STATISTICAL SCIENCE

Volume 39, Number 4 November 2024

Regular Papers

Short Communications

Statistical Science **[ISSN 0883-4237 (print); ISSN 2168-8745 (online)]**, Volume 39, Number 4, November 2024. Published quarterly by the Institute of Mathematical Statistics, 9760 Smith Road, Waite Hill, Ohio 44094, 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, PO Box 729, Middletown, Maryland 21769, USA.

Statistical Science Volume 39, Number 4 (519–653) November 2024

Volume 39 Number 4 November 2024

Statistical Science
2024, Vol. 39, No. 4, 519–554
https://doi.org/10.1214/24-STS945

Protocols for Observational Studies: Methods and Open Problems

Dylan S. Small

Abstract. For learning about the causal effect of a treatment, a randomized controlled trial (RCT) is considered the gold standard. However, randomizing treatment is sometimes unethical or infeasible, and instead an observational study may be conducted. While some aspects of a well-designed RCT cannot be replicated in an observational study, one aspect that can is to have a protocol with prespecified hypotheses about prespecified outcomes and a prespecified analysis. We illustrate the value of protocols for observational studies in three applications—the effect of playing high school football on later life mental functioning, the effect of police seizing a gun when arresting a domestic violence suspect on future domestic violence and the effect of mountaintop mining on health. We then discuss methodologies for observational study protocols. We discuss considerations for protocols that are similar between observational studies and RCTs, and considerations that are different. The considerations that are different include (i) whether the protocol should be specified before treatment assignment is known or after; (ii) how multiple outcomes should be incorporated into the planned analysis and (iii) how subgroups should be incorporated into the planned analysis. We conclude with discussion of a few open problems in the methodology of observational study protocols.

Key words and phrases: Causal inference, sensitivity analysis, matching, planned analysis.

REFERENCES

[1] ALIGHIERI, D. (1320). *Paradiso Canto XIII*, *the Divine Comedy*. Translated by Allen Mandelbaum, 1955.

[2] AUSTIN, P. C., GROOTENDORST, P. and ANDERSON, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study. *Stat. Med.* **26** 734–753. MR2339171 https://doi.org/10.1002/sim.2580

[3] BACON, F. (1620). Novum organum.

[4] BAIOCCHI, M., CHENG, J. and SMALL, D. S. (2014). Instrumental variable methods for causal inference. *Stat. Med.* **33** 2297–2340. MR3257582 https://doi.org/10.1002/sim.6128

[5] BALZER, L. B., VAN DER LAAN, M. J., PETERSEN, M. L. and COLLABORATION SEARCH (2016). Adaptive prespecification in randomized trials with and without pair-matching. *Stat. Med.* **35** 4528–4545. MR3554978 https://doi.org/10.1002/sim.7023

[6] BELANGER, H. G. and VANDERPLOEG, R. D. (2005). The neuropsychological impact of sports-related concussion: A meta-analysis. *J. Int. Neuropsychol. Soc.* **11** 345–357.

[7] BENJAMINI, Y. (2020). Selective inference: The silent killer of replicability. *Harv. Data Sci. Rev.* **2**.

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

[9] BONVINI, M. and KENNEDY, E. H. (2022). Sensitivity analysis via the proportion of unmeasured confounding. *J. Amer. Statist. Assoc.* **117** 1540–1550. MR4480730 https://doi.org/10.1080/01621459.2020.1864382

[10] BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. *Wadsworth Statistics/Probability Series*. Wadsworth Advanced Books and Software, Belmont, CA. MR0726392

[11] BRESLOW, N. E. and DAY, N. E. (1980). Statistical methods in cancer research. Volume I—the analysis of case-control studies. *IARC Sci. Publ.* **32** 5–338.

[12] BROOKHART, M. A., SCHNEEWEISS, S., ROTHMAN, K. J., GLYNN, R. J., AVORN, J. and STÜRMER, T. (2006). Variable selection for propensity score models. *Amer. J. Epidemiol.* **163** 1149–1156. https://doi.org/10.1093/aje/kwj149

[13] BUJA, A. and BROWN, L. (2014). Discussion: "A significance test for the lasso" [MR3210970]. *Ann. Statist.* **42** 509–517. MR3210976 https://doi.org/10.1214/14-AOS1175F

Dylan Small is Universal Furniture Professor, Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania, USA (e-mail: dsmall@wharton.upenn.edu).

[14] BUJA, A., COOK, D., HOFMANN, H., LAWRENCE, M., LEE, E.-K., SWAYNE, D. F. and WICKHAM, H. (2009). Statistical inference for exploratory data analysis and model diagnostics. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **367** 4361–4383. MR2546392 https://doi.org/10.1098/rsta.2009.0120

[15] CAFRI, G. and PAXTON, E. W. (2018). Mitigating reporting bias in observational studies using covariate balancing methods. *Obs. Stud.* **4** 292–296.

[16] CAMPBELL, D. T. Factors relevant to the validity of experiments in social settings. *Psychol. Bull.* **54** 297–312.

[17] CAMPBELL, D. T. (1969). Prospective: Artifact and control. In *Artifacts in Behavioral Research* (R. Rosenthal and R. L. Rosnow, eds.).

[18] CHAN, A.-W., HRÓBJARTSSON, A., HAAHR, M. T., GØTZSCHE, P. C. and ALTMAN, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *JAMA* **291** 2457–2465.

[19] CHAN, A.-W., TETZLAFF, J. M. and GØTZSCHE, P. C. et al. (2013). SPIRIT 2013 explanation and elaboration: Guidance for protocols of clinical trials. *BMJ* **346**.

[20] CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. and ROBINS, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21** C1–C68. MR3769544 https://doi.org/10.1111/ectj.12097

[21] CINELLI, C. and HAZLETT, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 39–67. MR4060976

[22] COCHRAN, W. G. (1965). The planning of observational studies of human populations. *J. R. Stat. Soc., A* **128** 234–266.

[23] COHEN, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Erlbaum, Hillsdale.

[24] COOK, T. D., CAMPBELL, D. T. and SHADISH, W. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* **1195**. Houghton Mifflin, Boston, MA.

[25] COX, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika* **62** 441–444. MR0378189 https://doi.org/10.1093/biomet/62.2.441

[26] COX, D. R. (2016). The design of empirical studies: Towards a unified view. *Eur. J. Epidemiol.* **31** 217–228. https://doi.org/10.1007/s10654-016-0135-9

[27] DANIEL (1953). *The Book of Daniel: In the Holy Bible*. Thomas Nelson & Sons.

[28] DESHPANDE, S. K., HASEGAWA, R. B., RABINOWITZ, A. R., WHYTE, J., ROAN, C. L., TABATABAEI, A., BAIOCCHI, M., KARLAWISH, J. H., MASTER, C. L. et al. (2016). Protocol for an Observational Study on the Effects of Playing High School Football on Later Life Cognitive Functioning and Mental Health. ArXiv preprint. Available at arXiv:1607.01756.

[29] DESHPANDE, S. K., HASEGAWA, R. B., RABINOWITZ, A. R., WHYTE, J., ROAN, C. L., TABATABAEI, A., BAIOCCHI, M., KARLAWISH, J. H., MASTER, C. L. et al. (2017). Association of playing high school football with cognition and mental health later in life. *JAMA Neurol.* **74** 909–918. https://doi.org/10.1001/jamaneurol.2017.1317

[30] DIACONIS, P. (1981). Magical thinking in the analysis of scientific data. *Ann. N.Y. Acad. Sci.* **364** 236–244.

[31] DING, P. and VANDERWEELE, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology* **27** 368.

[32] ELLENBERG, S. (2018). Multiple comparisons in clinical trials. Available at https://www.cdnetwork.org/wp-content/uploads/2018/05/The-Multiple-Comparisons-Problem-in-Clinical-Trials.pdf.

[33] FAY, M. P. and BRITTAIN, E. H. (2022). *Statistical Hypothesis Testing in Context—Reproducibility, Inference, and Science. Cambridge Series in Statistical and Probabilistic Mathematics* **52**. Cambridge Univ. Press, Cambridge. MR4586424

[34] FISHER, R. A. (1935). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

[35] FOGARTY, C. (2023). Sensitivity analysis. In *Handbook of Matching and Weighting Adjustments for Causal Inference* 553–582. CRC Press/CRC, Boca Raton, FL.

[36] GAUTRET, P., LAGIER, J.-C., PAROLA, P., MEDDEB, L., MAILHE, M., DOUDIER, B., COURJON, J., GIORDANENGO, V., VIEIRA, V. E. et al. (2020). Hydroxychloroquine and azithromycin as a treatment of COVID-19: Results of an open-label non-randomized clinical trial. *Int. J. Antimicrob. Agents* **56** 105949.

[37] GOEMAN, J. J., SOLARI, A. and STIJNEN, T. (2010). Three-sided hypothesis testing: Simultaneous testing of superiority, equivalence and inferiority. *Stat. Med.* **29** 2117–2125. MR2756559 https://doi.org/10.1002/sim.4002

[38] GRUBER, S., LOGAN, R. W., JARRÍN, I., MONGE, S. and HERNÁN, M. A. (2015). Ensemble learning of inverse probability weights for marginal structural modeling in large observational datasets. *Stat. Med.* **34** 106–117. MR3286242 https://doi.org/10.1002/sim.6322

[39] GUO, F. R. and ZHAO, Q. (2023). Confounder selection via iterative graph expansion.

[40] HAHN, P. R., MURRAY, J. S. and CARVALHO, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Anal.* **15** 965–1056. MR4154846 https://doi.org/10.1214/19-BA1195

[41] HAMMOND, E. C. (1964). Smoking in relation to mortality and morbidity. Findings in first thirty-four months of follow-up in a prospective study started in 1959. *J. Natl. Cancer Inst.* **32** 1161–1188.

[42] HANSEN, B. B. (2004). Full matching in an observational study of coaching for the SAT. *J. Amer. Statist. Assoc.* **99** 609–618. MR2086387 https://doi.org/10.1198/016214504000000647

[43] HANSEN, B. B. (2011). Propensity score matching to extract latent experiments from nonexperimental data: A case study. In *Looking Back* 149–181. Springer, Berlin.

[44] HANSEN, B. B., ROSENBAUM, P. R. and SMALL, D. S. (2014). Clustered treatment assignments and sensitivity to unmeasured biases in observational studies. *J. Amer. Statist. Assoc.* **109** 133–144. MR3180552 https://doi.org/10.1080/01621459.2013.863157

[45] HARDER, V. S., STUART, E. A. and ANTHONY, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychol. Methods* **15** 234.

[46] HASEGAWA, R. B., DESHPANDE, S. K., SMALL, D. S. and ROSENBAUM, P. R. (2020). Causal inference with two versions of treatment. *J. Educ. Behav. Stat.* **45** 426–445.

[47] HELLER, R., ROSENBAUM, P. R. and SMALL, D. S. (2009). Split samples and design sensitivity in observational studies. *J. Amer. Statist. Assoc.* **104** 1090–1101. MR2750238 https://doi.org/10.1198/jasa.2009.tm08338

[48] HENDRYX, M. and HOLLAND, B. (2016). Unintended consequences of the clean air act: Mortality rates in Appalachian coal mining communities. *Environ. Sci. Policy* **63** 1–6.

[49] HENG, S., KANG, H., SMALL, D. S. and FOGARTY, C. B. (2021). Increasing power for observational studies of aberrant response: An adaptive approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **83** 482–504. MR4294541 https://doi.org/10.1111/rssb.12424

[50] HERD, P., CARR, D. and ROAN, C. (2014). Cohort profile: Wisconsin longitudinal study (WLS). *Int. J. Epidemiol.* **43** 34–41. https://doi.org/10.1093/ije/dys194

[51] HERNÁN, M. A. and ROBINS, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *Amer. J. Epidemiol.* **183** 758–764.

[52] HERNÁN, M. A. and ROBINS, J. M. (2020). *Causal Inference: What If*. CRC Press/CRC, Boca Raton, FL.

[53] HERNÁN, M. A., SAUER, B. C., HERNÁNDEZ-DÍAZ, S., PLATT, R. and SHRIER, I. (2016). Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *J. Clin. Epidemiol.* **79** 70–75. https://doi.org/10.1016/j.jclinepi.2016.04.014

[54] HILL, J. L. (2011). Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Statist.* **20** 217–240. MR2816546 https://doi.org/10.1198/jcgs.2010.08162

[55] HILL, N., FRAPPIER-DAVIGNON, L. and MORRISON, B. (1979). The periodic health examination. *CMAJ, Can. Med. Assoc. J.* **121** 1193–254.

[56] HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6** 65–70. MR0538597

[57] HOSMAN, C. A., HANSEN, B. B. and HOLLAND, P. W. (2010). The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. *Ann. Appl. Stat.* **4** 849–870. MR2758424 https://doi.org/10.1214/09-AOAS315

[58] HOUSE, S. and HOWARD, J. (2009). *Something's Rising: Appalachians Fighting Mountaintop Removal*. Taylor & Francis, London.

