STATISTICAL SCIENCE

Volume 39, Number 2 May 2024

Statistical Science **[ISSN 0883-4237 (print); ISSN 2168-8745 (online)]**, Volume 39, Number 2, May 2024. Published quarterly by the Institute of Mathematical Statistics, 9760 Smith Road, Waite Hill, Ohio 44094, USA. Periodicals postage paid at Cleveland, Ohio and at additional mailing offices. **POSTMASTER:** Send address changes to *Statistical Science*, Institute of Mathematical Statistics, Dues and Subscriptions Office, PO Box 729, Middletown, Maryland 21769, USA.

Statistical Science Volume 39, Number 2 (221–373) May 2024

Volume 39 **Number 2** **May 2024**

Statistical Science
2024, Vol. 39, No. 2, 221–242
https://doi.org/10.1214/24-STS927

30 Years of Synthetic Data

Jörg Drechsler and Anna-Carolina Haensch

Abstract. The idea to generate synthetic data as a tool for broadening access to sensitive microdata has been proposed for the first time three decades ago. While first applications of the idea emerged around the turn of the century, the approach really gained momentum over the last ten years, stimulated at least in parts by some recent developments in computer science. We consider the 30th jubilee of Rubin's seminal paper on synthetic data (*J. Off. Stat.* **9** (1993) 462–468) as an opportunity to look back at the historical developments but also to offer a review of the diverse approaches and methodological underpinnings proposed over the years. We will also discuss the various strategies that have been suggested to measure the utility and remaining risk of disclosure of the generated data.

Key words and phrases: Access, confidentiality, data generation, disclosure, dissemination, privacy.

REFERENCES

[1] ABADI, M., CHU, A., GOODFELLOW, I., MCMAHAN, H. B., MIRONOV, I., TALWAR, K. and ZHANG, L. (2016). Deep learning with differential privacy. In *Proceedings of the* 2016 *ACM SIGSAC Conference on Computer and Communications Security* 308–318. ACM, Vienna, Austria.

[2] ABOWD, J., ASHMEAD, R., CUMINGS-MENON, R., GARFINKEL, S., HEINECK, M., HEISS, C., JOHNS, R., KIFER, D., LECLERC, P. et al. (2022). The 2020 census disclosure avoidance system TopDown algorithm. *Harv. Data Sci. Rev.* **2**. Special Issue.

[3] ABOWD, J., ASHMEAD, R., SIMSON, G., KIFER, D., LECLERC, P., MACHANAVAJJHALA, A. and SEXTON, W. (2019). Census topdown: Differentially private data, incremental schemas, and consistency with public knowledge. U.S. Census Bureau, Washington, DC.

[4] ABOWD, J. M., STINSON, M. and BENEDETTO, G. (2006). Final report to the social security administration on the SIPP/SSA/IRS public use file project Technical report, longitudinal employer–household dynamics program. U.S. Bureau of the Census, Washington, DC.

[5] ABOWD, J. M. and VILHUBER, L. (2008). How protective are synthetic data? In *Privacy in Statistical Databases* (J. Domingo-Ferrer and Y. Saygın, eds.) **5262** 239–246. Springer, Berlin.

[6] ABOWD, J. M. and WOODCOCK, S. D. (2001). Disclosure limitation in longitudinal linked data. In *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies* (P. Doyle, J. Lane, L. Zayatz and J. Theeuwes, eds.) 215–277. North-Holland, Amsterdam.

[7] ABOWD, J. M. and WOODCOCK, S. D. (2004). Multiply-imputing confidential characteristics and file links in longitudinal linked data. In *Privacy in Statistical Databases* (J. Domingo-Ferrer and V. Torra, eds.) 290–297. Springer, New York.

[8] ALAM, M. J., DOSTIE, B., DRECHSLER, J. and VILHUBER, L. (2020). Applying data synthesis for longitudinal business data across three countries. *Statist. Transition New Series* **21** 212–236.

[9] ALLKEN, V., HANDEGARD, N. O., ROSEN, S., SCHREYECK, T., MAHIOUT, T. and MALDE, K. (2018). Fish species identification using a convolutional neural network trained on synthetic data. *ICES J. Mar. Sci.* **76** 342–349.

[10] AN, D. and LITTLE, R. J. A. (2007). Multiple imputation: An alternative to top coding for statistical disclosure control. *J. Roy. Statist. Soc. Ser. A* **170** 923–940. MR2408985 https://doi.org/10.1111/j.1467-985X.2007.00492.x

[11] ARJOVSKY, M., CHINTALA, S. and BOTTOU, L. (2017). Wasserstein GAN. Available at arXiv:1701.07875 [stat.ML].

[12] ARNOLD, C. and NEUNHOEFFER, M. (2020). Really useful synthetic data–a framework to evaluate the quality of differentially private synthetic data. Available at arXiv:2004.07740.

[13] AUSTRALIAN BUREAU OF STATISTICS (2021). Methodological news, Dec 2021. Available at https://www.abs.gov.au/statistics/research/methodological-news-dec-2021. Last accessed on 2022-05-17.

[14] BAO, E., XIAO, X., ZHAO, J., ZHANG, D. and DING, B. (2021). Synthetic data generation with differential privacy via Bayesian networks. *J. Priv. Confid.* **11**.

[15] BAOWALY, M. K., LIN, C.-C., LIU, C.-L. and CHEN, K.-T. (2019). Synthesizing electronic health records using improved generative adversarial networks. *J. Amer. Med. Inform. Assoc.* **26** 228–241.

[16] BARAK, B., CHAUDHURI, K., DWORK, C., KALE, S., MCSHERRY, F. and TALWAR, K. (2007). Privacy, accuracy, and con-

Jörg Drechsler is head of the Department for Statistical Methods at the Institute for Employment Research, Germany, Professor, Ludwig-Maximilians-Universität, Munich, Germany, and Associate Research Professor, Joint Program in Survey Methodology, University of Maryland, College Park, Maryland 20742, USA (e-mail: joerg.drechsler@iab.de). Anna-Carolina Haensch is Lecturer, Ludwig-Maximilians-Universität, Munich, Germany, and Assistant Research Professor at the Joint Program in Survey Methodology, University of Maryland, College Park, Maryland 20742, USA (e-mail: anna-carolina.haensch@stat.uni-muenchen.de).

sistency too: A holistic solution to contingency table release. In *Proceedings of the Twenty-Sixth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems—PODS '07* 273–282. ACM, Beijing, China.

[17] BARRIENTOS, A. F., BOLTON, A., BALMAT, T., REITER, J. P., DE FIGUEIREDO, J. M., MACHANAVAJJHALA, A., CHEN, Y., KNEIFEL, C. and DELONG, M. (2018). Providing access to confidential research data through synthesis and verification: An application to data on employees of the U.S. federal government. *Ann. Appl. Stat.* **12** 1124–1156. MR3834297 https://doi.org/10.1214/18-AOAS1194

[18] BEAULIEU-JONES, B. K., WU, Z. S., WILLIAMS, C., LEE, R., BHAVNANI, S. P., BYRD, J. B. and GREENE, C. S. (2019). Privacy-preserving generative deep neural networks support clinical data sharing. *Circ. Cardiovasc. Qual. Outcomes* **12** e005122. https://doi.org/10.1161/CIRCOUTCOMES.118.005122

[19] BENEDETTO, G., STANLEY, J. C., TOTTY, E. et al. (2018). The creation and use of the SIPP synthetic beta version 7.0.

[20] BLUM, A., LIGETT, K. and ROTH, A. (2013). A learning theory approach to noninteractive database privacy. *J. ACM* **60** Art. 12, 25. MR3060810 https://doi.org/10.1145/2450142.2450148

[21] BONNÉRY, D., FENG, Y., HENNEBERGER, A. K., JOHNSON, T. L., LACHOWICZ, M., ROSE, B. A., SHAW, T., STAPLETON, L. M., WOOLLEY, M. E. et al. (2019). The promise and limitations of synthetic data as a strategy to expand access to state-level multi-agency longitudinal data. *J. Res. Educ. Eff.* **12** 616–647.

[22] BOWEN, C. M., BRYANT, V., BURMAN, L., CZAJKA, J., KHITATRAKUN, S., MACDONALD, G., MCCLELLAND, R., MUCCIOLO, L., PICKENS, M. et al. (2022). Synthetic individual income tax data: Methodology, utility, and privacy implications. In *International Conference on Privacy in Statistical Databases* 191–204. Springer, Berlin.

[23] BOWEN, C. M., BRYANT, V., BURMAN, L., KHITATRAKUN, S., MCCLELLAND, R., STALLWORTH, P., UEYAMA, K. and WILLIAMS, A. R. (2020). A synthetic supplemental public use file of low-income information return data: Methodology, utility, and privacy implications. In *International Conference on Privacy in Statistical Databases* 257–270. Springer, Berlin.

[24] BOWEN, C. M. and LIU, F. (2020). Comparative study of differentially private data synthesis methods. *Statist. Sci.* **35** 280–307. MR4106606 https://doi.org/10.1214/19-STS742

[25] BOWEN, C. M., LIU, F. and SU, B. (2021). Differentially private data release via statistical election to partition sequentially. *Metron* **79** 1–31. MR4239846 https://doi.org/10.1007/s40300-021-00201-0

[26] BOWEN, C. M. and SNOKE, J. (2021). Comparative study of differentially private synthetic data algorithms from the NIST PSCR differential privacy synthetic data challenge. *J. Priv. Confid.* **11**. https://doi.org/10.29012/jpc.748

[27] BURMAN, L. E., ENGLER, A., KHITATRAKUN, S., NUNNS, J. R., ARMSTRONG, S., ISELIN, J., MACDONALD, G. and STALLWORTH, P. (2019). Safely expanding research access to administrative tax data: creating a synthetic public use file and a validation server Technical report, Technical report US, Internal Revenue Service.

[28] BURRIDGE, J. (2003). Information preserving statistical obfuscation. *Stat. Comput.* **13** 321–327. MR2005433 https://doi.org/10.1023/A:1025658621216

[29] CAI, K., LEI, X., WEI, J. and XIAO, X. (2021). Data synthesis via differentially private Markov random fields. *Proc. VLDB Endow.* **14** 2190–2202.

[30] CAIOLA, G. and REITER, J. P. (2010). Random forests for generating partially synthetic, categorical data. *Trans. Data Priv.* **3** 27–42. MR2725418

[31] CAMINO, R., HAMMERSCHMIDT, C. and STATE, R. (2018). Generating multi-categorical samples with generative adversarial networks. Available at arXiv:1807.01202 [cs, stat].

[32] CANO, I., LADRA, S. and TORRA, V. (2010). Evaluation of information loss for privacy preserving data mining through comparison of fuzzy partitions. In *International Conference on Fuzzy Systems* 1–8 IEEE Press, Barcelona, Spain.

[33] CHALLENGE.GOV (2019). NIST differential privacy synthetic data challenge. Available at https://www.challenge.gov/?challenge=differential-privacy-synthetic-data-challenge. Last accessed on 2022-06-08.

[34] CHAREST, A.-S. (2011). How can we analyze differentially-private synthetic datasets? *J. Priv. Confid.* **2**.

[35] CHEN, J., CHUN, D., PATEL, M., CHIANG, E. and JAMES, J. (2019). The validity of synthetic clinical data: A validation study of a leading synthetic data generator (synthea) using clinical quality measures. *BMC Med. Inform. Decis. Mak.* **19** 1–9.

[36] CHEN, Y., ELLIOT, M. and SAKSHAUG, J. (2016). A genetic algorithm approach to synthetic data production. In *Proceedings of the 1st International Workshop on AI for Privacy and Security*. 1–4.

[37] CHEN, Y., ELLIOT, M. and SMITH, D. (2018). The application of genetic algorithms to data synthesis: A comparison of three crossover methods. In *International Conference on Privacy in Statistical Databases* 160–171. Springer, Berlin.

[38] CHIEN, C.-H., WELSH, A. H. and MOORE, J. D. (2020). Synthetic business microdata: An Australian example. *J. Priv. Confid.* **10**.

[39] CHOI, E., BISWAL, S., MALIN, B., DUKE, J., STEWART, W. F. and SUN, J. (2018). Generating multi-label discrete patient records using generative adversarial networks. Available at arXiv:1703.06490 [cs].

