STATISTICAL SCIENCE

Volume 33, Number 2 May 2018

Special Section on Missing Data

General Section

Statistical Science **[ISSN 0883-4237 (print); ISSN 2168-8745 (online)]**, Volume 33, Number 2, May 2018. Published quarterly by the Institute of Mathematical Statistics, 3163 Somerset Drive, Cleveland, OH 44122, USA. Periodicals postage paid at Cleveland, Ohio and at additional mailing offices.

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

Statistical Science **Volume 33, Number 2 (139–297) May 2018**

Volume 33 Number 2 May 2018

Statistical Science
2018, Vol. 33, No. 2, 139–141
https://doi.org/10.1214/18-STS332IN

Introduction to the Special Section on Missing Data

Julie Josse and Jerome P. Reiter

REFERENCES

BAKER, R., BRICK, J. M., BATES, N. A., BATTAGLIA, M., COUPER, M. P., DEVER, J. A., GILE, K. J. and TOURANGEAU, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology* **1** 90–143.

LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley, Hoboken, NJ. MR1925014

MADDEN, J. M., LAKOMA, M. D., RUSINAK, D., LU, C. Y. and SOUMERAI, S. B. (2016). Missing clinical and behavioral health data in a large electronic health record (ehr) system. *Journal of the American Medical Informatics Association* **23** 1143–1149.

MOHAN, K. and PEARL, J. (2018). Graphical models for processing missing data. Technical Report, Dept. Computer Science, Univ. California, Los Angeles.

NATIONAL RESEARCH COUNCIL (2013). *Nonresponse in Social Science Surveys*. Panel on a research agenda for the future of social science data collection. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. The National Academies Press, Washington, DC.

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

XIE, X. and MENG, X.-L. (2017). Dissecting multiple imputation from a multi-phase inference perspective: What happens when God's, imputer's and analyst's models are uncongenial? *Statist. Sinica* **27** 1485–1545. MR3701490

Julie Josse is Professor of Statistics, Ecole Polytechnique, route de Saclay, 91128 Palaiseau Cedex, France (e-mail: julie.josse@polytechnique.edu). Jerome P. Reiter is Professor of Statistical Science, Department of Statistical Science, Box 90251, Duke University, Durham, North Carolina 27705, USA (e-mail: jreiter@duke.edu).

Statistical Science
2018, Vol. 33, No. 2, 142–159
https://doi.org/10.1214/18-STS644

Multiple Imputation: A Review of Practical and Theoretical Findings

Jared S. Murray

Abstract. Multiple imputation is a straightforward method for handling missing data in a principled fashion. This paper presents an overview of multiple imputation, including important theoretical results and their practical implications for generating and using multiple imputations. A review of strategies for generating imputations follows, including recent developments in flexible joint modeling and sequential regression/chained equations/fully conditional specification approaches. Finally, we compare and contrast different methods for generating imputations on a range of criteria before identifying promising avenues for future research.

Key words and phrases: Missing data, proper imputation, congeniality, chained equations, fully conditional specification, sequential regression multivariate imputation.

REFERENCES

ABAYOMI, K., GELMAN, A. and LEVY, M. (2008). Diagnostics for multivariate imputations. *J. Roy. Statist. Soc. Ser. C* **57** 273–291. MR2440009

AKANDE, O., LI, F. and REITER, J. (2017). An empirical comparison of multiple imputation methods for categorical data. *Amer. Statist.* **71** 162–170. MR3668704

ANDRIDGE, R. R. and LITTLE, R. J. A. (2010). A review of hot deck imputation for survey non-response. *Int. Stat. Rev.* **78** 40–64.

ARNOLD, B. C., CASTILLO, E. and SARABIA, J. M. (2001). Conditionally specified distributions: An introduction. *Statist. Sci.* **16** 249–274. MR1874154

ARNOLD, B. C. and PRESS, J. S. (1989). Compatible conditional distributions. *J. Amer. Statist. Assoc.* **84** 152–156. MR0999673

AUDIGIER, V., HUSSON, F. and JOSSE, J. (2016). Multiple imputation for continuous variables using a Bayesian principal component analysis. *J. Stat. Comput. Simul.* **86** 2140–2156. MR3491013

AUDIGIER, V., HUSSON, F. and JOSSE, J. (2017). MIMCA: Multiple imputation for categorical variables with multiple correspondence analysis. *Stat. Comput.* **27** 501–518. MR3599686

BANERJEE, A., MURRAY, J. and DUNSON, D. B. (2013). Bayesian learning of joint distributions of objects. In *Proceedings of the* 16*th International Conference on Artificial Intelligence and Statistics* (*AISTATS*), Scottsdale, AZ.

BARNARD, J. and RUBIN, D. B. (1999). Miscellanea. Small-sample degrees of freedom with multiple imputation. *Biometrika* **86** 948–955.

BERNAARDS, C. A., BELIN, T. R. and SCHAFER, J. L. (2007). Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Stat. Med.* **26** 1368–1382. MR2345726

BLACKWELL, M., HONAKER, J. and KING, G. (2015). A unified approach to measurement error and missing data. *Sociol. Methods Res.* **46** 303–341.

BÖHNING, D., SEIDEL, W., ALFÓ, M., GAREL, B., PATILEA, V., WALTHER, G., DI ZIO, M., GUARNERA, U. and LUZI, O. (2007). Imputation through finite Gaussian mixture models. *Comput. Statist. Data Anal.* **51** 5305–5316. MR2370885

BONDARENKO, I. and RAGHUNATHAN, T. (2016). Graphical and numerical diagnostic tools to assess suitability of multiple imputations and imputation models. *Stat. Med.* **35** 3007–3020. MR3528239

BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.

BREIMAN, L., FRIEDMAN, J., STONE, C. J. and OLSHEN, R. A. (1984). *Classification and Regression Trees*. Wadsworth Advanced Books and Software, Belmont, CA. MR0726392

BURGETTE, L. F. and REITER, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *Am. J. Epidemiol.* **172** 1070–1076.

CARPENTER, J. and KENWARD, M. (2013). *Multiple Imputation and Its Application*, 1st ed. Wiley, New York.

CHEN, J. and SHAO, J. (2000). Nearest neighbor imputation for survey data. *J. Off. Stat.* **16** 113–131.

COLE, S. R., CHU, H. and GREENLAND, S. (2006). Multiple-imputation for measurement-error correction. *Int. J. Epidemiol.* **35** 1074–1081.

Jared S. Murray is Assistant Professor of Statistics, Department of Information, Risk, and Operations Management, and Department of Statistics and Data Science, University of Texas at Austin, 12110 Speedway B6500, Austin, Texas 78712, USA (e-mail: jared.murray@mccombs.utexas.edu).

COLLINS, L. M., SCHAFER, J. L. and KAM, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol. Methods* **6** 330–351.

DEYOREO, M., REITER, J. P. and HILLYGUS, D. S. (2017). Bayesian mixture models with focused clustering for mixed ordinal and nominal data. *Bayesian Anal.* **12** 679–730. MR3655872

DOOVE, L. L., VAN BUUREN, S. and DUSSELDORP, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Comput. Statist. Data Anal.* **72** 92–104. MR3139350

DRECHSLER, J. (2010). Multiple imputation of missing values in the wave 2007 of the IAB Establishment Panel. IAB Discussion Paper.

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

ELLIOTT, M. R. and STETTLER, N. (2007). Using a mixture model for multiple imputation in the presence of outliers: The "healthy for life" project. *J. Roy. Statist. Soc. Ser. C* **56** 63–78. MR2339163

FITHIAN, W. and JOSSE, J. (2017). Multiple correspondence analysis and the multilogit bilinear model. *J. Multivariate Anal.* **157** 87–102. MR3641738

FOSDICK, B. K., DEYOREO, M. and REITER, J. P. (2016). Categorical data fusion using auxiliary information. *Ann. Appl. Stat.* **10** 1907–1929. MR3592042

GEBREGZIABHER, M. and DESANTIS, S. M. (2010). Latent class based multiple imputation approach for missing categorical data. *J. Statist. Plann. Inference* **140** 3252–3262. MR2659852

GELMAN, A., CARLIN, J. B., RUBIN, D. B., VEHTARI, A., DUNSON, D. B. and STERN, H. S. (2014). *Bayesian Data Analysis*, 3rd ed. CRC Press, Boca Raton, FL. MR3235677

HE, Y. and ZASLAVSKY, A. M. (2012). Diagnosing imputation models by applying target analyses to posterior replicates of completed data. *Stat. Med.* **31** 1–18. MR2868986

HE, Y., ZASLAVSKY, A. M., LANDRUM, M. B., HARRINGTON, D. P. and CATALANO, P. (2010). Multiple imputation in a large-scale complex survey: A practical guide. *Stat. Methods Med. Res.* **19** 653–670. MR2744515

HEITJAN, D. F. and LITTLE, R. J. A. (1991). Multiple imputation for the fatal accident reporting system. *J. Roy. Statist. Soc. Ser. C* **40** 13–29.

HORTON, N. J., LIPSITZ, S. R. and PARZEN, M. (2003). A potential for bias when rounding in multiple imputation. *Amer. Statist.* **57** 229–232. MR2016255

HU, J., REITER, J. P. and WANG, Q. (2017). Dirichlet process mixture models for modeling and generating synthetic versions of nested categorical data. *Bayesian Anal.* **12** 679–703. MR3737948

HUGHES, R. A., WHITE, I. R., SEAMAN, S. R., CARPENTER, J. R., TILLING, K. and STERNE, J. A. C. (2014). Joint modelling rationale for chained equations. *BMC Med. Res. Methodol.* **14** 28.

IBRAHIM, J. G., LIPSITZ, S. R. and CHEN, M. H. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **61** 173–190. MR1664045

IBRAHIM, J. G., CHEN, M. H., LIPSITZ, S. R. and HERRING, A. H. (2005). Missing data methods for generalized linear models: A comparative review. *J. Amer. Statist. Assoc.* **100** 332–346. MR2166072

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

KIM, J. K. (2002). A note on approximate Bayesian bootstrap imputation. *Biometrika* **89** 470–477. MR1913974 MR1913974

KIM, J. K., BRICK, J. M., FULLER, W. A. and KALTON, G. (2006). On the bias of the multiple-imputation variance estimator in survey sampling. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 509–521. MR2278338

KIM, H. J., REITER, J. P., WANG, Q., COX, L. H. and KARR, A. F. (2014). Multiple imputation of missing or faulty values under linear constraints. *J. Bus. Econom. Statist.* **32** 375–386. MR3238592

KIM, H. J., COX, L. H., KARR, A. F., REITER, J. P. and WANG, Q. (2015). Simultaneous edit-imputation for continuous microdata. *J. Amer. Statist. Assoc.* **110** 987–999. MR3420678

KROPKO, J., GOODRICH, B., GELMAN, A. and HILL, J. (2014). Multiple imputation for continuous and categorical data: Comparing joint multivariate normal and conditional approaches. *Polit. Anal.* **22** 497–519.

LEE, M. C. and MITRA, R. (2016). Multiply imputing missing values in data sets with mixed measurement scales using a sequence of generalised linear models. *Comput. Statist. Data Anal.* **95** 24–38. MR3425936

LI, F., YU, Y. and RUBIN, D. B. (2012). Imputing missing data by fully conditional models: Some cautionary examples and guidelines. Dept. Statistical Science, Duke Univ., Durham, NC.

LI, F., BACCINI, M., MEALLI, F., ZELL, E. R., FRANGAKIS, C. E. and RUBIN, D. B. (2014). Multiple imputation by ordered monotone blocks with application to the anthrax vaccine research program. *J. Comput. Graph. Statist.* **23** 877–892. MR3224660

LIPSITZ, S. R. and IBRAHIM, J. G. (1996). A conditional model for incomplete covariates in parametric regression models. *Biometrika* **83** 916–922.

LITTLE, R. J. A. (1988). Missing-data adjustments in large surveys. *J. Bus. Econom. Statist.* **6** 287–296.

LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley-Interscience, Hoboken, NJ. MR1925014

LITTLE, R. J. A. and SCHLUCHTER, M. D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika* **72** 497–512. MR0817564

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

LIU, C. and RUBIN, D. B. (1998). Ellipsoidally symmetric extensions of the general location model for mixed categorical and continuous data. *Biometrika* **85** 673–688. MR1665830

LIU, J., GELMAN, A., HILL, J., SU, Y.-S. and KROPKO, J. (2014). On the stationary distribution of iterative imputations. *Biometrika* **101** 155–173. MR3180663

MANRIQUE-VALLIER, D. and REITER, J. P. (2014a). Bayesian estimation of discrete multivariate latent structure models with structural zeros. *J. Comput. Graph. Statist.* **23** 1061–1079. MR3270711

MANRIQUE-VALLIER, D. and REITER, J. P. (2014b). Bayesian multiple imputation for large-scale categorical data with structural zeros. *Surv. Methodol.* **40** 125–134.

MANRIQUE-VALLIER, D. and REITER, J. P. (2016). Bayesian simultaneous edit and imputation for multivariate categorical data. *J. Amer. Statist. Assoc.* **112** 1708–1719. MR3750893

MENG, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statist. Sci.* **9** 538–558.

MENG, X.-L. and ROMERO, M. (2003). Discussion: Efficiency and self-efficiency with multiple imputation inference. *Int. Stat. Rev.* **71** 607–618.

MORRIS, T. P., WHITE, I. R. and ROYSTON, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Med. Res. Methodol.* **14** 75.

MURRAY, J. S. and REITER, J. P. (2016). Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence. *J. Amer. Statist. Assoc.* **111** 1466–1479. MR3601702

NGUYEN, C. D., LEE, K. J. and CARLIN, J. B. (2015). Posterior predictive checking of multiple imputation models. *Biom. J.* **57** 676–694.

NIELSEN, S. F. (2003). Proper and improper multiple imputation. *Int. Stat. Rev.* **71** 593–607.

OLKIN, I. and TATE, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Ann. Math. Stat.* **32** 448–465.

PADDOCK, S. M. (2002). Bayesian nonparametric multiple imputation of partially observed data with ignorable nonresponse. *Biometrika* **89** 529–538. MR1929160

RAGHUNATHAN, T. E., REITER, J. P. and RUBIN, D. B. (2003). Multiple imputation for statistical disclosure limitation. *J. Off. Stat.* **19** 1–16.

RAGHUNATHAN, T. E., LEPKOWSKI, J. M., VAN HOEWYK, J. and SOLENBERGER, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv. Methodol.* **27** 85–96.

RÄSSLER, S. (2004). Data fusion: Identification problems, validity, and multiple imputation. *Aust. J. Stat.* **33** 153–171.

REITER, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *J. Off. Stat.* **18** 531.

REITER, J. P. (2005). Using CART to generate partially synthetic public use microdata. *J. Off. Stat.* **21** 441.

REITER, J. P. (2012). Bayesian finite population imputation for data fusion. *Statist. Sinica* **22** 795–811. MR2954362

REITER, J. (2017). Discussion: Dissecting multiple imputation from a multi-phase inference perspective: What happens when God's, imputer's and analyst's models are uncongenial? *Statist. Sinica*. **27** 1578–1583. MR3701498

REITER, J. P. and RAGHUNATHAN, T. E. (2007). The multiple adaptations of multiple imputation. *J. Amer. Statist. Assoc.* **102** 1462–1471. MR2372542

REITER, J. P., RAGHUNATHAN, T. E. and KINNEY, S. K. (2006). The importance of modeling the sampling design in multiple imputation for missing data. *Surv. Methodol.* **32** 143.