[59] HSU, J. Y., SMALL, D. S. and ROSENBAUM, P. R. (2013). Effect modification and design sensitivity in observational studies. *J. Amer. Statist. Assoc.* **108** 135–148. MR3174608 https://doi.org/10.1080/01621459.2012.742018

[60] HSU, J. Y., ZUBIZARRETA, J. R., SMALL, D. S. and ROSENBAUM, P. R. (2015). Strong control of the familywise error rate in observational studies that discover effect modification by exploratory methods. *Biometrika* **102** 767–782. MR3431552 https://doi.org/10.1093/biomet/asv034

[61] IMBENS, G. W. (2014). Instrumental variables: An econometrician's perspective. *Statist. Sci.* **29** 323–358. MR3264545 https://doi.org/10.1214/14-STS480

[62] IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference—for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge Univ. Press, New York. MR3309951 https://doi.org/10.1017/CBO9781139025751

[63] JOGDEO, K. (1977). Association and probability inequalities. *Ann. Statist.* **5** 495–504. MR0448703

[64] KARMAKAR, B., HELLER, R. and SMALL, D. S. (2018). False discovery rate control for effect modification in observational studies. *Electron. J. Stat.* **12** 3232–3253. MR3860111 https://doi.org/10.1214/18-EJS1476

[65] KARMAKAR, B. and SMALL, D. S. (2020). Assessment of the extent of corroboration of an elaborate theory of a causal hypothesis using partial conjunctions of evidence factors. *Ann. Statist.* **48** 3283–3311. MR4185809 https://doi.org/10.1214/19-AOS1929

[66] KOSOROK, M. R. and LABER, E. B. (2019). Precision medicine. *Annu. Rev. Stat. Appl.* **6** 263–286. MR3939521 https://doi.org/10.1146/annurev-statistics-030718-105251

[67] KUCHIBHOTLA, A. K., KOLASSA, J. E. and KUFFNER, T. A. (2022). Post-selection inference. *Annu. Rev. Stat. Appl.* **9** 505–527. MR4394918 https://doi.org/10.1146/annurev-statistics-100421-044639

[68] LACKLAND, D. T. (2014). Racial differences in hypertension: Implications for high blood pressure management. *Am. J. Med. Sci.* **348** 135–138. https://doi.org/10.1097/MAJ.0000000000000308

[69] LAWRENCE, R. (1989). US Preventive Services Task Force Edition. Guide to Clinical Preventive Services DIANE Publishing.

[70] LEE, K., SMALL, D. S. and DOMINICI, F. (2021). Discovering heterogeneous exposure effects using randomization inference in air pollution studies. *J. Amer. Statist. Assoc.* **116** 569–580. MR4270004 https://doi.org/10.1080/01621459.2020.1870476

[71] LEE, K., SMALL, D. S., HSU, J. Y., SILBER, J. H. and ROSENBAUM, P. R. (2018). Discovering effect modification in an observational study of surgical mortality at hospitals with superior nursing. *J. Roy. Statist. Soc. Ser. A* **181** 535–546. MR3749529 https://doi.org/10.1111/rssa.12298

[72] LEE, K., SMALL, D. S. and ROSENBAUM, P. R. (2018). A powerful approach to the study of moderate effect modification in observational studies. *Biometrics* **74** 1161–1170. MR3908134 https://doi.org/10.1111/biom.12884

[73] LEHMAN, D. R., WORTMAN, C. B. and WILLIAMS, A. F. (1987). Long-term effects of losing a spouse or child in a motor vehicle crash. *J. Pers. Soc. Psychol.* **52** 218.

[74] LIPKOVICH, I., DMITRIENKO, A. and D'AGOSTINO, R. B. SR. (2017). Tutorial in biostatistics: Data-driven subgroup identification and analysis in clinical trials. *Stat. Med.* **36** 136–196. MR3580950 https://doi.org/10.1002/sim.7064

[75] MARSH, J. L., HUTTON, J. L. and BINKS, K. (2002). Removal of radiation dose response effects: An example of overmatching. *BMJ* **325** 327–330. https://doi.org/10.1136/bmj.325.7359.327

[76] MILL, J. (1859). On liberty. Reprinted.

[77] MOLINEAUX, L., GRAMICCIA, G., ORGANIZATION, W. H. et al. (1980). *The Garki Project: Research on the Epidemiology and Control of Malaria in the Sudan Savanna of West Africa*. World Health Organization, Paris.

[78] MOSES, L. E. (1995). Measuring effects without randomized trials? Options, problems, challenges. *Med. Care* **33** AS8–AS14.

[79] MYERS, J. A., RASSEN, J. A., GAGNE, J. J., HUYBRECHTS, K. F., SCHNEEWEISS, S., ROTHMAN, K. J., JOFFE, M. M. and GLYNN, R. J. (2011). Effects of adjusting for instrumental variables on bias and precision of effect estimates. *Amer. J. Epidemiol.* **174** 1213–1222. https://doi.org/10.1093/aje/kwr364

[80] NAGGARA, O., RAYMOND, J., GUILBERT, F. and ALTMAN, D. G. (2011). The problem of subgroup analyses: An example from a trial on ruptured intracranial aneurysms. *Am. J. Neuroradiol.* **32** 633–636. https://doi.org/10.3174/ajnr.A2442

[81] NIKNAM, B. A., ARRIAGA, A. F., ROSENBAUM, P. R., HILL, A. S., ROSS, R. N., EVEN-SHOSHAN, O., ROMANO, P. S. and SILBER, J. H. (2018). Adjustment for atherosclerosis diagnosis distorts the effects of percutaneous coronary intervention and the ranking of hospital performance. *J. Amer. Heart Assoc.* **7**. https://doi.org/10.1161/JAHA.117.008366

[82] PACKARD, E. (1979). *The Cave of Time* **1**. Bantam Books, New York..

[83] PATRICK, A. R., SCHNEEWEISS, S., BROOKHART, M. A., GLYNN, R. J., ROTHMAN, K. J., AVORN, J. and STÜRMER, T. (2011). The implications of propensity score variable selection strategies in pharmacoepidemiology: An empirical illustration. *Pharmacoepidemiol. Drug Saf.* **20** 551–559. https://doi.org/10.1002/pds.2098

[84] PERICAK, A. A., THOMAS, C. J., KROODSMA, D. A., WASSON, M. F., ROSS, M. R., CLINTON, N. E., CAMPAGNA, D. J., FRANKLIN, Y., BERNHARDT, E. S. et al. (2018). Mapping the yearly extent of surface coal mining in central Appalachia using landsat and Google Earth engine. *PLoS ONE* **13** e0197758.

[85] PETO, R., COLLINS, R. and GRAY, R. (1995). Large-scale randomized evidence: Large, simple trials and overviews of trials. *J. Clin. Epidemiol.* **48** 23–40.

[86] PIMENTEL, S. D., KELZ, R. R., SILBER, J. H. and ROSENBAUM, P. R. (2015). Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. *J. Amer. Statist. Assoc.* **110** 515–527. MR3367244 https://doi.org/10.1080/01621459.2014.997879

[87] PIMENTEL, S. D., SMALL, D. S. and ROSENBAUM, P. R. (2016). Constructed second control groups and attenuation of unmeasured biases. *J. Amer. Statist. Assoc.* **111** 1157–1167. MR3561939 https://doi.org/10.1080/01621459.2015.1076342

[88] POCOCK, S. J. and STONE, G. W. (2016). The primary outcome fails—what next? *N. Engl. J. Med.* **375** 861–870.

[89] POPPER, K. R. (1972). *Objective Knowledge* **360**. Oxford Univ. Press, Oxford.

[90] RADLOFF, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Appl. Psychol. Meas.* **1** 385–401.

[91] RAMANI, R. V. (2012). Surface mining technology: Progress and prospects. *Proc. Eng.* **46** 9–21.

[92] RAMSEY, F. and SCHAFER, D. (2002). *The Statistical Sleuth: A Course in Methods of Data Analysis, Second Edition*. Duxbury, N. Scituate.

[93] ROSENBAUM, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *J. R. Stat. Soc., A* **147** 656–666.

[94] ROSENBAUM, P. R. (1987). The role of a second control group in an observational study. *Statist. Sci.* **2** 292–306.

[95] ROSENBAUM, P. R. (1991). A characterization of optimal designs for observational studies. *J. Roy. Statist. Soc. Ser. B* **53** 597–610. MR1125717

[96] ROSENBAUM, P. R. (1995). Quantiles in nonrandom samples and observational studies. *J. Amer. Statist. Assoc.* **90** 1424–1431. MR1379486

[97] ROSENBAUM, P. R. (1999). Choice as an alternative to control in observational studies. *Statist. Sci.* **14** 259–304.

[98] ROSENBAUM, P. R. (2002). *Observational Studies*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR1899138 https://doi.org/10.1007/978-1-4757-3692-2

[99] ROSENBAUM, P. R. (2002). Covariance adjustment in randomized experiments and observational studies. *Statist. Sci.* **17** 286–327. MR1962487 https://doi.org/10.1214/ss/1042727942

[100] ROSENBAUM, P. R. (2004). Design sensitivity in observational studies. *Biometrika* **91** 153–164. MR2050466 https://doi.org/10.1093/biomet/91.1.153

[101] ROSENBAUM, P. R. (2005). Sensitivity analysis in observational studies. *Encyclopedia Stat. Behav. Sci.*

[102] ROSENBAUM, P. R. (2008). Testing hypotheses in order. *Biometrika* **95** 248–252. MR2409727 https://doi.org/10.1093/biomet/asm085

[103] ROSENBAUM, P. R. (2010). *Design of Observational Studies*. *Springer Series in Statistics*. Springer, New York. MR2561612 https://doi.org/10.1007/978-1-4419-1213-8

[104] ROSENBAUM, P. R. (2012). An exact adaptive test with superior design sensitivity in an observational study of treatments for ovarian cancer. *Ann. Appl. Stat.* **6** 83–105. MR2951530 https://doi.org/10.1214/11-AOAS508

[105] ROSENBAUM, P. R. (2012). Testing one hypothesis twice in observational studies. *Biometrika* **99** 763–774. MR2999159 https://doi.org/10.1093/biomet/ass032

[106] ROSENBAUM, P. R. (2020). Modern algorithms for matching in observational studies. *Annu. Rev. Stat. Appl.* **7** 143–176. MR4104189 https://doi.org/10.1146/annurev-statistics-031219-041058

[107] ROSENBAUM, P. R. (2022). A new transformation of treated-control matched-pair differences for graphical display. *Amer. Statist.* **76** 346–352. MR4505940 https://doi.org/10.1080/00031305.2022.2063944

[108] ROSENBAUM, P. R. (2023). Propensity score. In *Handbook of Matching and Weighting Adjustments for Causal Inference* 21–38. CRC Press/CRC, Boca Raton, FL.

[109] ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974 https://doi.org/10.1093/biomet/70.1.41

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

[111] ROSENBAUM, P. R. and RUBIN, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Amer. Statist.* **39** 33–38.

[112] ROSENBAUM, P. R. and SILBER, J. H. (2001). Matching and thick description in an observational study of mortality after surgery. *Biostatistics* **2** 217–232.

[113] ROSENBAUM, P. R. and SILBER, J. H. (2009). Amplification of sensitivity analysis in matched observational studies. *J. Amer. Statist. Assoc.* **104** 1398–1405. MR2750570 https://doi.org/10.1198/jasa.2009.tm08470

[114] ROSENBAUM, P. R. and SMALL, D. S. (2017). An adaptive Mantel–Haenszel test for sensitivity analysis in observational studies. *Biometrics* **73** 422–430. MR3665959 https://doi.org/10.1111/biom.12591

[115] ROSENDAAL, F. R. (2020). Review of:"Hydroxychloroquine and azithromycin as a treatment of COVID-19: Results of an open-label non-randomized clinical trial Gautret et al 2010. *Int. J. Antimicrob. Agents* **56** 106063.

[116] RUBIN, D., DUDOIT, S. and VAN DER LAAN, M. (2006). A method to increase the power of multiple testing procedures through sample splitting. *Stat. Appl. Genet. Mol. Biol.* **5** Art. 19. MR2240850 https://doi.org/10.2202/1544-6115.1148

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

[118] RUBIN, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Ann. Intern. Med.* **127** 757–763. https://doi.org/10.7326/0003-4819-127-8_part_2-199710151-00064

[119] RUBIN, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Stat. Med.* **26** 20–36. MR2312697 https://doi.org/10.1002/sim.2739

[120] SCHNEEWEISS, S., RASSEN, J. A., GLYNN, R. J., AVORN, J., MOGUN, H. and BROOKHART, M. A. (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* **20** 512.

[121] SCHOPENHAUER, A. (1844). The world as will and representation. Volume 2, translated by David Carus and Richard Aquila. Translation published in 2011.

[122] SILBER, J. H., ROSENBAUM, P. R., CLARK, A. S., GIANTONIO, B. J., ROSS, R. N., TENG, Y., WANG, M., NIKNAM, B. A., LUDWIG, J. M. et al. (2013). Characteristics associated with differences in survival among black and white women with breast cancer. *JAMA* **310** 389–397.

