STATISTICAL SCIENCE

Volume 34, Number 2 May 2019

Statistical Science **[ISSN 0883-4237 (print); ISSN 2168-8745 (online)]**, Volume 34, Number 2, May 2019. 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 34, Number 2 (177–359) May 2019**

Volume 34 **Number 2** **May 2019**

Statistical Science
2019, Vol. 34, No. 2, 177–201
https://doi.org/10.1214/18-STS674

Bayes, Oracle Bayes and Empirical Bayes

Bradley Efron

Abstract. This article concerns the Bayes and frequentist aspects of empirical Bayes inference. Some of the ideas explored go back to Robbins in the 1950s, while others are current. Several examples are discussed, real and artificial, illustrating the two faces of empirical Bayes methodology: "oracle Bayes" shows empirical Bayes in its most frequentist mode, while "finite Bayes inference" is a fundamentally Bayesian application. In either case, modern theory and computation allow us to present a sharp finite-sample picture of what is at stake in an empirical Bayes analysis.

Key words and phrases: Finite Bayes inference, g-modeling, relevance, empirical Bayes regret.

REFERENCES

BROWN, L. D. and GREENSHTEIN, E. (2009). Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *Ann. Statist.* **37** 1685–1704. MR2533468

CARLIN, B. P. and GELFAND, A. E. (1991). A sample reuse method for accurate parametric empirical Bayes confidence intervals. *J. Roy. Statist. Soc. Ser. B* **53** 189–200.

DEELY, J. J. and LINDLEY, D. V. (1981). Bayes empirical Bayes. *J. Amer. Statist. Assoc.* **76** 833–841. MR0650894

EFRON, B. (1996). Empirical Bayes methods for combining likelihoods. *J. Amer. Statist. Assoc.* **91** 538–565. MR1395725

EFRON, B. (2010). *Large-Scale Inference*: *Empirical Bayes Methods for Estimation*, *Testing*, *and Prediction*. *Institute of Mathematical Statistics* (*IMS*) *Monographs* **1**. Cambridge Univ. Press, Cambridge. MR2724758

EFRON, B. (2011). Tweedie's formula and selection bias. *J. Amer. Statist. Assoc.* **106** 1602–1614. MR2896860

EFRON, B. (2014). Two modeling strategies for empirical Bayes estimation. *Statist. Sci.* **29** 285–301. MR3264543

EFRON, B. (2016). Empirical Bayes deconvolution estimates. *Biometrika* **103** 1–20. MR3465818

EFRON, B. and HASTIE, T. (2016). *Computer Age Statistical Inference*: *Algorithms*, *Evidence*, *and Data Science*. *Institute of Mathematical Statistics* (*IMS*) *Monographs* **5**. Cambridge Univ. Press, New York. MR3523956

EFRON, B. and MORRIS, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators. II. The empirical Bayes case. *J. Amer. Statist. Assoc.* **67** 130–139. MR0323015

FISHER, R. A., CORBET, A. S. and WILLIAMS, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* **12** 42–58.

GOOD, I. J. and TOULMIN, G. H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* **43** 45–63. MR0077039

GU, J. and KOENKER, R. (2016). On a problem of Robbins. *Int. Stat. Rev.* **84** 224–244. MR3537154

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

JIANG, W. and ZHANG, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist.* **37** 1647–1684. MR2533467

KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Stat.* **27** 887–906. MR0086464

KOENKER, R. and MIZERA, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *J. Amer. Statist. Assoc.* **109** 674–685. MR3223742

LAIRD, N. (1978). Nonparametric maximum likelihood estimation of a mixed distribution. *J. Amer. Statist. Assoc.* **73** 805–811. MR0521328

LAIRD, N. M. and LOUIS, T. A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *J. Amer. Statist. Assoc.* **82** 739–757. MR0909979

MORRIS, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *J. Amer. Statist. Assoc.* **78** 47–65. MR0696849

ROBBINS, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 1950 131–148. Univ. California Press, Berkeley. MR0044803

ROBBINS, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathemati-*

Bradley Efron is Professor, Department of Statistics, Stanford University, Sequoia Hall, 390 Serra Mall, Stanford, California 94305-4020, USA (e-mail: brad@stat.stanford.edu).

cal Statistics and Probability, 1954–1955, *Vol. I* 157–163. Univ. California Press, Berkeley. MR0084919

SCHWARTZMAN, A., DOUGHERTY, R. F. and TAYLOR, J. E. (2005). Cross-subject comparison of principal diffusion direction maps. *Magn. Reson. Med.* **53** 1423–1431.

ZHANG, C.-H. (2003). Compound decision theory and empirical Bayes methods. *Ann. Statist.* **31** 379–390. MR1983534

Statistical Science
2019, Vol. 34, No. 2, 202–205
https://doi.org/10.1214/19-STS704

Comment: Bayes, Oracle Bayes, and Empirical Bayes

Thomas A. Louis

REFERENCES

BELL, W. R., DATTA, G. S. and GHOSH, M. (2013). Benchmarking small area estimators. *Biometrika* **100** 189–202. MR3034332

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

BROWNSTEIN, N. C., LOUIS, T. A., O'HAGAN, A. and PENDERGAST, J. (2019). The role of expert judgment in statistical inference and evidence-based decision-making. *Amer. Statist.* **73** 56–68. MR3925709

CARLIN, B. P. and GELFAND, A. E. (1991). A sample reuse method for accurate parametric empirical Bayes confidence intervals. *J. Roy. Statist. Soc. Ser. B* **53** 189–200.

CARLIN, B. P. and LOUIS, T. A. (2009). *Bayesian Methods for Data Analysis*, 3rd ed. Chapman and Hall/CRC Press, Boca Raton, FL.

EFRON, B. (1986). Why isn't everyone a Bayesian? *Amer. Statist.* **40** 1–11. MR0828575

EFRON, B. (2014). Two modeling strategies for empirical Bayes estimation. *Statist. Sci.* **29** 285–301. MR3264543

LAIRD, N. M. (1982). Empirical Bayes Estimates Using the Nonparametric Maximum Likelihood Estimate for the Prior. *J. Stat. Comput. Simul.* **15** 211–220.

LAIRD, N. M. and LOUIS, T. A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *J. Amer. Statist. Assoc.* **82** 739–757. MR0909979

LAIRD, N. M. and LOUIS, T. A. (1991). Smoothing the nonparametric estimate of a prior distribution by roughening: A computational study. *Comput. Statist. Data Anal.* **12** 27–37. MR1131643

MORRIS, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *J. Amer. Statist. Assoc.* **78** 47–65. MR0696849

O'HAGAN, A. (2019). Expert knowledge elicitation: Subjective but scientific. *Amer. Statist.* **73** 69–81. MR3925710

PADDOCK, S. M. and LOUIS, T. A. (2011). Percentile-based empirical distribution function estimates for performance evaluation of healthcare providers. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **60** 575–589. MR2829191

PADDOCK, S. M., RIDGEWAY, G., LIN, R. and LOUIS, T. A. (2006). Flexible distributions of triple-goal estimates in two-stage hierarchical models. *Comput. Statist. Data Anal.* **50** 3243–3262. MR2239666

ROBBINS, H. (1983). Some thoughts on empirical Bayes estimation. *Ann. Statist.* **11** 713–723. MR0707923

SHEN, W. and LOUIS, T. A. (1998). Triple-goal estimates in two-stage hierarchical models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 455–471. MR1616061

SHEN, W. and LOUIS, T. A. (1999). Empirical Bayes estimation via the smoothing by roughening approach. *J. Comput. Graph. Statist.* **8** 800–823. MR1748968

Thomas A. Louis is Professor Emeritus, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, Maryland 21205, USA (e-mail: tlouis@jhu.edu).

Statistical Science
2019, Vol. 34, No. 2, 206–208
https://doi.org/10.1214/19-STS705

Comment: Bayes, Oracle Bayes, and Empirical Bayes

Nan Laird

REFERENCES

DERSIMONIAN, R. and LAIRD, N. M. (1983). Evaluating the effect of coaching on SAT scores: A meta-analysis. *Harv. Educ. Rev.* **53** 1–16.

EFRON, B. (1996). Empirical Bayes methods for combining likelihoods. *J. Amer. Statist. Assoc.* **91** 538–565. MR1395725

GILBERT, J. P., MCPEEK, B. and MOSTELLER, F. (1977). Progress in surgery and anesthesia: An evaluation of innovative therapy. In *Costs, Benefits and Risks of Surgery* (B. A. Barnes, J. P. Bunker and F. Mosteller, eds.) Oxford Univ. Press, New York.

LAIRD, N. M. and LOUIS, T. A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *J. Amer. Statist. Assoc.* **82** 739–757. MR0909979

MORRIS, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *J. Amer. Statist. Assoc.* **78** 47–65. MR0696849

MOSTELLER, F. and WALLACE, D. L. (1964). *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, Reading, MA. MR0175668

Nan Laird is the Harvey V. Fineberg Research Professor of Biostatistics at Harvard School of Public Health, 677 Huntington Ave, Boston Massachusetts 02115, USA (e-mail: laird@biostat.hsph.edu).

Statistical Science
2019, Vol. 34, No. 2, 209–213
https://doi.org/10.1214/19-STS706

Comment: Minimalist g-Modeling

Roger Koenker and Jiaying Gu

Abstract. Efron's elegant approach to g-modeling for empirical Bayes problems is contrasted with an implementation of the Kiefer–Wolfowitz nonparametric maximum likelihood estimator for mixture models for several examples. The latter approach has the advantage that it is free of tuning parameters and consequently provides a relatively simple complementary method.

Key words and phrases: Nonparametric maximum likelihood, mixture model, convex optimization.

REFERENCES

ANDERSEN, E. D. (2010). The Mosek Optimization Tools Manual, Version 6.0. Available from: http://www.mosek.com.

DEELY, J. J. and LINDLEY, D. V. (1981). Bayes empirical Bayes. *J. Amer. Statist. Assoc.* **76** 833–841. MR0650894

EFRON, B. (2010). *Large-Scale Inference*: *Empirical Bayes Methods for Estimation*, *Testing*, *and Prediction*. *Institute of Mathematical Statistics* (*IMS*) *Monographs* **1**. Cambridge Univ. Press, Cambridge. MR2724758

EFRON, B. (2016). Empirical Bayes deconvolution estimates. *Biometrika* **103** 1–20. MR3465818

FRIBERG, H. A. (2012). Users Guide to the R-to-Mosek Interface. Available at http://rmosek.r-forge.r-project.org.

HECKMAN, J. and SINGER, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* **52** 271–320. MR0735309

JIANG, W. and ZHANG, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist.* **37** 1647–1684. MR2533467

KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Stat.* **27** 887–906. MR0086464

KOENKER, R. and GU, J. (2015). REBayes: An R Package for Empirical Bayes Methods. Available from https://cran.r-project.org/package=REBayes.

KOENKER, R. and MIZERA, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *J. Amer. Statist. Assoc.* **109** 674–685. MR3223742

LAIRD, N. (1978). Nonparametric maximum likelihood estimation of a mixed distribution. *J. Amer. Statist. Assoc.* **73** 805–811. MR0521328

LINDLEY, D. V. and SMITH, A. (1995). A conversation with Dennis Lindley. *Statist. Sci.* **10** 305–319.

NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265. MR1823804

ROBBINS, H. (1950). A generalization of the method of maximum likelihood; estimating a mixing distribution (abstract). *Ann. Math. Stat.* **21** 314–315.

ROSS, G. J. and MARKWICK, D. (2018). Dirichletprocess: An R Package for Fitting Complex Bayesian Nonparametric Models. Available at https://cran.r-project.org/web/packages/dirichletprocess/vignettes/dirichletprocess.pdf.

STEFANSKI, L. and CARROLL, R. J. (1990). Deconvoluting kernel density estimators. *Statistics* **21** 169–184. MR1054861

Roger Koenker is Honorary Professor of Economics, Department of Economics, University College London, London, WC1H OAX, United Kingdom (e-mail: r.koenker@ucl.ac.uk). Jiaying Gu is Assistant Professor of Economics, Department of Economics, University of Toronto, Toronto, Ontario, M5S 3G7, Canada (e-mail: jiaying.gu@utoronto.ca).

Statistical Science
2019, Vol. 34, No. 2, 214–218
https://doi.org/10.1214/19-STS707

Comment: Bayes, Oracle Bayes and Empirical Bayes

Aad van der Vaart

REFERENCES

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

[2] BLACKWELL, D. (1973). Discreteness of Ferguson selections. *Ann. Statist.* **1** 356–358. MR0348905

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

[4] CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97** 465–480. MR2650751

[5] CASTILLO, I. (2012). A semiparametric Bernstein–von Mises theorem for Gaussian process priors. *Probab. Theory Related Fields* **152** 53–99. MR2875753

[6] CASTILLO, I. (2012). A semiparametric Bernstein–von Mises theorem for Gaussian process priors. *Probab. Theory Related Fields* **152** 53–99. MR2875753

[7] CASTILLO, I. and MISMER, R. (2018). Empirical Bayes analysis of spike and slab posterior distributions. *Electron. J. Stat.* **12** 3953–4001. MR3885271

[8] CASTILLO, I. and ROUSSEAU, J. (2015). A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *Ann. Statist.* **43** 2353–2383. MR3405597

[9] CASTILLO, I., SCHMIDT-HIEBER, J. and VAN DER VAART, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.* **43** 1986–2018. MR3375874

[10] CASTILLO, I. and VAN DER VAART, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Ann. Statist.* **40** 2069–2101. MR3059077

[11] CEREDA, G. (2017). Current challenges in statistical DNA evidence evaluation. Leiden Univ.

[12] COX, D. D. (1993). An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.* **21** 903–923. MR1232525

[13] DAVIDE, C. (2018). Statistical 'rock star' wins coveted international prize. *Nature*. Published online: 12 November 2018; DOI:10.1038/d41586-018-07395-w.

