States. *Epidemiology* **20** 205–213.

ARGO (2000). Argo float data and metadata from Global Data Assembly Centre (Argo GDAC).

ARRHENIUS, S. (1896). On the influence of carbonic acid in the air on the temperature on the ground. *Philos. Mag.* **41** 237–276.

ARRHENIUS, S. (1908). *Worlds in the Making*: *The Evolution of the Universe*. Harper & Brothers Publishers, New York.

BALSAMO, G., ALBERGEL, C., BELJAARS, A., BOUSSETTA, S., BRUN, E., CLOKE, H., DEE, D., DUTRA, E., MUÑOZ-SABATER, J. et al. (2015). ERA-Interim/Land: A global land surface reanalysis data set. *Hydrol. Earth Syst. Sci.* **19** 389–407.

BELL, T. L. (1987). A space-time stochastic model of rainfall for satellite remote-sensing studies. *J. Geophys. Res., Atmos.* **92** 9631–9643.

BERLINER, L. M., LEVINE, R. A. and SHEA, D. J. (2000). Bayesian climate change assessment. *J. Climate* **13** 3805–3820.

BERNER, J., ACHATZ, U., BATTÉ, L., BENGTSSON, L., DE LA CÁMARA, A., CHRISTENSEN, H. M., COLANGELI, M., COLEMAN, D. R. B., CROMMELIN, D. et al. (2017). Stochastic parameterization: Toward a new view of weather and climate models. *Bull. Am. Meteorol. Soc.* **98** 565–588.

BERROCAL, V. J., RAFTERY, A. E. and GNEITING, T. (2008). Probabilistic quantitative precipitation field forecasting using a two-stage spatial model. *Ann. Appl. Stat.* **2** 1170–1193. MR2655654 https://doi.org/10.1214/08-AOAS203

CASTRUCCIO, S., MCINERNEY, D. J., STEIN, M. L., LIU CROUCH, F., JACOB, R. L. and MOYER, E. J. (2014). Statistical emulation of climate model projections based on precomputed GCM runs. *J. Climate* **27** 1829–1844.

CHEN, D., OU, T. and GONG, L. (2010). Spatial interpolation of daily precipitation in China: 1951–2005. *Adv. Atmos. Sci.* **27** 1221–1232.

CHENG, L., AGHAKOUCHAK, A., GILLELAND, E. and KATZ, R. W. (2014). Non-stationary extreme value analysis in a changing climate. *Clim. Change* **127** 353–369.

Michael L. Stein is Distinguished Professor of Statistics, Department of Statistics, Rutgers University, Piscataway, New Jersey 08854, USA (e-mail: stein@galton.uchicago.edu).

COLLINS, M., BOOTH, B. B. B., BHASKARAN, B., HARRIS, G. R., MURPHY, J. M., SEXTON, D. M. H. and WEBB, M. J. (2011). Climate model errors, feedbacks and forcings: A comparison of perturbed physics and multi-model ensembles. *Clim. Dyn.* **36** 1737–1766.

COOLEY, D. and SAIN, S. R. (2010). Spatial hierarchical modeling of precipitation extremes from a regional climate model. *J. Agric. Biol. Environ. Stat.* **15** 381–402. MR2787265 https://doi.org/10.1007/s13253-010-0023-9

COOLEY, D., CISEWSKI, J., ERHARDT, R. J., JEON, S., MANNSHARDT, E., OMOLO, B. O. and SUN, Y. (2012). A survey of spatial extremes: Measuring spatial dependence and modeling spatial effects. *REVSTAT* **10** 135–165. MR2912374

CORTI, S., PALMER, T., BALMASEDA, M., WEISHEIMER, A., DRIJFHOUT, S., DUNSTONE, N., HAZELEGER, W., KRÖGER, J., POHLMANN, H. et al. (2015). Impact of initial conditions versus external forcing in decadal climate predictions: A sensitivity experiment. *J. Climate* **28** 4454–4470.

CRESSIE, N. and HUANG, H.-C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *J. Amer. Statist. Assoc.* **94** 1330–1340. MR1731494 https://doi.org/10.2307/2669946

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

DAVISON, A. C. and HUSER, R. (2015). Statistics of extremes. *Annu. Rev. Stat. Appl.* **2** 203–235.

DAVISON, A. C., PADOAN, S. A. and RIBATET, M. (2012). Statistical modeling of spatial extremes. *Statist. Sci.* **27** 161–186. MR2963980 https://doi.org/10.1214/11-STS376

DE HAAN, L. and FERREIRA, A. (2006). *Extreme Value Theory: An Introduction. Springer Series in Operations Research and Financial Engineering*. Springer, New York. MR2234156 https://doi.org/10.1007/0-387-34471-3

DIGGLE, P. J., TAWN, J. A. and MOYEED, R. A. (1998). Model-based geostatistics. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **47** 299–350. MR1626544 https://doi.org/10.1111/1467-9876.00113

DI LUZIO, M., JOHNSON, G. L., DALY, C., EISCHEID, J. K. and ARNOLD, J. G. (2008). Constructing retrospective gridded daily precipitation and temperature datasets for the conterminous United States. *J. Appl. Meteorol. Climatol.* **47** 475–497.

EDWARDS, P. N. (2011). History of climate modeling. *Wiley Interdiscip. Rev.: Clim. Change* **2** 128–139.

EINMAHL, J. H. J., DE HAAN, L. and ZHOU, C. (2016). Statistics of heteroscedastic extremes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 31–51. MR3453645 https://doi.org/10.1111/rssb.12099

FAWCETT, L. and WALSHAW, D. (2007). Improved estimation for temporally clustered extremes. *Environmetrics* **18** 173–188. MR2345653 https://doi.org/10.1002/env.810

FERRO, C. A. T. and SEGERS, J. (2003). Inference for clusters of extreme values. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 545–556. MR1983763 https://doi.org/10.1111/1467-9868.00401

FIELD, C., BARROS, V., STOCKER, D., QIN, D., DOKKEN, K., EBI, M., MASTRANDREA, K., MACH, G.-K., PLATTNER, S. et al., eds. (2012). *IPCC*, 2012: *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change*. Cambridge Univ. Press, Cambridge.

FORGET, G., CAMPIN, J. M., HEIMBACH, P., HILL, C. N., PONTE, R. M. and WUNSCH, C. (2015). ECCO version 4: An integrated framework for non-linear inverse modeling and global ocean state estimation. *Geosci. Model Dev.* **8** 3071–3104.

GASPARRINI, A. and ARMSTRONG, B. (2011). The impact of heat waves on mortality. *Epidemiology* **22** 68–73. https://doi.org/10.1097/EDE.0b013e3181fdcd99

GNEITING, T. (2002). Nonseparable, stationary covariance functions for space-time data. *J. Amer. Statist. Assoc.* **97** 590–600. MR1941475 https://doi.org/10.1198/016214502760047113

GRAY, A. R. and RISER, S. C. (2014). A global analysis of Sverdrup balance using absolute geostrophic velocities from Argo. *J. Phys. Oceanogr.* **44** 1213–1229.

HAUGEN, M. A., STEIN, M. L., MOYER, E. J. and SRIVER, R. L. (2018). Estimating changes in temperature distributions in a large ensemble of climate simulations using quantile regression. *J. Climate* **31** 8573–8588.

HOLLMANN, R., MERCHANT, C. J., SAUNDERS, R., DOWNY, C., BUCHWITZ, M., CAZENAVE, A., CHUVIECO, E., DEFOURNY, P., DE LEEUW, G. et al. (2013). The ESA climate change initiative: Satellite data records for essential climate variables. *Bull. Am. Meteorol. Soc.* **94** 1541–1552.

HOSKING, J. R. M., WALLIS, J. R. and WOOD, E. F. (1985). An appraisal of the regional flood frequency procedure in the UK Flood Studies Report. *Hydrol. Sci. J.* **30** 85–109.

HUANG, W. K., STEIN, M. L., MCINERNEY, D. J., SUN, S. and MOYER, E. J. (2016). Estimating changes in temperature extremes from millennial-scale climate simulations using generalized extreme value (GEV) distributions. *Adv. Stat. Climatol. Meteorol. Oceanogr.* **2** 79–103.

HUTCHINSON, M. F., MCKENNEY, D. W., LAWRENCE, K., PEDLAR, J. H., HOPKINSON, R. F., MILEWSKA, E. and PAPADOPOL, P. (2009). Development and testing of Canada-wide interpolated spatial models of daily minimum-maximum temperature and precipitation for 1961–2003. *J. Appl. Meteorol. Climatol.* **48** 725–741.

IPCC (2013). *Climate Change* 2013: *The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge Univ. Press, Cambridge.

JALBERT, J., FAVRE, A.-C., BÉLISLE, C. and ANGERS, J.-F. (2017). A spatiotemporal model for extreme precipitation simulated by a climate model, with an application to assessing changes in return levels over North America. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **66** 941–962. MR3715590 https://doi.org/10.1111/rssc.12212

JUN, M., KNUTTI, R. and NYCHKA, D. W. (2008). Spatial analysis to quantify numerical model bias and dependence: How many climate models are there? *J. Amer. Statist. Assoc.* **103** 934–947. MR2528820 https://doi.org/10.1198/016214507000001265

KATZ, R. W. and BROWN, B. G. (1992). Extreme events in a changing climate: Variability is more important than averages. *Clim. Change* **21** 289–302.

KATZ, R. W., PARLANGE, M. B. and NAVEAU, P. (2002). Statistics of extremes in hydrology. *Adv. Water Resour.* **25** 1287–1304.

KAY, J. E., DESER, C., PHILLIPS, A., MAI, A., HANNAY, C., STRAND, G., ARBLASTER, J. M., BATES, S. C., DANABASOGLU, G. et al. (2015). The Community Earth System Model (CESM) Large Ensemble Project: A community resource for studying climate change in the presence of internal climate variability. *Bull. Am. Meteorol. Soc.* **96** 1333–1349.

KHARIN, V. V. and ZWIERS, F. W. (2005). Estimating extremes in transient climate change simulations. *J. Climate* **18** 1156–1173.

KHARIN, V. V., ZWIERS, F. W., ZHANG, X. and WEHNER, M. (2013). Changes in temperature and precipitation extremes in the CMIP5 ensemble. *Clim. Change* **119** 345–357.

KNUTTI, R., FURRER, R., TEBALDI, C., CERMAK, J. and MEEHL, G. A. (2010). Challenges in combining projections from multiple climate models. *J. Climate* **23** 2739–2758.

KUUSELA, M. and STEIN, M. L. (2018). Locally stationary spatio-temporal interpolation of Argo profiling float data. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **474** 20180400.

López-Lopera, A. F., Bachoc, F., Durrande, N. and Roustant, O. (2018). Finite-dimensional Gaussian approximation with linear inequality constraints. *SIAM/ASA J. Uncertain. Quantificat.* **6** 1224–1255. MR3857898 https://doi.org/10.1137/17M1153157

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *J. Atmos. Sci.* **20** 130–141.

Mearns, L. O., Sain, S., Leung, L. R., Bukovsky, M. S., McGinnis, S., Biner, S., Caya, D., Arritt, R. W., Gutowski, W. et al. (2013). Climate change projections of the North American Regional Climate Change Assessment Program (NARCCAP). *Clim. Change* **120** 965–975.

Neelin, J. D., Bracco, A., Luo, H., McWilliams, J. C. and Meyerson, J. E. (2010). Considerations for parameter optimization and sensitivity in climate models. *Proc. Natl. Acad. Sci. USA* **107** 21349–21354.

Nordhaus, W. (2018). Projections and uncertainties about climate change in an era of minimal climate policies. *Am. Econ. J. Econ. Policy* **10** 333–360.

Institute of Hydrology (1975). *Flood Studies Report*. Natural Environment Research Council, London.

Olson, B. and Kleiber, W. (2017). Approximate Bayesian computation methods for daily spatiotemporal precipitation occurrence simulation. *Water Resour. Res.* **53** 3352–3372.

Opitz, T., Huser, R., Bakka, H. and Rue, H. (2018). INLA goes extreme: Bayesian tail regression for the estimation of high spatio-temporal quantiles. *Extremes* **21** 441–462. MR3855716 https://doi.org/10.1007/s10687-018-0324-x

Paynter, D., Frölicher, T. L., Horowitz, L. W. and Silvers, L. G. (2018). Equilibrium climate sensitivity obtained from multimillennial runs of two GFDL climate models. *J. Geophys. Res., Atmos.* **123** 1921–1941.

Peña-Arancibia, J. L., van Dijk, A. I. J. M., Renzullo, L. J. and Mulligan, M. (2013). Evaluation of precipitation estimation accuracy in reanalyses, satellite products, and an ensemble method for regions in Australia and South and East Asia. *J. Hydrometeorol.* **14** 1323–1333.

Pierrehumbert, R. T. (2010). *Principles of Planetary Climate*. Cambridge Univ. Press, Cambridge. MR2778154 https://doi.org/10.1017/CBO9780511780783

Rahmstorf, S., Cazenave, A., Church, J. A., Hansen, J. E., Keeling, R. F., Parker, D. E. and Somerville, R. C. J. (2007). Recent climate observations compared to projections. *Science* **316** 709–709.

Ramanathan, V. (1988). The greenhouse theory of climate change: A test by an inadvertent global experiment. *Science* **240** 293–299.

Ribes, A., Zwiers, F. W., Azaïs, J.-M. and Naveau, P. (2017). A new statistical approach to climate change detection and attribution. *Clim. Dyn.* **48** 367–386.

Riser, S. C., Freeland, H. J., Roemmich, D., Wijffels, S., Troisi, A., Belbeoch, M., Gilbert, D., Xu, J., Pouliquen, S. et al. (2016). Fifteen years of ocean observations with the global Argo array. *Nat. Clim. Change* **6** 145–153.

Rodda, J. C. and Dixon, H. (2012). Rainfall measurement revisited. *Weather* **67** 131–136.

Roth, M., Jongbloed, G. and Buishand, A. (2019). Monotone trends in the distribution of climate extremes. *Theor. Appl. Climatol.* **136** 1175–1184.

Rummukainen, M. (2012). Changes in climate and weather extremes in the 21st century. *Wiley Interdiscip. Rev.: Clim. Change* **3** 115–129.

Saha, S., Moorthi, S., Pan, H.-L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J. et al. (2010). The NCEP climate forecast system reanalysis. *Bull. Am. Meteorol. Soc.* **91** 1015–1058.

Schmit, T. J., Griffith, P., Gunshor, M. M., Daniels, J. M., Goodman, S. J. and Lebair, W. J. (2017). A closer look at the ABI on the GOES-R series. *Bull. Am. Meteorol. Soc.* **98** 681–698.

Sexton, D. M. H., Murphy, J. M., Collins, M. and Webb, M. J. (2012). Multivariate probabilistic projections using imperfect climate models part I: Outline of methodology. *Clim. Dyn.* **38** 2513–2542.

Sigrist, F., Künsch, H. R. and Stahel, W. A. (2015). Stochastic partial differential equation based modelling of large space-time data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 3–33. MR3299397 https://doi.org/10.1111/rssb.12061

Sriver, R. L., Forest, C. E. and Keller, K. (2015). Effects of initial conditions uncertainty on regional climate variability: An analysis using a low-resolution CESM ensemble. *Geophys. Res. Lett.* **42** 5468–5476.

Stammer, D., Balmaseda, M., Heimbach, P., Köhl, A. and Weaver, A. (2016). Ocean data assimilation in support of climate applications: Status and perspectives. *Annu. Rev. Mar. Sci.* **8** 491–518. https://doi.org/10.1146/annurev-marine-122414-034113

Stein, M. L. (2005). Space-time covariance functions. *J. Amer. Statist. Assoc.* **100** 310–321. MR2156840 https://doi.org/10.1198/016214504000000854

Stein, M. L. (2017). Should annual maximum temperatures follow a generalized extreme value distribution? *Biometrika* **104** 1–16. MR3626483 https://doi.org/10.1093/biomet/asw070

Sterl, A., Severijns, C., Dijkstra, H., Hazeleger, W., Jan van Oldenborgh, G., van den Broeke, M., Burgers, G., van den Hurk, B., Jan van Leeuwen, P. et al. (2012). When can we expect extremely high surface temperatures? *Geophys. Res. Lett.* **35** L14703.

Sun, Y. and Stein, M. L. (2015). A stochastic space-time model for intermittent precipitation occurrences. *Ann. Appl. Stat.* **9** 2110–2132. MR3456368 https://doi.org/10.1214/15-AOAS875

Taylor, K. E., Stouffer, R. J. and Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bull. Am. Meteorol. Soc.* **93** 485–498.

Tebaldi, C. and Knutti, R. (2007). The use of the multi-model ensemble in probabilistic climate projections. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **365** 2053–2075. MR2317897 https://doi.org/10.1098/rsta.2007.2076

Touma, D., Michalak, A. M., Swain, D. L. and Diffenbaugh, N. S. (2018). Characterizing the spatial scales of extreme daily precipitation in the United States. *J. Climate* **31** 8023–8037.

Troy, T. J., Kipgen, C. and Pal, I. (2015). The impact of climate extremes and irrigation on US crop yields. *Environ. Res. Lett.* **10** 054013.

Vardi, M. Y. (2010). Science has only two legs. *Commun. ACM* **53** 5.

Wallin, J. and Bolin, D. (2015). Geostatistical modelling using non-Gaussian Matérn fields. *Scand. J. Stat.* **42** 872–890. MR3391697 https://doi.org/10.1111/sjos.12141

Wang, X. and Berger, J. O. (2016). Estimating shape constrained functions using Gaussian processes. *SIAM/ASA J. Uncertain. Quantificat.* **4** 1–25. MR3452261 https://doi.org/10.1137/140955033

Westra, S., Alexander, L. V. and Zwiers, F. W. (2013). Global increasing trends in annual maximum daily precipitation. *J. Climate* **26** 3904–3918.

Wikle, C. K. (2015). Modern perspectives on statistics for spatio-temporal data. *Wiley Interdiscip. Rev.: Comput. Stat.* **7** 86–98. MR3348724 https://doi.org/10.1002/wics.1341

Wilks, D. S. and Wilby, R. L. (1999). The weather generation game: A review of stochastic weather models. *Prog. Phys. Geogr., Earth Environ.* **23** 329–357.

XU, G. and GENTON, M. G. (2017). Tukey g-and-h random fields. *J. Amer. Statist. Assoc.* **112** 1236–1249. MR3735373 https://doi.org/10.1080/01621459.2016.1205501

YATAGAI, A., KAMIGUCHI, K., ARAKAWA, O., HAMADA, A., YASUTOMI, N. and KITOH, A. (2012). APHRODITE: Constructing a long-term daily gridded precipitation dataset for Asia based on a dense network of rain gauges. *Bull. Am. Meteorol. Soc.* **93** 1401–1415.