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

[124] SMALL, D. S (2024). Supplement to "Protocols for Observational Studies: Methods and Open Problems." https://doi.org/10.1214/24-STS945SUPP
[125] SMALL, D. S., FIRTH, D., KEELE, L., HUBER, M., PASSARELLA, M., LORCH, S. and BURRIS, H. (2020). Protocol for a Study of the Effect of Surface Mining in Central Appalachia on Adverse Birth Outcomes. ArXiv preprint. Available at 2005.01873.
[126] SMALL, D. S., FIRTH, D. W., KEELE, L. J., HUBER, M., PASSARELLA, M., LORCH, S. A. and BURRIS, H. H. (2021). Surface mining and low birth weight in central Appalachia. *Environ. Res.* **196** 110340. https://doi.org/10.1016/j.envres.2020.110340
[127] SMALL, D. S., SORENSON, S. B. and BERK, R. A. (2019). After the gun: Examining police visits and intimate partner violence following incidents involving a firearm. *J. Behav. Med.* **42** 591–602. https://doi.org/10.1007/s10865-019-00013-8
[128] SMITH, L. H., GARCÍA-ALBÉNIZ, X., CHAN, J. M., ZHAO, S., COWAN, J. E., BROERING, J. M., COOPERBERG, M. R., CARROLL, P. R. and HERNÁN, M. A. (2022). Emulation of a target trial with sustained treatment strategies: An application to prostate cancer using both inverse probability weighting and the g-formula. *Eur. J. Epidemiol.* **37** 1205–1213.
[129] STANLEY, K. (2007). Design of randomized controlled trials. *Circulation* **115** 1164–1169. https://doi.org/10.1161/CIRCULATIONAHA.105.594945
[130] STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statist. Sci.* **25** 1–21. MR2741812 https://doi.org/10.1214/09-STS313
[131] SU, X., TSAI, C.-L., WANG, H., NICKERSON, D. M. and LI, B. (2009). Subgroup analysis via recursive partitioning. *J. Mach. Learn. Res.* **10**.
[132] TAUBES, G. (1995). Epidemiology faces its limits: The search for subtle links between diet, lifestyle, or environmental factors and disease is an unending source of fear—but often yields little certainty. *Science* **269** 164–169.
[133] TUKEY, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading.
[134] TUKEY, J. W. (1977). Some thoughts on clinical trials, especially problems of multiplicity. *Science* **198** 679–684. https://doi.org/10.1126/science.333584
[135] TUKEY, J. W. (1980). We need both exploratory and confirmatory. *Amer. Statist.* **34** 23–25.
[136] TUKEY, J. W. (1991). The philosophy of multiple comparisons. *Statist. Sci.* 100–116.
[137] VANDERWEELE, T. J. (2019). Principles of confounder selection. *Eur. J. Epidemiol.* **34** 211–219. https://doi.org/10.1007/s10654-019-00494-6
[138] VANDERWEELE, T. J. and SHPITSER, I. (2013). On the definition of a confounder. *Ann. Statist.* **41** 196–220. MR3059415 https://doi.org/10.1214/12-AOS1058
[139] VANSTEELANDT, S. and DANIEL, R. M. (2014). On regression adjustment for the propensity score. *Stat. Med.* **33** 4053–4072. MR3261062 https://doi.org/10.1002/sim.6207
[140] VAN DER LAAN, M. J. and ROSE, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. *Springer Series in Statistics*. Springer, New York. MR2867111 https://doi.org/10.1007/978-1-4419-9782-1
[141] WAGER, S. and ATHEY, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* **113** 1228–1242. MR3862353 https://doi.org/10.1080/01621459.2017.1319839
[142] WANG, S. V., POTTEGÅRD, A., CROWN, W., ARLETT, P., ASHCROFT, D. M., BENCHIMOL, E. I., BERGER, M. L., CRANE, G., GOETTSCH, W. et al. (2022). HARmonized protocol template to enhance reproducibility of hypothesis evaluating real-world evidence studies on treatment effects: A good practices report of a joint ISPE/ISPOR task force. *Value Health* **25** 1663–1672.
[143] WANG, S. V., SREEDHARA, S. and SCHNEEWEISS, S. et al. (2022). Reproducibility of real-world evidence studies using clinical practice data to inform regulatory and coverage decisions. *Nat. Commun.* **13** 1–11.
[144] WARD JR., K. Trump's interior department moves to stop mountaintop removal study. *Charleston Gazette-Mail*.
[145] WEI LOH, W. and VANSTEELANDT, S. (2021). Confounder selection strategies targeting stable treatment effect estimators. *Stat. Med.* **40** 607–630. MR4198435 https://doi.org/10.1002/sim.8792
[146] WILDE, E. A., WHITENECK, G. G., BOGNER, J., BUSHNIK, T., CIFU, D. X., DIKMEN, S., FRENCH, L., GIACINO, J. T., HART, T. et al. (2010). Recommendations for the use of common outcome measures in traumatic brain injury research. *Arch. Phys. Med. Rehabil.* **91** 1650–1660.
[147] WITTE, J. and DIDELEZ, V. (2019). Covariate selection strategies for causal inference: Classification and comparison. *Biom. J.* **61** 1270–1289. MR4013348 https://doi.org/10.1002/bimj.201700294
[148] WITTES, J. T. (2019). Commentary on randomization: The forgotten component of the randomized clinical trial. *Stat. Med.* **38** 14–16.
[149] WOOLDRIDGE, J. M. (2016). Should instrumental variables be used as matching variables? *Res. Econ.* **70** 232–237.
[150] YU, R., SILBER, J. H. and ROSENBAUM, P. R. (2020). Matching methods for observational studies derived from large administrative databases. *Statist. Sci.* **35** 338–355. MR4148206 https://doi.org/10.1214/19-STS699
[151] YU, R., SMALL, D. S., HARDING, D., AVELDANES, J. and ROSENBAUM, P. R. (2024). Optimal matching for observational studies that integrate quantitative and qualitative research. *Stat. Public Policy* **8** 45–52.
[152] YU, R., SMALL, D. S. and ROSENBAUM, P. R. (2021). The information in covariate imbalance in studies of hormone replacement therapy. *Ann. Appl. Stat.* **15** 2023–2042. MR4355087 https://doi.org/10.1214/21-aoas1448
[153] ZHANG, B. and SMALL, D. S. (2020). A calibrated sensitivity analysis for matched observational studies with application to the effect of second-hand smoke exposure on blood lead levels in children. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **69** 1285–1305. MR4166867 https://doi.org/10.1111/rssc.12443
[154] ZHANG, B., WEISS, J., SMALL, D. S. and ZHAO, Q. (2021). Selecting and ranking individualized treatment rules with unmeasured confounding. *J. Amer. Statist. Assoc.* **116** 295–308. MR4227695 https://doi.org/10.1080/01621459.2020.1736083
[155] ZHANG, H. and SINGER, B. H. (2010). *Recursive Partitioning and Applications*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2674991 https://doi.org/10.1007/978-1-4419-6824-1
[156] ZHANG, J., ZHANG, B. and SMALL, D. S. (2021). A method to aid statistical judgment on outliers: Comment on Hill's the statistician in medicine. *Stat. Med.* **40** 58–63.
[157] ZHANG, K., SMALL, D. S., LORCH, S., SRINIVAS, S. and ROSENBAUM, P. R. (2011). Using split samples and evidence factors in an observational study of neonatal outcomes. *J. Amer. Statist. Assoc.* **106** 511–524. MR2847966 https://doi.org/10.1198/jasa.2011.ap10604

[158] ZHAO, Q., SMALL, D. S. and ROSENBAUM, P. R. (2018). Cross-screening in observational studies that test many hypotheses. *J. Amer. Statist. Assoc.* **113** 1070–1084. MR3862340 https://doi.org/10.1080/01621459.2017.1407770

Statistical Science
2024, Vol. 39, No. 4, 555–559
https://doi.org/10.1214/24-STS942

Comment on “Protocols for Observational Studies”

Ben B. Hansen

REFERENCES

[1] AIKENS, R. C., GREAVES, D. and BAIOCCHI, M. (2020). A pilot design for observational studies: Using abundant data thoughtfully. *Stat. Med.* **39** 4821–4840. MR4190948 https://doi.org/10.1002/sim.8754

[2] BENJAMINI, Y. (2020). Selective inference: The silent killer of replicability. *Harv. Data Sci. Rev.* **2**.

[3] BERGER, R. L., BOOS, D. D. and GUESS, F. M. (1988). Tests and confidence sets for comparing two mean residual life functions. *Biometrics* **44** 103–115. MR0931630 https://doi.org/10.2307/2531899

[4] COCHRAN, W. G. (1972). Observational studies. In *Statistical Papers in Honor of George W. Snedecor* 77–90. Iowa State Univ. Press, Ames, IA. MR0451569

[5] COHEN, P. L., OLSON, M. A. and FOGARTY, C. B. (2020). Multivariate one-sided testing in matched observational studies as an adversarial game. *Biometrika* **107** 809–825. MR4186489 https://doi.org/10.1093/biomet/asaa024

[6] GELMAN, A. and LOKEN, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. Technical Report No. 348, Dept. Statistics, Columbia Univ.

[7] HANSEN, B. B. (2008). The essential role of balance tests in propensity-matched observational studies: Comments on ‘A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003’ by Peter Austin, *Statistics in Medicine* [MR2439882]. *Stat. Med.* **27** 2050–2054. MR2439883 https://doi.org/10.1002/sim.3208

[8] HANSEN, B. B. and SALES, A. C. (2015). Comments on ‘Observational studies,’ by William G. Cochran. *Obs. Stud.* **1** 184–193.

[9] HARWICKE, T. E. and VAZIRE, S. (2023). Editorial: Transparency is now the default at psychological science. *Psychol. Sci.* https://doi.org/10.1177/09567976231221573

[10] HERNÁN, M. A. and ROBINS, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *Amer. J. Epidemiol.* **183** 758–764.

[11] HO, D. E., IMAI, K., KING, G. and STUART, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit. Anal.* **15** 199–236.

[12] KERR, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personal. Soc. Psychol. Rev.* **2** 196–217. https://doi.org/10.1207/s15327957pspr0203_4

[13] LESSER, M. L., AKERMAN, M. B. and KOHN, N. (2016). Analogies for helping clinicians and investigators better understand the principles and practice of biostatistics. *Amer. Statist.* **70** 166–170. MR3511046 https://doi.org/10.1080/00031305.2015.1073625

[14] LYCURGUS, T., HANSEN, B. B. and WHITE, M. (2023). Conjuring power from a theory of change: The PWRD method for trials with anticipated variation in effects. *J. Res. Educ. Eff.* **16** 707–733.

[15] MANN, C. Z., HANSEN, B. B. and GAYDOSH, L. (2024). Early effects of 2014 U.S. medicaid expansions on mortality: Design-based inference for impacts on small subgroups despite small-cell suppression. *Ann. Appl. Stat.* to appear.

[16] MANN, C. Z., HANSEN, B. B., GAYDOSH, L. and LYCURGUS, T. (2021). Protocol—evaluating the effect of ACA medicaid expansion on mortality during the COVID-19 pandemic using county-level matching. *Obs. Stud.* **7** S1–S31. https://doi.org/10.1353/obs.2021.0034

[17] NIKNAM, B. A., ARRIAGA, A. F., ROSENBAUM, P. R., HILL, A. S., ROSS, R. N., EVEN-SHOSHAN, O., ROMANO, P. S. and SILBER, J. H. (2018). Adjustment for atherosclerosis diagnosis distorts the effects of percutaneous coronary intervention and the ranking of hospital performance. *J. Amer. Heart Assoc.* **7**. https://doi.org/10.1161/JAHA.117.008366

[18] ROSENBAUM, P. R. (2008). Testing hypotheses in order. *Biometrika* **95** 248–252. MR2409727 https://doi.org/10.1093/biomet/asm085

[19] ROSENBAUM, P. R. (2010). *Design of Observational Studies. Springer Series in Statistics*. Springer, New York. MR2561612 https://doi.org/10.1007/978-1-4419-1213-8

[20] RUBIN, D. B. (2008). For objective causal inference, design trumps analysis. *Ann. Appl. Stat.* **2** 808–840. MR2516795 https://doi.org/10.1214/08-AOAS187

[21] TUKEY, J. W. (1980). We need both exploratory and confirmatory. *Amer. Statist.* **34** 23–25.

[22] VANDENBROUCKE, J. P., ELM, E. V., ALTMAN, D. G., GØTZSCHE, P. C., MULROW, C. D., POCOCK, S. J., POOLE, C., SCHLESSELMAN, J. J., EGGER, M. et al. (2007). Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and elaboration. *Ann. Intern. Med.* **147** W-163–W-194.

Ben B. Hansen is Professor, Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109, USA (e-mail: bbh@umich.edu).

Statistical Science
2024, Vol. 39, No. 4, 560–565
https://doi.org/10.1214/24-STS943

Comment: Protocols for Observational Studies: An Application to Regression Discontinuity Designs

Matias D. Cattaneo and Rocío Titiunik

Abstract. In his 2022 IMS Medallion Lecture delivered at the Joint Statistical Meetings, Prof. Dylan S. Small eloquently advocated for the use of protocols in observational studies. We discuss his proposal and, inspired by his ideas, we develop a protocol for the regression discontinuity design.

Key words and phrases: Pre-registration plans, observational studies, causal inference, regression discontinuity designs.