[14] DE BLASI, P., LIJOI, A. and PRÜNSTER, I. (2013). An asymptotic analysis of a class of discrete nonparametric priors. *Statist. Sinica* **23** 1299–1321. MR3114715

[15] ESCOBAR, M. D. (1994). Estimating normal means with a Dirichlet process prior. *J. Amer. Statist. Assoc.* **89** 268–277. MR1266299

[16] FAVARO, S., LIJOI, A. and PRÜNSTER, I. (2012). Asymptotics for a Bayesian nonparametric estimator of species variety. *Bernoulli* **18** 1267–1283. MR2995795

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

[18] FERGUSON, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In *Recent Advances in Statistics* 287–302. Academic Press, New York. MR0736538

[19] GHOSAL, S. and VAN DER VAART, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.* **35** 697–723. MR2336864

[20] GHOSAL, S. and VAN DER VAART, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. *Cambridge Series in Statistical and Probabilistic Mathematics* **44**. Cambridge Univ. Press, Cambridge. MR3587782

[21] GHOSAL, S. and VAN DER VAART, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.* **29** 1233–1263. MR1873329

[22] GRIFFIN, J. E. and BROWN, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Anal.* **5** 171–188. MR2596440

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

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

[25] JAMES, L. F. (2008). Large sample asymptotics for the two-parameter Poisson–Dirichlet process. In *Pushing the Limits of Contemporary Statistics*: *Contributions in Honor of Jayanta K. Ghosh*. *Inst. Math. Stat.* (*IMS*) *Collect.* **3** 187–199. IMS, Beachwood, OH. MR2459225

[26] JARA, A. (2007). Applied Bayesian non-and semi-parametric inference using dppackage. *R News* **7** 17–26.

[27] JARA, A., HANSON, T., QUINTANA, F., MUELLER, P. and ROSNER, G. (2015). Package DPpackage.

[28] JIANG, W. and ZHANG, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist.* **37** 1647–1684. MR2533467

[29] JOHNSTONE, I. M. and SILVERMAN, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* **32** 1594–1649. MR2089135

[30] KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Stat.* **27** 887–906. MR0086464

Aad van der Vaart is Professor of Stochastics, Mathematical Institute, Niels Bohrweg 3, Leiden, Netherlands (e-mail: avdvaart@math.leidenuniv.nl; URL: www.math.leidenuniv.nl/~avdvaart).

[31] KIM, Y. (2006). The Bernstein–von Mises theorem for the proportional hazard model. *Ann. Statist.* **34** 1678–1700. MR2283713

[32] KNAPIK, B. T., SZABÓ, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2016). Bayes procedures for adaptive inference in inverse problems for the white noise model. *Probab. Theory Related Fields* **164** 771–813. MR3477780

[33] KOENKER, R. and MIZERA, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *J. Amer. Statist. Assoc.* **109** 674–685. MR3223742

[34] LIJOI, A., MENA, R. H. and PRÜNSTER, I. (2007). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika* **94** 769–786. MR2416792

[35] LIJOI, A., PRÜNSTER, I. and WALKER, S. G. (2008). Bayesian nonparametric estimators derived from conditional Gibbs structures. *Ann. Appl. Probab.* **18** 1519–1547. MR2434179

[36] LINDSAY, B. (1995). Mixture models: Theory, geometry and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics* i–163. IMS, Hayward, CA.

[37] LO, A. Y. (1983). Weak convergence for Dirichlet processes. *Sankhyā Ser. A* **45** 105–111. MR0749358

[38] MCCLOSKEY, J. W. T. (1965). *A Model for the Distribution of Individuals by Species in an Environment*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.), Michigan State Univ. MR2615013

[39] MILLER, J. W. and HARRISON, M. T. (2014). Inconsistency of Pitman–Yor process mixtures for the number of components. *J. Mach. Learn. Res.* **15** 3333–3370. MR3277163

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

[41] MURPHY, S. A. and VAN DER VAART, A. W. (2000). On profile likelihood. *J. Amer. Statist. Assoc.* **95** 449–485. MR1803168

[42] NEAL, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* **9** 249–265. MR1823804

[43] PERMAN, M., PITMAN, J. and YOR, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probab. Theory Related Fields* **92** 21–39. MR1156448

[44] PETRONE, S., RIZZELLI, S., ROUSSEAU, J. and SCRICCIOLO, C. (2014). Empirical Bayes methods in classical and Bayesian inference. *Metron* **72** 201–215. MR3233149

[45] PFANZAGL, J. (1988). Consistency of maximum likelihood estimators for certain nonparametric families, in particular: Mixtures. *J. Statist. Plann. Inference* **19** 137–158. MR0944202

[46] PITMAN, J. (1995). Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields* **102** 145–158. MR1337249

[47] PITMAN, J. (1996). Some developments of the Blackwell–MacQueen urn scheme. In *Statistics, Probability and Game Theory. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **30** 245–267. IMS, Hayward, CA. MR1481784

[48] POLSON, N. G. and SCOTT, J. G. (2011). Shrink globally, act locally: Sparse Bayesian regularization and prediction. In *Bayesian Statistics* 9 (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) 501–538. Oxford Univ. Press, Oxford. MR3204017

[49] RAY, K. (2017). Adaptive Bernstein–von Mises theorems in Gaussian white noise. *Ann. Statist.* **45** 2511–2536. MR3737900

[50] RIVOIRARD, V. and ROUSSEAU, J. (2012). Bernstein–von Mises theorem for linear functionals of the density. *Ann. Statist.* **40** 1489–1523. MR3015033

[51] ROČKOVÁ, V. (2018). Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *Ann. Statist.* **46** 401–437. MR3766957

[52] SCRICCIOLO, C. (2014). Adaptive Bayesian density estimation in L^p-metrics with Pitman–Yor or normalized inverse-Gaussian process kernel mixtures. *Bayesian Anal.* **9** 475–520. MR3217004

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

[54] SHEN, W., TOKDAR, S. T. and GHOSAL, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika* **100** 623–640. MR3094441

[55] SNIEKERS, S. and VAN DER VAART, A. (2015). Adaptive Bayesian credible sets in regression with a Gaussian process prior. *Electron. J. Stat.* **9** 2475–2527. MR3425364

[56] SZABÓ, B., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *Ann. Statist.* **43** 1391–1428. MR3357861

[57] VAN DER PAS, S., SZABÓ, B. and VAN DER VAART, A. (2017). Adaptive posterior contraction rates for the horseshoe. *Electron. J. Stat.* **11** 3196–3225. MR3705450

[58] VAN DER PAS, S., SZABÓ, B. and VAN DER VAART, A. (2017). Uncertainty quantification for the horseshoe (with discussion). *Bayesian Anal.* **12** 1221–1274. MR3724985

[59] VAN DER VAART, A. (1991). On differentiable functionals. *Ann. Statist.* **19** 178–204. MR1091845

[60] VAN DER VAART, A. (2002). Semiparametric statistics. In *Lectures on Probability Theory and Statistics* (*Saint-Flour*, 1999). *Lecture Notes in Math.* **1781** 331–457. Springer, Berlin. MR1915446

[61] WALKER, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Comm. Statist. Simulation Comput.* **36** 45–54. MR2370888

Statistical Science
2019, Vol. 34, No. 2, 219–223
https://doi.org/10.1214/19-STS708

Comment: Empirical Bayes Interval Estimation

Wenhua Jiang

Abstract. This is a contribution to the discussion of the enlightening paper by Professor Efron. We focus on empirical Bayes interval estimation. We discuss the oracle interval estimation rules, the empirical Bayes estimation of the oracle rule and the computation. Some numerical results are reported.

Key words and phrases: Empirical Bayes, interval estimation, oracle rule, generalized MLE.

REFERENCES

EFRON, B. (2014). Two modeling strategies for empirical Bayes estimation. *Statist. Sci.* **29** 285–301. MR3264543

EFRON, B. (2016). Empirical Bayes deconvolution estimates. *Biometrika* **103** 1–20. MR3465818

HANNAN, J. F. and ROBBINS, H. (1955). Asymptotic solutions of the compound decision problem for two completely specified distributions. *Ann. Math. Stat.* **26** 37–51. MR0067444

JIANG, W. and ZHANG, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist.* **37** 1647–1684. MR2533467

JIANG, W. and ZHANG, C.-H. (2016). Generalized likelihood ratio test for normal mixtures. *Statist. Sinica* **26** 955–978. MR3559938

KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Stat.* **27** 887–906. MR0086464

KOENKER, R. and MIZERA, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *J. Amer. Statist. Assoc.* **109** 674–685. MR3223742

LAIRD, N. M. and LOUIS, T. A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *J. Amer. Statist. Assoc.* **82** 739–757. MR0909979

MORRIS, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *J. Amer. Statist. Assoc.* **78** 47–65. MR0696849

ROBBINS, H. (1951). Asymptotically subminimax solutions of compound decision problems. In *Proc. Second Berkeley Symp. Math. Statist. Probab.* **1** 131–148. Univ. California Press, Berkeley, CA. MR0044803

ROBBINS, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955 **1** 157–163. Univ. California Press, Berkeley and Los Angeles. MR0084919

VARDI, Y., SHEPP, L. A. and KAUFMAN, L. (1985). A statistical model for positron emission tomography. *J. Amer. Statist. Assoc.* **80** 8–20. MR0786595

Wenhua Jiang is Young Investigator, School of Mathematical Sciences, Fudan University and Shanghai Center for Mathematical Sciences, 220 Handan Road, Shanghai 200433, China (e-mail: wenhuajiang@yahoo.com).

Statistical Science
2019, Vol. 34, No. 2, 224–228
https://doi.org/10.1214/19-STS709

Comment: Empirical Bayes, Compound Decisions and Exchangeability

Eitan Greenshtein and Ya'acov Ritov

Abstract. We present some personal reflections on empirical Bayes/compound decision (EB/CD) theory following Efron (2019). In particular, we consider the role of exchangeability in the EB/CD theory and how it can be achieved when there are covariates. We also discuss the interpretation of EB/CD confidence interval, the theoretical efficiency of the CD procedure, and the impact of sparsity assumptions.

Key words and phrases: f-modeling, g-modeling, sparsity.

REFERENCES

BROWN, L. D. and GREENSHTEIN, E. (2009). Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *Ann. Statist.* **37** 1684–1704. MR2533468

BROWN, L. D., GREENSHTEIN, E. and RITOV, Y. (2013). The Poisson compound decision problem revisited. *J. Amer. Statist. Assoc.* **108** 741–749. MR3174656

COHEN, N., GREENSHTEIN, E. and RITOV, Y. (2013). Empirical Bayes in the presence of explanatory variables. *Statist. Sinica* **23** 333–357. MR3076170

EFRON, B. and MORRIS, C. (1973). Stein's estimation rule and its competitors—An empirical Bayes approach. *J. Amer. Statist. Assoc.* **68** 117–130. MR0388597

FAY, R. E. III and HERRIOT, R. A. (1979). Estimates of income for small places: An application of James–Stein procedures to census data. *J. Amer. Statist. Assoc.* **74** 269–277. MR0548019

GREENSHTEIN, E., MANSURA, A. and RITOV, Y. (2018). Nonparametric empirical Bayes improvement of common shrikage estimators. Submitted.

GREENSHTEIN, E., PARK, J. and RITOV, Y. (2008). Estimating the mean of high valued observations in high dimensions. *J. Stat. Theory Pract.* **2** 407–418. MR2528789

GREENSHTEIN, E. and RITOV, Y. (2009). Asymptotic efficiency of simple decisions for the compound decision problem. In *Optimality*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **57** 266–275. IMS, Beachwood, OH. MR2681676

HANNAN, J. F. and ROBBINS, H. (1955). Asymptotic solutions of the compound decision problem for two completely specified distributions. *Ann. Math. Stat.* **26** 37–51. MR0067444

JIANG, W. and ZHANG, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist.* **37** 1647–1684.

KOENKER, R. and MIZERA, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *J. Amer. Statist. Assoc.* **109** 674–685.

SPEARMAN, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology* **15** 72–101.

WEINSTEIN, A., MA, Z., BROWN, L. D. and ZHANG, C.-H. (2018). Group-linear empirical Bayes estimates for a heteroscedastic normal mean. *J. Amer. Statist. Assoc.* **113** 698–710. MR3832220

Eitan Greenshtein, Ph.D., is with the Israel Central Bureau of Statistics, Kanfei Nesharim 66, 9546456 Jerusalem, Israel (e-mail: eitan.greenshtein@gmail.com). Ya'acov Ritov is Professor, Department of Statistics, University of Michigan, 1085 South University, Ann Arbor, Michigan 48109-1107, USA, and The Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem, Edmund J. Safra Campus, 91904 Jerusalem, Israel (e-mail: yritov@umich.edu).

Statistical Science
2019, Vol. 34, No. 2, 229–233
https://doi.org/10.1214/19-STS710

Comment: Variational Autoencoders as Empirical Bayes

Yixin Wang, Andrew C. Miller and David M. Blei

REFERENCES

BENGIO, Y., LAUFER, E., ALAIN, G. and YOSINSKI, J. (2014). Deep generative stochastic networks trainable by backprop. In *International Conference on Machine Learning* 226–234.

BERGLUND, M., RAIKO, T., HONKALA, M., KÄRKKÄINEN, L., VETEK, A. and KARHUNEN, J. T. (2015). Bidirectional recurrent neural networks as generative models. In *Advances in Neural Information Processing Systems* 856–864.

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

EFRON, B. (2011). Tweedie's formula and selection bias. *J. Amer. Statist. Assoc.* **106** 1602–1614. MR2896860

EFRON, B. (2019). Bayes, oracle Bayes, and empirical Bayes. *Statist. Sci.* **34** 177–201.

JIANG, W., ZHANG, C.-H. et al. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist.* **37** 1647–1684. MR2533467

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

KINGMA, D. P. and WELLING, M. (2013). Auto-encoding variational Bayes. Arxiv preprint. Available at arXiv:1312.6114.

LAROCHELLE, H. and MURRAY, I. (2011). The neural autoregressive distribution estimator. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* 29–37.

REZENDE, D. J. and MOHAMED, S. (2015). Variational inference with normalizing flows. Arxiv preprint. Available at arXiv:1505.05770.

SCHUSTER, M. and PALIWAL, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **45** 2673–2681.

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

WANG, Y. and BLEI, D. M. (2018). Frequentist consistency of variational Bayes. *J. Amer. Statist. Assoc.* 1–15.

Yixin Wang is Ph.D. student, Department of Statistics, Columbia University, New York, New York 10027, USA (e-mail: yixin.wang@columbia.edu). Andrew C. Miller is Postdoctoral Research Scientist, Data Science Institute, Columbia University, New York, New York 10027, USA (e-mail: am5171@columbia.edu). David M. Blei is Professor, Department of Statistics, Department of Computer Science and Data Science Institute, Columbia University, New York, New York 10027, USA (e-mail: david.blei@columbia.edu).