YOKOHATA, T., WEBB, M. J., COLLINS, M., WILLIAMS, K. D., YOSHIMORI, M., HARGREAVES, J. C. and ANNAN, J. D. (2010). Structural similarities and differences in climate responses to CO_2 increase between two perturbed physics ensembles. *J. Climate* **23** 1392–1410.

ZWIERS, F. W., ALEXANDER, L. V., HEGERL, G. C., KNUTSON, T. R., KOSSIN, J. P., NAVEAU, P., NICHOLLS, N., SCHÄR, C., SENEVIRATNE, S. I. et al. (2013). Climate extremes: Challenges in estimating and understanding recent changes in the frequency and intensity of extreme climate and weather events. In *Climate Science for Serving Society* (G. Asrar and J. Hurrell, eds.) 339–389. Springer, Dordrecht.

Statistical Science
2020, Vol. 35, No. 1, 42–50
https://doi.org/10.1214/19-STS738

A Tale of Two Parasites: Statistical Modelling to Support Disease Control Programmes in Africa

Peter J. Diggle, Emanuele Giorgi, Julienne Atsame, Sylvie Ntsame Ella, Kisito Ogoussan and Katherine Gass

Abstract. Vector-borne diseases have long presented major challenges to the health of rural communities in the wet tropical regions of the world, but especially in sub-Saharan Africa. In this paper, we describe the contribution that statistical modelling has made to the global elimination programme for one vector-borne disease, onchocerciasis.

We explain why information on the spatial distribution of a second vector-borne disease, Loa loa, is needed before communities at high risk of onchocerciasis can be treated safely with mass distribution of ivermectin, an antifiarial medication.

We show how a model-based geostatistical analysis of Loa loa prevalence survey data can be used to map the predictive probability that each location in the region of interest meets a WHO policy guideline for safe mass distribution of ivermectin and describe two applications: one is to data from Cameroon that assesses prevalence using traditional blood-smear microscopy; the other is to Africa-wide data that uses a low-cost questionnaire-based method.

We describe how a recent technological development in image-based microscopy has resulted in a change of emphasis from prevalence alone to the bivariate spatial distribution of prevalence and the intensity of infection among infected individuals. We discuss how statistical modelling of the kind described here can contribute to health policy guidelines and decision-making in two ways. One is to ensure that, in a resource-limited setting, prevalence surveys are designed, and the resulting data analysed, as efficiently as possible. The other is to provide an honest quantification of the uncertainty attached to any binary decision by reporting predictive probabilities that a policy-defined condition for action is or is not met.

Key words and phrases: Neglected tropical diseases, predictive disease mapping, spatial statistics.

REFERENCES

AMAZIGO, U. (2008). The African programme for onchocerciasis control (APOC). *Annals of Tropical Medicine and Parasitology* **102** 19–22.

AMOAH, B., GIORGI, E. and DIGGLE, P. J. (2018). A geostatistical framework for combining spatially referenced disease prevalence data from multiple diagnostics. *Biometrics*. https://doi.org/10.1111/biom.13142

BOUSSINESQ, M., GARDON, J., GARDON-WENDEL, N., KAMGNO, J., NGOUMOU, P. and CHIPPAUX, J. P. (1998). Three probable cases of Loa loa encephalopathy following Ivermectin treatment for Onchocerciasis. *American Journal of Tropical Medicine and Hygiene* **58** 461–469.

Peter J. Diggle is Distinguished University Professor, CHICAS, Lancaster University, Lancaster, LA1 4YF, UK and PhD Programme Director, Health Data Research UK, London, UK (e-mail: p.diggle@lancaster.ac.uk). Emanuele Giorgi is Lecturer, CHICAS, Lancaster University, Lancaster, LA1 4YF, UK. Julienne Atsame is Director, Control Program of Parasitic Diseases, Ministry of Health, Libreville, Gabon. Sylvie Ntsame Ella is Laboratory Technician, Control Program of Parasitic Diseases, Ministry of Health, Libreville, Gabon. Kisito Ogoussan is Implementation Management Lead, USAID's Act to End Neglected Tropical Diseases West, FHI360, Washington, DC, USA. Katherine Gass is Director of Research, Task Force for Global Health, Decatur, Georgia, USA.

BOUSSINESQ, M., GARDON, J., KAMGNO, J., PION, S. D. S., GARDON-WENDEL, N. and CHIPPAUX, J. P. (2001). Relationships between the prevalence and intensity of Loa loa infection in the Central Province of Cameroon. *Annals of Tropical Medicine and Parasitology* **95** 495–507.

BOUSSINESQ, M., GARDON, J., GARDON-WENDEL, N. and CHIPPAUX, J. (2003). Clinical picture, epidemiology and outcome of Loa-associated serious adverse events related to mass ivermectin treatment of onchocerciasis in Cameroon. *Filaria Journal* **2** 1–13.

BRESLOW, N. E. and CLAYTON, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88** 9–25.

CHIPETA, M. G., TERLOUW, D. J., PHIRI, K. S. and DIGGLE, P. J. (2016). Adaptive geostatistical design and analysis for prevalence surveys. *Spat. Stat.* **15** 70–84. MR3457669 https://doi.org/10.1016/j.spasta.2015.12.004

D'AMBROSIO, M. V., BAKALAR, M., BENNURU, S., REBER, C., SKANDARAJAH, A., NILSSON, L., SWITZ, N., KAMGNO, J., PION, S. et al. (2015). Point-of-care quantification of blood-borne filarial parasites with a mobile phone microscope. *Science Translational Medicine* **7** 286re4. https://doi.org/10.1126/scitranslmed.aaa3480

DIGGLE, P. J., TAWN, J. A. and MOYEED, R. A. (1998). Model-based geostatistics. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **47** 299–350. With discussion and a reply by the authors. MR1626544 https://doi.org/10.1111/1467-9876.00113

DIGGLE, P. J., THOMSON, M. C., CHRISTENSEN, O. F., ROWLINGSON, B., OBSOMER, V., GARDON, J., WANJI, S., TAKOUGANG, I., ENYONG, P. et al. (2007). Spatial modelling and prediction of Loa loa risk: Decision making under uncertainty. *Annals of Tropical Medicine and Parasitology* **101** 499–509.

CENTERS FOR DISEASE CONTROL AND PREVENTION (1993). Recommendations of the international task force for disease eradication. *MMWR Recommendations and Reports* **42** 1–38.

GARDON, J., GARDON-WENDEL, N., DENMARGA-NGANGUE, KAMGNO, J., CHIPPAUX, J. P. and BOUSSINESQ, M. (1997). Serious reactions after mass treatment of onchocerciasis with ivermectin in an area endemic for Loa loa infection. *Lancet* **350** 18–22.

GIORGI, E., SCHLÜTER, D. K. and DIGGLE, P. J. (2018). Bivariate geostatistical modelling of the relationship between *Loa loa* prevalence and intensity of infection. *Environmetrics* **29** e2447, 10. MR3830867 https://doi.org/10.1002/env.2447

HOMEIDA, M., BRAIDE, E., ELHASSAN, E., AMAZIGO, U. V., LIESE, B., BENTON, B., NOMA, M., ETYA'ALÉ, D., DADZIE, K. Y. et al. (2002). APOC's strategy of community-directed treatment with ivermectin (CDTI) and its potential for providing additional health services to the poorest populations. African Programme for Onchocerciasis Control. *Annals of Tropical Medicine and Parasitology* **96** S93–104.

KAMGNO, J., PION, S. D., CHESNAIS, C. B., BAKALAR, M. H., D'AMBROSIO, M. V., MACKENZIE, C. D., NANA-DJEUNGA, H. C., GOUNOUE-KAMKUMO, R., NJITCHOUANG, G.-R. et al. (2017). A test-and-not-treat strategy for Onchocerciasis in Loa loa-endemic areas. *N. Engl. J. Med.* **377** 2044–2052.

SCHLÜTER, D. K., NDEFFO-MBAH, M. L., TAKOUGANG, I., UKETY, T., WANJI, S., GALVANI, A. P. and DIGGLE, P. J. (2016). Using community-level prevalence of Loa loa infection to predict the proportion of highly-infected individuals: Statistical modelling to support lymphatic filariasis elimination programs. *PLoS Neglected Tropical Diseases* **10** 12, e0005157. https://doi.org/10.1371/journal.pntd.0005157

TAKOUGANG, I., MEREMIKWU, M., WANJI, S., YENSHU, E. V., ARIPKO, B., LAMLENN, S. B., EKA, B. L., ENYONG, P., MELI, J. et al. (2002). Rapid assessment method for prevalence and intensity of L. loa infection. *Bulletin of the World Health Organisation* **80** 852–858.

THIELE, E. A., CAMA, V. A., LAKWO, T., MEKASHA, S., ABANYIE, F., SLESHI, M., KEBEDE, A. and CANTEY, P. T. (2016). Detection of Onchocerca volvulus in skin snips by microscopy and real-time polymerase chain reaction: Implications for monitoring and evaluation activities. *American Journal of Tropical Medicine and Hygiene* **94** 906–911.

THOMSON, M. C., OBSOMER, V., DUNNE, M., CONNOR, S. J. and MOLYNEUX, D. H. (2000). Satellite mapping of Loa loa prevalence in relation to ivermectin use in west and central Africa. *Lancet* **356** 1077–1078.

THOMSON, M. C., OBSOMER, V., KAMGNO, J., GARDON, J., WANJI, S., TAKOUGANG, I., ENYONG, P., REMME, J. H., MOLYNEUX, D. H. et al. (2004). Mapping the distribution of Loa loa in Cameroon in support of the African Programme for Onchocerciasis Control. *Filaria J.* **3** 7. https://doi.org/10.1186/1475-2883-3-7

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

WANJI, S., AKOTSHI, D. O., MUTRO, M. N., TEPAGE, F., UKETY, T., DIGGLE, P. J., and REMME, J. H. (2012). The validation of the rapid assessment procedures for loiasis (RAPLOA) in the Democratic Republic of Congo: Health policy implications. *Parasites and Vectors* **5** 25. https://doi.org/10.1186/1756-3305-5-25

WORLD HEALTH ORGANISATION (2012). *Accelerating Work to Overcome the Global Impact of Neglected Tropical Diseases—a Roadmap for Implementation*. World Health Organization, Geneva.

WORLD HEALTH ORGANISATION (2013). Progress towards eliminating onchocerciasis in the WHO region of the Americas: Verification by WHO of elimination of transmission in Colombia. *Weekly Epidemiological Record* **88** 381–385.

WORLD HEALTH ORGANISATION (2014). Elimination of onchocerciasis in the WHO region of the Americas: Ecuador's progress towards verification of elimination. *Weekly Epidemiological Record* **89** 401–405.

WORLD HEALTH ORGANISATION (2015). Progress toward eliminating onchocerciasis in the WHO region of the Americas: Verification of elimination of transmission in Mexico. *Weekly Epidemiological Record* **90** 577–581.

WORLD HEALTH ORGANISATION (2018). Weekly epidemiological report. *Weekly Epidemiological Record* **93** 633–648.

ZOURE, H., WANJI, S., NOMA, M., AMAZIGO, U., DIGGLE, P. J., TEKLE, A. and REMME, J. H. (2011). The geographic distribution of Loa loa in Africa: Results of large-scale implementation of the Rapid Assessment Procedure for Loiasis (RAPLOA). *Public Library of Science*: *Neglected Tropical Diseases* **5** e1210. https://doi.org/10.1371/journal.pntd.0001210

Statistical Science
2020, Vol. 35, No. 1, 51–74
https://doi.org/10.1214/19-STS745

Quantum Science and Quantum Technology

Yazhen Wang and Xinyu Song

Abstract. Quantum science and quantum technology are of great current interest in multiple frontiers of many scientific fields ranging from computer science to physics and chemistry, and from engineering to mathematics and statistics. Their developments will likely lead to a new wave of scientific revolutions and technological innovations in a wide range of scientific studies and applications. This paper provides a brief review on quantum communication, quantum information, quantum computation, quantum simulation, and quantum metrology. We present essential quantum properties, illustrate relevant concepts of quantum science and quantum technology, and discuss their scientific developments. We point out the need for statistical analysis in their developments, as well as their potential applications to and impacts on statistics and data science.

Key words and phrases: Quantum communication, quantum information, quantum computation, quantum simulation, quantum annealing, quantum sensing and quantum metrology, quantum bit (qubit).

REFERENCES

AARONSON, S. and CHEN, L. (2017). Complexity-theoretic foundations of quantum supremacy experiments. In 32*nd Computational Complexity Conference*. *LIPIcs. Leibniz Int. Proc. Inform.* **79** Art. No. 22, 67. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern. MR3691147

ABRAMS, D. S. and LLOYD, S. (1997). Simulation of many-body Fermi systems on a universal quantum computer. *Phys. Rev. Lett.* **79** 2586.

ACÍN, A. and MASANES, L. (2016). Certified randomness in quantum physics. *Nature* **540** 213.

ADACHI, S. H. and HENDERSON, M. P. (2015). Application of quantum annealing to training of deep neural networks. Preprint. Available at arXiv:1510.06356.

AHARONOV, D. and TA-SHMA, A. (2003). Adiabatic quantum state generation and statistical zero knowledge. In *Proceedings of the Thirty-Fifth Annual ACM Symposium on Theory of Computing* 20–29. ACM, New York. MR2121066 https://doi.org/10.1145/780542.780546

AHARONOV, D., VAN DAM, W., KEMPE, J., LANDAU, Z., LLOYD, S. and REGEV, O. (2008). Adiabatic quantum computation is equivalent to standard quantum computation. *SIAM Rev.* **50** 755–787. MR2460803 https://doi.org/10.1137/080734479

ALBASH, T. and LIDAR, D. A. (2018). Adiabatic quantum computation. *Rev. Modern Phys.* **90** 015002, 64. MR3788424 https://doi.org/10.1103/RevModPhys.90.015002

ALBASH, T., RØNNOW, T. F., TROYER, M. and LIDAR, D. A. (2015). Reexamining classical and quantum models for the d-wave one processor. *The European Physical Journal Special Topics* **224** 111–129.

AMIN, M. H., ANDRIYASH, E., ROLFE, J., KULCHYTSKYY, B. and MELKO, R. (2018). Quantum Boltzmann machine. *Phys. Rev. X* **8** 021050.

ARODZ, T. and SAEEDI, S. (2019). Quantum sparse support vector machines. Available at arXiv:1902.01879v2.

ARTILES, L. M., GILL, R. D. and GUŢĂ, M. I. (2005). An invitation to quantum tomography. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 109–134. MR2136642 https://doi.org/10.1111/j.1467-9868.2005.00491.x

ARUNACHALAM, S. and DE WOLF, R. (2018). Optimal quantum sample complexity of learning algorithms. *J. Mach. Learn. Res.* **19** Paper No. 71, 36. MR3899773

ARUTE, F., ARYA, K., BABBUSH, R., BACON, D., BARDIN, J. C., BARENDS, R., BISWAS, R., BOIXO, S., BRANDAO, F. G. S. L. et al. (2019). Quantum supremacy using a programmable superconducting processor. *Nature* **574** 505–510.

ASPURU-GUZIK, A. and WALTHER, P. (2012). Photonic quantum simulators. *Nature Physics* **8** 285.

ASPURU-GUZIK, A., DUTOI, A. D., LOVE, P. J. and HEAD-GORDON, M. (2005). Simulated quantum computation of molecular energies. *Science* **309** 1704–1707.

BENEDETTI, M., REALPE-GÓMEZ, J., BISWAS, R. and PERDOMO-ORTIZ, A. (2016). Estimation of effective temperatures in quantum annealers for sampling applications: A case study with possible applications in deep learning. *Phys. Rev. A* **94** 022308.

BENNETT, C. H. and BRASSARD, G. (2014). Quantum cryptography: Public key distribution and coin tossing. *Theoret. Comput. Sci.* **560** 7–11. MR3283256 https://doi.org/10.1016/j.tcs.2014.05.025

BERNSTEIN, D. J. and LANGE, T. (2017). Post-quantum cryptography-dealing with the fallout of physics success. *IACR Cryptology ePrint Archive* **2017** 314.

Yazhen Wang is Professor, Department of Statistics, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA (e-mail: yzwang@stat.wisc.edu). Xinyu Song is Assistant Professor, School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, China (e-mail: song.xinyu@mail.shufe.edu.cn).

BERTSIMAS, D. and TSITSIKLIS, J. (1993). Simulated annealing. *Statist. Sci.* **8** 10–15.

BIAMONTE, J., WITTEK, P., PANCOTTI, N., REBENTROST, P., WIEBE, N. and LLOYD, S. (2017). Quantum machine learning. *Nature* **549** 195–202. https://doi.org/10.1038/nature23474

BLATT, R. and ROOS, C. F. (2012). Quantum simulations with trapped ions. *Nature Physics* **8** 277.

BLOCH, I., DALIBARD, J. and NASCIMBENE, S. (2012). Quantum simulations with ultracold quantum gases. *Nature Physics* **8** 267.

BOGHOSIAN, B. M. and TAYLOR, W. IV (1998). Simulating quantum mechanics on a quantum computer *Phys. D, Nonlinear Phenom.* **120** 30–42. MR1679863 https://doi.org/10.1016/S0167-2789(98)00042-6

BOIXO, S., RØNNOW, T. F., ISAKOV, S. V., WANG, Z., WECKER, D., LIDAR, D. A., MARTINIS, J. M. and TROYER, M. (2014). Evidence for quantum annealing with more than one hundred qubits. *Nature Physics* **10** 218.

BOIXO, S., SMELYANSKIY, V. N., SHABANI, A., ISAKOV, S. V., DYKMAN, M., DENCHEV, V. S., AMIN, M. H., SMIRNOV, A. Y., MOHSENI, M. et al. (2016). Computational multiqubit tunnelling in programmable quantum annealers. *Nat. Commun.* **7** 10327. https://doi.org/10.1038/ncomms10327

BOIXO, S., ISAKOV, S. V., SMELYANSKIY, V. N., BABBUSH, R., DING, N., JIANG, Z., BREMNER, M. J., MARTINIS, J. M. and NEVEN, H. (2018). Characterizing quantum supremacy in near-term devices. *Nature Physics* **14** 595.

BOULAND, A., FEFFERMAN, B., NIRKHE, C. and VAZIRANI, U. (2018). Quantum supremacy and the complexity of random circuit sampling. Preprint. Available at arXiv:1803.04402.

BRADY, L. T. and VAN DAM, W. (2016). Quantum Monte Carlo simulations of tunneling in quantum adiabatic optimization. *Phys. Rev. A* **93** 032304.