[40] COMMISSION, E. (2022). European data strategy. Available at https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en. Last accessed on 2022-05-03.

[41] DE MONTJOYE, Y.-A., HIDALGO, C. A., VERLEYSEN, M. and BLONDEL, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Sci. Rep.* **3** 1–5.

[42] DE MONTJOYE, Y.-A., RADAELLI, L., SINGH, V. K. and PENTLAND, A. S. (2015). Identity and privacy. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science* **347** 536–539. https://doi.org/10.1126/science.1256297

[43] DE WOLF, P.-P. (2015). Public use files of EU-SILC and EU-LFS data. Joint UNECE/Eurostat work session on statistical data confidentiality Helsinki, Finland, 1–10.

[44] DENTON, E. L., CHINTALA, S., FERGUS, R. et al. (2015). Deep generative image models using a Laplacian pyramid of adversarial networks. *Adv. Neural Inf. Process. Syst.* **28**.

[45] DEPARTMENT FOR DIGITAL, CULTURE, MEDIA & SPORT (2022). National data strategy. Available at https://www.gov.uk/government/publications/uk-national-data-strategy/national-data-strategy. Last accessed on 2022-05-03.

[46] DING, B., KULKARNI, J. and YEKHANIN, S. (2017). Collecting telemetry data privately. *Adv. Neural Inf. Process. Syst.* 3571–3580.

[47] DONG, Q., ELLIOTT, M. R. and RAGHUNATHAN, T. E. (2014). A nonparametric method to generate synthetic populations to adjust for complex sampling design features. *Surv. Methodol.* **40** 29–46.

[48] DONG, Q., ELLIOTT, M. R. and RAGHUNATHAN, T. E. (2014). Combining information from multiple complex surveys. *Surv. Methodol*. **40** 347–354.

[49] DRECHSLER, J. (2010). Using support vector machines for generating synthetic datasets. In *International Conference on Privacy in Statistical Databases* 148–161. Springer, Berlin.

[50] DRECHSLER, J. (2011). *Synthetic Datasets for Statistical Disclosure Control*: *Theory and Implementation*. *Lecture Notes in Statistics* **201**. Springer, New York. MR2809912 https://doi.org/10.1007/978-1-4614-0326-5

[51] DRECHSLER, J. (2011). Improved variance estimation for fully synthetic datasets. Proceedings of the joint UNECE/EUROSTAT work session on statistical data confidentiality.

[52] DRECHSLER, J. (2012). New data dissemination approaches in old Europe—synthetic datasets for a German establishment survey. *J. Appl. Stat*. **39** 243–265. MR2879819 https://doi.org/10.1080/02664763.2011.584523

[53] DRECHSLER, J. (2018). Some clarifications regarding fully synthetic data. In *International Conference on Privacy in Statistical Databases* 109–121. Springer, Berlin.

[54] DRECHSLER, J. (2022). Challenges in measuring utility for fully synthetic data. In *International Conference on Privacy in Statistical Databases* 220–233. Springer, Berlin.

[55] DRECHSLER, J. and HU, J. (2021). Synthesizing geocodes to facilitate access to detailed geographical information in large-scale administrative data. *J. Surv. Stat. Methodol*. **9** 523–548.

[56] DRECHSLER, J. and REITER, J. P. (2008). Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. In *Privacy in Statistical Databases* (J. Domingo-Ferrer and Y. Saygin, eds.) 227–238. Springer, New York.

[57] DRECHSLER, J. and REITER, J. P. (2009). Disclosure risk and data utility for partially synthetic data: An empirical study using the German IAB establishment survey. *J. Off. Stat*. **25** 589–603.

[58] DRECHSLER, J. and REITER, J. P. (2010). Sampling with synthesis: A new approach for releasing public use census microdata. *J. Amer. Statist. Assoc*. **105** 1347–1357. Supplementary materials available online. MR2796555 https://doi.org/10.1198/jasa.2010.ap09480

[59] DRECHSLER, J. and REITER, J. P. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Comput. Statist. Data Anal*. **55** 3232–3243. MR2825406 https://doi.org/10.1016/j.csda.2011.06.006

[60] DRECHSLER, J. and REITER, J. P. (2012). Combining synthetic data with subsampling to create public use microdata files for large scale surveys. *Surv. Methodol*. **38** 73–79.

[61] DRECHSLER, J. and VILHUBER, L. (2014). Synthetic longitudinal business databases for international comparisons. In *International Conference on Privacy in Statistical Databases* 243–252. Springer, Berlin.

[62] DRECHSLER, J. and VILHUBER, L. (2014). A first step towards a German SynLBD: Constructing a German longitudinal business database. *Stat. J. IAOS* **30** 137–142.

[63] DUNCAN, G. T., ELLIOT, M. and SALAZAR-GONZÁLEZ, J.-J. (2011). *Statistical Confidentiality*: *Principles and Practice*. *Statistics for Social and Behavioral Sciences*. Springer, New York. MR3186259 https://doi.org/10.1007/978-1-4419-7802-8

[64] DWORK, (2008). Differential privacy: A survey of results. In *Theory and Applications of Models of Computation* (M. Agrawal, D. Du, Z. Duan and A. Li, eds.) 1–19. Springer, Berlin.

[65] DWORK, C., MCSHERRY, F., NISSIM, K. and SMITH, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*. *Lecture Notes in Computer Science* **3876** 265–284. Springer, Berlin. MR2241676 https://doi.org/10.1007/11681878_14

[66] DWORK, C. and ROTH, A. (2013). The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci*. **9** 211–487. MR3254020 https://doi.org/10.1561/0400000042

[67] ENO, J. and THOMPSON, C. W. (2008). Generating synthetic data to match data mining patterns. *IEEE Internet Comput*. **12** 78–82.

[68] ERLINGSSON, Ú., PIHUR, V. and KOROLOVA, A. (2014). Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the* 2014 *ACM SIGSAC Conference on Computer and Communications Security* 1054–1067.

[69] ESTEBAN, C., HYLAND, S. L. and RÄTSCH, G. (2017). Real-valued (medical) time series generation with recurrent conditional gans. Available at arXiv:1706.02633.

[70] EUROPEAN COMMISSION (2024). How contact tracing and warning apps helped during the COVID-19 pandemic. Available at https://commission.europa.eu/strategy-and-policy/coronavirus-response/travel-during-coronavirus-pandemic/contact-tracing-and-warning-apps-during-covid-19_en. Last accessed on 2024-01-12.

[71] EUROSTAT (2022). Statistics on income and living conditions. Available at https://ec.europa.eu/eurostat/web/microdata/statistics-on-income-and-living-conditions. Last accessed on 2022-05-16.

[72] FOOTE, A. D., MACHANAVAJJHALA, A. and MCKINNEY, K. (2019). Releasing earnings distributions using differential privacy: Disclosure avoidance system for post-secondary employment outcomes (PSEO). *J. Priv. Confid*. **9**.

[73] FORBES, S. and ZEALAND, S. N. (2008). Raising statistical capability: Statistics New Zealand's contribution. In *Government Statistical Offices and Statistical Literacy* 1–18.

[74] FRID-ADAR, M., KLANG, E., AMITAI, M., GOLDBERGER, J. and GREENSPAN, H. (2018). Synthetic data augmentation using GAN for improved liver lesion classification. In 2018 *IEEE* 15*th International Symposium on Biomedical Imaging* (*ISBI* 2018) 289–293.

[75] FRIGERIO, L., DE OLIVEIRA, A. S., GOMEZ, L. and DUVERGER, P. (2019). Differentially private generative adversarial networks for time series, continuous, and discrete open data.

[76] GABOARDI, M., ARIAS, E. J. G., HSU, J., ROTH, A. and WU, Z. S. (2014). Dual query: Practical private query release for high dimensional data. In *Proceedings of the* 31*st International Conference on Machine Learning* (E. P. Xing and T. Jebara, eds.). *Proceedings of Machine Learning Research* **32** 1170–1178. PMLR, Bejing, China.

[77] GAL, Y., CHEN, Y. and GHAHRAMANI, Z. (2015). Latent Gaussian processes for distribution estimation of multivariate categorical data. In *International Conference on Machine Learning* 645–654. PMLR.

[78] GHORBANI, A., NATARAJAN, V., COZ, D. and LIU, Y. (2020). DermGAN: Synthetic generation of clinical skin images with pathology. In *Proceedings of the Machine Learning for Health NeurIPS Workshop* (A. V. Dalca, M. B. A. McDermott, E. Alsentzer, S. G. Finlayson, M. Oberst, F. Falck and B. Beaulieu-Jones, eds.). *Proceedings of Machine Learning Research* **116** 155–170. PMLR.

[79] GOLDSTEIN, R., WOOLLEY, M. E., STAPLETON, L. M., BONNÉRY, D., LACHOWICZ, M., SHAW, T. V., HENNEBERGER, A. K., JOHNSON, T. L. and FENG, Y. (2020). Expanding MLDS data access and research capacity with synthetic data sets.

[80] GOMATAM, S. and KARR, A. F. (2003). Distortion measures for categorical data swapping Technical report, National Institute of Statistical Sciences, Research Triangle Park, NC.

[81] GONCALVES, A., RAY, P., SOPER, B., STEVENS, J., COYLE, L. and SALES, A. P. (2020). Generation and evaluation of synthetic patient data. *BMC Med. Res. Methodol.* **20** 1–40.

[82] GOODFELLOW, I. J., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A. and BENGIO, Y. (2014). Generative adversarial networks. Available at arXiv:1406.2661 [cs, stat].

[83] GULRAJANI, I., AHMED, F., ARJOVSKY, M., DUMOULIN, V. and COURVILLE, A. (2017). Improved training of Wasserstein GANs.

[84] HARDT, M., LIGETT, K. and MCSHERRY, F. (2012). A simple and practical algorithm for differentially private data release. Available at arXiv:1012.4763 [cs].

[85] HAWALA, S. (2008). Producing partially synthetic data to avoid disclosure. In *Proceedings of the Joint Statistical Meetings* Amer. Statist. Assoc., Alexandria, VA.

[86] HOMER, N., SZELINGER, S., REDMAN, M., DUGGAN, D., TEMBE, W., MUEHLING, J., PEARSON, J. V., STEPHAN, D. A., NELSON, S. F. et al. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* **4** e1000167. https://doi.org/10.1371/journal.pgen.1000167

[87] HORNBY, R. and HU, J. (2021). Identification risks evaluation of partially synthetic data with the IdentificationRiskCalculation R package. *Trans. Data Priv.* **14** 37–52.

[88] HU, J. (2019). Bayesian estimation of attribute and identification disclosure risks in synthetic data. *Trans. Data Priv.* **12** 61–89.

[89] HU, J., AKANDE, O. and WANG, Q. (2021). Multiple imputation and synthetic data generation with NPBayesImputeCat. *R J.* **13**.

[90] HU, J. and HOSHINO, N. (2018). The quasi-multinomial synthesizer for categorical data. In *International Conference on Privacy in Statistical Databases* 75–91. Springer, Berlin.

[91] HU, J., REITER, J. P. and WANG, Q. (2014). Disclosure risk evaluation for fully synthetic categorical data. In *Privacy in Statistical Databases* (J. Domingo-Ferrer, ed.). *Lecture Notes in Computer Science* **8744** 185–199. Springer, Heidelberg.

[92] HU, J., REITER, J. P. and WANG, Q. (2018). Dirichlet process mixture models for modeling and generating synthetic versions of nested categorical data. *Bayesian Anal.* **13** 183–200. MR3737948 https://doi.org/10.1214/16-BA1047

[93] HU, J., SAVITSKY, T. D. and WILLIAMS, M. R. (2021). Risk-efficient Bayesian data synthesis for privacy protection. *J. Surv. Stat. Methodol.* (online-first).

[94] HU, J., SAVITSKY, T. D. and WILLIAMS, M. R. (2022). Private tabular survey data products through synthetic microdata generation. *J. Surv. Stat. Methodol.* **10** 720–752.