ROBINS, J. M. and WANG, N. (2000). Inference for imputation estimators. *Biometrika* **87** 113–124. MR1766832

ROUSSEAU, J. (2016). On the frequentist properties of Bayesian nonparametric methods. *Annual Review of Statistics and Its Application* **3** 211–231.

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

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley. New York. MR0899519

RUBIN, D. B. (1993). Discussion: Statistical disclosure limitation. *J. Off. Stat.* **9** 461–468.

RUBIN, D. B. (1996). Multiple imputation after 18+ years. *J. Amer. Statist. Assoc.* **91** 473–489.

RUBIN, D. B. (2003a). Discussion on multiple imputation. *Int. Stat. Rev.* **71** 619–625.

RUBIN, D. B. (2003b). Nested multiple imputation of NMES via partially incompatible MCMC. *Stat. Neerl.* **57** 3–18. MR2055518

RUBIN, D. B. and SCHAFER, J. L. (1990). Efficiently creating multiple imputations for incomplete multivariate normal data. In *Proc. Statistical Computing Section of the American Statistical Association* 83–88. Amer. Statist. Assoc., Alexandria, VA.

RUBIN, D. B. and SCHENKER, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *J. Amer. Statist. Assoc.* **81** 366–374. MR0845877

SCHAFER, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London. MR1692799

SCHAFER, J. L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Stat. Neerl.* **57** 19–35. MR2055519

SCHENKER, N. and TAYLOR, J. M. G. (1996). Partially parametric techniques for multiple imputation. *Comput. Statist. Data Anal.* **22** 425–446.

SCHIFELING, T. A. and REITER, J. P. (2016). Incorporating marginal prior information in latent class models. *Bayesian Anal.* **11** 499–518. MR3472000

SEAMAN, S. R. and HUGHES, R. A. (2016). Relative efficiency of joint-model and full-conditional-specification multiple imputation when conditional models are compatible: The general location model. *Stat. Methods Med. Res.* DOI:10.1177/0962280216665872.

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

SHAH, A. D., BARTLETT, J. W., CARPENTER, J., NICHOLAS, O. and HEMINGWAY, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. *Am. J. Epidemiol.* **179** 764–774.

SI, Y. and REITER, J. P. (2013). Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *J. Educ. Behav. Stat.* **38** 499–521.

STUART, E. A., AZUR, M., FRANGAKIS, C. and LEAF, P. (2009). Multiple imputation with large data sets: A case study of the children's mental health initiative. *Am. J. Epidemiol.* **169** 1133–1139.

SU, Y.-S., GELMAN, A., HILL, J., YAJIMA, M. et al. (2011). Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *J. Stat. Softw.* **45** 1–31.

VAN BUUREN, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Stat. Methods Med. Res.* **16** 219–42.

VAN BUUREN, S. (2012). *Flexible Imputation of Missing Data*. CRC Press, Boca Raton, FL.

VAN BUUREN, S. and GROOTHUIS-OUDSHOORN, K. (2011). mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **45** 1–67.

VAN BUUREN, S. and OUDSHOORN, K. (1999). *Flexible Multivariate Imputation by MICE*. TNO Prevention Center, Leiden, The Netherlands.

VAN BUUREN, S., BRAND, J. P. L., GROOTHUIS-OUDSHOORN, C. G. M. and RUBIN, D. B. (2006). Fully conditional specification in multivariate imputation. *J. Stat. Comput. Simul.* **76** 1049–1064. MR2307507

VERMUNT, J. K., VAN GINKEL, J. R., VAN DER ARK, L. A. and SIJTSMA, K. (2008). Multiple imputation of incomplete categorial data using latent class analysis. *Sociol. Method.* **38** 369–397.

VIDOTTO, D., VERMUNT, J. K. and KAPTEIN, M. C. (2015). Multiple imputation of missing categorical data using latent class models: State of art. *Psychol. Test Assess. Model.* **57** 542–576.

VINK, G., FRANK, L. E., PANNEKOEK, J. and VAN BUUREN, S. (2014). Predictive mean matching imputation of semicontinuous variables. *Stat. Neerl.* **68** 61–90. DOI:10.1111/stan.12023. MR3168318

WANG, N. and ROBINS, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika* **85** 935–948. MR1666715

XIE, X. and MENG, X.-L. (2017). Dissecting multiple imputation from a multi-phase inference perspective: What happens when God's, imputer's and analyst's models are uncongenial? *Statist. Sinica.* **27** 1485–1545. MR3701490

XU, D., DANIELS, M. J. and WINTERSTEIN, A. G. (2016). Sequential BART for imputation of missing covariates. *Biostatistics* **17** 589–602. MR3603956

ZHU, J. and RAGHUNATHAN, T. E. (2015). Convergence properties of a sequential regression multiple imputation algorithm. *J. Amer. Statist. Assoc.* **110** 1112–1124. MR3420688

Statistical Science
2018, Vol. 33, No. 2, 160–183
https://doi.org/10.1214/18-STS646

Multiple Imputation for Multilevel Data with Continuous and Binary Variables

Vincent Audigier, Ian R. White, Shahab Jolani, Thomas P. A. Debray, Matteo Quartagno, James Carpenter, Stef van Buuren and Matthieu Resche-Rigon

Abstract. We present and compare multiple imputation methods for multilevel continuous and binary data where variables are systematically and sporadically missing. The methods are compared from a theoretical point of view and through an extensive simulation study motivated by a real dataset comprising multiple studies. The comparisons show that these multiple imputation methods are the most appropriate to handle missing values in a multilevel setting and why their relative performances can vary according to the missing data pattern, the multilevel structure and the type of missing variables. This study shows that valid inferences can only be obtained if the dataset includes a large number of clusters. In addition, it highlights that heteroscedastic multiple imputation methods provide more accurate inferences than homoscedastic methods, which should be reserved for data with few individuals per cluster. Finally, guidelines are given to choose the most suitable multiple imputation method according to the structure of the data.

Key words and phrases: Missing data, systematically missing values, multilevel data, mixed data, multiple imputation, joint modelling, fully conditional specification.

REFERENCES

ALBERT, A. and ANDERSON, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71** 1–10. MR0738319

ALLISON, P. (2002). *Missing Data*. Sage, Thousand Oaks, CA.

ANDRIDGE, R. R. (2011). Quantifying the impact of fixed effects modeling of clusters in multiple imputation for cluster randomized trials. *Biom. J.* **53** 57–74. MR2767378

ASPAROUHOV, T. and MUTHÉN, B. (2010). Multiple imputation with Mplus. Technical report. Available at http://www.statmodel.com/download/Imputations7.pdf.

AUDIGIER, V. and RESCHE-RIGON, M. (2017). micemd: Multiple imputation by chained equations with multilevel data. R package version 1.2.0.

Vincent Audigier is Associate Professor, CNAM, Cedric MSDMA, Paris, France (e-mail: vincent.audigier@cnam.fr). Ian R. White is Professor, MRC Biostatistics Unit, Cambridge Institute of Public Health, United Kingdom; MRC Clinical Trials Unit at UCL, London, United Kingdom (e-mail: ian.white@ucl.ac.uk). Shahab Jolani is Assistant Professor, Department of Methodology and Statistics, School CAPHRI, Care and Public Health Research Institute, Maastricht University, The Netherlands (e-mail: s.jolani@maastrichtuniversity.nl). Thomas P. A. Debray is Assistant Professor, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands; Cochrane Netherlands, University Medical Center Utrecht, Utrecht, The Netherlands (e-mail: T.Debray@umcutrecht.nl). Matteo Quartagno is Research Fellow, Department of Medical Statistics, London School of Hygiene & Tropical Medicine, London, United Kingdom; MRC Clinical Trials Unit at UCL, London, United Kingdom (e-mail: Matteo.Quartagno@lshtm.ac.uk). James Carpenter is Professor, Department of Medical Statistics, London School of Hygiene & Tropical Medicine, London, United Kingdom; MRC Clinical Trials Unit at UCL, London, United Kingdom (e-mail: james.carpenter@lshtm.ac.uk). Stef van Buuren is Professor, Department of Methodology & Statistics, FSS, University of Utrecht, The Netherlands; Netherlands Organisation for Applied Scientific Research TNO, Leiden, The Netherlands (e-mail: S.vanBuuren@uu.nl). Matthieu Resche-Rigon is Professor, Service de Biostatistique et Information Médicale, Hôpital Saint-Louis, AP-HP, Paris, France; Université Paris Diderot—Paris 7, Sorbonne Paris Cité, UMR-S 1153, Paris, France; INSERM, UMR 1153, Equipe ECSTRA, Hôpital Saint-Louis, Paris, France (e-mail: matthieu.resche-rigon@univ-paris-diderot.fr).

AUDIGIER, V., WHITE, I. R., JOLANI, S., DEBRAY, T. P. A., QUARTAGNO, M., CARPENTER, J., VAN BUUREN, S. and RESCHE-RIGON, M. (2018). Supplement to "Multiple imputation for multilevel data with continuous and binary variables." DOI:10.1214/18-STS646SUPPA, DOI:10.1214/18-STS646SUPPB.

BARTLETT, J. W., SEAMAN, S. R., WHITE, I. R. and CARPENTER, J. R. (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Stat. Methods Med. Res.* **24** 462–487. MR3372102

BATES, D., MÄCHLER, M., BOLKER, B. and WALKER, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67** 1–48.

BLOSSFELD, H.-P., GÜNTHER ROŠBACH, H. and VON MAURICE, J., eds. (2011). *Education as a Lifelong Process*: *The German National Educational Panel Study* (*NEPS*). VS Verlag für Sozialwissenschaften, Wiesbaden, Germany.

BOS, W., LANKES, E.-M., PRENZEL, M., SCHWIPPERT, K. and VALTIN, R., eds. (2003). *Erste Ergebnisse aus IGLU*: *Schülerleistungen Am Ende der Vierten Jahrgangsstufe Im Internationalen Vergleich* [*the First*]. Waxmann, Münster, Germany.

CARPENTER, J. and KENWARD, M. (2013). *Multiple Imputation and Its Application*, 1st ed. Wiley, New York.

CARRIG, M. M., MANRIQUE-VALLIER, D., RANBY, K. W., REITER, J. and HOYLE, R. H. (2015). A nonparametric, multiple imputation-based method for the retrospective integration of data sets. *Multivar. Behav. Res.* **50** 383–397.

CURRAN, P. J. and HUSSONG, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychol. Methods* **14** 81–100.

CURRAN, P. J., HUSSONG, A. M., CAI, L., HUANG, W., CHASSIN, L., SHER, K. J. and ZUCKER, R. A. (2008). Pooling data from multiple longitudinal studies: The role of item response theory in integrative data analysis. *Dev. Psychol.* **44** 365–380.

DEBRAY, T., RILEY, R., ROVERS, M., REITSMA, J., MOONS, K. and ON BEHALF OF THE COCHRANE IPD META-ANALYSIS METHODS GROUP (2015b). Individual participant data (IPD) meta-analyses of diagnostic and prognostic modeling studies: Guidance on their use. *PLoS Med.* **12** e1001886.

DEBRAY, T., MOONS, K., VAN VALKENHOEF, G., EFTHIMIOU, O., HUMMEL, N., GROENWOLD, R. and REITSMA, J. O. (2015a). Get real in individual participant data (IPD) meta-analysis: A review of the methodology. *Res. Synth. Methods* **6** 293–309.

DERSIMONIAN, R. and LAIRD, N. (1986). Meta-analysis in clinical trials. *Control. Clin. Trials* **7** 177–188.

DRECHSLER, J. (2015). Multiple imputation of multilevel missing data—rigor versus simplicity. *J. Educ. Behav. Stat.* **40** 69–95.

ENDERS, C. (2010). *Applied Missing Data Analysis*. Guilford Press, New York.

ENDERS, C. K., KELLER, B. T. and LEVY, R. (2017). A fully conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychol. Methods*.

ENDERS, C., MISTLER, S. and KELLER, B. (2016). Multilevel multiple imputation: A review and evaluation of joint modeling and chained equations imputation. *Psychological Methods* **21** 222–240.

ERLER, N. S., RIZOPOULOS, D., VAN ROSMALEN, J., JADDOE, V. W. V., FRANCO, O. H. and LESAFFRE, E. M. E. H. (2016). Dealing with missing covariates in epidemiologic studies: A comparison between multiple imputation and a full Bayesian approach. *Stat. Med.* **35** 2955–2974. MR3528236

FIRTH, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80** 27–38. MR1225212

GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1** 515–533. MR2221284

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

GLOBAL RESEARCH ON ACUTE CONDITIONS TEAM (GREAT) NETWORK (2013). Managing acute heart failure in the ED—case studies from the acute heart failure academy. Available at http://www.greatnetwork.org.

GOLDSTEIN, H., BONNET, G. and ROCHER, T. (2007). Multilevel structural equation models for the analysis of comparative data on educational performance. *J. Educ. Behav. Stat.* **32** 252–286.

GOLDSTEIN, H., CARPENTER, J., KENWARD, M. G. and LEVIN, K. A. (2009). Multilevel models with multivariate mixed response types. *Stat. Model.* **9** 173–197. MR2756416

GRAHAM, J. W. (2012). *Missing Data*: *Analysis and Design*. Springer, New York. MR2952499

GRUND, S., LÜDTKE, O. and ROBITZSCH, A. (2016). Multiple imputation of missing covariate values in multilevel models with random slopes: A cautionary note. *Behav. Res. Methods* **48** 640–649.

HUGHES, R. A., WHITE, I. R., SEAMAN, S., CARPENTER, J., TILLING, K. and STERNE, J. (2014). Joint modelling rationale for chained equations. *BMC Med. Res. Methodol.* **14** 28.

JACKSON, D., WHITE, I. R. and RILEY, R. D. (2013). A matrix-based method of moments for fitting the multivariate random effects model for meta-analysis and meta-regression. *Biom. J.* **55** 231–245. MR3045843

JOLANI, S. (2018). Hierarchical imputation of systematically and sporadically missing data: An approximate Bayesian approach using chained equations. *Biom. J.* **60** 333–351.

JOLANI, S., DEBRAY, T. P. A., KOFFIJBERG, H., VAN BUUREN, S. and MOONS, K. G. M. (2015). Imputation of systematically missing predictors in an individual participant data meta-analysis: A generalized approach using MICE. *Stat. Med.* **34** 1841–1863. MR3334696

KROPKO, J., GOODRICH, B., GELMAN, A. and HILL, J. (2014). Multiple imputation for continuous and categorical data: Comparing joint multivariate normal and conditional approaches. *Polit. Anal.* **22** 497–519.

KUNKEL, D. and KAIZAR, E. E. (2017). A comparison of existing methods for multiple imputation in individual participant data meta-analysis. *Stat. Med.* **36** 3507–3532. MR3696506

LANGAN, D., HIGGINS, J. P. T. and SIMMONDS, M. (2017). Comparative performance of heterogeneity variance estimators in meta-analysis: A review of simulation studies. *Res. Synth. Methods* **8** 181–198.

LASSUS, J., GAYAT, E., MUELLER, C., PEACOCK, W., SPINAR, J., HARJOLA, V., VAN KIMMENADE, R., PATHAK, A., MUELLER, T. et al. (2013). Incremental value of biomarkers to clinical variables for mortality prediction in acutely decompensated heart failure: The multinational observational cohort on acute heart failure (MOCA) study. *Int. J. Cardiol.* **168** 2186–2194.

LEE, K. and CARLIN, J. (2010). Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. *Am. J. Epidemiol.* **171** 624–632.