BRANDÃO, F. G. S. L., KALEV, A., LI, T., LIN, C. Y.-Y., SVORE, K. M. and WU, X. (2019). Quantum SDP solvers: Large speed-ups, optimality, and applications to quantum learning. In 46*th International Colloquium on Automata, Languages, and Programming. LIPIcs. Leibniz Int. Proc. Inform.* **132** Art. No. 27, 14. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern. MR3984844

BRAVYI, S., GOSSET, D. and KÖNIG, R. (2018). Quantum advantage with shallow circuits. *Science* **362** 308–311. MR3839777 https://doi.org/10.1126/science.aar3106

BROOKE, J., BITKO, D. and AEPPLI, G. (1999). Quantum annealing of a disordered magnet. *Science* **284** 779–781.

BROWNE, D. (2014). Quantum computation: Model versus machine. *Nature Physics* **10** 179.

BUHRMAN, H., CLEVE, R., MASSAR, S. and DE WOLF, R. (2010). Nonlocality and communication complexity. *Rev. Modern Phys.* **82** 665.

CAI, T., KIM, D., WANG, Y., YUAN, M. and ZHOU, H. H. (2016). Optimal large-scale quantum state tomography with Pauli measurements. *Ann. Statist.* **44** 682–712. MR3476614 https://doi.org/10.1214/15-AOS1382

CAMPBELL, E. T., TERHAL, B. M. and VUILLOT, C. (2017). Roads towards fault-tolerant universal quantum computation. *Nature* **549** 172–179. https://doi.org/10.1038/nature23460

CHILDS, A. M., MASLOV, D., NAM, Y., ROSS, N. J. and SU, Y. (2018). Toward the first quantum simulation with quantum speedup. *Proc. Natl. Acad. Sci. USA* **115** 9456–9461. MR3859035 https://doi.org/10.1073/pnas.1801723115

CHONG, F. T., FRANKLIN, D. and MARTONOSI, M. (2017). Programming languages and compiler design for realistic quantum hardware. *Nature* **549** 180–187. https://doi.org/10.1038/nature23459

CHOWDHURY, A. N. and SOMMA, R. D. (2017). Quantum algorithms for Gibbs sampling and hitting-time estimation. *Quantum Inf. Comput.* **17** 41–64. MR3676655

CILIBERTO, C., HERBSTER, M., IALONGO, A. D., PONTIL, M., ROCCHETTO, A., SEVERINI, S. and WOSSNIG, L. (2018). Quantum machine learning: A classical perspective. *Proc. A.* **474** 20170551, 26. MR3762887 https://doi.org/10.1098/rspa.2017.0551

CIRAC, J. I. and ZOLLER, P. (2012). Goals and opportunities in quantum simulation. *Nature Physics* **8** 264.

CROSS, A. W., SMITH, G. and SMOLIN, J. A. (2015). Quantum learning robust against noise. *Phys. Rev. A* **92** 012327.

CROSSON, E. and HARROW, A. W. (2016). Simulated quantum annealing can be exponentially faster than classical simulated annealing. In 57*th Annual IEEE Symposium on Foundations of Computer Science—FOCS* 2016 714–723. IEEE Computer Soc., Los Alamitos, CA. MR3631034

DAS, A. and CHAKRABARTI, B. K. (2005). *Quantum Annealing and Related Optimization Methods* **679**. Springer, Berlin.

DAS, A. and CHAKRABARTI, B. K. (2008). Colloquium: Quantum annealing and analog quantum computation. *Rev. Modern Phys.* **80** 1061–1081. MR2443721 https://doi.org/10.1103/RevModPhys.80.1061

DEGEN, C. L., REINHARD, F. and CAPPELLARO, P. (2017). Quantum sensing. *Rev. Modern Phys.* **89** 035002, 39. MR3713686 https://doi.org/10.1103/RevModPhys.89.035002

DENCHEV, V. S., BOIXO, S., ISAKOV, S. V., DING, N., BABBUSH, R., SMELYANSKIY, V., MARTINIS, J. and NEVEN, H. (2016). What is the computational value of finite-range tunneling? *Phys. Rev. X* **6** 031015.

DEUTSCH, D. (1985). Quantum theory, the Church–Turing principle and the universal quantum computer. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **400** 97–117. MR0801665

DICARLO, L., CHOW, J. M., GAMBETTA, J. M., BISHOP, L. S., JOHNSON, B. R., SCHUSTER, D. I., MAJER, J., BLAIS, A., FRUNZIO, L. et al. (2009). Demonstration of two-qubit algorithms with a superconducting quantum processor. *Nature* **460** 240–244. https://doi.org/10.1038/nature08121

DIVINCENZO, D. P. (1995). Quantum computation. *Science* **270** 255–261. MR1355956 https://doi.org/10.1126/science.270.5234.255

DUNJKO, V. and BRIEGEL, H. J. (2018). Machine learning & artificial intelligence in the quantum domain: A review of recent progress. *Rep. Progr. Phys.* **81** 074001, 67. MR3827116 https://doi.org/10.1088/1361-6633/aab406

DUNJKO, V., TAYLOR, J. M. and BRIEGEL, H. J. (2016). Quantum-enhanced machine learning. *Phys. Rev. Lett.* **117** 130501, 6. MR3636529 https://doi.org/10.1103/PhysRevLett.117.130501

FARHI, E., GOLDSTONE, J. and GUTMANN, S. (2002). Quantum adiabatic evolution algorithms versus simulated annealing. Preprint. Available at quant-ph/0201031.

FARHI, E., GOLDSTONE, J., GUTMANN, S. and SIPSER, M. (1998). Limit on the speed of quantum computation in determining parity. *Phys. Rev. Lett.* **81** 5442–5444.

FARHI, E., GOLDSTONE, J., GUTMANN, S. and SIPSER, M. (2000). Quantum computation by adiabatic evolution. Preprint. Available at quant-ph/0001106.

FARHI, E., GOLDSTONE, J., GUTMANN, S., LAPAN, J., LUNDGREN, A. and PREDA, D. (2001). A quantum adiabatic evolution algorithm applied to random instances of an NP-complete problem. *Science* **292** 472–476. MR1838761 https://doi.org/10.1126/science.1057726

FEYNMAN, R. P. (1981/82). Simulating physics with computers. *Internat. J. Theoret. Phys.* **21** 467–488. MR0658311 https://doi.org/10.1007/BF02650179

GRANADE, C. E., FERRIE, C., WIEBE, N. and CORY, D. G. (2012). Robust online Hamiltonian learning. *New J. Phys*. **14** 103013, 31. MR3036977 https://doi.org/10.1088/1367-2630/14/10/103013

GRILO, A. B., KERENIDIS, I. and ZIJLSTRA, T. (2018). Learning with errors is easy with quantum samples. Available at arXiv:1702.08255v2.

GROVER, L. K. (1997). Quantum mechanics helps in searching for a needle in a haystack. *Phys. Rev. Lett*. **79** 325.

HARROW, A. W. and MONTANARO, A. (2017). Quantum computational supremacy. *Nature* **549** 203–209. https://doi.org/10.1038/nature23458

HAVLÍČEK, V., CÓRCOLES, A. D., TEMME, K., HARROW, A. W., KANDALA, A., CHOW, J. M. and GAMBETTA, J. M. (2019). Supervised learning with quantum-enhanced feature spaces. *Nature* **567** 209–212.

HAYASHI, M. (2006). *Quantum Information*. Springer, Berlin. MR2228302

HOLEVO, A. S. (1998). The capacity of the quantum channel with general signal states. *IEEE Trans. Inform. Theory* **44** 269–273. MR1486663 https://doi.org/10.1109/18.651037

HORODECKI, R., HORODECKI, P., HORODECKI, M. and HORODECKI, K. (2009). Quantum entanglement. *Rev. Modern Phys*. **81** 865–942. MR2515619 https://doi.org/10.1103/RevModPhys.81.865

HOUCK, A. A., TÜRECI, H. E. and KOCH, J. (2012). On-chip quantum simulation with superconducting circuits. *Nature Physics* **8** 292.

ISAKOV, S. V., MAZZOLA, G., SMELYANSKIY, V. N., JIANG, Z., BOIXO, S., NEVEN, H. and TROYER, M. (2016). Understanding quantum tunneling through quantum Monte-Carlo simulations. *Phys. Rev. Lett*. **117** 180402. https://doi.org/10.1103/PhysRevLett.117.180402

JANÉ, E., VIDAL, G., DÜR, W., ZOLLER, P. and CIRAC, J. I. (2003). Simulation of quantum dynamics with quantum optical systems. *Quantum Inf. Comput*. **3** 15–37. MR1965173

JIANG, Z., SMELYANSKIY, V. N., ISAKOV, S. V., BOIXO, S., MAZZOLA, G., TROYER, M. and NEVEN, H. (2017). Scaling analysis and instantons for thermally assisted tunneling and quantum Monte Carlo simulations. *Phys. Rev. A* **95** 012322.

JOHNSON, M. W., AMIN, M. H. S., GILDERT, S., LANTING, T., HAMZE, F., DICKSON, N., HARRIS, R., BERKLEY, A. J., JOHANSSON, J. et al. (2011). Quantum annealing with manufactured spins. *Nature* **473** 194–198. https://doi.org/10.1038/nature10012

JORDAN, S. P. (2005). Fast quantum algorithm for numerical gradient estimation. *Phys. Rev. Lett*. **95** 050501. https://doi.org/10.1103/PhysRevLett.95.050501

JÖRG, T., KRZAKALA, F., KURCHAN, J. and MAGGS, A. C. (2010). Quantum annealing of hard problems. *Progr. Theoret. Phys. Suppl*. **184** 290–303.

KADOWAKI, T. and NISHIMORI, H. (1998). Quantum annealing in the transverse Ising model. *Phys. Rev. E* (3) **58** 5355.

KASSAL, I., JORDAN, S. P., LOVE, P. J., MOHSENI, M. and ASPURU-GUZIK, A. (2008). Polynomial-time quantum algorithm for the simulation of chemical dynamics. *Proc. Natl. Acad. Sci. USA* pnas–0808245105.

KASSAL, I., WHITFIELD, J. D., PERDOMO-ORTIZ, A., YUNG, M.-H. and ASPURU-GUZIK, A. (2011). Simulating chemistry using quantum computers. *Annu Rev Phys Chem* **62** 185–207. https://doi.org/10.1146/annurev-physchem-032210-103512

KIEFEROVA, M. and WIEBE, N. (2016). Tomography and generative data modeling via quantum boltzmann training. Preprint. Available at arXiv:1612.05204.

KIRKPATRICK, S., GELATT, C. D. JR. and VECCHI, M. P. (1983). Optimization by simulated annealing. *Science* **220** 671–680. MR0702485 https://doi.org/10.1126/science.220.4598.671

KRENN, M., MALIK, M., SCHEIDL, T., URSIN, R. and ZEILINGER, A. (2017). Quantum communication with photons. Preprint. Available at arXiv:1701.00989.

KRUSE, I., LANGE, K., PEISE, J., LÜCKE, B., PEZZÈ, L., ARLT, J., ERTMER, W., LISDAT, C., SANTOS, L. et al. (2016). Improvement of an atomic clock using squeezed vacuum. *Phys. Rev. Lett*. **117** 143004. https://doi.org/10.1103/PhysRevLett.117.143004

LANYON, B. P., WHITFIELD, J. D., GILLETT, G. G., GOGGIN, M. E., ALMEIDA, M. P., KASSAL, I., BIAMONTE, J. D., MOHSENI, M., POWELL, B. J. et al. (2010). Towards quantum chemistry on a quantum computer. *Nat Chem* **2** 106–111. https://doi.org/10.1038/nchem.483

LLOYD, S. (1996). Universal quantum simulators. *Science* **273** 1073–1078. MR1407944 https://doi.org/10.1126/science.273.5278.1073

LLOYD, S., MOHSENI, M. and REBENTROST, P. (2014). Quantum principal component analysis. *Nature Physics* **10** 631.

LUND, A., BREMNER, M. J. and RALPH, T. (2017). Quantum sampling problems, bosonsampling and quantum supremacy. *Npj Quantum Information* **3** 15.

MALLEY, J. D. and HORNSTEIN, J. (1993). Quantum statistical inference. *Statist. Sci*. **8** 433–457. MR1250150

MARIANTONI, M., WANG, H., YAMAMOTO, T., NEELEY, M., BIALCZAK, R. C., CHEN, Y., LENANDER, M., LUCERO, E., OACONNELL, A. D. et al. (2011). Implementing the quantum von Neumann architecture with superconducting circuits. *Science* 1208517.

MARKOV, I. L., FATIMA, A., ISAKOV, S. V. and BOIXO, S. (2018). Quantum supremacy is both closer and farther than it appears. Preprint. Available at arXiv:1807.10749.

MARVIAN, I. and LLOYD, S. (2016). Universal quantum emulator. Preprint. Available at arXiv:1606.02734.

MCGEOCH, C. C. (2014). Adiabatic quantum computation and quantum annealing: Theory and practice. *Synthesis Lectures on Quantum Computing* **5** 1–93.

MOHSENI, M., READ, P., NEVEN, H., BOIXO, S., DENCHEV, V., BABBUSH, R., FOWLER, A., SMELYANSKIY, V. and MARTINIS, J. (2017). Commercialize quantum technologies in five years. *Nature News* **543** 171.

MONTANARO, A. (2016). Quantum algorithms: An overview. *Npj Quantum Information* **2** 15023.

NEILL, C., ROUSHAN, P., KECHEDZHI, K. et al. (2018). A blueprint for demonstrating quantum supremacy with superconducting qubits. *Science* **360** 195–199. MR3792641 https://doi.org/10.1126/science.aao4309

NIELSEN, M. A. and CHUANG, I. L. (2010). *Quantum Computation and Quantum Information*. Cambridge Univ. Press, Cambridge. MR1796805

O'GORMAN, B., BABBUSH, R., PERDOMO-ORTIZ, A., ASPURU-GUZIK, A. and SMELYANSKIY, V. (2015). Bayesian network structure learning using quantum annealing. *The European Physical Journal Special Topics* **224** 163–188.

PAESANI, S., GENTILE, A. A., SANTAGATI, R., WANG, J., WIEBE, N., TEW, D. P., O'BRIEN, J. L. and THOMPSON, M. G. (2017). Experimental Bayesian quantum phase estimation on a silicon photonic chip. *Phys. Rev. Lett*. **118** 100503. https://doi.org/10.1103/PhysRevLett.118.100503

PEZZÈ, L., SMERZI, A., OBERTHALER, M. K., SCHMIED, R. and TREUTLEIN, P. (2018). Quantum metrology with nonclassical states of atomic ensembles. *Rev. Modern Phys*. **90** 035005, 70. MR3861238 https://doi.org/10.1103/RevModPhys.90.035005

REBENTROST, P., MOHSENI, M. and LLOYD, S. (2014). Quantum support vector machine for big data classification. *Phys. Rev. Lett*. **113** 130503. https://doi.org/10.1103/PhysRevLett.113.130503

RICHTER, P. C. (2006). Quantum speedup of classical mixing processes. *Phys. Rev. A* **76** 042306.

RIVEST, R. L., SHAMIR, A. and ADLEMAN, L. (1978). A method for obtaining digital signatures and public-key cryptosystems. *Commun. ACM* **21** 120–126. MR0700103 https://doi.org/10.1145/359340.359342

RØNNOW, T. F., WANG, Z., JOB, J., BOIXO, S., ISAKOV, S. V., WECKER, D., MARTINIS, J. M., LIDAR, D. A. and TROYER, M. (2014). Defining and detecting quantum speedup. *Science* **345** 420–424.

SAKURAI, J. and NAPOLITANO, J. (2017). *Modern Quantum Mechanics*. *Modern Quantum Mechanics*, *by JJ Sakurai*, *Jim Napolitano*. Cambridge Univ. Press, Cambridge.

SALAKHUTDINOV, R. and HINTON, G. (2009). Deep Boltzmann machines. In *Artificial Intelligence and Statistics* 448–455.

SANGOUARD, N., SIMON, C., DE RIEDMATTEN, H. and GISIN, N. (2011). Quantum repeaters based on atomic ensembles and linear optics. *Rev. Modern Phys.* **83** 33.

SAYRIN, C., DOTSENKO, I., ZHOU, X., PEAUDECERF, B., RYBARCZYK, T., GLEYZES, S., ROUCHON, P., MIRRAHIMI, M., AMINI, H. et al. (2011). Real-time quantum feedback prepares and stabilizes photon number states. *Nature* **477** 73–77. https://doi.org/10.1038/nature10376

SCHUMACHER, B. (1995). Quantum coding. *Phys. Rev. A* (3) **51** 2738–2747. MR1328824 https://doi.org/10.1103/PhysRevA.51.2738

SCHUMACHER, B. and WESTMORELAND, M. D. (1997). Sending classical information via noisy quantum channels. *Phys. Rev. A* **56** 131.

SHANKAR, R. (2012). *Principles of Quantum Mechanics*. Springer, New York.

SHENVI, N., KEMPE, J. and WHALEY, K. B. (2003). Quantum random-walk search algorithm. *Phys. Rev. A* **67** 052307.

SHOR, P. W. (1994). Algorithms for quantum computation: Discrete logarithms and factoring. In 35*th Annual Symposium on Foundations of Computer Science* (*Santa Fe*, *NM*, 1994) 124–134. IEEE Comput. Soc. Press, Los Alamitos, CA. MR1489242 https://doi.org/10.1109/SFCS.1994.365700

SVORE, K. M., HASTINGS, M. B. and FREEDMAN, M. (2014). Faster phase estimation. *Quantum Inf. Comput.* **14** 306–328. MR3186297

SZEGEDY, M. (2004). Quantum speed-up of Markov chain based algorithms. In *Proceedings-Annual IEEE Symposium on Foundations of Computer Science*, *FOCS* 32–41.

TEMME, K., OSBORNE, T. J., VOLLBRECHT, K. G., POULIN, D. and VERSTRAETE, F. (2011). Quantum Metropolis sampling. *Nature* **471** 87–90. https://doi.org/10.1038/nature09770

WANG, Y. (2011). Quantum Monte Carlo simulation. *Ann. Appl. Stat.* **5** 669–683. MR2840170 https://doi.org/10.1214/10-AOAS406

WANG, Y. (2012). Quantum computation and quantum information. *Statist. Sci.* **27** 373–394. MR3012432 https://doi.org/10.1214/11-STS378

WANG, Y. (2013). Asymptotic equivalence of quantum state tomography and noisy matrix completion. *Ann. Statist.* **41** 2462–2504. MR3127872 https://doi.org/10.1214/13-AOS1156

WANG, Y., WU, S. and ZOU, J. (2016). Quantum annealing with Markov chain Monte Carlo simulations and D-wave quantum computers. *Statist. Sci.* **31** 362–398. MR3552740 https://doi.org/10.1214/16-STS560

WANG, Y. and XU, C. (2015). Density matrix estimation in quantum homodyne tomography. *Statist. Sinica* **25** 953–973. MR3409732

WIEBE, N. and GRANADE, C. (2016). Efficient Bayesian phase estimation. *Phys. Rev. Lett.* **117** 010503. https://doi.org/10.1103/PhysRevLett.117.010503

WIEBE, N., GRANADE, C. and CORY, D. G. (2015). Quantum bootstrapping via compressed quantum Hamiltonian learning. *New J. Phys.* **17** 022005.