Statistical Science
2019, Vol. 34, No. 2, 234–235
https://doi.org/10.1214/19-STS674REJ

Rejoinder: Bayes, Oracle Bayes, and Empirical Bayes

Bradley Efron

REFERENCES

[1] EFRON, B. (2011). Tweedie's formula and selection bias. *J. Amer. Statist. Assoc.* **106** 1602–1614. MR2896860

[2] EFRON, B. (2016). Empirical Bayes deconvolution estimates. *Biometrika* **103** 1–20. MR3465818

[3] EFRON, B. and HASTIE, T. (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science. Institute of Mathematical Statistics (IMS) Monographs* **5**. Cambridge Univ. Press, New York. MR3523956

[4] EFRON, B. and NARASIMHAN, B. (2019). A *g*-modeling program for deconvolution and empirical Bayes estimation. *J. Stat. Softw.* To appear.

[5] GOOD, I. J. and TOULMIN, G. H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* **43** 45–63. MR0077039

[6] JIANG, W. and ZHANG, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist.* **37** 1647–1684. MR2533467

[7] KOENKER, R. and MIZERA, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *J. Amer. Statist. Assoc.* **109** 674–685. MR3223742

[8] LAIRD, N. (1978). Nonparametric maximum likelihood estimation of a mixed distribution. *J. Amer. Statist. Assoc.* **73** 805–811. MR0521328

Bradley Efron is Professor, Department of Statistics, Stanford University, Sequoia Hall, 390 Serra Mall, Stanford, California 94305-4020, USA (e-mail: brad@stat.stanford.edu).

Statistical Science
2019, Vol. 34, No. 2, 236–252
https://doi.org/10.1214/18-STS675

A Kernel Regression Procedure in the 3D Shape Space with an Application to Online Sales of Children's Wear

Gregorio Quintana-Ortí and Amelia Simó

Abstract. This paper is focused on kernel regression when the response variable is the shape of a 3D object represented by a configuration matrix of landmarks. Regression methods on this shape space are not trivial because this space has a complex finite-dimensional Riemannian manifold structure (non-Euclidean). Papers about it are scarce in the literature, the majority of them are restricted to the case of a single explanatory variable, and many of them are based on the approximated tangent space. In this paper, there are several methodological innovations. The first one is the adaptation of the general method for kernel regression analysis in manifold-valued data to the three-dimensional case of Kendall's shape space. The second one is its generalization to the multivariate case and the addressing of the curse-of-dimensionality problem. Finally, we propose bootstrap confidence intervals for prediction. A simulation study is carried out to check the goodness of the procedure, and a comparison with a current approach is performed. Then, it is applied to a 3D database obtained from an anthropometric survey of the Spanish child population with a potential application to online sales of children's wear.

Key words and phrases: Shape space, statistical shape analysis, Kernel regression, Fréchet mean, children's wear.

REFERENCES

AFSARI, B., TRON, R. and VIDAL, R. (2013). On the convergence of gradient descent for finding the Riemannian center of mass. *SIAM J. Control Optim.* **51** 2230–2260. MR3057324

ALVAREZ, F., BOLTE, J. and MUNIER, J. (2008). A unifying local convergence result for Newton's method in Riemannian manifolds. *Found. Comput. Math.* **8** 197–226. MR2407031

AZENCOTT, R. (1994). Deterministic and random deformations; applications to shape recognition. In *Conference at HSSS Workshop in Cortona, Italy*.

AZENCOTT, R., COLDEFY, F. and YOUNES, L. (1996). A distance for elastic matching in object recognition. In *Proceedings of* 12*th ICPR*.

BADDELEY, A. and MOLCHANOV, I. (1998). Averaging of random sets based on their distance functions. *J. Math. Imaging Vision* **8** 79–92. MR1612209

BALLESTER, A., PARRILLA, E., URIEL, J., PIEROLA, A., ALEMANY, S., NACHER, B., GONZALEZ, J. and GONZALEZ, J. C. (2014). 3D-based resources fostering the analysis, use, and exploitation of available body anthropometric data. In 5*th International Conference on* 3*D Body Scanning Technologies*.

BAUER, M., HARMS, P. and MICHOR, P. W. (2012). Sobolev metrics on shape space, II: Weighted Sobolev metrics and almost local metrics. *J. Geom. Mech.* **4** 365–383. MR3011892

BHATTACHARYA, R. and PATRANGENARU, V. (2002). Nonparametic estimation of location and dispersion on Riemannian manifolds. *J. Statist. Plann. Inference* **108** 23–35. MR1947389

BHATTACHARYA, R. and PATRANGENARU, V. (2003). Large sample theory of intrinsic and extrinsic sample means on manifolds. I. *Ann. Statist.* **31** 1–29. MR1962498

BOOKSTEIN, F. L. (1978). Lecture notes in biomathematics. In *The Measurement of Biological Shape and Shape Change* Springer, Berlin.

Gregorio Quintana-Ortí is Associate Professor in Computer Science, Department of Computer Science and Engineering, Universitat Jaume I, Avda. de Sos Baynat, s/n. 12071-Castellón de la Plana, Spain (e-mail: gquintan@uji.es). Amelia Simó is Professor of Statistics, Department of Mathematics-IMAC, Universitat Jaume I, Avda. de Sos Baynat, s/n. 12071-Castellón de la Plana, Spain (e-mail: simo@uji.es).

BOOKSTEIN, F. L. (1986). Size and shape spaces for landmark data in two dimensions. *Statist. Sci.* **2** 181–222.

COX, T. F. and COX, M. A. (2000). *Multidimensional Scaling*. CRC Press, Boca Raton, FL.

DAVIS, B. C., FLETCHER, P. T., BULLITT, E. and JOSHI, S. (2007). Population shape regression from random design data. In *Computer Vision*, 2007. *ICCV* 2007. *IEEE* 11*th International Conference on* 1–7. IEEE, New York.

DEDIEU, J.-P., PRIOURET, P. and MALAJOVICH, G. (2003). Newton's method on Riemannian manifolds: Convariant alpha theory. *IMA J. Numer. Anal.* **23** 395–419. MR1987937

DEL BIMBO, A., DE MARSICO, M., LEVIALDI, S. and PERITORE, G. (1998). Query by dialog: An interactive approach to pictorial querying. *Image Vis. Comput.* **16** 557–569.

DRYDEN, I. L. (2012). `shapes` package. R Foundation for Statistical Computing, Vienna, Austria. Contributed package.

DRYDEN, I. L. and MARDIA, K. V. (1998). *Statistical Shape Analysis*. *Wiley Series in Probability and Statistics*: *Probability and Statistics*. Wiley, Chichester. MR1646114

DRYDEN, I. L. and MARDIA, K. V. (2016). *Statistical Shape Analysis*: *With Applications in R*. Wiley, Chichester. MR3559734

FLETCHER, T. (2011). Geodesic regression on Riemannian manifolds. In *Proceedings of the Third International Workshop on Mathematical Foundations of Computational Anatomy-Geometrical and Statistical Methods for Modelling Biological Shape Variability* 75–86.

FLETCHER, P. T. and ZHANG, M. (2016). Probabilistic geodesic models for regression and dimensionality reduction on Riemannian manifolds. In *Riemannian Computing in Computer Vision* 101–121. Springer, Cham. MR3444348

FRÉCHET, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. Henri Poincaré* **10** 215–310. MR0027464

GONZÁLEZ-RODRÍGUEZ, G., TRUTSCHNIG, W. and COLUBI, A. (2009). Confidence regions for the mean of a fuzzy random variable. In *IFSA-EUSFLAT*, *Lisbon*, *Portugal* (J. P. Carvalho, D. Dubois, U. Kaymak and J. M. da Costa Sousa, eds.) 1433–1438.

GOODALL, C. (1991). Procrustes methods in the statistical analysis of shape. *J. Roy. Statist. Soc. Ser. B* **53** 285–339. MR1108330

GUAL-ARNAU, X., HEROLD-GARCÍA, S. and SIMÓ, A. (2013). Shape description from generalized support functions. *Pattern Recogn. Lett.* **34** 619–626.

GUAL-ARNAU, X., HEROLD-GARCÍA, S. and SIMÓ, A. (2015). Geometric analysis of planar shapes with applications to cell deformations. *Image Anal. Stereol.* **34** 171–182. MR3513425

GYÖRFI, L., KOHLER, M., KRZYŻAK, A. and WALK, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. *Springer Series in Statistics*. Springer, New York. MR1920390

HÄRDLE, W., MÜLLER, M., SPERLICH, S. and WERWATZ, A. (2004). *Nonparametric and Semiparametric Models*. *Springer Series in Statistics*. Springer, New York. MR2061786

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

JERMYN, I. H., KURTEK, S., LAGA, H. and SRIVASTAVA, A. (2017). *Elastic Shape Analysis of Three-Dimensional Objects*. *Synthesis Lectures on Computer Vision* **12** 1–185.

JUPP, P. E. and KENT, J. T. (1987). Fitting smooth paths to spherical data. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **36** 34–46. MR0887825

KARCHER, H. (1977). Riemannian center of mass and mollifier smoothing. *Comm. Pure Appl. Math.* **30** 509–541. MR0442975

KENDALL, D. G. (1977). The diffusion of shape. *Adv. in Appl. Probab.* **9** 428–430.

KENDALL, D. G. (1984). Shape manifolds, Procrustean metrics, and complex projective spaces. *Bull. Lond. Math. Soc.* **16** 81–121. MR0737237

KENDALL, W. S. (1990). Probability, convexity, and harmonic maps with small image. I. Uniqueness and fine existence. *Proc. Lond. Math. Soc.* (3) **61** 371–406. MR1063050

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

KENT, J. T. (1994). The complex Bingham distribution and shape analysis. *J. Roy. Statist. Soc. Ser. B* **56** 285–299. MR1281934

KIM, H., ADLURU, N., COLLINS, M., CHUNG, M., BENDLIN, B., JOHNSON, S., DAVIDSON, R. and SINGH, V. (2014). Multivariate general linear models (MGLM) on Riemannian manifolds with applications to statistical analysis of diffusion weighted images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2705–2712.

KINDRATENKO, V. V. (2003). On using functions to describe the shape. *J. Math. Imaging Vision* **18** 225–245. MR1971180

KLASSEN, E., SRIVASTAVA, A., MIO, M. and JOSHI, S. H. (2004). Analysis of planar shapes using geodesic paths on shape spaces. *IEEE Trans. Pattern Anal. Mach. Intell.* **26** 372–383.

KLEMELÄ, J. (2014). *Multivariate Nonparametric Regression and Visualization*. *Wiley Series in Computational Statistics*. Wiley, Hoboken, NJ. With R and applications to finance. MR3222314

KOBAYASHI, S. and NOMIZU, K. (1969). *Foundations of Differential Geometry. Vol. II. Interscience Tracts in Pure and Applied Mathematics*, *No.* 15 *Vol. II*. Interscience Publishers Wiley, New York. MR0238225

LE, H. (2001). Locating Fréchet means with application to shape spaces. *Adv. in Appl. Probab.* **33** 324–338. MR1842295

LE, H. L. and KENDALL, D. G. (1993). The Riemannian structure of Euclidean shape spaces: A novel environment for statistics. *Ann. Statist.* **21** 1225–1271. MR1241266

LONCARIC, S. (1998). A survey of shape analysis techniques. *Pattern Recognit.* **31** 983–1001.

MAMMEN, E. (2000). Resampling methods for nonparametric regression. In *Smoothing and Regression*: *Approaches*, *Computation*, *and Application* 425–450. Wiley, New York.

MARDIA, K. V. and JUPP, P. E. (2000). *Directional Statistics*. *Wiley Series in Probability and Statistics*. Wiley, Chichester. Revised reprint of ıt Statistics of directional data by Mardia [MR0336854 (49 #1627)]. MR1828667

MATHERON, G. (1975). *Random Sets and Integral Geometry*. Wiley, New York. MR0385969

MOLCHANOV, I. (2005). *Theory of Random Sets*. *Probability and Its Applications* (*New York*). Springer, London. MR2132405

NADARAYA, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **9** 141–142.

PENNEC, X. (2006). Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. *J. Math. Imaging Vision* **25** 127–154. MR2254442

PRINCE, J. L. and WILLSKY, A. S. (1990). Reconstructing convex sets from support line measurements. *IEEE Trans. Pattern Anal. Mach. Intell.* **12** 377–389.

R DEVELOPMENT CORE TEAM (2014). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

SERRA, J. (1984). *Image Analysis and Mathematical Morphology*. Academic Press, London. English version revised by Noel Cressie. MR0753649

SHI, X., STYNER, M., LIEBERMAN, J., IBRAHIM, J. G., LIN, W. and ZHU, H. (2009). Intrinsic regression models for manifold-valued data. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI* 2009 192–199.

SIMÓ, A., DE VES, E. and AYALA, G. (2004). Resuming shapes with applications. *J. Math. Imaging Vision* **20** 209–222. MR2060144

SMALL, C. G. (1996). *The Statistical Theory of Shape*. *Springer Series in Statistics*. Springer, New York. MR1418639

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

STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705. MR0790566

STOYAN, D. and STOYAN, H. (1994). *Fractals, Random Shapes and Point Fields*. Wiley, Chichester. Translated from the 1992 German original by N. Bamber and R. B. Johnson. MR1297125

VINUÉ, G., SIMÓ, A. and ALEMANY, S. (2016). The k-means algorithm for 3D shapes with an application to apparel design. *Adv. Data Anal. Classif.* **10** 103–132. MR3464302

WOODS, R. P. (2003). Characterizing volume and surface deformations in an atlas framework: Theory, applications, and implementation. *NeuroImage* **18** 769–788.

