STATISTICAL SCIENCE

Volume 32, Number 3 August 2017

Statistical Science **[ISSN 0883-4237 (print); ISSN 2168-8745 (online)]**, Volume 32, Number 3, August 2017. Published quarterly by the Institute of Mathematical Statistics, 3163 Somerset Drive, Cleveland, OH 44122, USA. Periodicals postage paid at Cleveland, Ohio and at additional mailing offices.

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

Statistical Science
2017, Vol. 32, No. 3, 331–345
DOI: 10.1214/16-STS571

A Paradox from Randomization-Based Causal Inference

Peng Ding

Abstract. Under the potential outcomes framework, causal effects are defined as comparisons between potential outcomes under treatment and control. To infer causal effects from randomized experiments, Neyman proposed to test the null hypothesis of zero average causal effect (Neyman's null), and Fisher proposed to test the null hypothesis of zero individual causal effect (Fisher's null). Although the subtle difference between Neyman's null and Fisher's null has caused a lot of controversies and confusions for both theoretical and practical statisticians, a careful comparison between the two approaches has been lacking in the literature for more than eighty years. We fill this historical gap by making a theoretical comparison between them and highlighting an intriguing paradox that has not been recognized by previous researchers. Logically, Fisher's null implies Neyman's null. It is therefore surprising that, in actual completely randomized experiments, rejection of Neyman's null does not imply rejection of Fisher's null for many realistic situations, including the case with constant causal effect. Furthermore, we show that this paradox also exists in other commonly-used experiments, such as stratified experiments, matched-pair experiments and factorial experiments. Asymptotic analyses, numerical examples and real data examples all support this surprising phenomenon. Besides its historical and theoretical importance, this paradox also leads to useful practical implications for modern researchers.

Key words and phrases: Average null hypothesis, Fisher randomization test, potential outcome, randomized experiment, repeated sampling property, sharp null hypothesis.

REFERENCES

AGRESTI, A. and MIN, Y. (2004). Effects and non-effects of paired identical observations in comparing proportions with binary matched-pairs data. *Stat. Med.* **23** 65–75.

ANGRIST, J. D. and PISCHKE, J. S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton Univ. Press, Princeton, NJ.

ANSCOMBE, F. J. (1948). The validity of comparative experiments. *J. Roy. Statist. Soc. Ser. A* **111** 181–200; discussion, 200–211. MR0030181

ARONOW, P. M., GREEN, D. P. and LEE, D. K. K. (2014). Sharp bounds on the variance in randomized experiments. *Ann. Statist.* **42** 850–871. MR3210989

BARNARD, G. A. (1947). Significance tests for 2×2 tables. *Biometrika* **34** 123–138. MR0019285

BOX, G. E. P. (1992). Teaching engineers experimental design with a paper helicopter. *Qual. Eng.* **4** 453–459.

CHUNG, E. and ROMANO, J. P. (2013). Exact and asymptotically robust permutation tests. *Ann. Statist.* **41** 484–507. MR3099111

COX, D. R. (1958). The interpretation of the effects of non-additivity in the Latin square. *Biometrika* **45** 69–73.

COX, D. R. (1970). *The Analysis of Binary Data*. Methuen & Co., Ltd., London. MR0282453

COX, D. R. (1992). *Planning of Experiments*. Wiley, New York. Reprint of the 1958 original. MR1175752

COX, D. R. (2012). Statistical causality: Some historical remarks. In *Causality: Statistical Perspectives and Applications* (C. Berzuini, P. Dawid and L. Bernardinelli, eds.) 1–5. Wiley, New York.

DASGUPTA, T., PILLAI, N. S. and RUBIN, D. B. (2015). Causal inference from 2^K factorial designs by using potential out-

Peng Ding is Assistant Professor, Department of Statistics, University of California, Berkeley, 425 Evans Hall, Berkeley, California 94720, USA (e-mail: pengdingpku@berkeley.edu).

comes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 727–753. MR3382595

DING, P. (2017). Supplement to "A paradox from randomization-based causal inference." DOI:10.1214/16-STS571SUPP.

DING, P. and DASGUPTA, T. (2016). A potential tale of two-by-two tables from completely randomized experiments. *J. Amer. Statist. Assoc.* **111** 157–168. MR3494650

DING, P., FELLER, A. and MIRATRIX, L. W. (2016). Randomization inference for treatment effect variation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 655–671.

EBERHARDT, K. R. and FLIGNER, M. A. (1977). Comparison of two tests for equality of two proportions. *Amer. Statist.* **31** 151–155. MR0488444

EDEN, T. and YATES, F. (1933). On the validity of Fisher's z-test when applied to an actual example of non-normal data. *J. Agric. Sci.* **23** 6–17.

EDGINGTON, E. S. and ONGHENA, P. (2007). *Randomization Tests*, 4th ed. Chapman & Hall/CRC, Boca Raton, FL. With 1 CD-ROM (Windows). MR2291573

FIENBERG, S. E. and TANUR, J. M. (1996). Reconsidering the fundamental contributions of Fisher and Neyman on experimentation and sampling. *Int. Stat. Rev.* **64** 237–253.

FISHER, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain* **33** 503–513.

FISHER, R. A. (1935a). *The Design of Experiments*, 1st ed. Oliver and Boyd, Edinburgh.

FISHER, R. A. (1935b). Comment on "Statistical problems in agricultural experimentation". *Suppl. J. R. Stat. Soc.* **2** 154–157. 173.

FREEDMAN, D. A. (2008). On regression adjustments to experimental data. *Adv. in Appl. Math.* **40** 180–193. MR2388610

GAIL, M. H., MARK, S. D., CARROLL, R. J., GREEN, S. B. and PEE, D. (1996). On design considerations and randomization-based inference for community intervention trials. *Stat. Med.* **15** 1069–1092.

GREENLAND, S. (1991). On the logical justification of conditional tests for two-by-two contingency tables. *Amer. Statist.* **45** 248–251.

HÁJEK, J. (1960). Limiting distributions in simple random sampling from a finite population. *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **5** 361–374. MR0125612

HINKELMANN, K. and KEMPTHORNE, O. (2008). *Design and Analysis of Experiments, Vol.* 1: *Introduction to Experimental Design*, 2nd ed. Wiley, New York. MR2363107

HODGES, J. L. JR. and LEHMANN, E. L. (1964). *Basic Concepts of Probability and Statistics*. Holden-Day, Inc., San Francisco, CA–London–Amsterdam. MR0185709

HOEFFDING, W. (1952). The large-sample power of tests based on permutations of observations. *Ann. Math. Stat.* **23** 169–192. MR0057521

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

IMAI, K. (2008). Variance identification and efficiency analysis in randomized experiments under the matched-pair design. *Stat. Med.* **27** 4857–4873. MR2528770

IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*: *An Introduction*. Cambridge Univ. Press, New York. MR3309951

JANSSEN, A. (1997). Studentized permutation tests for non-i.i.d. hypotheses and the generalized Behrens–Fisher problem. *Statist. Probab. Lett.* **36** 9–21. MR1491070

KEMPTHORNE, O. (1952). *The Design and Analysis of Experiments*. Wiley, New York; Chapman & Hall, London. MR0045368

KEMPTHORNE, O. (1955). The randomization theory of experimental inference. *J. Amer. Statist. Assoc.* **50** 946–967. MR0071696

LANG, J. B. (2015). A closer look at testing the "no-treatment-effect" hypothesis in a comparative experiment. *Statist. Sci.* **30** 352–371. MR3383885

LEHMANN, E. L. (1999). *Elements of Large-Sample Theory*. Springer, New York. MR1663158

LI, X. and DING, P. (2016). Exact confidence intervals for the average causal effect on a binary outcome. *Stat. Med.* **35** 957–960.

LI, X. and DING, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *J. Amer. Statist. Assoc.* To appear.

LIN, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Ann. Appl. Stat.* **7** 295–318. MR3086420

LIN, W., HALPERN, S. D., PRASAD KERLIN, M. and SMALL, D. S. (2017). A "placement of death" approach for studies of treatment effects on ICU length of stay. *Stat. Methods Med. Res.* **26** 292–311. MR3592727

NELSEN, R. B. (2006). *An Introduction to Copulas*, 2nd ed. Springer, New York.

NEUHAUS, G. (1993). Conditional rank tests for the two-sample problem under random censorship. *Ann. Statist.* **21** 1760–1779. MR1245767

NEYMAN, J. (1935). Statistical problems in agricultural experimentation (with discussion). *Suppl. J. R. Stat. Soc.* **2** 107–180.

NEYMAN, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Sci.* **236** 333–380.

NEYMAN, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–472. Translated from the 1923 Polish original and edited by D. M. Dabrowska and T. P. Speed. MR1092986

PAULY, M., BRUNNER, E. and KONIETSCHKE, F. (2015). Asymptotic permutation tests in general factorial designs. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 461–473. MR3310535

PITMAN, E. J. G. (1937). Significance tests which may be applied to samples from any populations. *Suppl. J. R. Stat. Soc.* **4** 119–130.

PITMAN, E. J. G. (1938). Significance tests which can be applied to samples from any populations. III. The analysis of variance test. *Biometrika* **29** 322–335.

REID, C. (1982). *Neyman—From Life*. Springer, New York. MR0680939

RIGDON, J. and HUDGENS, M. G. (2015). Randomization inference for treatment effects on a binary outcome. *Stat. Med.* **34** 924–935. MR3310672

ROBBINS, H. (1977). A fundamental question of practical statistics. *Amer. Statist.* **31** 97.

ROBINS, J. M. (1988). Confidence intervals for causal parameters. *Stat. Med.* **7** 773–785.

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

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688–701.

RUBIN, D. B. (1980). Comment on "Randomization analysis of experimental data: The Fisher randomization test" by D. Basu. *J. Amer. Statist. Assoc.* **75** 591–593.

RUBIN, D. B. (1990). Comment on J. Neyman and causal inference in experiments and observational studies: "On the application of probability theory to agricultural experiments. Essay on principles. Section 9" [*Ann. Agric. Sci.* **10** (1923), 1–51]. *Statist. Sci.* **5** 472–480. MR1092987

RUBIN, D. B. (2004). Teaching statistical inference for causal effects in experiments and observational studies. *J. Educ. Behav. Stat.* **29** 343–367.

SABBAGHI, A. and RUBIN, D. B. (2014). Comments on the Neyman–Fisher controversy and its consequences. *Statist. Sci.* **29** 267–284. MR3264542

SAMII, C. and ARONOW, P. M. (2012). On equivalencies between design-based and regression-based variance estimators for randomized experiments. *Statist. Probab. Lett.* **82** 365–370. MR2875224

SCHEFFÉ, H. (1959). *The Analysis of Variance*. Wiley, New York; Chapman & Hall, London. MR0116429

SCHOCHET, P. Z. (2010). Is regression adjustment supported by the Neyman model for causal inference? *J. Statist. Plann. Inference* **140** 246–259. MR2568136

WELCH, B. L. (1937). On the z-test in randomized blocks and Latin squares. *Biometrika* **29** 21–52.

WHITE, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48** 817–838. MR0575027

WILK, M. B. and KEMPTHORNE, O. (1957). Non-additivities in a Latin square design. *J. Amer. Statist. Assoc.* **52** 218–236. MR0088137

WU, C. F. J. and HAMADA, M. S. (2009). *Experiments: Planning, Analysis, and Optimization*, 2nd ed. Wiley, Hoboken, NJ. MR2583259

YATES, F. (1937). The design and analysis of factorial experiments. Technical communication 35, Imperial Bureau of Soil Sciences, Harpenden.

Statistical Science
2017, Vol. 32, No. 3, 346–348
DOI: 10.1214/16-STS582

Understanding Ding's Apparent Paradox

Peter M. Aronow and Molly R. Offer-Westort

REFERENCES

ARONOW, P. M. (2013). Model assisted causal inference. Ph.D. thesis, Dept. Political Science, Yale Univ., New Haven, CT.

COPAS, J. B. (1973). Randomization models for the matched and unmatched 2×2 tables. *Biometrika* **60** 467–476. MR0448746

DING, P. (2017). A paradox from randomization-based causal inference. *Statist*. *Sci*. **32** 331–345.

ENGLE, R. F. (1984). Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. In *Handbook of Econometrics*, *Vol*. 2 775–826. Elsevier, Amsterdam.

FREEDMAN, D. A. (2008). On regression adjustments to experimental data. *Adv. in Appl*. *Math*. **40** 180–193. MR2388610

GERBER, A. S. and GREEN, D. P. (2012). *Field Experiments*: *Design*, *Analysis*, *and Interpretation*. Norton, New York.

LIN, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Ann*. *Appl*. *Stat*. **7** 295–318. MR3086420

MOSTELLER, F. and ROURKE, R. E. (1973). *Sturdy Statistics*: *Nonparametrics and Order Statistics*. Addison-Wesley, Reading, MA.