REFERENCES

[1] CALONICO, S., CATTANEO, M. D. and FARRELL, M. H. (2020). Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs. *Econom. J.* **23** 192–210. MR4108924 https://doi.org/10.1093/ectj/utz022

[2] CATTANEO, M. D., FRANDSEN, B. R. and TITIUNIK, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the U.S. Senate. *J. Causal Inference* **3** 1–24. MR4289424 https://doi.org/10.1515/jci-2013-0010

[3] CATTANEO, M. D., IDROBO, N. and TITIUNIK, R. (2020). *A Practical Introduction to Regression Discontinuity Designs*: *Foundations*. *Cambridge Elements*: *Quantitative and Computational Methods for Social Science*. Cambridge Univ. Press, Cambridge.

[4] CATTANEO, M. D., IDROBO, N. and TITIUNIK, R. (2024). *A Practical Introduction to Regression Discontinuity Designs*: *Extensions*. *Cambridge Elements*: *Quantitative and Computational Methods for Social Science*. Cambridge Univ. Press, Cambridge.

[5] CATTANEO, M. D., KEELE, L. and TITIUNIK, R. (2023). A guide to regression discontinuity designs in medical applications. *Stat. Med.* **42** 4484–4513. MR4653504 https://doi.org/10.1002/sim.9861

[6] CATTANEO, M. D., KEELE, L. and TITIUNIK, R. (2023). Covariate adjustment in regression discontinuity designs. In *Handbook of Matching and Weighting in Causal Inference* (D. S. S. J. R. Zubizarreta, E. A. Stuart and P. R. Rosenbaum, eds.) **8** 153–168. CRC Press, Boca Raton, FL.

[7] CATTANEO, M. D. and TITIUNIK, R. (2022). Regression discontinuity designs. *Ann. Rev. Econ.* **14** 821–851.

[8] CATTANEO, M. D., TITIUNIK, R. and VAZQUEZ-BARE, G. (2017). Comparing inference approaches for RD designs: A reexamination of the effect of head start on child mortality. *J. Policy Anal. Manage.* **36** 643–681.

[9] HAHN, J., TODD, P. and VAN DER KLAAUW, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* **69** 201–209.

[10] SEKHON, J. S. and TITIUNIK, R. (2016). Understanding regression discontinuity designs as observational studies. *Obs. Stud.* **2** 174–182.

[11] SMALL, D. S. (2024). Protocols for observational studies: Methods and open problems. *Statist. Sci.*

[12] TITIUNIK, R. (2021). Natural experiments. In *Advances in Experimental Political Science* (J. N. Druckman and D. P. Green, eds.) **6** 103–129. Cambridge Univ. Press, Cambridge.

Matias D. Cattaneo is Professor, Department of Operations Research and Financial Engineering, Princeton University, Princeton, New Jersey 08540, USA (e-mail: cattaneo@princeton.edu). Rocío Titiunik is Professor, Department of Politics, Princeton University, Princeton, New Jersey 08544, USA (e-mail: titiunik@princeton.edu).

Statistical Science
2024, Vol. 39, No. 4, 566–567
https://doi.org/10.1214/24-STS956

Rejoinder: Protocols for Observational Studies: Methods and Open Problems

Dylan S. Small

REFERENCES

[1] AIKENS, R. C., GREAVES, D. and BAIOCCHI, M. (2020). A pilot design for observational studies: Using abundant data thoughtfully. *Stat. Med.* **39** 4821–4840. MR4190948 https://doi.org/10.1002/sim.8754

[2] CALONICO, S., CATTANEO, M. D. and FARRELL, M. H. (2020). Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs. *Econom. J.* **23** 192–210. MR4108924 https://doi.org/10.1093/ectj/utz022

[3] CALONICO, S., CATTANEO, M. D., FARRELL, M. H. and TITIUNIK, R. (2019). Regression discontinuity designs using covariates. *Rev. Econ. Stat.* **101** 442–451.

[4] CALONICO, S., CATTANEO, M. D. and TITIUNIK, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica* **82** 2295–2326. MR3301169 https://doi.org/10.3982/ECTA11757

[5] CATTANEO, M. D., IDROBO, N. and TITIUNIK, R. (2019). *A Practical Introduction to Regression Discontinuity Designs: Foundations*. Cambridge Univ. Press, Cambridge.

[6] CATTANEO, M. D., IDROBO, N. and TITIUNIK, R. (2024). *A Practical Introduction to Regression Discontinuity Designs: Extensions*. Cambridge Univ. Press, Cambridge.

[7] CATTANEO, M. D., KEELE, L., TITIUNIK, R. and VAZQUEZ-BARE, G. (2021). Extrapolating treatment effects in multi-cutoff regression discontinuity designs. *J. Amer. Statist. Assoc.* **116** 1941–1952. MR4353724 https://doi.org/10.1080/01621459.2020.1751646

[8] CATTANEO, M. D. and TITIUNIK, R. (2022). Regression discontinuity designs. *Ann. Rev. Econ.* **14** 821–851.

[9] LYCURGUS, T., MANN, C. and HANSEN, B. (2021). Protocol-evaluating the effect of ACA medicaid expansion on 2015–2018 mortality through matching and weighting. *Obs. Stud.* **7**.

[10] MANN, C. Z., HANSEN, B. B., GAYDOSH, L. and LYCURGUS, T. (2021). Protocol—evaluating the effect of ACA medicaid expansion on mortality during the COVID-19 pandemic using county-level matching. *Obs. Stud.* **7** S1–S31.

[11] WANG, S. V., POTTEGÅRD, A., CROWN, W., ARLETT, P., ASHCROFT, D. M., BENCHIMOL, E. I., BERGER, M. L., CRANE, G., GOETTSCH, W. et al. (2022). HARmonized protocol template to enhance reproducibility of hypothesis evaluating real-world evidence studies on treatment effects: A good practices report of a joint ISPE/ISPOR task force. *Value Health* **25** 1663–1672.

Dylan Small is Universal Furniture Professor, Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA (e-mail: dsmall@wharton.upenn.edu).

Statistical Science
2024, Vol. 39, No. 4, 568–588
https://doi.org/10.1214/24-STS925

Generally Altered, Inflated, Truncated and Deflated Regression

Thomas W. Yee and Chenchen Ma

Abstract. Models such as the zero-inflated and zero-altered Poisson and zero-truncated binomial are now well established. We review important ideas behind the three operators involved (as well as deflation and modification) and other incremental extensions. We propose a super mixture model that unifies alteration, inflation, truncation and deflation for counts, given a 1- or 2-parameter parent or base distribution. Since all of the operators except for truncation have both parametric and nonparametric variants, it is necessary to consider a total of seven different scenarios. Highlights of this paper include the following: (i) the mixture distribution is exceeding flexible, accommodating up to seven modes; (ii) under-, equi- and over-dispersion can be handled using a negative binomial (NB) parent, with underdispersion handled by a novel Generally-Truncated-Expansion method; (iii) under- and over-dispersion are studied holistically in terms of the the four operators previously referred to; (iv) while generally-altered regression explains why observations are there, generally-inflated regression accounts for why they are there *in excess*, and generally-deflated regression explains why observations are *not* there. The important application of heaped and seeped data from retrospective self-reported surveys is briefly mentioned. The GAITD-NB has potential to become a Swiss army knife for analyzing count responses.

Key words and phrases: Finite mixture distribution, multinomial logit model, overdispersion and underdispersion, vector generalized linear model.

REFERENCES

[1] AGRESTI, A. (2013). *Categorical Data Analysis*, 3rd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR3087436

[2] AGRESTI, A. (2015). *Foundations of Linear and Generalized Linear Models*. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR3308143

[3] ALANKO, T. (1984). Overview of techniques and problems in the measurement of alcohol consumption. *Res*. *Adv*. *Alcohol Drug Probl*. **8** 209–226.

[4] ANGERS, J.-F. and BISWAS, A. (2003). A Bayesian analysis of zero-inflated generalized Poisson model. *Comput*. *Statist*. *Data Anal*. **42** 37–46. MR1963008 https://doi.org/10.1016/S0167-9473(02)00154-8

[5] BERĘSEWICZ, M. and PAWLUKIEWICZ, K. (2020). Estimation of the number of irregular foreigners in Poland using non-linear count regression models. Available at arXiv:2008.09407.

[6] BERGER, M. and TUTZ, G. (2021). Transition models for count data: A flexible alternative to fixed distribution models. *Stat*. *Methods Appl*. **30** 1259–1283. MR4324411 https://doi.org/10.1007/s10260-021-00558-6

[7] BILDER, C. M. and LOUGHIN, T. M. (2015). *Analysis of Categorical Data with R*. CRC Press, Boca Raton, FL.

[8] BLASCO-MORENO, A., PÉREZ-CASANY, M., PUIG, P., MORANTE, M. and CASTELLS, E. (2019). What does a zero mean? Understanding false, random and structural zeros in ecology. *Methods Ecol*. *Evol*. **10** 949–959.

[9] BOCCI, C., GRASSINI, L. and ROCCO, E. (2021). A multiple inflated negative binomial hurdle regression model: Analysis of the Italians' tourism behaviour during the Great Recession. *Stat*. *Methods Appl*. **30** 1109–1133. MR4324404 https://doi.org/10.1007/s10260-020-00542-6

[10] BÖHNING, D. (2023). On the equivalence of one-inflated zero-truncated and zero-truncated one-inflated count data likelihoods. *Biom*. *J*. **65** Paper No. 2100343, 10. MR4555959

[11] BÖHNING, D. and OGDEN, H. E. (2021). General flation models for count data. *Metrika* **84** 245–261. MR4207513 https://doi.org/10.1007/s00184-020-00786-y

[12] BÜRKNER, P. C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *J*. *Stat*. *Softw*. **80** 1–28.

[13] BÜRKNER, P. C. (2021). Bayesian item response modeling in R with brms and Stan. *J*. *Stat*. *Softw*. **100** 1–54.

[14] CAI, T., XIA, Y. and ZHOU, Y. (2021). Generalized inflated discrete models: A strategy to work with multimodal discrete

Thomas Yee is a Senior Lecturer, Department of Statistics, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand (e-mail: t.yee@auckland.ac.nz). Chenchen Ma is a PhD student at the School of Mathematical Sciences and Center for Statistical Science, Peking University, China (e-mail: ChenchenMa@pku.edu.cn).

distributions. *Sociol. Methods Res.* **50** 365–400. MR4198562 https://doi.org/10.1177/0049124118782535

[15] CAMERON, A. C. and TRIVEDI, P. K. (2013). *Regression Analysis of Count Data*, 2nd ed. *Econometric Society Monographs* **53**. Cambridge Univ. Press, Cambridge. MR3155491 https://doi.org/10.1017/CBO9781139013567

[16] CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A. and CRAINICEANU, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd ed. *Monographs on Statistics and Applied Probability* **105**. CRC Press, Boca Raton, FL. MR2243417 https://doi.org/10.1201/9781420010138

[17] CHEN, J. and LI, H. (2013). Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Ann. Appl. Stat.* **7** 418–442. MR3086425 https://doi.org/10.1214/12-AOAS592

[18] CONSUL, P. C. and JAIN, G. C. (1973). A generalization of the Poisson distribution. *Technometrics* **15** 791–799. MR0391332 https://doi.org/10.1080/00401706.1973.10489112

[19] COX, N. J. (2004). Speaking stata: Graphing distributions. *Stata J.* **4** 66–88.

[20] CRAGG, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica* **39** 829–844. MR0573314 https://doi.org/10.2307/1913124

[21] DIETZ, E. and BÖHNING, D. (2000). On estimation of the Poisson parameter in zero-modified Poisson models. *Comput. Statist. Data Anal.* **34** 441–459.

[22] FILZMOSER, P., HRON, K. and TEMPL, M. (2018). Applied compositional data analysis: With worked examples in R. Springer, Nature, Switzerland.

[23] GHOSH, S. K., MUKHOPADHYAY, P. and LU, J.-C. (2006). Bayesian analysis of zero-inflated regression models. *J. Statist. Plann. Inference* **136** 1360–1375. MR2253768 https://doi.org/10.1016/j.jspi.2004.10.008

[24] GILES, D. E. (2007). Modeling inflated count data. In *MODSIM* 2007 *International Congress on Modelling and Simulation* (L. Oxley and D. Kulasiri, eds.) 919–925. Modelling and Simulation Society of Australia and New Zealand, New Zealand.

[25] GMEL, G. and REHM, J. (2004). Measuring alcohol consumption. *Contemp. Drug Probl.* **31** 467–540.

[26] GODWIN, R. T. (2017). One-inflation and unobserved heterogeneity in population size estimation. *Biom. J.* **59** 79–93. MR3593722 https://doi.org/10.1002/bimj.201600063

[27] GODWIN, R. T. and BÖHNING, D. (2017). Estimation of the population size by using the one-inflated positive Poisson model. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **66** 425–448. MR3611695 https://doi.org/10.1111/rssc.12192

[28] GREEN, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *J. Roy. Statist. Soc. Ser. B* **46** 149–192. With discussion. MR0781879

[29] GREENE, W. H. (2012). *Econometric Analysis*, 7th ed. Prentice Hall, Upper Saddle River, NJ.

[30] GUPTA, P. L., GUPTA, R. C. and TRIPATHI, R. C. (1996). Analysis of zero-adjusted count data. *Comput. Statist. Data Anal.* **23** 207–218.