[95] HUNDEPOOL, A., DOMINGO-FERRER, J., FRANCONI, L., GIESSING, S., NORDHOLT, E. S., SPICER, K. and DE WOLF, P.-P. (2012). *Statistical Disclosure Control. Wiley Series in Survey Methodology*. Wiley, Chichester. MR3026260 https://doi.org/10.1002/9781118348239

[96] JACKSON, J., MITRA, R., FRANCIS, B. and DOVE, I. (2022). On integrating the number of synthetic data sets m into the a priori synthesis approach. In *Privacy in Statistical Databases* (J. Domingo-Ferrer and M. Laurent, eds.) 205–219. Springer, Cham.

[97] JACKSON, J., MITRA, R., FRANCIS, B. and DOVE, I. (2022). Using saturated count models for user-friendly synthesis of large confidential administrative database. *J. Roy. Statist. Soc. Ser. A* **185** 1613–1643. MR4537790 https://doi.org/10.1111/rssa.12876

[98] JANICKI, R., HOLAN, S. H., IRIMATA, K. M., LIVSEY, J. and RAIM, A. (2023). Spatial change of support models for differentially private decennial census counts of persons by detailed race and ethnicity. *J. Stat. Theory Pract.* **17** Paper No. 31, 20. MR4565882 https://doi.org/10.1007/s42519-023-00328-5

[99] KAMTHE, S., ASSEFA, S. and DEISENROTH, M. (2021). Copula flows for synthetic data generation. Available at arXiv:2101.00598 [cs, stat].

[100] KARR, A. F., KOHNEN, C. N., OGANIAN, A., REITER, J. P. and SANIL, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *Amer. Statist.* **60** 224–232. MR2246755 https://doi.org/10.1198/000313006X124640

[101] KEEGAN, A. and TIDESWELL, A. (2013). Enabling learners to discover real stories in official statistics with a new synthetic unit record file of the New Zealand Income Survey 2011. Contributed paper to satellite: Statistics education for progress: Youth and official statistics.

[102] KENNICKELL, A. B. (1997). Multiple imputation and disclosure protection: The case of the 1995 survey of consumer finances. In *Record Linkage Techniques*, 1997 (W. Alvey and B. Jamerson, eds.) 248–267. National Academy Press, Washington, DC.

[103] KIFER, D. and MACHANAVAJJHALA, A. (2011). No free lunch in data privacy. In *Proceedings of the* 2011 *ACM SIGMOD International Conference on Management of Data* 193–204.

[104] KIM, H. J., DRECHSLER, J. and THOMPSON, K. J. (2021). Synthetic microdata for establishment surveys under informative sampling. *J. Roy. Statist. Soc. Ser. A* **184** 255–281. MR4204919 https://doi.org/10.1111/rssa.12622

[105] KIM, H. J., REITER, J. P. and KARR, A. F. (2018). Simultaneous edit-imputation and disclosure limitation for business establishment data. *J. Appl. Stat.* **45** 63–82. MR3736858 https://doi.org/10.1080/02664763.2016.1267123

[106] KINGMA, D. P. and WELLING, M. (2014). Auto-encoding variational bayes. Available at arXiv:1312.6114 [cs, stat].

[107] KINNEY, S. K. and REITER, J. P. (2010). Tests of multivariate hypotheses when using multiple imputation for missing data and disclosure limitation. *J. Off. Stat.* **26** 301–315.

[108] KINNEY, S. K., REITER, J. P. and MIRANDA, J. (2014). Synlbd 2.0: Improving the synthetic longitudinal business database. *Stat. J. IAOS* **30** 129–135.

[109] KINNEY, S. K., REITER, J. P., REZNEK, A. P., MIRANDA, J., JARMIN, R. S. and ABOWD, J. M. (2011). Towards unrestricted public use business microdata: The synthetic longitudinal business database. *Int. Stat. Rev.* **79** 362–384.

[110] KLEIN, M. and SINHA, B. (2015). Likelihood based finite sample inference for singly imputed synthetic data under the multivariate normal and multiple linear regression models. *J. Priv. Confid.* **7**.

[111] KOIVU, A., SAIRANEN, M., AIROLA, A. and PAHIKKALA, T. (2020). Synthetic minority oversampling of vital statistics data with generative adversarial networks. *J. Amer. Med. Inform. Assoc.* **27** 1667–1674. https://doi.org/10.1093/jamia/ocaa127

[112] LEE, J. H., KIM, I. Y. and O'KEEFE, C. M. (2013). On regression-tree-based synthetic data methods for business data. *J. Priv. Confid.* **5**.

[113] LI, H., XIONG, L. and JIANG, X. (2014). Differentially private synthesization of multi-dimensional data using Copula functions.

[114] LI, N., LI, T. and VENKATASUBRAMANIAN, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. In 2007 *IEEE* 23*rd International Conference on Data Engineering* 106–115.

[115] LIEW, C. K., CHOI, U. J. and LIEW, C. J. (1985). A data distortion by probability distribution. *ACM Trans. Database Syst.* **10** 395–411. MR0794552

[116] LITTLE, C., ELLIOT, M., ALLMENDINGER, R. and SAMANI, S. S. (2021). Generative adversarial networks for synthetic data generation: A comparative study. Available at arXiv:2112.01925.

[117] LITTLE, R. J. and RAGHUNATHAN, T. (1997). Should imputation of missing data condition on all observed variables. In *Proceedings of the Section on Survey Research Methods* 617–622. Amer. Statist. Assoc., Alexandria, VA.

[118] LITTLE, R. J. A. (1993). Statistical analysis of masked data. *J. Off. Stat.* **9** 407–426.

[119] LITTLE, R. J. A. and RUBIN, D. B. (1987). *Statistical Analysis with Missing Data. Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR0890519

[120] LIU, T., VIETRI, G., STEINKE, T., ULLMAN, J. and WU, S. (2021). Leveraging public data for practical private query release. In *International Conference on Machine Learning* 6968–6977. PMLR.

[121] MA, C., TSCHIATSCHEK, S., HERNÁNDEZ-LOBATO, J. M., TURNER, R. and ZHANG, C. (2020). VAEM: A deep generative model for heterogeneous mixed type data. Available at arXiv:2006.11941 [cs, stat].

[122] MACHANAVAJJHALA, A., KIFER, D., ABOWD, J. M., GEHRKE, J. and VILHUBER, L. (2008). Privacy: Theory meets practice on the map. In *IEEE* 24*th International Conference on Data Engineering* 277–286.

[123] MACHANAVAJJHALA, A., KIFER, D., GEHRKE, J. and VENKITASUBRAMANIAM, M. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data* **1** 3–es.

[124] MAHMOOD, F., BORDERS, D., CHEN, R. J., MCKAY, G. N., SALIMIAN, K. J., BARAS, A. and DURR, N. J. (2019). Deep adversarial training for multi-organ nuclei segmentation in histopathology images. *IEEE Trans. Med. Imag.* **39** 3257–3267.

[125] MANRIQUE-VALLIER, D. and HU, J. (2018). Bayesian nonparametric generation of fully synthetic multivariate categorical data in the presence of structural zeros. *J. Roy. Statist. Soc. Ser. A* **181** 635–647. MR3807501 https://doi.org/10.1111/rssa.12352

[126] MCCLURE, D. and REITER, J. P. (2012). Differential privacy and statistical disclosure risk measures: An investigation with binary synthetic data. *Trans. Data Priv.* **5** 535–552. MR3018910

[127] MCCLURE, D. and REITER, J. P. (2016). Assessing disclosure risks for synthetic data with arbitrary intruder knowledge. *Stat. J. IAOS* **32** 109–126.

[128] MCCLURE, D. R. and REITER, J. P. (2012). Towards providing automated feedback on the quality of inferences from synthetic datasets. *J. Priv. Confid.* **4**.

[129] MCKENNA, R., MIKLAU, G. and SHELDON, D. (2021). Winning the NIST contest: A scalable and general approach to differentially private synthetic data. *J. Priv. Confid.* **11**.

[130] MCKENNA, R., SHELDON, D. and MIKLAU, G. (2019). Graphical-model based estimation and inference for differential privacy.

[131] MENG, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input (Disc: P558-573). *Statist. Sci.* **9** 538–558.

[132] MIRZA, M. and OSINDERO, S. (2014). Conditional generative adversarial nets. CoRR. Available at arXiv:1411.1784.

[133] MITRA, R., BLANCHARD, S., DOVE, I., TUDOR, C. and SPICER, K. (2020). Confidentiality challenges in releasing longitudinally linked data. *Trans. Data Priv.* **13** 151–170.

[134] MITRA, R. and REITER, J. P. (2006). Adjusting survey weights when altering identifying design variables via synthetic data. In *International Conference on Privacy in Statistical Databases* 177–188. Springer, Berlin.

[135] MOTTINI, A., LHERITIER, A. and ACUNA-AGOST, R. (2018). Airline passenger name record generation using generative adversarial networks. Available at arXiv:1807.06657 [cs, stat].

[136] NEUNHOEFFER, M., WU, Z. S. and DWORK, C. (2021). Private post-GAN boosting. Available at arXiv:2007.11934 [cs, stat].

[137] NICHOLSON CONSULTING & KŌTĀTĀ INSIGHT (2021). He Ara Poutama Mō te reo Māori Technical report.

[138] NOWOK, B., RAAB, G. M. and DIBBEN, C. (2016). Synthpop: Bespoke creation of synthetic data in R. *J. Stat. Softw.* **74** 1–26.

[139] NOWOK, B., RAAB, G. M. and DIBBEN, C. (2017). Providing bespoke synthetic data for the UK longitudinal studies and other sensitive data with the synthpop package for R. *Stat. J. IAOS* **33** 785–796.

[140] O'DONOGHUE, C. (2014). *Handbook of Microsimulation Modelling*. Emerald Group Publishing, Leeds, England.

[141] OHM, P. (2009). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Rev.* **57** 1701–1776.

[142] OSINSKI, B., JAKUBOWSKI, A., ZIECINA, P., MIŁOŚ, P., GALIAS, C., HOMOCEANU, S. and MICHALEWSKI, H. (2020). Simulation-based reinforcement learning for real-world autonomous driving. In 2020 *IEEE International Conference on Robotics and Automation* (*ICRA*) 6411–6418.

[143] PAIVA, T., CHAKRABORTY, A., REITER, J. and GELFAND, A. (2014). Imputation of confidential data sets with spatial locations using disease mapping models. *Stat. Med.* **33** 1928–1945. MR3256912 https://doi.org/10.1002/sim.6078

[144] PAPERNOT, N., SONG, S., MIRONOV, I., RAGHUNATHAN, A., TALWAR, K. and ERLINGSSON, Ú. (2018). Scalable private learning with PATE.

[145] PARK, N., MOHAMMADI, M., GORDE, K., JAJODIA, S., PARK, H. and KIM, Y. (2018). Data synthesis based on generative adversarial networks. *Proc. VLDB Endow.* **11** 1071–1083.

[146] PATKI, N., WEDGE, R. and VEERAMACHANENI, K. (2016). The synthetic data vault. In 2016 *IEEE International Conference on Data Science and Advanced Analytics* (*DSAA*) 399–410. IEEE Press, New York.

[147] PISTNER, M., SLAVKOVIĆ, A. and VILHUBER, L. (2018). Synthetic data via quantile regression for heavy-tailed and heteroskedastic data. In *International Conference on Privacy in Statistical Databases* 92–108. Springer, Berlin.

[148] PUBLICATIONS OFFICE OF THE EUROPEAN UNION (2022). data.europa.eu. Available at https://data.europa.eu/en. Last accessed on 2022-05-04.

[149] QUICK, H. (2021). Generating Poisson-distributed differentially private synthetic data. *J. Roy. Statist. Soc. Ser. A* **184** 1093–1108. MR4305573 https://doi.org/10.1111/rssa.12711

[150] QUICK, H. (2021). Improving the utility of Poisson-distributed, differentially private synthetic data via prior predictive truncation with an application to cdc wonder. *J. Surv. Stat. Methodol.* **10** 596–617. MR4305573 https://doi.org/10.1111/rssa.12711

[151] QUICK, H., HOLAN, S. H. and WIKLE, C. K. (2018). Generating partially synthetic geocoded public use data with decreased disclosure risk by using differential smoothing. *J. Roy. Statist. Soc. Ser. A* **181** 649–661. MR3807502 https://doi.org/10.1111/rssa.12360

[152] QUICK, H., HOLAN, S. H., WIKLE, C. K. and REITER, J. P. (2015). Bayesian marked point process modeling for generating fully synthetic public use data with point-referenced geography. *Spat. Stat.* **14** 439–451. MR3431050 https://doi.org/10.1016/j.spasta.2015.07.008

[153] RAAB, G. M., NOWOK, B. and DIBBEN, C. (2016). Practical data synthesis for large samples. *J. Priv. Confid.* **7** 67–97.