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

PRATT, J. W. (1964). Robustness of some procedures for the two-sample location problem. *J. Amer. Statist*. *Assoc*. **59** 665–680. MR0166871

REICHARDT, C. S. and GOLLOB, H. F. (1999). Justifying the use and increasing the power of a t test for a randomized experiment with a convenience sample. *Psychol*. *Methods* **4** 117–128.

ROMANO, J. P. (1990). On the behavior of randomization tests without a group invariance assumption. *J. Amer. Statist*. *Assoc*. **85** 686–692. MR1138350

SAMII, C. and ARONOW, P. M. (2012). On equivalencies between design-based and regression-based variance estimators for randomized experiments. *Statist*. *Probab*. *Lett*. **82** 365–370. MR2875224

Peter M. Aronow is Associate Professor, Departments of Political Science and Biostatistics, Yale University, 77 Prospect Street, New Haven, Connecticut 06520, USA (e-mail: peter.aronow@yale.edu). Molly R. Offer-Westort is a graduate student, Departments of Political Science and Statistics and Data Science, Yale University, 77 Prospect Street, New Haven, Connecticut 06520, USA (e-mail: molly.offer-westort@yale.edu).

Statistical Science
2017, Vol. 32, No. 3, 349–351
DOI: 10.1214/16-STS590

Randomization-Based Tests for "No Treatment Effects"

EunYi Chung

Abstract. Although both Fisher's and Neyman's tests are for testing "no treatment effects," they both test fundamentally different null hypotheses. While Neyman's null concerns the average casual effect, Fisher's null focuses on the individual causal effect. When conducting a test, researchers need to understand what is really being tested and what underlying assumptions are being made. If these fundamental issues are not fully appreciated, dubious conclusions regarding causal effects can be made.

Key words and phrases: Fisher's randomization test, Neyman's randomization test, treatment effect, Wilcoxon–Mann–Whitney rank sum test.

REFERENCES

CHUNG, E. and ROMANO, J. P. (2013). Exact and asymptotically robust permutation tests. *Ann. Statist.* **41** 484–507. MR3099111

CHUNG, E. and ROMANO, J. P. (2016). Asymptotically valid and exact permutation tests based on two-sample U-statistics. *J. Statist. Plann. Inference* **168** 97–105. MR3412224

ROMANO, J. P. (1990). On the behavior of randomization tests without a group invariance assumption. *J. Amer. Statist. Assoc.* **85** 686–692. MR1138350

EunYi Chung is Assistant Professor of Economics, Department of Economics, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA (e-mail: eunyi@illinois.edu).

Statistical Science
2017, Vol. 32, No. 3, 352–355
DOI: 10.1214/16-STS600

Inference from Randomized (Factorial) Experiments

R. A. Bailey

Abstract. This is a contribution to the discussion of the interesting paper by Ding [*Statist. Sci.* **32** (2017) 331–345], which contrasts approaches attributed to Neyman and Fisher. I believe that Fisher's usual assumption was unit-treatment additivity, rather than the "sharp null hypothesis" attributed to him. Fisher also developed the notion of interaction in factorial experiments. His explanation leads directly to the concept of marginality, which is essential for the interpretation of data from any factorial experiment.

Key words and phrases: Factorial design, marginality, randomisation, unit-treatment additivity.

REFERENCES

BAILEY, R. A. (1981). A unified approach to design of experiments. *J. Roy. Statist. Soc. Ser. A* **144** 214–223. MR0625801

BAILEY, R. A. (1991). Strata for randomised experiments. *J. Roy. Statist. Soc. Ser. B* **53** 27–78. With discussion and a reply by the author. MR1094275

BAILEY, R. A. (2008). *Design of Comparative Experiments*. *Cambridge Series in Statistical and Probabilistic Mathematics* **25**. Cambridge Univ. Press, Cambridge. MR2422352

BAILEY, R. A. (2015). Structures defined by factors. In *Handbook of Design and Analysis of Experiments* (A. Dean, M. Morris, J. Stufken and D. Bingham, eds.) 371–414. Chapman & Hall/CRC Press, Boca Raton.

BAILEY, R. A. and BRIEN, C. J. (2016). Randomization-based models for multitiered experiments: I. A chain of randomizations. *Ann. Statist.* **44** 1131–1164. MR3485956

BENNETT, J. H. (ed.) (1990). *Statistical Inference and Analysis*: *Selected Correspondence of R. A. Fisher*. The Clarendon Press, Oxford. With a preface by J. H. Bennett. MR1076366

COX, D. R. (1958a). The interpretation of the effects of non-additivity in the Latin square. *Biometrika* **45** 69–73.

COX, D. R. (1958b). *Planning of Experiments*. Wiley, New York; Chapman & Hall, London. MR0095561

DASGUPTA, T., PILLAI, N. S. and RUBIN, D. B. (2015). Causal inference from 2^K factorial designs by using potential outcomes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 727–753. MR3382595

DING, P. (2017). A paradox from randomization-based causal inference. *Statist. Sci.* **32** 331–345.

FISHER, R. A. (1925). *Statistical Methods for Research Workers*, 1st ed. Oliver and Boyd, Edinburgh.

FISHER, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain* **33** 503–513.

FISHER, R. A. (1935a). *The Design of Experiments*, 1st ed. Oliver and Boyd, Edinburgh.

FISHER, R. A. (1935b). Contribution to discussion of "Statistical problems in agricultural experimentation" by J. Neyman, with the help of K. Iwaskiewicz and St. Kołodziecyk. *J. Roy. Statist. Soc. Suppl.* **2** 154–157.

FISHER, R. A. (1937). *The Design of Experiments*, 2nd ed. Oliver and Boyd, Edinburgh.

FISHER, R. A. (1960). *The Design of Experiments*, 7th ed. Oliver and Boyd, Edinburgh.

FISHER BOX, J. (1978). *R. A. Fisher*: *The Life of a Scientist*. Wiley, New York. MR0500579

HURLBERT, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecol. Monogr.* **54** 187–211.

HURLBERT, S. H. (2009). The ancient black art and transdisciplinary extent of pseudoreplication. *Journal of Comparative Psychology* **123** 436–443.

KEMPTHORNE, O. (1975a). Fixed and mixed models in the analysis of variance. *Biometrics* **31** 473–486. MR0373183

KEMPTHORNE, O. (1975b). Inference from experiments and randomisation. In *A Survey of Statistical Design and Linear Models* (*Proc. Internat. Sympos.*, *Colorado State Univ.*, *Ft. Collins*, *Colo.*, 1973) (J. N. Srivastava, ed.) 303–331. North-Holland, Amsterdam. MR0375664

NELDER, J. A. (1965). The analysis of randomised experiments with orthogonal block structure. II. Treatment structure and the

R. A. Bailey is Professor of Mathematics and Statistics, School of Mathematics and Statistics, University of St. Andrews, St. Andrews, Fife KY16 9SS, United Kingdom, and Professor emerita of Statistics, School of Mathematical Sciences, Queen Mary University of London, Mile End Road, London E1 4NS, United Kingdom (e-mail: rab24@st-andrews.ac.uk).

general analysis of variance. *Proc. Roy. Soc. Ser. A* **283** 163–178. MR0174156

NELDER, J. A. (1977). A reformulation of linear models. *J. Roy. Statist. Soc. Ser. A* **140** 48–76. With discussion. MR0458743

NELDER, J. A. (1994). The statistics of linear models: Back to basics. *Stat. Comput.* **4** 221–234.

NELDER, J. A. (1998). The great mixed-model muddle is alive and flourishing, alas! *Food Qual. Prefer.* **9** 157–159.

NEYMAN, J. (1935). Statistical problems in agricultural experimentation. *J. Roy. Statist. Soc. Suppl.* **2** 107–154. With the help of K. Iwaskiewicz and St. Kołodziecyk.

PAULY, M., BRUNNER, E. and KONIETSCHKE, F. (2015). Asymptotic permutation tests in general factorial designs. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 461–473. MR3310535

REID, C. (1982). *Neyman—From Life*. Springer, New York. MR0680939

SENN, S. (2004). Added values. Controversies concerning randomisation and additivity in clinical trials. *Stat. Med.* **23** 3729–3753.

SPARKS, T. H., BAILEY, R. A. and ELSTON, D. A. (1997). Pseudoreplication: Common (mal)practice. *SETAC* (*Society of Environmental Toxicology and Chemistry*) *News* **17**(3) 12–13.

WILK, M. B. and KEMPTHORNE, O. (1957). Non-additivities in a Latin square design. *J. Amer. Statist. Assoc.* **52** 218–236. MR0088137

YATES, F. (1933). The formation of Latin squares for use in field experiments. *Empire Journal of Experimental Agriculture* **1** 235–244.

YATES, F. (1935a). Contribution to discussion of "Statistical problems in agricultural experimentation" by J. Neyman, with the help of K. Iwaskiewicz and St. Kołodziecyk. *J. Roy. Statist. Soc. Suppl.* **2** 161–166.

YATES, F. (1935b). Complex experiments. *J. Roy. Statist. Soc. Suppl.* **2** 181–223.

YATES, F. (1964). Sir Ronald Fisher and the design of experiments. *Biometrics* **20** 307–321.

YATES, F. (1965). A fresh look at the basic principles of the design and analysis of experiments. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol.* 4 777–790. Univ. California Press, Berkeley, CA.

Statistical Science
2017, Vol. 32, No. 3, 356–361
DOI: 10.1214/17-STS610

An Apparent Paradox Explained

Wen Wei Loh, Thomas S. Richardson and James M. Robins

REFERENCES

ARONOW, P. M., GREEN, D. P. and LEE, D. K. K. (2014). Sharp bounds on the variance in randomized experiments. *Ann. Statist.* **42** 850–871. MR3210989

BAYARRI, M. J. and BERGER, J. O. (2000). *P* values for composite null models. *J. Amer. Statist. Assoc.* **95** 1127–1142, 1157–1170. With comments and a rejoinder by the authors. MR1804239

CHIBA, Y. (2015). Exact tests for the weak causal null hypothesis on a binary outcome in randomized trials. *J. Biom. Biostat.* **6** 244.

CHUNG, E. and ROMANO, J. P. (2013). Exact and asymptotically robust permutation tests. *Ann. Statist.* **41** 484–507. MR3099111

DING, P. (2016). Personal communication.

DING, P. (2017). A paradox from randomization-based causal inference. *Statist. Sci.* **32** 331–345.

DING, P. and DASGUPTA, T. (2016). A potential tale of two-by-two tables from completely randomized experiments. *J. Amer. Statist. Assoc.* **111** 157–168. MR3494650

LANG, J. B. (2015). A closer look at testing the "no-treatment-effect" hypothesis in a comparative experiment. *Statist. Sci.* **30** 352–371. MR3383885

LOH, W. W. and RICHARDSON, T. S. (2017). Likelihood analysis for the finite population Neyman–Rubin binary causal model. In preparation.

PRÆSTGAARD, J. T. (1995). Permutation and bootstrap Kolmogorov–Smirnov tests for the equality of two distributions. *Scand. J. Stat.* **22** 305–322. MR1363215

RIGDON, J. and HUDGENS, M. G. (2015). Randomization inference for treatment effects on a binary outcome. *Stat. Med.* **34** 924–935. MR3310672

ROBINS, J. M. (1988). Confidence intervals for causal parameters. *Stat. Med.* **7** 773–785.

ROBINS, J. M., VAN DER VAART, A. and VENTURA, V. (2000). Asymptotic distribution of p values in composite null models. *J. Amer. Statist. Assoc.* **95** 1143–1167, 1171–1172. MR1804240

Wen Wei Loh is Postdoctoral Research Associate, Department of Biostatistics, University of North Carolina, CB #7420, Chapel Hill, North Carolina 27599, USA (e-mail: wloh@u.washington.edu). Thomas S. Richardson is Professor and Chair, Department of Statistics, University of Washington, Box 354322, Seattle, Washington 98195, USA (e-mail: thomasr@u.washington.edu). James M. Robins is Mitchell L. and Robin LaFoley Dong Professor of Epidemiology, Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Boston, Massachusetts 02115, USA (e-mail: robins@hsph.harvard.edu).

Statistical Science
2017, Vol. 32, No. 3, 362–366
DOI: 10.1214/17-STS571REJ

Rejoinder: A Paradox from Randomization-Based Causal Inference

Peng Ding

REFERENCES

ARONOW, P. M., GREEN, D. P. and LEE, D. K. K. (2014). Sharp bounds on the variance in randomized experiments. *Ann. Statist.* **42** 850–871. MR3210989

CHUNG, E. and ROMANO, J. P. (2016). Asymptotically valid and exact permutation tests based on two-sample U-statistics. *J. Statist. Plann. Inference* **168** 97–105. MR3412224

DING, P. (2014). A paradox from randomization-based causal inference. Technical report. Available at arXiv:1402.0142v1.

DING, P. (2017). A paradox from randomization-based causal inference. *Statist. Sci.* **32** 331–345.