WIEBE, N., KAPOOR, A. and SVORE, K. M. (2015). Quantum algorithms for nearest-neighbor methods for supervised and unsupervised learning. *Quantum Inf. Comput.* **15** 316–356. MR3328494

WIEBE, N., KAPOOR, A. and SVORE, K. M. (2016). Quantum deep learning. *Quantum Inf. Comput.* **16** 541–587. MR3559656

WIEBE, N., GRANADE, C., FERRIE, C. and CORY, D. G. (2014). Hamiltonian learning and certification using quantum resources. *Phys. Rev. Lett.* **112** 190501.

WITTEK, P. (2014). *Quantum Machine Learning*: *What Quantum Computing Means to Data Mining*. Academic Press, San Diego.

YIN, J., CAO, Y., LI, Y.-H., LIAO, S.-K., ZHANG, L., REN, J.-G., CAI, W.-Q., LIU, W.-Y., LI, B. et al. (2017). Satellite-based entanglement distribution over 1200 kilometers. *Science* **356** 1140–1144. https://doi.org/10.1126/science.aan3211

Statistical Science
2020, Vol. 35, No. 1, 75–91
https://doi.org/10.1214/19-STS752

Statistical Methodology in Single-Molecule Experiments

Chao Du and S. C. Kou

Abstract. Toward the last quarter of the 20th century, the emergence of single-molecule experiments enabled scientists to track and study individual molecules' dynamic properties in real time. Unlike macroscopic systems' dynamics, those of single molecules can only be properly described by stochastic models even in the absence of external noise. Consequently, statistical methods have played a key role in extracting hidden information about molecular dynamics from data obtained through single-molecule experiments. In this article, we survey the major statistical methodologies used to analyze single-molecule experimental data. Our discussion is organized according to the types of stochastic models used to describe single-molecule systems as well as major experimental data collection techniques. We also highlight challenges and future directions in the application of statistical methodologies to single-molecule experiments.

Key words and phrases: Autocorrelation, continuous-time Markov chain, diffusion process, heterogeneity, hidden Markov model, molecular dynamics.

REFERENCES

ANDREC, M., LEVY, R. M. and TALAGA, D. S. (2003). Direct determination of kinetic rates from single-molecule photon arrival trajectories using hidden Markov models. *J. Phys. Chem. A* **107** 7454–7464. https://doi.org/10.1021/jp035514+

ARCIZET, D., MEIER, B., SACKMANN, E., RÄDLER, J. O. and HEINRICH, D. (2008). Temporal analysis of active and passive transport in living cells. *Phys. Rev. Lett.* **101** 248103. https://doi.org/10.1103/PhysRevLett.101.248103

BALL, F. G. and SANSOM, M. S. P. (1989). Ion-channel gating mechanisms: Model identification and parameter estimation from single channel recordings. *Proc. R. Soc. Lond., B Biol. Sci.* **236** 385–416.

BALL, F. G., CAI, Y., KADANE, J. B. and O'HAGAN, A. (1999). Bayesian inference for ion-channel gating mechanisms directly from single-channel recordings, using Markov chain Monte Carlo. *Proc. R. Soc. Lond. A Mat.* **455** 2879–2932.

BARRY, D. and HARTIGAN, J. A. (1993). A Bayesian analysis for change point problems. *J. Amer. Statist. Assoc.* **88** 309–319. MR1212493

BAUER, M., LI, C., MÜLLEN, K., BASCHÉ, T. and HINZE, G. (2018). State transition identification in multivariate time series (STIMTS) applied to rotational jump trajectories from single molecules. *J. Chem. Phys.* **149** 164104.

BEREZHKOVSKII, A. M., SZABO, A. and WEISS, G. H. (2000). Theory of the fluorescence of single molecules undergoing multistate conformational dynamics. *J. Phys. Chem. B* **104** 3776–3780.

BERGLUND, A. J. (2010). Statistics of camera-based single-particle tracking. *Phys. Rev. E* **82** 011917.

BERNSTEIN, J. and FRICKS, J. (2016). Analysis of single particle diffusion with transient binding using particle filtering. *J. Theoret. Biol.* **401** 109–121.

BLAINEY, P. C., LUO, G., KOU, S. C., MANGEL, W. F., VERDINE, G. L., BAGCHI, B. and XIE, S. X. (2009). Nonspecifically bound proteins spin while diffusing along DNA. *Nat. Struct. Mol. Biol.* **16** 1224–1229.

BLANCO, M. and WALTER, N. G. (2010). Analysis of complex single-molecule FRET time trajectories. In *Methods in Enzymology* **472** 153–178. Elsevier, Amsterdam.

BLANCO, M. R., MARTIN, J. S., KAHLSCHEUER, M. L., KRISHNAN, R., ABELSON, J., LAEDERACH, A. and WALTER, N. G. (2015). Single molecule cluster analysis dissects splicing pathway conformational dynamics. *Nat. Methods* **12** 1077–1084. https://doi.org/10.1038/nmeth.3602

BOKINSKY, G., RUEDA, D., MISRA, V. K., RHODES, M. M., GORDUS, A., BABCOCK, H. P., WALTER, N. G. and ZHUANG, X. (2003). Single-molecule transition-state analysis of RNA folding. *Proc. Natl. Acad. Sci. USA* **100** 9302–9307.

BOSCH, P. J., KANGER, J. S. and SUBRAMANIAM, V. (2014). Classification of dynamical diffusion states in single molecule tracking microscopy. *Biophys. J.* **107** 588–598. https://doi.org/10.1016/j.bpj.2014.05.049

BOYSEN, L., KEMPE, A., LIEBSCHER, V., MUNK, A. and WITTICH, O. (2009). Consistencies and rates of convergence of jump-penalized least squares estimators. *Ann. Statist.* **37** 157–183. MR2488348 https://doi.org/10.1214/07-AOS558

BRAUN, J. V., BRAUN, R. K. and MÜLLER, H.-G. (2000). Multiple changepoint fitting via quasilikelihood, with application to DNA

Chao Du is Assistant Professor in Statistics, University of Virginia, Halsey Hall, Charlottesville, Virginia 22903, USA (e-mail: cd2wb@virginia.edu). S. C. Kou is Professor of Statistics and Professor of Biostatistics, Harvard University, Science Center, 1 Oxford Street, Cambridge, Massachusetts 02138, USA (e-mail: kou@stat.harvard.edu).

sequence segmentation. *Biometrika* **87** 301–314. MR1782480 https://doi.org/10.1093/biomet/87.2.301

BROCK, R., HINK, M. A. and JOVIN, T. M. (1998). Fluorescence correlation microscopy of cells in the presence of autofluorescence. *Biophys. J.* **75** 2547–2557.

BRONSON, J. E., FEI, J., HOFMAN, J. M., GONZALEZ JR, R. L. and WIGGINS, C. H. (2009). Learning rates and states from biophysical time series: A Bayesian approach to model selection and single-molecule FRET data. *Biophys. J.* **97** 3196–3205.

CHEN, Y., FUH, C.-D., KAO, C.-L. and KOU, S. C. (2019). Determine the number of states in hidden Markov models via marginal likelihood. *Preprint*.

CHEN, Y., MÜLLER, J. D., SO, P. T. C. and GRATTON, E. (1999). The photon counting histogram in fluorescence fluctuation spectroscopy. *Biophys. J.* **77** 553–567.

CHEN, Y., SHEN, K., SHAN, S.-O. and KOU, S. C. (2016). Analyzing single-molecule protein transportation experiments via hierarchical hidden Markov models. *J. Amer. Statist. Assoc.* **111** 951–966. MR3561922 https://doi.org/10.1080/01621459.2016.1140050

CHERNOFF, H. and ZACKS, S. (1964). Estimating the current mean of a normal distribution which is subjected to changes in time. *Ann. Math. Stat.* **35** 999–1018. MR0179874 https://doi.org/10.1214/aoms/1177700517

CHIB, S. (1998). Estimation and comparison of multiple change-point models. *J. Econometrics* **86** 221–241. MR1649222 https://doi.org/10.1016/S0304-4076(97)00115-2

CHUNG, H. S. and GOPICH, I. V. (2014). Fast single-molecule FRET spectroscopy: Theory and experiment. *Phys. Chem. Chem. Phys.* **16** 18644–18657. https://doi.org/10.1039/c4cp02489c

CHUNG, S. H. and KENNEDY, R. A. (1991). Forward-backward non-linear filtering technique for extracting small biological signals from noise. *J. Neurosci. Methods* **40** 71–86. https://doi.org/10.1016/0165-0270(91)90118-j

CHUNG, H. S., LOUIS, J. M. and GOPICH, I. V. (2016). Analysis of fluorescence lifetime and energy transfer efficiency in single-molecule photon trajectories of fast-folding proteins. *J. Phys. Chem. B* **120** 680–699.

CHUNG, S.-H., MOORE, J. B., XIA, L., PREMKUMAR, L. S. and GAGE, P. W. (1990). Characterization of single channel currents using digital signal processing techniques based on hidden Markov models. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **329** 265–285.

CHUNG, I., AKITA, R., VANDLEN, R., TOOMRE, D., SCHLESSINGER, J. and MELLMAN, I. (2010). Spatial control of EGF receptor activation by reversible dimerization on living cells. *Nature* **464** 783–787.

CLAUSEN, M. P. and LAGERHOLM, C. B. (2013). Visualization of plasma membrane compartmentalization by high-speed quantum dot tracking. *Nano Lett.* **13** 2332–2337.

COLQUHOUN, D. and HAWKES, A. G. (1981). On the stochastic properties of single ion channels. *Proc. R. Soc. Lond., B Biol. Sci.* **211** 205–235.

DAS, R., CAIRO, C. W. and COOMBS, D. (2009). A hidden Markov model for single particle tracks quantifies dynamic interactions between LFA-1 and the actin cytoskeleton. *PLoS Comput. Biol.* **5** e1000556, 16. MR2577427 https://doi.org/10.1371/journal.pcbi.1000556

DE GUNST, M. C. M. and SCHOUTEN, B. (2003). Model selection for hidden Markov models of ion channel data by reversible jump Markov chain Monte Carlo. *Bernoulli* **9** 373–393. MR1997489 https://doi.org/10.3150/bj/1065444810

DEL CASTILLO, J. and KATZ, B. (1957). Interaction at end-plate receptors between different choline derivatives. *Proc. R. Soc. Lond., B Biol. Sci.* **146** 369–381.

DU, C., KAO, C.-L. M. and KOU, S. C. (2016). Stepwise signal extraction via marginal likelihood. *J. Amer. Statist. Assoc.* **111** 314–330. MR3494662 https://doi.org/10.1080/01621459.2015.1006365

DU, C. and KOU, S. C. (2012). Correlation analysis of enzymatic reaction of a single protein molecule. *Ann. Appl. Stat.* **6** 950–976. MR3012516 https://doi.org/10.1214/12-AOAS541

DU, C. and KOU, S. C. (2020). Supplement to "Statistical methodology in single-molecule experiments." https://doi.org/10.1214/19-STS752SUPP.

ELSON, E. L. (2011). Fluorescence correlation spectroscopy: Past, present, future. *Biophys. J.* **101** 2855–2870. https://doi.org/10.1016/j.bpj.2011.11.012

ELSON, E. L. and MAGDE, D. (1974). Fluorescence correlation spectroscopy. I. Conceptual basis and theory. *Biopolymers: Original Research on Biomolecules* **13** 1–27.

ENDERLEIN, J., GREGOR, I., PATRA, D. and FITTER, J. (2005). Statistical analysis of diffusion coefficient determination by fluorescence correlation spectroscopy. *J. Fluoresc.* **15** 415–422.

ENGLISH, B. P., MIN, W., VAN OIJEN, A. M., LEE, K. T., LUO, G., SUN, H., CHERAYIL, B. J., KOU, S. C. and XIE, X. S. (2006). Ever-fluctuating single enzyme molecules: Michaelis–Menten equation revisited. *Nat. Chem. Biol.* **2** 87–94.

EPSTEIN, M., CALDERHEAD, B., GIROLAMI, M. A. and SIVILOTTI, L. G. (2016). Bayesian statistical inference in ion-channel models with exact missed event correction. *Biophys. J.* **111** 333–348.

FEARNHEAD, P. and LIU, Z. (2007). On-line inference for multiple changepoint problems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 589–605. MR2370070 https://doi.org/10.1111/j.1467-9868.2007.00601.x

FLOYD, D. L., HARRISON, S. C. and VAN OIJEN, A. M. (2010). Analysis of kinetic intermediates in single-particle dwell-time distributions. *Biophys. J.* **99** 360–366.

FREDKIN, D. R. and RICE, J. A. (1992a). Bayesian restoration of single-channel patch clamp recordings. *Biometrics* 427–448.

FREDKIN, D. R. and RICE, J. A. (1992b). Maximum likelihood estimation and identification directly from single-channel recordings. *Proc. R. Soc. Lond., B Biol. Sci.* **249** 125–132.

FRICK, K., MUNK, A. and SIELING, H. (2014). Multiscale change point inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 495–580. MR3210728 https://doi.org/10.1111/rssb.12047

GASSIAT, E. and ROUSSEAU, J. (2014). About the posterior distribution in hidden Markov models with unknown number of states. *Bernoulli* **20** 2039–2075. MR3263098 https://doi.org/10.3150/13-BEJ550

GENNERICH, A. and SCHILD, D. (2000). Fluorescence correlation spectroscopy in small cytosolic compartments depends critically on the diffusion model used. *Biophys. J.* **79** 3294–3306.

GIN, E., FALCKE, M., WAGNER, L. E., YULE, D. I. and SNEYD, J. (2009). Markov chain Monte Carlo fitting of single-channel data from inositol trisphosphate receptors. *J. Theoret. Biol.* **257** 460–474.

GLOTER, A. and JACOD, J. (2001). Diffusions with measurement errors. I. Local asymptotic normality. *ESAIM Probab. Stat.* **5** 225–242. MR1875672 https://doi.org/10.1051/ps:2001110

HA, T., ENDERLE, T., OGLETREE, D. F., CHEMLA, D. S., SELVIN, P. R. and WEISS, S. (1996). Probing the interaction between two single molecules: Fluorescence resonance energy transfer between a single donor and a single acceptor. *Proc. Natl. Acad. Sci. USA* **93** 6264–6268.

HARAN, G. (2004). Noise reduction in single-molecule fluorescence trajectories of folding proteins. *Chem. Phys.* **307** 137–145.

HE, J., GUO, S.-M. and BATHE, M. (2012). Bayesian approach to the analysis of fluorescence correlation spectroscopy data I: Theory. *Anal. Chem.* **84** 3871–3879.

HINES, K. E., BANKSTON, J. R. and ALDRICH, R. W. (2015). Analyzing single-molecule time series via nonparametric Bayesian inference. *Biophys. J.* **108** 540–556.

HINKLEY, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika* **57** 1–17. MR0273727 https://doi.org/10.1093/biomet/57.1.1

HODGSON, M. E. A. and GREEN, P. J. (1999). Bayesian choice among Markov models of ion channels using Markov chain Monte Carlo. *Proc. R. Soc. Lond. A Mat.* **455** 3425–3448.

HORN, R. (1987). Statistical methods for model discrimination. Applications to gating kinetics and permeation of the acetylcholine receptor channel. *Biophys. J.* **51** 255–263.

HUET, S., KARATEKIN, E., TRAN, V. S., FANGET, I., CRIBIER, S. and HENRY, J.-P. (2006). Analysis of transient behavior in complex trajectories: Application to secretory vesicle dynamics. *Biophys. J.* **91** 3542–3559. https://doi.org/10.1529/biophysj.105.080622

JEON, J.-H. and METZLER, R. (2010). Fractional Brownian motion and motion governed by the fractional Langevin equation in confined geometries. *Phys. Rev. E* **81** 021103, 11. MR2610879 https://doi.org/10.1103/PhysRevE.81.021103

KASK, P., PALO, K., FAY, N., BRAND, L., METS, Ü., ULLMANN, D., JUNGMANN, J., PSCHORR, J. and GALL, K. (2000). Two-dimensional fluorescence intensity distribution analysis: Theory and applications. *Biophys. J.* **78** 1703–1713.

KELLER, B. G., KOBITSKI, A., JÄSCHKE, A., NIENHAUS, G. U. and NOÉ, F. (2014). Complex RNA folding kinetics revealed by single-molecule FRET and hidden Markov models. *J. Am. Chem. Soc.* **136** 4534–4543.

KEPTEN, E., BRONSHTEIN, I. and GARINI, Y. (2013). Improved estimation of anomalous diffusion exponents in single-particle tracking experiments. *Phys. Rev. E* **87** 052713.

KEPTEN, E., WERON, A., SIKORA, G., BURNECKI, K. and GARINI, Y. (2015). Guidelines for the fitting of anomalous diffusion mean square displacement graphs from single particle tracking experiments. *PLoS ONE* **10** e0117722. https://doi.org/10.1371/journal.pone.0117722

KOO, P. K. and MOCHRIE, S. G. J. (2016). Systems-level approach to uncovering diffusive states and their transitions from single-particle trajectories. *Phys. Rev. E* **94** 052412. https://doi.org/10.1103/PhysRevE.94.052412

KOO, P. K., WEITZMAN, M., SABANAYGAM, C. R., VAN GOLEN, K. L. and MOCHRIE, S. G. J. (2015). Extracting diffusive states of Rho GTPase in live cells: Towards in vivo biochemistry. *PLoS Comput. Biol.* **11** e1004297.

KOPPEL, D. E. (1974). Statistical accuracy in fluorescence correlation spectroscopy. *Phys. Rev. A* **10** 1938.

KOU, S. C. (2008a). Stochastic modeling in nanoscale biophysics: Subdiffusion within proteins. *Ann. Appl. Stat.* **2** 501–535. MR2524344 https://doi.org/10.1214/07-AOAS149

KOU, S. C. (2008b). Stochastic networks in nanoscale biophysics: Modeling enzymatic reaction of a single protein. *J. Amer. Statist. Assoc.* **103** 961–975. MR2462886 https://doi.org/10.1198/016214507000001021

KOU, S. C. and XIE, X. S. (2004). Generalized Langevin equation with fractional Gaussian noise: Subdiffusion within a single protein molecule. *Phys. Rev. Lett.* **93** 180603.