[154] RAAB, G. M., NOWOK, B. and DIBBEN, C. (2021). Assessing, visualizing and improving the utility of synthetic data. Available at arXiv:2109.12717.

[155] RAGHUNATHAN, T. E. (2021). Synthetic data. *Annu. Rev. Stat. Appl.* **8** 129–140. MR4243543 https://doi.org/10.1146/annurev-statistics-040720-031848

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

[157] RASHID, S., DRECHSLER, J. and MITRA, R. (2021). Accounting for longitudinal data structures when disseminating synthetic data to the public. In *UNECE Expert Meeting on Statistical Data Confidentiality* 2021.

[158] REITER, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *J. Off. Stat.* **18** 531–544.

[159] REITER, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Surv. Methodol.* **30** 235–242.

[160] REITER, J. P. (2005). Inference for partially synthetic, public use microdata sets. *Surv. Methodol.* **29** 181–189.

[161] REITER, J. P. (2005). Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study. *J. Roy. Statist. Soc. Ser. A* **168** 185–205. MR2113234 https://doi.org/10.1111/j.1467-985X.2004.00343.x

[162] REITER, J. P. (2005). Significance tests for multi-component estimands from multiply imputed, synthetic microdata. *J. Statist. Plann. Inference* **131** 365–377. MR2139378 https://doi.org/10.1016/j.jspi.2004.02.003

[163] REITER, J. P. (2005). Using CART to generate partially synthetic, public use microdata. *J. Off. Stat.* **21** 441–462.

[164] REITER, J. P. (2005). Estimating risks of identification disclosure in microdata. *J. Amer. Statist. Assoc.* **100** 1103–1112. MR2236926 https://doi.org/10.1198/016214505000000619

[165] REITER, J. P. and DRECHSLER, J. (2010). Releasing multiply-imputed synthetic data generated in two stages to protect confidentiality. *Statist. Sinica* **20** 405–421. MR2640701

[166] REITER, J. P. and KINNEY, S. K. (2012). Inferentially valid, partially synthetic data: Generating from posterior predictive distributions not necessary. *J. Off. Stat.* **28** 583–590.

[167] REITER, J. P. and MITRA, R. (2009). Estimating risks of identification disclosure in partially synthetic data. *J. Priv. Confid.* **1** 99–110.

[168] REITER, J. P., OGANIAN, A. and KARR, A. F. (2009). Verification servers: Enabling analysts to assess the quality of inferences from public use data. *Comput. Statist. Data Anal.* **53** 1475–1482. MR2657106 https://doi.org/10.1016/j.csda.2008.10.006

[169] REITER, J. P. and RAGHUNATHAN, T. E. (2007). The multiple adaptations of multiple imputation. *J. Amer. Statist. Assoc.* **102** 1462–1471. MR2372542 https://doi.org/10.1198/016214507000000932

[170] REITER, J. P., WANG, Q. and ZHANG, B. (2014). Bayesian estimation of disclosure risks for multiply imputed, synthetic data. *J. Priv. Confid.* **6**.

[171] ROCHER, L., HENDRICKX, J. M. and DE MONTJOYE, Y.-A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nat. Commun.* **10**. https://doi.org/10.1038/s41467-019-10933-3

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

[173] RUBIN, D. B. (1978). Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. In *Proceedings of the Survey Research Methods Section of the American Statistical Association* **1** 20–34 Amer. Statist. Assoc., Alexandria, VA, USA.

[174] RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR0899519 https://doi.org/10.1002/9780470316696

[175] RUBIN, D. B. (1993). Discussion: Statistical disclosure limitation. *J. Off. Stat.* **9** 462–468.

[176] SAKSHAUG, J. W. and RAGHUNATHAN, T. E. (2010). Synthetic data for small area estimation. In *Privacy in Statistical Databases* (J. Domingo-Ferrer and E. Magkos, eds.) 162–173. Springer, Heidelberg.

[177] SAKSHAUG, J. W. and RAGHUNATHAN, T. E. (2014). Generating synthetic data to produce public-use microdata for small geographic areas based on complex sample survey data with application to the National Health Interview Survey. *J. Appl. Stat.* **41** 2103–2122. MR3292662 https://doi.org/10.1080/02664763.2014.909778

[178] SALLIER, K. (2020). Toward more user-centric data access solutions: Producing synthetic data of high analytical value by data synthesis. *Stat. J. IAOS* **36** 1059–1066.

[179] SHLOMO, N. (2014). Probabilistic record linkage for disclosure risk assessment. In *International Conference on Privacy in Statistical Databases* 269–282. Springer, Berlin.

[180] SIWICKI, B. (2021). Synthetic data boosts accuracy and speed of brain tumor surgery CDS. Available at https://www.healthcareitnews.com/news/synthetic-data-boosts-accuracy-and-speed-brain-tumor-surgery-cds. Last accessed on 2022-05-04.

[181] SKINNER, C. and SHLOMO, N. (2008). Assessing identification risk in survey microdata using log-linear models. *J. Amer. Statist. Assoc.* **103** 989–1001. MR2462887 https://doi.org/10.1198/016214507000001328

[182] SNOKE, J., RAAB, G. M., NOWOK, B., DIBBEN, C. and SLAVKOVIC, A. (2018). General and specific utility measures for synthetic data. *J. Roy. Statist. Soc. Ser. A* **181** 663–688. MR3807503 https://doi.org/10.1111/rssa.12358

[183] SRIVASTAVA, A., VALKOV, L., RUSSELL, C., GUTMANN, M. U. and SUTTON, C. (2017). VEEGAN: Reducing mode collapse in GANs using implicit variational learning.

[184] STADLER, T., OPRISANU, B. and TRONCOSO, C. (2021). Synthetic data—anonymisation groundhog day. Available at arXiv:2011.07018.

[185] SWEENEY, L. (2002). k-anonymity: A model for protecting privacy. *Internat. J. Uncertain. Fuzziness Knowledge-Based Systems* **10**. Aggregation and security assessment for inference control in statistical databases. MR1948199 https://doi.org/10.1142/S0218488502001648

[186] SWEENEY, L. (2013). Matching known patients to health records in Washington state data. Available at arXiv:1307.1370.

[187] TAUB, J. and ELLIOT, M. (2019). The synthetic data challenge. Joint UNECE/Eurostat work session on statistical data confidentiality, The Hague, The Netherlands.

[188] THOMPSON, K. and KIM, H. J. (2022). Incorporating economic conditions in synthetic microdata for business programs. *J. Surv. Stat. Methodol.* **10** 830–859.

[189] THOMPSON, S. A. and WARZEL, C. (2019). Twelve million phones, one dataset, zero privacy. Available at https://www.nytimes.com/interactive/2019/12/19/opinion/location-tracking-cell-phone.html. Last accessed on 2023-06-20.

[190] TORFI, A. (2020). Privacy-preserving synthetic medical data generation with deep learning. Virginia Tech.

[191] TORFI, A. and FOX, E. A. (2020). COR-GAN: Correlation-capturing convolutional neural networks for generating synthetic healthcare records. CoRR. Available at arXiv:2001.09346.

[192] TORKZADEHMAHANI, R., KAIROUZ, P. and PATEN, B. (2020). DP-CGAN: Differentially private synthetic data and label generation. Available at arXiv:2001.09700 [cs, stat].

[193] U. S. GENERAL SERVICES ADMINISTRATION (2022). Data.gov. Available at https://data.gov/. Last accessed on 2022-05-04.

[194] VADHAN, S. (2017). The complexity of differential privacy. In *Tutorials on the Foundations of Cryptography*. *Inf. Secur. Cryptography* 347–450. Springer, Cham. MR3837668

[195] VARDHAN, L. V. H. and KOK, S. (2020). Generating privacy-preserving synthetic tabular data using oblivious variational autoencoders. In *Proceedings of the Workshop on Economics of Privacy and Data Labor at the 37th International Conference on Machine Learning*.

[196] VOAS, D. and WILLIAMSON, P. (2001). Evaluating goodness-of-fit measures for synthetic microdata. *Geogr. Environ. Model.* **5** 177–200.

[197] WAHEED, A., GOYAL, M., GUPTA, D., KHANNA, A., AL-TURJMAN, F. and PINHEIRO, P. R. (2020). CovidGAN: Data augmentation using auxiliary classifier GAN for improved Covid-19 detection. *IEEE Access* **8** 91916–91923. https://doi.org/10.1109/ACCESS.2020.2994762

[198] WANG, H. and REITER, J. P. (2012). Multiple imputation for sharing precise geographies in public use data. *Ann. Appl. Stat.* **6** 229–252. MR2951536 https://doi.org/10.1214/11-AOAS506

[199] WEI, L. and REITER, J. P. (2016). Releasing synthetic magnitude microdata constrained to fixed marginal totals. *Stat. J. IAOS* **32** 93–108.

[200] WEN, B., COLON, L. O., SUBBALAKSHMI, K. P. and CHANDRAMOULI, R. (2021). Causal-TGAN: Generating tabular data using causal generative adversarial networks.

[201] WIESE, M., KNOBLOCH, R., KORN, R. and KRETSCHMER, P. (2020). Quant GANs: Deep generation of financial time series. *Quant. Finance* **20** 1419–1440. MR4149599 https://doi.org/10.1080/14697688.2020.1730426

[202] WOO, M. J., REITER, J. P., OGANIAN, A. and KARR, A. F. (2009). Global measures of data utility for microdata masked for disclosure limitation. *J. Priv. Confid.* **1** 111–124.

[203] XIAO, X., WANG, G. and GEHRKE, J. (2011). Differential privacy via wavelet transforms. *IEEE Trans. Knowl. Data Eng.* **23** 1200–1214.

[204] XIE, L., LIN, K., WANG, S., WANG, F. and ZHOU, J. (2018). Differentially private generative adversarial network. Available at arXiv:1802.06739 [cs, stat].

[205] XU, L., SKOULARIDOU, M., CUESTA-INFANTE, A. and VEERAMACHANENI, K. (2019). Modeling tabular data using conditional GAN. In *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. D. Alché-Buc, E. Fox and R. Garnett, eds.). **32**. Curran Associates, Red Hook.

[206] YAHI, A., VANGURI, R., ELHADAD, N. and TATONETTI, N. P. (2017). Generative adversarial networks for electronic health records: A framework for exploring and evaluating methods for predicting drug-induced laboratory test trajectories. Available at arXiv:1712.00164.

[207] YOON, J., JORDON, J. and SCHAAR, M. V. D. (2019). PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*.

[208] YU, H. and REITER, J. P. (2018). Differentially private verification of regression predictions from synthetic data. *Trans. Data Priv.* **11** 279–297.

[209] ZHANG, J., CORMODE, G., PROCOPIUC, C. M., SRIVASTAVA, D. and XIAO, X. (2014). PrivBayes: Private data release via Bayesian networks. In *Proceedings of the* 2014 *ACM SIGMOD International Conference on Management of Data*. 1423–1434.

[210] ZHANG, J., CORMODE, G., PROCOPIUC, C. M., SRIVASTAVA, D. and XIAO, X. (2017). PrivBayes: Private data release via Bayesian networks. *ACM Trans. Database Syst.* **42** Art. 25, 41. MR3730676 https://doi.org/10.1145/3134428

[211] ZHAO, Z., KUNAR, A., VAN DER SCHEER, H., BIRKE, R. and CHEN, L. Y. (2021). CTAB-GAN: Effective table data synthesizing. Available at arXiv:2102.08369 [cs].

[212] ZHOU, H., ELLIOTT, M. R. and RAGHUNATHAN, T. E. (2016). Synthetic multiple-imputation procedure for multistage complex samples. *J. Off. Stat.* **32** 231–256. https://doi.org/10.1515/JOS-2016-0011

[213] (2017). Learning with privacy at scale. *Apple Mach. Learn. J.* **1** 8.