LEE, Y., NELDER, J. A. and PAWITAN, Y. (2006). *Generalized Linear Models with Random Effects*: *Unified Analysis via H-Likelihood*. *Monographs on Statistics and Applied Probability* **106**. Chapman & Hall/CRC, Boca Raton, FL. With 1 CD-ROM (Windows). MR2259540

LITTLE, R. (1988). Missing-data adjustments in large surveys. *J. Bus. Econom. Statist.* **6** 287–296.

LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley-Interscience, Hoboken, NJ. MR1925014

LIU, J., GELMAN, A., HILL, J., SU, Y.-S. and KROPKO, J. (2014). On the stationary distribution of iterative imputations. *Biometrika* **101** 155–173. MR3180663

LONGFORD, N. T. (2008). Missing data. In *Handbook of Multilevel Analysis* 377–399. Springer, New York. MR2412943

MATHEW, T. and NORDSTRÖM, K. (2010). Comparison of one-step and two-step meta-analysis models using individual patient data. *Biom. J.* **52** 271–287. MR2756877

MCNEISH, D. and STAPLETON, L. M. (2016). Modeling clustered data with very few clusters. *Multivar. Behav. Res.* **51** 495–518.

MEBAZAA, A., GAYAT, E., LASSUS, J., MEAS, T., MUELLER, C. et al. (2013). Association between elevated blood glucose and outcome in acute heart failure: Results from an international observational cohort. *J. Am. Coll. Cardiol.* **61** 820–829.

MENG, X. (1994). Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statist. Sci.* **10** 538–573.

MORRIS, T. P., WHITE, I. R. and CROWTHER, M. J. (2017). Using simulation studies to evaluate statistical methods. ArXiv e-prints.

MULLIS, I., MARTIN, M., GONZALEZ, E. and KENNEDY, A. (2003). Pirls 2001 international report: Iea's study of reading literacy achievement in primary school in 35 countries. Available at: https://timssandpirls.bc.edu/pirls2001i/pdf/p1_IR_book.pdf.

NOH, M. and LEE, Y. (2007). REML estimation for binary data in GLMMs. *J. Multivariate Anal.* **98** 896–915. MR2325413

PINHEIRO, J. and BATES, D. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer, New York.

PINHEIRO, J., BATES, D., DEBROY, S. and SARKAR, D. (2016). nlme: Linear and nonlinear mixed effects models. R package version 3.1-128.

QUARTAGNO, M. and CARPENTER, J. R. (2016a). Multiple imputation for IPD meta-analysis: Allowing for heterogeneity and studies with missing covariates. *Stat. Med.* **35** 2938–2954. MR3528235

QUARTAGNO, M. and CARPENTER, J. (2016b). jomo: A package for multilevel joint modelling multiple imputation. R package version 2.2-0.

RAGHUNATHAN, T., LEPKOWSKI, J. M., VAN HOEWYK, J. and SOLENBERGER, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv. Methodol.* **27** 85–96.

REITER, J., RAGHUNATHAN, T. E. and KINNEY, S. K. (2006). The importance of modeling the sampling design in multiple imputation for missing data. *Surv. Methodol.* **32** 143.

RESCHE-RIGON, M. and WHITE, I. (2016). Multiple imputation by chained equations for systematically and sporadically missing multilevel data. *Stat. Methods Med. Res.* DOI:10.1177/0962280216666564.

RESCHE-RIGON, M., WHITE, I. R., BARTLETT, J. W., PETERS, S. A. E., THOMPSON, S. G. and GROUP, P. S. (2013). Multiple imputation for handling systematically missing confounders in meta-analysis of individual participant data. *Stat. Med.* **32** 4890–4905. MR3127183

RILEY, R. D., LAMBERT, P. C., STAESSEN, J. A., WANG, J., GUEYFFIER, F., THIJS, L. and BOUTITIE, F. (2008). Meta-analysis of continuous outcomes combining individual patient data and aggregate data. *Stat. Med.* **27** 1870–1893. MR2420350

RILEY, R. D., ENSOR, J., SNELL, K. I. E., DEBRAY, T. P. A., ALTMAN, D. G., MOONS, K. G. M. and COLLINS, G. S. (2016). External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: Opportunities and challenges. *BMJ* **353** i3140.

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

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

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York. MR0899519

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

SCHAFER, J. L. and YUCEL, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *J. Comput. Graph. Statist.* **11** 437–457. MR1938143

SIMMONDS, M., HIGGINS, J., STEWART, L., TIERNEY, J., CLARKE, M. and THOMPSON, S. (2005). Meta-analysis of individual patient data from randomized trials: A review of methods used in practice. *Clin. Trials* **2** 209–217.

TANNER, M. A. and WONG, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82** 528–550. MR0898357

R CORE TEAM (2016). *R*: *A Language and Environment for Statistical Computing*. *Version* 3.3.0. R Foundation for Statistical Computing, Vienna, Austria.

VAN BUUREN, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Stat. Methods Med. Res.* **16** 219–242. MR2371007

VAN BUUREN, S. (2011). Multiple imputation of multilevel data. In *The Handbook of Advanced Multilevel Analysis* (J. J. Hox, ed.) 173–196. Routledge, New York.

VAN BUUREN, S. (2012). *Flexible Imputation of Missing Data* (*Chapman & Hall/CRC Interdisciplinary Statistics*). Chapman & Hall/CRC, London.

VAN BUUREN, S. and GROOTHUIS-OUDSHOORN, K. (2011). mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **45** 1–67.

VAN BUUREN, S., BRAND, J. P. L., GROOTHUIS-OUDSHOORN, C. G. M. and RUBIN, D. B. (2006). Fully conditional specification in multivariate imputation. *J. Stat. Comput. Simul.* **76** 1049–1064. MR2307507

VINK, G., LAZENDIC, G. and VAN BUUREN, S. (2015). Partitioned predictive mean matching as a multilevel imputation technique. *Psychol. Test Assess. Model.* **57** 577–594.

WAGSTAFF, D. and HAREL, O. (2011). A closer examination of three small-sample approximations to the multiple-imputation degrees of freedom. *Stata J.* **11** 403–419.

YUCEL, R. M. (2011). Random covariances and mixed-effects models for imputing multivariate multilevel continuous data. *Stat. Model.* **11** 351–370. MR2906705

ZHAO, Y. and LONG, Q. (2016). Multiple imputation in the presence of high-dimensional data. *Stat. Methods Med. Res.* **25** 2021–2035. MR3553324

ZHAO, E. and YUCEL, R. (2009). Performance of sequential imputation method in multilevel applications. In *Proceedings of the Survey Research Methods Section* (*JSM* 2009) 2800–2810. Amer. Statist. Assoc., Alexandria, VA.

ZHU, J. and RAGHUNATHAN, T. E. (2015). Convergence properties of a sequential regression multiple imputation algorithm. *J. Amer. Statist. Assoc.* **110** 1112–1124. MR3420688

Statistical Science
2018, Vol. 33, No. 2, 184–197
https://doi.org/10.1214/18-STS647

Introduction to Double Robust Methods for Incomplete Data

Shaun R. Seaman and Stijn Vansteelandt

Abstract. Most methods for handling incomplete data can be broadly classified as inverse probability weighting (IPW) strategies or imputation strategies. The former model the occurrence of incomplete data; the latter, the distribution of the missing variables given observed variables in each missingness pattern. Imputation strategies are typically more efficient, but they can involve extrapolation, which is difficult to diagnose and can lead to large bias. Double robust (DR) methods combine the two approaches. They are typically more efficient than IPW and more robust to model misspecification than imputation. We give a formal introduction to DR estimation of the mean of a partially observed variable, before moving to more general incomplete-data scenarios. We review strategies to improve the performance of DR estimators under model misspecification, reveal connections between DR estimators for incomplete data and "design-consistent" estimators used in sample surveys, and explain the value of double robustness when using flexible data-adaptive methods for IPW or imputation.

Key words and phrases: Augmented inverse probability weighting, calibration estimators, data-adaptive methods, doubly robust, empirical likelihood, imputation, inverse probability weighting, missing data, semiparametric methods.

REFERENCES

[1] BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–972. MR2216189

[2] BELLONI, A. and CHERNOZHUKOV, V. (2011). l_1-Penalized quantile regression in high-dimensional sparse models. *Ann. Statist.* **39** 82–130. MR2797841

[3] BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2016). Lasso methods for Gaussian instrumental variables models. Preprint. Available at arXiv:1012.1297.

[4] BROOKHART, M. A. and VAN DER LAAN, M. J. (2006). A semiparametric model selection criterion with applications to the marginal structural model. *Comput. Statist. Data Anal.* **50** 475–498. MR2201874

[5] CAO, W., TSIATIS, A. A. and DAVIDIAN, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* **96** 723–734. MR2538768

[6] CASSEL, C. M., SARNDAL, C. E. and WRETMAN, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika* **63** 615–620. MR0445666

[7] CHENG, G., YU, Z. and HUANG, J. Z. (2013). The cluster bootstrap consistency in generalized estimating equations. *J. Multivariate Anal.* **115** 33–47. MR3004543

[8] CHERNOZHUKOV, V., ESCANCIANO, J. C., ICHIMURA, H. and NEWEY, W. K. (2016). Locally robust semiparametric estimation. Preprint. Available at arXiv:1608.00033.

[9] FARRELL, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *J. Econometrics* **189** 1–23.

[10] GRUBER, S. and VAN DER LAAN, M. J. (2010). A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *Int. J. Biostat.* **6** Article 26.

[11] HORVITZ, D. G. and THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite

Shaun R. Seaman is Senior Research Associate, Medical Research Council Biostatistics Unit, University of Cambridge, Cambridge CB2 0SR, United Kingdom (e-mail: shaun.seaman@mrc-bsu.cam.ac.uk). Stijn Vansteelandt is Professor of Statistics, Department of Applied Mathematics, Computer Science and Statistics, Ghent University, 9000 Gent, Belgium, and Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London WC1E 7HT, United Kingdom.

universe. *J. Amer. Statist. Assoc.* **47** 663–685. MR0053460 MR0053460

[12] 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. MR2420458

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

[14] LEEB, H. and PÖTSCHER, B. M. (2006). Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some general lower risk-bound results. *Econometric Theory* **22** 69–97.

[15] LIANG, K.-Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalised linear models. *Biometrika* **73** 13–22.

[16] LITTLE, R. and AN, H. (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statist. Sinica* **14** 949–968. MR2089342

[17] LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley, New York.

[18] LONG, Q., ZHANG, X. and JOHNSON, B. A. (2011). Robust estimation of area under ROC curve using auxiliary variables in the presence of missing biomarker values. *Biometrics* **67** 559–567.

[19] MENG, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statist. Sci.* **9** 538–573.

[20] NEWEY, W. K., HSIEH, F. and ROBINS, J. M. (2004). Twicing kernels and a small bias property of semiparametric estimators. *Econometrica* **72** 947–962.

[21] PAIK, M. C. (1997). The generalized estimating equations approach when data are not missing completely at random. *J. Amer. Statist. Assoc.* **92** 1320–1329.

[22] PORTER, K. E., GRUBER, S., VAN DER LAAN, M. J. and SEKHON, J. S. (2011). The relative performance of targeted maximum likelihood estimators. *Int. J. Biostat.* **7** Article 31.

[23] QI, L., WANG, C. Y. and PRENTICE, R. L. (2005). Weighted estimators for proportional hazards regression with missing covariates. *J. Amer. Statist. Assoc.* **100** 1250–1263. MR2236439

[24] ROBINS, J. and ROTNITZKY, A. (1998). Discussion on the paper by Firth and Bennett. *J. Roy. Statist. Soc. Ser. B* **60** 51–52.

[25] ROBINS, J., SUED, M., LEI-GOMEZ, Q. and ROTNITZKY, A. (2007). Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable [MR2420458]. *Statist. Sci.* **22** 544–559. MR2420460

[26] ROBINS, J. M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association Section on Bayesian Statistical Science* 1999 6–10. Amer. Statist. Assoc., Alexandria, VA.

[27] ROBINS, J. M. and GILL, R. D. (1997). Non-response models for the analysis of non-monotone ignorable missing data. *Stat. Med.* **16** 39–56.

[28] ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866.

[29] ROTNITZKY, A., FARAGGI, D. and SCHISTERMAN, E. (2006). Doubly robust estimation of the area under the receiver-operating characteristic curve in the presence of verification bias. *J. Amer. Statist. Assoc.* **101** 1276–1288. MR2328313

[30] ROTNITZKY, A., LEI, Q. H., SUED, M. and ROBINS, J. M. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika* **99** 439–456. MR2931264

[31] ROTNITZKY, A. and VANSTEELANDT, S. (2014). Double-robust methods. In *Handbook of Missing Data Methodology* (G. Molenberghs, G. Fitzmaurice, M. G. Kenward, A. Tsiatis and G. Verbeke, eds.) 185–212. CRC Press, Boca Raton, FL.

[32] SCHARFSTEIN, D. O., ROTNITZKY, A. and ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models: Rejoinder. *J. Amer. Statist. Assoc.* **94** 1135–1146.

[33] SCHNITZER, M. E., LOK, J. J. and BOSCH, R. J. (2016). Double robust and efficient estimation of a prognostic model for events in the presence of dependent censoring. *Biostatistics* **17** 165–177. MR3449858

[34] SEAMAN, S. and COPAS, A. (2009). Doubly robust generalized estimating equations for longitudinal data. *Stat. Med.* **28** 937–955. MR2518358

[35] SEAMAN, S. R., GALATI, J., JACKSON, D. and CARLIN, J. (2013). What is meant by "missing at random"? *Statist. Sci.* **28** 257–268. MR3112409

[36] SEAMAN, S. R. and VANSTEELANDT, S. (2018). Supplement to "Introduction to double robust methods for incomplete data." DOI:10.1214/18-STS647SUPP.

[37] TAN, Z. (2006). A distributional approach for causal inference using propensity scores. *J. Amer. Statist. Assoc.* **101** 1619–1637. MR2279484

[38] TAN, Z. (2008). Comment: Improved local efficiency and double robustness. *Int. J. Biostat.* **4** Article 10. MR2426120

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

[40] TSIATIS, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer, New York. MR2233926

[41] TSIATIS, A. A. and DAVIDIAN, M. (2014). Missing data methods: A semi-parametric perspective. In *Handbook of Missing Data Methodology* (G. Molenberghs, G. Fitzmaurice, M. G. Kenward, A. Tsiatis and G. Verbeke, eds.) Chapter 8. CRC Press, Boca Raton, FL.

[42] TSIATIS, A. A., DAVIDIAN, M. and CAO, W. (2011). Improved doubly robust estimation when data are monotonely coarsened, with application to longitudinal studies with dropout. *Biometrics* **67** 536–545. MR2829022

[43] VAN DER LAAN, M. J. and RUBIN, D. B. (2006). Targeted maximum likelihood learning. *Int. J. Biostat.* **2** Art. 11. MR2306500

[44] VANSTEELANDT, S., CARPENTER, J. and KENWARD, M. G. (2015). Analysis of incomplete data using inverse probability weighting and doubly robust estimators. *Methodology* **6** 37–48.

[45] VAN DER LAAN, M. J. and GRUBER, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *Int. J. Biostat.* **6** Article 17. MR2653848

[46] VERMEULEN, K. and VANSTEELANDT, S. (2015). Bias-reduced doubly robust estimation. *J. Amer. Statist. Assoc.* **110** 1024–1036. MR3420681

[47] WILSON, A. and REICH, B. J. (2014). Confounder selection via penalized credible regions. *Biometrics* **70** 852–861. MR3295746

[48] WIRTH, K. E., TCHETGEN TCHETGEN, E. J. and MURRAY, M. (2010). Adjustment for missing data in complex surveys using doubly robust estimation. *Epidemiology* **21** 863–871.