[31] GURMU, S. (1991). Tests for detecting overdispersion in the positive Poisson model. *J. Bus. Econom. Statist.* **9** 215–222.

[32] HALL, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: A case study. *Biometrics* **56** 1030–1039. MR1815581 https://doi.org/10.1111/j.0006-341X.2000.01030.x

[33] HASSANZADEH, F. and KAZEMI, I. (2017). Regression modeling of one-inflated positive count data. *Statist. Papers* **58** 791–809. MR3686851 https://doi.org/10.1007/s00362-015-0726-7

[34] HAUCK, W. W. JR. and DONNER, A. (1977). Wald's test as applied to hypotheses in logit analysis. *J. Amer. Statist. Assoc.* **72** 851–853. MR0471146

[35] HE, H., WANG, W., CRITS-CHRISTOPH, P., GALLOP, R., TANG, W., CHEN, D. D. and TU, X. M. (2014). On the implication of structural zeros as independent variables in regression analysis: Applications to alcohol research. *J. Data Sci.* **12** 439–460.

[36] HEILBRON, D. C. (1994). Zero-altered and other regression for count data with added zeros. *Biom. J.* **36** 531–547.

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

[38] HILBE, J. M. (2016). COUNT: Functions, data and code for count data R package version 1.3.4.

[39] HIRSHKOWITZ, M., WHITON, K. et al. (2015). National sleep foundation's sleep time duration recommendations: Methodology and results summary. *Sleep Health* **1** 40–43.

[40] HORTON, N. J., KIM, E. and SAITZ, R. (2007). A cautionary note regarding count models of alcohol consumption in randomized controlled trials. *BMC Med. Res. Methodol.* **7** 1–9.

[41] HUANG, A. (2017). Mean-parametrized Conway–Maxwell–Poisson regression models for dispersed counts. *Stat. Model.* **17** 359–380. MR3720937 https://doi.org/10.1177/1471082X17697749

[42] HUANG, A. (2023). On arbitrarily underdispersed discrete distributions. *Amer. Statist.* **77** 29–34. MR4541812 https://doi.org/10.1080/00031305.2022.2106305

[43] HUGGINS, R. M. (1989). On the statistical analysis of capture experiments. *Biometrika* **76** 133–140. MR0991431 https://doi.org/10.1093/biomet/76.1.133

[44] KAUL, A., MANDAL, S., DAVIDOV, O. and PEDDADA, S. D. (2017). Analysis of microbiome data in the presence of excess zeros. *Front. Microbiol.* **8** 1–10.

[45] KLEIBER, C. and ZEILEIS, A. (2008). *Applied Econometrics with R*. Springer, New York.

[46] KUHNERT, P. M., MARTIN, T. G., MENGERSEN, K. and POSSINGHAM, H. P. (2005). Assessing the impacts of grazing levels on bird density in woodland habitat: A Bayesian approach using expert opinion. *Environmetrics* **16** 717–747. MR2196281 https://doi.org/10.1002/env.732

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

[48] LI, H. (2015). Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annu. Rev. Stat. Appl.* **2** 73–94.

[49] LIU, F. and EUGENIO, E. C. (2018). A review and comparison of Bayesian and likelihood-based inferences in beta regression and zero-or-one-inflated beta regression. *Stat. Methods Med. Res.* **27** 1024–1044. MR3770133 https://doi.org/10.1177/0962280216650699

[50] LIU, L., SHIH, Y.-C. T., STRAWDERMAN, R. L., ZHANG, D., JOHNSON, B. A. and CHAI, H. (2019). Statistical analysis of zero-inflated nonnegative continuous data: A review. *Statist. Sci.* **34** 253–279. MR3983328 https://doi.org/10.1214/18-STS681

[51] MACKENZIE, D. I., NICHOLS, J. D., LACHMAN, G. B., DROEGE, S., ROYLE, J. and LANGTIMM, C. A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology* **83** 2248–55.

[52] MACMAHON, S., NORTON, R., JACKSON, R., MACKIE, M. J., CHENG, A., HOORN, S. V., MILNE, A. and MCCULLOCH, A. (1995). Fletcher Challenge—Univ. Auckland Heart & Health Study: Design and baseline findings. *N. Z. Med. J.* **108** 499–502.

[53] Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., Tyre, A. J. and Possingham, H. P. (2005). Zero tolerance ecology: Improving ecological inference by modelling the source of zero observations. *Ecol. Lett.* **8** 1235–1246.

[54] McCullagh, P. (1980). Regression models for ordinal data. *J. Roy. Statist. Soc. Ser. B* **42** 109–142. MR0583347

[55] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. *Monographs on Statistics and Applied Probability*. CRC Press, London. MR3223057 https://doi.org/10.1007/978-1-4899-3242-6

[56] Miranda-Soberanis, V. F. and Yee, T. W. (2023). Two-parameter link functions, with applications to negative binomial, Weibull and quantile regression. *Comput. Statist.* **38** 1463–1485. MR4617303

[57] Moran, P. A. P. (1971). Maximum-likelihood estimation in non-standard conditions. *Math. Proc. Cambridge Philos. Soc.* **70** 441–450. MR0290493 https://doi.org/10.1017/s0305004100050088

[58] Moriña, D., Puig, P. and Navarro, A. (2021). Analysis of zero inflated dichotomous variables from a Bayesian perspective: Application to occupational health. *BMC Med. Res. Methodol.* **21** 1–9.

[59] Mullahy, J. (1986). Specification and testing of some modified count data models. *J. Econometrics* **33** 341–365. MR0867980 https://doi.org/10.1016/0304-4076(86)90002-3

[60] Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *J. Roy. Statist. Soc. Ser. A* **135** 370–384. MR0375592 https://doi.org/10.1093/biomet/61.3.439

[61] Olshen, R. A., Biden, E. N., Wyatt, M. P. and Sutherland, D. H. (1989). Gait analysis and the bootstrap. *Ann. Statist.* **17** 1419–1440. MR1026290 https://doi.org/10.1214/aos/1176347372

[62] Ospina, R. and Ferrari, S. L. P. (2012). A general class of zero-or-one inflated beta regression models. *Comput. Statist. Data Anal.* **56** 1609–1623. MR2892364 https://doi.org/10.1016/j.csda.2011.10.005

[63] Otis, D. L., Burnham, K. P., White, G. C. and Anderson, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildl. Monogr.* **62** 3–135.

[64] Paul, S. R., Balasooriya, U. and Banerjee, T. (2005). Fisher information matrix of the Dirichlet-multinomial distribution. *Biom. J.* **47** 230–236. MR2137237 https://doi.org/10.1002/bimj.200410103

[65] Ridout, M., Hinde, J. and Demétrio, C. G. B. (2001). A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics* **57** 219–223. MR1833310 https://doi.org/10.1111/j.0006-341X.2001.00219.x

[66] Schwartz, J. and Giles, D. E. (2016). Bias-reduced maximum likelihood estimation of the zero-inflated Poisson distribution. *Comm. Statist. Theory Methods* **45** 465–478. MR3447927 https://doi.org/10.1080/03610926.2013.824590

[67] Seber, G. A. F. and Schofield, M. R. (2019). *Capture–Recapture: Parameter Estimation for Open Animal Populations*. *Statistics for Biology and Health*. Springer, Cham. MR3967717 https://doi.org/10.1007/978-3-030-18187-1

[68] Self, S. G. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Amer. Statist. Assoc.* **82** 605–610. MR0898365

[69] Sellers, K. F. (2023). *The Conway–Maxwell–Poisson Distribution*. *Institute of Mathematical Statistics* (*IMS*) *Monographs* **8**. Cambridge Univ. Press, Cambridge. MR4692991 https://doi.org/10.1017/9781108646437

[70] Sellers, K. F., Borle, S. and Shmueli, G. (2012). The COM-Poisson model for count data: A survey of methods and applications. *Appl. Stoch. Models Bus. Ind.* **28** 104–116. MR2911711 https://doi.org/10.1002/asmb.918

[71] Sellers, K. F. and Morris, D. S. (2017). Underdispersion models: Models that are "under the radar". *Comm. Statist. Theory Methods* **46** 12075–12086. MR3740852 https://doi.org/10.1080/03610926.2017.1291976

[72] Sellers, K. F. and Raim, A. (2016). A flexible zero-inflated model to address data dispersion. *Comput. Statist. Data Anal.* **99** 68–80. MR3473082 https://doi.org/10.1016/j.csda.2016.01.007

[73] 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 https://doi.org/10.1002/sim.6263

[74] Smithson, M. and Merkle, E. C. (2014). *Generalized Linear Models for Categorical and Continuous Limited Dependent Variables*. CRC Press, London.

[75] Stan Development Team (2022). RStan: The R interface to Stan. R package version 2.21.7.

[76] Su, X., Fan, J., Levine, R. A., Tan, X. and Tripathi, A. (2013). Multiple-inflation Poisson model with L_1 regularization. *Statist. Sinica* **23** 1071–1090. MR3114705

[77] Tang, W., Lu, N., Chen, T., Wang, W., Gunzler, D. D., Han, Y. and Tu, X. M. (2015). On performance of parametric and distribution-free models for zero-inflated and over-dispersed count responses. *Stat. Med.* **34** 3235–3245. MR3412629 https://doi.org/10.1002/sim.6560

[78] Tian, G.-L., Ding, X., Liu, Y. and Tang, M.-L. (2019). Some new statistical methods for a class of zero-truncated discrete distributions with applications. *Comput. Statist.* **34** 1393–1426. MR3982868 https://doi.org/10.1007/s00180-018-00860-0

[79] Tian, G.-L., Ma, H., Zhou, Y. and Deng, D. (2015). Generalized endpoint-inflated binomial model. *Comput. Statist. Data Anal.* **89** 97–114. MR3349670 https://doi.org/10.1016/j.csda.2015.03.009

[80] Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica* **26** 24–36. MR0090462 https://doi.org/10.2307/1907382

[81] Tutz, G. (2012). *Regression for Categorical Data*. *Cambridge Series in Statistical and Probabilistic Mathematics* **34**. Cambridge Univ. Press, Cambridge. MR2856629

[82] Umbach, D. (1981). On inference for a mixture of a Poisson and a degenerate distribution. *Comm. Statist. Theory Methods* **10** 299–306. MR0612561 https://doi.org/10.1080/03610928108828039

[83] Vieira, A. M. C., Hinde, J. P. and Demétrio, C. G. B. (2000). Zero-inflated proportion data models applied to a biological control assay. *J. Appl. Stat.* **27** 373–389.

[84] Vuong, Q. H. (1989). Likelihood ratio tests for model selection and nonnested hypotheses. *Econometrica* **57** 307–333. MR0996939 https://doi.org/10.2307/1912557

[85] Wang, H. and Heitjan, D. F. (2008). Modeling heaping in self-reported cigarette counts. *Stat. Med.* **27** 3789–3804. MR2526609 https://doi.org/10.1002/sim.3281

[86] Welsh, A. H., Cunningham, R. B., Donnelly, C. F. and Lindenmayer, D. B. (1996). Modelling the abundances of rare species: Statistical models for counts with extra zeros. *Ecol. Model.* **88** 297–308.

[87] Wilson, P. (2015). The misuse of the Vuong test for nonnested models to test for zero-inflation. *Econom. Lett.* **127** 51–53. MR3305593 https://doi.org/10.1016/j.econlet.2014.12.029

[88] WOLFF, J. and AUGUSTIN, T. (2003). Heaping and its consequences for duration analysis: A simulation study. *Allg. Stat. Arch.* **87** 59–86. MR1963225

[89] WOOD, S. N. (2017). *Generalized Additive Models: An Introduction with R*, 2nd ed. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. MR3726911

[90] XIE, F.-C., LIN, J.-G. and WEI, B.-C. (2014). Bayesian zero-inflated generalized Poisson regression model: Estimation and case influence diagnostics. *J. Appl. Stat.* **41** 1383–1392. MR3268901 https://doi.org/10.1080/02664763.2013.871508

[91] YANG, Z., HARDIN, J. W. and ADDY, C. L. (2009). Testing overdispersion in the zero-inflated Poisson model. *J. Statist. Plann. Inference* **139** 3340–3353. MR2538085 https://doi.org/10.1016/j.jspi.2009.03.016

[92] YEE, T. W. (2010). The VGAM package for categorical data analysis. *J. Stat. Softw.* **32** 1–34.

[93] YEE, T. W. (2015). *Vector Generalized Linear and Additive Models with an Implementation in R. Springer Series in Statistics*. Springer, New York. MR3408425 https://doi.org/10.1007/978-1-4939-2818-7

[94] YEE, T. W. (2022). On the Hauck–Donner effect in Wald tests: Detection, tipping points, and parameter space characterization. *J. Amer. Statist. Assoc.* **117** 1763–1774. MR4528469 https://doi.org/10.1080/01621459.2021.1886936

[95] YEE, T. W. (2023). VGAM: Vector generalized linear and additive models R package version 1.1-9.

[96] YEE, T. W. and GRAY, J. (2023). VGAMdata: Data supporting the 'VGAM' package R package version 1.1-8.