DING, P. and DASGUPTA, T. (2016). A randomization-based perspective of analysis of variance: A test statistic robust to treatment effect heterogeneity. Available at arXiv:1608.01787.

FAN, Y. and PARK, S. S. (2010). Sharp bounds on the distribution of treatment effects and their statistical inference. *Econometric Theory* **26** 931–951. MR2646485

LI, X. and DING, P. (2016). Exact confidence intervals for the average causal effect on a binary outcome. *Stat. Med.* **35** 957–960. MR3457618

LU, J., DING, P. and DASGUPTA, T. (2015a). Construction of alternative hypotheses for randomization tests with ordinal outcomes. *Statist. Probab. Lett.* **107** 348–355. MR3412795

LU, J., DING, P. and DASGUPTA, T. (2015b). Treatment effects on ordinal outcomes: Causal estimands and sharp bounds. Available at arXiv:1507.01542.

RIGDON, J. and HUDGENS, M. G. (2015). Randomization inference for treatment effects on a binary outcome. *Stat. Med.* **34** 924–935. MR3310672

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

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York. MR0595165

Peng Ding is Assistant Professor, Department of Statistics, University of California, Berkeley, 425 Evans Hall, Berkeley, California 94720, USA (e-mail: pengdingpku@berkeley.edu).

Statistical Science
2017, Vol. 32, No. 3, 367–384
DOI: 10.1214/16-STS602

Logistic Regression: From Art to Science

Dimitris Bertsimas and Angela King

Abstract. A high quality logistic regression model contains various desirable properties: predictive power, interpretability, significance, robustness to error in data and sparsity, among others. To achieve these competing goals, modelers incorporate these properties iteratively as they hone in on a final model. In the period 1991–2015, algorithmic advances in Mixed-Integer Linear Optimization (MILO) coupled with hardware improvements have resulted in an astonishing 450 billion factor speedup in solving MILO problems. Motivated by this speedup, we propose modeling logistic regression problems algorithmically with a mixed integer nonlinear optimization (MINLO) approach in order to explicitly incorporate these properties in a joint, rather than sequential, fashion. The resulting MINLO is flexible and can be adjusted based on the needs of the modeler. Using both real and synthetic data, we demonstrate that the overall approach is generally applicable and provides high quality solutions in realistic timelines as well as a guarantee of suboptimality. When the MINLO is infeasible, we obtain a guarantee that imposing distinct statistical properties is simply not feasible.

Key words and phrases: Logistic regression, computational statistics, mixed integer nonlinear optimization.

REFERENCES

[1] BACH, F. R. (2008). Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.* **9** 1179–1225. MR2417268

[2] BACHE, K. and LICHMAN, M. (2014). UCI machine learning repository. Available at http://archive.ics.uci.edu/ml. Accessed: 2014-08-20.

[3] BEN-TAL, A., EL GHAOUI, L. and NEMIROVSKI, A. (2009). *Robust Optimization*. Princeton Univ. Press, Princeton, NJ. MR2546839

[4] BERK, R., BROWN, L., BUJA, A., ZHANG, K. and ZHAO, L. (2013). Valid post-selection inference. *Ann. Statist.* **41** 802–837. MR3099122

[5] BERTSIMAS, D., BROWN, D. B. and CARAMANIS, C. (2011). Theory and applications of robust optimization. *SIAM Rev.* **53** 464–501. MR2834084

[6] BERTSIMAS, D., DUNN, J., PAWLOWSKI, C. and ZHUO, Y. D. (2017). Robust classification. *J. Mach. Learn. Res.* To appear.

[7] BERTSIMAS, D. and KING, A. (2017). Supplement to "Logistic Regression: From Art to Science." DOI:10.1214/16-STS602SUPP.

[8] BERTSIMAS, D., KING, A. and MAZUMDER, R. (2016). Best subset selection via a modern optimization lens. *Ann. Statist.* **44** 813–852. MR3476618

[9] BEZANSON, J., KARPINSKI, S., SHAH, V. B. and EDELMAN, A. (2012). Julia: A fast dynamic language for technical computing. Preprint. Available at https://arxiv.org/abs/1209.5145.

[10] BIANCO, A. M. and YOHAI, V. J. (1996). Robust estimation in the logistic regression model. In *Robust Statistics*, *Data Analysis*, *and Computer Intensive Methods* (*Schloss Thurnau*, 1994). *Lect. Notes Stat.* **109** 17–34. Springer, New York. MR1491394

[11] BONAMI, P., KILINÇ, M. and LINDEROTH, J. (2012). Algorithms and software for convex mixed integer nonlinear programs. In *Mixed Integer Nonlinear Programming* 1–39. Springer, Berlin.

[12] BOX, G. E. P. and TIDWELL, P. W. (1962). Transformation of the independent variables. *Technometrics* **4** 531–550. MR0184313

[13] BUSSIECK, M. R. and VIGERSKE, S. (2010). *Minlp Solver Software*. In *Wiley Encyclopedia of Operations Research and Management Science*. Wiley Online Library.

[14] CARROLL, R. J. and PEDERSON, S. (1993). On robustness in the logistic regression model. *J. R. Stat. Soc. Ser. B. Stat.*

Dimitris Bertsimas is Professor, Operations Research Center and Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA (e-mail: dbertsim@mit.edu). Angela King is Data Scientist, End to End Analytics, San Francisco, California, USA (e-mail: aking10@mit.edu).

Methodol. **55** 693–706. MR1223937

[15] CHATTERJEE, S., HADI, A. S. and PRICE, B. (2012). *Regression Analysis by Example*, 5th ed. Wiley, New York.

[16] CRAMER, J. S. (2002). The origins of logistic regression. Technical report, Tinbergen Institute.

[17] CROUX, C. and HAESBROECK, G. (2003). Implementing the Bianco and Yohai estimator for logistic regression. *Comput*. *Statist*. *Data Anal*. **44** 273–295. Special issue in honour of Stan Azen: a birthday celebration. MR2020151

[18] CZYZYK, J., MESNIER, M. P. and MORÉ, J. J. (1998). The neos server. *J*. *Comput*. *Sci*. *Eng*. **5** 68–75.

[19] DOBSON, A. J. and BARNETT, A. G. (2008). *An Introduction to Generalized Linear Models*, 3rd ed. CRC Press, Boca Raton, FL. MR2459739

[20] DOLAN, E. D. (2001). Neos server 4.0 administrative guide. Preprint. Available at arXiv:cs/0107034.

[21] DURAN, M. A. and GROSSMANN, I. E. (1986). An outer-approximation algorithm for a class of mixed-integer nonlinear programs. *Math*. *Program*. **36** 307–339.

[22] EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann*. *Statist*. **7** 1–26. MR0515681

[23] ELDAR, Y. C. and KUTYNIOK, G. (2012). *Compressed Sensing*: *Theory and Applications*. Cambridge Univ. Press, London.

[24] FIGUEIREDO, M. A. T. (2003). Adaptive sparseness for supervised learning. *IEEE Trans*. *Pattern Anal*. *Mach*. *Intell*. **25** 1150–1159.

[25] FITHIAN, W., SUN, D. and TAYLOR, J. (2014). Optimal inference after model selection. Preprint. Available at arXiv:1410.2597.

[26] FREE SOFTWARE FOUNDATION (2015). GNU linear programming kit. Available at http://www.gnu.org/software/glpk/glpk.html. Accessed: 2015-03-06.

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

[28] FURNIVAL, G. M. and WILSON, R. W. (1974). Regressions by leaps and bounds. *Technometrics* **16** 499–511.

[29] GROPP, W. and MORÉ, J. (1997). Optimization environments and the neos server. In *Approximation Theory and Optimization* 167–182. Cambridge Univ. Press, Cambridge, UK.

[30] HILBE, J. M. (2011). *Logistic Regression Models*. CRC Press, Boca Raton, FL.

[31] HOSMER, D. W., JOVANOVIC, B. and LEMESHOW, S. (1989). Best subsets logistic regression. *Biometrics* **45** 1265–1270.

[32] HOSMER JR., D. W. and LEMESHOW, S. (2013). *Applied Logistic Regression*. Wiley, Hoboken, NJ.

[33] IBM ILOG CPLEX OPTIMIZATION STUDIO (2015). Cplex optimizer. Available at http://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/index.html. Accessed: 2015-03-06.

[34] GUROBI INC. (2014). Gurobi optimizer reference manual. Available at http://www.gurobi.com. Accessed: 2014-08-20.

[35] KIM, Y., KIM, J. and KIM, Y. (2006). Blockwise sparse regression. *Statist*. *Sinica* **16** 375–390. MR2267240

[36] KOH, K., KIM, S.-J. and BOYD, S. P. (2007). An interior-point method for large-scale l_1-regularized logistic regression. *J*. *Mach*. *Learn*. *Res*. **8** 1519–1555. MR2332440

[37] KRISHNAPURAM, B., CARIN, L., FIGUEIREDO, M. A. T. and HARTEMINK, A. J. (2005). Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Trans*. *Pattern Anal*. *Mach*. *Intell*. **27** 957–968.

[38] KRISHNAPURAM, B., HARTERNINK, A. J., CARIN, L. and FIGUEIREDO, M. A. T. (2004). A Bayesian approach to joint feature selection and classifier design. *IEEE Trans*. *Pattern Anal*. *Mach*. *Intell*. **26** 1105–1111.

[39] LEE, S.-I., LEE, H., ABBEEL, P. and NG, A. Y. (2006). Efficient ℓ_1 regularized logistic regression. In *Proceedings of the National Conference on Artificial Intelligence* **21** 401. AAAI Press, Menlo Park, CA.

[40] LOCKHART, R., TAYLOR, J., TIBSHIRANI, R. J. and TIBSHIRANI, R. (2014). A significance test for the Lasso. *Ann*. *Statist*. **42** 413–468. MR3210970

[41] LUBIN, M. and DUNNING, I. (2015). Computing in operations research using Julia. *INFORMS J*. *Comput*. **27** 238–248.

[42] MA, S., SONG, X. and HUANG, J. (2007). Supervised group lasso with applications to microarray data analysis. *BMC Bioinformatics* **8** 60.

[43] MARONNA, R., MARTIN, R. D. and YOHAI, V. (2006). *Robust Statistics*. Wiley, Chichester.

[44] MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2008). The group lasso for logistic regression. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **70** 53–71.

[45] MENARD, S. (2002). *Applied Logistic Regression Analysis* **106**. Sage, Thousand Oaks, CA.

[46] PREGIBON, D. (1981). Logistic regression diagnostics. *Ann*. *Statist*. **9** 705–724.

[47] RYAN, T. P. (2009). *Modern Regression Methods*, 2nd ed. Wiley, Hoboken, NJ. MR2459755

[48] SATO, T., TAKANO, Y., MIYASHIRO, R. and YOSHISE, A. (2016). Feature subset selection for logistic regression via mixed integer optimization. *Comput*. *Optim*. *Appl*. **64** 865–880. MR3506236

[49] SHAFIEEZADEH-ABADEH, S., MOHAJERIN, P. and KUHN, D. Distributionally robust logistic regression. In *Proceedings of the* 28*th International Conference on Neural Information Processing Systems* (*NIPS*'15), *Montreal*, *Canada*, *December* 07–12, 2015 (C. Cortes, D. D. Lee, M. Sugiyama and R. Garnett, eds.) 1576–1584. MIT Press, Cambridge, MA.

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

[51] TABACHNICK, B. G., FIDELL, L. S. et al. (2001). *Using Multivariate Statistics*. Allyn and Bacon, Boston, MA.

[52] TIPPING, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *J*. *Mach*. *Learn*. *Res*. **1** 211–244.

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

[54] ZHAO, P., ROCHA, G. and YU, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Ann*. *Statist*. **37** 3468–3497.

[55] ZOU, H. (2006). The adaptive lasso and its oracle properties. *J*. *Amer*. *Statist*. *Assoc*. **101** 1418–1429.

Statistical Science
2017, Vol. 32, No. 3, 385–404
DOI: 10.1214/16-STS604

Principles of Experimental Design for Big Data Analysis

Christopher C. Drovandi, Christopher C. Holmes, James M. McGree, Kerrie Mengersen, Sylvia Richardson and Elizabeth G. Ryan

Abstract. Big Datasets are endemic, but are often notoriously difficult to analyse because of their size, heterogeneity and quality. The purpose of this paper is to open a discourse on the potential for modern decision theoretic optimal experimental design methods, which by their very nature have traditionally been applied prospectively, to improve the analysis of Big Data through retrospective designed sampling in order to answer particular questions of interest. By appealing to a range of examples, it is suggested that this perspective on Big Data modelling and analysis has the potential for wide generality and advantageous inferential and computational properties. We highlight current hurdles and open research questions surrounding efficient computational optimisation in using retrospective designs, and in part this paper is a call to the optimisation and experimental design communities to work together in the field of Big Data analysis.