YOUNES, L. (1998). Computable elastic distances between shapes. *SIAM J. Appl. Math.* **58** 565–586. MR1617630

YOUNES, L., MICHOR, P. W., SHAH, J. and MUMFORD, D. (2008). A metric on shape space with explicit geodesics. *Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei* (9) *Mat. Appl.* **19** 25–57. MR2383560

Statistical Science
2019, Vol. 34, No. 2, 253–279
https://doi.org/10.1214/18-STS681

Statistical Analysis of Zero-Inflated Nonnegative Continuous Data: A Review

Lei Liu, Ya-Chen Tina Shih, Robert L. Strawderman, Daowen Zhang, Bankole A. Johnson and Haitao Chai

Abstract. Zero-inflated nonnegative continuous (or semicontinuous) data arise frequently in biomedical, economical, and ecological studies. Examples include substance abuse, medical costs, medical care utilization, biomarkers (e.g., CD4 cell counts, coronary artery calcium scores), single cell gene expression rates, and (relative) abundance of microbiome. Such data are often characterized by the presence of a large portion of zero values and positive continuous values that are skewed to the right and heteroscedastic. Both of these features suggest that no simple parametric distribution may be suitable for modeling such type of outcomes. In this paper, we review statistical methods for analyzing zero-inflated nonnegative outcome data. We will start with the cross-sectional setting, discussing ways to separate zero and positive values and introducing flexible models to characterize right skewness and heteroscedasticity in the positive values. We will then present models of correlated zero-inflated nonnegative continuous data, using random effects to tackle the correlation on repeated measures from the same subject and that across different parts of the model. We will also discuss expansion to related topics, for example, zero-inflated count and survival data, nonlinear covariate effects, and joint models of longitudinal zero-inflated nonnegative continuous data and survival. Finally, we will present applications to three real datasets (i.e., microbiome, medical costs, and alcohol drinking) to illustrate these methods. Example code will be provided to facilitate applications of these methods.

Key words and phrases: Two-part model, Tobit model, health econometrics, semiparametric regression, joint model, cure rate, frailty model, splines.

REFERENCES

AITCHISON, J. (1955). On the distribution of a positive random variable having a discrete probability mass at the origin. *J. Amer. Statist. Assoc.* **50** 901–908. MR0071685

ALBERT, P. S. (2005). Letter to the editor. *Biometrics* **61** 879–881. MR2196179

AMEMIYA, T. (1994). *Introduction to Statistics and Econometrics*. Harvard Univ. Press, Boston, MA.

BANG, H. and TSIATIS, A. A. (2002). Median regression with censored cost data. *Biometrics* **58** 643–649. MR1926117

BASU, A. and MANNING, W. G. (2006). A test for proportional hazards assumption within the exponential conditional mean framework. *Health Serv. Outcomes Res. Methodol.* **6** 81–100.

Lei Liu is Professor of Biostatistics, Division of Biostatistics, Washington University in St. Louis, St. Louis, Missouri 63110, USA (e-mail: lei.liu@wustl.edu). Ya-Chen Tina Shih is Professor, Department of Health Services Research, University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA (e-mail: yashih@mdanderson. org). Robert L. Strawderman is the Donald M. Foster M.D. Distinguished Professor of Biostatistics and Chair, Department of Biostatistics and Computational Biology, University of Rochester, Rochester, New York 14642, USA (e-mail: robert_strawderman@urmc.rochester.edu). Daowen Zhang is Professor, Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, USA (e-mail: daowen_zhang@ncsu.edu). Bankole A. Johnson is Professor and Chair, Department of Psychiatry, University of Maryland, Baltimore, Maryland 21201, USA (e-mail: bjohnson@psych.umaryland.edu). Haitao Chai is Ph.D. student, Institute for Financial Studies, Shandong University, Jinan, Shandong 250100, China (e-mail: cht0816@163.com).

BASU, A., MANNING, W. G. and MULLAHY, J. (2004). Comparing alternative models: Log vs Cox proportional hazard? *Health Econ*. **13** 749–765.

BASU, A. and RATHOUS, P. J. (2005). Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. *Biostatistics* **6** 93–109.

BERK, K. N. and LACHENBRUCH, P. A. (2002). Repeated measures with zeros. *Stat. Methods Med. Res*. **11** 303–316.

BJERRE, B., MARQUES, P., SELEN, J. and THORSSON, U. (2007). Swedish alcohol ignition interlock programme for drink-drivers: Effects on hospital care utilization and sick leave. *Addiction* **102** 560–570.

BLOUGH, D. K., MADDEN, C. W. and HORNBROOK, M. C. (1999). Modeling risk using generalized linear models. *J. Health Econ*. **18** 153–171.

BOAG, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J. Roy. Statist. Soc*. **11** 15–53.

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

BRETON, C. V., KILE, M. L., CATALANO, P. J., HOFFMAN, E., QUAMRUZZAMAN, Q., RAHMAN, M., MAHIUDDIN, G. and CHRISTIANI, D. C. (2007). GSTM1 and APE1 genotypes affect arsenic-induced oxidative stress: A repeated measures study. *Environ. Health* **6** 39.

CHAI, H. S. and BAILEY, K. R. (2008). Use of log-skew-normal distribution in analysis of continuous data with a discrete component at zero. *Stat. Med*. **27** 3643–3655. MR2523977

CHAI, H., JIANG, H., LIN, L. and LIU, L. (2018). A marginalized two-part Beta regression model for microbiome compositional data. *PLoS Comput. Biol*. **14** e1006329.

CHEN, E. Z. and LI, H. (2016). A two-part mixed-effect model for analyzing longitudinal microbiome compositional data. *Bioinformatics* **32** 2611–2617.

CHEN, J., JOHNSON, B. A., WANG, X. Q., O'QUIGLEY, J., ISAAC, M., ZHANG, D. and LIU, L. (2012). Trajectory analyses in alcohol treatment research. *Alcohol. Clin. Exp. Res*. **36** 1442–1448.

CHEN, J., LIU, L., JOHNSON, B. A. and O'QUIGLEY, J. (2013a). Penalized likelihood estimation for semiparametric mixed models, with application to alcohol treatment research. *Stat. Med*. **32** 335–346. MR3041871

CHEN, J., LIU, L., ZHANG, D. and SHIH, Y.-C. T. (2013b). A flexible model for the mean and variance functions, with application to medical cost data. *Stat. Med*. **32** 4306–4318. MR3118356

CHEN, J., LIU, L., SHIH, Y.-C. T., ZHANG, D. and SEVERINI, T. A. (2016). A flexible model for correlated medical costs, with application to medical expenditure panel survey data. *Stat. Med*. **35** 883–894. MR3457613

COOPER, N. J., LAMBERT, P. C., ABRAMS, K. R. and SUTTON, A. J. (2007). Predicting costs over time using Bayesian Markov chain Monte Carlo methods: An application to early inflammatory polyarthritis. *Health Econ*. **16** 37–56.

COTTER, D., THAMER, M., NARASIMHAN, K., ZHANG, Y. and BULLOCK, K. (2006). Translating epoetin research into practice: The role of government and the use of scientific evidence. *Health Aff*. **25** 1249–1259.

DOMINICI, F. and ZEGER, S. L. (2005). Smooth quantile ratio estimation with regression: Estimating medical expenditures for smoking-attributable diseases. *Biostatistics* **6** 505–519.

DOMINICI, F., COPE, L., NAIMAN, D. Q. and ZEGER, S. L. (2005). Smooth quantile ratio estimation. *Biometrika* **92** 543–557. MR2202645

DOW, W. H. and NORTON, E. C. (2003). Choosing between and interpreting the heckit and two-part models for corner solutions. *Health Serv. Outcomes Res. Methodol*. **4** 5–18.

DUAN, N. (1983). Smearing estimate: A nonparametric retransformation method. *J. Amer. Statist. Assoc*. **78** 605–610. MR0721207

DUAN, N., MANNING, W. G., MORRIS, C. and NEWHOUSE, J. P. (1983). A comparison of alternative models for the demand for medical care. *J. Bus. Econom. Statist*. **1** 115–126.

DUDLEY, R. A., HARRELL, F. E. JR, SMITH, L. R., MARK, D. B., CALIFF, R. M., PRYOR, D. B., GLOWER, D., LIPSCOMB, J. and HLATKY, M. (1993). Comparison of analytic models for estimating the effect of clinical factors on the cost of coronary artery bypass graft surgery. *J. Clin. Epidemiol*. **46** 261–271.

FALK, D., WANG, X. Q., LIU, L., FERTIG, J., MATTSON, M., RYAN, M., JOHNSON, B., STOUT, R. and LITTEN, R. Z. (2010). Percentage of subjects with no heavy drinking days: Evaluation as an efficacy endpoint for alcohol clinical trials. *Alcohol. Clin. Exp. Res*. **34** 2022–2034.

FAREWELL, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* **38** 1041–1046.

FINAK, G., MCDAVID, A., YAJIMA, M., DENG, J., GERSUK, V., SHALEK, A. K., SLICHTER, C. K., MILLER, H. W., MCELRATH, M. J., PRLIC, M., LINSLEY, P. S. and GOTTARDO, R. (2015). MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol*. **16** 278.

FOOD AND DRUG ADMINISTRATION (2006). *Medical Review of Vivitrol* 21-897. U.S. Government, Rockville, MD.

GATSONIS, C., EPSTEIN, A. M., NEWHOUSE, J. P., NORMAND, S. L. and MCNEIL, B. J. (1995). Variations in the utilization of coronary angiography for elderly patients with an acute myocaridal infaction: An analysis using hierarchical logistic regression. *Med. Care* **33** 625–642.

GHOSH, P. and ALBERT, P. S. (2009). A Bayesian analysis for longitudinal semicontinuous data with an application to an acupuncture clinical trial. *Comput. Statist. Data Anal*. **53** 699–706. MR2654581

HALL, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: A case study. *Biometrics* **56** 1030–1039. MR1815581

HALL, D. B. and SEVERINI, T. A. (1998). Extended generalized estimating equations for clustered data. *J. Amer. Statist. Assoc*. **93** 1365–1375. MR1666633

HAN, D., LIU, L., SU, X., JOHNSON, B. and SUN, L. (2018). Variable selection for random effects two-part model. *Stat. Methods Med. Res*. DOI:10.1177/0962280218784712.

HECKMAN, J. J. (1979). Sample selection bias as a specification error. *Econometrica* **47** 153–161. MR0518832

HEITJAN, D. F., KIM, C. Y. and LI, H. (2004). Bayesian estimation of cost-effectiveness from censored data. *Stat. Med*. **23** 1297–1309.

HENDERSON, R., DIGGLE, P. and DOBSON, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics* **1** 465–480.

HYNDMAN, R. and GRUNWALD, G. (2000). Generalized additive modelling of mixed distribution Markov models with application to Melbourne's rainfall. *Aust. N. Z. J. Stat.* **42** 145–158.

JAIN, A. K. and STRAWDERMAN, R. L. (2002). Flexible hazard regression modeling for medical cost data. *Biostatistics* **3** 101–118.

JAMES, G., WITTEN, D., HASTIE, T. and TIBSHIRANI, R. (2013). *An Introduction to Statistical Learning*: *With Applications in R. Springer Texts in Statistics* **103**. Springer, New York. MR3100153

JHA, A. K., VAROSY, P. D., KANAYA, A. K., HUNNINGHAKE, D. B., HLATKY, M. A., WATERS, D. D., FURBERG, C. D. and SHLIPAK, M. G. (2003). Differences in medical care and disease outcomes among black and white women with heart disease. *Circulation* **108** 1089–1094.

JOHNSON, B. A., ROSENTHAL, N., CAPECE, J. A., WIEGAND, F., MAO, L., BAYERS, K., MCKAY, A., AIT-DAOUD, N., ANTON, R. F., CIRAULO, D. A., KRANZLER, H. R., MANN, K., O'MALLEY, S. S. and SWIFT, R. M. (2007). Topiramate for treating alcohol dependence—a randomized controlled trial. *J. Am. Med. Assoc.* **298** 1641–1651.

JOHNSON, B. A., AIT-DAOUD, N., WANG, X.-Q., PENBERTHY, J. K., JAVORS, M. A., SENEVIRATNE, C. and LIU, L. (2013). Topiramate for the treatment of cocaine addiction: A randomized clinical trial. *J. Am. Med. Dir. Assoc. Psychiatr.* **70** 1338–1346.

KALBFLEISCH, J. D. and PRENTICE, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR1924807

KUK, A. Y. C. and CHEN, C. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika* **79** 531–541.

LAMBERT, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34** 1–14.

LEUNG, S. F. and YU, S. (1996). On the choice between sample selection and two-part models. *J. Econometrics* **72** 197–229.

LEWIS, J. D., CHEN, E. Z., BALDASSANO, R. N., OTLEY, A. R., GRIFFITHS, A. M., LEE, D., BITTINGER, K., BAILEY, A., FRIEDMAN, E. S., HOFFMANN, C., ALBENBERG, L., SINHA, R., COMPHER, C., GILROY, E., NESSEL, L., GRANT, A., CHEHOUD, C., LI, H., WU, G. D. and BUSHMAN, F. D. (2015). Inflammation, antibiotics, and diet as environmental stressors of the gut microbiome in pediatric Crohn's disease. *Cell Host Microbe* **18** 489–500.

LI, P., SCHNEIDER, J. E. and WARD, M. M. (2007). Effect of critical access hospital conversion on patient safety. *Health Serv. Res.* **42** 2089–2108; discussion 2294–2323.

LI, C.-S. and TAYLOR, J. M. G. (2002). A semi-parametric accelerated failure time cure model. *Stat. Med.* **21** 3235–3247.

LIN, D. Y., ETZIONI, R., FEUER, E. J. and WAX, Y. (1997). Estimating medical costs from incomplete follow-up data. *Biometrics* **53** 419–434.

LIPSCOMB, J., ANCUKIEWICZ, M., PARMIGIANI, G., HASSELBLAD, V., SAMSA, G. and MATCHAR, D. B. (1998). Predicting the cost of illness: A comparison of alternative models applied to stroke. *Med. Decis. Mak.* **18** S39–S56.

LITTELL, R. C., MILLIKEN, G. A., STROUP, W. W., WOLFINGER, R. D. and SCHABERNBERGER, O. (2006). *SAS for Mixed Model*, 2nd ed. SAS Institute Inc., Cary, NC.

LIU, L. (2009). Joint modeling longitudinal semi-continuous data and survival, with application to longitudinal medical cost data. *Stat. Med.* **28** 972–986. MR2518360

LIU, L. and HUANG, X. (2008). The use of Gaussian quadrature for estimation in frailty proportional hazards models. *Stat. Med.* **27** 2665–2683. MR2440058

LIU, Y. and LIU, L. (2015). Joint models for longitudinal data and time-to-event occurrence. In *Routledge International Handbook of Advanced Quantitative Methods in Nursing Research* (S. J. Henly, ed.) 253–263. Taylor and Francis, London.