[214] (2021). Exposure notification privacy-preserving analytics. White paper, available at https://covid19-static.cdn-apple.com/applications/covid19/current/static/contact-tracing/pdf/ENPA_White_Paper.pdf. Last accessed on 2023-06-21.

Statistical Science
2024, Vol. 39, No. 2, 243–261
https://doi.org/10.1214/23-STS895

Statistical Frameworks for Oncology Dose-Finding Designs with Late-Onset Toxicities: A Review

Tianjian Zhou and Yuan Ji

Abstract. In oncology dose-finding trials, due to staggered enrollment, it might be desirable to make dose-assignment decisions in real time in the presence of pending toxicity outcomes, for example, when the dose-limiting toxicity is late onset. Patients' time-to-event information may be utilized to facilitate such decisions. We review statistical frameworks for time-to-event modeling in dose-finding trials and summarize existing designs into two classes: TITE designs and POD designs. TITE designs are based on inference about toxicity probabilities, while POD designs are based on probabilities of dose-assignment decisions. These two classes of designs contain existing individual designs as special cases and also give rise to new designs. We discuss and study the theoretical properties of these designs, including large-sample convergence properties, coherence principles and the underlying decision rules. To facilitate the use of these designs in practice, we introduce efficient computational algorithms and review common practical considerations, such as safety rules and suspension rules. Finally, the operating characteristics of several designs are evaluated and compared through computer simulations.

Key words and phrases: Clinical trial design, maximum tolerated dose, missing data, survival analysis, time-to-event modeling.

REFERENCES

[1] ANDRILLON, A., CHEVRET, S., LEE, S. M. and BIARD, L. (2020). Dose-finding design and benchmark for a right censored endpoint. *J. Biopharm. Statist.* **30** 948–963.

[2] BABB, J., ROGATKO, A. and ZACKS, S. (1998). Cancer phase I clinical trials: Efficient dose escalation with overdose control. *Stat. Med.* **17** 1103–1120. https://doi.org/10.1002/(sici)1097-0258(19980530)17:10<1103::aid-sim793>3.0.co;2-9

[3] CHEUNG, Y. K. (2005). Coherence principles in dose-finding studies. *Biometrika* **92** 863–873. MR2234191 https://doi.org/10.1093/biomet/92.4.863

[4] CHEUNG, Y. K. and CHAPPELL, R. (2000). Sequential designs for phase I clinical trials with late-onset toxicities. *Biometrics* **56** 1177–1182. MR1815616 https://doi.org/10.1111/j.0006-341X.2000.01177.x

[5] CHEUNG, Y. K. and CHAPPELL, R. (2002). A simple technique to evaluate model sensitivity in the continual reassessment method. *Biometrics* **58** 671–674. MR1933538 https://doi.org/10.1111/j.0006-341X.2002.00671.x

[6] CLERTANT, M. and O'QUIGLEY, J. (2017). Semiparametric dose finding methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1487–1508. MR3731672 https://doi.org/10.1111/rssb.12229

[7] CLERTANT, M. and O'QUIGLEY, J. (2019). Semiparametric dose finding methods: Special cases. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **68** 271–288. MR3902994

[8] DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

[9] DOMENICANO, I., VENTZ, S., CELLAMARE, M., MAK, R. H. and TRIPPA, L. (2019). Bayesian uncertainty-directed dose finding designs. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **68** 1393–1410. MR4022818 https://doi.org/10.1111/rssc.12355

[10] FOOD AND DRUG ADMINISTRATION (2023). Project Optimus: Reforming the dose optimization and dose selection paradigm in oncology. Available at https://www.fda.gov/about-fda/oncology-center-excellence/project-optimus.

[11] GASPARINI, M. and EISELE, J. (2000). A curve-free method for phase I clinical trials. *Biometrics* **56** 609–615. MR1795024 https://doi.org/10.1111/j.0006-341X.2000.00609.x

[12] GOODMAN, S. N., ZAHURAK, M. L. and PIANTADOSI, S. (1995). Some practical improvements in the continual reassessment method for phase I studies. *Stat. Med.* **14** 1149–1161. https://doi.org/10.1002/sim.4780141102

[13] GUO, W., JI, Y. and LI, D. (2019). R-TPI: Rolling toxicity probability interval design to shorten the duration and maintain safety of phase I trials. *J. Biopharm. Statist.* **29** 411–424.

Tianjian Zhou is Assistant Professor, Department of Statistics, Colorado State University, Fort Collins, Colorado 80523-1877, USA (e-mail: tianjian.zhou@colostate.edu). Yuan Ji is Professor, Department of Public Health Sciences, University of Chicago, Chicago, Illinois 60637, USA (e-mail: yji@bsd.uchicago.edu).

[14] GUO, W., WANG, S.-J., YANG, S., LYNN, H. and JI, Y. (2017). A Bayesian interval dose-finding design addressing Ockham's razor: MTPI-2. *Contemp. Clin. Trials* **58** 23–33.

[15] IVANOVA, A., FLOURNOY, N. and CHUNG, Y. (2007). Cumulative cohort design for dose-finding. *J. Statist. Plann. Inference* **137** 2316–2327. MR2325437 https://doi.org/10.1016/j.jspi.2006.07.009

[16] IVANOVA, A., WANG, Y. and FOSTER, M. C. (2016). The rapid enrollment design for Phase I clinical trials. *Stat. Med.* **35** 2516–2524. MR3513702 https://doi.org/10.1002/sim.6886

[17] JI, Y., LIU, P., LI, Y. and NEBIYOU BEKELE, B. (2010). A modified toxicity probability interval method for dose-finding trials. *Clin. Trials* **7** 653–663.

[18] JI, Y. and WANG, S.-J. (2013). Modified toxicity probability interval design: A safer and more reliable method than the 3+3 design for practical phase I trials. *J. Clin. Oncol.* **31** 1785–1791.

[19] JIN, I. H., LIU, S., THALL, P. F. and YUAN, Y. (2014). Using data augmentation to facilitate conduct of phase I–II clinical trials with delayed outcomes. *J. Amer. Statist. Assoc.* **109** 525–536. MR3223730 https://doi.org/10.1080/01621459.2014.881740

[20] KANJANAPAN, Y., DAY, D., BUTLER, M., WANG, L., JOSHUA, A., HOGG, D., LEIGHL, N., RAZAK, A. A., HANSEN, A. et al. (2019). Delayed immune-related adverse events in assessment for dose-limiting toxicity in early phase immunotherapy trials. *Eur. J. Cancer* **107** 1–7.

[21] KLEIN, J. P. and MOESCHBERGER, M. L. (2006). *Survival Analysis*: *Techniques for Censored and Truncated Data*. Springer, Berlin.

[22] LEE, J. J. and LIU, D. D. (2008). A predictive probability design for phase II cancer clinical trials. *Clin. Trials* **5** 93–106.

[23] LEE, S. M., URSINO, M., CHEUNG, Y. K. and ZOHAR, S. (2019). Dose-finding designs for cumulative toxicities using multiple constraints. *Biostatistics* **20** 17–29. MR3892280 https://doi.org/10.1093/biostatistics/kxx059

[24] LIN, R. and YUAN, Y. (2020). Time-to-event model-assisted designs for dose-finding trials with delayed toxicity. *Biostatistics* **21** 807–824. MR4164059 https://doi.org/10.1093/biostatistics/kxz007

[25] LIU, J., YUAN, S., BEKELE, B. N. and JI, Y. (2023). The backfill i3+3 design for dose-finding trials in oncology. Preprint. Available at arXiv:2303.15798.

[26] LIU, M., JI, Y. and LIN, J. (2021). PoD-BIN: A probability of decision Bayesian interval design for time-to-event dose-finding trials with multiple toxicity grades. Preprint. Available at arXiv:2103.06368.

[27] LIU, M., WANG, S.-J. and JI, Y. (2020). The i3+3 design for phase I clinical trials. *J. Biopharm. Statist.* **30** 294–304.

[28] LIU, S. and NING, J. (2013). A Bayesian dose-finding design for drug combination trials with delayed toxicities. *Bayesian Anal.* **8** 703–722. MR3102231 https://doi.org/10.1214/13-BA839

[29] LIU, S., YIN, G. and YUAN, Y. (2013). Bayesian data augmentation dose finding with continual reassessment method and delayed toxicity. *Ann. Appl. Stat.* **7** 2138–2156. MR3161716 https://doi.org/10.1214/13-AOAS661

[30] LIU, S. and YUAN, Y. (2015). Bayesian optimal interval designs for phase I clinical trials. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **64** 507–523. MR3325461 https://doi.org/10.1111/rssc.12089

[31] MANDER, A. P. and SWEETING, M. J. (2015). A product of independent beta probabilities dose escalation design for dual-agent phase I trials. *Stat. Med.* **34** 1261–1276. MR3322767 https://doi.org/10.1002/sim.6434

[32] MOZGUNOV, P., GASPARINI, M. and JAKI, T. (2020). A surface-free design for phase I dual-agent combination trials. *Stat. Methods Med. Res.* **29** 3093–3109. MR4136580 https://doi.org/10.1177/0962280220919450

[33] NEUENSCHWANDER, B., BRANSON, M. and GSPONER, T. (2008). Critical aspects of the Bayesian approach to phase I cancer trials. *Stat. Med.* **27** 2420–2439. MR2432497 https://doi.org/10.1002/sim.3230

[34] NORMOLLE, D. and LAWRENCE, T. (2006). Designing dose-escalation trials with late-onset toxicities using the time-to-event continual reassessment method. *J. Clin. Oncol.* **24** 4426–4433. https://doi.org/10.1200/JCO.2005.04.3844

[35] O'QUIGLEY, J., PEPE, M. and FISHER, L. (1990). Continual reassessment method: A practical design for phase 1 clinical trials in cancer. *Biometrics* **46** 33–48. MR1059105 https://doi.org/10.2307/2531628

[36] O'QUIGLEY, J. and SHEN, L. Z. (1996). Continual reassessment method: A likelihood approach. *Biometrics* **52** 673–684.

[37] ORON, A. P., AZRIEL, D. and HOFF, P. D. (2011). Dose-finding designs: The role of convergence properties. *Int. J. Biostat.* **7** Art. 39, 19 pp. MR2873999 https://doi.org/10.2202/1557-4679.1298

[38] SAVILLE, B. R., CONNOR, J. T., AYERS, G. D. and ALVAREZ, J. (2014). The utility of Bayesian predictive probabilities for interim monitoring of clinical trials. *Clin. Trials* **11** 485–493. https://doi.org/10.1177/1740774514531352

[39] SERTKAYA, A., WONG, H.-H., JESSUP, A. and BELECHE, T. (2016). Key cost drivers of pharmaceutical clinical trials in the United States. *Clin. Trials* **13** 117–126.

[40] SHEN, L. Z. and O'QUIGLEY, J. (1996). Consistency of continual reassessment method under model misspecification. *Biometrika* **83** 395–405. MR1439791 https://doi.org/10.1093/biomet/83.2.395

[41] SKOLNIK, J. M., BARRETT, J. S., JAYARAMAN, B., PATEL, D. and ADAMSON, P. C. (2008). Shortening the timeline of pediatric phase I trials: The rolling six design. *J. Clin. Oncol.* **26** 190–195. https://doi.org/10.1200/JCO.2007.12.7712

[42] STORER, B. E. (1989). Design and analysis of phase I clinical trials. *Biometrics* **45** 925–937. MR1029610 https://doi.org/10.2307/2531693

[43] TAKEDA, K., MORITA, S. and TAGURI, M. (2020). TITE-BOIN-ET: Time-to-event Bayesian optimal interval design to accelerate dose-finding based on both efficacy and toxicity outcomes. *Pharm. Stat.* **19** 335–349.

[44] TAKEDA, K., XIA, Q., LIU, S. and RONG, A. (2022). TITE-gBOIN: Time-to-event Bayesian optimal interval design to accelerate dose-finding accounting for toxicity grades. *Pharm. Stat.* **21** 496–506.