Key words and phrases: Active learning, Big Data, dimension reduction, experimental design, sub-sampling.

REFERENCES

AMZAL, B., BOIS, F. Y., PARENT, E. and ROBERT, C. P. (2006). Bayesian-optimal design via interacting particle systems. *J. Amer. Statist. Assoc.* **101** 773–785. MR2281248

AUSTIN, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar. Behav. Res.* **46** 399–424.

BARDENET, R., DOUCET, A. and HOLMES, C. (2014). Towards scaling up Markov chain Monte Carlo: An adaptive subsampling approach. In *Proceedings of the* 31*st International Conference on Machine Learning* (*ICML*-14) 405–413.

BARDENET, R., DOUCET, A. and HOLMES, C. (2015). On Markov chain Monte Carlo methods for tall data. Preprint. Available at arXiv:1505.02827 [stat.ME].

BOUVEYRON, C. and BRUNET-SAUMARD, C. (2014). Model-based clustering of high-dimensional data: A review. *Comput. Statist. Data Anal.* **71** 52–78. MR3131954

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

BRICK, J. M. and MONTAQUILA, J. M. (2009). Nonresponse and weighting. In *Sample Surveys: Design, Methods and Applications. Handbook of Statist.* **29** 163–185. Elsevier, Amsterdam. MR2654638

CHAMBERS, R. (1988). Design-adjusted regression with selectivity bias. *Appl. Stat.* **37** 323–334.

CHEN, C., GRENNAN, K., BADNER, J., ZHANG, D., JIN, E. G. L. and LI, C. (2011). Removing batch effects in analysis of expression microarray data: An evaluation of six batch adjustment methods. *PLoS ONE* **6** e17238.

CICHOSZ, P. (2015). *Data Mining Algorithms: Explained Using R*. Wiley, United Kingdom.

DAGOSTINO, R. B. (1998). Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat. Med.* **17** 2265–2281.

Christopher C. Drovandi is Senior Lecturer, James M. McGree is Associate Professor, Kerrie Mengersen is Professor, School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia, 4000 (e-mail: c.drovandi@qut.edu.au, james.mcgree@qut.edu.au, k.mengersen@qut.edu.au). Christopher C. Holmes is Professor of Biostatistics, Department of Statistics, and Nuffield Department of Clinical Medicine, University of Oxford, Oxford, United Kingdom, OX1 3TG (e-mail: cholmes@stats.ox.ac.uk). Sylvia Richardson is Professor, MRC Biostatistics Unit, Cambridge Institute of Public Health, Cambridge, United Kingdom, CB2 0SR (e-mail: sylvia.richardson@mrc-bsu.cam.ac.uk). Elizabeth G. Ryan is Statistician, Biostatistics & Health Informatics Department, King's College London, United Kingdom, SE5 8AF (e-mail: elizabeth.ryan@kcl.ac.uk).

DROVANDI, C. C., MCGREE, J. M. and PETTITT, A. N. (2013). Sequential Monte Carlo for Bayesian sequentially designed experiments for discrete data. *Comput. Statist. Data Anal.* **57** 320–335. MR2981091

DROVANDI, C. C. and TRAN, M.-N. (2016). Improving the efficiency of fully Bayesian optimal design of experiments using randomised quasi-Monte Carlo. Available at http://eprints.qut.edu.au/97889.

DUFFULL, S. B., GRAHAM, G., MENGERSEN, K. and ECCLESTON, J. (2012). Evaluation of the pre-posterior distribution of optimized sampling times for the design of pharmacokinetic studies. *J. Biopharm. Statist.* **22** 16–29. MR2872632

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499.

ELGAMAL, T. and HEFEEDA, M. (2015). Analysis of PCA algorithms in distributed environments. Preprint. Available at arXiv:1503.05214v2 [cs.DC].

ESPIRO-HERNANDEZ, G., GUSTAFSON, P. and BURSTYN, I. (2011). Bayesian adjustment for measurement error in continuous exposures in an individually matched case-control study. *BMC Med. Res. Methodol.* **11** 67–77.

FAN, J., FENG, Y. and RUI SONG, R. (2011). Nonparametric independence screening in sparse ultra-high dimensional additive models. *J. Amer. Statist. Assoc.* **106** 544–557.

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

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911.

FEDOROV, V. V. (1972). *Theory of Optimal Experiments*. Academic Press, New York.

FOUSKAKIS, D., NTZOUFRAS, I. and DRAPER, D. (2009). Bayesian variable selection using cost-adjusted BIC, with application to cost-effective measurement of quality of health care. *Ann. Appl. Stat.* **3** 663–690.

GAMA, J., ŽLIOBAITĖ, I., BIFET, A., PECHENIZKIY, M. and BOUCHACHIA, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys (CSUR)* **46** Article Number 44.

GANDOMI, A. and HAIDER, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *Internat. J. Inform. Management Sci.* **35** 137–144.

GELMAN, A. (2007). Struggles with survey weighting and regression modeling (with discussion). *Statist. Sci.* **22** 153–164.

GUHAA, S., HAFEN, R., ROUNDS, J., XIA, J., LI, J., XI, B. and CLEVELAND, W. S. (2012). Large complex data: Divide and recombine (D&R) with RHIPE. *Stat.* **1** 53–67.

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

KARVANEN, J., KULATHINAL, S. and GASBARRA, D. (2009). Optimal designs to select individuals for genotyping conditional on observed binary or survival outcomes and non-genetic covariates. *Comput. Statist. Data Anal.* **53** 1782–1793. MR2649544

KETTANEHA, N., BERGLUND, A. and WOLD, S. (2005). PCA and PLS with very large data sets. *Comput. Statist. Data Anal.* **48** 68–85.

KISH, L. and HESS, I. (1950). On noncoverage of sample dwellings. *J. Amer. Statist. Assoc.* **53** 509–524.

KLEINER, A., TALWALKAR, A., SARKAR, P. and JORDAN, M. I. (2014). A scalable bootstrap for massive data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 795–816.

KÜCK, H., DE FREITAS, N. and DOUCET, A. (2006). SMC samplers for Bayesian optimal nonlinear design. Technical Report, Univ. British Columbia, Vancouver, BC.

LEHMANN, H. P. and GOODMAN, S. N. (2000). Bayesian communication: A clinically significiant paradigm for electronic communication. *J. Am. Med. Inform. Assoc.* **7** 254–266.

LESKOVEC, J., RAJARAMAN, A. and ULLMAN, J. D. (2014). *Mining of Massive Datasets*. Cambridge Univ. Press, Cambridge.

LESSLER, J. T. and KALSBEEK, W. D. (1992). *Nonsampling Error in Surveys*. Wiley, New York. MR1193229

LEVY, P. S. and LEMESHOW, S. (1999). *Sampling of Populations: Methods and Applications*, 3rd ed. Wiley, New York.

LIANG, F., CHENG, Y., SONG, Q., PARK, J. and YANG, P. (2013). A resampling-based stochastic approximation method for analysis of large geostatistical data. *J. Amer. Statist. Assoc.* **108** 325–339. MR3174623

LIBERTY, E. (2013). Simple and deterministic matrix sketching. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 581–588. ACM, New York.

LONG, Q., SCAVINO, M., TEMPONE, R. and WANG, S. (2013). Fast estimation of expected information gains for Bayesian experimental designs based on Laplace approximations. *Comput. Methods Appl. Mech. Engrg.* **259** 24–39.

MASON, A., BEST, N., PLEWIS, I. and RICHARDSON, S. (2012). Strategy for modelling nonrandom missing data mechanisms in observational studies using Bayesian methods. *J. Off. Stat.* **28** 279–302.

MCCARRON, C. E., PULLENAYEGUM, E. M., THABANE, L., GOEREE, R. and TARRIDE, J.-E. (2011). Bayesian hierarchical models combining different study types and adjusting for covariate imbalances: A simulation study to assess model performance. *PLoS ONE* **6** e25635.

MENTRÉ, F., MALLET, A. and BACCAR, D. (1997). Optimal design in random-effects regression models. *Biometrika* **84** 429–442.

MUFF, S., RIEBLER, A., HELD, L., RUE, H. and SANER, P. (2015). Bayesian analysis of measurement error models using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **64** 231–252.

MÜLLER, P. (1999). Simulation-based optimal design. In *Bayesian Statistics*, 6 (*Alcoceber*, 1998) 459–474. Oxford Univ. Press, New York. MR1723509

MYERS, R. H., MONTGOMERY, D. C. and ANDERSON-COOK, C. M. (2009). *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, 3rd ed. Wiley, Hoboken, NJ. MR2464113

NAWARATHNA, L. S. and CHOUDHARY, P. K. (2015). A heteroscedastic measurement error model for method comparison data with replicate measurements. *Stat. Med.* **34** 1242–1258. MR3322747

OGUNGBENRO, K. and AARONS, L. (2007). Design of population pharmacokinetic experiments using prior information. *Xenobiotica* **37** 1311–1330.

OLESON, J. J., HE, C., SUN, D. and SHERIFF, S. (2007). Bayesian estimation in small areas when the sampling design strata differ from the study domains. *Surv. Methodol.* **33** 173–185.

OSWALD, F. L. and PUTKA, D. J. (2015). Statistical methods for big data. In *Big Data at Work*: *The Data Science Revolution and Organisational Psychology*. Routledge, New York.

PITCHFORTH, J. and MENGERSEN, K. (2012). Bayesian meta-analysis. In *Case Studies in Bayesian Statistics* 121–144. Wiley, New York.

PUKELSHEIM, F. (1993). *Optimal Design of Experiments*. Wiley, New York. MR1211416

REINIKAINEN, J., KARVANEN, J. and TOLONEN, H. (2016). Optimal selection of individuals for repeated covariate measurements in follow-up studies. *Stat*. *Methods Med*. *Res*. **25** 2420–2433. MR3572861

RICHARDSON, S. and GILKS, S. (1993). A Bayesian approach to measurement error problems in epidemiology using conditional independence models. *Am*. *J*. *Epidemiol*. **138** 430–442.

RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **71** 319–392.

RYAN, E. G., DROVANDI, C. C. and PETTITT, A. N. (2015). Simulation-based fully Bayesian experimental design for mixed effects models. *Comput*. *Statist*. *Data Anal*. **92** 26–39. MR3384249

SAVAGE, L. J. (1972). *The Foundations of Statistics*, revised ed. Dover Publications, New York. MR0348870

SCHIFANO, E. D., WU, J., WANG, C., YAN, J. and CHEN, M.-H. (2016). Online updating of statistical inference in the big data setting. *Technometrics* **58** 393–403. MR3520668

SCHMID, C. H. and MENGERSEN, K. (2013). Handbook of Meta-analysis in Ecology and Evolution. 145–173 Bayesian meta-analysis. Princeton Univ. Press, Princeton.

SCOTT, S. L., BLOCKER, A. W. and BONASSI, F. V. (2013). Bayes and big data: The consensus Monte Carlo algorithm. In *Bayes* 250.

SI, Y., PILLAI, N. and GELMAN, A. (2015). Bayesian nonparametric weighted sampling inference. *Bayesian Anal*. **10** 605–625.

SUYKENS, J. A. K., SIGNORETTO, M. and ARGYRIOU, A. (2015). *Regularization*, *Optimization*, *Kernels*, *and Support Vector Machines*. Chapman and Hall/CRC, Boca Raton, FL.

TAN, F. E. S. and BERGER, M. P. F. (1999). Optimal allocation of time points for the random effects models. *Comm*. *Statist*. **28** 517–540.

TOULIS, P., AIROLDI, E. and RENNI, J. (2014). Statistical analysis of stochastic gradient methods for generalized linear models. In *Proceedings of the* 31*st International Conference on Machine Learning* 667–675.

TROST, S. G., LOPRINZI, P. D., MOORE, R. and PFEIFFER, K. A. (2011). Comparison of accelerometer cut-points for predicting activity intensity in Youth. *Med*. *Sci*. *Sports Exerc*. **43** 1360–1368.

WANG, C., CHEN, M. H., SCHIFANO, E., WU, J. and YAN, J. (2015). A survey of statistical methods and computing for big data. Preprint. Available at arXiv:1502.07989v1 [stat.CO].

WOLPERT, R. L. and MENGERSEN, K. L. (2004a). Adjusted likelihoods for synthesizing empirical evidence from studies that differ in quality and design: Effects of environmental tobacco smoke. *Statist*. *Sci*. **3** 450–471.

WOLPERT, R. L. and MENGERSEN, K. L. (2004b). Adjusted likelihoods for synthesizing empirical evidence from studies that differ in quality and design: Effects of environmental tobacco smoke. *Statist*. *Sci*. **19** 450–471. MR2185626

WOODS, D. C., LEWIS, S. M., ECCLESTON, J. A. and RUSSELL, K. G. (2006). Designs for generalized linear models with several variables and model uncertainty. *Technometrics* **48** 284–292.