LIU, L., MA, J. Z. and JOHNSON, B. A. (2008). A multi-level two-part random effects model, with application to an alcohol-dependence study. *Stat. Med.* **27** 3528–3539. MR2523969

LIU, L., WOLFE, R. A. and HUANG, X. (2004). Shared frailty models for recurrent events and a terminal event. *Biometrics* **60** 747–756. MR2089451

LIU, L., WOLFE, R. A. and KALBFLEISCH, J. D. (2007). A shared random effects model for censored medical costs and mortality. *Stat. Med.* **26** 139–155. MR2312704

LIU, L., CONAWAY, M. R., KNAUS, W. A. and BERGIN, J. D. (2008). A random effects four-part model, with application to correlated medical costs. *Comput. Statist. Data Anal.* **52** 4458–4473. MR2432473

LIU, L., STRAWDERMAN, R. L., COWEN, M. E. and SHIH, Y. C. T. (2010). A flexible two-part random effects model for correlated medical costs. *J. Health Econ.* **29** 110–123.

LIU, L., HUANG, X., YAROSHINSKY, A. and CORMIER, J. N. (2016a). Joint frailty models for zero-inflated recurrent events in the presence of a terminal event. *Biometrics* **72** 204–214. MR3500589

LIU, L., STRAWDERMAN, R. L., JOHNSON, B. A. and O'QUIGLEY, J. M. (2016b). Analyzing repeated measures semi-continuous data, with application to an alcohol dependence study. *Stat. Methods Med. Res.* **25** 133–152. MR3460432

LU, S.-E., LIN, Y. and SHIH, W.-C. J. (2004). Analyzing excessive no changes in clinical trials with clustered data. *Biometrics* **60** 257–267. MR2044122

MAHMUD, S., LOU, W. W. and JOHNSTON, N. W. (2010). A probit- log- skew-normal mixture model for repeated measures data with excess zeros, with application to a cohort study of paediatric respiratory symptoms. *BMC Med. Res. Methodol.* **10** 55.

MANNING, W. G. (1998). The logged dependent variable, heteroscedasticity, and the retransformation problem. *J. Health Econ.* **17** 283–295.

MANNING, W. G., BASU, A. and MULLAHY, J. (2005). Generalized modeling approaches to risk adjustment of skewed outcomes data. *J. Health Econ.* **20** 465–488.

MANNING, W. G., DUAN, N. and ROGERS, W. H. (1987). Monte-Carlo evidence on the choice between sample selection and 2-part models. *J. Econometrics* **35** 59–82.

MANNING, W. G. and MULLAHY, J. (2001). Estimating log models: To transform or not to transform? *J. Health Econ.* **20** 461–494.

MANNING, W., MORRIS, C., NEWHOUSE, J. et al. (1981). A two-part model of the demand for medical care: Preliminary results from the health insurance study. In *Health, Economics, and*

Health Economics (J. van der Gaag and M. Perlman, eds.) 103–123. North-Holland, Amsterdam.

MARTINUSSEN, T. and SCHEIKE, T. H. (2006). *Dynamic Regression Models for Survival Data*. *Statistics for Biology and Health*. Springer, New York. MR2214443

MCDAVID, A., FINAK, G., CHATTOPADYAY, P. K., DOMINGUEZ, M., LAMOREAUX, L., MA, S. S., ROEDERER, M. and GOTTARDO, R. (2013). Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics* **29** 461–467.

MIN, Y. and AGRESTI, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Stat. Model.* **5** 1–19. MR2133525

MOULTON, L. and HALSEY, N. (1995). A mixture model with detection limits for regression analyses of antibody response to vaccine. *Biometrics* **51** 1570–1578.

MULLAHY, J. (1998). Much ado about two: Reconsidering retransformation and the two-part model in health econometrics. *J. Health Econ.* **17** 247–281.

NEELON, B., O'MALLEY, A. J. and NORMAND, S.-L. T. (2011). A Bayesian two-part latent class model for longitudinal medical expenditure data: Assessing the impact of mental health and substance abuse parity. *Biometrics* **67** 280–289. MR2898840

NEELON, B., O'MALLEY, A. J. and SMITH, V. A. (2016). Modeling zero-modified count and semicontinuous data in health services research part 1: Background and overview. *Stat. Med.* **35** 5070–5093. MR3569914

NEELON, B., ZHU, L. and NEELON, S. E. B. (2015). Bayesian two-part spatial models for semicontinuous data with application to emergency department expenditures. *Biostatistics* **16** 465–479. MR3365440

NEELON, B., CHANG, H. H., LING, Q. and HASTINGS, N. S. (2016). Spatiotemporal hurdle models for zero-inflated count data: Exploring trends in emergency department visits. *Stat. Methods Med. Res.* **25** 2558–2576. MR3572870

OLSEN, M. K. and SCHAFER, J. L. (2001). A two-part random-effects model for semicontinuous longitudinal data. *J. Amer. Statist. Assoc.* **96** 730–745. MR1946438

OTHUS, M., BARLOGIE, B., LEBLANC, M. L. and CROWLEY, J. J. (2012). Cure models as a useful statistical tool for analyzing survival. *Clin. Cancer Res.* **18** 3731–3736.

PARK, R. E. (1966). Estimation with heteroscedastic error terms. *Econometrica* **34** 888.

PENG, Y. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics* **56** 237–243.

PENG, Y. (2003). Fitting semiparametric cure models. *Comput. Statist. Data Anal.* **41** 481–490. MR1973725

PENG, Y., TAYLOR, J. M. G. and YU, B. (2007). A marginal regression model for multivariate failure time data with a surviving fraction. *Lifetime Data Anal.* **13** 351–369. MR2409955

PULLENAYEGUM, E. M. and WILLAN, A. R. (2007). Semiparametric regression models for cost-effectiveness analysis: Improving the efficiency of estimation from censored data. *Stat. Med.* **26** 3274–3299. MR2380581

RAUDENBUSH, S. W., YANG, M.-L. and YOSEF, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *J. Comput. Graph. Statist.* **9** 141–157. MR1826278

RIGBY, R. A. and STASINOPOULOS, D. M. (2005). Generalized additive models for location, scale and shape. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **54** 507–554. MR2137253

ROBERT, C. P. (2007). *The Bayesian Choice*: *From Decision-Theoretic Foundations to Computational Implementation*, 2nd ed. *Springer Texts in Statistics*. Springer, New York. MR2723361

RONDEAU, V., SCHAFFNER, E., CORBIÈRE, F., GONZALEZ, J. R. and MATHOULIN-PÉLISSIER, S. (2013). Cure frailty models for survival data: Application to recurrences for breast cancer and to hospital readmissions for colorectal cancer. *Stat. Methods Med. Res.* **22** 243–260. MR3190656

SCHOENFELD, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika* **69** 239–241.

SMITH, V. A., PREISSER, J. S., NEELON, B. and MACIEJEWSKI, M. L. (2014). A marginalized two-part model for semicontinuous data. *Stat. Med.* **33** 4891–4903. MR3276507

SMITH, V. A., NEELON, B., MACIEJEWSKI, M. L. and PREISSER, J. S. (2017a). Two parts are better than one. *Health Serv. Outcomes Res. Methodol.* **17** 198–218.

SMITH, V. A., NEELON, B., PREISSER, J. S. and MACIEJEWSKI, M. L. (2017b). A marginalized two-part model for longitudinal semicontinuous data. *Stat. Methods Med. Res.* **26** 1949–1968. MR3687189

SOBELL, L. C. and SOBELL, M. B. (1992). Timeline follow-back: A technique for assessing self-reported alcohol consumption. In *Measuring Alcohol Consumption*: *Psychosocial and Biochemical Methods* (R. Z. Litten and J. P. Allen, eds.) 41–72. Humana Press Inc., Totowa, NJ.

SPOSTO, R. (2002). Cure model analysis in cancer: An application to data from the children's cancer group. *Stat. Med.* **21** 293–312.

STRAM, D. O. and LEE, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics* **50** 1171–1177.

STUKEL, T. A., LUCAS, F. L. and WENNBERG, D. E. (2005). Long-term outcomes of regional variations in intensity of invasive vs medical management of medicare patients with acute myocardial infarction. *J. Am. Med. Assoc.* **293** 1329–1337.

STUKEL, T. A., FISHER, E. S., WENNBERG, D. E., ALTER, D. A., GOTTLIEB, D. J. and VERMEULEN, M. J. (2007). Analysis of observational studies in the presence of treatment selection bias effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. *J. Am. Med. Assoc.* **297** 278–285.

SU, L., TOM, B. D. M. and FAREWELL, V. T. (2009). Bias in 2-part mixed models for longitudinal semicontinuous data. *Biostatistics* **10** 374–389.

SU, X., WIJAYASINGHE, C. S., FAN, J. and ZHANG, Y. (2016). Sparse estimation of Cox proportional hazards models via approximated information criteria. *Biometrics* **72** 751–759. MR3545668

SY, J. P. and TAYLOR, J. M. G. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics* **56** 227–236. MR1767631

THERNEAU, T. M. and GRAMBSCH, P. M. (2000). *Modeling Survival Data*: *Extending the Cox Model*. *Statistics for Biology and Health*. Springer, New York. MR1774977

TIAN, L., ZUCKER, D. and WEI, L. J. (2005). On the Cox model with time-varying regression coefficients. *J. Amer. Statist. Assoc.* **100** 172–183. MR2156827

TOBIN, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica* **26** 24–36. MR0090462

TOOZE, J. A., GRUNWALD, G. K. and JONES, R. H. (2002). Analysis of repeated measures data with clumping at zero. *Stat. Methods Med. Res.* **11** 341–355.

TOOZE, J. A., MIDTHUNE, D., DODD, K. W., FREEDMAN, L. S., KREBS-SMITH, S. M., SUBAR, A. F., GUENTHER, P. M., CARROLL, R. J. and KIPNIS, V. (2006). A new statistical method for estimating the usual intake of episodically consumed foods with application to their distribution. *J. Am. Diet. Assoc.* **106** 1575–1587.

TSIATIS, A. A. and DAVIDIAN, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. *Statist. Sinica* **14** 809–834. MR2087974

TWISK, J. and RIJMEN, F. (2009). Longitudinal tobit regression: A new approach to analyze outcome variables with floor or ceiling effects. *J. Clin. Epidemiol.* **62** 953–958.

TYLER, A. D., SMITH, M. I. and SILVERBERG, M. S. (2014). Analyzing the human microbiome: A how to guide for physicians. *Am. J. Gastroenterol.* **109** 983–993.

VONESH, E. F., GREENE, T. and SCHLUCHTER, M. D. (2006). Shared parameter models for the joint analysis of longitudinal data and event times. *Stat. Med.* **25** 143–163. MR2222079

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

WANG, M.-C., QIN, J. and CHIANG, C.-T. (2001). Analyzing recurrent event data with informative censoring. *J. Amer. Statist. Assoc.* **96** 1057–1065. MR1947253

WILLIAMSON, J. M., DATTA, S. and SATTEN, G. A. (2003). Marginal analyses of clustered data when cluster size is informative. *Biometrics* **59** 36–42. MR1978471

WOOLDRIDGE, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA.

WULFSOHN, M. S. and TSIATIS, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53** 330–339. MR1450186

XIE, H., MCHUGO, G., SENGUPTA, A., CLARK, R. and DRAKE, R. (2004). A method for analyzing long longitudinal outcomes with many zeros. *Ment. Health Serv. Res.* **6** 239–246.

YABROFF, K. R., WARREN, J. L., SCHRAG, D., MARIOTTO, A., MEEKINS, A., TOPOR, M. and BROWN, M. L. (2009). Comparison of approaches for estimating incidence costs of care for colorectal cancer patients. *Med. Care* **47** S56–S63.

YAMAGUCHI, K. (1992). Accelerated failure-time regression models with a regression model of surviving fraction: An application to the analysis of "Permanent Employment" in Japan. *J. Amer. Statist. Assoc.* **87** 284–292.

YU, Z., LIU, L., BRAVATA, D. M., WILLIAMS, L. S. and TEPPER, R. S. (2013). A semiparametric recurrent events model with time-varying coefficients. *Stat. Med.* **32** 1016–1026. MR3042854

ZHANG, M., STRAWDERMAN, R. L., COWEN, M. E. and WELLS, M. T. (2006). Bayesian inference for a two-part hierarchical model: An application to profiling providers in managed health care. *J. Amer. Statist. Assoc.* **101** 934–945. MR2324094

ZHOU, X. H. and TU, W. (1999). Comparison of several independent population means when their samples contain log-normal and possibly zero observations. *Biometrics* **55** 645–651.

Statistical Science
2019, Vol. 34, No. 2, 280–300
https://doi.org/10.1214/18-STS686

The Importance of Being Clustered: Uncluttering the Trends of Statistics from 1970 to 2015

Laura Anderlucci, Angela Montanari and Cinzia Viroli

Abstract. In this paper, we retrace the recent history of statistics by analyzing all the papers published in five prestigious statistical journals since 1970, namely: *The Annals of Statistics*, *Biometrika*, *Journal of the American Statistical Association*, *Journal of the Royal Statistical Society, Series B* and *Statistical Science*. The aim is to construct a kind of "taxonomy" of the statistical papers by organizing and clustering them in main themes. In this sense being identified in a cluster means being important enough to be uncluttered in the vast and interconnected world of the statistical research. Since the main statistical research topics naturally born, evolve or die during time, we will also develop a dynamic clustering strategy, where a group in a time period is allowed to migrate or to merge into different groups in the following one. Results show that statistics is a very dynamic and evolving science, stimulated by the rise of new research questions and types of data.

Key words and phrases: Model-based clustering, cosine distance, textual data analysis.

REFERENCES

AMBROISE, C. and GOVAERT, G. (2000). EM Algorithm for Partially Known Labels. In *Data analysis*, *classification*, *and related methods*, 161–166. Springer, Berlin.

BANERJEE, A., DHILLON, I. S., GHOSH, J. and SRA, S. (2005). Clustering on the unit hypersphere using von Mises–Fisher distributions. *J. Mach. Learn. Res.* **6** 1345–1382. MR2249858

BEN-ISRAEL, A. and IYIGUN, C. (2008). Probabilistic D-clustering. *J. Classification* **25** 5–26. MR2429670

BLEI, D. M. and LAFFERTY, J. D. (2006). Dynamic topic models. In *ICML '06 Proceedings of the 23rd international conference on Machine learning* 113–120. ACM, New York.

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

BOUVEYRON, C., LATOUCHE, P. and ZREIK, R. (2018). The stochastic topic block model for the clustering of vertices in networks with textual edges. *Stat. Comput.* **28** 11–31. MR3741634

CHANG, J. and BLEI, D. M. (2009). Relational topic models for document networks. In *International Conference on Artificial Intelligence and Statistics* 81–88. Avaialble at http://proceedings.mlr.press/v5/chang09a/chang09a.pdf.