[97] YEE, T. W. and HASTIE, T. J. (2003). Reduced-rank vector generalized linear models. *Stat. Model.* **3** 15–41. MR1977163 https://doi.org/10.1191/1471082X03st045oa

[98] YEE, T. W. and MA, C. (2024). Supplement to "Generally altered, inflated, truncated and deflated regression." https://doi.org/10.1214/24-STS925SUPPA, https://doi.org/10.1214/24-STS925SUPPB, https://doi.org/10.1214/24-STS925SUPPC, https://doi.org/10.1214/24-STS925SUPPD

[99] YEE, T. W., FRIGAU, L. and MA, C. (2024). Heaping and seeping, GAITD regression and doubly constrained reduced rank vector generalized linear models, in smoking studies. Unpublished manuscript.

[100] YEE, T. W., STOKLOSA, J. and HUGGINS, R. M. (2015). The VGAM package for capture–recapture data using the conditional likelihood. *J. Stat. Softw.* **65** 1–33.

[101] ZHANG, X., PEI, Y. F., ZHANG, L., GUO, B., PENDEGRAFT, A. H., ZHUANG, W. and YI, N. (2018). Negative binomial mixed models for analyzing longitudinal microbiome data. *Front. Microbiol.* **9** 1–10. MR3844772

Statistical Science
2024, Vol. 39, No. 4, 589–600
https://doi.org/10.1214/24-STS935

A Bayesian "Sandwich" for Variance Estimation

Kendrick Li and Kenneth Rice

Abstract. Large-sample Bayesian analogs exist for many frequentist methods, but are less well known for the widely-used "sandwich" or "robust" variance estimators. We review existing approaches to Bayesian analogs of sandwich variance estimators and propose a new analog, as the Bayes rule under a form of balanced loss function, that combines elements of standard parametric inference with fidelity of the data to the model. Our development is general, for essentially any regression setting with independent outcomes. Being the large-sample equivalent of its frequentist counterpart, we show by simulation that Bayesian robust standard error estimates can faithfully quantify the variability of parameter estimators even under model misspecification—thus retaining the major attraction of the original frequentist version. We demonstrate our Bayesian analog of standard error estimates when studying the association between age and systolic blood pressure in NHANES.

Key words and phrases: Bayesian statistics, robust standard errors, parametric inference.

REFERENCES

ABADIE, A., IMBENS, G. W. and ZHENG, F. (2014). Inference for misspecified models with fixed regressors. *J. Amer. Statist. Assoc.* **109** 1601–1614. MR3293613 https://doi.org/10.1080/01621459.2014.928218

AGRESTI, A. (2013). *Categorical Data Analysis*, 3rd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR3087436

BARRON, J. T. (2019). A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 4331–4339.

BERK, R. H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *Ann. Math. Stat.* **37** 51–58; correction, ibid. 745–746. MR0189176 https://doi.org/10.1214/aoms/1177699477

BERK, R. H. (1970). Consistency a posteriori. *Ann. Math. Stat.* **41** 894–906. MR0266356 https://doi.org/10.1214/aoms/1177696967

BUJA, A., BROWN, L., BERK, R., GEORGE, E., PITKIN, E., TRASKIN, M., ZHANG, K. and ZHAO, L. (2019a). Models as approximations I: Consequences illustrated with linear regression. *Statist. Sci.* **34** 523–544. MR4048582 https://doi.org/10.1214/18-STS693

BUJA, A., BROWN, L., KUCHIBHOTLA, A. K., BERK, R., GEORGE, E. and ZHAO, L. (2019b). Models as approximations II: A model-free theory of parametric regression. *Statist. Sci.* **34** 545–565. MR4048583 https://doi.org/10.1214/18-STS694

BUNKE, O. and MILHAUD, X. (1998). Asymptotic behavior of Bayes estimates under possibly incorrect models. *Ann. Statist.* **26** 617–644. MR1626075 https://doi.org/10.1214/aos/1028144851

CHATURVEDI, A. and SHALABH (2014). Bayesian estimation of regression coefficients under extended balanced loss function. *Comm. Statist. Theory Methods* **43** 4253–4264. MR3269272 https://doi.org/10.1080/03610926.2012.725498

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

DAWID, A. P. and SEBASTIANI, P. (1999). Coherent dispersion criteria for optimal experimental design. *Ann. Statist.* **27** 65–81. MR1701101 https://doi.org/10.1214/aos/1018031101

DEY, D. K., GHOSH, M. and STRAWDERMAN, W. E. (1999). On estimation with balanced loss functions. *Statist. Probab. Lett.* **45** 97–101. MR1718467 https://doi.org/10.1016/S0167-7152(99)00047-4

HJORT, N. L., HOLMES, C., MÜLLER, P. and WALKER, S. G. (2010). *Bayesian Nonparametrics* **28**. Cambridge Univ. Press.

HOFF, P. and WAKEFIELD, J. (2013). Bayesian sandwich posteriors for pseudo-true parameters. A discussion of "Bayesian inference with misspecified models" by Stephen Walker [MR3082220]. *J. Statist. Plann. Inference* **143** 1638–1642. MR3082222 https://doi.org/10.1016/j.jspi.2013.05.014

HOLLAND, M. J. and IKEDA, K. (2016). Minimum proper loss estimators for parametric models. *IEEE Trans. Signal Process.* **64** 704–713. MR3453346 https://doi.org/10.1109/TSP.2015.2489608

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability* (*Berkeley, Calif.*, 1965/66), *Vol. I*: *Statistics* 221–233. Univ. California Press, Berkeley, CA. MR0216620

HUBER, P. J. (1981). *Robust Statistics*. *Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. MR0606374

JAFARI JOZANI, M., MARCHAND, É. and PARSIAN, A. (2012). Bayesian and robust Bayesian analysis under a general class of

Kendrick Li is Assistant Member, Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, Tennessee 38105, USA (e-mail: kendrick.li@stjude.org). Kenneth Rice is Professor, Department of Biostatistics, University of Washington, Seattle, Washington 98195, USA (e-mail: kenrice@uw.edu).

balanced loss functions. *Statist. Papers* **53** 51–60. MR2878590 https://doi.org/10.1007/s00362-010-0307-8

KIM, J.-Y. (2002). Limited information likelihood and Bayesian analysis. *J. Econometrics* **107** 175–193. Information and entropy econometrics. MR1889958 https://doi.org/10.1016/S0304-4076(01)00119-1

KLEIJN, B. J. K., VAN DER VAART, A. W. et al. (2012). The Bernstein–von-Mises theorem under misspecification. *Electron. J. Stat.* **6** 354–381.

LANCASTER, T. (2003). A note on bootstraps and robustness. Brown Univ. Dept. Economics Working Paper 2006-06.

LI, K. and RICE, K. (2024). Supplement to "A Bayesian 'Sandwich' for Variance Estimation." https://doi.org/10.1214/24-STS935SUPP

MCCULLAGH, P. (1983). Quasilikelihood functions. *Ann. Statist.* **11** 59–67. MR0684863

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models. Monographs on Statistics and Applied Probability*. CRC Press, London. Second edition [of MR0727836]. MR3223057 https://doi.org/10.1007/978-1-4899-3242-6

MILLER, J. W. and DUNSON, D. B. (2019). Robust Bayesian inference via coarsening. *J. Amer. Statist. Assoc.* **114** 1113–1125. MR4011766 https://doi.org/10.1080/01621459.2018.1469995

MÜLLER, U. K. (2013). Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica* **81** 1805–1849. MR3117034 https://doi.org/10.3982/ECTA9097

NOTT, D. J., DROVANDI, C. and FRAZIER, D. T. (2024). Bayesian inference for misspecified generative models. *Annu. Rev. Stat. Appl.* **11** 179–202. MR4742281

O'HAGAN, A. (2013). Bayesian inference with misspecified models: Inference about what?. *J. Statist. Plann. Inference* **143** 1643–1648. MR3082223 https://doi.org/10.1016/j.jspi.2013.05.016

PARMIGIANI, G. and INOUE, L. Y. T. (2009). *Decision Theory: Principles and Approaches. Wiley Series in Probability and Statistics*. Wiley, Chichester. With contributions by Hedibert F. Lopes. MR2604978 https://doi.org/10.1002/9780470746684

PLUMMER, M. et al. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* **124** 1–10. Vienna, Austria.

ROYALL, R. M. (1986). Model robust confidence intervals using maximum likelihood estimators. *Int. Stat. Rev.* **54** 221–226. MR0962937 https://doi.org/10.2307/1403146

RUBIN, D. B. (1981). The Bayesian bootstrap. *Ann. Statist.* **9** 130–134. MR0600538

SAVAGE, L. J. (1951). The theory of statistical decision. *J. Amer. Statist. Assoc.* **46** 55–67.

STARTZ, R. (2012). Bayesian heteroskedasticity-robust standard errors. Technical Report, Dept. Economics, UC Santa Barbara.

SZPIRO, A. A., RICE, K. M. and LUMLEY, T. (2010). Model-robust regression and a Bayesian "sandwich" estimator. *Ann. Appl. Stat.* **4** 2099–2113. MR2829948 https://doi.org/10.1214/10-AOAS362

VAN DER VAART, A. W. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

WALD, A. (1949). Statistical decision functions. *Ann. Math. Stat.* **20** 165–205. MR0044802 https://doi.org/10.1214/aoms/1177730030

WALKER, S. G. (2013). Bayesian inference with misspecified models. *J. Statist. Plann. Inference* **143** 1621–1633. MR3082220 https://doi.org/10.1016/j.jspi.2013.05.013

YU, Z. (2021). *Few-Shot Learning: Contributions to Deep Learning with Limited Data*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.), Univ. Wisconsin–Madison. MR4690732

ZELLNER, A. (1994). Bayesian and non-Bayesian estimation using balanced loss functions. In *Statistical Decision Theory and Related Topics, V (West Lafayette, IN*, 1992) 377–390. Springer, New York. MR1286314

Statistical Science
2024, Vol. 39, No. 4, 601–622
https://doi.org/10.1214/24-STS936

Scalable Empirical Bayes Inference and Bayesian Sensitivity Analysis

Hani Doss and Antonio Linero

Abstract. Consider a Bayesian setup in which we observe Y, whose distribution depends on a parameter θ, that is, $Y \mid \theta \sim \pi_{Y \mid \theta}$. The parameter θ is unknown and treated as random, and a prior distribution chosen from some parametric family $\{\pi_\theta(\cdot; h), h \in \mathcal{H}\}$, is to be placed on it. For the subjective Bayesian there is a single prior in the family which represents his or her beliefs about θ, but determination of this prior is very often extremely difficult. In the empirical Bayes approach, the latent distribution on θ is estimated from the data. This is usually done by choosing the value of the hyperparameter h that maximizes some criterion. Arguably the most common way of doing this is to let $m(h)$ be the marginal likelihood of h, that is, $m(h) = \int \pi_{Y \mid \theta} \nu_h(\theta)\, d\theta$, and choose the value of h that maximizes $m(\cdot)$. Unfortunately, except for a handful of textbook examples, analytic evaluation of $\arg\max_h m(h)$ is not feasible. The purpose of this paper is two-fold. First, we review the literature on estimating it and find that the most commonly used procedures are either potentially highly inaccurate or don't scale well with the dimension of h, the dimension of θ, or both. Second, we present a method for estimating $\arg\max_h m(h)$, based on Markov chain Monte Carlo, that applies very generally and scales well with dimension. Let g be a real-valued function of θ, and let $I(h)$ be the posterior expectation of $g(\theta)$ when the prior is ν_h. As a byproduct of our approach, we show how to obtain point estimates and globally-valid confidence bands for the family $I(h), h \in \mathcal{H}$. To illustrate the scope of our methodology we provide three detailed examples, having different characters.

Key words and phrases: Bayesian model selection, Donsker class, geometric ergodicity, hyperparameter selection, Markov chain Monte Carlo, regenerative simulation.

REFERENCES

ATCHADÉ, Y. F. (2011). A computational framework for empirical Bayes inference. *Stat. Comput.* **21** 463–473. MR2826685 https://doi.org/10.1007/s11222-010-9182-3

ATHREYA, K. B. and NEY, P. (1978). A new approach to the limit theory of recurrent Markov chains. *Trans. Amer. Math. Soc.* **245** 493–501. MR0511425 https://doi.org/10.2307/1998882

BEAL, M. J. and GHAHRAMANI, Z. (2003). The variational Bayesian EM algorithm for incomplete data: With application to scoring graphical model structures. In *Bayesian Statistics*, 7 (*Tenerife*, 2002) M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, M. West et al., eds.) 453–463. Oxford Univ. Press, New York. MR2003189

BLEI, D. M., KUCUKELBIR, A. and MCAULIFFE, J. D. (2017). Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* **112** 859–877. MR3671776 https://doi.org/10.1080/01621459.2017.1285773

BREHENY, P. and HUANG, J. (2009). Penalized methods for bi-level variable selection. *Stat. Interface* **2** 369–380. MR2540094 https://doi.org/10.4310/SII.2009.v2.n3.a10

BYRD, R. H., LU, P., NOCEDAL, J. and ZHU, C. Y. (1995). A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.* **16** 1190–1208. MR1346301 https://doi.org/10.1137/0916069

CHEN, M.-H. (1994). Importance-weighted marginal Bayesian posterior density estimation. *J. Amer. Statist. Assoc.* **89** 818–824. MR1294727

CHIB, S. (1995). Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc.* **90** 1313–1321. MR1379473

CHIB, S. and JELIAZKOV, I. (2001). Marginal likelihood from the Metropolis–Hastings output. *J. Amer. Statist. Assoc.* **96** 270–281. MR1952737 https://doi.org/10.1198/016214501750332848

Hani Doss is Professor, Department of Statistics, University of Florida, Gainesville, Florida 32603, USA (e-mail: doss@stat.ufl.edu). Antonio Linero is Associate Professor, Department of Statistics and Data Science, University of Texas at Austin, Austin, Texas 78705, USA (e-mail: antonio.linero@austin.utexas.edu).