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

[46] WAGES, N. A., BRAUN, T. M. and CONAWAY, M. R. (2023). Isotonic design for phase I cancer clinical trials with late-onset toxicities. *J. Biopharm. Statist.* **33** 357–370. https://doi.org/10.1080/10543406.2022.2162068

[47] WEBER, J. S., YANG, J. C., ATKINS, M. B. and DISIS, M. L. (2015). Toxicities of immunotherapy for the practitioner. *J. Clin. Oncol.* **33** 2092–2099.

[48] WHEELER, G. M., SWEETING, M. J. and MANDER, A. P. (2019). A Bayesian model-free approach to combination therapy phase I trials using censored time-to-toxicity data. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **68** 309–329. MR3902996 https://doi.org/10.1111/rssc.12323

[49] XU, Z. and LIN, X. (2022). Probability-of-decision interval 3+3 (POD-i3+3) design for phase I dose finding trials with late-onset toxicity. *Stat. Methods Med. Res.* **31** 534–548. MR4388400 https://doi.org/10.1177/09622802211052746

[50] YAN, F., MANDREKAR, S. J. and YUAN, Y. (2017). Keyboard: A novel Bayesian toxicity probability interval design for phase I clinical trials. *Clin. Cancer Res.* **23** 3994–4003.

[51] YIN, G., ZHENG, S. and XU, J. (2013). Fractional dose-finding methods with late-onset toxicity in Phase I clinical trials. *J. Biopharm. Statist.* **23** 856–870. MR3196080 https://doi.org/10.1080/10543406.2013.789892

[52] YUAN, Y., LIN, R., LI, D., NIE, L. and WARREN, K. E. (2018). Time-to-event Bayesian optimal interval design to accelerate phase I trials. *Clin. Cancer Res.*. Article No. 0246.

[53] YUAN, Y. and YIN, G. (2011). Robust EM continual reassessment method in oncology dose finding. *J. Amer. Statist. Assoc.* **106** 818–831. MR2894740 https://doi.org/10.1198/jasa.2011.ap09476

[54] ZHOU, T., GUO, W. and JI, Y. (2020). PoD-TPI: Probability-of-decision toxicity probability interval design to accelerate phase I trials. *Stat. Biosci.* **12** 124–145.

[55] ZHOU, T. and JI, Y. (2020). Emerging methods for oncology clinical trials. *Chance* **33** 39–48.

[56] ZHOU, T. and JI, Y. (2024). Supplement to "Statistical frameworks for oncology dose-finding designs with late-onset toxicities: A review." https://doi.org/10.1214/23-STS895SUPPA, https://doi.org/10.1214/23-STS895SUPPB

[57] ZHOU, Y., LIN, R., LEE, J. J., LI, D., WANG, L., LI, R. and YUAN, Y. (2022). TITE-BOIN12: A Bayesian phase I/II trial design to find the optimal biological dose with late-onset toxicity and efficacy. *Stat. Med.* **41** 1918–1931. MR4411866 https://doi.org/10.1002/sim.9337

Statistical Science
2024, Vol. 39, No. 2, 262–285
https://doi.org/10.1214/23-STS898

ANOVA for Metric Spaces, with Applications to Spatial Data

Raoul Müller, Dominic Schuhmacher and Jorge Mateu

Abstract. We give a review of some recent ANOVA-like procedures for testing group differences based on data in a metric space and present a new such procedure. Our statistic is derived from the classic Levene's test for detecting differences in dispersion. It uses only pairwise distances of data points and can be computed quickly and precisely in situations where the computation of barycenters ("generalized means") in the data space is slow, only by approximation or even infeasible. It also satisfies asymptotic normality.

We discuss the relative merits of the various procedures based on simulation studies for spatial point patterns and image data in a 1-way ANOVA setting. As applications, we perform 1- and 2-way ANOVAs on a data set of bubbles in a mineral flotation process and a data set of local pest counts in Madrid.

Key words and phrases: ANOVA, images, Levene's test, metric spaces, spatial point patterns.

REFERENCES

[1] ALEKSEYENKO, A. V. (2016). Multivariate Welch t-test on distances. *Bioinformatics* **32** 3552–3558. https://doi.org/10.1093/bioinformatics/btw524

[2] ANDERSON, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology* **26** 32–46.

[3] ANDERSON, M. J. (2006). Distance-based tests for homogeneity of multivariate dispersions. *Biometrics* **62** 245–253. MR2226579 https://doi.org/10.1111/j.1541-0420.2005.00440.x

[4] ANDERSON, M. J. (2017). Permutational multivariate analysis of variance (PERMANOVA). *Wiley Statsref*: *Statistics Reference Online* 1–15.

[5] ANDERSON, M. J., WALSH, D. C. I., CLARKE, K. R., GORLEY, R. N. and GUERRA-CASTRO, E. (2017). Some solutions to the multivariate Behrens–Fisher problem for dissimilarity-based analyses. *Aust. N. Z. J. Stat.* **59** 57–79. MR3635167 https://doi.org/10.1111/anzs.12176

[6] BERTSEKAS, D. P. (1988). The auction algorithm: A distributed relaxation method for the assignment problem. *Ann. Oper. Res.* **14** 105–123. MR0963896 https://doi.org/10.1007/BF02186476

[7] BILLERA, L. J., HOLMES, S. P. and VOGTMANN, K. (2001). Geometry of the space of phylogenetic trees. *Adv. in Appl. Math.* **27** 733–767. MR1867931 https://doi.org/10.1006/aama.2001.0759

[8] BORGWARDT, S. and PATTERSON, S. (2020). Improved linear programs for discrete barycenters. *INFORMS J. Optim.* **2** 14–33. MR4172566 https://doi.org/10.1287/ijoo.2019.0020

[9] BORGWARDT, S. and PATTERSON, S. (2021). On the computational complexity of finding a sparse Wasserstein barycenter. *J. Comb. Optim.* **41** 736–761. MR4228512 https://doi.org/10.1007/s10878-021-00713-5

[10] BROWN, M. B. and FORSYTHE, A. B. (1974). Robust tests for the equality of variances. *J. Amer. Statist. Assoc.* **69** 364–367.

[11] BROWN, M. B. and FORSYTHE, A. B. (1974). The small sample behavior of some statistics which test the equality of several means. *Technometrics* **16** 129–132. MR0334368 https://doi.org/10.2307/1267501

[12] CUEVAS, A., FEBRERO, M. and FRAIMAN, R. (2004). An anova test for functional data. *Comput. Statist. Data Anal.* **47** 111–122. MR2087932 https://doi.org/10.1016/j.csda.2003.10.021

[13] DALEY, D. J. and VERE-JONES, D. (2003). *An Introduction to the Theory of Point Processes. Vol. I*: *Elementary Theory and Methods*, 2nd ed. *Probability and Its Applications* (*New York*). Springer, New York. MR1950431

[14] DALEY, D. J. and VERE-JONES, D. (2008). *An Introduction to the Theory of Point Processes. Vol. II*: *General Theory and Structure*, 2nd ed. *Probability and Its Applications* (*New York*). Springer, New York. MR2371524 https://doi.org/10.1007/978-0-387-49835-5

[15] DENKER, M. and KELLER, G. (1983). On U-statistics and v. Mises' statistics for weakly dependent processes. *Z. Wahrsch. Verw. Gebiete* **64** 505–522. MR0717756 https://doi.org/10.1007/BF00534953

[16] DUBEY, P. and MÜLLER, H.-G. (2019). Fréchet analysis of variance for random objects. *Biometrika* **106** 803–821. MR4031200 https://doi.org/10.1093/biomet/asz052

[17] FISHER, R. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd.

Raoul Müller is a Research Assistant, Institute for Mathematical Stochastics, University of Göttingen, 37077 Göttingen, Germany (e-mail: raoul.mueller@uni-goettingen.de). Dominic Schuhmacher is a Professor, Institute for Mathematical Stochastics, University of Göttingen, 37077 Göttingen, Germany. Jorge Mateu is a Professor, Department of Mathematics, University Jaume I, 12071 Castellón, Spain.

[18] GASTWIRTH, J. L., GEL, Y. R. and MIAO, W. (2009). The impact of Levene's test of equality of variances on statistical theory and practice. *Statist. Sci.* **24** 343–360. MR2757435 https://doi.org/10.1214/09-STS301

[19] GE, D., WANG, H., XIONG, Z. and YE, Y. (2019). Interior-point methods strike back: Solving the Wasserstein barycenter problem. *Adv. Neural Inf. Process. Syst.* **32**.

[20] GINESTET, C. E., LI, J., BALACHANDRAN, P., ROSENBERG, S. and KOLACZYK, E. D. (2017). Hypothesis testing for network data in functional neuroimaging. *Ann. Appl. Stat.* **11** 725–750. MR3693544 https://doi.org/10.1214/16-AOAS1015

[21] GONZÁLEZ, J. A., LAGOS-ÁLVAREZ, B. M. and MATEU, J. (2021). Two-way layout factorial experiments of spatial point pattern responses in mineral flotation. *TEST* **30** 1046–1075. MR4346817 https://doi.org/10.1007/s11749-021-00768-w

[22] GRAYBILL, F. A. and MARSAGLIA, G. (1957). Idempotent matrices and quadratic forms in the general linear hypothesis. *Ann. Math. Stat.* **28** 678–686. MR0092307 https://doi.org/10.1214/aoms/1177706879

[23] GRETTON, A., BORGWARDT, K. M., RASCH, M. J., SCHÖLKOPF, B. and SMOLA, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.* **13** 723–773. MR2913716

[24] HAMIDI, B., WALLACE, K., VASU, C. and ALEKSEYENKO, A. V. (2019). W_d^*-Test: Robust distance-based multivariate analysis of variance. *Microbiome* **7** 1–9.

[25] HEINEMANN, F. (2021). WSGeometry: Geometric Tools Based on Balanced/Unbalanced Optimal Transport. R package version 1.2.1. Available at https://CRAN.R-project.org/package=WSGeometry.

[26] HEINEMANN, F., KLATT, M. and MUNK, A. (2023). Kantorovich–Rubinstein distance and barycenter for finitely supported measures: Foundations and algorithms. *Appl. Math. Optim.* **87** Paper No. 4. MR4506758 https://doi.org/10.1007/s00245-022-09911-x

[27] HEINEMANN, F., MUNK, A. and ZEMEL, Y. (2022). Randomized Wasserstein barycenter computation: Resampling with statistical guarantees. *SIAM J. Math. Data Sci.* **4** 229–259. MR4386483 https://doi.org/10.1137/20M1385263

[28] HOEFFDING, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Stat.* **19** 293–325. MR0026294 https://doi.org/10.1214/aoms/1177730196

[29] HOEFFDING, W. (1961). The strong law of large numbers for U-statistics. Technical Report, Mimeograph Series No. 302. Dept. Statistics, Univ. North Carolina.

[30] HUCKEMANN, S., HOTZ, T. and MUNK, A. (2009). Intrinsic MANOVA for Riemannian manifolds with an application to Kendall's space of planar shapes. *IEEE Trans. Pattern Anal. Mach. Intell.* **32** 593–603.

[31] LEE, Y. T. and SIDFORD, A. (2014). Path-finding methods for linear programming: Solving linear programs in $\widetilde{O}(\sqrt{rank})$ iterations and faster algorithms for maximum flow. In 55*th Annual IEEE Symposium on Foundations of Computer Science—FOCS* 2014 424–433. IEEE Computer Soc., Los Alamitos, CA. MR3344892 https://doi.org/10.1109/FOCS.2014.52

[32] LEVENE, H. (1960). Robust tests for equality of variances. In *Contributions to Probability and Statistics. Stanford Studies in Mathematics and Statistics* **2** 278–292. Stanford Univ. Press, Stanford, CA. MR0120709

[33] MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1979). *Multivariate Analysis. Probability and Mathematical Statistics: A Series of Monographs and Textbooks*. Academic Press [Harcourt Brace Jovanovich, Publishers], London-New York-Toronto. MR0560319

[34] MÜLLER, R. and SCHUHMACHER, D. (2019–2022). ttbary: Barycenter Methods for Spatial Point Patterns. R package version 0.3-0. Available at https://CRAN.R-project.org/package=ttbary.