XI, B., CHEN, H., CLEVELAND, W. S. and TELKAMP, T. (2010). Statistical analysis and modelling of Internet VoIP traffic for network engineering. *Electron*. *J*. *Stat*. **4** 58–116.

YOO, C., RAMIREZ, L. and JUAN LIUZZI, J. (2014). Big data analysis using modern statistical and machine learning methods in medicine. *International Neurology Journal* **18** 50–57.

Statistical Science
2017, Vol. 32, No. 3, 405–431
DOI: 10.1214/17-STS611

Importance Sampling: Intrinsic Dimension and Computational Cost

S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso and A. M. Stuart

Abstract. The basic idea of importance sampling is to use independent samples from a proposal measure in order to approximate expectations with respect to a target measure. It is key to understand how many samples are required in order to guarantee accurate approximations. Intuitively, some notion of distance between the target and the proposal should determine the computational cost of the method. A major challenge is to quantify this distance in terms of parameters or statistics that are pertinent for the practitioner. The subject has attracted substantial interest from within a variety of communities. The objective of this paper is to overview and unify the resulting literature by creating an overarching framework. A general theory is presented, with a focus on the use of importance sampling in Bayesian inverse problems and filtering.

Key words and phrases: Importance sampling, notions of dimension, small noise, absolute continuity, inverse problems, filtering.

REFERENCES

[1] ACHUTEGUI, K., CRISAN, D., MIGUEZ, J. and RIOS, G. (2014). A simple scheme for the parallelization of particle filters and its application to the tracking of complex stochastic systems. Preprint. Available at arXiv:1407.8071.

[2] AGAPIOU, S., LARSSON, S. and STUART, A. M. (2013). Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems. *Stochastic Process. Appl.* **123** 3828–3860. MR3084161

[3] AGAPIOU, S. and MATHÉ, P. (2014). Preconditioning the prior to overcome saturation in Bayesian inverse problems. Preprint. Available at arXiv:1409.6496.

[4] AGAPIOU, S., PAPASPILIOPOULOS, O., SANZ-ALONSO, D. and STUART, A. M. (2017). Supplement to "Importance sampling: Intrinsic dimension and computational cost." DOI:10.1214/17-STS611SUPP.

[5] AGAPIOU, S., STUART, A. M. and ZHANG, Y.-X. (2014). Bayesian posterior contraction rates for linear severely ill-posed inverse problems. *J. Inverse Ill-Posed Probl.* **22** 297–321. MR3215928

[6] BAIN, A. and CRISAN, D. (2009). *Fundamentals of Stochastic Filtering* **3**. Springer, Berlin.

[7] BENDER, C. M. and ORSZAG, S. A. (1999). *Advanced Mathematical Methods for Scientists and Engineers. I. Asymptotic Methods and Perturbation Theory*. Springer, New York. MR1721985

[8] BENGTSSON, T., BICKEL, P., LI, B. et al. (2008). Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems. In *Probability and Statistics*: *Essays in Honor of David A. Freedman*. *Inst. Math. Stat.* (*IMS*) *Collect.* 316–334. IMS, Beachwood, OH.

[9] BESKOS, A., CRISAN, D. and JASRA, A. (2014). On the stability of sequential Monte Carlo methods in high dimensions. *Ann. Appl. Probab.* **24** 1396–1445. MR3211000

[10] BESKOS, A., CRISAN, D., JASRA, A., KAMATANI, K. and ZHOU, Y. (2014). A stable particle filter in high-dimensions. Preprint. Available at 1412.3501.

[11] BICKEL, P., LI, B. and BENGTSSON, T. (2008). Sharp failure rates for the bootstrap particle filter in high dimensions. In *Pushing the Limits of Contemporary Statistics*: *Contributions in Honor of Jayanta K. Ghosh*. *Inst. Math. Stat.* (*IMS*) *Collect.* **3** 318–329. IMS, Beachwood, OH. MR2459233

[12] BISHOP, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York. MR2247587

S. Agapiou is Lecturer, Department of Mathematics and Statistics, University of Cyprus, 1 University Avenue, 2109 Nicosia, Cyprus (e-mail: agapiou.sergios@ucy.ac.cy). O. Papaspiliopoulos is ICREA Research Professor, ICREA & Department of Economics, Universitat Pompeu Fabra, Ramon Trias Fargas 25-27, 08005 Barcelona, Spain (e-mail: omiros.papaspiliopoulos@upf.edu). D. Sanz-Alonso is Postdoctoral Research Associate, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02906, USA (e-mail: daniel_sanz-alonso1@brown.edu). A. M. Stuart is Bren Professor of Computing and Mathematical Sciences, Computing and Mathematical Sciences, California Institute of Technology, Pasadena, California 91125, USA (e-mail: astuart@caltech.edu).

[13] BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities*. Oxford Univ. Press, Oxford. MR3185193

[14] BUI-THANH, T. and GHATTAS, O. (2012). Analysis of the Hessian for inverse scattering problems: I. Inverse shape scattering of acoustic waves. *Inverse Problems* **28** 055001, 32. MR2915289

[15] BUI-THANH, T., GHATTAS, O., MARTIN, J. and STADLER, G. (2013). A computational framework for infinite-dimensional Bayesian inverse problems part I: The linearized case, with application to global seismic inversion. *SIAM J. Sci. Comput.* **35** A2494–A2523.

[16] CAFLISCH, R. E., MOROKOFF, W. J. and OWEN, A. B. (1997). *Valuation of Mortgage Backed Securities Using Brownian Bridges to Reduce Effective Dimension*. Dept. Mathematics, Univ. California, Los Angeles.

[17] CAPONNETTO, A. and DE VITO, E. (2007). Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.* **7** 331–368.

[18] CAVALIER, L. and TSYBAKOV, A. (2002). Sharp adaptation for inverse problems with random noise. *Probab. Theory Related Fields* **123** 323–354.

[19] CHATTERJEE, S. and DIACONIS, P. (2015). The sample size required in importance sampling. Preprint. Available at arXiv:1511.01437.

[20] CHEN, Y. (2005). Another look at rejection sampling through importance sampling. *Statist. Probab. Lett.* **72** 277–283. MR2153124

[21] CHOPIN, N. (2004). Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *Ann. Statist.* **32** 2385–2411. MR2153989

[22] CHOPIN, N. and PAPASPILIOPOULOS, O. (2016). A Concise Introduction to Sequential Monte Carlo.

[23] CHORIN, A. J. and MORZFELD, M. (2013). Conditions for successful data assimilation. *Journal of Geophysical Research: Atmospheres* **118** 11–522.

[24] CONSTANTINE, P. G. (2015). *Active Subspaces: Emerging Ideas for Dimension Reduction in Parameter Studies* **2**. SIAM Spotlights, Philadelphia, PA.

[25] COTTER, S. L., ROBERTS, G. O., STUART, A. M. and WHITE, D. (2013). MCMC methods for functions: Modifying old algorithms to make them faster. *Statist. Sci.* **28** 424–446. MR3135540

[26] CRISAN, D., DEL MORAL, P. and LYONS, T. (1999). Discrete filtering using branching and interacting particle systems. *Markov Process. Related Fields* **5** 293–318.

[27] CRISAN, D. and DOUCET, A. (2002). A survey of convergence results on particle filtering methods for practitioners. *IEEE Trans. Signal Process.* **50** 736–746.

[28] CRISAN, D. and ROZOVSKIĬ, B., eds. (2011). *The Oxford Handbook of Nonlinear Filtering*. Oxford Univ. Press, Oxford. MR2882749

[29] CUI, T., MARTIN, J., MARZOUK, Y. M., SOLONEN, A. and SPANTINI, A. (2014). Likelihood-informed dimension reduction for nonlinear inverse problems. *Inverse Problems* **30** 114015, 28. MR3274599

[30] DASHTI, M. and STUART, A. M. (2016). The Bayesian approach to inverse problems. In *Handbook of Uncertainty Quantification* (R. Ghanem, D. Higdon and H. Owhadi, eds.). http://arxiv.org/abs/1302.6989.

[31] DEL MORAL, P. (2004). *Feynman–Kac Formulae*. Springer, New York. MR2044973

[32] DEL MORAL, P. and MICLO, L. (2000). Branching and interacting particle systems approximations of Feynman–Kac formulae with applications to non-linear filtering. In *Séminaire de Probabilités, XXXIV. Lecture Notes in Math.* **1729** 1–145. Springer, Berlin. MR1768060

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

[34] DOUCET, A., GODSILL, S. and ANDRIEU, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Stat. Comput.* **10** 197–208.

[35] DOUKHAN, P. and LANG, G. (2009). Evaluation for moments of a ratio with application to regression estimation. *Bernoulli* **15** 1259–1286.

[36] DOWNEY, P. J. and WRIGHT, P. E. (2007). The ratio of the extreme to the sum in a random sequence. *Extremes* **10** 249–266.

[37] DUPUIS, P., SPILIOPOULOS, K. and WANG, H. (2012). Importance sampling for multiscale diffusions. *Multiscale Model. Simul.* **10** 1–27. MR2902596

[38] ENGL, H. W., HANKE, M. and NEUBAUER, A. (1996). *Regularization of Inverse Problems. Mathematics and Its Applications* **375**. Kluwer Academic, Dordrecht. MR1408680

[39] FRANKLIN, J. N. (1970). Well-posed stochastic extensions of ill-posed linear problems. *J. Math. Anal. Appl.* **31** 682–716. MR0267654

[40] FREI, M. and KÜNSCH, H. R. (2013). Bridging the ensemble Kalman and particle filters. *Biometrika* **100** 781–800.

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

[42] GELMAN, A., ROBERTS, G. O. and GILKS, W. R. (1996). Efficient Metropolis jumping rules. In *Bayesian Statistics*, 5 (*Alicante*, 1994). 599–607. Oxford Univ. Press, New York. MR1425429

[43] GIBBS, A. L. and SU, F. E. (2002). On choosing and bounding probability metrics. *Int. Stat. Rev.* **70** 419–435.

[44] GOODMAN, J., LIN, K. K. and MORZFELD, M. (2015). Small-noise analysis and symmetrization of implicit Monte Carlo samplers. *Comm. Pure Appl. Math.*

[45] HAN, W. (2013). On the Numerical Solution of the Filtering Problem Ph.D. thesis, Dept. Mathematics, Imperial College, London.

[46] JOHANSEN, A. M. and DOUCET, A. (2008). A note on auxiliary particle filters. *Statist. Probab. Lett.* **78** 1498–1504.

[47] KAHN, H. (1955). *Use of Different Monte Carlo Sampling Techniques*. Rand Corporation, Santa Monica, CA.

[48] KAHN, H. and MARSHALL, A. W. (1953). Methods of reducing sample size in Monte Carlo computations. *Journal of the Operations Research Society of America* **1** 263–278.

[49] KAIPIO, J. and SOMERSALO, E. (2005). *Statistical and Computational Inverse Problems. Applied Mathematical Sciences* **160**. Springer, New York. MR2102218

[50] KALLENBERG, O. (2002). *Foundations of Modern Probability. Probability and Its Applications*, 2nd ed. Springer, Berlin.

[51] KALMAN, R. E. (1960). A new approach to linear filtering and prediction problems. *J. Fluids Eng.* **82** 35–45.

[52] KALNAY, E. (2003). *Atmospheric Modeling, Data Assimilation, and Predictability*. Cambridge Univ. Press, Cambridge.

[53] KEKKONEN, H., LASSAS, M. and SILTANEN, S. (2016). Posterior consistency and convergence rates for Bayesian inversion with hypoelliptic operators. *Inverse Probl.* **32** 085005, 31. MR3535664

[54] KNAPIK, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2011). Bayesian inverse problems with Gaussian priors. *Ann. Statist.* **39** 2626–2657.

[55] KNAPIK, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2013). Bayesian recovery of the initial condition for the heat equation. *Comm. Statist. Theory Methods* **42** 1294–1313.

[56] KONG, A. (1992). A note on importance sampling using standardized weights. Technical Report 348, Dept. Statistics, Univ. Chicago, Chicago, IL.

[57] KONG, A., LIU, J. S. and WONG, W. H. (1994). Sequential imputations and Bayesian missing data problems. *J. Amer. Statist. Assoc.* **89** 278–288.

[58] KUO, F. Y. and SLOAN, I. H. (2005). Lifting the curse of dimensionality. *Notices Amer. Math. Soc.* **52** 1320–1329. MR2183869

[59] LANCASTER, P. and RODMAN, L. (1995). *Algebraic Riccati Equations*. Oxford Univ. Press, London.

[60] LASANEN, S. (2007). Measurements and infinite-dimensional statistical inverse theory. *PAMM* **7** 1080101–1080102.

[61] LASANEN, S. (2012). Non-Gaussian statistical inverse problems. Part I: Posterior distributions. *Inverse Probl. Imaging* **6** 215–266. MR2942739

[62] LASANEN, S. (2012). Non-Gaussian statistical inverse problems. Part II: Posterior convergence for approximated unknowns. *Inverse Probl. Imaging* **6** 267–287.