CÔME, E., OUKHELLOU, L., DENŒUX, T. and AKNIN, P. (2009). Learning from partially supervised data using mixture models and belief functions. *Pattern Recognition* **42** 334–348.

DEERWESTER, S., DUMAIS, S., FURNAS, G., LANDAUER, T. and HARSHMAN, R. (1990). Indexing by latent semantic analysis. *J. Amer. Soc. Inform. Sci.* **41** 391–407.

DHILLON, I. S. and MODHA, D. S. (2001). Concept decompositions for large sparse text data using clustering. *Mach. Learn.* **42** 143–175.

DIACONIS, P. (1988). *Group Representations in Probability and Statistics*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **11**. IMS, Hayward, CA. MR0964069

FLIGNER, M. A. and VERDUCCI, J. S. (1986). Distance based ranking models. *J. Roy. Statist. Soc. Ser. B* **48** 359–369. MR0876847

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

Laura Anderlucci is Senior Assistant Professor, Department of Statistical Sciences, University of Bologna, via Belle Arti 41, 40126 Bologna, Italy (e-mail: laura.anderlucci@unibo.it). Angela Montanari is Professor of Statistics, Department of Statistical Sciences, University of Bologna, via Belle Arti 41, 40126 Bologna, Italy (e-mail: angela.montanari@unibo.it). Cinzia Viroli is Professor of Statistics, Department of Statistical Sciences, University of Bologna, via Belle Arti 41, 40126 Bologna, Italy (e-mail: cinzia.viroli@unibo.it).

HOFMANN, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the* 22*nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* 50–57. ACM, New York.

JI, P. and JIN, J. (2016). Coauthorship and citation networks for statisticians. *Ann. Appl. Stat.* **10** 1779–1812. MR3592033

KOLAR, M. and TADDY, M. (2016). Discussion of "Coauthorship and citation networks for statisticians" [MR3592033]. *Ann. Appl. Stat.* **10** 1835–1841. MR3592037

MAITRA, R. and RAMLER, I. P. (2010). A k-mean-directions algorithm for fast clustering of data on the sphere. *J. Comput. Graph. Statist.* **19** 377–396. MR2758308

MALLOWS, C. L. (1957). Non-null ranking models. I. *Biometrika* **44** 114–130. MR0087267

MARDIA, K. V. and JUPP, P. E. (2000). *Directional Statistics*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR1828667

MCLACHLAN, G. and PEEL, D. (2000). *Finite Mixture Models*. *Wiley Series in Probability and Statistics*: *Applied Probability and Statistics*. Wiley Interscience, New York. MR1789474

MURPHY, T. B. and MARTIN, D. (2003). Mixtures of distance-based models for ranking data. *Comput. Statist. Data Anal.* **41** 645–655. MR1973732

NIGAM, K., MCCALLUM, A., THRUN, S. and MITCHELL, T. (2000). Text classification from labeled and unlabeled documents using em. *Mach. Learn.* **39** 103–134.

SALTON, G. and MCGILL, M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

SHANNON, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.* **27** 379–423, 623–656. MR0026286

SUN, Y., HAN, J., GAO, J. and YU, Y. (2009). Itopicmodel: Information network-integrated topic modeling. In *Ninth IEEE International Conference on Data Mining* 493–502.

VANDEWALLE, V., BIERNACKI, C., CELEUX, G. and GOVAERT, G. (2013). A predictive deviance criterion for selecting a generative model in semi-supervised classification. *Comput. Statist. Data Anal.* **64** 220–236. MR3061900

VARIN, C., CATTELAN, M. and FIRTH, D. (2016). Statistical modelling of citation exchange between statistics journals. *J. Roy. Statist. Soc. Ser. A* **179** 1–63. MR3461568

ZHONG, S. and GHOSH, J. (2005). Generative model-based document clustering: A comparative study. *Knowledge and Information Systems* **8** 374–384.

ZHU, X., GOLDBERG, A. B., BRACHMAN, R. and DIETTERICH, T. (2009). *Introduction to Semi-Supervised Learning*. Morgan and Claypool, Williston, VT.

Statistical Science
2019, Vol. 34, No. 2, 301–316
https://doi.org/10.1214/18-STS687

Producing Official County-Level Agricultural Estimates in the United States: Needs and Challenges

Nathan B. Cruze, Andreea L. Erciulescu, Balgobin Nandram, Wendy J. Barboza and Linda J. Young

Abstract. In the United States, county-level estimates of crop yield, production, and acreage published by the United States Department of Agriculture's National Agricultural Statistics Service (USDA NASS) play an important role in determining the value of payments allotted to farmers and ranchers enrolled in several federal programs. Given the importance of these official county-level crop estimates, NASS continually strives to improve its crops county estimates program in terms of accuracy, reliability and coverage. In 2015, NASS engaged a panel of experts convened under the auspices of the National Academies of Sciences, Engineering, and Medicine Committee on National Statistics (CNSTAT) for guidance on implementing models that may synthesize multiple sources of information into a single estimate, provide defensible measures of uncertainty, and potentially increase the number of publishable county estimates. The final report titled *Improving Crop Estimates by Integrating Multiple Data Sources* was released in 2017. This paper discusses several needs and requirements for NASS county-level crop estimates that were illuminated during the activities of the CNSTAT panel. A motivating example of planted acreage estimation in Illinois illustrates several challenges that NASS faces as it considers adopting any explicit model for official crops county estimates.

Key words and phrases: Agricultural surveys, auxiliary data, benchmarking, official statistics, small area estimation.

REFERENCES

ADRIAN, D. W. (2012). A model-based approach to forecasting corn and soybean yields. In *Proceedings of the Fourth International Conference on Establishment Surveys* Amer. Statist. Assoc., Montreal, QC. https://ww2.amstat.org/meetings/ices/2012/papers/302190.pdf [Accessed: 2019-01-31].

BAILEY, J. T. and KOTT, P. S. (1997). An application of multiple list frame sampling for multi-purpose surveys. In *JSM Proceedings*, *Survey Research Methods Section* 496–500. Amer. Statist. Assoc., Alexandria, VA.

BATTESE, G. E., HARTER, R. M. and FULLER, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *J. Amer. Statist. Assoc.* **83** 28–36.

Nathan B. Cruze is Mathematical Statistician in Research and Development Division, USDA National Agricultural Statistics Service (NASS), 1400 Independence Avenue, SW, Washington, DC 20250, USA (e-mail: nathan.cruze@nass.usda.gov). Andreea L. Erciulescu is Senior Statistician, Westat, 1600 Research Boulevard, Rockville, Maryland 20850, USA (e-mail: AndreeaErciulescu@westat.com). Balgobin Nandram is Senior Professor of Statistics, Worcester Polytechnic Institute, Department of Mathematical Sciences, Stratton Hall, 100 Institute Road, Worcester, Massachusetts 01609, USA (e-mail: balnan@wpi.edu). Wendy J. Barboza (retired) is former Deputy Director of Research and Development Division, USDA National Agricultural Statistics Service, 1400 Independence Avenue, SW, Washington, DC 20250, USA (e-mail: wbarboza@cox.net). Linda J. Young is Chief Mathematical Statistician and Director of Research and Development Division, USDA National Agricultural Statistics Service, 1400 Independence Avenue, SW, Washington, DC 20250, USA (e-mail: linda.young@nass.usda.gov).

BELL, J. and BARBOZA, W. (2012). Evaluation of using CVs as a publication standard. In *Proceedings of the Fourth International Conference on Establishment Surveys* Amer. Statist. Assoc., Montreal, QC. https://ww2.amstat.org/meetings/ices/2012/papers/302002.pdf [Accessed: 2019-01-31].

BELL, W., BASEL, W. W. and MAPLES, J. J. (2016). An overview of the U.S. census bureau's small area income and poverty estimates program. In *Analysis of Poverty Data by Small Area Estimation* (M. Pratesi, ed.) 349–378 19. Wiley, Hoboken, NJ.

BELLOW, M. E. and LAHIRI, P. (2010). Empirical Bayes methodology for the NASS county estimation program. In *JSM Proceedings*, *Survey Research Methods Section* 343–355. Amer. Statist. Assoc., Alexandria, VA.

BELLOW, M. E. and LAHIRI, P. (2011). An empirical best linear unbiased prediction approach to small area estimation of crop parameters. In *JSM Proceedings*, *Survey Research Methods Section* 3976–3986. Amer. Statist. Assoc., Alexandria, VA.

BELLOW, M. E. and LAHIRI, P. (2012). Evaluation of methods for county level estimation of crop harvested area that employ mixed models. In *Proceedings of the DC-AAPOR/WSS Summer Conference*, *Bethesda*, *MD* Amer. Statist. Assoc., Alexandria, VA.

BORYAN, C. (2010). The USDA NASS Cropland Data Layer Program: Transition from Research to Operations (2006–2009. USDA NASS Education and Outreach: Research Reports https://www.nass.usda.gov/Education_and_Outreach/Reports,_Presentations_and_Conferences/reports/Boryan_CDL_Program_Research_to_Operations_2006-2009_final.pdf [Accessed: 2019-01-31].

BORYAN, C., YANG, Z., MUELLER, R. and CRAIG, M. (2011). Monitoring US agriculture: The US Department of Agriculture, National Agricultural Statistics Service, cropland data layer program. *Geocarto Int*. **26** 341–358.

CRUZE, N. B. (2015). Integrating survey data with auxiliary sources of information to estimate crop yields. In *JSM Proceedings*, *Survey Research Methods Section* 565–578. Amer. Statist. Assoc., Alexandria, VA.

CRUZE, N. B. (2016). A Bayesian hierarchical model for combining several crop yield indications. In *JSM Proceedings*, *Survey Research Methods Section* 2045–2053. Amer. Statist. Assoc., Alexandria, VA.

CRUZE, N. B. and BENECHA, H. K. (2017). A model-based approach to crop yield forecasting. In *JSM Proceedings*, *Survey Research Methods Section* 2724–2733. Amer. Statist. Assoc., Alexandria, VA.

CRUZE, N. B., ERCIULESCU, A. L., NANDRAM, B., BARBOZA, W. J. and YOUNG, L. J. (2019). Supplement to "Producing Official County-Level Agricultural Estimates in the United States: Needs and Challenges." DOI:10.1214/18-STS687SUPP.

CZAPLEWSKI, R. L. (1992). Misclassification bias in areal estimates. *Photogramm. Eng. Remote Sens*. **58** 189–192.

ERCIULESCU, A. L., CRUZE, N. B. and NANDRAM, B. (2016). Small area estimates incorporating auxiliary soucres of information. In *JSM Proceedings*, *Survey Research Methods Section* 3591–3605. Amer. Statist. Assoc., Alexandria, VA.

ERCIULESCU, A. L., CRUZE, N. B. and NANDRAM, B. (2017). Small area estimates of end-of-season agricultural quantities. In *JSM Proceedings*, *Survey Research Methods Section* 541–560. Amer. Statist. Assoc., Alexandria, VA.

ERCIULESCU, A. L., CRUZE, N. B. and NANDRAM, B. (2018). Benchmarking a triplet of official estimates. *Environ. Ecol. Stat*. **25** 523–547. MR3878890

ERCIULESCU, A. L., CRUZE, N. B. and NANDRAM, B. (2019). Model-based county level crop estimates incorporating auxiliary sources of information. *J. Roy. Statist. Soc. Ser. A* **182** 283–303. MR3902644

FAY, R. E. III and HERRIOT, R. A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *J. Amer. Statist. Assoc*. **74** 269–277. MR0548019

FULLER, W. A. and BATTESE, G. E. (1973). Transformations for estimation of linear models with nested-error structure. *J. Amer. Statist. Assoc*. **68** 626–632. MR0359188

GALLEGO, F. J. (2004). Remote sensing and land cover area estimation. *Int. J. Remote Sens*. **25** 3019–3047.

IWIG, W. (1996). The national agricultural statistics service county estimates program. In *Indirect Esitmators in U.S. Federal Programs* (W. Schaible, ed.) 129–144 7. Springer, New York.

JOHNSON, D. M. (2014). An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. *Remote Sens. Environ*. **141** 116–128.

JOHNSON, D. M. (2016). A comprehensive assessment of the correlations between field crop yields and commonly used MODIS products. *Int. J. Appl. Earth Obs. Geoinf*. **52** 65–81.

KIM, J. K., WANG, Z., ZHU, Z. and CRUZE, N. B. (2018). Combining survey and non-survey data for improved sub-area prediction using a multi-level model. *J. Agric. Biol. Environ. Stat*. **23** 175–189. MR3805267

KOTT, P. S. (1989). Robust small domain estimation using random effects modeling. *Surv. Methodol*. **15** 3–12.

NANDRAM, B., BERG, E. and BARBOZA, W. (2014). A hierarchical Bayesian model for forecasting state-level corn yield. *Environ. Ecol. Stat*. **21** 507–530. MR3248537

NATIONAL ACADEMIES OF SCIENCES ENGINEERING, AND MEDICINE (2017). *Improving Crop Estimates by Integrating Multiple Data Sources*. The National Academies Press, Washington, DC.

NATIONAL RESEARCH COUNCIL (1997). *Small-Area Estimates of School-Age Children in Poverty: Interim Report 1, Evaluation of 1993 County Estimates for Title I Allocations*. The National Academies Press, Washington, DC.

NATIONAL RESEARCH COUNCIL (1998). *Small-Area Estimates of School-Age Children in Poverty: Interim Report 2, Evaluation of Revised 1993 County Estimates for Title I Allocations*. The National Academies Press, Washington, DC.

NATIONAL RESEARCH COUNCIL (1999). *Small-Area Estimates of School-Age Children in Poverty: Interim Report 3*. The National Academies Press, Washington, DC.

NATIONAL RESEARCH COUNCIL (2000a). *Small-Area Estimates of School-Age Children in Poverty: Evaluation of Current Methodology*. The National Academies Press, Washington, DC.

NATIONAL RESEARCH COUNCIL (2000b). *Small-Area Income and Poverty Estimates: Priorities for 2000 and Beyond*. The National Academies Press, Washington, DC.

NATIONAL RESEARCH COUNCIL (2007). *Using the American Community Survey: Benefits and Challenges*. The National Academies Press, Washington, DC.

NATIONAL RESEARCH COUNCIL (2008). *Understanding American Agriculture: Challenges for the Agricultural Resource Management Survey*. The National Academies Press, Washington, DC.