[35] MÜLLER, R., SCHUHMACHER, D. and MATEU, J. (2020). Metrics and barycenters for point pattern data. *Stat. Comput.* **30** 953–972. MR4108686 https://doi.org/10.1007/s11222-020-09932-y

[36] RAMÓN, P., DE LA CRUZ, M., CHACÓN-LABELLA, J. and ESCUDERO, A. (2016). A new non-parametric method for analyzing replicated point patterns in ecology. *Ecography* **39** 1109–1117.

[37] RIZZO, M. L. and SZÉKELY, G. J. (2010). DISCO analysis: A nonparametric extension of analysis of variance. *Ann. Appl. Stat.* **4** 1034–1055. MR2758432 https://doi.org/10.1214/09-AOAS245

[38] SCHEFFÉ, H. (1967). *The Analysis of Variance*, 1st ed. John Wiley & Sons.

[39] SCHUHMACHER, D., BÄHRE, B., BONNEEL, N., GOTTSCHLICH, C., HARTMANN, V., HEINEMANN, F., SCHMITZER, B. and SCHRIEBER, J. (2014–2022). transport: Computation of Optimal Transport Plans and Wasserstein Distances. R package version 0.13-0. Available at https://CRAN.R-project.org/package=transport.

[40] SONG, H. and CHEN, H. (2022). New graph-based multi-sample tests for high-dimensional and non-Euclidean data. Preprint. Available at https://arxiv.org/abs/2205.13787.

[41] TAMAYO-URIA, I., MATEU, J. and DIGGLE, P. J. (2014). Modelling of the spatio-temporal distribution of rat sightings in an urban environment. *Spat. Stat.* **9** 192–206. MR3326839 https://doi.org/10.1016/j.spasta.2014.03.005

[42] WELCH, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika* **38** 330–336. MR0046617 https://doi.org/10.1093/biomet/38.3-4.330

[43] WOOLDRIDGE, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*, 2nd ed. MIT Press, Cambridge, MA. MR2768559

[44] ZHANG, J.-T., GUO, J. and ZHOU, B. (2022). Testing equality of several distributions in separable metric spaces: A maximum mean discrepancy based approach. *J. Econometrics*. https://doi.org/10.1016/j.jeconom.2022.03.007.

[45] ZHANG, Q., MAHDI, G., TINKER, J. and CHEN, H. (2020). A graph-based multi-sample test for identifying pathways associated with cancer progression. *Comput. Biol. Chem.* **87** 107285.

Statistical Science
2024, Vol. 39, No. 2, 286–304
https://doi.org/10.1214/23-STS900

Variable Selection Using Bayesian Additive Regression Trees

Chuji Luo and Michael J. Daniels

Abstract. Variable selection is an important statistical problem. This problem becomes more challenging when the candidate predictors are of mixed type (e.g., continuous and binary) and impact the response variable in nonlinear and/or nonadditive ways. In this paper, we review existing variable selection approaches for the Bayesian additive regression trees (BART) model, a nonparametric regression model, which is flexible enough to capture the interactions between predictors and nonlinear relationships with the response. An emphasis of this review is on the ability to identify relevant predictors. We also propose two variable importance measures, which can be used in a permutation-based variable selection approach, and a backward variable selection procedure for BART. We introduce these variations as a way of illustrating limitations and opportunities for improving current approaches and assess these via simulations.

Key words and phrases: BART, feature selection, nonparametric regression.

REFERENCES

ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679. MR1224394

ALTMANN, A., TOLOŞI, L. and SANDER, O. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics* **26** 1340–1347. https://doi.org/10.1093/bioinformatics/btq134

BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32. MR3874153

BARBIERI, M. M. and BERGER, J. O. (2004). Optimal predictive model selection. *Ann. Statist.* **32** 870–897. MR2065192 https://doi.org/10.1214/009053604000000238

BHATTACHARYA, A., PATI, D., PILLAI, N. S. and DUNSON, D. B. (2015). Dirichlet–Laplace priors for optimal shrinkage. *J. Amer. Statist. Assoc.* **110** 1479–1490. MR3449048 https://doi.org/10.1080/01621459.2014.960967

BLEICH, J., KAPELNER, A., GEORGE, E. I. and JENSEN, S. T. (2014). Variable selection for BART: An application to gene regulation. *Ann. Appl. Stat.* **8** 1750–1781. MR3271352 https://doi.org/10.1214/14-AOAS755

CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97** 465–480. MR2650751 https://doi.org/10.1093/biomet/asq017

CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (1998). Bayesian CART model search. *J. Amer. Statist. Assoc.* **93** 935–948. https://doi.org/10.1080/01621459.1998.10473750

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

EFROYMSON, M. A. (1960). Multiple regression analysis. In *Mathematical Methods for Digital Computers* 191–203. Wiley, New York. MR0117923

FRIEDMAN, J. H. (1991). Multivariate adaptive regression splines. *Ann. Statist.* **19** 1–141. With discussion and a rejoinder by the author. MR1091842 https://doi.org/10.1214/aos/1176347963

FRIEDMAN, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **29** 1189–1232. MR1873328 https://doi.org/10.1214/aos/1013203451

FRIEDMAN, J. H. (2002). Stochastic gradient boosting. *Comput. Statist. Data Anal.* **38** 367–378. MR1884869 https://doi.org/10.1016/S0167-9473(01)00065-2

GELFAND, A. E. and DEY, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *J. Roy. Statist. Soc. Ser. B* **56** 501–514. MR1278223

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-6** 721–741. https://doi.org/10.1109/TPAMI.1984.4767596

GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889. https://doi.org/10.1080/01621459.1993.10476353

HASTIE, T. and TIBSHIRANI, R. (2000). Bayesian backfitting. *Statist. Sci.* **15** 196–223. With comments and a rejoinder by the authors. MR1820768 https://doi.org/10.1214/ss/1009212815

HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109. MR3363437 https://doi.org/10.1093/biomet/57.1.97

Chuji Luo is Data Scientist, Google LLC, Mountain View, California 94043, USA (e-mail: chujiluo0212@gmail.com). Michael J. Daniels is Professor, Chair, and Andrew Banks Family Endowed Chair, Department of Statistics, University of Florida, Gainesville, Florida 32611, USA (e-mail: daniels@ufl.edu).

KAPELNER, A. and BLEICH, J. (2016). bartMachine: Machine learning with Bayesian additive regression trees. *J. Stat. Softw.* **70** 1–40. https://doi.org/10.18637/jss.v070.i04

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

LIU, Y., ROČKOVÁ, V. and WANG, Y. (2021). Variable selection with ABC Bayesian forests. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **83** 453–481. MR4294540 https://doi.org/10.1111/rssb.12423

LOUPPE, G. (2014). Understanding random forests. Cornell Univ. Library.

LUO, C. and DANIELS, M. J. (2021). The BartMixVs R package.

LUO, C. and DANIELS, M. J. (2024). Supplement to "Variable selection using Bayesian additive regression trees." https://doi.org/10.1214/23-STS900SUPPA, https://doi.org/10.1214/23-STS900SUPPB, https://doi.org/10.1214/23-STS900SUPPC

ROČKOVÁ, V. and VAN DER PAS, S. (2020). Posterior concentration for Bayesian regression trees and forests. *Ann. Statist.* **48** 2108–2131. MR4134788 https://doi.org/10.1214/19-AOS1879

SPARAPANI, R., SPANBAUER, C. and MCCULLOCH, R. (2021). Nonparametric machine learning and efficient computation with Bayesian additive regression trees: The BART R package. *J. Stat. Softw.* **97** 1–66.

STROBL, C., BOULESTEIX, A. and ZEILEIS, A. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinform.* **8**. https://doi.org/10.1186/1471-2105-8-25

TADESSE, M. G. and VANNUCCI, M. (2021). *Handbook of Bayesian Variable Selection*.

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

VEHTARI, A., GELMAN, A. and GABRY, J. (2017). Erratum to: Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC [MR3647105]. *Stat. Comput.* **27** 1433. MR3647106 https://doi.org/10.1007/s11222-016-9709-3

WANG, C., PARMIGIANI, G. and DOMINICI, F. (2012). Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics* **68** 680–686. MR3055172 https://doi.org/10.1111/j.1541-0420.2011.01735.x

ZHU, R., ZENG, D. and KOSOROK, M. R. (2015). Reinforcement learning trees. *J. Amer. Statist. Assoc.* **110** 1770–1784. MR3449072 https://doi.org/10.1080/01621459.2015.1036994

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 768. MR2210692 https://doi.org/10.1111/j.1467-9868.2005.00527.x

Statistical Science
2024, Vol. 39, No. 2, 305–321
https://doi.org/10.1214/23-STS905

Bayesian Sample Size Determination for Causal Discovery

Federico Castelletti and Guido Consonni

Abstract. Graphical models based on Directed Acyclic Graphs (DAGs) are widely used to answer causal questions across a variety of scientific and social disciplines. However, observational data alone cannot distinguish in general between DAGs representing the same conditional independence assertions (Markov equivalent DAGs); as a consequence, the orientation of some edges in the graph remains indeterminate. Interventional data, produced by exogenous manipulations of variables in the network, enhance the process of structure learning because they allow to distinguish among equivalent DAGs, thus sharpening causal inference. Starting from an equivalence class of DAGs, a few procedures have been devised to produce a collection of variables to be manipulated in order to identify a causal DAG. Yet, these algorithmic approaches do not determine the sample size of the interventional data required to obtain a desired level of statistical accuracy. We tackle this problem from a Bayesian experimental design perspective, taking as input a sequence of target variables to be manipulated to identify edge orientation. We then propose a method to determine, at each intervention, the optimal sample size to produce an experiment which, with high assurance, will deliver an overall probability of decisive and correct evidence.

Key words and phrases: Active learning, Bayes factor, Bayesian experimental design, directed acyclic graph, intervention.

REFERENCES

[1] ADCOCK, C. J. (1997). Sample size determination: A review. *J. R. Stat. Soc., Ser. D, Stat.* **46** 261–283.

[2] ANDERSSON, S. A., MADIGAN, D. and PERLMAN, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *Ann. Statist.* **25** 505–541. MR1439312 https://doi.org/10.1214/aos/1031833662

[3] ANDERSSON, S. A., MADIGAN, D. and PERLMAN, M. D. (2001). Alternative Markov properties for chain graphs. *Scand. J. Stat.* **28** 33–85. MR1844349 https://doi.org/10.1111/1467-9469.00224

[4] BANDYOPADHYAY, P. S. and FORSTER, M. R., eds. (2011). Posterior model probabilities. In *Philosophy of Statistics. Handbook of the Philosophy of Science* **7**. Elsevier/North-Holland, Amsterdam. MR3295937 https://doi.org/10.1016/B978-0-444-51862-0.50001-0

[5] CASTELLETTI, F. and CONSONNI, G. (2019). Objective Bayes model selection of Gaussian interventional essential graphs for the identification of signaling pathways. *Ann. Appl. Stat.* **13** 2289–2311. MR4037431 https://doi.org/10.1214/19-aoas1275

[6] CASTELLETTI, F. and CONSONNI, G. (2020). Discovering causal structures in Bayesian Gaussian directed acyclic graph models. *J. Roy. Statist. Soc. Ser. A* **183** 1727–1745. MR4157833

[7] CASTELLETTI, F. and CONSONNI, G. (2021). Bayesian inference of causal effects from observational data in Gaussian graphical models. *Biometrics* **77** 136–149. MR4229727 https://doi.org/10.1111/biom.13281

[8] CASTELLETTI, F., CONSONNI, G., DELLA VEDOVA, M. L. and PELUSO, S. (2018). Learning Markov equivalence classes of directed acyclic graphs: An objective Bayes approach. *Bayesian Anal.* **13** 1231–1256. MR3855370 https://doi.org/10.1214/18-BA1101

[9] CASTELLETTI, F. and PELUSO, S. (2021). Equivalence class selection of categorical graphical models. *Comput. Statist. Data Anal.* **164** Paper No. 107304. MR4280200 https://doi.org/10.1016/j.csda.2021.107304

[10] CASTELLETTI, F. and PELUSO, S. (2023). Bayesian learning of network structures from interventional experimental data. *Biometrika* asad032.