Statistical Science
2018, Vol. 33, No. 2, 198–213
https://doi.org/10.1214/17-STS630

Bayesian Approaches for Missing Not at Random Outcome Data: The Role of Identifying Restrictions

Antonio R. Linero and Michael J. Daniels

Abstract. Missing data is almost always present in real datasets, and introduces several statistical issues. One fundamental issue is that, in the absence of strong uncheckable assumptions, effects of interest are typically not nonparametrically identified. In this article, we review the generic approach of the use of identifying restrictions from a likelihood-based perspective, and provide points of contact for several recently proposed methods. An emphasis of this review is on restrictions for nonmonotone missingness, a subject that has been treated sparingly in the literature. We also present a general, fully Bayesian, approach which is widely applicable and capable of handling a variety of identifying restrictions in a uniform manner.

Key words and phrases: Missing data, MNAR, mixture models, multiple imputation, nonignorable missingness, nonparametric Bayes.

REFERENCES

BIRMINGHAM, J., ROTNITZKY, A. and FITZMAURICE, G. M. (2003). Pattern-mixture and selection models for analysing longitudinal data with monotone missing patterns. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 275–297. MR1959827

BUNOUF, P., MOLENBERGHS, G., GROUIN, J.-M. and THIJS, H. (2015). A SAS program combining R functionalities to implement pattern-mixture models. *J. Stat. Softw.* **68** 1–26.

CARPENTER, J. and KENWARD, M. (2012). *Multiple Imputation and Its Application*. Wiley, New York.

CARPENTER, J., POCOCK, S. and JOHAN LAMM, C. (2002). Coping with missing data in clinical trials: A model-based approach applied to asthma trials. *Stat. Med.* **21** 1043–1066.

COX, D. and DONNELLY, C. A. (2011). *Principles of Applied Statistics*, 1st ed. Cambridge Univ. Press, Cambridge. MR2817147

CREEMERS, A., HENS, N., AERTS, M., MOLENBERGHS, G., VERBEKE, G. and KENWARD, M. G. (2010). A sensitivity analysis for shared-parameter models for incomplete longitudinal outcomes. *Biom. J.* **52** 111–125. MR2756597

CREEMERS, A., HENS, N., AERTS, M., MOLENBERGHS, G., VERBEKE, G. and KENWARD, M. G. (2011). Generalized shared-parameter models and missingness at random. *Stat. Model.* **11** 279–310. MR2906703

DANIELS, M. J. (1999). A prior for the variance in hierarchical models. *Canad. J. Statist.* **27** 567–578. MR1745822

DANIELS, M. J. and HOGAN, J. W. (2000). Reparameterizing the pattern mixture model for sensitivity analyses under informative dropout. *Biometrics* **56** 1241–1248. MR1815627

DANIELS, M. J. and HOGAN, J. W. (2008). *Missing Data in Longitudinal Studies*, 1st ed. Chapman & Hall/CRC, Boca Raton, FL.

DANIELS, M. J. and LINERO, A. R. (2015). Bayesian nonparametrics for missing data in longitudinal clinical trials. In *Nonparametric Bayesian Inference in Biostatistics* 423–446. Springer, Cham.

DANIELS, M. J. and LUO, X. (2017). A note on "congeniality" for missing data in the presence of auxiliary covariates. Technical report.

DANIELS, M. J., WANG, C. and MARCUS, B. H. (2014). Fully Bayesian inference under ignorable missingness in the presence of auxiliary covariates. *Biometrics* **70** 62–72. MR3251667

DIGGLE, P. and KENWARD, M. G. (1994). Informative drop-out in longitudinal data analysis. *Appl. Stat.* **43** 49–73.

DUNSON, D. B. and PERREAULT, S. D. (2001). Factor analytic models of clustered multivariate data with informative censoring. *Biometrics* **57** 302–308. MR1833321

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

GASKINS, J., DANIELS, M. and MARCUS, B. (2016). Bayesian methods for nonignorable dropout in joint models in smok-

Antonio R. Linero is Assistant Professor, Department of Statistics, Florida State University, Tallahassee, Florida 32306, USA (e-mail: arlinero@stat.fsu.edu). Michael J. Daniels is Professor and Chair, Department of Statistics, University of Florida, Gainesville, Florida 32611, USA (e-mail: mdaniels@stat.ufl.edu).

ing cessation studies. *J. Amer. Statist. Assoc.* **111** 1454–1465. MR3601701

HAREL, O. and SCHAFER, J. L. (2009). Partial and latent ignorability in missing-data problems. *Biometrika* **96** 37–50. MR2482133

HAREL, O. and ZHOU, X.-H. (2007). Multiple imputation: Review of theory, implementation and software. *Stat. Med.* **26** 3057–3077. MR2380504

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

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

HOGAN, J. W. and LAIRD, N. M. (1997). Mixture models for the joint distribution of repeated measures and event times. *Stat. Med.* **16** 239–257.

IBRAHIM, J. G., LIPSITZ, S. R. and CHEN, M.-H. (1999). Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **61** 173–190. MR1664045

IBRAHIM, J. G. and MOLENBERGHS, G. (2009). Missing data methods in longitudinal studies: A review. *Test* **18** 1–43. MR2495958

KENWARD, M. G. (1998). Selection models for repeated measurements with non-random dropout: An illustration of sensitivity. *Stat. Med.* **17** 2723–2732.

KENWARD, M. G., MOLENBERGHS, G. and THIJS, H. (2003). Pattern-mixture models with proper time dependence. *Biometrika* **90** 53–71. MR1966550

KIM, C., DANIELS, M. J., MARCUS, B. H. and ROY, J. A. (2017). A framework for Bayesian nonparametric inference for causal effects of mediation. *Biometrics* **73** 401–409. MR3665957

LEACY, F. P., FLOYD, S., YATES, T. A. and WHITE, I. R. (2017). Analyses of sensitivity to the missing-at-random assumption using multiple imputation with delta adjustment: Application to a tuberculosis/HIV prevalence survey with incomplete HIV-status data. *Am. J. Epidemiol.* **185** 304–315.

LIN, H., LIU, D. and ZHOU, X.-H. (2010). A correlated random-effects model for normal longitudinal data with nonignorable missingness. *Stat. Med.* **29** 236–247. MR2750513

LINERO, A. R. (2017). Bayesian nonparametric analysis of longitudinal studies in the presence of informative missingness. *Biometrika* **104** 327–341. MR3698257

LINERO, A. R. and DANIELS, M. J. (2015). A flexible Bayesian approach to monotone missing data in longitudinal studies with informative dropout with application to a schizophrenia clinical trial. *J. Amer. Statist. Assoc.* **110** 45–55.

LINERO, A. R. and DANIELS, M. J. (2018). Supplement to "Bayesian approaches for missing not at random outcome data: The role of identifying restrictions." DOI:10.1214/17-STS630SUPP.

LITTLE, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *J. Amer. Statist. Assoc.* **88** 125–134.

LITTLE, R. J. A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika* **81** 471–483. MR1311091

LITTLE, R. J. (1995). Modeling the drop-out mechanism in repeated-measures studies. *J. Amer. Statist. Assoc.* **90** 1112–1121. MR1354029

LIUBLINSKA, V. and RUBIN, D. B. (2014). Sensitivity analysis for a partially missing binary outcome in a two-arm randomized clinical trial. *Stat. Med.* **33** 4170–4185. MR3267402

MENG, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statist. Sci.* **9** 538–558.

MOLENBERGHS, G., MICHIELS, B., KENWARD, M. G. and DIGGLE, P. J. (1998). Monotone missing data and pattern-mixture models. *Stat. Neerl.* **52** 153–161. MR1649081

MURRAY, J. S. and REITER, J. P. (2016). Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence. *J. Amer. Statist. Assoc.* **111** 1466–1479. MR3601702

NATIONAL RESEARCH COUNCIL (2010). *The Prevention and Treatment of Missing Data in Clinical Trials*. The National Academies Press, Washington, DC.

ROBINS, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—Application to control of the healthy worker survivor effect. *Math. Model.* **7** 1393–1512. MR0877758

ROBINS, J. M. (1997). Non-response models for the analysis of non-monotone non-ignorable missing data. *Stat. Med.* **16** 21–37.

ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Amer. Statist. Assoc.* **90** 106–121. MR1325118

ROTNITZKY, A., ROBINS, J. M. and SCHARFSTEIN, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *J. Amer. Statist. Assoc.* **93** 1321–1339. MR1666631

ROY, J. (2003). Modeling longitudinal data with nonignorable dropouts using a latent dropout class model. *Biometrics* **59** 441–456. MR2025106

ROY, J., LUM, K. J. and DANIELS, M. J. (2017). A Bayesian nonparametric approach to marginal structural models for point treatments and a continuous or survival outcome. *Biostatistics* **18** 32–47. MR3612272

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

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York. MR0899519

RUBIN, D. B. (1996). Multiple imputation after 18+ years. *J. Amer. Statist. Assoc.* **91** 473–489.

SADINLE, M. and REITER, J. P. (2017a). Itemwise conditionally independent nonresponse modeling for incomplete multivariate data. *Biometrika* **104** 207–220.

SADINLE, M. and REITER, J. R. (2017b). Sequential identification of nonignorable missing data mechanisms. *Statist. Sinica* To appear.

SCHARFSTEIN, D. O., ROTNITZKY, A. and ROBINS, J. M. (1999). Adjusting for nonignorable dropout using semiparametric nonresponse models. *J. Amer. Statist. Assoc.* **94** 1135–1146.

SCHARFSTEIN, D., MCDERMOTT, A., OLSON, W. and WIEGAND, F. (2014). Global sensitivity analysis for repeated measures studies with informative dropout: A fully parametric approach. *Stat. Biopharm. Res.* **6** 338–348.

SCHARFSTEIN, D., MCDERMOTT, A., DIAZ, I., CARONE, M., LUNARDON, N. and TURKOZ, I. (2018). Global sensitivity analysis for repeated measres studies with informative dropout: A semiparametric approach. *Biometrics* **74** 207–219.

SEAMAN, S., GALATI, J., JACKSON, D. and CARLIN, J. (2013). What is meant by "missing at random"? *Statist. Sci.* **28** 257–268. MR3112409

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

SHPITSER, I. (2016). Consistent estimation of functions of data missing non-monotonically and not at random. In *Advances in Neural Information Processing Systems* 3144–3152.

SI, Y. and REITER, J. P. (2013). Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *J. Educ. Behav. Stat.* **38** 499–521.

TCHETGEN TCHETGEN, E. J., WANG, L. and SUN, B. (2016). Discrete choice models for nonmonotone nonignorable missing data: Identification and inference. Preprint. Available at arXiv:1607.02631.

THIJS, H., MOLENBERGHS, G., MICHIELS, B., VERBEKE, G. and CURRAN, D. (2002). Strategies to fit pattern-mixture models. *Biostatistics* **3** 245–265.

TSIATIS, A. A. (2007). *Semiparametric Theory and Missing Data*. Springer, New York.

VANSTEELANDT, S., ROTNITZKY, A. and ROBINS, J. M. (2007). Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. *Biometrika* **94** 841–860. MR2416795

VANSTEELANDT, S., GOETGHEBEUR, E., KENWARD, M. G. and MOLENBERGHS, G. (2006). Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statist. Sinica* **16** 953–979. MR2281311

VAN BUUREN, S. (2012). *Flexible Imputation of Missing Data*. Chapman and Hall/CRC, Boca Raton, FL.

WANG, C. and DANIELS, M. J. (2011). A note on MAR, identifying restrictions, model comparison, and sensitivity analysis in pattern mixture models with and without covariates for incomplete data. *Biometrics* **67** 810–818. MR2829255

WANG, C., DANIES, M. J., SCHARFSTEIN, D. O. and LAND, S. (2010). A Bayesian shrinkage model for incomplete longitudinal binary data with application to the breast cancer prevention trial. *J. Amer. Statist. Assoc.* **105** 1333–1346. MR2796554

WU, M. C. and CARROLL, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* **45** 175–188. MR0931633

XU, D., DANIELS, M. J. and WINTERSTEIN, A. G. (2016). Sequential BART for imputation of missing covariates. *Biostatistics* **17** 589–602. MR3603956

Statistical Science
2018, Vol. 33, No. 2, 214–237
https://doi.org/10.1214/18-STS645

Causal Inference: A Missing Data Perspective

Peng Ding and Fan Li

Abstract. Inferring causal effects of treatments is a central goal in many disciplines. The potential outcomes framework is a main statistical approach to causal inference, in which a causal effect is defined as a comparison of the potential outcomes of the same units under different treatment conditions. Because for each unit at most one of the potential outcomes is observed and the rest are missing, causal inference is inherently a missing data problem. Indeed, there is a close analogy in the terminology and the inferential framework between causal inference and missing data. Despite the intrinsic connection between the two subjects, statistical analyses of causal inference and missing data also have marked differences in aims, settings and methods. This article provides a systematic review of causal inference from the missing data perspective. Focusing on ignorable treatment assignment mechanisms, we discuss a wide range of causal inference methods that have analogues in missing data analysis, such as imputation, inverse probability weighting and doubly robust methods. Under each of the three modes of inference—Frequentist, Bayesian and Fisherian randomization—we present the general structure of inference for both finite-sample and super-population estimands, and illustrate via specific examples. We identify open questions to motivate more research to bridge the two fields.

Key words and phrases: Assignment mechanism, ignorability, imputation, missing data mechanism, observational studies, potential outcome, propensity score, randomization, weighting.

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. (2011). Bias corrected matching estimators for average treatment effects. *J. Bus. Econom. Statist.* **29** 1–11. MR2789386

ANDREWS, D. W. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica* **68** 399–405. MR1748009

ANGRIST, J. D., IMBENS, G. W. and RUBIN, D. B. (1996). Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* **91** 444–455. MR3042387

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

ATHEY, S. and IMBENS, G. W. (2015). Machine learning methods for estimating heterogeneous causal effects. Available at arXiv:1504.01132.

ATHEY, S., IMBENS, G. W. and WAGER, S. (2018). Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* To appear. Available at https://arxiv.org/abs/1604.07125.

ATHEY, S., IMBENS, G., PHAM, T. and WAGER, S. (2017). Estimating average treatment effects: Supplementary analyses and remaining challenges. *Am. Econ. Rev.* **107** 278–281.

BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–972. MR2216189

BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.* **81** 608–650. MR3207983

Peng Ding is Assistant Professor, Department of Statistics, University of California, Berkeley, 425 Evans Hall, Berkeley, California 94720, USA (e-mail: pengdingpku@berkeley.edu). Fan Li is Associate Professor, Department of Statistical Science, Duke University, 122 Old Chemistry Bldg, Durham, North Carolina 27708, USA (e-mail: fl35@duke.edu).

BELLONI, A., CHERNOZHUKOV, V., FERNÁNDEZ-VAL, I. and HANSEN, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica* **85** 233–298. MR3611771

BICKEL, P. J. and DOKSUM, K. A. (2015). *Mathematical Statistics*: *Basic Ideas and Selected Topics*, *Volume I*, 2nd ed. CRC Press, Boca Raton, FL. MR3445928

BLONIARZ, A., LIU, H., ZHANG, C.-H., SEKHON, J. S. and YU, B. (2016). Lasso adjustments of treatment effect estimates in randomized experiments. *Proc. Natl. Acad. Sci. USA* **113** 7383–7390. MR3531136

CARPENTER, B., GELMAN, A., HOFFMAN, M. D., LEE, D., GOODRICH, B., BETANCOURT, M., BRUBAKER, M., GUO, J., LI, P. and RIDDELL, A. (2017). Stan: A probabilistic programming language. *J. Stat. Softw.* **76** 1–32.