KOU, S. C., XIE, X. S. and LIU, J. S. (2005). Bayesian analysis of single-molecule experimental data. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **54** 469–506. MR2137252 https://doi.org/10.1111/j.1467-9876.2005.00509.x

KOU, S. C., CHERAYIL, B. J., MIN, W., ENGLISH, B. P. and XIE, X. S. (2005). Single-molecule Michaelis–Menten equations. *J. Phys. Chem. B* **109** 19068–19081.

KUSUMI, A., SAKO, Y. and YAMAMOTO, M. (1993). Confined lateral diffusion of membrane receptors as studied by single particle tracking (nanovid microscopy). Effects of calcium-induced differentiation in cultured epithelial cells. *Biophys. J.* **65** 2021–2040.

LIU, Y., PARK, J., DAHMEN, K. A., CHEMLA, Y. R. and HA, T. (2010). A comparative study of multivariate and univariate hidden Markov modelings in time-binned single-molecule FRET data analysis. *J. Phys. Chem. B* **114** 5386–5403.

LU, H. P., XUN, L. and XIE, X. S. (1998). Single-molecule enzymatic dynamics. *Science* **282** 1877–1882.

MAGDE, D., ELSON, E. and WEBB, W. W. (1972). Thermodynamic fluctuations in a reacting systemmeasurement by fluorescence correlation spectroscopy. *Phys. Rev. Lett.* **29** 705–708.

MCKINNEY, S. A., JOO, C. and HA, T. (2006). Analysis of single-molecule FRET trajectories using hidden Markov modeling. *Biophys. J.* **91** 1941–1951.

MCKINNEY, S. A., DÉCLAIS, A.-C., LILLEY, D. M. J. and HA, T. (2003). Structural dynamics of individual holliday junctions. *Nat. Struct. Mol. Biol.* **10** 93–97.

MEILHAC, N., LE GUYADER, L., SALOME, L. and DESTAINVILLE, N. (2006). Detection of confinement and jumps in single-molecule membrane trajectories. *Phys. Rev. E* **73** 011915.

MELNYKOV, A. V. and HALL, K. B. (2009). Revival of high-order fluorescence correlation analysis: Generalized theory and biochemical applications. *J. Phys. Chem. B* **113** 15629–15638. https://doi.org/10.1021/jp906539k

MESETH, U., WOHLAND, T., RIGLER, R. and VOGEL, H. (1999). Resolution of fluorescence correlation measurements. *Biophys. J.* **76** 1619–1631.

METZLER, R. and KLAFTER, J. (2000). The random walk's guide to anomalous diffusion: A fractional dynamics approach. *Phys. Rep.* **339** 77. MR1809268 https://doi.org/10.1016/S0370-1573(00)00070-3

METZLER, R., JEON, J.-H., CHERSTVY, A. G. and BARKAI, E. (2014). Anomalous diffusion models and their properties: Non-stationarity, non-ergodicity, and ageing at the centenary of single particle tracking. *Phys. Chem. Chem. Phys.* **16** 24128–24164.

MICHALET, X. (2010). Mean square displacement analysis of single-particle trajectories with localization error: Brownian motion in an isotropic medium. *Phys. Rev. E* **82** 041914, 13. MR2788037 https://doi.org/10.1103/PhysRevE.82.041914

MIN, W., ENGLISH, B., LUO, G., CHERAYIL, B., KOU, S. C. and XIE, X. S. (2005). Fluctuating enzymes: Lessons from single-molecule studies. *Acc. Chem. Res.* **38** 923–931.

MONNIER, N., GUO, S.-M., MORI, M., HE, J., LÉNÁRT, P. and BATHE, M. (2012). Bayesian approach to MSD-based analysis of particle motion in live cells. *Biophys. J.* **103** 616–626.

MONNIER, N., BARRY, Z., PARK, H. Y., SU, K.-C., KATZ, Z., ENGLISH, B. P., DEY, A., PAN, K., CHEESEMAN, I. M. et al. (2015). Inferring transient particle transport dynamics in live cells. *Nat. Methods* **12** 838–840.

MÜLLER, J. D. (2004). Cumulant analysis in fluorescence fluctuation spectroscopy. *Biophys. J.* **86** 3981–3992.

MÜLLER, J. D., CHEN, Y. and GRATTON, E. (2000). Resolving heterogeneity on the single molecular level with the photon-counting histogram. *Biophys. J.* **78** 474–486.

NEHER, E. and SAKMANN, B. (1976). Single-channel currents recorded from membrane of denervated frog muscle fibres. *Nature* **260** 799–802. https://doi.org/10.1038/260799a0

OKAMOTO, K. and SAKO, Y. (2012). Variational Bayes analysis of a photon-based hidden Markov model for single-molecule FRET trajectories. *Biophys. J.* **103** 1315–1324.

ORRIT, M., HA, T. and SANDOGHDAR, V. (2014). Single-molecule optical spectroscopy. *Chem. Soc. Rev.* **43** 973–976.

OTT, M., SHAI, Y. and HARAN, G. (2013). Single-particle tracking reveals switching of the HIV fusion peptide between two diffusive modes in membranes. *J. Phys. Chem. B* **117** 13308–13321. https://doi.org/10.1021/jp4039418

PEIN, F., SIELING, H. and MUNK, A. (2017). Heterogeneous change point inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1207–1227. MR3689315 https://doi.org/10.1111/rssb.12202

PERSSON, F., LINDÉN, M., UNOSON, C. and ELF, J. (2013). Extracting intracellular diffusive states and transition rates from single-molecule tracking data. *Nat. Methods* **10** 265–269. https://doi.org/10.1038/nmeth.2367

PIATT, S. and PRICE, A. C. (2019). Analyzing dwell times with the generalized method of moments. *PLoS ONE* **14** e0197726. https://doi.org/10.1371/journal.pone.0197726

QIAN, H. and ELSON, E. L. (1990a). Distribution of molecular aggregation by analysis of fluctuation moments. *Proc. Natl. Acad. Sci. USA* **87** 5479–5483.

QIAN, H. and ELSON, E. L. (1990b). On the analysis of high order moments of fluorescence fluctuations. *Biophys. J.* **57** 375–380.

QIAN, H. and KOU, S. C. (2014). Statistics and related topics in single-molecule biophysics. *Annu. Rev. Stat. Appl.* **1** 465–492. https://doi.org/10.1146/annurev-statistics-022513-115535

QIAN, H., SHEETZ, M. P. and ELSON, E. L. (1991). Single particle tracking. Analysis of diffusion and flow in two-dimensional systems. *Biophys. J.* **60** 910–921.

QIN, F. (2004). Restoration of single-channel currents using the segmental k-means method based on hidden Markov modeling. *Biophys. J.* **86** 1488–1501.

QIN, F., AUERBACH, A. and SACHS, F. (2000a). A direct optimization approach to hidden Markov modeling for single channel kinetics. *Biophys. J.* **79** 1915–1927.

QIN, F., AUERBACH, A. and SACHS, F. (2000b). Hidden Markov modeling for single channel kinetics with filtering and correlated noise. *Biophys. J.* **79** 1928–1944.

RENNER, M., WANG, L., LEVI, S., HENNEKINNE, L. and TRILLER, A. (2017). A simple and powerful analysis of lateral subdiffusion using single particle tracking. *Biophys. J.* **113** 2452–2463.

ROSALES, R. A. (2004). MCMC for hidden Markov models incorporating aggregation of states and filtering. *Bull. Math. Biol.* **66** 1173–1199. MR2253818 https://doi.org/10.1016/j.bulm.2003.12.001

ROSALES, R., STARK, J. A., FITZGERALD, W. J. and HLADKY, S. B. (2001). Bayesian restoration of ion channel records using hidden Markov models. *Biophys. J.* **80** 1088–1103.

SAKMANN, B. (2013). *Single-Channel Recording*. Springer, Berlin.

SAXTON, M. J. (1993). Lateral diffusion in an archipelago. Single-particle diffusion. *Biophys. J.* **64** 1766–1780.

SAXTON, M. J. (1994). Single-particle tracking: Models of directed transport. *Biophys. J.* **67** 2110–2119.

SAXTON, M. J. (1997). Single-particle tracking: The distribution of diffusion coefficients. *Biophys. J.* **72** 1744–1753.

SAXTON, M. J. and JACOBSON, K. (1997). Single-particle tracking: Applications to membrane dynamics. *Annu. Rev. Biophys. Biomol. Struct.* **26** 373–399.

SBALZARINI, I. F. and KOUMOUTSAKOS, P. (2005). Feature point tracking and trajectory analysis for video imaging in cell biology. *J. Struct. Biol.* **151** 182–195.

SCHENTER, G. K., LU, H. P. and XIE, X. S. (1999). Statistical analyses and theoretical models of single-molecule enzymatic dynamics. *J. Phys. Chem. A* **103** 10477–10488.

SCHMID, S. and HUGEL, T. (2018). Efficient use of single molecule time traces to resolve kinetic rates, models and uncertainties. *J. Chem. Phys.* **148** 123312. https://doi.org/10.1063/1.5006604

SCHMIDT, T., SCHÜTZ, G. J., BAUMGARTNER, W., GRUBER, H. J. and SCHINDLER, H. (1996). Imaging of single molecule diffusion. *Proc. Natl. Acad. Sci. USA* **93** 2926–2929.

SCHRÖDER, G. F. and GRUBMÜLLER, H. (2003). Maximum likelihood trajectories from single molecule fluorescence resonance energy transfer experiments. *J. Chem. Phys.* **119** 9920–9924.

SIEKMANN, I., SNEYD, J. and CRAMPIN, E. J. (2014). Statistical analysis of modal gating in ion channels. *Proc. R. Soc. A* **470** 20140030.

SIEKMANN, I., WAGNER II, L. E., YULE, D., FOX, C., BRYANT, D., CRAMPIN, E. J. and SNEYD, J. (2011). MCMC estimation of Markov models for ion channels. *Biophys. J.* **100** 1919–1929.

SIKORA, G., KEPTEN, E., WERON, A., BALCEREK, M. and BURNECKI, K. (2017a). An efficient algorithm for extracting the magnitude of the measurement error for fractional dynamics. *Phys. Chem. Chem. Phys.* **19** 26566–26581.

SIKORA, G., TEUERLE, M., WYŁOMAŃSKA, A. and GREBENKOV, D. (2017b). Statistical properties of the anomalous scaling exponent estimator based on time-averaged mean-square displacement. *Phys. Rev. E* **96** 022132. https://doi.org/10.1103/PhysRevE.96.022132

SIMSON, R., SHEETS, E. D. and JACOBSON, K. (1995). Detection of temporary lateral confinement of membrane proteins using single-particle tracking analysis. *Biophys. J.* **69** 989–993.

SLATOR, P. J. and BURROUGHS, N. J. (2018). A hidden Markov model for detecting confinement in single-particle tracking trajectories. *Biophys. J.* **115** 1741–1754. https://doi.org/10.1016/j.bpj.2018.09.005

SLATOR, P. J., CAIRO, C. W. and BURROUGHS, N. J. (2015). Detection of diffusion heterogeneity in single particle tracking trajectories using a hidden Markov model with measurement noise propagation. *PLoS ONE* **10** e0140759. https://doi.org/10.1371/journal.pone.0140759

SMITH, A. F. M. (1975). A Bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika* **62** 407–416. MR0381115 https://doi.org/10.1093/biomet/62.2.407

SUH, J., CHOY, K.-L., LAI, S. K., SUK, J. S., TANG, B. C., PRABHU, S. and HANES, J. (2007). PEGylation of nanoparticles improves their cytoplasmic transport. *Int. J. Nanomed.* **2** 735–741.

SUN, G., GUO, S.-M., TEH, C., KORZH, V., BATHE, M. and WOHLAND, T. (2015). Bayesian model selection applied to the analysis of fluorescence correlation spectroscopy data of fluorescent proteins in vitro and in vivo. *Anal. Chem.* **87** 4326–4333.

TAYLOR, J. N., MAKAROV, D. E. and LANDES, C. F. (2010). Denoising single-molecule FRET trajectories with wavelets and Bayesian inference. *Biophys. J.* **98** 164–173.

TSEKOURAS, K., CUSTER, T. C., JASHNSAZ, H., WALTER, N. G. and PRESSÉ, S. (2016). A novel method to accurately locate and count large numbers of steps by photobleaching. *Mol. Biol. Cell* **27** 3601–3615. https://doi.org/10.1091/mbc.E16-06-0404

VANDONGEN, A. M. (1996). A new algorithm for idealizing single ion channel data containing multiple unknown conductance levels. *Biophys. J.* **70** 1303–1315.

VAN DE MEENT, J.-W., BRONSON, J. E., WIGGINS, C. H. and GONZALEZ JR, R. L. (2014). Empirical Bayes methods enable advanced population-level analyses of single-molecule FRET experiments. *Biophys. J.* **106** 1327–1337.

VENKATARAMANAN, L., KUC, R. and SIGWORTH, F. J. (1998). Identification of hidden Markov models for ion channel currents. II. State-dependent excess noise. *IEEE Trans. Signal Process.* **46** 1916–1929.

VENKATARAMANAN, L., WALSH, J. L., KUC, R. and SIGWORTH, F. J. (1998). Identification of hidden Markov models for ion channel currents. I. Colored background noise. *IEEE Trans. Signal Process.* **46** 1901–1915.

VESTERGAARD, C. L., BLAINEY, P. C. and FLYVBJERG, H. (2014). Optimal estimation of diffusion coefficients from single-particle trajectories. *Phys. Rev. E* **89** 022726.

WAGNER, T., KROLL, A., HARAMAGATTI, C. R., LIPINSKI, H.-G. and WIEMANN, M. (2017). Classification and segmentation of nanoparticle diffusion trajectories in cellular micro environments. *PLoS ONE* **12** e0170165. https://doi.org/10.1371/journal.pone.0170165

WOHLAND, T., RIGLER, R. and VOGEL, H. (2001). The standard deviation in fluorescence correlation spectroscopy. *Biophys. J.* **80** 2987–2999.

WU, B. and MÜLLER, J. D. (2005). Time-integrated fluorescence cumulant analysis in fluorescence fluctuation spectroscopy. *Biophys. J.* **89** 2721–2735.

WU, Z., BI, H., PAN, S., MENG, L. and ZHAO, X. S. (2016). Determination of equilibrium constant and relative brightness in fluorescence correlation spectroscopy by considering third-order correlations. *J. Phys. Chem. B* **120** 11674–11682. https://doi.org/10.1021/acs.jpcb.6b07953

YANG, S. and CAO, J. (2001). Two-event echos in single-molecule kinetics: A signature of conformational fluctuations. *J. Phys. Chem. B* **105** 6536–6549.

YANG, H. and XIE, X. S. (2002). Statistical approaches for probing single-molecule dynamics photon-by-photon. *Chem. Phys.* **284** 423–437.

YANG, H., LUO, G., KARNCHANAPHANURACH, P., LOUIE, T.-M., RECH, I., COVA, S., XUN, L. and XIE, X. S. (2003). Protein conformational dynamics probed by single-molecule electron transfer. *Science* **302** 262–266.

YAO, Y.-C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statist. Probab. Lett.* **6** 181–189. MR0919373 https://doi.org/10.1016/0167-7152(88)90118-6

YIN, S., SONG, N. and YANG, H. (2018). Detection of velocity and diffusion coefficient change points in single-particle trajectories. *Biophys. J.* **115** 217–229.

ZHANG, T. and KOU, S. C. (2010). Nonparametric inference of doubly stochastic Poisson process data via the kernel method. *Ann. Appl. Stat.* **4** 1913–1941. MR2829941 https://doi.org/10.1214/10-AOAS352

ZHANG, N. R. and SIEGMUND, D. O. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* **63** 22–32, 309. MR2345571 https://doi.org/10.1111/j.1541-0420.2006.00662.x

ZHUANG, X., BARTLEY, L. E., BABCOCK, H. P., RUSSELL, R., HA, T., HERSCHLAG, D. and CHU, S. (2000). A single-molecule study of RNA catalysis and folding. *Science* **288** 2048–2051.

ZHUANG, X., KIM, H., PEREIRA, M. J. B., BABCOCK, H. P., WALTER, N. G. and CHU, S. (2002). Correlating structural dynamics and function in single ribozyme molecules. *Science* **296** 1473–1476.

ZWANZIG, R. (1992). Dynamical disorder: Passage through a fluctuating bottleneck. *J. Chem. Phys.* **97** 3587–3589.

Statistical Science
2020, Vol. 35, No. 1, 92–111
https://doi.org/10.1214/19-STS753

Statistical Molecule Counting in Super-Resolution Fluorescence Microscopy: Towards Quantitative Nanoscopy

Thomas Staudt, Timo Aspelmeier, Oskar Laitenberger, Claudia Geisler, Alexander Egner and Axel Munk

Abstract. Super-resolution microscopy is rapidly gaining importance as an analytical tool in the life sciences. A compelling feature is the ability to label biological units of interest with fluorescent markers in (living) cells and to observe them with considerably higher resolution than conventional microscopy permits. The images obtained this way, however, lack an absolute intensity scale in terms of numbers of fluorophores observed. In this article, we discuss state of the art methods to count such fluorophores and statistical challenges that come along with it. In particular, we suggest a modeling scheme for time series generated by single-marker-switching (SMS) microscopy that makes it possible to quantify the number of markers in a statistically meaningful manner from the raw data. To this end, we model the entire process of photon generation in the fluorophore, their passage through the microscope, detection and photoelectron amplification in the camera, and extraction of time series from the microscopic images. At the heart of these modeling steps is a careful description of the fluorophore dynamics by a novel hidden Markov model that operates on two timescales (HTMM). Besides the fluorophore number, information about the kinetic transition rates of the fluorophore's internal states is also inferred during estimation. We comment on computational issues that arise when applying our model to simulated or measured fluorescence traces and illustrate our methodology on simulated data.

Key words and phrases: Molecule counting, super-resolution microscopy, quantitative nanoscopy, biophysics and computational biology, inhomogeneous hidden Markov models, statistical thinning.

REFERENCES

[1] Aspelmeier, T., Egner, A. and Munk, A. (2015). Modern statistical challenges in high-resolution fluorescence microscopy. *Annu. Rev. Stat. Appl.* **2** 163–202.