CHIPMAN, H., GEORGE, E. I., GRAMACY, R. B. and MCCULLOCH, R. (2013). Bayesian treed response surface models. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **3** 298–305.

CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4** 266–298. MR2758172 https://doi.org/10.1214/09-AOAS285

CLYDE, M., DESIMONE, H. and PARMIGIANI, G. (1996). Prediction via orthogonalized model mixing. *J. Amer. Statist. Assoc.* **91** 1197–1208.

DEVROYE, L. (1986). *Nonuniform Random Variate Generation*. Springer, New York. MR0836973 https://doi.org/10.1007/978-1-4613-8643-8

DONOHO, D. L. and LIU, R. C. (1991). Geometrizing rates of convergence. II. *Ann. Statist.* **19** 633–667. MR1105839 https://doi.org/10.1214/aos/1176348114

DOSS, H. and LINERO, A. (2024). Supplement to "Scalable empirical Bayes inference and Bayesian sensitivity analysis." https://doi.org/10.1214/24-STS936SUPP

DOSS, H. and PARK, Y. (2018). An MCMC approach to empirical Bayes inference and Bayesian sensitivity analysis via empirical processes. *Ann. Statist.* **46** 1630–1663. MR3819112 https://doi.org/10.1214/17-AOS1597

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

FLEGAL, J. M., HARAN, M. and JONES, G. L. (2008). Markov chain Monte Carlo: Can we trust the third significant figure? *Statist. Sci.* **23** 250–260. MR2516823 https://doi.org/10.1214/08-STS257

FLEGAL, J. M. and JONES, G. L. (2010). Batch means and spectral variance estimators in Markov chain Monte Carlo. *Ann. Statist.* **38** 1034–1070. MR2604704 https://doi.org/10.1214/09-AOS735

FORT, G. and MOULINES, E. (2003). Convergence of the Monte Carlo expectation maximization for curved exponential families. *Ann. Statist.* **31** 1220–1259. MR2001649 https://doi.org/10.1214/aos/1059655912

FRIEDMAN, J. H. (1991). Multivariate adaptive regression splines. *Ann. Statist.* **19** 1–67. MR1091842 https://doi.org/10.1214/aos/1176347963

GELMAN, A. and MENG, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statist. Sci.* **13** 163–185. MR1647507 https://doi.org/10.1214/ss/1028905934

GEORGE, C. P. and DOSS, H. (2018). Principled selection of hyperparameters in the latent Dirichlet allocation model. *J. Mach. Learn. Res.* **18** 1–38. MR3813811

HOFFMAN, M. D., BLEI, D. M., WANG, C. and PAISLEY, J. (2013). Stochastic variational inference. *J. Mach. Learn. Res.* **14** 1303–1347. MR3081926

HUANG, J., MA, S., XIE, H. and ZHANG, C.-H. (2009). A group bridge approach for variable selection. *Biometrika* **96** 339–355. MR2507147 https://doi.org/10.1093/biomet/asp020

IBRAGIMOV, I. A. and LINNIK, Y. V. (1971). *Independent and Stationary Sequences of Random Variables*. Wolters-Noordhoff Publishing, Groningen. MR0322926

JAMES, W. and STEIN, C. (1960). Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I* 361–379. Univ. California Press, Berkeley-Los Angeles, CA. MR0133191

JONES, G. L., HARAN, M., CAFFO, B. S. and NEATH, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *J. Amer. Statist. Assoc.* **101** 1537–1547. MR2279478 https://doi.org/10.1198/016214506000000492

KOSOROK, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. *Springer Series in Statistics*. Springer, New York. MR2724368 https://doi.org/10.1007/978-0-387-74978-5

LIANG, F., PAULO, R., MOLINA, G., CLYDE, M. A. and BERGER, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *J. Amer. Statist. Assoc.* **103** 410–423. MR2420243 https://doi.org/10.1198/016214507000001337

LINERO, A. R. (2017). A review of tree-based Bayesian methods. *Commun. Stat. Appl. Methods* **24** 543–559.

LINERO, A. R. (2018). Bayesian regression trees for high-dimensional prediction and variable selection. *J. Amer. Statist. Assoc.* **113** 626–636. MR3832214 https://doi.org/10.1080/01621459.2016.1264957

LINERO, A. R. and YANG, Y. (2018). Bayesian regression tree ensembles that adapt to smoothness and sparsity. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 1087–1110. MR3874311 https://doi.org/10.1111/rssb.12293

LIU, J. S. (1994). The fraction of missing information and convergence rate for data augmentation. In *Computationally Intensive Statistical Methods: Proceedings of the 26th Symposium Interface* (J. Sall and A. Lehman, eds.) 490–497. Interface Foundation, Fairfax Station, VA.

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

MENG, X.-L. and WONG, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statist. Sinica* **6** 831–860. MR1422406

MEYN, S. P. and TWEEDIE, R. L. (1993). *Markov Chains and Stochastic Stability*. *Communications and Control Engineering Series*. Springer London, Ltd., London. MR1287609 https://doi.org/10.1007/978-1-4471-3267-7

MILLER, J. W. (2019). Fast and accurate approximation of the full conditional for gamma shape parameters. *J. Comput. Graph. Statist.* **28** 476–480. MR3974896 https://doi.org/10.1080/10618600.2018.1537929

MYKLAND, P., TIERNEY, L. and YU, B. (1995). Regeneration in Markov chain samplers. *J. Amer. Statist. Assoc.* **90** 233–241. MR1325131

NEAL, R. M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo* A. Gelman, G. L. Jones and X. L. Meng, eds.). *Chapman & Hall/CRC Handb. Mod. Stat. Methods* 113–162. CRC Press, Boca Raton, FL. MR2858447

NEWTON, M. A. and RAFTERY, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *J. Roy. Statist. Soc. Ser. B* **56** 3–48. MR1257793

PETRONE, S., ROUSSEAU, J. and SCRICCIOLO, C. (2014). Bayes and empirical Bayes: Do they merge? *Biometrika* **101** 285–302. MR3215348 https://doi.org/10.1093/biomet/ast067

RAFTERY, A. E., MADIGAN, D. and HOETING, J. A. (1997). Bayesian model averaging for linear regression models. *J. Amer. Statist. Assoc.* **92** 179–191. MR1436107 https://doi.org/10.2307/2291462

ROY, V. and EVANGELOU, E. (2024). Selection of proposal distributions for multiple importance sampling. *Statist. Sinica* **34** 27–46. MR4683562

SCHEIPL, F., FAHRMEIR, L. and KNEIB, T. (2012). Spike-and-slab priors for function selection in structured additive regression models. *J. Amer. Statist. Assoc.* **107** 1518–1532. MR3036413 https://doi.org/10.1080/01621459.2012.737742

SCOTT, J. G. and BERGER, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* **38** 2587–2619. MR2722450 https://doi.org/10.1214/10-AOS792

SMITH, M. and KOHN, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics* **75** 317–343.

STARK, P. C., RYAN, L. M., MCDONALD, J. L. and BURGE, H. A. (1997). Using meteorologic data to model and predict daily ragweed pollen levels. *Aerobiologia* **13** 177–184.

STEIN, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135–1151. MR0630098

STORLIE, C. and HELTON, J. (2008). Multiple predictor smoothing methods for sensitivity analysis: Example results. *Reliab. Eng. Syst. Saf.* **93** 55–77.

STORLIE, C. B., BONDELL, H. D., REICH, B. J. and ZHANG, H. H. (2011). Surface estimation, variable selection, and the nonparametric oracle property. *Statist. Sinica* **21** 679–705. MR2829851 https://doi.org/10.5705/ss.2011.030a

TSYBAKOV, A. B. (1990). Recurrent estimation of the mode of a multidimensional distribution. *Problemy Peredachi Informatsii* **26** 38–45. MR1051586

VAN DYK, D. A. and MENG, X.-L. (2001). The art of data augmentation. *J. Comput. Graph. Statist.* **10** 1–50. MR1936358 https://doi.org/10.1198/10618600152418584

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

WOLPERT, R. L. and SCHMIDLER, S. C. (2012). α-stable limit laws for harmonic mean estimators of marginal likelihoods. *Statist. Sinica* **22** 1233–1251. MR2987490 https://doi.org/10.5705/ss.2010.221

YANG, Y., WAINWRIGHT, M. J. and JORDAN, M. I. (2016). On the computational complexity of high-dimensional Bayesian variable selection. *Ann. Statist.* **44** 2497–2532. MR3576552 https://doi.org/10.1214/15-AOS1417

Statistical Science
2024, Vol. 39, No. 4, 623–636
https://doi.org/10.1214/24-STS937

Feature Importance: A Closer Look at Shapley Values and LOCO

Isabella Verdinelli and Larry Wasserman

Abstract. There is much interest lately in explainability in statistics and machine learning. One aspect of explainability is to quantify the importance of various features (or covariates). Two popular methods for defining variable importance are LOCO (Leave Out COvariates) and Shapley Values. We take a look at the properties of these methods and their advantages and disadvantages. We are particularly interested in the effect of correlation between features which can obscure interpretability. Contrary to some claims, Shapley values do not eliminate feature correlation. We critique the game theoretic axioms for Shapley values and we question their relevance for assessing feature importance. We propose new, more statistically oriented axioms for feature importance and some measures that satisfy these axioms. However, correcting for correlation is a Faustian bargain: removing the effect of correlation creates other forms of bias. Ultimately, we recommend a slightly modified version of LOCO. We briefly consider how to modify Shapley values to better address feature correlation.

Key words and phrases: Feature importance, shapley values, LOCO, interpretability.

REFERENCES

BARBER, R. F. and CANDÈS, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.* **43** 2055–2085. MR3375876 https://doi.org/10.1214/15-AOS1337

BÉNARD, C., DA VEIGA, S. and SCORNET, E. (2022). Mean decrease accuracy for random forests: Inconsistency, and a practical solution via the Sobol-MDA. *Biometrika* **109** 881–900. MR4519106 https://doi.org/10.1093/biomet/asac017

CHEN, H., COVERT, I. C., LUNDBERG, S. M. and LEE, S.-I. (2022). Algorithms to estimate Shapley value feature attributions. ArXiv preprint. Available at arXiv:2207.07605.

COVERT, I., LUNDBERG, S. and LEE, S.-I. (2020). Understanding global feature contributions through additive importance measures. *Adv. Neural Inf. Process. Syst.* **33** 17212–17223.

COVERT, I. C., LUNDBERG, S. and LEE, S.-I. (2021). Explaining by removing: A unified framework for model explanation. *J. Mach. Learn. Res.* **22** Paper No. 209. MR4329788

DAI, B., SHEN, X. and PAN, W. (2024). Significance tests of feature relevance for a black-box learner. *IEEE Trans. Neural Netw. Learn. Syst.* **35** 1898–1911. MR4710268 https://doi.org/10.1109/tnnls.2022.3185742

GAN, L., ZHENG, L. and ALLEN, G. I. (2022). Inference for interpretable machine learning: Fast, model-agnostic confidence intervals for feature importance. ArXiv preprint. Available at arXiv:2206.02088.

HAMA, N., MASE, M. and OWEN, A. B. (2022). Model free Shapley values for high dimensional data. ArXiv preprint. Available at arXiv:2211.08414.

HOOKER, G., MENTCH, L. and ZHOU, S. (2021). Unrestricted permutation forces extrapolation: Variable importance requires at least one more model, or there is no free variable importance. *Stat. Comput.* **31** Paper No. 82. MR4334494 https://doi.org/10.1007/s11222-021-10057-z

HUDSON, A. (2023). Nonparametric inference on non-negative dissimilarity measures at the boundary of the parameter space. ArXiv preprint. Available at arXiv:2306.07492.

IOOSS, B. and PRIEUR, C. (2019). Shapley effects for sensitivity analysis with correlated inputs: Comparisons with Sobol′ indices, numerical estimation and applications. *Int. J. Uncertain. Quantif.* **9** 493–514. MR4041329 https://doi.org/10.1615/Int.J.UncertaintyQuantification.2019028372

KOBAK, D., LOMOND, J. and SANCHEZ, B. (2020). The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *J. Mach. Learn. Res.* **21** Paper No. 169. MR4209455

KUCHIBHOTLA, A. K., BALAKRISHNAN, S. and WASSERMAN, L. (2021). The hulc: Confidence regions from convex hulls. ArXiv preprint. Available at arXiv:2105.14577.

KUMAR, I. E., VENKATASUBRAMANIAN, S., SCHEIDEGGER, C. and FRIEDLER, S. (2020). Problems with Shapley-value-based expla-

Isabella Verdinelli is Professor in Residence, Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA (e-mail: isabella@stat.cmu.edu). Larry Wasserman is Professor, Department of Statistics and Data Science, and Professor, Machine Learning Department, Carnegie Mellon University, Pennsylvania 15213, USA (e-mail: larry@stat.cmu.edu).

nations as feature importance measures. In *International Conference on Machine Learning* 5491–5500. PMLR.