CHAPIN, F. S. (1947). *Experimental Designs in Sociological Research*. Harper, New York.

CHEN, H., GENG, Z. and ZHOU, X.-H. (2009). Identifiability and estimation of causal effects in randomized trials with noncompliance and completely nonignorable missing data. *Biometrics* **65** 675–682. MR2649840

CHENG, J. and SMALL, D. S. (2006). Bounds on causal effects in three-arm trials with non-compliance. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 815–836. MR2301296

CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. and ROBINS, J. (2016). Double/debiased machine learning for treatment and causal parameters. Preprint. Available at arXiv:1608.00060.

CHIB, S. and JACOBI, L. (2016). Bayesian fuzzy regression discontinuity analysis and returns to compulsory schooling. *J. Appl. Econometrics* **31** 1026–1047. MR3556650

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

COCHRAN, W. G. (1953). *Sampling Techniques*, 1st ed. Wiley, New York. MR0054199

COCHRAN, W. G. (1957). Analysis of covariance: Its nature and uses. *Biometrics* **13** 261–281. MR0090952

COCHRAN, W. G. (2007). *Sampling Techniques*, 3rd ed. Wiley, New York.

CORNFIELD, J., HAENSZEL, W., HAMMOND, E. et al. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *J. Natl. Cancer Inst.* **22** 173–203.

CRUMP, R. K., HOTZ, V. J., IMBENS, G. W. and MITNIK, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96** 187–199. MR2482144

DAWID, A. P. (2000). Causal inference without counterfactuals. *J. Amer. Statist. Assoc.* **95** 407–424. MR1803167

DAWID, A. P. MUSIO, M. and MURTAS, R. (2017). The probability of causation. *Law, Probability and Risk* **16** 163–179.

DEMPSTER, A., LAIRD, N. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **39** 1–38. MR0501537

DING, P. (2014). Three occurrences of the hyperbolic-secant distribution. *Amer. Statist.* **68** 32–35. MR3303831

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. (2016). Randomization inference for treatment effect variation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 655–671. MR3506797

DING, P. and GENG, Z. (2014). Identifiability of subgroup causal effects in randomized experiments with nonignorable missing covariates. *Stat. Med.* **33** 1121–1133. MR3247784

DING, P. and LU, J. (2017). Principal stratification analysis using principal scores. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 757–777. MR3641406

DING, P. and VANDERWEELE, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology* **27** 368–377.

DING, P., GENG, Z., YAN, W. and ZHOU, X.-H. (2011). Identifiability and estimation of causal effects by principal stratification with outcomes truncated by death. *J. Amer. Statist. Assoc.* **106** 1578–1591. MR2896858

DING, W. and SONG, P. X.-K. (2016). EM algorithm in Gaussian copula with missing data. *Comput. Statist. Data Anal.* **101** 1–11. MR3504831

ELLIOTT, M., RAGHUNATHAN, T. and LI, Y. (2010). Bayesian inference for causal mediation effects using principal stratification with dichotomous mediators and outcomes. *Biostatistics* **11** 353–372.

FAN, Y., GUERRE, E. and ZHU, D. (2017). Partial identification of functionals of the joint distribution of "potential outcomes". *J. Econometrics* **197** 42–59. MR3598644

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

FELLER, A., GREIF, E., MIRATRIX, L. and PILLAI, N. (2016). Principal stratification in the twilight zone: Weakly separated components in finite mixture models. Preprint. Available at arXiv:1602.06595.

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

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

FRANGAKIS, C. and RUBIN, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika* **86** 365–378. MR1705410

FRANGAKIS, C. and RUBIN, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58** 21–29. MR1891039

FRUMENTO, P., MEALLI, F., PACINI, B. and RUBIN, D. B. (2012). Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data. *J. Amer. Statist. Assoc.* **107** 450–466. MR2980057

FRUMENTO, P., MEALLI, F., PACINI, B. and RUBIN, D. B. (2016). The fragility of standard inferential approaches in principal stratification models relative to direct likelihood approaches. *Stat. Anal. Data Min.* **9** 58–70. MR3465093

GALLOP, R., SMALL, D., LIN, J., ELLIOT, M., JOFFE, M. and HAVE, T. T. (2009). Mediation analysis with principal stratification. *Stat. Med.* **28** 1108–1130. MR2662200

GELFAND, A. and SMITH, A. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409. MR1141740

GELMAN, A., MENG, X.-L. and STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statist. Sinica* **6** 733–807. MR1422404

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2014). *Bayesian Data Analysis*, 3nd ed. Chapman & Hall/CRC, Boca Raton, FL. MR2027492

GILBERT, P. and HUDGENS, M. (2008). Evaluating candidate principal surrogate endpoints. *Biometrics* **64** 1146–1154. MR2522262

GRAHAM, B. S., DE XAVIER PINTO, C. C. and EGEL, D. (2012). Inverse probability tilting for moment condition models with missing data. *Rev. Econ. Stud.* **79** 1053–1079. MR2986390

GRILLI, L. and MEALLI, F. (2008). Nonparametric bounds on the causal effect of university studies on job opportunities using principal stratification. *J. Educ. Behav. Stat.* **33** 111–130.

GUSTAFSON, P. (2009). What are the limits of posterior distributions arising from nonidentified models, and why should we care? *J. Amer. Statist. Assoc.* **104** 1682–1695. MR2750585

GUSTAFSON, P. (2015). *Bayesian Inference for Partially Identified Models*: *Exploring the Limits of Limited Data*. CRC Press, Boca Raton, FL.

HAHN, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66** 315–331. MR1612242

HAINMUELLER, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Polit. Anal.* **20** 25–46.

HÁJEK, J. (1971). Comment on a paper by D. Basu. In *Foundations of Statistical Inference* (V. P. Godambe and D. A. Sprott, eds.) 236. Holt, Rinehart and Winston, Toronto.

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

HECKMAN, J., LOPES, H. and PIATEK, R. (2014). Treatment effects: A Bayesian perspective. *Econometric Rev.* **33** 36–67. MR3170840

HIRANO, K. and IMBENS, G. W. (2004). The propensity score with continuous treatments. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. 73–84. Wiley, Chichester. MR2134803

HIRANO, K., IMBENS, G. W. and RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71** 1161–1189. MR1995826

HO, D. E., IMAI, K., KING, G. and STUART, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *J. Stat. Softw.* **42** 1–28.

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

HOLLAND, P. (1986). Statistics and causal inference (with discussion). *J. Amer. Statist. Assoc.* **81** 945–970. MR0867618

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

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability* **1** 221–233. Univ. California Press, Berkeley, CA. MR0216620

ICHINO, A., MEALLI, F. and NANNICINI, T. (2008). From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity? *J. Appl. Econometrics* **23** 305–327. MR2420362

IMAI, K. (2008). Sharp bounds on the causal effects in randomized experiments with "truncation-by-death". *Statist. Probab. Lett.* **78** 144–149. MR2382067

IMAI, K. and RATKOVIC, M. (2014). Covariate balancing propensity score. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 243–263. MR3153941

IMAI, K. and VAN DYK, D. (2004). Causal treatment 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. (2003). Sensitivity to exogeneity assumptions in program evaluation. *Am. Econ. Rev.* **93** 126–132.

IMBENS, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev. Econ. Stat.* **86** 4–29.

IMBENS, G. W. and ANGRIST, J. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62** 467–476.

IMBENS, G. W. and RUBIN, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *Ann. Statist.* **25** 305–327. MR1429927

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

KANG, J. D. 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. MR2420458

LI, X. and DING, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *J. Amer. Statist. Assoc.* **112** 1759–1769. MR3750897

LI, F., MATTEI, A. and MEALLI, F. (2015). Evaluating the causal effect of university grants on student dropout: Evidence from a regression discontinuity design using principal stratification. *Ann. Appl. Stat.* **9** 1906–1931. MR3456358

LI, F., MORGAN, K. and ZASLAVSKY, A. (2018). Balancing covariates via propensity score weighting. *J. Amer. Statist. Assoc.* To appear. Available at https://doi.org/10.1080/01621459.2016.1260466.

LI, F., BACCINI, M., MEALLI, F., ZELL, E. R., FRANGAKIS, C. E. and RUBIN, D. B. (2014). Multiple imputation by ordered monotone blocks with application to the anthrax vaccine research program. *J. Comput. Graph. Statist.* **23** 877–892. MR3224660

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

LINDLEY, D. V. (1972). *Bayesian Statistics*: *A Review*. SIAM, Philadelphia, PA. MR0329081

LITTLE, R. J. (1988). Missing-data adjustments in large surveys. *J. Bus. Econom. Statist.* **6** 287–296.

LITTLE, R. and AN, H. (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statist. Sinica* **14** 949–968. MR2089342

LITTLE, R. J. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley-Interscience, Hoboken, NJ. MR1925014

LIUBLINSKA, V. and RUBIN, D. B. (2014). Sensitivity analysis for a partially missing binary outcome in a two-arm randomized clinical trial. *Stat. Med.* **33** 4170–4185. MR3267402

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

LUNCEFORD, J. K. and DAVIDIAN, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Stat. Med.* **23** 2937–2960.

MANSKI, C. F. (1990). Nonparametric bounds on treatment effects. *Am. Econ. Rev.* **80** 319–323.

MATTEI, A. and MEALLI, F. (2011). Augmented designs to assess principal strata direct effects. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 729–752. MR2867456

MATTEI, A., MEALLI, F. and PACINI, B. (2014). Identification of causal effects in the presence of nonignorable missing outcome values. *Biometrics* **70** 278–288. MR3258033

MEALLI, F. and RUBIN, D. B. (2015). Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika* **102** 995–1000. MR3431570

MEALLI, F., IMBENS, G. W., FERRO, S. and BIGGERI, A. (2004). Analyzing a randomized trial on breast self-examination with noncompliance and missing outcomes. *Biostatistics* **5** 207–222.

MEBANE, W. R. JR and POAST, P. (2013). Causal inference without ignorability: Identification with nonrandom assignment and missing treatment data. *Polit. Anal.* **21** 233–251.

MENG, X.-L. (1994). Posterior predictive p-values. *Ann. Statist.* **22** 1142–1160. MR1311969

MERCATANTI, A. (2004). Analyzing a randomized experiment with imperfect compliance and ignorable conditions for missing data: Theoretical and computational issues. *Comput. Statist. Data Anal.* **46** 493–509. MR2061964

MERCATANTI, A. and LI, F. (2014). Do debit cards increase household spending? Evidence from a semiparametric causal analysis of a survey. *Ann. Appl. Stat.* **8** 2405–2508. MR3292506

MERCATANTI, A. and LI, F. (2017). Do debit cards decrease cash demand? Causal inference and sensitivity analysis using principal stratification. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **66** 759–776. MR3670416

MIRATRIX, L. W., SEKHON, J. S. and YU, B. (2013). Adjusting treatment effect estimates by post-stratification in randomized experiments. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 369–396. MR3021392

MITRA, R. and REITER, J. P. (2011). Estimating propensity scores with missing covariate data using general location mixture models. *Stat. Med.* **30** 627–641. MR2767460

MITRA, R. and REITER, J. P. (2016). A comparison of two methods of estimating propensity scores after multiple imputation. *Stat. Methods Med. Res.* **25** 188–204.

MOLINARI, F. (2010). Missing treatments. *J. Bus. Econom. Statist.* **28** 82–95. MR2650602

MURRAY, J. S. and REITER, J. P. (2016). Multiple imputation of missing categorical and continuous values via Bayesian mixture models with local dependence. *J. Amer. Statist. Assoc.* **111** 1466–1479. MR3601702

NEWEY, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *J. Econometrics* **79** 147–168. MR1457700

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

NEYMAN, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–472. MR1092986

NOLEN, T. L. and HUDGENS, M. G. (2011). Randomization-based inference within principal strata. *J. Amer. Statist. Assoc.* **106** 581–593. MR2847972

QIN, J. (2017). *Biased Sampling, Over-Identified Parameter Problems and Beyond*. Springer, Singapore. MR3675467

RICHARDSON, T. S., EVANS, R. J. and ROBINS, J. M. (2010). Transparent parameterizations of models for potential outcomes. In *Bayesian Statistics* **9** 569–610. Oxford Univ. Press, Oxford.

RIDGEWAY, G., MCCAFFREY, D., MORRAL, A., GRIFFIN, B. A. and BURGETTE, L. (2017). twang: Toolkit for Weighting and Analysis of Nonequivalent Groups. R package version 1.5.

ROBINS, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods—Application to control of the healthy worker survivor effect. *Math. Modelling* **7** 1393–1512. MR0877758

ROBINS, J. M. and RITOV, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semiparametric models. *Stat. Med.* **16** 285–319.

ROBINS, J. M., ROTNITZKY, A. and SCHARFSTEIN, D. O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials* 1–94. Springer, New York. MR1731681

ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Amer. Statist. Assoc.* **90** 106–121. MR1325118

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** 1143–1156. MR1804240

ROSENBAUM, P. R. (1984a). Conditional permutation tests and the propensity score in observational studies. *J. Amer. Statist. Assoc.* **79** 565–574. MR0763575

ROSENBAUM, P. R. (1984b). The consquences of adjustment for a concomitant variable that has been affected by the treatment. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **147** 656–666.

ROSENBAUM, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika* **74** 13–26. MR0885915

ROSENBAUM, P. R. (2002a). Covariance adjustment in randomized experiments and observational studies. *Statist. Sci.* **17** 286–327. MR1962487

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

ROSENBAUM, P. R. (2010). *Design of Observational Studies*. Springer, New York. MR2561612

ROSENBAUM, P. R. and RUBIN, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **45** 212–218.

ROSENBAUM, P. R. and RUBIN, D. B. (1983b). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974

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.

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

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

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

RUBIN, D. B. (1977). Assignment to a treatment group on the basis of a covariate. *Journal of Educational Statistics* **2** 1–26.

RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58. MR0472152

RUBIN, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *J. Amer. Statist. Assoc.* **74** 318–324.

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. (1984). Bayesianly justifiable and relevant frequency calculations for the applies statistician. *Ann. Statist.* **12** 1151–1172.

RUBIN, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *J. Bus. Econom. Statist.* **4** 87–94.

RUBIN, D. B. (1998). More powerful randomization-based p-values in double-blind trials with non-compliance. *Stat. Med.* **17** 371–385.

RUBIN, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* **100** 322–331. MR2166071

RUBIN, D. B. (2006a). Causal inference through potential outcomes and principal stratification: Application to studies with "censoring" due to death. *Statist. Sci.* **91** 299–321. MR2339125

RUBIN, D. B. (2006b). *Matched Sampling for Causal Effects*. Cambridge Univ. Press, Cambridge. MR2307965

RUBIN, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Stat. Med.* **26** 20–36. MR2312697

RUBIN, D. B. (2008). For objective causal inference, design trumps analysis. *Ann. Appl. Stat.* **2** 808–840. MR2516795

SCHARFSTEIN, D., ROTNITZKY, A. and ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). *J. Amer. Statist. Assoc.* **94** 1096–1146. MR1731478

SCHWARTZ, S., LI, F. and REITER, J. P. (2012). Sensitivity analysis for unmeasured confounding in principal stratification settings with binary variables. *Stat. Med.* **31** 949–962. MR2913871

SEAMAN, S., GALATI, J., JACKSON, D. and CARLIN, J. (2013). What is meant by "missing at random"? *Statist. Sci.* **28** 257–268. MR3112409

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

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

TANNER, M. and WONG, W. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82** 528–540. MR0898357

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **58** 267–288. MR1379242

TSIATIS, A. A., DAVIDIAN, M., ZHANG, M. and LU, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Stat. Med.* **27** 4658–4677. MR2528575

TUKEY, J. W. (1993). Tightening the clinical trial. *Controlled Clinical Trials* **14** 266–285.