[2] Bakshi, S., Siryaporn, A., Goulian, M. and Weisshaar, J. C. (2012). Superresolution imaging of ribosomes and RNA polymerase in live Escherichia coli cells. *Mol. Microbiol.* **85** 21–38. https://doi.org/10.1111/j.1365-2958.2012.08081.x

[3] Balzarotti, F., Eilers, Y., Gwosch, K. C., Gynnå, A. H., Westphal, V., Stefani, F. D., Elf, J. and Hell, S. W. (2017). Nanometer resolution imaging and tracking of fluorescent molecules with minimal photon fluxes. *Science* **355** 606–612.

[4] Berlier, J. E., Rothe, A., Buller, G., Bradford, J., Gray, D. R., Filanoski, B. J., Telford, W. G., Yue, S., Liu, J. et al. (2003). Quantitative comparison of long-wavelength Alexa Fluor dyes to Cy dyes: Fluorescence of the

Thomas Staudt is a PhD student, Institute for Mathematical Stochastics, Georg-August University of Göttingen, Goldschmidtstraße 7, 37077 Göttingen, Germany (e-mail: thomas.staudt@uni-goettingen.de). Timo Aspelmeier is Assistant Professor, Institute for Mathematical Stochastics, Georg-August University of Göttingen, Goldschmidtstraße 7, 37077 Göttingen, Germany (e-mail: timo.aspelmeier@mathematik.uni-goettingen.de). Oskar Laitenberger is a researcher, Laser-Laboratorium Göttingen e.V., Hans-Adolf-Krebs-Weg 1, 37077 Göttingen, Germany (e-mail: oskar.laitenberger@llg-ev.de). Claudia Geisler is a researcher, Laser-Laboratorium Göttingen e.V., Hans-Adolf-Krebs-Weg 1, 37077 Göttingen, Germany (e-mail: claudia.geisler@llg-ev.de). Alexander Egner is Director, Laser-Laboratorium Göttingen e.V., Hans-Adolf-Krebs-Weg 1, 37077 Göttingen, Germany (e-mail: alexander.egner@llg-ev.de). Axel Munk is Felix-Bernstein Professor of Statistics, Institute for Mathematical Stochastics, Georg-August University of Göttingen, Goldschmidtstraße 7, 37077 Göttingen, Germany (e-mail: munk@math.uni-goettingen.de).

dyes and their bioconjugates. *J. Histochem. Cytochem.* **51** 1699–1712.

[5] BERNING, S., WILLIG, K. I., STEFFENS, H., DIBAJ, P. and HELL, S. W. (2012). Nanoscopy in a living mouse brain. *Science* **335** 551–551.

[6] BETZIG, E., PATTERSON, G. H., SOUGRAT, R., LINDWASSER, O. W., OLENYCH, S., BONIFACINO, J. S., DAVIDSON, M. W., LIPPINCOTT-SCHWARTZ, J. and HESS, H. F. (2006). Imaging intracellular fluorescent proteins at nanometer resolution. *Science* **313** 1642–1645.

[7] BORN, M. and WOLF, E. (1999). *Principles of Optics*, 7th ed. Cambridge Univ. Press, Cambridge.

[8] BRAKEMANN, T., STIEL, A. C., WEBER, G., ANDRESEN, M., TESTA, I., GROTJOHANN, T., LEUTENEGGER, M., PLESSMANN, U., URLAUB, H. et al. (2011). A reversibly photoswitchable GFP-like protein with fluorescence excitation decoupled from switching. *Nat. Biotechnol.* **29** 942–947.

[9] CHEN, C., ZONG, S., WANG, Z., LU, J., ZHU, D., ZHANG, Y. and CUI, Y. (2016). Imaging and intracellular tracking of cancer-derived exosomes using single-molecule localization-based super-resolution microscope. *ACS Appl. Mater. Interfaces* **8** 25825–25833.

[10] CHOJNACKI, J., STAUDT, T., GLASS, B., BINGEN, P., ENGELHARDT, J., ANDERS, M., SCHNEIDER, J., MÜLLER, B., HELL, S. W. et al. (2012). Maturation-dependent HIV-1 surface protein redistribution revealed by fluorescence nanoscopy. *Science* **338** 524–528.

[11] D'ESTE, E., KAMIN, D., GÖTTFERT, F., EL-HADY, A. and HELL, S. W. (2015). STED nanoscopy reveals the ubiquity of subcortical cytoskeleton periodicity in living neurons. *Cell Rep.* **10** 1246–1251.

[12] EGNER, A., GEISLER, C., VON MIDDENDORFF, C., BOCK, H., WENZEL, D., MEDDA, R., ANDRESEN, M., STIEL, A. C., JAKOBS, S. et al. (2007). Fluorescence nanoscopy in whole cells by asynchronous localization of photoswitching emitters. *Biophys. J.* **93** 3285–3290.

[13] EILERS, Y., TA, H., GWOSCH, K. C., BALZAROTTI, F. and HELL, S. W. (2018). MINFLUX monitors rapid molecular jumps with superior spatiotemporal resolution. *Proc. Natl. Acad. Sci. USA* **115** 6117–6122.

[14] FELLER, W. (2008). *An Introduction to Probability Theory and Its Applications*, **2**. Wiley, New York, NY.

[15] FÖLLING, J., BOSSI, M., BOCK, H., MEDDA, R., WURM, C. A., HEIN, B., JAKOBS, S., EGGELING, C. and HELL, S. W. (2008). Fluorescence nanoscopy by ground-state depletion and single-molecule return. *Nat. Methods* **5** 943–945.

[16] FRAHM, L., KELLER-FINDEISEN, J., ALT, P., SCHNORRENBERG, S., RUIZ, M. D. Á., ASPELMEIER, T., MUNK, A., JAKOBS, S. and HELL, S. W. (2019). Molecular contribution function in RESOLFT nanoscopy. *Opt. Express* **27** 21956–21987. https://doi.org/10.1364/OE.27.021956

[17] GOODMAN, J. W. (1996). *Introduction to Fourier Optics*, 2nd ed. McGraw-Hill, New York.

[18] GROTJOHANN, T., TESTA, I., LEUTENEGGER, M., BOCK, H., URBAN, N. T., LAVOIE-CARDINAL, F., WILLIG, K. I., EGGELING, C., JAKOBS, S. et al. (2011). Diffraction-unlimited all-optical imaging and writing with a photochromic GFP. *Nature* **478** 204–208.

[19] HARREMOËS, P., JOHNSON, O. and KONTOYIANNIS, I. (2010). Thinning, entropy, and the law of thin numbers. *IEEE Trans. Inform. Theory* **56** 4228–4244. MR2807322 https://doi.org/10.1109/TIT.2010.2053893

[20] HARTMANN, A., HUCKEMANN, S., DANNEMANN, J., LAITENBERGER, O., GEISLER, C., EGNER, A. and MUNK, A. (2016). Drift estimation in sparse sequential dynamic imaging, with application to nanoscale fluoresence microscopy. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 563–587. MR3506793 https://doi.org/10.1111/rssb.12128

[21] HELL, S. W. (2008). Microscopy and its focal switch. *Nat. Methods* **6** 24–32.

[22] HELL, S. W., SAHL, S. J., BATES, M., ZHUANG, X., HEINTZMANN, R., BOOTH, M. J., BEWERSDORF, J., SHTENGEL, G., HESS, H. et al. (2015). The 2015 super-resolution microscopy roadmap. *J. Phys. D, Appl. Phys.* **48** 443001.

[23] HELL, S. W. and WICHMANN, J. (1994). Breaking the diffraction resolution limit by stimulated emission: Stimulated-emission-depletion fluorescence microscopy. *Opt. Lett.* **19** 780–782.

[24] HESS, S. T., GIRIRAJAN, T. P. K. and MASON, M. D. (2006). Ultra-high resolution imaging by fluorescence photoactivation localization microscopy. *Biophys. J.* **91** 4258–4272.

[25] HIRSCH, M., WAREHAM, R. J., MARTIN-FERNANDEZ, M. L., HOBSON, M. P. and ROLFE, D. J. (2013). A stochastic model for electron multiplication charge-coupled devices—from theory to practice. *PLoS ONE* **8** e53671. https://doi.org/10.1371/journal.pone.0053671

[26] HOFMANN, M., EGGELING, C., JAKOBS, S. and HELL, S. W. (2005). Breaking the diffraction barrier in fluorescence microscopy at low light intensities by using reversibly photoswitchable proteins. *Proc. Natl. Acad. Sci. USA* **102** 17565–17569.

[27] HUMMER, G., FRICKE, F. and HEILEMANN, M. (2016). Model-independent counting of molecules in single-molecule localization microscopy. *Mol. Biol. Cell* **27** 3637–3644. https://doi.org/10.1091/mbc.E16-07-0525

[28] KOENIG, M., BERLAGE, C., REISCH, P., OELSNER, C., KOBERLING, F., TA, H. and ERDMANN, R. (2019). Molecular counting by photon statistics in confocal fluorescence imaging. *Biophys. J.* **116** 134a–135a.

[29] KOMIS, G., MISTRIK, M., ŠAMAJOVÁ, O., OVEČKA, M., BARTEK, J. and ŠAMAJ, J. (2015). Superresolution live imaging of plant cells using structured illumination microscopy. *Nat. Protoc.* **10** 1248–1263.

[30] LAITENBERGER, O., ASPELMEIER, T., GEISLER, C., MUNK, A. and EGNER, A. (2019). Towards unbiased molecule counting in superresolution fluorescence microscopy. Preprint.

[31] LAPLANTE, C., HUANG, F., TEBBS, I. R., BEWERSDORF, J. and POLLARD, T. D. (2016). Molecular organization of cytokinesis nodes and contractile rings by super-resolution fluorescence microscopy of live fission yeast. *Proc. Natl. Acad. Sci. USA* **113** E5876–E5885.

[32] LEE, S.-H., SHIN, J. Y., LEE, A. and BUSTAMANTE, C. (2012). Counting single photoactivatable fluorescent molecules by photoactivated localization microscopy (PALM). *Proc. Natl. Acad. Sci. USA* **109** 17436–17441.

[33] LIN, Y., LONG, J. J., HUANG, F., DUIM, W. C., KIRSCHBAUM, S., ZHANG, Y., SCHROEDER, L. K., REBANE, A. A., VELASCO, M. G. M. et al. (2015). Quantifying and optimizing single-molecule switching nanoscopy at high speeds. *PLoS ONE* **10** e0128135.

[34] MAGLIONE, M. and SIGRIST, S. J. (2013). Seeing the forest tree by tree: Super-resolution light microscopy meets the neurosciences. *Nat. Neurosci.* **16** 790–797.

[35] MESSINA, T. C., KIM, H., GIURLEO, J. T. and TALAGA, D. S. (2006). Hidden Markov model analysis of multichromophore photobleaching. *J. Phys. Chem. B* **110** 16366–16376.

[36] MURANYI, W., MALKUSCH, S., MÜLLER, B., HEILEMANN, M. and KRÄUSSLICH, H.-G. (2013). Super-resolution microscopy reveals specific recruitment of HIV-1 envelope proteins

to viral assembly sites dependent on the envelope C-terminal tail. *PLoS Pathog*. **9** e1003198. https://doi.org/10.1371/journal.ppat.1003198

[37] NELDER, J. A. and MEAD, R. (1965). A simplex method for function minimization. *Comput. J.* **7** 308–313. MR3363409 https://doi.org/10.1093/comjnl/7.4.308

[38] PAGEON, S. V., CORDOBA, S.-P., OWEN, D. M., ROTHERY, S. M., OSZMIANA, A. and DAVIS, D. M. (2013). Superresolution microscopy reveals nanometer-scale reorganization of inhibitory natural killer cell receptors upon activation of NKG2D. *Sci. Signal*. **6** ra62. https://doi.org/10.1126/scisignal.2003947

[39] PATEL, L., GUSTAFSSON, N., LIN, Y., OBER, R., HENRIQUES, R. and COHEN, E. (2019). A hidden Markov model approach to characterizing the photo-switching behavior of fluorophores. *Ann. Appl. Stat*. **13** 1397–1429. MR4019144 https://doi.org/10.1214/19-AOAS1240

[40] PRESCHER, J., BAUMGÄRTEL, V., IVANCHENKO, S., TORRANO, A. A., BRÄUCHLE, C., MÜLLER, B. and LAMB, D. C. (2015). Super-resolution imaging of ESCRT-proteins at HIV-1 assembly sites. *PLoS Pathog*. **11** e1004677. https://doi.org/10.1371/journal.ppat.1004677

[41] ROBBINS, M. S. and HADWEN, B. J. (2003). The noise performance of electron multiplying charge-coupled devices. *IEEE Trans. Electron Devices* **50** 1227–1232.

[42] ROLLINS, G. C., SHIN, J. Y., BUSTAMANTE, C. and PRESSÉ, S. (2015). Stochastic approach to the molecular counting problem in superresolution microscopy. *Proc. Natl. Acad. Sci. USA* **112** E110–E118.

[43] RUST, M. J., BATES, M. and ZHUANG, X. (2006). Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nat. Methods* **3** 793–796.

[44] SAHL, S. J., HELL, S. W. and JAKOBS, S. (2017). Fluorescence nanoscopy in cell biology. *Nat. Rev., Mol. Cell Biol.* **18** 685–701. https://doi.org/10.1038/nrm.2017.71

[45] SCHNEIDER, L. F., SCHMIDT-HIEBER, J., STAUDT, T., KRAJINA, A., ASPELMEIER, T. and MUNK, A. (2018). Posterior consistency for n in the binomial (n, p) problem with both parameters unknown—with applications to quantitative nanoscopy. ArXiv Preprint. Available at arXiv:1809.02443.

[46] SCHNORRENBERG, S., GROTJOHANN, T., VORBRÜGGEN, G., HERZIG, A., HELL, S. W. and JAKOBS, S. (2016). In vivo super-resolution RESOLFT microscopy of Drosophila melanogaster. *eLife* **5**. https://doi.org/10.7554/eLife.15567

[47] SHARMA, S., SANTISKULVONG, C., BENTOLILA, L. A., RAO, J., DORIGO, O. and GIMZEWSKI, J. K. (2012). Correlative nanomechanical profiling with super-resolution F-actin imaging reveals novel insights into mechanisms of cisplatin resistance in ovarian cancer cells. *Nanomedicine* **8** 757–766.

[48] STAUDT, T., ASPELMEIER, T., LAITENBERGER, O., GEISLER, C., EGNER, A. and MUNK, A. (2020). Supplement to "Statistical Molecule Counting in Super-Resolution Fluorescence Microscopy: Towards Quantitative Nanoscopy." https://doi.org/10.1214/19-STS753SUPP.

[49] SYDOR, A. M., CZYMMEK, K. J., PUCHNER, E. M. and MENNELLA, V. (2015). Super-resolution microscopy: From single molecules to supramolecular assemblies. *Trends Cell Biol*. **25** 730–748. https://doi.org/10.1016/j.tcb.2015.10.004

[50] TA, H., KELLER, J., HALTMEIER, M., SAKA, S. K., SCHMIED, J., OPAZO, F., TINNEFELD, P., MUNK, A. and HELL, S. W. (2015). Mapping molecules in scanning far-field fluorescence nanoscopy. *Nat. Commun*. **6** 7977. https://doi.org/10.1038/ncomms8977

[51] TSEKOURAS, K., CUSTER, T. C., JASHNSAZ, H., WALTER, N. G. and PRESSÉ, S. (2016). A novel method to accurately locate and count large numbers of steps by photobleaching. *Mol. Biol. Cell* **27** 3601–3615. https://doi.org/10.1091/mbc.E16-06-0404

[52] VAN DE LINDE, S., LÖSCHBERGER, A., KLEIN, T., HEIDBREDER, M., WOLTER, S., HEILEMANN, M. and SAUER, M. (2011). Direct stochastic optical reconstruction microscopy with standard fluorescent probes. *Nat. Protoc*. **6** 991–1009. https://doi.org/10.1038/nprot.2011.336

[53] VOGELSANG, J., STEINHAUER, C., FORTHMANN, C., STEIN, I. H., PERSON-SKEGRO, B., CORDES, T. and TINNEFELD, P. (2010). Make them blink: Probes for super-resolution microscopy. *ChemPhysChem* **11** 2475–2490.

[54] WILLIAMSON, D. J., OWEN, D. M., ROSSY, J., MAGENAU, A., WEHRMANN, M., GOODING, J. J. and GAUS, K. (2011). Pre-existing clusters of the adaptor Lat do not participate in early T cell signaling events. *Nat. Immunol*. **12** 655–662.

Statistical Science
2020, Vol. 35, No. 1, 112–128
https://doi.org/10.1214/19-STS7560

Data Denoising and Post-Denoising Corrections in Single Cell RNA Sequencing

Divyansh Agarwal, Jingshu Wang and Nancy R. Zhang

Abstract. Single cell sequencing technologies are transforming biomedical research. However, due to the inherent nature of the data, single cell RNA sequencing analysis poses new computational and statistical challenges. We begin with a survey of a selection of topics in this field, with a gentle introduction to the biology and a more detailed exploration of the technical noise. We consider in detail the problem of single cell data denoising, sometimes referred to as "imputation" in the relevant literature. We discuss why this is not a typical statistical imputation problem, and review current approaches to this problem. We then explore why the use of denoised values in downstream analyses invites novel statistical insights, and how denoising uncertainty should be accounted for to yield valid statistical inference. The utilization of denoised or imputed matrices in statistical inference is not unique to single cell genomics, and arises in many other fields. We describe the challenges in this type of analysis, discuss some preliminary solutions, and highlight unresolved issues.

Key words and phrases: Single cell biology, RNA sequencing, imputation, post-denoising inference, empirical Bayes, deep learning.

REFERENCES

ANDREWS, T. S. and HEMBERG, M. (2018). False signals induced by single-cell imputation. *F*1000*Res* **7**.

AGARWAL, D., WANG, J. and ZHANG, N. R (2020). Supplement to "Data Denoising and Post-Denoising orrections in Single Cell RNA Sequencing." https://doi.org/10.1214/19-STS7560SUPP.

ARKIN, A., ROSS, J. and MCADAMS, H. H. (1998). Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected Escherichia coli cells. *Genetics* **149** 1633–1648.

BADSHA, M. B., LI, R., LIU, B., LI, Y. I., XIAN, M., BANOVICH, N. E. and FU, A. Q. (2018). Imputation of single-cell gene expression with an autoencoder neural network. *BioRxiv* 504977.

BARON, M., VERES, A., WOLOCK, S. L., FAUST, A. L., GAUJOUX, R., VETERE, A., RYU, J. H., WAGNER, B. K., SHEN-ORR, S. S. et al. (2016). A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell Systems* **3** 346–360.

BARROSO, G. V., PUZOVIC, N. and DUTHEIL, J. Y. (2018). The evolution of gene-specific transcriptional noise is driven by selection at the pathway level. *Genetics* **208** 173–189.