RAO, J. N. K. and MOLINA, I. (2015). *Small Area Estimation*, 2nd ed. Wiley, Hoboken, NJ. MR3380626

STASNY, E. A., GOEL, P. K. and RUMSEY, D. J. (1991). County estimates of wheat production. *Surv. Methodol.* **17** 211–225.

UNITED STATES CENSUS BUREAU (2017). Small Area Income and Poverty Estimates (SAIPE) Program. https://www.census.gov/programs-surveys/saipe/about.html. [Accessed: 2019-01-31].

UNITED STATES GOVERNMENT PUBLISHING OFFICE (2015). Big Data and Agriculture: Innovation and Implications–Hearing before the Committee on Agriculture, House of Representatives, 114th Congress. https://www.govinfo.gov/content/pkg/CHRG-114hhrg97412/pdf/CHRG-114hhrg97412.pdf. [Accessed: 2019-01-31].

UNITED STATES GOVERNMENT PUBLISHING OFFICE (2016). Big Data and Agriculture: Innovation in the Air–Hearing before the Subcommittee on General Farm Commodities and Risk Management of the Committee on Agriculture, House of Representatives, 114th Congress. https://www.govinfo.gov/content/pkg/CHRG-114hhrg20574/pdf/CHRG-114hhrg20574.pdf. [Accessed: 2019-01-31].

USDA NASS (2010). *Field Crops: Usual Planting and Harvesting Dates*. Agricultural Handbook No. 628. https://downloads.usda.library.cornell.edu/usda-esmis/files/vm40xr56k/dv13zw65p/w9505297d/planting-10-29-2010.pdf [Accessed: 2019-01-31].

USDA NASS (2014). *Small Grains 2014 Summary*. https://downloads.usda.library.cornell.edu/usda-esmis/files/5t34sj573/2r36v149p/qj72p9909/SmalGraiSu-09-30-2014.pdf [Accessed: 2019-01-31].

USDA NASS (2015). *Crop Production 2014 Summary*. https://downloads.usda.library.cornell.edu/usda-esmis/files/k3569432s/5q47rr167/1n79h650b/CropProdSu-01-12-2015_revision.pdf [Accessed: 2019-01-31].

VOSE, R. S., APPLEQUIST, S., SQUIRES, M., DURRE, I., MENNE, M. J., WILLIAMS, C. N., FENIMORE, C., GLEASON, K. and ARNDT, D. (2014). Improved historical temperature and precipitation time series for U.S. climate divisions. *J. Appl. Meteorol. Climatol.* **53** 1232–1251.

WALKER, G. and SIGMAN, R. (1984). The use of LANDSAT for county estimates of crop areas: Evaluation of the huddleston-ray and the battese-fuller estimators for the case of stratified sampling. *Commun. Statist. Theory Methods* **13** 2975–2996.

WANG, J. C., HOLAN, S. H., NANDRAM, B., BARBOZA, W., TOTO, C. and ANDERSON, E. (2012). A Bayesian approach to estimating agricultural yield based on multiple repeated surveys. *J. Agric. Biol. Environ. Stat.* **17** 84–106. MR2912556

WILLIAMS, M. (2013). Small area modeling of county estimates for corn and soybean yields in the US. Presentation at the Federal Committee on Statistical Methodology Research Conference. http://www.copafs.org/UserFiles/file/fcsm/C2_Williams_2013FCSM.pdf [Accessed: 2019-01-31].

YOUNG, L. J. (2019). Agricultural crop forecasting for large geographical areas. *Ann. Rev. Stat. Appl.* **6**. 173–196.

Statistical Science
2019, Vol. 34, No. 2, 317–333
https://doi.org/10.1214/18-STS692

Two-Sample Instrumental Variable Analyses Using Heterogeneous Samples

Qingyuan Zhao, Jingshu Wang, Wes Spiller, Jack Bowden and Dylan S. Small

Abstract. Instrumental variable analysis is a widely used method to estimate causal effects in the presence of unmeasured confounding. When the instruments, exposure and outcome are not measured in the same sample, Angrist and Krueger (*J. Amer. Statist. Assoc.* **87** (1992) 328–336) suggested to use two-sample instrumental variable (TSIV) estimators that use sample moments from an instrument-exposure sample and an instrument-outcome sample. However, this method is biased if the two samples are from heterogeneous populations so that the distributions of the instruments are different. In linear structural equation models, we derive a new class of TSIV estimators that are robust to heterogeneous samples under the key assumption that the structural relations in the two samples are the same. The widely used two-sample two-stage least squares estimator belongs to this class. It is generally not asymptotically efficient, although we find that it performs similarly to the optimal TSIV estimator in most practical situations. We then attempt to relax the linearity assumption. We find that, unlike one-sample analyses, the TSIV estimator is not robust to misspecified exposure model. Additionally, to nonparametrically identify the magnitude of the causal effect, the noise in the exposure must have the same distributions in the two samples. However, this assumption is in general untestable because the exposure is not observed in one sample. Nonetheless, we may still identify the sign of the causal effect in the absence of homogeneity of the noise.

Key words and phrases: Generalized method of moments, linkage disequilibrium, local average treatment effect, Mendelian randomization, two stage least squares.

REFERENCES

1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* **526** 68–74.

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *J. Econometrics* **113** 231–263. MR1960380

Anderson, T. W. and Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *Ann. Math. Stat.* **20** 46–63. MR0028546

Angrist, J. D., Graddy, K. and Imbens, G. W. (2000). The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *Rev. Econ. Stud.* **67** 499–527.

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

Angrist, J. D. and Krueger, A. B. (1992). The effect of age at school entry on educational attainment: An application of instrumental variables with moments from two samples. *J. Amer.*

Qingyuan Zhao is Postdoctoral Fellow, Department of Statistics, The Wharton School, University of Pennsylvania, USA (e-mail: qyzhao@wharton.upenn.edu). Jingshu Wang is Postdoctoral Fellow, Department of Statistics, The Wharton School, University of Pennsylvania, USA (e-mail: jingshuw@wharton.upenn.edu). Wes Spiller is Ph.D. Student, MRC Integrative Epidemiology Unit, University of Bristol, United Kingdom (e-mail: wes.spiller@bristol.ac.uk). Jack Bowden is Senior Lecturer, MRC Integrative Epidemiology Unit, University of Bristol, United Kingdom (e-mail: jack.bowden@bristol.ac.uk). Dylan S. Small is Professor, Department of Statistics, The Wharton School, University of Pennsylvania, USA (e-mail: dsmall@wharton.upenn.edu).

Statist. Assoc. **87** 328–336.

ANGRIST, J. D. and KRUEGER, A. B. (1995). Split-sample instrumental variables estimates of the return to schooling. *J. Bus. Econom. Statist.* **13** 225–235.

BAIOCCHI, M., CHENG, J. and SMALL, D. S. (2014). Instrumental variable methods for causal inference. *Stat. Med.* **33** 2297–2340. MR3257582

BAKER, S. G. and LINDEMAN, K. S. (1994). The paired availability design: A proposal for evaluating epidural analgesia during labor. *Stat. Med.* **13** 2269–2278.

BALKE, A. and PEARL, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *J. Amer. Statist. Assoc.* **92** 1171–1176.

BARBEIRA, A., DICKINSON, S. P., BONAZZOLA, R., ZHENG, J., WHEELER, H. E., TORRES, J. M., TORSTENSON, E. S., SHAH, K. P., GARCIA, T. et al. (2018). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9** 1825.

BOWDEN, J., DAVEY SMITH, G. and BURGESS, S. (2015). Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44** 512–525.

BOWDEN, J., DAVEY SMITH, G., HAYCOCK, P. C. and BURGESS, S. (2016). Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genet. Epidemiol.* **40** 304–314.

BUJA, A., BERK, R., BROWN, L., GEORGE, E., PITKIN, E., TRASKIN, M., ZHAO, L. and ZHANG, K. (2014). Models as approximations, part I: A conspiracy of nonlinearity and random regressors in linear regression. *Statist. Sci.* Available at arXiv:1404.1578.

BURGESS, S., SMALL, D. S. and THOMPSON, S. G. (2017). A review of instrumental variable estimators for Mendelian randomization. *Stat. Methods Med. Res.* **26** 2333–2355. MR3712236

BURGESS, S., SCOTT, R. A., TIMPSON, N. J., SMITH, G. D., THOMPSON, S. G. and EPIC-INTERACT CONSORTIUM (2015). Using published data in Mendelian randomization: A blueprint for efficient identification of causal risk factors. *Eur. J. Epidemiol.* **30** 543–552.

CHOI, J., GU, J. and SHEN, S. (2018). Weak-instrument robust inference for two-sample instrumental variables regression. *J. Appl. Econometrics* **33** 109–125. MR3771577

CURRIE, J. and YELOWITZ, A. (2000). Are public housing projects good for kids? *J. Public Econ.* **75** 99–124.

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

DAVEY SMITH, G. and HEMANI, G. (2014). Mendelian randomization: Genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* **23** R89–98.

DAVIDSON, R. and MACKINNON, J. G. (1993). *Estimation and Inference in Econometrics*. Oxford University Press, New York. MR1350531

FULLER, W. A. (1977). Some properties of a modification of the limited information estimator. *Econometrica* **45** 939–953. MR0451608

GAMAZON, E. R., WHEELER, H. E., SHAH, K. P., MOZAFFARI, S. V., AQUINO-MICHAELS, K., CARROLL, R. J., EYLER, A. E., DENNY, J. C., NICOLAE, D. L. et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47** 1091–1098.

GRAHAM, B. S., PINTO, C. C. X. and EGEL, D. (2016). Efficient estimation of data combination models by the method of auxiliary-to-study tilting (AST). *J. Bus. Econom. Statist.* **34** 288–301. MR3475879

HAAVELMO, T. (1944). The probability approach in econometrics. *Econometrica* **12** S1–S115. MR0010953

HANSEN, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50** 1029–1054. MR0666123

HANSEN, C., HAUSMAN, J. and NEWEY, W. (2008). Estimation with many instrumental variables. *J. Bus. Econom. Statist.* **26** 398–422. MR2459342

HEMANI, G., ZHENG, J., ELSWORTH, B., WADE, K. H., HABERLAND, V., BAIRD, D., LAURIN, C., BURGESS, S., BOWDEN, J. et al. (2018). The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **7** e34408.

HERNÁN, M. A. and ROBINS, J. M. (2006). Instruments for causal inference: An epidemiologist's dream? *Epidemiology* 360–372.

IMBENS, G. W. (2007). Nonadditive models with endogenous regressors. In *Advances in Economics and Econometrics* (R. Blundell, W. Newey and T. Persson, eds.) **3** 17–46. Cambridge Univ. Press, Cambridge.

IMBENS, G. and ANGRIST, J. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62** 467–475.

INOUE, A. and SOLON, G. (2010). Two-sample instrumental variables estimators. *Rev. Econ. Stat.* **92** 557–561.

JAPPELLI, T., PISCHKE, J.-S. and SOULELES, N. S. (1998). Testing for liquidity constraints in Euler equations with complementary data sources. *Rev. Econ. Stat.* **80** 251–262.

KANG, H., ZHANG, A., CAI, T. T. and SMALL, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *J. Amer. Statist. Assoc.* **111** 132–144. MR3494648

KLEVMARKEN, A. (1982). Missing variables and two-stage least-squares estimation from more than one data set. Working Paper Series 62, Research Institute of Industrial Economics, Stockholm.

LAWLOR, D. A. (2016). Commentary: Two-sample Mendelian randomization: Opportunities and challenges. *Int. J. Epidemiol.* **45** 908–915.

LAWLOR, D. A., HARBORD, R. M., STERNE, J. A. C., TIMPSON, N. and SMITH, G. D. (2008). Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Stat. Med.* **27** 1133–1163. MR2420151

LOCKE, A. E., KAHALI, B., BERNDT, S. I., JUSTICE, A. E., PERS, T. H., DAY, F. R., POWELL, C., VEDANTAM, S., BUCHKOVICH, M. L. et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518** 197–206.

MACHIELA, M. J. and CHANOCK, S. J. (2015). LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31** 3555–3557.

MANOLIO, T. A., COLLINS, F. S., COX, N. J., GOLDSTEIN, D. B., HINDORFF, L. A., HUNTER, D. J., MCCARTHY, M. I., RAMOS, E. M., CARDON, L. R. et al. (2009). Finding the missing heritability of complex diseases. *Nature* **461** 747–753.

OGBURN, E. L., ROTNITZKY, A. and ROBINS, J. M. (2015). Doubly robust estimation of the local average treatment effect curve. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 373–396. MR3310531

PACINI, D. (2018). The two-sample linear regression model with interval-censored covariates. *J. Appl. Econometrics* **34** 66–81.

PACINI, D. and WINDMEIJER, F. (2016). Robust inference for the two-sample 2SLS estimator. *Econom. Lett.* **146** 50–54. MR3542584

PEARL, J. (2009). *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2548166

PETERS, J., BÜHLMANN, P. and MEINSHAUSEN, N. (2016). Causal inference by using invariant prediction: Identification and confidence intervals. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 947–1012. MR3557186

PIERCE, B. L. and BURGESS, S. (2013). Efficient design for Mendelian randomization studies: Subsample and 2-sample instrumental variable estimators. *Am. J. Epidemiol.* **178** 1177–1184.

RIDDER, G. and MOFFITT, R. (2007). The econometrics of data combination. *Handb. Econom.* **6** 5469–5547.

SHERRY, S. T., WARD, M.-H., KHOLODOV, M., BAKER, J., PHAN, L., SMIGIELSKI, E. M. and SIROTKIN, K. (2001). dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.* **29** 308–311.

STOCK, J. H., WRIGHT, J. H. and YOGO, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *J. Bus. Econom. Statist.* **20** 518–529. MR1973801

THEIL, H. (1958). *Economic Forecasts and Policy*. North-Holland, Amsterdam.

VANSTEELANDT, S. and DIDELEZ, V. (2015). Robustness and efficiency of covariate adjusted linear instrumental variable estimators. Preprint. Available at arXiv:1510.01770.

WALD, A. (1940). The fitting of straight lines if both variables are subject to error. *Ann. Math. Stat.* **11** 285–300. MR0002739

WANG, L. and TCHETGEN TCHETGEN, E. (2018). Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 531–550. MR3798877

WHITE, H. (1980). Using least squares to approximate unknown regression functions. *Internat. Econom. Rev.* **21** 149–170. MR0572464

WRIGHT, P. G. (1928). *Tariff on Animal and Vegetable Oils*. Macmillan, New York.