VAN BUUREN, S. (2012). *Flexible Imputation of Missing Data*. CRC press, Boca Raton, FL.

VAN DER LAAN, M. J. and ROSE, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York. MR2867111

VANDERWEELE, T. (2008). Simple relations between principal stratification and direct and indirect effects. *Statist. Probab. Lett.* **78** 2957–2962. MR2516810

WAGER, S., DU, W., TAYLOR, J. and TIBSHIRANI, R. J. (2016). High-dimensional regression adjustments in randomized experiments. *Proc. Natl. Acad. Sci. USA* **113** 12673–12678. MR3576188

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

YANG, S. and DING, P. (2018). Asymptotic causal inference with observational studies trimmed by the estimated propensity scores. *Biometrika*. To appear. Available at https://arxiv.org/abs/1604.07125.

YANG, F. and SMALL, D. S. (2016). Using post-outcome measurement information in censoring-by-death problems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 299–318. MR3453657

YANG, S., WANG, L. and DING, P. (2017). Nonparametric identification of causal effects with confounders subject to instrumental missingness. Preprint. Available at arXiv:1702.03951.

ZHANG, G. and LITTLE, R. J. (2009). Extensions of the penalized spline of propensity prediction method of imputation. *Biometrics* **65** 911–918. MR2649864

ZHANG, J. and RUBIN, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *J. Educ. Behav. Stat.* **28** 353–358.

ZHANG, J., RUBIN, D. B. and MEALLI, F. (2009). Likelihood-based analysis of the causal effects of job-training programs using principal stratification. *J. Amer. Statist. Assoc.* **104** 166–176.

ZHANG, Z., LIU, W., ZHANG, B., TANG, L. and ZHANG, J. (2016). Causal inference with missing exposure information: Methods and applications to an obstetric study. *Stat. Methods Med. Res.* **25** 2053–2066. MR3553326

ZHOU, J., ZHANG, Z., LI, Z. and ZHANG, J. (2015). Coarsened propensity scores and hybrid estimators for missing data and causal inference. *Int. Stat. Rev.* **83** 449–471. MR3429286

ZIGLER, C. and BELIN, T. (2012). A Bayesian approach to improved estimation of causal effect predictiveness for a principal surrogate endpoint. *Biometrics* **68** 922–932. MR3055197

ZUBIZARRETA, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *J. Amer. Statist. Assoc.* **110** 910–922. MR3420672

Statistical Science
2018, Vol. 33, No. 2, 238–260
https://doi.org/10.1214/18-STS642

Flexible Low-Rank Statistical Modeling with Missing Data and Side Information

William Fithian and Rahul Mazumder

Abstract. We explore a general statistical framework for low-rank modeling of matrix-valued data, based on convex optimization with a generalized nuclear norm penalty. We study several related problems: the usual low-rank matrix completion problem with flexible loss functions arising from generalized linear models; reduced-rank regression and multi-task learning; and generalizations of both problems where side information about rows and columns is available, in the form of features or smoothing kernels. We show that our approach encompasses maximum a posteriori estimation arising from Bayesian hierarchical modeling with latent factors, and discuss ramifications of the missing-data mechanism in the context of matrix completion. While the above problems can be naturally posed as rank-constrained optimization problems, which are nonconvex and computationally difficult, we show how to relax them via generalized nuclear norm regularization to obtain convex optimization problems. We discuss algorithms drawing inspiration from modern convex optimization methods to address these large scale convex optimization computational tasks. Finally, we illustrate our flexible approach in problems arising in functional data reconstruction and ecological species distribution modeling.

Key words and phrases: Matrix completion, nuclear norm regularization, matrix factorization, convex optimization, missing data.

REFERENCES

ABERNETHY, J., BACH, F., EVGENIOU, T. and VERT, J.-P. (2009). A new approach to collaborative filtering: Operator estimation with spectral regularization. *J. Mach. Learn. Res.* **10** 803–826.

AGGARWAL, C. C. and CHEN, B.-C. (2009). Regression-based latent factor models. In *Proceedings of the* 15*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 19–28. ACM, New York.

AGARWAL, D. K. and CHEN, B.-C. (2015). *Statistical Methods for Recommender Systems*. Cambridge Univ. Press, Cambridge.

AGARWAL, D., ZHANG, L. and MAZUMDER, R. (2011). Modeling item–item similarities for personalized recommendations on Yahoo! front page. *Ann. Appl. Stat.* **5** 1839–1875. MR2884924

ANDERSON, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Ann. Math. Stat.* **22** 327–351. MR0042664

ANGST, R., ZACH, C. and POLLEFEYS, M. (2011). The generalized trace-norm and its application to structure-from-motion problems. In 2011 *IEEE International Conference on Computer Vision* (*ICCV*) 2502–2509. IEEE, Los Alamitos, CA.

ATCHADÉ, Y. F., MAZUMDER, R. and CHEN, J. (2015). Scalable computation of regularized precision matrices via stochastic optimization. Preprint. Available at arXiv:1509.00426.

AUDIGIER, V., HUSSON, F. and JOSSE, J. (2016). A principal component method to impute missing values for mixed data. *Adv. Data Anal. Classif.* **10** 5–26. MR3464297

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

BELL, R. M. and KOREN, Y. (2007). Lessons from the Netflix Prize Challenge. *ACM SIGKDD Explor. Newsl.* **9** 75–79.

BENNETT, J. and LANNING, S. (2007). The Netflix Prize. In *Proceedings of KDD Cup and Workshop* 3–6. ACM New York.

William Fithian is Assistant Professor, Department of Statistics, University of California, Berkeley, 301 Evans Hall, Berkeley, California 94720, USA (e-mail: wfithian@berkeley.edu). Rahul Mazumder is Assistant Professor, Sloan School of Management, Operations Research Center and MIT Center for Statistics, Massachusetts Institute of Technology, Building E62-583, 100 Main Street, Cambridge, Massachusetts 02142, USA (e-mail: rahulmaz@mit.edu).

BERTSEKAS, D. P. (1999). *Nonlinear Programming*, 2nd ed. Athena Scientific, Belmont, MA. MR3444832

BERTSIMAS, D., COPENHAVER, M. S. and MAZUMDER, R. (2017). Certifiably optimal low rank factor analysis. *J. Mach. Learn. Res.* **18** Paper No. 29. MR3634896

BOTTOU, L. and BOUSQUET, O. (2008). The trade-offs of large scale learning. In *Advances in Neural Information Processing Systems* **20** (J. C. Platt, D. Koller, Y. Singer and S. T. Roweis, eds.) 161–168. MIT Press, Cambridge, MA.

BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge Univ. Press, Cambridge. MR2061575

BOYD, S., PARIKH, N., CHU, E., PELEATO, B. and ECKSTEIN, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **3** 1–122.

BURER, S. and MONTEIRO, R. D. C. (2005). Local minima and convergence in low-rank semidefinite programming. *Math. Program.* **103** 427–444. MR2166543

CAI, T. and ZHOU, W.-X. (2013). A max-norm constrained minimization approach to 1-bit matrix completion. *J. Mach. Learn. Res.* **14** 3619–3647. MR3159403

CANDES, E. and PLAN, Y. (2010). Matrix completion with noise. *Proc. IEEE* **98** 925–936.

CANDÈS, E. J. and RECHT, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.* **9** 717–772. MR2565240

CANDÈS, E. J. and TAO, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory* **56** 2053–2080. MR2723472

CANDÈS, E. J., LI, X., MA, Y. and WRIGHT, J. (2011). Robust principal component analysis? *J. ACM* **58** Art. ID 11. MR2811000

CANDÈS, E. J., ELDAR, Y. C., STROHMER, T. and VORONINSKI, V. (2015). Phase retrieval via matrix completion [reprint of MR3032952]. *SIAM Rev.* **57** 225–251. MR3345342

CARPENTIER, A., KLOPP, O., LÖFFLER, M. and NICKL, R. (2016). Adaptive confidence sets for matrix completion. Preprint. Available at arXiv:1608.04861.

CHEN, Y. and WAINWRIGHT, M. J. (2015). Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. Preprint. Available at arXiv:1509.03025.

COTTET, V. and ALQUIER, P. (2018). 1-Bit matrix completion: PAC-Bayesian analysis of a variational approximation. *Mach. Learn.* **107** 579–603. MR3761297

DAVENPORT, M. A., PLAN, Y., VAN DEN BERG, E. and WOOTTERS, M. (2014). 1-bit matrix completion. *Inf. Inference* **3** 189–223. MR3311452

DE LEEUW, J. and VAN DER HEIJDEN, P. G. M. (1988). Correspondence analysis of incomplete contingency tables. *Psychometrika* **53** 223–233. MR0955467

DEVOLDER, O., GLINEUR, F. and NESTEROV, Y. (2014). First-order methods of smooth convex optimization with inexact oracle. *Math. Program.* **146** 37–75. MR3232608

FAZEL, M. (2002). Matrix rank minimization with applications. Ph.D. thesis, Stanford Univ.

FITHIAN, W. and MAZUMDER, R. (2013). Scalable convex methods for flexible low-rank matrix modeling. Preprint. Available at arXiv:1308.4211.

FITHIAN, W., ELITH, J., HASTIE, T. and KEITH, D. A. (2015). Bias correction in species distribution models: Pooling survey and collection data for multiple species. *Methods Ecol. Evol.* **6** 424–438.

FOYGEL, R. and SREBRO, N. (2011). Concentration-based guarantees for low-rank matrix reconstruction. In *COLT* 315–340.

FRANK, M. and WOLFE, P. (1956). An algorithm for quadratic programming. *Nav. Res. Logist. Q.* **3** 95–110. MR0089102

FREUND, R. M., GRIGAS, P. and MAZUMDER, R. (2017). An extended Frank–Wolfe method with "in-face" directions, and its application to low-rank matrix completion. *SIAM J. Optim.* **27** 319–346. MR3615468

GERRISH, S. and BLEI, D. M. (2011). Predicting legislative roll calls from text. In *Proceedings of the* 28*th International Conference on Machine Learning* (*ICML*-11) 489–496.

GOLUB, G. H. and VAN LOAN, C. F. (1983). *Matrix Computations*. *Johns Hopkins Series in the Mathematical Sciences* **3**. Johns Hopkins Univ. Press, Baltimore, MD. MR0733103

HASTIE, T., BUJA, A. and TIBSHIRANI, R. (1995). Penalized discriminant analysis. *Ann. Statist.* **23** 73–102. MR1331657

HASTIE, T., TIBSHIRANI, R. and WAINWRIGHT, M. (2015). *Statistical Learning with Sparsity*: *The Lasso and Generalizations*. *Monographs on Statistics and Applied Probability* **143**. CRC Press, Boca Raton, FL. MR3616141

HASTIE, T., MAZUMDER, R., LEE, J. D. and ZADEH, R. (2015). Matrix completion and low-rank SVD via fast alternating least squares. *J. Mach. Learn. Res.* **16** 3367–3402. MR3450542

HUBER, P. J. (2011). *Robust Statistics*. Springer, New York.

JAGGI, M. and SULOVSK, M. (2010). A simple algorithm for nuclear norm regularized problems. In *Proceedings of the* 27*th International Conference on Machine Learning* (*ICML*-10) 471–478.

JAIN, P., NETRAPALLI, P. and SANGHAVI, S. (2013). Low-rank matrix completion using alternating minimization (extended abstract). In *STOC*'13*—Proceedings of the* 2013 *ACM Symposium on Theory of Computing* 665–674. ACM, New York. MR3210828

JOSSE, J. and HUSSON, F. (2012). Handling missing values in exploratory multivariate data analysis methods. *J. SFdS* **153** 79–99. MR3008600

JOSSE, J., WAGER, S. and HUSSON, F. (2016). Confidence areas for fixed-effects PCA. *J. Comput. Graph. Statist.* **25** 28–48. MR3474035

JOURNÉE, M., BACH, F., ABSIL, P.-A. and SEPULCHRE, R. (2010). Low-rank optimization on the cone of positive semidefinite matrices. *SIAM J. Optim.* **20** 2327–2351. MR2678395

KESHAVAN, R. H., MONTANARI, A. and OH, S. (2010). Matrix completion from a few entries. *IEEE Trans. Inform. Theory* **56** 2980–2998. MR2683452

KLOPP, O., LAFOND, J., MOULINES, É. and SALMON, J. (2015). Adaptive multinomial matrix completion. *Electron. J. Stat.* **9** 2950–2975. MR3439190

KOREN, Y. (2010). Collaborative filtering with temporal dynamics. *Commun. ACM* **53** 89–97.

KOREN, Y., BELL, R. and VOLINSKY, C. (2009). Matrix factorization techniques for recommender systems. *Computer* **42** 30–37.

LAFOND, J. (2015). Low rank matrix completion with exponential family noise. Preprint. Available at arXiv:1502.06919.

LARSEN, R. M. (2004). PROPACK—Software for large and sparse SVD calculations. Available at http://sun.stanford.edu/~rmunk/PROPACK.

LEE, J., RECHT, B., SREBRO, N., TROPP, J. and SALAKHUTDINOV, R. (2010). Practical large-scale optimization for max-norm regularization. In *Advances in Neural Information Processing Systems* 1297–1305.

LESIEUR, T., KRZAKALA, F. and ZDEBOROVÁ, L. (2015). MMSE of probabilistic low-rank matrix estimation: Universality with respect to the output channel. In 2015 53*rd Annual Allerton Conference on Communication, Control, and Computing* (*Allerton*) 680–687. IEEE, Los Alamitos, CA.

MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1979). *Multivariate Analysis*. Academic Press, London. MR0560319

MARTIN, A. D. and QUINN, K. M. (2002). Dynamic ideal point estimation via Markov chain Monte Carlo for the US Supreme Court, 1953–1999. *Polit. Anal.* **10** 134–153.

MAZUMDER, R., HASTIE, T. and TIBSHIRANI, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.* **11** 2287–2322. MR2719857

MAZUMDER, R., RADCHENKO, P. and DEDIEU, A. (2017). Subset selection with shrinkage: Sparse linear modeling when the SNR is low. Preprint. Available at arXiv:1708.03288.

MENON, A. K. and ELKAN, C. (2010). A log-linear model with latent features for dyadic prediction. In 2010 *IEEE* 10*th International Conference on Data Mining* (*ICDM*) 364–373. IEEE, Los Alamitos, CA.

NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *J. Mach. Learn. Res.* **13** 1665–1697. MR2930649

NESTEROV, Y. (2004). *Introductory Lectures on Convex Optimization*: *A Basic Course*. *Applied Optimization* **87**. Kluwer Academic, Boston, MA. MR2142598

NESTEROV, YU. (2005). Smooth minimization of non-smooth functions. *Math. Program.* **103** 127–152. MR2166537

PARKER, J. T., SCHNITER, P. and CEVHER, V. (2014a). Bilinear generalized approximate message passing—Part I: Derivation. *IEEE Trans. Signal Process.* **62** 5839–5853. MR3281527

PARKER, J. T., SCHNITER, P. and CEVHER, V. (2014b). Bilinear generalized approximate message passing—Part II: Applications. *IEEE Trans. Signal Process.* **62** 5854–5867. MR3281528

REINSEL, G. C. and VELU, R. P. (1998). *Multivariate Reduced-Rank Regression*: *Theory and Applications*. *Lecture Notes in Statistics* **136**. Springer, New York. MR1719704

RENNIE, J. and SREBRO, N. (2005). Fast maximum margin matrix factorization for collaborative prediction. In *ICML*.

ROWEIS, S. (1998). EM algorithms for PCA and SPCA. In *Advances in Neural Information Processing Systems* **10** 626–632. MIT Press, Cambridge, MA.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. With comments by R. J. A. Little and a reply by the author. MR0455196

SALAKHUTDINOV, R. and MNIH, A. (2008a). Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of the* 25*th International Conference on Machine Learning* 880–887. ACM, New York.

SALAKHUTDINOV, R. and MNIH, A. (2008b). Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems* **20** (J. C. Platt, D. Koller, Y. Singer and S. T. Roweis, eds.) 1257–1264. MIT Press, Cambridge, MA.

SALAKHUTDINOV, R. and SREBRO, N. (2010). Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. Preprint. Available at arXiv:1002.2780.

SREBRO, N., RENNIE, J. and JAAKKOLA, T. (2005). Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems* **17** 1329–1336. MIT Press, Cambridge, MA.

SREBRO, N. and SHRAIBMAN, A. (2005). Rank, trace-norm and max-norm. In *Learning Theory*. *Lecture Notes in Computer Science* **3559** 545–560. Springer, Berlin. MR2203286

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TIPPING, M. E. and BISHOP, C. M. (1999). Probabilistic principal component analysis. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **61** 611–622. MR1707864

TODESCHINI, A., CARON, F. and CHAVENT, M. (2013). Probabilistic low-rank matrix completion with adaptive spectral regularization algorithms. In *Advances in Neural Information Processing Systems* 845–853.

UDELL, M., HORN, C., ZADEH, R. and BOYD, S. (2016). Generalized low rank models. *Found. Trends Mach. Learn.* **9** 1–118.

YANG, Y., MA, J. and OSHER, S. (2013). Seismic data reconstruction via matrix completion. *Inverse Probl. Imaging* **7** 1379–1392. MR3180685

YEE, T. W. and HASTIE, T. J. (2003). Reduced-rank vector generalized linear models. *Stat. Model.* **3** 15–41. MR1977163

YUAN, M., EKICI, A., LU, Z. and MONTEIRO, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 329–346. MR2323756

Statistical Science
2018, Vol. 33, No. 2, 261–276
https://doi.org/10.1214/17-STS638

Missing Information Principle: A Unified Approach for General Truncated and Censored Survival Data Problems

Yifei Sun, Jing Qin and Chiung-Yu Huang

Abstract. It is well known that truncated survival data are subject to sampling bias, where the sampling weight depends on the underlying truncation time distribution. Recently, there has been a rising interest in developing methods to better exploit the information about the truncation time, thus the sampling weight function, to obtain more efficient estimation. In this paper, we propose to treat truncation and censoring as "missing data mechanism" and apply the missing information principle to develop a unified framework for analyzing left-truncated and right-censored data with unspecified or known truncation time distributions. Our framework is structured in a way that is easy to understand and enjoys a great flexibility for handling different types of models. Moreover, a new test for checking the independence between the underlying truncation time and survival time is derived along the same line. The proposed hypothesis testing procedure utilizes all observed data and hence can yield a much higher power than the conditional Kendall's tau test that only involves comparable pairs of observations under truncation. Simulation studies with practical sample sizes are conducted to compare the performance of the proposed method with its competitors. The proposed methodologies are applied to a dementia study and a nursing house study for illustration.

Key words and phrases: Kendall's tau, inverse probability weighted estimator, outcome-dependent sampling, prevalent sampling, self-consistency algorithm.

REFERENCES

ADDONA, V. and WOLFSON, D. B. (2006). A formal test for the stationarity of the incidence rate using data from a prevalent cohort study with follow-up. *Lifetime Data Anal*. **12** 267–284. MR2328577

ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. and KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York. MR1198884

ASGHARIAN, M., M'LAN, C. E. and WOLFSON, D. B. (2002). Length-biased sampling with right censoring: An unconditional approach. *J. Amer. Statist. Assoc*. **97** 201–209. MR1947280

BARTLETT, M. S. (1937). Some examples of statistical methods of research in agriculture and applied biology. *J. Roy. Statist. Soc. Ser. B* **4** 137–183.

BEGUN, J. M., HALL, W. J., HUANG, W.-M. and WELLNER, J. A. (1983). Information and asymptotic efficiency in parametric–nonparametric models. *Ann. Statist*. **11** 432–452. MR0696057

BHATTACHARYA, P. K., CHERNOFF, H. and YANG, S. S. (1983). Nonparametric estimation of the slope of a truncated regression. *Ann. Statist*. **11** 505–514. MR0696063

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Max-

Yifei Sun is Assistant Professor, Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, New York 10032, USA (e-mail: ys3072@cumc.columbia.edu). Jing Qin is Mathematical Statistician, Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA (e-mail: jingqin@niaid.nih.gov). Chiung-Yu Huang is Professor, Department of Epidemiology and Biostatistics, School of Medicine, University of California, San Francisco, San Francisco, California 94158, USA (e-mail: ChiungYu.Huang@ucsf.edu).

imum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

HUANG, C.-Y., NING, J. and QIN, J. (2015). Semiparametric likelihood inference for left-truncated and right-censored data. *Biostatistics* **16** 785–798. MR3449843

HUANG, C.-Y. and QIN, J. (2012). Composite partial likelihood estimation under length-biased sampling, with application to a prevalent cohort study of dementia. *J. Amer. Statist. Assoc.* **107** 946–957. MR3010882

HYDE, J. (1977). Testing survival under right censoring and left truncation. *Biometrika* **64** 225–230. MR0494775

KENDALL, M. and GIBBONS, J. D. (1990). *Rank Correlation Methods*, 5th ed. Edward Arnold, London. MR1079065

LANCASTER, T. (1990). *The Econometric Analysis of Transition Data*. *Econometric Society Monographs* **17**. Cambridge Univ. Press, Cambridge. MR1167199

LUO, X. and TSAI, W. Y. (2009). Nonparametric estimation for right-censored length-biased data: A pseudo-partial likelihood approach. *Biometrika* **96** 873–886. MR2767276

LYNDEN-BELL, D. (1971). Article navigation a method of allowing for known observational selection in small samples applied to 3CR quasars. *Mon. Not. R. Astron. Soc.* **155** 95–118.

MARTIN, E. C. and BETENSKY, R. A. (2005). Testing quasi-independence of failure and truncation times via conditional Kendall's tau. *J. Amer. Statist. Assoc.* **100** 484–492. MR2160552

MCDOWELL, I., HILL, G. and LINDSAY, J. (2001). An overview of the Canadian study of health and aging. *Int. Psychogeriatr.* **13** 1–18.

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

NING, J., QIN, J. and SHEN, Y. (2014). Score estimating equations from embedded likelihood functions under accelerated failure time model. *J. Amer. Statist. Assoc.* **109** 1625–1635. MR3293615

OAKES, D. (2008). On consistency of Kendall's tau under censoring. *Biometrika* **95** 997–1001. MR2461227

ORCHARD, T. and WOODBURY, M. A. (1972). A missing information principle: Theory and applications. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (*Univ. California, Berkeley, Calif.*, 1970/1971), *Vol. I*: *Theory of Statistics* 697–715. Univ. California Press, Berkeley, CA. MR0400516

QIN, J., NING, J., LIU, H. and SHEN, Y. (2011). Maximum likelihood estimations and EM algorithms with length-biased data. *J. Amer. Statist. Assoc.* **106** 1434–1449. MR2896847

SHEN, Y., NING, J. and QIN, J. (2017). Nonparametric and semiparametric regression estimation for length-biased survival data. *Lifetime Data Anal.* **23** 3–24. MR3601682

TSAI, W.-Y. (1990). Testing the assumption of independence of truncation time and failure time. *Biometrika* **77** 169–177. MR1049418

TSAI, W. Y. (2009). Pseudo-partial likelihood for proportional hazards models with biased-sampling data. *Biometrika* **96** 601–615. MR2538760

TSAI, W.-Y., JEWELL, N. P. and WANG, M.-C. (1987). A note on the product-limit estimator under right censoring and left truncation. *Biometrika* **74** 883–886.

TURNBULL, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. Roy. Statist. Soc. Ser. B* **38** 290–295. MR0652727

VARDI, Y. (1989). Multiplicative censoring, renewal processes, deconvolution and decreasing density: Nonparametric estimation. *Biometrika* **76** 751–761. MR1041420

VARDI, Y. and ZHANG, C.-H. (1992). Large sample study of empirical distributions in a random-multiplicative censoring model. *Ann. Statist.* **20** 1022–1039. MR1165604

WANG, M.-C. (1991). Nonparametric estimation from cross-sectional survival data. *J. Amer. Statist. Assoc.* **86** 130–143. MR1137104

WANG, M.-C. (1996). Hazards regression analysis for length-biased data. *Biometrika* **83** 343–354. MR1439788

WANG, M.-C., BROOKMEYER, R. and JEWELL, N. P. (1993). Statistical models for prevalent cohort data. *Biometrics* **49** 1–11. MR1221402

YATES, F. (1933). The analysis of replicated experiments when the field results are incomplete. *Emp. J. Exp. Agric.* **1** 129–142.

Statistical Science
2018, Vol. 33, No. 2, 277–284
https://doi.org/10.1214/17-STS639

Marie-France Bru and Bernard Bru on Dice Games and Contracts

Glenn Shafer

Abstract. This note introduces Marie-France and Bernard Bru's forthcoming book on the history of probability, especially its chapter on dice games, translated in this issue of *Statistical Science*, and its commentary on the history of fair price in the settlement of contracts.

As the Brus remind us, the traditions of counting chances in dice games and estimating fair price came together in the correspondence between Pascal and Fermat in 1654. To solve the problem of dividing the stakes in a prematurely halted game, Fermat used combinatorial principles that had been used for centuries to analyze dice games, while Pascal used principles that had been proposed in previous centuries by students of commercial arithmetic.

Key words and phrases: Dice games, emergence of probability, *De vetula*, expectation.

REFERENCES

[1] AL-KADI, I. A. Origins of cryptology: The Arab contributions. *Cryptologia* **16** 97–126.

[2] AL KHALILI, J. (2011). *The House of Wisdom*: *How Arabic Science Saved Ancient Knowledge and Gave Us the Renaissance*. Penguin, New York.

[3] AL KINDI (2004). *Al-Kindi's Treatise on Cyptanalysis*. KFCRIS & KACST, Riyadh. Book One of the Series on Arabic Origins of Cryptology.

[4] BELLHOUSE, D. R. (2000). *De vetula*: A medieval manuscript containing probability calculations. *Int. Stat. Rev.* **68** 123–136.

[5] BERNOULLI, J. (1713). *Ars Conjectandi*, *Opus Posthumum*. *Accedit Tractatus de Seriebus Infinitis et Epistola Gallice Scripta de Ludo Pilae Reticularis*. Impensis Thurnisiorum fratrum, Basel. French translation by Jean Peyroux, Paris, A. Blanchard, 1998. English translation with notes by Edith Dudley Sylla, Johns Hopkins University Press, Baltimore, MD, 2006.

[6] BERNOULLI, J. (1969–1999). *Die Werke Von Jakob Bernoulli*. Birkhäuser, Basel. 6 volumes.

[7] BRU, B., BRU, M.-F. and BIENAYMÉ, O. (1997). La statistique critiquée par le calcul des probabilités: Deux manuscrits inédits d'Irenée Jules Bienaymé. *Rev. Histoire Math.* **3** 137–239.

[8] BRU, B., BRU, M.-F. and CHUNG, K. L. (1999). Borel et la martingale de Saint-Pétersbourg. *Rev. Histoire Math.* **5** 181–247. Translated into English as "Borel and the Saint-Petersburg paradox" in *Electronic Journal for History of Probability and Statistics*, 5(1), 2009.

[9] CARDANO, G. (1539). *Practica arithmeticæ, & mensurandi singularis*. Castellioneus, Milan.

[10] COUMET, E. (1965). Le problème des partis avant Pascal. *Arch. Int. Hist. Sci.* **18** 245–272. Reprinted on pages 73–95 of *Œuvres d'Ernest Coumet*, volume 1, *Presses universitaire de Franche-Comté, 2016*.

[11] DAVID, F. N. (1962). *Games, Gods, and Gambling*: *The Origins and History of Probability and Statistical Ideas from the Earliest Times to the Newtonian Era*. Griffin, London.

[12] EDWARDS, A. W. F. (2002). *Pascal's Arithmetic Triangle*: *The Story of a Mathematical Idea*, 2nd ed.

[13] FRANCI, R. (2002). Una soluzione esatta del problema delle parti in un manoscritto della prima metà del Quattrocento. *Boll. Stor. Sci. Mat.* **XXII** 253–266.

[14] FRANKLIN, J. (2001). *The Science of Conjecture*: *Evidence and Probability Before Pascal*. Johns Hopkins Univ. Press, Baltimore, MD. Second edition 2015.

[15] HACKING, I. (1975). *The Emergence of Probability*. Cambridge Univ. Press, New York. Second edition 2006.

[16] HALD, A. (1990). *A History of Probability and Statistics and Their Applications Before* 1750. Wiley, New York.

[17] HØYRUP, J. (2014). Fibonacci—Protagonist or witness? Who taught Catholic Christian Europe about Mediterranean commercial arithmetic? *J. Transcult. Mediev. Stud.* **1** 219–247.

[18] HUYGENS, C. (1657). De Ratiociniis in Ludo Aleae. In *Exercitationum Mathematicarum*, *Liber V* (F. V. Schooten, ed.) 511–534. Elsevier, Leiden. The Dutch original, written in 1656, was published in Amsterdam in 1659 and reprinted along with a translation into French in Huygens's *Œuvres*, volume 14, pages 1–91.

Glenn Shafer is University Professor, Rutgers University, 1 Washington Park, Newark, New Jersey 07102, USA (e-mail: gshafer@rutgers.edu).

[19] KENDALL, M. G. (1956). Studies in the history of probability and statistics: II. The beginnings of a probability calculus. *Biometrika* **43** 1–14. Reprinted in Pearson and Kendall, *Studies in the History of Statistics and Probability* I: 19–34.

[20] KLOPSCH, P. (1967). *Pseudo-Ovidius De Vetula*: *Untersuchung und Text*. Brill, Leiden.

[21] LACROIX, S. F. (1816). *Traité élémentaire du Calcul des Probabilités*. Courcier, Paris. Second edition 1822.

[22] LAPLACE, P. S. (1814). *Essai Philosophique sur les Probabilités*, 1st ed. Courcier, Paris. The fifth and definitive edition appeared in 1825 and was reprinted in 1986 (Christian Bourgois, Paris) with a commentary by Bernard Bru. Multiple English translations have appeared.

[23] LEWIS, D. L. (2008). *God's Crucible*: *Islam and the Making of Europe*, 570–1215. Norton, New York.

[24] LUBBOCK, J. W. and DRINKWATER-BETHUNE, J. E. (1830). *On Probability*. Baldwin & Craddock, London.