[63] LAX, P. D. (2002). *Functional Analysis*. Wiley, New York.

[64] LEHTINEN, M. S., PAIVARINTA, L. and SOMERSALO, E. (1989). Linear inverse problems for generalised random variables. *Inverse Probl.* **5** 599.

[65] LE GLAND, F., MONBET, V. and TRAN, V. (2009). Large sample asymptotics for the ensemble Kalman filter Ph.D. thesis, INRIA.

[66] LI, W., TAN, Z. and CHEN, R. (2013). Two-stage importance sampling with mixture proposals. *J. Amer. Statist. Assoc.* **108** 1350–1365.

[67] LIANG, F., LIU, C. and CARROLL, R. J. (2007). Stochastic approximation in Monte Carlo computation. *J. Amer. Statist. Assoc.* **102** 305–320.

[68] LIN, K., LU, S. and MATHÉ, P. (2015). Oracle-type posterior contraction rates in Bayesian inverse problems. *Inverse Probl. Imaging* **9**.

[69] LINDLEY, D. V. and SMITH, A. F. M. (1972). Bayes estimates for the linear model. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **34** 1–41. MR0415861

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

[71] LIU, J. S. (2008). *Monte Carlo Strategies in Scientific Computing*. Springer, New York. MR2401592

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

[73] LU, S. and MATHÉ, P. (2014). Discrepancy based model selection in statistical inverse problems. *J. Complexity* **30** 290–308. MR3183335

[74] MANDELBAUM, A. (1984). Linear estimators and measurable linear transformations on a Hilbert space. *Z. Wahrsch. Verw. Gebiete* **65** 385–397. MR0731228

[75] MARTIN, J., WILCOX, L. C., BURSTEDDE, C. and GHATTAS, O. (2012). A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. *SIAM J. Sci. Comput.* **34** A1460–A1487. MR2970260

[76] MCLEISH, D. L. and O'BRIEN, G. L. (1982). The expected ratio of the sum of squares to the square of the sum. *Ann. Probab.* **10** 1019–1028. MR0672301

[77] MÍGUEZ, J., CRISAN, D. and DJURIĆ, P. M. (2013). On the convergence of two sequential Monte Carlo methods for maximum a posteriori sequence estimation and stochastic global optimization. *Stat. Comput.* **23** 91–107.

[78] MORZFELD, M., TU, X., WILKENING, J. and CHORIN, A. J. (2015). Parameter estimation by implicit sampling. *Commun. Appl. Math. Comput. Sci.* **10** 205–225.

[79] MOSKOWITZ, B. and CAFLISCH, R. E. (1996). Smoothness and dimension reduction in quasi-Monte Carlo methods. *Math. Comput. Modelling* **23** 37–54. MR1398000

[80] MUELLER, J. L. and SILTANEN, S. (2012). *Linear and Nonlinear Inverse Problems with Practical Applications. Computational Science & Engineering* **10**. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. MR2986262

[81] OLIVER, D. S., REYNOLDS, A. C. and LIU, N. (2008). *Inverse Theory for Petroleum Reservoir Characterization and History Matching*. Cambridge Univ. Press, Cambridge.

[82] OWEN, A. and ZHOU, Y. (2000). Safe and effective importance sampling. *J. Amer. Statist. Assoc.* **95** 135–143. MR1803146

[83] PITT, M. K. and SHEPHARD, N. (1999). Filtering via simulation: Auxiliary particle filters. *J. Amer. Statist. Assoc.* **94** 590–599. MR1702328

[84] RAY, K. (2013). Bayesian inverse problems with non-conjugate priors. *Electron. J. Stat.* **7** 2516–2549.

[85] REBESCHINI, P. and VAN HANDEL, R. (2015). Can local particle filters beat the curse of dimensionality? *Ann. Appl. Probab.* **25** 2809–2866. MR3375889

[86] REN, Y.-F. and LIANG, H.-Y. (2001). On the best constant in Marcinkiewicz–Zygmund inequality. *Statist. Probab. Lett.* **53** 227–233. MR1841623

[87] SANZ-ALONSO, D. (2016). Importance sampling and necessary sample size: An information theory approach. Preprint. Available at arXiv:1608.08814.

[88] SLIVINSKI, L. and SNYDER, C. Practical estimates of the ensemble size necessary for particle filters.

[89] SNYDER, C. (2011). Particle filters, the "optimal" proposal and high-dimensional systems. In *Proceedings of the ECMWF Seminar on Data Assimilation for Atmosphere and Ocean*.

[90] SNYDER, C., BENGTSSON, T., BICKEL, P. and ANDERSON, J. (2008). Obstacles to high-dimensional particle filtering. *Mon. Weather Rev.* **136** 4629–4640.

[91] SNYDER, C., BENGTSSON, T. and MORZFELD, M. (2015). Performance bounds for particle filters using the optimal proposal. *Mon. Weather Rev.* **143** 4750–4761.

[92] SPANTINI, A., SOLONEN, A., CUI, T., MARTIN, J., TENORIO, L. and MARZOUK, Y. (2015). Optimal low-rank approximations of Bayesian linear inverse problems. *SIAM J. Sci. Comput.* **37** A2451–A2487.

[93] SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 583–639. MR1979380

[94] SPILIOPOULOS, K. (2013). Large deviations and importance sampling for systems of slow-fast motion. *Appl. Math. Optim.* **67** 123–161.

[95] STUART, A. M. (2010). Inverse problems: A Bayesian perspective. *Acta Numer.* **19** 451–559. MR2652785

[96] TAN, Z. (2004). On a likelihood approach for Monte Carlo integration. *J. Amer. Statist. Assoc.* **99** 1027–1036. MR2109492

[97] TIERNEY, L. (1998). A note on Metropolis–Hastings kernels for general state spaces. *Ann. Appl. Probab.* **8** 1–9. MR1620401

[98] VAN LEEUWEN, P. J. (2010). Nonlinear data assimilation in geosciences: An extremely efficient particle filter. *Q. J. R. Meteorol. Soc.* **136** 1991–1999.

[99] VANDEN-EIJNDEN, E. and WEARE, J. (2012). Data assimilation in the low noise, accurate observation regime with application to the Kuroshio current. *Mon. Weather Rev.* **141**.

[100] VOLLMER, S. J. (2013). Posterior consistency for Bayesian inverse problems through stability and regression results. *Inverse Probl.* **29** 125011.

[101] WHITELEY, N., LEE, A. and HEINE, K. (2016). On the role of interaction in sequential Monte Carlo algorithms. *Bernoulli* **22** 494–529. MR3449791

[102] ZHANG, T. (2002). Effective dimension and generalization of kernel learning. In *Advances in Neural Information Processing Systems* 454–461.

[103] ZHANG, T. (2005). Learning bounds for kernel regression using effective data dimensionality. *Neural Comput.* **17** 2077–2098.

[104] ZHANG, W., HARTMANN, C., WEBER, M. and SCHÜTTE, C. (2013). Importance sampling in path space for diffusion processes. *Multiscale Model. Simul.*

Statistical Science
2017, Vol. 32, No. 3, 432–454
DOI: 10.1214/17-STS612

Estimation of Causal Effects with Multiple Treatments: A Review and New Ideas

Michael J. Lopez and Roee Gutman

Abstract. The propensity score is a common tool for estimating the causal effect of a binary treatment in observational data. In this setting, matching, subclassification, imputation or inverse probability weighting on the propensity score can reduce the initial covariate bias between the treatment and control groups. With more than two treatment options, however, estimation of causal effects requires additional assumptions and techniques, the implementations of which have varied across disciplines. This paper reviews current methods, and it identifies and contrasts the treatment effects that each one estimates. Additionally, we propose possible matching techniques for use with multiple, nominal categorical treatments, and use simulations to show how such algorithms can yield improved covariate similarity between those in the matched sets, relative the pre-matched cohort. To sum, this manuscript provides a synopsis of how to notate and use causal methods for categorical treatments.

Key words and phrases: Causal inference, propensity score, multiple treatments, matching, observational data.

REFERENCES

ABADIE, A. and IMBENS, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica* **74** 235–267. MR2194325

ABADIE, A. and IMBENS, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica* **76** 1537–1557.

ARMSTRONG, C. S., JAGOLINZER, A. D. and LARCKER, D. F. (2010). Chief executive officer equity incentives and accounting irregularities. *J. Acc. Res.* **48** 225–271.

AUSTIN, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat. Med.* **28** 3083–3107. MR2750408

AUSTIN, P. C. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharm. Stat.* **10** 150–161.

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

AUSTIN, P. C. and SMALL, D. S. (2014). The use of bootstrapping when using propensity-score matching without replacement: A simulation study. *Stat. Med.* **33** 4306–4319.

BEZDEK, J. C., EHRLICH, R. and FULL, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Comput. Geosci.* **10** 191–203.

BRYSON, A., DORSETT, R. and PURDON, S. (2002). The use of propensity score matching in the evaluation of active labour market policies.

CALIENDO, M. and KOPEINIG, S. (2008). Some practical guidance for the implementation of propensity score matching. *J. Econ. Surv.* **22** 31–72.

CANGUL, M. Z., CHRETIEN, Y. R., GUTMAN, R. and RUBIN, D. B. (2009). Testing treatment effects in unconfounded studies under model misspecification: Logistic regression, discretization, and their combination. *Stat. Med.* **28** 2531–2551. MR2750307

CHERTOW, G. M., NORMAND, S. L. T. and MCNEIL, B. J. (2004). "Renalism": Inappropriately low rates of coronary angiography in elderly individuals with renal insufficiency. *J. Am. Soc. Nephrol.* **15** 2462–2468.

CRUMP, R. K., HOTZ, V. J., IMBENS, G. W. and MITNIK, O. A. (2009). Dealing with limited overlap in estimation of average

Michael J. Lopez is Assistant Professor of Statistics, Department of Mathematics, Skidmore College, 815 N. Broadway, Saratoga Springs, New York 12866, USA (e-mail: mlopez1@skidmore.edu). Roee Gutman is Assistant Professor of Biostatistics, Department of Biostatistics, Brown University, 121 S. Main Street, Providence, Rhode Island 02912, USA (e-mail: roee_gutman@brown.edu).