YANG, J., FERREIRA, T., MORRIS, A. P., MEDLAND, S. E., MADDEN, P. A., HEATH, A. C., MARTIN, N. G., MONTGOMERY, G. W., WEEDON, M. N. et al. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44** 369–375.

ZHAO, Q., WANG, J., BOWDEN, J. and SMALL, D. S. (2019). Statistical inference in two-sample summary-data mendelian randomization using robust adjusted profile score. *Ann. Statist.* To appear. Available at arXiv:1801.09652.

Statistical Science
2019, Vol. 34, No. 2, 334–348
https://doi.org/10.1214/18-STS691

A Conversation with Robert E. Kass

Sam Behseta

Abstract. Rob Kass has been been on the faculty of the Department of Statistics at Carnegie Mellon since 1981; he joined the Center for the Neural Basis of Cognition (CNBC) in 1997, and the Machine Learning Department (in the School of Computer Science) in 2007. He served as Department Head of Statistics from 1995 to 2004 and served as Interim Co-Director of the CNBC 2015–2018. He became the Maurice Falk Professor of Statistics and Computational Neuroscience in 2016.

Kass has served as Chair of the Section for Bayesian Statistical Science of the American Statistical Association, Chair of the Statistics Section of the American Association for the Advancement of Science, founding Editor-in-Chief of the journal *Bayesian Analysis* and Executive Editor of *Statistical Science*. He is an elected Fellow of the American Statistical Association, the Institute of Mathematical Statistics and the American Association for the Advancement of Science. He has been recognized by the Institute for Scientific Information as one of the 10 most highly cited researchers, 1995–2005, in the category of mathematics. Kass is the recipient of the 2017 Fisher Award and lectureship by the Committee of the Presidents of the Statistical Societies. This interview took place at Carnegie Mellon University in November 2017.

Key words and phrases: Statistical training, Bayesian statistics, statistics in neuroscience, academic life, statistical narrative.

REFERENCES

BERGER, J. O. and DELAMPADY, M. (1987). Testing precise hypotheses. *Statist. Sci.* **3** 317–352. MR0920141

BERGER, J. O. and SELKE, T. (1987). Testing a point null hypothesis: irreconcilability of P values and evidence. With comments and a rejoinder by the authors. *J. Amer. Statist. Assoc.* **82** 112–139. MR0883340

CRAMÉR, H. (1973). *The Elements of Probability Theory and Some of Its Applications*. Wiley, New York. MR0067379

DEGROOT, M. H. (1986). A conversation with David Blackwell. *Statist. Sci.* **1** 40–53. MR0833274

FELLER, W. (1950). *An Introduction to Probability Theory and Its Applications. Vol. I*. Wiley, New York, NY. MR0038583

FELLER, W. (1957). *An Introduction to Probability Theory and Its Applications. Vol. II*, 2nd ed. Wiley, New York. MR0270403

FERGUSON, T. S. (1967). *Mathematical Statistics: A Decision Theoretic Approach. Probability and Mathematical Statistics, Vol.* 1. Academic Press, New York. MR0215390

JAROSIEWICZ, B. CHASE, S. M. FRASER, G. W. VELLISTE, M. KASS, R. E. and SCHWARTZ, A. B. (2008). Functional network reorganization during learning in a brain-machine interface paradigm. *Proc. Natl. Acad. Sci. USA* **105** 19486–19491.

KASS, R. E. (1989). The geometry of asymptotic inference. *Statist. Sci.* **4** 188–234. With comments and a rejoinder by the author. MR1015274

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

KASS, R. E., KELLY, R. C. and LOH, W.-L. (2011). Assessment of synchrony in multiple neural spike trains using log-linear point process models. *Ann. Appl. Stat.* **5** 1262–1292. MR2849774

KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795. MR3363402

KASS, R. E. and STEFFEY, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *J. Amer. Statist. Assoc.* **84** 717–726. MR1132587

KASS, R., VENTURA, V. and BROWN, E. N. (2005). Statistical issues in the analysis of neuronal data. *J. Neurophysiol.* **1** 8–25.

KASS, R. E. and VOS, P. W. (1997). *Geometrical Foundations of Asymptotic Inference. Wiley Series in Probability and Statistics: Probability and Statistics*. Wiley, New York. MR1461540

KASS, R. E. and WASSERMAN, L. A. (1996). The selection of prior distributions by formal rules. *J. Amer. Statist. Assoc.* **91**

Sam Behseta is Professor of Mathematics, California State University, Fullerton, 800 North College Blvd., Fullerton, California 92831, USA (e-mail: sbehseta@fullerton.edu).

1343–1370.

KASS, R. E., CAFFO, B. S., DAVIDIAN, M., MENG, X.-L., YU, B. and REID, N. (2016). Ten simple rules for effective statistical practice. *PLoS Comput. Biol.* **12** e1004961.

KENDALL, M. and STUART, A. (1977). *The Advanced Theory of Statistics*: *Distribution Theory. Vol.* 1, 4th ed. Macmillan Publishing, New York. MR0467977

MOSTELLER, F. and TUKEY, J. (1968). Data analysis, including statistics. In *Handbook of Social Psychology*, 2nd ed. (G. Lindzey and E. Aronson, eds.) **2**. Wiley, New York.

MOSTELLER, F. and WALLACE, D. L. (1964). *Inference and Disputed Authorship*: *The Federalist*. Addison-Wesley, Reading, MA. MR0175668

RAFTERY, A. (2001). *Statistics in the Twenty First Century*, 1st ed. CRC Press, New York.

RAO, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. Wiley, New York. MR0346957

SELLKE, T., BAYARRI, M. J. and BERGER, J. O. (2001). Calibration of p values for testing precise null hypotheses. *Amer. Statist.* **55** 62–71. MR1818723

TIERNEY, L. (1994). Markov chains for exploring posterior distributions. *Ann. Statist.* **22** 1701–1762. MR1329166

Statistical Science
2019, Vol. 34, No. 2, 349–359
https://doi.org/10.1214/19-STS695

A Conversation with Noel Cressie

Christopher K. Wikle and Jay M. Ver Hoef

Abstract. Noel Cressie, FAA is Director of the Centre for Environmental Informatics in the National Institute for Applied Statistics Research Australia (NIASRA) and Distinguished Professor in the School of Mathematics and Applied Statistics at the University of Wollongong, Australia. He is also Adjunct Professor at the University of Missouri (USA), Affiliate of Org 398, Science Data Understanding, at NASA's Jet Propulsion Laboratory (USA), and a member of the Science Team for NASA's Orbiting Carbon Observatory-2 (OCO-2) satellite. Cressie was awarded a B.Sc. with First Class Honours in Mathematics in 1972 from the University of Western Australia, and an M.A. and Ph.D. in Statistics in 1973 and 1975, respectively, from Princeton University (USA). Two brief postdoctoral periods followed, at the Centre de Morphologie Mathématique, ENSMP, in Fontainebleau (France) from April 1975–September 1975, and at Imperial College, London (UK) from September 1975–January 1976. His past appointments have been at The Flinders University of South Australia from 1976–1983, at Iowa State University (USA) from 1983–1998, and at The Ohio State University (USA) from 1998–2012. He has authored or coauthored four books and more than 280 papers in peer-reviewed outlets, covering areas that include spatial and spatio-temporal statistics, environmental statistics, empirical-Bayesian and Bayesian methods including sequential design, goodness-of-fit, and remote sensing of the environment. Many of his papers also address important questions in the sciences. Cressie is a Fellow of the Australian Academy of Science, the American Statistical Association, the Institute of Mathematical Statistics, and the Spatial Econometrics Association, and he is an Elected Member of the International Statistical Institute. Noel Cressie's refereed, unrefereed, and other publications are available at: https://niasra.uow.edu.au/cei/people/UOW232444.html.

Key words and phrases: Bayesian statistics, geostatistics, spatial statistics, spatio-temporal models.

REFERENCES

BERLINER, L. M. (1996). Hierarchical Bayesian time series models. In *Maximum Entropy and Bayesian Methods* (*Santa Fe, NM*, 1995). *Fund. Theories Phys.* **79** 15–22. Kluwer Academic, Dordrecht. MR1446713

BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc. Ser. B* **36** 192–236. MR0373208

BESAG, J. (1981). On resistant techniques and statistical analysis. *Biometrika* **68** 463–469. MR0626408

CRESSIE, N. (1982). A useful empirical Bayes identity. *Ann. Statist.* **10** 625–629. MR0653538

CRESSIE, N. A. C. (1993). *Statistics for Spatial Data*, rev. ed. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR1239641

CRESSIE, N. (2018). Mission CO_2ntrol: A statistical scientist's role in remote sensing of atmospheric carbon dioxide. *J. Amer. Statist. Assoc.* **113** 152–168. MR3803446

CRESSIE, N. and HAWKINS, D. M. (1980). Robust estimation

Christopher K. Wikle is Curators' Distinguished Professor and Chair, Department of Statistics, University of Missouri, Columbia, Missouri 65211, USA (e-mail: wiklec@missouri.edu) Jay M. Ver Hoef is Senior Scientist and Statistician, Marine Mammal Laboratory, Alaska Fisheries Science Center, NOAA Fisheries, 7600 Sand Point Way NE, Seattle, Washington 98115, USA (e-mail: jay.verhoef@noaa.gov).

of the variogram. I. *J. Int. Assoc. Math. Geol.* **12** 115–125. MR0595404

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

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6** 721–741.

MATHERON, G. (1975). *Random Sets and Integral Geometry*. Wiley, New York. MR0385969

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

RIPLEY, B. D. (1977). Modelling spatial patterns (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 172–212. MR0488279

RIPLEY, B. D. (1981). *Spatial Statistics*. Wiley, New York. MR0624436

VER HOEF, J. M. and CRESSIE, N. (1993). Multivariable spatial prediction. *Math. Geol.* **25** 219–240. MR1206187

WIKLE, C. K., BERLINER, L. M. and CRESSIE, N. (1998). Hierarchical Bayesian space–time models. *Environ. Ecol. Stat.* **5** 117–154.

WIKLE, C. K., ZAMMIT-MANGION, A. and CRESSIE, N. (2019). *Spatio-Temporal Statistics with R*. Chapman & Hall/CRC Press, Boca Raton, FL.

INSTITUTE OF MATHEMATICAL STATISTICS

(Organized September 12, 1935)

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

IMS OFFICERS

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

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

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

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

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

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

IMS 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

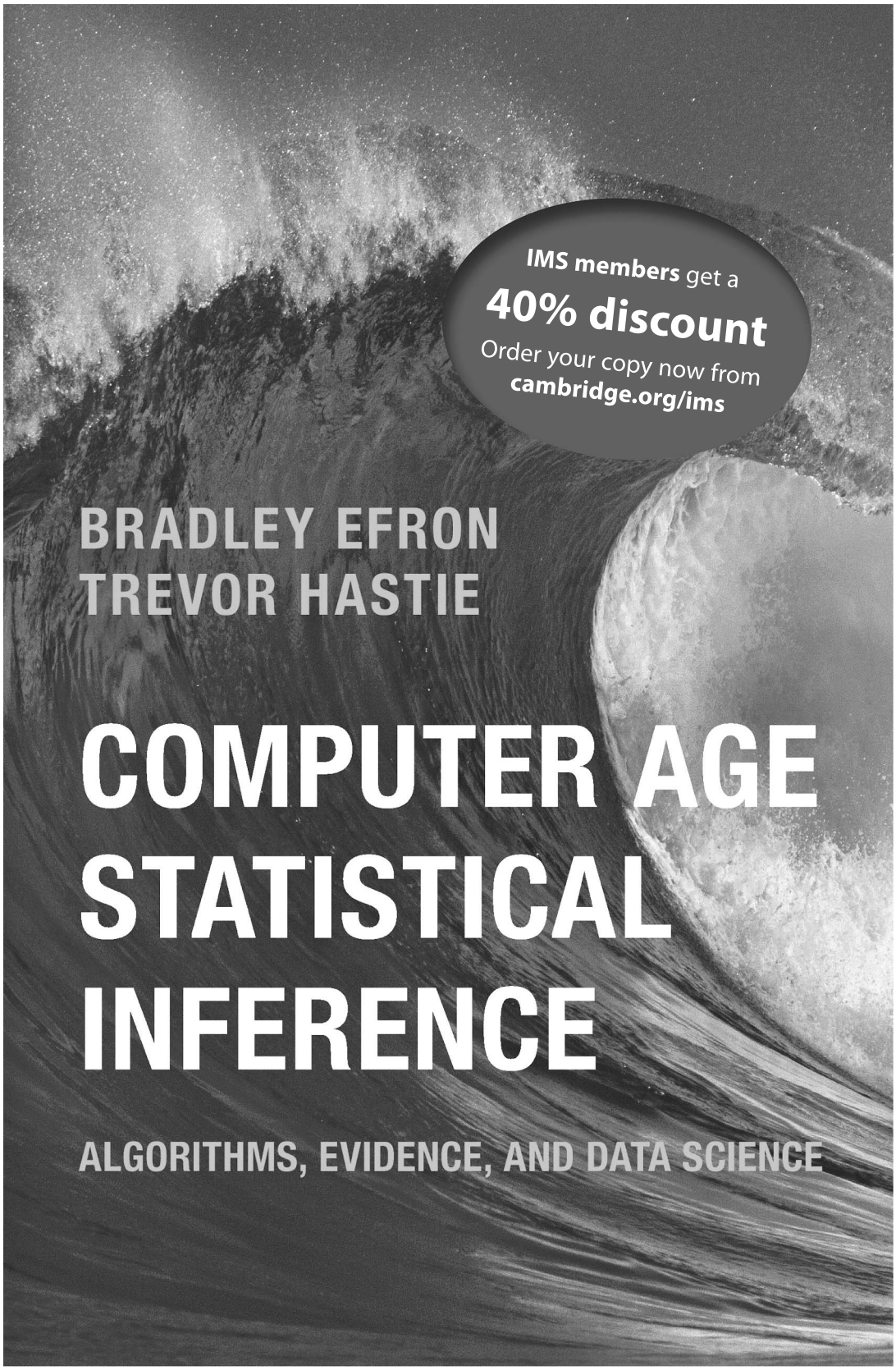
The Annals of Applied Statistics. *Editor-in-Chief*: Karen Kafadar, Department of Statistics, University of Virginia, Heidelberg Institute for Theoretical Studies, Charlottesville, VA 22904-4135, USA

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