[25] MEUSNIER, N. (2007). Le problème des partis bouge... de plus en plus. *Electron. J. Hist. Probab. Stat.* **3**.

[26] PACIOLI, L. (1494). *Summa de arithmetica*, *geometria*, *proportioni et proportionalità*. Paganino de Paganini, Venice.

[27] PASCAL, B. (1665). *Traité du Triangle Artimétique et Traités Connexes*. Desprez, Paris.

[28] ROBATHAN, D. M. (1968). *The Pseudo-Ovidian De Vetula*: *Text*, *Introduction*, *and Notes*. Hakkert, Amsterdam.

[29] SALIBA, G. (2007). *Islamic Science and the Making of the European Renaissance*. MIT Press, Cambridge, MA.

[30] SCHNEIDER, I. (1985). Luca Pacioli und das Teilungsproblem: Hintergrund und Lösungsversuch. In *Mathemata* (M. Folkerts and U. Lindgren, eds.). Steiner, Wiesbaden and Stuttgart.

[31] SCHNEIDER, I., ed. (1988). *Die Entwicklung der Wahrscheinlichkeitstheorie Von Den Anfängen Bis* 1933: *Einführungen und Texte*. Wissenschaftliche Buchgesellschaft, Darmstadt.

[32] SCHNEIDER, I. (1988). The market place and games of chance in the fifteenth and sixteenth centuries. In *Mathematics from Manuscript to Print*, 1300–1600 (C. Hay, ed.) 220–235. Oxford.

[33] SHAFER, G. and VOVK, V. (2001). *Probability and Finance*: *It's Only a Game!* Wiley, New York.

[34] SHAFER, G. and VOVK, V. (2006). The sources of Kolmogorov's *Grundbegriffe*. *Statist. Sci.* **21** 70–98. MR2275967

[35] SHAFER, G., VOVK, V. and TAKEMURA, A. (2012). Lévy's zero-one law in game-theoretic probability. *J. Theoret. Probab.* **25** 1–24. MR2886375

[36] STIGLER, S. M. (2014). Soft question, hard answer: Jacob Bernoulli's probability in historical context. *Int. Stat. Rev.* **82** 1–16.

[37] SWAN, E. J. (2000). *Building the Global Market*: *A* 4000 *Year History of Derivatives*. Kluwer Law International, The Hague.

[38] SYLLA, E. (2003). Business ethics, commercial mathematics, and the origins of mathematical probability. *Hist. Polit. Econ.* **35** 309–337.

[39] TARTAGLIA, N. (1556). *General Trattato di Numeri e Misure*. C. Troiano dei Navo, Venice.

[40] TODHUNTER, I. (1865). *A History of the Mathematical Theory of Probability from the Time of Pascal to That of Laplace*. Macmillan, London.

[41] TOTI RIGATELLI, L. (1985). Il 'problema delle parti' in manoscritti del XIV e XV secolo. In *Mathemata*: *Festschrift Für Helmuth Gericke* (M. Folkerts and U. Lindgren, eds.) 229–236. Steiner, Stuttgart.

Statistical Science
2018, Vol. 33, No. 2, 285–297
https://doi.org/10.1214/17-STS640

Dice Games

Marie-France Bru and Bernard Bru

Abstract. Translated from the French by Glenn Shafer, the French text will appear as Chapter 1 of Volume 2 of *Les jeux de l'infini et du hasard*, by Marie-France and Bernard Bru, to be published by the Presses universitaires de Franche-Comté. The translation is published here with the permission of the publisher and the surviving author. The text has been edited to omit most references to other parts of the book. The authors extensive notes, which provide many additional references and historical details, have also been omitted.

Key words and phrases: *De vetula*, dice games, history of probability, Huygens, Jacob Bernoulli, Laplace, Montmort, normal approximation.

REFERENCES

[1] BERNELIN (STUDENT OF GERBERT D'AURILLAC C. 1000) (1999). *Libre d'abaque*. Princi Néguer, Pau. Latin and French text established by B. Bakhouche with complementary notes by J. Cassinet.

[2] ALLARD, A. (1992). *Muhammad Ibn Musa Al-Khwarizmi. Le Calcul Indien* (*Algorismus*). A. Blanchard, Paris. Edition, with French translation and commentary, of the oldest Latin versions going back to the XIIth century.

[3] ALLARD, A. (1994). *L'enseignement du calcul arithmétique à partir des XIIe et XIIIe siècles*: *L'exemple de la multiplication*. Publications de l'Institut d'Etudes Médiévales, Université Catholique de Louvain, No. 16.

[4] ALLARD, A. (1997). L'influence des mathématiques arabes dans l'occident médiéval. In *Histoire des Sciences Arabes* (R. Rashed and R. Morelon, eds.) **2** 199–230. Seuil, Paris.

[5] BEAUJOUAN, G. (1991). *Par Raison de Nombres*, *L'art du Calcul et des Savoirs Scientifiques Médiévaux*. Gower, Aldershot.

[6] BEAUJOUAN, G., CARBONNE, P., CASSINET, J. et al. (1995). *Huit Siècles de Mathématiques en Occitanie*. CIHSO, Toulouse. Reprinted by Monein, Pyrémonde, 2008.

[7] BELLHOUSE, D. R. (2011). *Abraham De Moivre*: *Setting the Stage for Classical Probability and Its Applications*. CRC, Boca Raton, FL.

[8] BERNOULLI, J. (1713). *Ars conjectandi*, *opus posthumum*. *Accedit Tractatus de seriebus infinitis et epistola Gallice scripta De ludo pilae reticularis*. Impensis Thurnisiorum Fratrum, Basel. French translation by J. Peyroux, A. Blanchard, Paris, 1998. English translation with notes by E. D. Sylla, Johns Hopkins Univ. Press, Baltimore, 2006.

[9] BERNOULLI, N. (1709). De usu artis conjectandi in jure. Doctoral thesis, Basel. Reprinted in *Die Werke von Jakob Bernoulli* **3** (1975) 287–326. French translation with notes by N. Meusnier, Paris, CAMS, 1992.

[10] BIGGS, N. L. (1979). The roots of combinatorics. *Historia Mathematica* **6** 109–136. DOI:10.1016/0315-0860(79)90074-0.

[11] BOETHIUS, A. M. S. (1995). *De Institutione Arithmetica*. Les Belles Lettres, Paris. French translation and notes by J.-Y. Guillaumin.

[12] BOREL, É. (1909). *Éléments de la Théorie des Probabilités*. Hermann, Paris. Further editions in 1910 and 1924, and 1950. The 1950 edition, published by Albin Michel, was translated into English by J. E. Freund and published by Prentice Hall, Englewood Cliffs, 1965.

[13] BOREL, É. (1920). Radioactivité, probabilité et déterminisme. *Rev. Mois* **21** 33–40. Reproduced as note I of the 1924 edition of [12] and in volume 4 of [14], pages 2189–2196.

[14] BOREL, É. (1972). *Œuvres D'Émile Borel*. CNRS, Paris. 4 volumes.

[15] COURNOT, A. (1828). De la théorie des probabilités considérée comme la matière d'un enseignement. *Le Lycée* **2** 447–453. Reprinted in Volume 11 of [17].

[16] COURNOT, A. (1843). *Exposition de la Théorie des Chances et des Probabilités*. Hachette, Paris. Reprinted with notes and index as Volume 1 of [17].

[17] COURNOT, A. (1973–2010). *Œuvres Complètes*. Vrin and Presses Universitaires de Franche-Comté, Paris and Besançon. 11 titles in 13 volumes.

[18] DE MOIVRE, A. (1710–1712) De mensura sortis, seu de probabilitate eventuum in ludis a casu fortuito pendentibus. *Philos*. *Trans*. *R*. *Soc*. *Lond*. **27** 213–264.

[19] DE MOIVRE, A. (1718). *The Doctrine of Chances*: *Or*, *A Method of Calculating the Probability of Events in Play*. W. Pearson, London. Later editions in 1738 and 1756.

Marie-France Bru was formerly at UFR de mathématiques, Université Paris Diderot, France. She died on January 30, 2012. Bernard Bru is retired from MAP5, Université Paris Descartes, 45 rue des Saints-Pères, 75006 Paris, France (e-mail: leslogesb@gmail.com).

[20] DE MONTMORT, P. R. (1708). *Essay d'analyse sur les jeux de hazard*. Quillau, Paris. Published anonymously. Second edition in 1713.

[21] DJEBBAR, A. (1985). *L'analyse combinatoire au Maghreb*: *L'exemple d'Ibn Mun'im* (*XIIe–XIIIe s.*). Publications Mathématiques D'Orsay, number 85-01.

[22] GUILLAUMIN, J.-Y. (2012). Boethius's *De institutione arithmetica* and its influence on posterity. In *A Companion to Boethius in the Middle Ages* (N. H. Kaylor and P. E. Phillips, eds.) 135–162. Brill, Leiden.

[23] HAGSTROEM, K. G. (1932). *Les préludes antiques de la théorie des probabilités*. Fritze, Stockholm.

[24] HALD, A. (1990). *A History of Probability and Statistics and Their Applications Before* 1750. Wiley, New York.

[25] HUYGENS, C. (1657). De ratiociniis in ludo aleae. In *Exercitationum Mathematicarum*, *Liber V* (F. van Schooten, ed.) 511–534. Elsevier, Leiden. The Dutch original, written in 1656, was published in Amsterdam in 1660 and reprinted along with a translation into French in Huygens's *Œuvres*, volume 14, pages 1–91.

[26] LAPLACE, P.-S. (1878–1912). *Œuvres complètes*. Gauthier-Villars, Paris. 14 volumes.

[27] LAPLACE, P. S. (1812). *Théorie Analytique des Probabilités*. Courcier, Paris. Later editions in 1814, 1820, and 1825. The 1825 edition was reprinted in Volume 7 of [26].

[28] MONTUCLA, J.-É. (1802). *Histoire des mathématiques*. H. Agasse, Paris. Completed and published by J. de Lalande, 4 volumes.

[29] PACIOLI, L. (1498). *Divina Proportione*. Manuscript, Milan. First published by Paganini, Venice, 1509. Facsimile and French translation by G. Duchesne and M. Giraud, with a historical introduction by M.-T. Sarrade, Librairie du Compagnonnage, Paris, 1988.

[30] STIGLER, S. M. (2014). Soft question, hard answer: Jacob Bernoulli's probability in historical context. *Int. Stat. Rev.* **82** 1–16.

[31] STIGLER, S. M. (2016). *The Seven Pillars of Statistical Wisdom*. Harvard Univ. Press, Cambridge, MA.

INSTITUTE OF MATHEMATICAL STATISTICS

(Organized September 12, 1935)

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

IMS OFFICERS

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

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

Past President: Jon Wellner, Department of Statistics, University of Washington, Seattle, Washington 98195-4322, USA

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

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

Program Secretary: Judith Rousseau, Université Paris Dauphine, Place du Maréchal DeLattre de Tassigny, 75016 Paris, France

IMS EDITORS

The Annals of Statistics. *Editors*: Edward I. George, Department of Statistics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA; Tailen Hsing, Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109-1107, USA

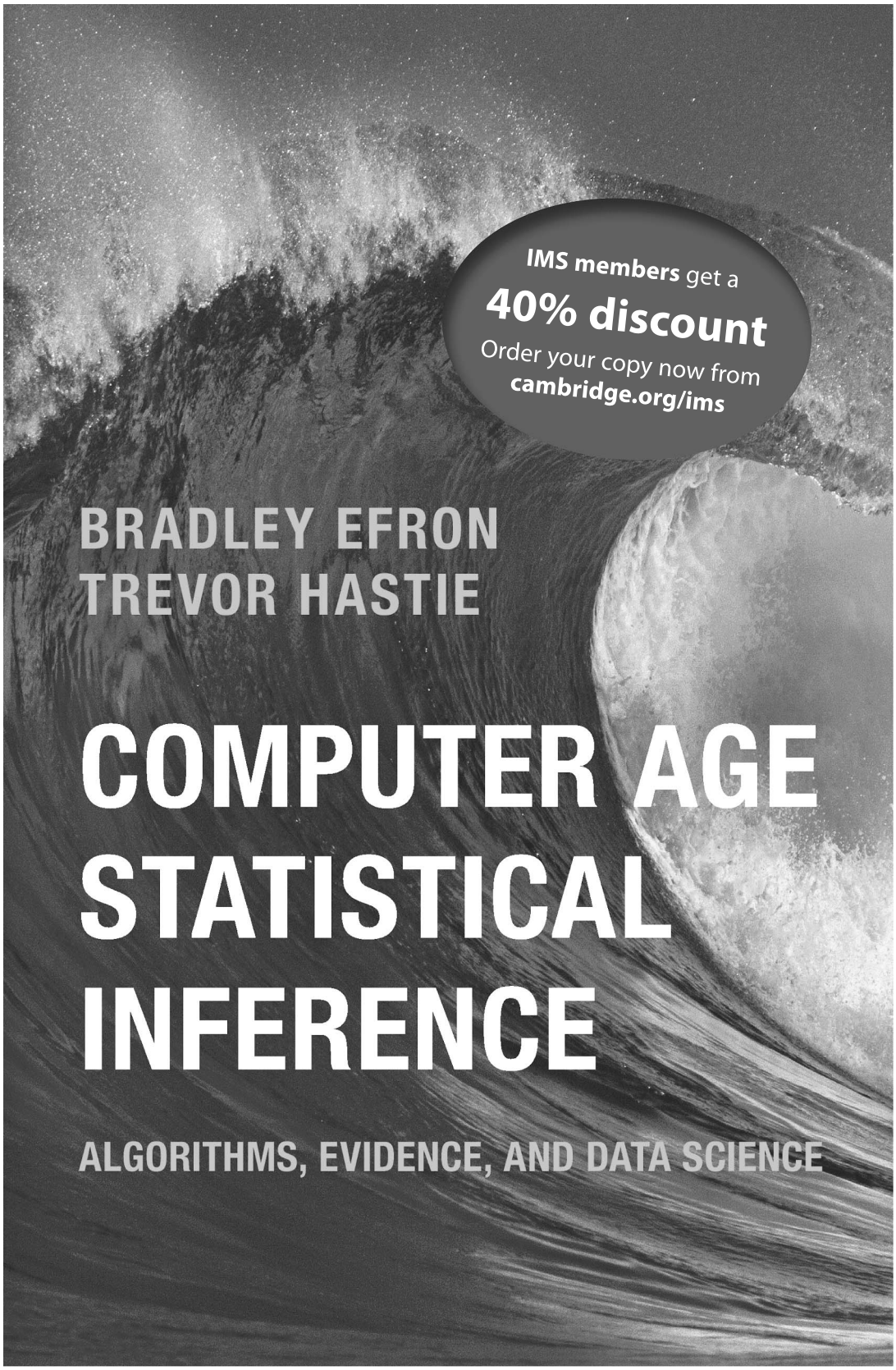
The Annals of Applied Statistics. *Editor-in-Chief*: Tilmann Gneiting, Heidelberg Institute for Theoretical Studies, HITS gGmbH, Schloss-Wolfsbrunnenweg 35, 69118 Heidelberg, Germany

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

The Annals of Applied Probability. *Editor*: Bálint Tóth, School of Mathematics, University of Bristol, University Walk, BS8 1TW, Bristol, United Kingdom, and Alfréd Rényi Institute of Mathematics, Hungarian Academy of Sciences, Budapest, Hungary

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

The IMS Bulletin. *Editor*: Vlada Limic, DR2, CNRS, Université Paris Sud 11, UMR 8628, Département de Mathématiques, 91405 Orsay, France


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