treatment effects. *Biometrika* **96** 187–199. MR2482144
D'AGOSTINO, R. B. (1998). Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat. Med.* **17** 2265–2281.
DAVIDSON, M. B., HIX, J. K., VIDT, D. G. and BROTMAN, D. J. (2006). Association of impaired diurnal blood pressure variation with a subsequent decline in glomerular filtration rate. *Arch. Intern. Med.* **166** 846–852.
DEARING, E., MCCARTNEY, K. and TAYLOR, B. A. (2009). Does higher quality early child care promote low-income children's math and reading achievement in middle childhood? *Child Dev.* **80** 1329–1349.
DEHEJIA, R. H. and WAHBA, S. (1998). Causal effects in nonexperimental studies: Re-evaluating the evaluation of training programs. Technical report, National Bureau of Economic Research.
DEHEJIA, R. H. and WAHBA, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Rev. Econ. Stat.* **84** 151–161.
DORE, D. D., SWAMINATHAN, S., GUTMAN, R., TRIVEDI, A. N. and MOR, V. (2013). Different analyses estimate different parameters of the effect of erythropoietin stimulating agents on survival in end stage renal disease: A comparison of payment policy analysis, instrumental variables, and multiple imputation of potential outcomes. *J. Clin. Epidemiol.* **66** S42–S50.
DORSETT, R. (2006). The new deal for young people: Effect on the labour market status of young men. *Labour Econ.* **13** 405–422.
DRICHOUTIS, A. C., LAZARIDIS, P. and NAYGA JR., R. M. (2005). Nutrition knowledge and consumer use of nutritional food labels. *Eur. Rev. Agricult. Econ.* **32** 93–118.
EFRON, B. and TIBSHIRANI, R. J. (1994). *An Introduction to the Bootstrap*. CRC Press, Boca Raton.
FENG, P., ZHOU, X.-H., ZOU, Q.-M., FAN, M.-Y. and LI, X.-S. (2012). Generalized propensity score for estimating the average treatment effect of multiple treatments. *Stat. Med.* **31** 681–697.
FILARDO, G., HAMILTON, C., HAMMAN, B. and GRAYBURN, P. (2007). Obesity and stroke after cardiac surgery: The impact of grouping body mass index. *Ann. Thorac. Surg.* **84** 720–722.
FILARDO, G., HAMILTON, C., HAMMAN, B., HEBELER JR., R. F. and GRAYBURN, P. A. (2009). Relation of obesity to atrial fibrillation after isolated coronary artery bypass grafting. *Am. J. Cardiol.* **103** 663–666.
FRANK, R., AKRESH, I. R. and LU, B. (2010). Latino immigrants and the US racial order. *Am. Sociol. Rev.* **75** 378–401.
GUTMAN, R. and RUBIN, D. B. (2013). Robust estimation of causal effects of binary treatments in unconfounded studies with dichotomous outcomes. *Stat. Med.* **32** 1795–1814. MR3067363
GUTMAN, R. and RUBIN, D. B. (2015). Estimation of causal effects of binary treatments in unconfounded studies. *Stat. Med.* **34** 3381–3398. MR3412639
HADE, E. M. (2012). Propensity score adjustment in multiple group observational studies: Comparing matching and alternative methods. Ph.D. thesis, Ohio State University.
HADE, E. M. and LU, B. (2014). Bias associated with using the estimated propensity score as a regression covariate. *Stat. Med.* **33** 74–87. MR3141554
HEDMAN, L. and VAN HAM, M. (2012). *Understanding Neighbourhood Effects*: *Selection Bias and Residential Mobility*. Springer, Berlin.
HILL, J. and REITER, J. P. (2006). Interval estimation for treatment effects using propensity score matching. *Stat. Med.* **25** 2230–2256. MR2240098
HOLLAND, P. W. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81** 945–970. MR0867618
HOTT, J. R., BRUNELLE, N. and MYERS, J. A. (2012). KD-tree algorithm for propensity score matching with three or more treatment groups. Division of Pharmacoepidemiology and Pharmacoeconomics, Technical Report Series.
IACUS, S. M., KING, G. and PORRO, G. (2011). Causal inference without balance checking: Coarsened exact matching. *Polit. Anal.* mpr013.
IMAI, K. and RATKOVIC, M. (2014). Covariate balancing propensity score. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 243–263.
IMAI, K. and VAN DYK, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *J. Amer. Statist. Assoc.* **99** 854–866. MR2090918
IMBENS, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **87** 706–710. MR1789821
IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference in Statistics*, *Social*, *and Biomedical Sciences*. Cambridge Univ. Press, Cambridge.
JOFFE, M. M. and ROSENBAUM, P. R. (1999). Invited commentary: Propensity scores. *Am. J. Epidemiol.* **150** 327–333.
JOHNSON, R. A., WICHERN, D. W. et al. (1992). *Applied Multivariate Statistical Analysis* **4**. Prentice Hall, Englewood Cliffs, NJ.
KANG, J. D. Y. and SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22** 523–539.
KARP, R. M. (1972). Reducibility among combinatorial problems. In *Complexity of Computer Computations* (*Proc. Sympos.*, *IBM Thomas J. Watson Res. Center*, *Yorktown Heights*, *N.Y.*, 1972) 85–103. Plenum, New York. MR0378476
KILPATRICK, R. D., GILBERTSON, D., BROOKHART, M. A., POLLEY, E., ROTHMAN, K. J. and BRADBURY, B. D. (2013). Exploring large weight deletion and the ability to balance confounders when using inverse probability of treatment weighting in the presence of rare treatment decisions. *Pharmacoepidemiol. Drug Saf.* **22** 111–121.
KOSTEAS, V. D. (2010). The effect of exercise on earnings: Evidence from the NLSY. *J. Labor Res.* 1–26.
LECHNER, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. *Econom. Evaluation Labour Mark. Polic.* 43–58.
LECHNER, M. (2002). Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies. *Rev. Econ. Stat.* **84** 205–220.
LEE, B. K., LESSLER, J. and STUART, E. A. (2011). Weight trimming and propensity score weighting. *PLoS ONE* **6** e18174.
LEVIN, I. and ALVAREZ, R. M. (2009). Measuring the effects of voter confidence on political participation: An application to the 2006 Mexican election. VTP Working Paper 75, Caltech/MIT Voting Technology Project.
LITTLE, R. J. A. (1988). Missing-data adjustments in large surveys. *J. Bus. Econom. Statist.* 287–296.

LOPEZ, M. J. and GUTMAN, R. (2014). Estimating the average treatment effects of nutritional label use using subclassification with regression adjustment. *Stat. Methods Med. Res.* DOI:10.1177/0962280214560046.

LU, B., ZANUTTO, E., HORNIK, R. and ROSENBAUM, P. R. (2001). Matching with doses in an observational study of a media campaign against drug abuse. *J. Amer. Statist. Assoc.* **96** 1245–1253.

MCCAFFREY, D. F., RIDGEWAY, G. and MORRAL, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol. Methods* **9** 403–425.

MCCAFFREY, D. F., GRIFFIN, B. A., ALMIRALL, D., SLAUGHTER, M. E., RAMCHAND, R. and BURGETTE, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat. Med.* **32** 3388–3414. MR3074364

MCCULLAGH, P. (1980). Regression models for ordinal data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **42** 109–142. MR0583347

MOORE, A. W. (1991). An introductory tutorial on kd-trees. Extract from PhD thesis. Technical report.

QUADE, D. (1979). Using weighted rankings in the analysis of complete blocks with additive block effects. *J. Amer. Statist. Assoc.* **74** 680–683.

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

RASSEN, J. A., SOLOMON, D. H., GLYNN, R. J. and SCHNEEWEISS, S. (2011). Simultaneously assessing intended and unintended treatment effects of multiple treatment options: A pragmatic "matrix design." *Pharmacoepidemiol. Drug Saf.* **20** 675–683.

RASSEN, J. A., SHELAT, A. A., FRANKLIN, J. M., GLYNN, R. J., SOLOMON, D. H. and SCHNEEWEISS, S. (2013). Matching by propensity score in cohort studies with three treatment groups. *Epidemiology* **24** 401–409.

ROBINS, J. M., HERNAN, M. A. and BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11** 550–560.

ROSENBAUM, P. R. (1991). A characterization of optimal designs for observational studies. *J. R. Stat. Soc., B* **53** 597–610.

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

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

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

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.

ROYSTON, P., ALTMAN, D. G. and SAUERBREI, W. (2006). Dichotomizing continuous predictors in multiple regression: A bad idea. *Stat. Med.* **25** 127–141. MR2222078

RUBIN, D. B. (1973). Matching to remove bias in observational studies. *Biometrics* **29** 159–183.

RUBIN, D. B. (1975). Bayesian inference for causality: The importance of randomization. In *The Proceedings of the Social Statistics Section of the American Statistical Association* 233–239.

RUBIN, D. B. (1976). Multivariate matching methods that are equal percent bias reducing. II. Maximums on bias reduction for fixed sample sizes. *Biometrics* **32** 121–132. MR0400556

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

RUBIN, D. B. (1980). Discussion of Basu's paper. *J. Amer. Statist. Assoc.* **75** 591–593.

RUBIN, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Serv. Outcomes Res. Methodol.* **2** 169–188.

RUBIN, D. B. and THOMAS, N. (1992a). Affinely invariant matching methods with ellipsoidal distributions. *Ann. Statist.* **20** 1079–1093.

RUBIN, D. B. and THOMAS, N. (1992b). Characterizing the effect of matching using linear propensity score methods with normal distributions. *Biometrika* **79** 797–809. MR1209479

RUBIN, D. B. and THOMAS, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics* **52** 249–264.

SAKIA, R. M. (1992). The Box–Cox transformation technique: A review. *Statistician* **42** 169–178.

SAS INSTITUTE INC. (2003). *SAS/STAT Software*. SAS Institute Inc., Cary, NC.

SCHNEEWEISS, S., SETOGUCHI, S., BROOKHART, A., DORMUTH, C. and WANG, P. S. (2007). Risk of death associated with the use of conventional versus atypical antipsychotic drugs among elderly patients. *CMAJ, Can. Med. Assoc. J.* **176** 627–632.

SEKHON, J. (2011). Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *J. Stat. Softw.* **42**, 1–52.

SNODGRASS, G., BLOKLAND, A. A. J., HAVILAND, A., NIEUWBEERTA, P. and NAGIN, D. S. (2011). Does the time cause the crime? An examination of the relationship between time served and reoffending in the Netherlands. *Criminology* **49** 1149–1194.

SPLAWA-NEYMAN, J., DABROWSKA, D. M. and SPEED, T. P. (1990 [1923]). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–472.

SPREEUWENBERG, M. D., BARTAK, A., CROON, M. A., HAGENAARS, J. A., BUSSCHBACH, J. J. V., ANDREA, H., TWISK, J. and STIJNEN, T. (2010). The multiple propensity score as control for bias in the comparison of more than two treatment arms: An introduction from a case study in mental health. *Med. Care* **48** 166.

SPRENT, P. and SMEETON, N. C. (2007). *Applied Nonparametric Statistical Methods*. CRC Press, Boca Raton, FL.

STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statist. Sci.* **25** 1–21. MR2741812

STUART, E. A. and RUBIN, D. B. (2008). Best practices in quasi-experimental designs. *Best Pract. Quant. Methods* 155–176.

TAN, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika* **97** 661–682. MR2672490

TCHERNIS, R., HORVITZ-LENNON, M. and NORMAND, S. L. T. (2005). On the use of discrete choice models for causal inference. *Stat. Med.* **24** 2197–2212.

TU, C., JIAO, S. and KOH, W. Y. (2012). Comparison of clustering algorithms on generalized propensity score in observational studies: A simulation study. *J. Stat. Comput. Simul.* **83** 2206–2218.

VERMORKEN, J. B., PARMAR, M. K., BRADY, M. F., EISENHAUER, E. A., HOGBERG, T., OZOLS, R. F., ROCHON, J.,RUSTIN, G. J., SAGAE, S., VERHEIJEN, R. H. et al. (2005). Clinical trials in ovarian carcinoma: Study methodology. *Ann*. *Oncol*. **16** viii20.

YANOVITZKY, I., ZANUTTO, E. and HORNIK, R. (2005). Estimating causal effects of public health education campaigns using propensity score methodology. *Eval*. *Program Plann*. **28** 209–220.

ZANUTTO, E., LU, B. and HORNIK, R. (2005). Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. *J*. *Educ*. *Behav*. *Stat*. **30** 59–73.

ZUBIZARRETA, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *J*. *Amer*. *Statist*. *Assoc*. **107** 1360–1371. MR3036400

Statistical Science
2017, Vol. 32, No. 3, 455–468
DOI: 10.1214/17-STS613

On the Choice of Difference Sequence in a Unified Framework for Variance Estimation in Nonparametric Regression

Wenlin Dai, Tiejun Tong and Lixing Zhu

Abstract. Difference-based methods do not require estimating the mean function in nonparametric regression and are therefore popular in practice. In this paper, we propose a unified framework for variance estimation that combines the linear regression method with the higher-order difference estimators systematically. The unified framework has greatly enriched the existing literature on variance estimation that includes most existing estimators as special cases. More importantly, the unified framework has also provided a smart way to solve the challenging difference sequence selection problem that remains a long-standing controversial issue in nonparametric regression for several decades. Using both theory and simulations, we recommend to use the ordinary difference sequence in the unified framework, no matter if the sample size is small or if the signal-to-noise ratio is large. Finally, to cater for the demands of the application, we have developed a unified R package, named VarED, that integrates the existing difference-based estimators and the unified estimators in nonparametric regression and have made it freely available in the R statistical program http://cran.r-project.org/web/packages/.

Key words and phrases: Difference-based estimator, nonparametric regression, optimal difference sequence, ordinary difference sequence, residual variance.

REFERENCES

BENKO, M., HÄRDLE, W. and KNEIP, A. (2009). Common functional principal components. *Ann. Statist.* **37** 1–34. MR2488343

BERKEY, C. S. (1982). Bayesian approach for a nonlinear growth model. *Biometrics* **38** 953–961.

BLIZNYUK, N., CARROLL, R. J., GENTON, M. G. and WANG, Y. (2012). Variogram estimation in the presence of trend. *Stat. Interface* **5** 159–168. MR2928067

BROWN, L. D. and LEVINE, M. (2007). Variance estimation in nonparametric regression via the difference sequence method. *Ann. Statist.* **35** 2219–2232. MR2363969

CHARNIGO, R., HALL, B. and SRINIVASAN, C. (2011). A generalized C_p criterion for derivative estimation. *Technometrics* **53** 238–253. MR2857702

CHENG, M.-Y., PENG, L. and WU, J.-S. (2007). Reducing variance in univariate smoothing. *Ann. Statist.* **35** 522–542. MR2336858

COOK, J. R. and STEFANSKI, L. A. (1995). Simulation-extrapolation estimation in parametric measurement error models. *J. Amer. Statist. Assoc.* **89** 1314–1328.

DAI, W., TONG, T. and GENTON, M. G. (2016). Optimal estimation of derivatives in nonparametric regression. *J. Mach. Learn. Res.* **17(164)** 1–25.

DAI, W., TONG, T. and ZHU, L. (2017). Supplement to "On the Choice of Difference Sequence in a Unified Framework for Variance Estimation in Nonparametric Regression." DOI:10.1214/17-STS613SUPP.