LEI, J., G'SELL, M., RINALDO, A., TIBSHIRANI, R. J. and WASSERMAN, L. (2018). Distribution-free predictive inference for regression. *J. Amer. Statist. Assoc.* **113** 1094–1111. MR3862342 https://doi.org/10.1080/01621459.2017.1307116

LUNDBERG, S. M. and LEE, S.-I. (2017). A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30**.

MASE, M., OWEN, A. B. and SEILER, B. B. (2024). Variable importance without impossible data. *Annu. Rev. Stat. Appl.* **11** 153–178. MR4742280

MENTCH, L. and ZHOU, S. (2022). Getting better from worse: Augmented bagging and a cautionary tale of variable importance. *J. Mach. Learn. Res.* **23** Paper No. 224. MR4577663

OWEN, A. B. and PRIEUR, C. (2017). On Shapley value for measuring importance of dependent inputs. *SIAM/ASA J. Uncertain. Quantificat.* **5** 986–1002. MR3702874 https://doi.org/10.1137/16M1097717

RIBEIRO, M. T., SINGH, S. and GUESTRIN, C. (2016). Model-agnostic interpretability of machine learning. ArXiv preprint. Available at arXiv:1606.05386.

RINALDO, A., WASSERMAN, L. and G'SELL, M. (2019). Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *Ann. Statist.* **47** 3438–3469. MR4025748 https://doi.org/10.1214/18-AOS1784

ROBINS, J. M. and GREENLAND, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3** 143–155. https://doi.org/10.1097/00001648-199203000-00013

SAMAROV, M. (1993). Exploring regression structure using nonparametric functional estimation. *J. Amer. Statist. Assoc.* **88** 836–847. MR1242934 https://doi.org/10.1080/01621459.1993.10476348

SHAPLEY, L. S. (1953). A value for n-person games. In *Contributions to the Theory of Games*, *Vol.* 2. *Ann. of Math. Stud.*, *No.* 28 307–317. Princeton Univ. Press, Princeton, NJ. MR0053477

SOBOL', I. M. (1990). Estimation of the sensitivity of nonlinear mathematical models. *Mat. Model.* **2** 112–118. MR1052836

SOBOL', I. M. (1993). Sensitivity estimates for nonlinear mathematical models. *Math. Model. Comput. Exp.* **1** 407–414. MR1335161

VERDINELLI, I. and WASSERMAN, L. (2024). Decorrelated variable importance. *J. Mach. Learn. Res.* **25** Paper No. 7. MR4723857

WILLIAMSON, B. and FENG, J. (2020). Efficient nonparametric statistical inference on population feature importance using Shapley values. In *International Conference on Machine Learning* 10282–10291. PMLR.

WILLIAMSON, B. D., GILBERT, P. B., CARONE, M. and SIMON, N. (2021). Nonparametric variable importance assessment using machine learning techniques. *Biometrics* **77** 9–22. MR4229718 https://doi.org/10.1111/biom.13392

WILLIAMSON, B. D., GILBERT, P. B., SIMON, N. R. and CARONE, M. (2020). A unified approach for inference on algorithm-agnostic variable importance. ArXiv preprint. Available at arXiv:2004.03683.

ZHANG, L. and JANSON, L. (2020). Floodgate: Inference for model-free variable importance. ArXiv preprint. Available at arXiv:2007.01283.

Statistical Science
2024, Vol. 39, No. 4, 637–643
https://doi.org/10.1214/24-STS940

No Need for an Oracle: The Nonparametric Maximum Likelihood Decision in the Compound Decision Problem Is Minimax

Ya'acov Ritov

Abstract. We discuss the asymptotics of the nonparametric maximum likelihood estimator (NPMLE) in the normal mixture model. We then prove the convergence rate of the NPMLE decision in the empirical Bayes problem with normal observations. We point to (and heavily use) the connection between the NPMLE decision and Stein unbiased risk estimator (SURE). Next, we prove that the same solution is optimal in the compound decision problem where the unobserved parameters are not assumed to be random.

Similar results are usually claimed using an oracle-based argument. However, we contend that the standard oracle argument is not valid. It was only partially proved that it can be fixed, and the existing proofs of these partial results are tedious. Our approach, on the other hand, is straightforward and short.

Key words and phrases: Empirical Bayes, compound decision, minimax, oracle, nonparametric maximum likelihood.

REFERENCES

BICKEL, P. J. (1981). Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *Ann. Statist.* **9** 1301–1309. MR0630112

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models. Johns Hopkins Series in the Mathematical Sciences*. Johns Hopkins Univ. Press, Baltimore, MD. MR1245941

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 https://doi.org/10.1214/08-AOS630

CARROLL, R. J. and HALL, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.* **83** 1184–1186. MR0997599

DONOHO, D. L. and LOW, M. G. (1992). Renormalization exponents and optimal pointwise rates of convergence. *Ann. Statist.* **20** 944–970. MR1165601 https://doi.org/10.1214/aos/1176348665

DUBKOV, A. and MALAKHOV, A. (1976). Properties and interdependence of the cumulants of a random variable. *Radiophys. Quantum Electron.* **19** 833–839.

EFRON, B. (2011). Tweedie's formula and selection bias. *J. Amer. Statist. Assoc.* **106** 1602–1614. MR2896860 https://doi.org/10.1198/jasa.2011.tm11181

EFRON, B. (2019). Bayes, oracle Bayes and empirical Bayes. *Statist. Sci.* **34** 177–201. MR3983318 https://doi.org/10.1214/18-STS674

FAN, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.* **19** 1257–1272. MR1126324 https://doi.org/10.1214/aos/1176348248

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 https://doi.org/10.1214/09-LNMS5716

GREENSHTEIN, E. and RITOV, Y. (2019). Comment: Empirical Bayes, compound decisions and exchangeability [MR3983318]. *Statist. Sci.* **34** 224–228. MR3983324 https://doi.org/10.1214/19-STS709

GREENSHTEIN, E. and RITOV, Y. (2022). Generalized maximum likelihood estimation of the mean of parameters of mixtures. With applications to sampling and to observational studies. *Electron. J. Stat.* **16** 5934–5954. MR4515714 https://doi.org/10.1214/22-ejs2082

JAMESON, G. J. O. (2006). Counting zeros of generalised polynomials: Descartes' rule of signs and Laguerre's extensions. *Math. Gaz.* **90** 223–234. https://doi.org/10.1017/S0025557200179628

JEWELL, N. P. (1982). Mixtures of exponential distributions. *Ann. Statist.* **10** 479–484. MR0653523

JIANG, W. and ZHANG, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist.* **37** 1647–1684. MR2533467 https://doi.org/10.1214/08-AOS638

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 https://doi.org/10.1214/aoms/1177728066

KOENKER, R. and MIZERA, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *J. Amer. Statist. Assoc.* **109** 674–685. MR3223742 https://doi.org/10.1080/01621459.2013.869224

Ya'acov Ritov is Professor, Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109, USA (e-mail: yritov@umich.edu).

POLLARD, D. (1984). *Convergence of Stochastic Processes*. *Clinical Perspectives in Obstetrics and Gynecology*. Springer, New York.

PÓLYA, G. and SZEGÖ, G. (1945). *Aufgaben und Lehrsätze aus der Analysis*. Dover, New York. MR0015435

RICE, J. and ROSENBLATT, M. (1983). Smoothing splines: Regression, derivatives and deconvolution. *Ann*. *Statist*. **11** 141–156. MR0684872

RITOV, Y. (1983). Rates of convergence for Gaussian deconvolution.

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

SAHA, S. and GUNTUBOYINA, A. (2020). On the nonparametric maximum likelihood estimator for Gaussian location mixture densities with application to Gaussian denoising. *Ann*. *Statist*. **48** 738–762. MR4102674 https://doi.org/10.1214/19-AOS1817

STEFANSKI, L. and CARROLL, R. J. (1990). Deconvoluting kernel density estimators. *Statistics* **21** 169–184. MR1054861 https://doi.org/10.1080/02331889008802238

STUTE, W. (1983). Empirical processes indexed by smooth functions. *Stochastic Process*. *Appl*. **14** 55–66. MR0676273 https://doi.org/10.1016/0304-4149(83)90046-7

ZHANG, C.-H. (2003). Compound decision theory and empirical Bayes methods *Ann*. *Statist*. **31** 379–390. Dedicated to the memory of Herbert E. Robbins. MR1983534 https://doi.org/10.1214/aos/1051027872

Statistical Science
2024, Vol. 39, No. 4, 644–653
https://doi.org/10.1214/24-STS941

The van Trees Inequality in the Spirit of Hájek and Le Cam

Elisabeth Gassiat and Gilles Stoltz

Abstract. In honor of the 100th birth anniversary of Lucien Le Cam (November 18, 1924–April 24, 2000), we work out a version of the van Trees inequality in a Hájek–Le Cam spirit, that is, under minimal assumptions that, in particular, involve no direct pointwise regularity assumptions on densities but rather almost-everywhere differentiability in quadratic mean of the model. Surprisingly, it suffices that the latter differentiability holds along canonical directions—not along all directions. Also, we identify a (slightly stronger) version of the van Trees inequality as a very instance of a Cramér–Rao bound, that is, the van Trees inequality is not just a Bayesian analog of the Cramér–Rao bound. We provide, as an illustration, an elementary proof of the local asymptotic minimax theorem for quadratic loss functions, again assuming differentiability in quadratic mean only along canonical directions.

Key words and phrases: van Trees inequality, Cramér–Rao bound, differentiability in quadratic mean, local asymptotic minimax theorem.

REFERENCES

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. *Johns Hopkins Series in the Mathematical Sciences*. Johns Hopkins Univ. Press, Baltimore, MD. MR1245941

BOBROVSKY, B. Z., MAYER-WOLF, E. and ZAKAI, M. (1987). Some classes of global Cramér–Rao bounds. *Ann. Statist.* **15** 1421–1438. MR0913566 https://doi.org/10.1214/aos/1176350602

EVANS, L. C. and GARIEPY, R. F. (1992). *Measure Theory and Fine Properties of Functions*. *Studies in Advanced Mathematics*. CRC Press, Boca Raton, FL. MR1158660

GASSIAT, E., ROUSSEAU, J. and VERNET, E. (2018). Efficient semiparametric estimation and model selection for multidimensional mixtures. *Electron. J. Stat.* **12** 703–740. MR3769193 https://doi.org/10.1214/17-EJS1387

GILL, R. D. and LEVIT, B. Y. (1995). Applications of the van Trees inequality: A Bayesian Cramér–Rao bound. *Bernoulli* **1** 59–79. MR1354456 https://doi.org/10.2307/3318681

HÁJEK, J. (1970). A characterization of limiting distributions of regular estimates. *Z. Wahrsch. Verw. Gebiete* **14** 323–330. MR0283911 https://doi.org/10.1007/BF00533669

HÁJEK, J. (1972). Local asymptotic minimax and admissibility in estimation. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (*Univ. California, Berkeley, Calif.*, 1970/1971), *Vol. I*: *Theory of Statistics* 175–194. Univ. California Press, Berkeley, CA. MR0400513

IBRAGIMOV, I. A. and HAS'MINSKII, R. Z. (1981). *Statistical Estimation*: *Asymptotic Theory*. Springer, Berlin.

JUPP, P. E. (2010). A van Trees inequality for estimators on manifolds. *J. Multivariate Anal.* **101** 1814–1825. MR2651957 https://doi.org/10.1016/j.jmva.2010.03.007

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

LENSTRA, A. J. (2005). Cramér–Rao revisited. *Bernoulli* **11** 263–282. MR2132726 https://doi.org/10.3150/bj/1116340294

LETAC, G. (2022). The randomization by Wishart laws and the Fisher information. Preprint. Available at arXiv:2211.14137.

MOSS, D. and ROUSSEAU, J. (2024). Efficient Bayesian estimation and use of cut posterior in semiparametric hidden Markov models. *Electron. J. Stat.* **18** 1815–1886. MR4736272 https://doi.org/10.1214/23-ejs2201

POLLARD, D. (2001; 2005). Asymptotia (book in progress), chapter on Hellinger differentiability. Lecture notes from the Paris 2001 statistics semester at IHP, with a final edit in 2005. Available at http://www.stat.yale.edu/~pollard/Courses/607.spring05/handouts/DQM.pdf.

TAKATSU, K. and KUCHIBHOTLA, A. K. (2024). Generalized van Trees inequality: Local minimax bounds for non-smooth functionals and irregular statistical models. Preprint. Available at arXiv:2405.06437.

TITCHMARSH, E. C. (1939). *The Theory of Functions*, 2nd ed. Oxford Univ. Press, Oxford. MR3728294

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

VAN TREES, H. L. (1968). *Detection*, *Estimation and Modulation Theory*. Wiley, New York.

Elisabeth Gassiat is Professor, Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France (e-mail: elisabeth.gassiat@universite-paris-saclay.fr). Gilles Stoltz is CNRS Senior research fellow, Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France (e-mail: gilles.stoltz@universite-paris-saclay.fr).

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.