BRENNECKE, P., ANDERS, S., KIM, J. K., KOŁODZIEJCZYK, A. A., ZHANG, X., PROSERPIO, V., BAYING, B., BENES, V., TEICHMANN, S. A. et al. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10** 1093–1095.

BUTLER, A., HOFFMAN, P., SMIBERT, P., PAPALEXI, E. and SATIJA, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36** 411–420.

CHEN, M. and ZHOU, X. (2018). VIPER: Variability-preserving imputation for accurate gene expression recovery in single-cell RNA sequencing studies. *Genome Biol.* **19** 196.

CHEN, R., WU, X., JIANG, L. and ZHANG, Y. (2017). Single-cell RNA-seq reveals hypothalamic cell diversity. *Cell Reports* **18** 3227–3241.

CHIRON, L., VAN AGTHOVEN, M. A., KIEFFER, B., ROLANDO, C. and DELSUC, M.-A. (2014). Efficient denoising algorithms for large experimental datasets and their applications in Fourier transform ion cyclotron resonance mass spectrometry. *Proc. Natl. Acad. Sci. USA* **111** 1385–1390.

CLEVERS, H., RAFELSKI, S. and ELOWITZ, M. et al. (2017). What is your conceptual definition of 'cell type' in the context of a mature organism? *Cell Systems* **4** 255–259.

DEGRELLE, S. A., HENNEQUET-ANTIER, C., CHIAPELLO, H., PIOT-KAMINSKI, K., PIUMI, F., ROBIN, S., RENARD, J.-P. and HUE, I. (2008). Amplification biases: Possible differences among deviating gene expressions. *BMC Genomics* **9** 46.

Divyansh Agarwal is Ph.D. and M.D. candidate, Graduate Program in Genomics and Computational Biology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA (e-mail: divyansh.agarwal@pennmedicine.upenn.edu). Jingshu Wang is Assistant Professor, Department of Statistics, The University of Chicago, Chicago, Illinois 60637, USA (e-mail: jingshuw@uchicago.edu). Nancy R. Zhang is Professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA (e-mail: nzh@wharton.upenn.edu).

DI GREGORIO, A., BOWLING, S. and RODRIGUEZ, T. A. (2016). Cell competition and its role in the regulation of cell fitness from development to cancer. *Developmental Cell* **38** 621–634.

EBERWINE, J., YEH, H., MIYASHIRO, K., CAO, Y., NAIR, S., FINNELL, R., ZETTEL, M. and COLEMAN, P. (1992). Analysis of gene expression in single live neurons. *Proc. Natl. Acad. Sci. USA* **89** 3010–3014.

ELDAR, A. and ELOWITZ, M. B. (2010). Functional roles for noise in genetic circuits. *Nature* **467** 167–173.

ELOWITZ, M. B., LEVINE, A. J., SIGGIA, E. D. and SWAIN, P. S. (2002). Stochastic gene expression in a single cell. *Science* **297** 1183–1186.

ENGE, M., ARDA, H. E., MIGNARDI, M., BEAUSANG, J., BOTTINO, R., KIM, S. K. and QUAKE, S. R. (2017). Single-cell analysis of human pancreas reveals transcriptional signatures of aging and somatic mutation patterns. *Cell* **171** 321–330.

ERASLAN, G., SIMON, L. M., MIRCEA, M., MUELLER, N. S. and THEIS, F. J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* **10** 390.

GONG, W., KWAK, I.-Y., POTA, P., KOYANO-NAKAGAWA, N. and GARRY, D. J. (2018). DrImpute: Imputing dropout events in single cell RNA sequencing data. *BMC Bioinform.* **19** 220.

GOSSETT, D. R., HENRY, T., LEE, S. A., YING, Y., LINDGREN, A. G., YANG, O. O., RAO, J., CLARK, A. T. and DI CARLO, D. (2012). Hydrodynamic stretching of single cells for large population mechanical phenotyping. *Proc. Natl. Acad. Sci. USA* **109** 7630–7635.

HAFEMEISTER, C. and SATIJA, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. BioRxiv 576827.

HAGHVERDI, L., LUN, A. T. L., MORGAN, M. D. and MARIONI, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36** 421–427.

HAN, X., WANG, R., ZHOU, Y., FEI, L., SUN, H., LAI, S., SAADATPOUR, A., ZHOU, Z., CHEN, H. et al. (2018). Mapping the mouse cell atlas by microwell-seq. *Cell* **172** 1091–1107.

HEDLUND, E. and DENG, Q. (2018). Single-cell RNA sequencing: Technical advancements and biological applications. *Mol. Aspects Med.* **59** 36–46.

HICKS, S. C., TOWNES, F. W., TENG, M. and IRIZARRY, R. A. (2018). Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* **19** 562–578.

HSU, L., SELF, S. G., GROVE, D., RANDOLPH, T., WANG, K., DELROW, J. J., LOO, L. and PORTER, P. (2005). Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics* **6** 211–226.

HUANG, M., WANG, J., TORRE, E., DUECK, H., SHAFFER, S., BONASIO, R., MURRAY, J. I., RAJ, A., LI, M. et al. (2018). SAVER: Gene expression recovery for single-cell RNA sequencing. *Nat. Methods* **15** 539.

HWANG, B., LEE, J. H. and BANG, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & Molecular Medicine* **50** 1–14.

ISLAM, S., ZEISEL, A., JOOST, S., MANNO, G. L., ZAJAC, P., KASPER, M., LÖNNERBERG, P. and LINNARSSON, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11** 163–166.

KIM, J. K., KOLODZIEJCZYK, A. A., ILICIC, T., ILLICIC, T., TEICHMANN, S. A. and MARIONI, J. C. (2015). Characterizing noise structure in single-cell RNA-seq distinguishes genuine from technical stochastic allelic expression. *Nat. Commun.* **6** 8687.

KIM, T., CHEN, I. R., LIN, Y., WANG, A. Y.-Y., YANG, J. Y. H. and YANG, P. (2019). Impact of similarity metrics on single-cell RNA-seq data clustering. *Brief. Bioinform.* **20** 2316–2326.

KLEIN, A. M., MAZUTIS, L., AKARTUNA, I., TALLAPRAGADA, N., VERES, A., LI, V., PESHKIN, L., WEITZ, D. A. and KIRSCHNER, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161** 1187–1201.

KOLODZIEJCZYK, A. A., KIM, J. K., SVENSSON, V., MARIONI, J. C. and TEICHMANN, S. A. (2015). The technology and biology of single-cell RNA sequencing. *Molecular Cell* **58** 610–620.

LA MANNO, G., GYLLBORG, D., CODELUPPI, S., NISHIMURA, K., SALTO, C., ZEISEL, A., BORM, L. E., STOTT, S. R., TOLEDO, E. M. et al. (2016). Molecular diversity of midbrain development in mouse, human, and stem cells. *Cell* **167** 566–580.

LA MANNO, G., SOLDATOV, R., ZEISEL, A., BRAUN, E., HOCHGERNER, H., PETUKHOV, V., LIDSCHREIBER, K., KASTRITI, M. E., LÖNNERBERG, P. et al. (2018). RNA velocity of single cells. *Nature* **560** 494–498.

LI, W. V. and LI, J. J. (2018). An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.* **9** 997.

LINDERMAN, G. C., ZHAO, J. and KLUGER, Y. (2018). Zero-preserving imputation of scRNA-seq data using low-rank approximation. *BioRxiv* 397588.

LOPEZ, R., REGIER, J., COLE, M. B., JORDAN, M. I. and YOSEF, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15** 1053–1058.

LOSICK, R. and DESPLAN, C. (2008). Stochasticity and cell fate. *Science* **320** 65–68.

MARTINEZ-JIMENEZ, C. P., ELING, N., CHEN, H.-C., VALLEJOS, C. A., KOLODZIEJCZYK, A. A., CONNOR, F., STOJIC, L., RAYNER, T. F., STUBBINGTON, M. J. T. et al. (2017). Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science* **355** 1433–1436.

MCADAMS, H. H. and ARKIN, A. (1997). Stochastic mechanisms in gene expression. *Proc. Natl. Acad. Sci. USA* **94** 814–819.

NOVICK, A. and WEINER, M. (1957). Enzyme induction as an all-or-none phenomenon. *Proc. Natl. Acad. Sci. USA* **43** 553–566.

PAPALEXI, E. and SATIJA, R. (2018). Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev., Immunol.* **18** 35–45.

PAREKH, S., ZIEGENHAIN, C., VIETH, B., ENARD, W. and HELLMANN, I. (2016). The impact of amplification on differential expression analyses by RNA-seq. *Sci. Rep.* **6** 25533. https://doi.org/10.1038/srep25533

PARK, J., SHRESTHA, R., QIU, C., KONDO, A., HUANG, S., WERTH, M., LI, M., BARASCH, J. and SUSZTÁK, K. (2018). Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. *Science* **360** 758–763.

PEARSON, K. (1982). *The Grammer of Science*. Cambridge Univ. Press, Cambridge.

RAJ, A. and VAN OUDENAARDEN, A. (2008). Nature, nurture, or chance: Stochastic gene expression and its consequences. *Cell* **135** 216–226.

REGEV, A., TEICHMANN, S. A., LANDER, E. S., AMIT, I., BENOIST, C., BIRNEY, E., BODENMILLER, B., CAMPBELL, P., CARNINCI, P. et al. (2017). Science forum: The human cell atlas. *eLife* **6** e27041.

ROZENBLATT-ROSEN, O., STUBBINGTON, M. J., REGEV, A. and TEICHMANN, S. A. (2017). The human cell atlas: From vision to reality. *Nature News* **550** 451.

SAELENS, W., CANNOODT, R., TODOROV, H. and SAEYS, Y. (2019). A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37** 547–554.

SKINNIDER, M. A., SQUAIR, J. W. and FOSTER, L. J. (2019). Evaluating measures of association for single-cell transcriptomics. *Nat. Methods* **16** 381–386.

SONESON, C. and ROBINSON, M. D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15** 255–261.

SONG, R., SARNOSKI, E. A. and ACAR, M. (2018). The systems biology of single-cell aging. *IScience* **7** 154–169.

STUART, T. and SATIJA, R. (2019). Integrative single-cell analysis. *Nat. Rev. Genet.* **20** 257–272.

STUART, T., BUTLER, A., HOFFMAN, P., HAFEMEISTER, C., PAPALEXI, E., MAUCK, W. M., HAO, Y., STOECKIUS, M., SMIBERT, P. et al. (2019). Comprehensive integration of single-cell data. *Cell* **177** 1888–1902.

SVENSSON, V., VENTO-TORMO, R. and TEICHMANN, S. A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nat Protoc* **13** 599–604.

SVENSSON, V., NATARAJAN, K. N., LY, L.-H., MIRAGAIA, R. J., LABALETTE, C., MACAULAY, I. C., CVEJIC, A. and TEICHMANN, S. A. (2017). Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* **14** 381–387.

THE TABULA MURIS CONSORTIUM (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562** 367.

TANG, F., BARBACIORU, C., WANG, Y., NORDMAN, E., LEE, C., XU, N., WANG, X., BODEAU, J., TUCH, B. B. et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6** 377.

TESCHENDORFF, A. E. and ENVER, T. (2017). Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. *Nat. Commun.* **8** 15599.

TIAN, L., DONG, X., FREYTAG, S., LÊ CAO, K.-A., SU, S., JALALABADI, A., AMANN-ZALCENSTEIN, D., WEBER, T. S., SEIDI, A. et al. (2019). Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat. Methods* **16** 479–487.

TRAPNELL, C. (2015). Defining cell types and states with single-cell genomics. *Genome Res.* **25** 1491–1498.

TUNG, P.-Y., BLISCHAK, J. D., HSIAO, C. J., KNOWLES, D. A., BURNETT, J. E., PRITCHARD, J. K. and GILAD, Y. (2017). Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.* **7** 39921.

VAN DIJK, D., SHARMA, R., NAINYS, J., YIM, K., KATHAIL, P., CARR, A. J., BURDZIAK, C., MOON, K. R., CHAFFER, C. L. et al. (2018). Recovering gene interactions from single-cell data using data diffusion. *Cell* **174** 716–729.

VAN GELDER, R. N., VON ZASTROW, M. E., YOOL, A., DEMENT, W. C., BARCHAS, J. D. and EBERWINE, J. H. (1990). Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *Proc. Natl. Acad. Sci. USA* **87** 1663–1667.

WAGNER, F., YAN, Y. and YANAI, I. (2017). K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data. *BioRxiv* 217737.

WANG, J., HUANG, M., TORRE, E., DUECK, H., SHAFFER, S., MURRAY, J., RAJ, A., LI, M. and ZHANG, N. R. (2018). Gene expression distribution deconvolution in single-cell RNA sequencing. *Proc. Natl. Acad. Sci. USA* **115** E6437–E6446.

WANG, J., AGARWAL, D., HUANG, M., HU, G., ZHOU, Z., YE, C. and ZHANG, N. R. (2019). Data denoising with transfer learning in single-cell transcriptomics. *Nat. Methods* **16** 875–878.

ZAPPIA, L., PHIPSON, B. and OSHLACK, A. (2018). Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS Comput. Biol.* **14** e1006245.

ZEISEL, A., MUÑOZ-MANCHADO, A. B., CODELUPPI, S., LÖNNERBERG, P., LA MANNO, G., JURÉUS, A., MARQUES, S., MUNGUBA, H., HE, L. et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347** 1138–1142.

ZHANG, L. and ZHANG, S. (2018). Comparison of computational methods for imputing single-cell RNA-sequencing data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* https://doi.org/10.1109/TCBB.2018.2848633

ZHENG, G. X., TERRY, J. M., BELGRADER, P., RYVKIN, P., BENT, Z. W., WILSON, R., ZIRALDO, S. B., WHEELER, T. D., MCDERMOTT, G. P. et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8** 14049.

ZIEGENHAIN, C., VIETH, B., PAREKH, S., REINIUS, B., GUILLAUMET-ADKINS, A., SMETS, M., LEONHARDT, H., HEYN, H., HELLMANN, I. et al. (2017). Comparative analysis of single-cell RNA sequencing methods. *Molecular Cell* **65** 631–643.

Statistical Science
2020, Vol. 35, No. 1, 129–144
https://doi.org/10.1214/19-STS7561

Statistical Inference for the Evolutionary History of Cancer Genomes

Khanh N. Dinh, Roman Jaksik, Marek Kimmel, Amaury Lambert and Simon Tavaré

Abstract. Recent years have seen considerable work on inference about cancer evolution from mutations identified in cancer samples. Much of the modeling work has been based on classical models of population genetics, generalized to accommodate time-varying cell population size. Reverse-time, genealogical views of such models, commonly known as coalescents, have been used to infer aspects of the past of growing populations. Another approach is to use branching processes, the simplest scenario being the classical linear birth-death process. Inference from evolutionary models of DNA often exploits summary statistics of the sequence data, a common one being the so-called Site Frequency Spectrum (SFS). In a bulk tumor sequencing experiment, we can estimate for each site at which a novel somatic point mutation has arisen, the proportion of cells that carry that mutation. These numbers are then grouped into collections of sites which have similar mutant fractions. We examine how the SFS based on birth-death processes differs from those based on the coalescent model. This may stem from the different sampling mechanisms in the two approaches. However, we also show that despite this, they are quantitatively comparable for the range of parameters typical for tumor cell populations. We also present a model of tumor evolution with selective sweeps, and demonstrate how it may help in understanding the history of a tumor as well as the influence of data pre-processing. We illustrate the theory with applications to several examples from The Cancer Genome Atlas tumors.

Key words and phrases: Cancer evolution, coalescents, birth-death processes, site frequency spectrum, tumor heterogeneity, clonal selection, ploidy, bulk sequencing.

REFERENCES

ABRAMOWITZ, M. and STEGUN, I. A. (1964). *Handbook of Mathematical Functions*. *Applied Mathematics Series* **55**. National Bureau of Standards.

CHEEK, D. and ANTAL, T. (2018). Mutation frequencies in a birth-death branching process. *Ann. Appl. Probab.* **28** 3922–3947. MR3861830 https://doi.org/10.1214/18-AAP1413

DEL MONTE, U. (2009). Does the cell number 10^9 still really fit one gram of tumor tissue? *Cell Cycle* **8** 505–506. https://doi.org/10.4161/cc.8.3.7608

DINH, K. N, JAKSIK, R., KIMMEL, M., LAMBERT, A. and TAVARÉ, S. (2020). Supplement to "Statistical inference for the evolutionary history of cancer genomes." https://doi.org/10.1214/19-STS7561SUPP.

DURRETT, R. (2013). Population genetics of neutral mutations in exponentially growing cancer cell populations. *Ann. Appl. Probab.* **23** 230–250. MR3059234 https://doi.org/10.1214/11-AAP824

EWENS, W. J. (1972). The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* **3** 87–112. MR0325177 https://doi.org/10.1016/0040-5809(72)90035-4

FU, Y. X. (1995). Statistical properties of segregating sites. *Theor.*

Khanh Dinh is a Postdoctoral Researcher, Department of Statistics, Columbia University, New York, New York 10027, USA (e-mail: knd2127@columbia.edu). Roman Jaksik is Assistant Professor, Silesian University of Technology, Gliwice, Poland (e-mail: Roman.Jaksik@polsl.pl). Marek Kimmel is a Professor, Statistics and Bioengineering, Rice University with an adjunct appointment at the Silesian University of Technology, Gliwice, Poland (e-mail: kimmel@rice.edu). Amaury Lambert is Full Professor, Lab of Probability, Statistics & Modeling, Sorbonne Université, Paris, France. He also runs a research group at the Center for Interdisciplinary Research in Biology, Collège de France, Paris, France. (e-mail: amaury.lambert@college-de-france.fr). Simon Tavaré is the Herbert and Florence Irving Director of the Irving Institute for Cancer Dynamics, and a professor in Statistics and Biological Sciences at Columbia University, New York, New York 10027, USA, and a Senior Associate Core Member, New York Genome Center, New York, New York 10013, USA (e-mail: st3193@columbia.edu).

Popul. Biol. **48** 172–197.

GERNHARD, T. (2008). The conditioned reconstructed process. *J. Theoret. Biol.* **253** 769–778. MR2964590 https://doi.org/10.1016/j.jtbi.2008.04.005

GREAVES, M. and MALEY, C. C. (2012). Clonal evolution in cancer. *Nature* **481** 306–313.