DAI, W., MA, Y., TONG, T. and ZHU, L. (2015). Difference-based variance estimation in nonparametric regression with repeated measurement data. *J. Statist. Plann. Inference* **163** 1–20.

DETTE, H. and HETZLER, B. (2009). A simple test for the parametric form of the variance function in nonparametric regression. *Ann. Inst. Statist. Math.* **61** 861–886. MR2556768

DETTE, H., MUNK, A. and WAGNER, T. (1998). Estimating the

Wenlin Dai is Postdoctoral Fellow, CEMSE Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia (e-mail: wenlin.dai@kaust.edu.sa). Tiejun Tong is Associate Professor and Corresponding Author, Department of Mathematics, Hong Kong Baptist University, Hong Kong (e-mail: tongt@hkbu.edu.hk). Lixing Zhu is Chair Professor, Department of Mathematics, Hong Kong Baptist University, Hong Kong (e-mail: lzhu@hkbu.edu.hk).

variance in nonparametric regression—what is a reasonable choice? *J. Roy Statist. Soc. Ser. B.* **60** 751–764. MR1649480

DE BRABANTER, K., DE BRABANTER, J., DE MOOR, B. and GIJBELS, I. (2013). Derivative estimation with local polynomial fitting. *J. Mach. Learn. Res.* **14** 281–301.

EINMAHL, J. H. J. and VAN KEILEGOM, I. (2008). Tests for independence in nonparametric regression. *Statist. Sinica* **18** 601–615. MR2411617

EUBANK, R. L. and SPIEGELMAN, C. H. (1990). Testing the goodness of fit of a linear model via nonparametric regression techniques. *J. Amer. Statist. Assoc.* **85** 387–392. MR1141739

GASSER, T., KNEIP, A. and KÖHLER, W. (1991). A flexible and fast method for automatic smoothing. *J. Amer. Statist. Assoc.* **86** 643–652. MR1147088

GASSER, T., SROKA, L. and JENNEN-STEINMETZ, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* **73** 625–633. MR0897854

HALL, P. and HECKMAN, N. E. (2000). Testing for monotonicity of a regression mean by calibrating for linear functions. *Ann. Statist.* **28** 20–39. MR1762902

HALL, P., KAY, J. W. and TITTERINGTON, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* **77** 521–528. MR1087842

HALL, P. and KEILEGOM, I. V. (2003). Using difference-based methods for inference in nonparametric regression with time series errors. *J. Roy. Statist. Soc. Ser. B* **65** 443–456. MR1983757

HALL, P. and MARRON, J. S. (1990). On variance estimation in nonparametric regression. *Biometrika* **77** 415–419. MR1064818

HÄRDLE, W. (1990). *Applied Nonparametric Regression*. Cambridge Univ. Press, Cambridge.

HÄRDLE, W. and TSYBAKOV, A. (1997). Local polynomial estimators of the volatility function in nonparametric autoregression. *J. Econometrics* **81** 223–242.

MÜLLER, H.-G. and STADTMÜLLER, U. (1999). Discontinuous versus smooth regression. *Ann. Statist.* **27** 299–337. MR1701113

MUNK, A., BISSANTZ, N., WAGNER, T. and FREITAG, G. (2005). On difference-based variance estimation in nonparametric regression when the covariate is high dimensional. *J. Roy. Statist. Soc. Ser. B* **67** 19–41. MR2136637

PAIGE, R. L., SUN, S. and WANG, K. (2009). Variance reduction in smoothing splines. *Scand. J. Stat.* **36** 112–126. MR2508334

PARK, C., KIM, I. and LEE, Y. (2012). Error variance estimation via least squares for small sample nonparametric regression. *J. Statist. Plann. Inference* **142** 2369–2385. MR2911851

PENDAKUR, K. and SPERLICH, S. (2010). Semiparametric estimation of consumer demand systems in real expenditure. *J. Appl. Econometrics* **25** 420–457. MR2752121

RICE, J. A. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12** 1215–1230. MR0760684

RUPPERT, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *J. Amer. Statist. Assoc.* **92** 1049–1062. MR1482136

SHEN, H. and BROWN, L. D. (2006). Non-parametric modelling of time-varying customer service times at a bank call centre. *Appl. Stoch. Models Bus. Ind.* **22** 297–311.

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

STEFANSKI, L. A. and COOK, J. R. (1995). Simulation-extrapolation: the measurement error jackknife. *J. Amer. Statist. Assoc.* **90** 1247–1256. MR1379467

TABAKAN, G. and AKDENIZ, F. (2010). Difference-based ridge estimator of parameters in partial linear model. *Statist. Papers* **51** 357–368. MR2665357

TONG, T., MA, Y. and WANG, Y. (2013). Optimal variance estimation without estimating the mean function. *Bernoulli* **19** 1839–1854. MR3129036

TONG, T. and WANG, Y. (2005). Estimating residual variance in nonparametric regression using least squares. *Biometrika* **92** 821–830. MR2234188

WAHBA, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B* **45** 133–150. MR0701084

WANG, Y. (2011). *Smoothing Splines*: *Methods and Applications*. Chapman & Hall, New York.

WANG, L., BROWN, L. D. and CAI, T. (2011). A difference based approach to the semiparametric partial linear model. *Electron. J. Stat.* **5** 619–641. MR2813557

WANG, W. W. and LIN, L. (2015). Derivative estimation based on difference sequence via locally weighted least squares regression. *J. Mach. Learn. Res.* **16** 2617–2641. MR3450519

YE, J. (1998). On measuring and correcting the effects of data mining and model selection. *J. Amer. Statist. Assoc.* **93** 120–131. MR1614596

ZHOU, Y., CHENG, Y., WANG, L. and TONG, T. (2015). Optimal difference-based variance estimation in heteroscedastic nonparametric regression. *Statist. Sinica* **25** 1377–1397. MR3409072

Statistical Science
2017, Vol. 32, No. 3, 469–485
DOI: 10.1214/17-STS616

Models for the Assessment of Treatment Improvement: The Ideal and the Feasible

P. C. Álvarez-Esteban, E. del Barrio, J. A. Cuesta-Albertos and C. Matrán

Abstract. Comparisons of different treatments or production processes are the goals of a significant fraction of applied research. Unsurprisingly, two-sample problems play a main role in statistics through natural questions such as "Is the the new treatment significantly better than the old?" However, this is only partially answered by some of the usual statistical tools for this task. More importantly, often practitioners are not aware of the real meaning behind these statistical procedures. We analyze these troubles from the point of view of the order between distributions, the stochastic order, showing evidence of the limitations of the usual approaches, paying special attention to the classical comparison of means under the normal model. We discuss the unfeasibility of statistically proving stochastic dominance, but show that it is possible, instead, to gather statistical evidence to conclude that slightly relaxed versions of stochastic dominance hold.

Key words and phrases: Stochastic dominance, similarity, two-sample comparison, trimmed distributions, winsorized distributions, Behrens–Fisher problem, index of stochastic dominance.

REFERENCES

ÁLVAREZ-ESTEBAN, P. C., DEL BARRIO, E., CUESTA-ALBERTOS, J. A. and MATRÁN, C. (2008). Trimmed comparison of distributions. *J. Amer. Statist. Assoc.* **103** 697–704. MR2435470

ÁLVAREZ-ESTEBAN, P. C., DEL BARRIO, E., CUESTA-ALBERTOS, J. A. and MATRÁN, C. (2012). Similarity of samples and trimming. *Bernoulli* **18** 606–634. MR2922463

ÁLVAREZ-ESTEBAN, P. C., DEL BARRIO, E., CUESTA-ALBERTOS, J. A. and MATRÁN, C. (2014). A contamination model for approximate stochastic order: Extended version. Available at http://arxiv.org/abs/1412.1920.

ÁLVAREZ-ESTEBAN, P. C., DEL BARRIO, E., CUESTA-ALBERTOS, J. A. and MATRÁN, C. (2016). A contamination model for the stochastic order. *TEST* **25** 751–774. MR3554414

ANDERSON, G. (1996). Nonparametric tests for stochastic dominance. *Econometrica* **64** 1183–1193.

BARRETT, G. F. and DONALD, S. G. (2003). Consistent tests for stochastic dominance. *Econometrica* **71** 71–104. MR1956856

BERGER, R. L. (1988). A nonparametric, intersection-union test for stochastic order. In *Statistical Decision Theory and Related Topics*, *IV*, *Vol*. 2 (*West Lafayette*, *Ind*., 1986) 253–264. Springer, New York. MR0927137

CHEN, G., CHEN, J. and CHEN, Y. (2002). Statistical inference on comparing two distribution functions with a possible crossing point. *Statist. Probab. Lett.* **60** 329–341. MR1942937

DAVIDSON, R. and DUCLOS, J.-Y. (2000). Statistical inference for stochastic dominance and for the measurement of poverty and inequality. *Econometrica* **68** 1435–1464. MR1793365

DAVIDSON, R. and DUCLOS, J.-Y. (2013). Testing for restricted stochastic dominance. *Econometric Rev.* **32** 84–125. MR2988921

DEL BARRIO, E., GINÉ, E. and MATRÁN, C. (1999). Central limit theorems for the Wasserstein distance between the empirical and the true distributions. *Ann. Probab.* **27** 1009–1071. MR1698999

DETTE, H. and MUNK, A. (1998). Validation of linear regression models. *Ann. Statist.* **26** 778–800. MR1626016

P. C. Álvarez-Esteban is Professor, Departamento de Estadística e Investigación Operativa and IMUVA, Universidad de Valladolid, Paseo de Belén, 7. 47011-Valladolid, Spain (e-mail: pedroc@eio.uva.es). E. del Barrio is Professor, Departamento de Estadística e Investigación Operativa and IMUVA, Universidad de Valladolid, Paseo de Belén, 7. 47011-Valladolid, Spain (e-mail: tasio@eio.uva.es). J. A. Cuesta-Albertos is Professor, Departamento de Matemáticas, Estadística y Computación, Universidad de Cantabria, Avda. de los Castros, 48. 39005-Santander, Spain (e-mail: cuestaj@unican.es). C. Matrán is Professor, Departamento de Estadística e Investigación Operativa and IMUVA, Universidad de Valladolid, Paseo de Belén, 7. 47011-Valladolid, Spain (e-mail: carlos.matran@uva.es).

HAWKINS, D. L. and KOCHAR, S. C. (1991). Inference for the crossing point of two continuous CDFs. *Ann. Statist.* **19** 1626–1638. MR1126342

HODGES, J. L. JR. and LEHMANN, E. L. (1954). Testing the approximate validity of statistical hypotheses. *J. Roy. Statist. Soc. Ser. B* **16** 261–268. MR0069461

LEHMANN, E. L. (1955). Ordered families of distributions. *Ann. Math. Stat.* **26** 399–419. MR0071684

LEHMANN, E. L. and ROJO, J. (1992). Invariant directional orderings. *Ann. Statist.* **20** 2100–2110. MR1193328

LESHNO, M. and LEVY, H. (2002). Preferred by "All" and preferred by "Most" decision makers: Almost stochastic dominance. *Manage. Sci.* **48** 1074–1085.

LINTON, O., MAASOUMI, E. and WHANG, Y.-J. (2005). Consistent testing for stochastic dominance under general sampling schemes. *Rev. Econ. Stud.* **72** 735–765. MR2148141

LINTON, O., SONG, K. and WHANG, Y.-J. (2010). An improved bootstrap test of stochastic dominance. *J. Econometrics* **154** 186–202. MR2558960

LIU, J. and LINDSAY, B. G. (2009). Building and using semiparametric tolerance regions for parametric multinomial models. *Ann. Statist.* **37** 3644–3659. MR2549573

MANN, H. B. and WHITNEY, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **18** 50–60. MR0022058

MCFADDEN, D. (1989). Testing for stochastic dominance. In *Studies in the Economics of Uncertainty*: *In Honor of Josef Hadar* (T. B. Fomby and T. K. Seo, eds.) Springer, New York.

MÜLLER, A. and STOYAN, D. (2002). *Comparison Methods for Stochastic Models and Risks*. Wiley, Chichester. MR1889865

MUNK, A. and CZADO, C. (1998). Nonparametric validation of similar distributions and assessment of goodness of fit. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 223–241. MR1625620

RAGHAVACHARI, M. (1973). Limiting distributions of Kolmogorov–Smirnov type statistics under the alternative. *Ann. Statist.* **1** 67–73. MR0346976

RUDAS, T., CLOGG, C. C. and LINDSAY, B. G. (1994). A new index of fit based on mixture methods for the analysis of contingency tables. *J. Roy. Statist. Soc. Ser. B* **56** 623–639. MR1293237

SHAKED, M. and SHANTHIKUMAR, J. G. (2007). *Stochastic Orders*. Springer, New York. MR2265633

TSETLIN, I., WINKLER, R. L., HUANG, R. J. and TZENG, L. Y. (2015). Generalized almost stochastic dominance. *Oper. Res.* **63** 363–377. MR3338587

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York. MR1385671