GRIFFITHS, R. C. and TAVARÉ, S. (1998). The age of a mutation in a general coalescent tree. *Stoch. Models* **14** 273–295. MR1617552 https://doi.org/10.1080/15326349808807471

HACCOU, P., JAGERS, P. and VATUTIN, V. A. (2007). *Branching Processes*: *Variation*, *Growth*, *and Extinction of Populations*. *Cambridge Studies in Adaptive Dynamics* **5**. Cambridge Univ. Press, Cambridge. MR2429372

JAGERS, P. (1975). *Branching Processes with Biological Applications*. Wiley, London–New York–Sydney. MR0488341

KIMMEL, M. and AXELROD, D. E. (2015). *Branching Processes in Biology*, 2nd ed. *Interdisciplinary Applied Mathematics* **19**. Springer, New York. MR3310028 https://doi.org/10.1007/978-1-4939-1559-0

KINGMAN, J. F. C. (1982a). On the genealogy of large populations. *J. Appl. Probab.* **19A** 27–43. MR0633178

KINGMAN, J. F. C. (1982b). The coalescent. *Stochastic Process. Appl.* **13** 235–248. MR0671034 https://doi.org/10.1016/0304-4149(82)90011-4

KLASSMAN, A. and FERRETTI, L. (2017). The third moments of the site frequency spectrum. *BioArxiv*. https://doi.org/10.1101/109579.

KUIPERS, J., JAHN, K., RAPHAEL, B. J. and BEERENWINKEL, N. (2017). Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. *Genome Res.* **27** 1885–1894. https://doi.org/10.1101/gr.220707.117

LAKS, E., MCPHERSON, A., ZAHN, H., LAI, D., STEIF, A., BRIMHALL, J., BIELE, J., WANG, B., MASUD, T. et al. (2019). Clonal decomposition and DNA replication states defined by scaled single-cell genome sequencing. *Cell* **179** 1207–1221.e22.

LAMBERT, A. (2009). The allelic partition for coalescent point processes. *Markov Process. Related Fields* **15** 359–386. MR2554367

LAMBERT, A. (2010). The contour of splitting trees is a Lévy process. *Ann. Probab.* **38** 348–395. MR2599603 https://doi.org/10.1214/09-AOP485

LAMBERT, A. and STADLER, T. (2013). Birth-death models and coalescent point processes: The shape and probability of reconstructed phylogenies. *Theor. Popul. Biol.* **90** 113–128.

LING, S., HU, Z., YANG, Z., YANG, F., LI, Y., LIN, P., CHEN, K., DONG, L., CAO, L. et al. (2015). Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution. *Proc. Natl. Acad. Sci. USA* **112** E6496–E6505.

MORAN, P. A. P. (1958). Random processes in genetics. *Proc. Camb. Philos. Soc.* **54** 60–71. MR0127989 https://doi.org/10.1017/s0305004100033193

MORAN, P. A. P. (1962). *The Statistical Processes of Evolutionary Theory*. Clarendon Press, Oxford.

MURA, M., FEILLET, C., BERTOLUSSO, R., DELAUNAY, F. and KIMMEL, M. (2019). Mathematical modelling reveals unexpected inheritance and variability patterns of cell cycle parameters in mammalian cells. *PLoS Comput. Biol.* **15** e1007054. https://doi.org/10.1371/journal.pcbi.1007054

NEE, S., MAY, R. M. and HARVEY, P. H. (1994). The reconstructed evolutionary process. *Philos. Trans. R. Soc. Lond. B* **344** 305–311.

NOWELL, P. C. (1976). The clonal evolution of tumor cell populations. *Science* **194** 23–28. https://doi.org/10.1126/science.959840

POPOVIC, L. (2004). Asymptotic genealogy of a critical branching process. *Ann. Appl. Probab.* **14** 2120–2148. MR2100386 https://doi.org/10.1214/105051604000000486

RANNALA, B. (1997). Gene genealogy in a population of variable size. *Heredity* **78** 417–423.

SARGSYAN, O. (2015). An analytical framework in the general coalescent tree setting for analyzing polymorphisms created by two mutations. *J. Math. Biol.* **70** 913–956. MR3306621 https://doi.org/10.1007/s00285-014-0785-8

SOTTORIVA, A., KANG, H., MA, Z., GRAHAM, T. A., SALOMON, M. P., ZHAO, J., MARJORAM, P., SIEGMUND, K., PRESS, M. F. et al. (2015). A Big Bang model of human colorectal tumor growth. *Nat. Genet.* **47** 209–216.

SOTTORIVA, A., SPITERI, I., PICCIRILLO, S. G. M., TOULOUMIS, A., COLLINS, V. P., MARIONI, J. C., CURTIS, C., WATTS, C. and TAVARÉ, S. (2013). Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc. Natl. Acad. Sci. USA* **110** 4009–4014.

THOMPSON, E. A. (1975). *Human Evolutionary Trees*. Cambridge Univ. Press, Cambridge.

TOMASETTI, C., VOGELSTEIN, B. and PARMIGIANI, G. (2013). Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proc. Natl. Acad. Sci. USA* **110** 1999–2004.

TURAJLIC, S., SOTTORIVA, A., GRAHAM, T. and SWANTON, C. (2019). Resolving genetic heterogeneity in cancer. *Nat. Rev. Genet.* **20** 404–416.

WATTERSON, G. A. (1996). Motoo Kimura's use of diffusion theory in population genetics. *Theor. Popul. Biol.* **49** 154–188.

WILLIAMS, M. J., WERNER, B., HEIDE, T., CURTIS, C., BARNES, C. P., SOTTORIVA, A. and GRAHAM, T. A. (2018). Quantification of subclonal selection in cancer from bulk sequencing data. *Nat. Genet.* **50** 895–903.

ZAHN, H., STEIF, A., LAKS, E., EIREW, P., VANINSBERGHE, M., SHAH, S. P., APARICIO, S. and HANSEN, C. L. (2017). Scalable whole-genome single-cell library preparation without preamplification. *Nat. Methods* **14** 167–173. https://doi.org/10.1038/nmeth.4140

Statistical Science
2020, Vol. 35, No. 1, 145–157
https://doi.org/10.1214/19-STS763

Maximum Independent Component Analysis with Application to EEG Data

Ruosi Guo, Chunming Zhang and Zhengjun Zhang

Abstract. In many scientific disciplines, finding hidden influential factors behind observational data is essential but challenging. The majority of existing approaches, such as the independent component analysis (ICA), rely on linear transformation, that is, true signals are linear combinations of hidden components. Motivated from analyzing nonlinear temporal signals in neuroscience, genetics, and finance, this paper proposes the "maximum independent component analysis" (MaxICA), based on max-linear combinations of components. In contrast to existing methods, MaxICA benefits from focusing on significant major components while filtering out ignorable components. A major tool for parameter learning of MaxICA is an augmented genetic algorithm, consisting of three schemes for the elite weighted sum selection, randomly combined crossover, and dynamic mutation. Extensive empirical evaluations demonstrate the effectiveness of MaxICA in either extracting max-linearly combined essential sources in many applications or supplying a better approximation for nonlinearly combined source signals, such as EEG recordings analyzed in this paper.

Key words and phrases: Blind source separation, genetic algorithm, image analysis, nonlinear time series, optimization.

REFERENCES

[1] ANDRZEJAK, R. G., SCHINDLER, K. and RUMMEL, C. (2012). Nonrandomness, nonlinear dependence, and nonstationarity of electroencephalographic recordings from epilepsy patients. *Phys. Rev. E* **86** 046206.

[2] ARTONI, F., DELORME, A. and MAKEIG, S. (2018). Applying dimension reduction to EEG data by principal component analysis reduces the quality of its subsequent independent component decomposition. *NeuroImage* **175** 176–187. https://doi.org/10.1016/j.neuroimage.2018.03.016

[3] BRONKHORST, A. W. (2000). The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. *Acta Acust. Acust.* **86** 117–128.

[4] CHEN, A. and BICKEL, P. J. (2006). Efficient independent component analysis. *Ann. Statist.* **34** 2825–2855. MR2329469 https://doi.org/10.1214/009053606000000939

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

[6] CRYER, J. D. and CHAN, K.-S. (2008). Time series regression models. In *Time Series Analysis with Applications in R*, Chapter 11 249–276.

[7] DELORME, A., MAKEIG, S., FABRE-THORPE, M. and SEJNOWSKI, T. (2002). From single-trial eeg to brain area dynamics. *Neurocomputing* **44** 1057–1064.

[8] DELORME, A., ROUSSELET, G. A., MACE, M. J.-M. and FABRE-THORPE, M. (2004). Interaction of top-down and bottom-up processing in the fast visual analysis of natural scenes. *Cogn. Brain Res.* **19** 103–113.

[9] DRISS, I., MOUSS, K. N. and LAGGOUN, A. (2015). A new genetic algorithm for exible job-shop scheduling problems. *J. Mech. Sci. Technol.* **29** 1273–1281.

[10] FABRE-THORPE, M., DELORME, A., MARLOT, C. and THORPE, S. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *J. Cogn. Neurosci.* **13** 171–180.

[11] FAN, J. and YAO, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods. Springer Series in Statistics*. Springer, New York. MR1964455 https://doi.org/10.1007/b97702

[12] GAO, S., SHI, L. and ZHANG, Z. (2018). A peak-over-threshold search method for global optimization. *Automatica J. IFAC* **89** 83–91. MR3762035 https://doi.org/10.1016/j.automatica.2017.12.002

[13] GOLDBERG, D. E. and DEB, K. (1991). A comparative analysis of selection schemes used in genetic algorithms. In *Foundations of Genetic Algorithms* (*Bloomington, IN*, 1990) 69–93. Morgan Kaufmann, San Mateo, CA. MR1147425

Ruosi Guo is a Ph.D. graduate, Department of Statistics, University of Wisconsin, Madison, Wisconsin 53706, USA (e-mail: rguo25@wisc.edu). Chunming Zhang is Professor, Department of Statistics, University of Wisconsin, Madison, Wisconsin 53706, USA (e-mail: cmzhang@stat.wisc.edu). Zhengjun Zhang is Professor, Department of Statistics, University of Wisconsin, Madison, Wisconsin 53706, USA (e-mail: zjz@stat.wisc.edu).

[14] GUO, R., ZHANG, C. and ZHANG, Z. (2020). Supplement to "Maximum independent component analysis with application to EEG data." https://doi.org/10.1214/19-STS763SUPP.

[15] GUO, Y. and TANG, L. (2013). A hierarchical model for probabilistic independent component analysis of multi-subject fMRI studies. *Biometrics* **69** 970–981. MR3146792 https://doi.org/10.1111/biom.12068

[16] HASSAN, I. I. (2015). Combined crossover operator. *Res. J. Appl. Sci.* **10** 75–79.

[17] HAUPT, R. L. and HAUPT, S. E. (1998). *Practical Genetic Algorithms. A Wiley-Interscience Publication*. Wiley, New York. MR1491878

[18] HUANG, H., LU, J., WU, J., DING, Z., CHEN, S., DUAN, L., CUI, J., CHEN, F., KANG, D. et al. (2018). Tumor tissue detection using blood-oxygen-level-dependent functional mri based on independent component analysis. *Sci. Rep.* **8**. Article number: 1223.

[19] HYVÄRINEN, A., KARHUNEN, J. and OJA, E. (2001). *Independent Component Analysis*. Wiley, New York.

[20] HYVÄRINEN, A. and OJA, E. (2000). Independent component analysis: Algorithms and applications. *Neural Netw.* **13** 411–430.

[21] JEBARI, K. and MADIAFI, M. (2013). Selection methods for genetic algorithms. *Int. J. Emerg. Sci.* **3** 333–344.

[22] KASSOUF, A., BOUVERESSE, D. J.-R. and RUTLEDGE, D. N. (2018). Determination of the optimal number of components in independent components analysis. *Talanta* **179** 538–545. https://doi.org/10.1016/j.talanta.2017.11.051

[23] NASCIMENTO, M., SILVA, F., SAFADI, T., NASCIMENTO, A. C. C., FERREIRA, T. E. M., BARROSO, L. M. A., AZEVEDO, C. F., GUIMAR AES, S. E. and SER AO, N. V. (2017). Independent component analysis (ica) based-clustering of temporal rna-seq data. *PLoS ONE* **12** e0181195.

[24] NAVEAU, P., ZHANG, Z. and ZHU, B. (2011). An extension of max autoregressive models. *Stat. Interface* **4** 253–266. MR2812820 https://doi.org/10.4310/SII.2011.v4.n2.a19

[25] OBITKO, M., SLAVIK, P. and WALTER, P. (1998). *Introduction to genetic algorithms*. http://www.obitko.com/tutorials/genetic-algorithms/index.php.

[26] PAES, F. G., PESSOA, A. A. and VIDAL, T. (2017). A hybrid genetic algorithm with decomposition phases for the unequal area facility layout problem. *European J. Oper. Res.* **256** 742–756. MR3549773 https://doi.org/10.1016/j.ejor.2016.07.022

[27] SHOEB, A. H. (2009). Application of machine learning to epileptic seizure onset detection and treatment Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

[28] SOKOL, A., MAATHUIS, M. H. and FALKEBORG, B. (2014). Quantifying identifiability in independent component analysis. *Electron. J. Stat.* **8** 1438–1459. MR3263128 https://doi.org/10.1214/14-EJS932

[29] THIERENS, D. (2002). Adaptive mutation rate control schemes in genetic algorithms. In *Evolutionary Computation*, 2002. *CEC'*02. *Proceedings of the* 2002 *Congress on* **1** 980–985. IEEE, New York.

[30] UMBARKAR, A. and SHETH, P. (2015). Crossover operators in genetic algorithms: A review. *ICTACT J. Soft Comput.* **06** 1083–1092.

[31] ZHANG, C., CHAI, Y., GUO, X., GAO, M., DEVILBISS, D. and ZHANG, Z. (2016). Statistical learning of neuronal functional connectivity. *Technometrics* **58** 350–359. MR3520664 https://doi.org/10.1080/00401706.2016.1142904

INSTITUTE OF MATHEMATICAL STATISTICS

(Organized September 12, 1935)

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

IMS OFFICERS

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

President-Elect: Regina Y. Liu, Department of Statistics, Rutgers University, Piscataway, New Jersey 08854-8019, USA

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

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

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

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

IMS EDITORS

The Annals of Statistics. *Editors*: Richard J. Samworth, Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, Cambridge, CB3 0WB, UK. Ming Yuan, Department of Statistics, Columbia University, New York, NY 10027, USA

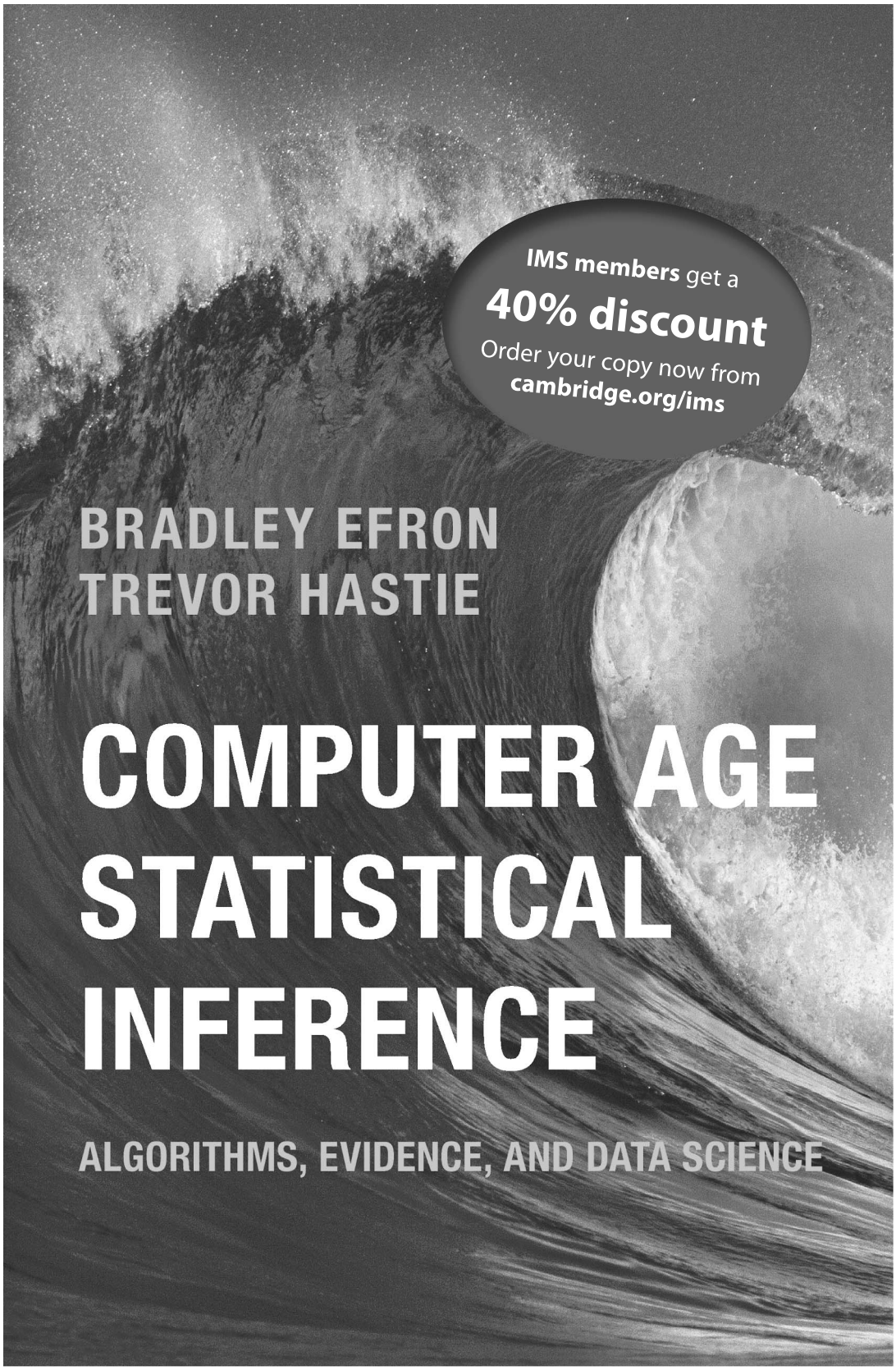
The Annals of Applied Statistics. *Editor-in-Chief*: Karen Kafadar, Department of Statistics, University of Virginia, Heidelberg Institute for Theoretical Studies, Charlottesville, VA 22904-4135, USA

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

The Annals of Applied Probability. *Editors*: François Delarue, Laboratoire J. A. Dieudonné, Université de Nice Sophia-Antipolis, France-06108 Nice Cedex 2. Peter Friz, Institut für Mathematik, Technische Universität Berlin, 10623 Berlin, Germany and Weierstrass-Institut für Angewandte Analysis und Stochastik, 10117 Berlin, Germany

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

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


IMS members get a
40% discount
Order your copy now from
cambridge.org/ims
BRADLEY EFRON
TREVOR HASTIE
COMPUTER AGE
STATISTICAL
INFERENCE
ALGORITHMS, EVIDENCE, AND DATA SCIENCE