[11] CASTELO, R. and PERLMAN, M. D. (2004). Learning essential graph Markov models from data. In *Advances in Bayesian Networks. Stud. Fuzziness Soft Comput.* **146** 255–269. Springer, Berlin. MR2090887 https://doi.org/10.1007/978-3-540-39879-0_14

[12] CHALONER, K. and VERDINELLI, I. (1995). Bayesian experimental design: A review. *Statist. Sci.* **10** 273–304. MR1390519

Federico Castelletti is Assistant Professor, Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milan, Italy (e-mail: federico.castelletti@unicatt.it). Guido Consonni is Professor, Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milan, Italy (e-mail: guido.consonni@unicatt.it).

[13] CHICKERING, D. M. (1995). A transformational characterization of equivalent Bayesian network structures. In *Uncertainty in Artificial Intelligence* (*Montreal*, *PQ*, 1995) 87–98. Morgan Kaufmann, San Francisco, CA. MR1615012

[14] CHICKERING, D. M. (2002). Learning equivalence classes of Bayesian-network structures. *J. Mach. Learn. Res.* **2** 445–498. MR1929415 https://doi.org/10.1162/153244302760200696

[15] CONSONNI, G. and LA ROCCA, L. (2012). Objective Bayes factors for Gaussian directed acyclic graphical models. *Scand. J. Stat.* **39** 743–756. MR3000846 https://doi.org/10.1111/j.1467-9469.2011.00785.x

[16] CONSONNI, G. and VERONESE, P. (2008). Compatibility of prior specifications across linear models. *Statist. Sci.* **23** 332–353. MR2483907 https://doi.org/10.1214/08-STS258

[17] COWELL, R. G., DAWID, A. P., LAURITZEN, S. L. and SPIEGELHALTER, D. J. (1999). *Probabilistic Networks and Expert Systems*. *Statistics for Engineering and Information Science*. Springer, New York. MR1697175

[18] DASGUPTA, A. (1996). Review of optimal Bayes designs. In *Design and Analysis of Experiments*. *Handbook of Statist.* **13** 1099–1147. North-Holland, Amsterdam. MR1492591 https://doi.org/10.1016/S0169-7161(96)13031-5

[19] DAWID, A. P. (1992). Prequential analysis, stochastic complexity and Bayesian inference. In *Bayesian Statistics*, 4 (*Peñíscola*, 1991) 109–125. Oxford Univ. Press, New York. MR1380273

[20] DAWID, A. P. (2010). Beware of the DAG!. In *Proceedings of Workshop on Causality*: *Objectives and Assessment at NIPS* 2008 (I. Guyon, D. Janzing and B. Schölkopf, eds.). *Proceedings of Machine Learning Research* **6** 59–86. PMLR, Whistler, Canada.

[21] DAWID, A. P. and LAURITZEN, S. L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21** 1272–1317. MR1241267 https://doi.org/10.1214/aos/1176349260

[22] DE SANTIS, F. (2004). Statistical evidence and sample size determination for Bayesian hypothesis testing. *J. Statist. Plann. Inference* **124** 121–144. MR2066230 https://doi.org/10.1016/S0378-3758(03)00198-8

[23] EBERHARDT, F. (2008). Almost optimal intervention sets for causal discovery. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*. *UAI'*08 161–168. AUAI Press, Arlington, VA, USA.

[24] ETZIONI, R. and KADANE, J. B. (1993). Optimal experimental design for another's analysis. *J. Amer. Statist. Assoc.* **88** 1404–1411. MR1245377

[25] FRIEDMAN, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science* **303** 799–805. https://doi.org/10.1126/science.1094068

[26] FROT, B., NANDY, P. and MAATHUIS, M. H. (2019). Robust causal structure learning with some hidden variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 459–487. MR3961495 https://doi.org/10.1111/rssb.12315

[27] GEIGER, D. and HECKERMAN, D. (2002). Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *Ann. Statist.* **30** 1412–1440. MR1936324 https://doi.org/10.1214/aos/1035844981

[28] HAO, W., SUO, F., LIN, Q., CHEN, Q., ZHOU, L., LIU, Z., CUI, W. and ZHOU, Z. (2020). Design and construction of portable CRISPR-Cpf1-mediated genome editing in bacillus subtilis 168 oriented toward multiple utilities. *Front. Bioeng. Biotechnol.* **8**.

[29] HAUSER, A. and BÜHLMANN, P. (2012). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *J. Mach. Learn. Res.* **13** 2409–2464. MR2973606

[30] HAUSER, A. and BÜHLMANN, P. (2014). Two optimal strategies for active learning of causal models from interventional data. *Internat. J. Approx. Reason.* **55** 926–939. MR3178409 https://doi.org/10.1016/j.ijar.2013.11.007

[31] HAUSER, A. and BÜHLMANN, P. (2015). Jointly interventional and observational data: Estimation of interventional Markov equivalence classes of directed acyclic graphs. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 291–318. MR3299409 https://doi.org/10.1111/rssb.12071

[32] HE, Y.-B. and GENG, Z. (2008). Active learning of causal networks with intervention experiments and optimal designs. *J. Mach. Learn. Res.* **9** 2523–2547. MR2460892

[33] HYTTINEN, A., EBERHARDT, F. and HOYER, P. O. (2013). Experiment selection for causal discovery. *J. Mach. Learn. Res.* **14** 3041–3071. MR3138909

[34] IMBENS, G. W. (2020). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *J. Econ. Lit.* **58** 1129–1179.

[35] JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Clarendon Press, Oxford. MR0187257

[36] JOHNSON, V. E. and ROSSELL, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 143–170. MR2830762 https://doi.org/10.1111/j.1467-9868.2009.00730.x

[37] KALISCH, M. and BÜHLMANN, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.* **8** 613–636.

[38] KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795. MR3363402 https://doi.org/10.1080/01621459.1995.10476572

[39] KOLLER, D. and FRIEDMAN, N. (2009). *Probabilistic Graphical Models*: *Principles and Techniques*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR2778120

[40] LAURITZEN, S. L. (1996). *Graphical Models*. *Oxford Statistical Science Series* **17**. Oxford University Press, New York. MR1419991

[41] LINDLEY, D. V. (1971). *Bayesian Statistics*, *a Review*. *Conference Board of the Mathematical Sciences Regional Conference Series in Applied Mathematics*, *No.* 2. SIAM, Philadelphia, PA. MR0329081

[42] LINDLEY, D. V. (1997). The choice of sample size. *J. R. Stat. Soc.*, *Ser. D*, *Stat.* **46** 129–138.

[43] MAATHUIS, M. H., KALISCH, M. and BÜHLMANN, P. (2009). Estimating high-dimensional intervention effects from observational data. *Ann. Statist.* **37** 3133–3164. MR2549555 https://doi.org/10.1214/09-AOS685

[44] MEGANCK, S., LERAY, P. and MANDERICK, B. (2006). Learning causal Bayesian networks from observations and experiments: A decision theoretic approach. In *Modeling Decisions for Artificial Intelligence* (V. Torra, Y. Narukawa, A. Valls and J. Domingo-Ferrer, eds.) 58–69. Springer Berlin Heidelberg, Berlin, Heidelberg.

[45] MUIRHEAD, R. J. (1982). *Aspects of Multivariate Statistical Theory*. *Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. MR0652932

[46] NAGARAJAN, R., SCUTARI, M. and LÈBRE, S. (2013). *Bayesian Networks in R with Applications in Systems Biology*. *Use R!* Springer, New York. MR3059206 https://doi.org/10.1007/978-1-4614-6446-4

[47] O'HAGAN, A. (1995). Fractional Bayes factors for model comparison. *J. Roy. Statist. Soc. Ser. B* **57** 99–138. MR1325379

[48] O'HAGAN, A. and STEVENS, J. W. (2001). Bayesian assessment of sample size for clinical trials of cost-effectiveness.

Med. Decis. Mak. **21** 219–230. https://doi.org/10.1177/0272989X0102100307

[49] PAN, J. and BANERJEE, S. (2021). A unifying Bayesian approach for sample size determination using design and analysis priors. ArXiv preprint. Available at arXiv:2112.03509.

[50] PEARL, J. (2000). *Causality*: *Models*, *Reasoning*, *and Inference*. Cambridge Univ. Press, Cambridge. MR1744773

[51] PEARL, J. (2003). Statistics and causal inference: A review. *TEST* **12** 281–318. MR2044313 https://doi.org/10.1007/BF02595718

[52] PENG, S., SHEN, X. and PAN, W. (2020). Reconstruction of a directed acyclic graph with intervention. *Electron. J. Stat.* **14** 4133–4164. MR4175391 https://doi.org/10.1214/20-EJS1767

[53] PETERS, J. and BÜHLMANN, P. (2014). Identifiability of Gaussian structural equation models with equal error variances. *Biometrika* **101** 219–228. MR3180667 https://doi.org/10.1093/biomet/ast043

[54] PRESS, S. J. (1982). *Applied Multivariate Analysis*: *Using Bayesian and Frequentist Methods of Inference*. Krieger Publishing Company, Malabar, FL.

[55] RAIFFA, H. and SCHLAIFER, R. (1961). *Applied Statistical Decision Theory*. *Harvard Business School Publications*. Division of Research, Graduate School of Business Adminitration, Harvard Univ. MR0117844

[56] ROYALL, R. (2000). On the probability of observing misleading statistical evidence. *J. Amer. Statist. Assoc.* **95** 760–780. MR1803877 https://doi.org/10.2307/2669456

[57] ROYALL, R. M. (1997). *Statistical Evidence*: *A Likelihood Paradigm*. *Monographs on Statistics and Applied Probability* **71**. CRC Press, London. MR1629481

[58] SACHS, K., PEREZ, O., PE'ER, D., LAUFFENBURGER, D. A. and NOLAN, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308** 523–529. https://doi.org/10.1126/science.1105809

[59] SCHÖNBRODT, F. D. and WAGENMAKERS, E. J. (2017). Bayes factor design analysis: Planning for compelling evidence. *Psychon. Bull. Rev.* **25** 128–142.

[60] SHOJAIE, A. and MICHAILIDIS, G. (2009). Analysis of gene sets based on the underlying regulatory network. *J. Comput. Biol.* **16** 407–426. MR2487566 https://doi.org/10.1089/cmb.2008.0081

[61] SPIEGELHALTER, D. J., ABRAMS, K. R. and MYLES, J. P. (2003). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Wiley, New York.

[62] SPIEGELHALTER, D. J. and FREEDMAN, L. S. (1986). A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Stat. Med.* **5** 1–13. https://doi.org/10.1002/sim.4780050103

[63] SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation*, *Prediction*, *and Search*, 2nd ed. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR1815675

[64] SQUIRES, C., MAGLIACANE, S., GREENEWALD, K., KATZ, D., KOCAOGLU, M. and SHANMUGAM, K. (2020). Active structure learning of causal DAGs via directed clique trees. In *Proceedings of the* 34*th International Conference on Neural Information Processing Systems*. *NIPS'*20. Curran Associates Inc., Red Hook, NY, USA.

[65] STEFAN, A. M., SCHÖNBRODT, F. D., EVANS, N. J. and WAGENMAKERS, E. J. (2022). Efficiency in sequential testing: Comparing the sequential probability ratio test and the sequential Bayes factor test. *Behav. Res. Methods* **54** 1554–3528.

[66] R CORE TEAM (2021). *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[67] TONG, S. and KOLLER, D. (2001). Active learning for structure in Bayesian networks. In *Proceedings of the* 17*th International Joint Conference on Artificial Intelligence - Volume* 2. *IJCAI'*01 863–869. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[68] VERMA, T. and PEARL, J. (1990). Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*. *UAI* 90 255–270. Elsevier Science Inc., New York, NY, USA.

[69] VON KÜGELGEN, J., RUBENSTEIN, P. K., SCHÖLKOPF, B. and WELLER, A. (2019). Optimal experimental design via Bayesian optimization: Active causal structure learning for Gaussian process networks. In *NeurIPS* 2019 *Workshop do the Right Thing*: *Machine Learning and Causal Inference for Improved Decision Making*.

[70] WANG, F. and GELFAND, A. E. (2002). A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statist. Sci.* **17** 193–208. MR1925941 https://doi.org/10.1214/ss/1030550861

[71] WEISS, R. (1997). Bayesian sample size calculations for hypothesis testing. *J. R. Stat. Soc., Ser. D, Stat.* **46** 185–191.

[72] YANG, K., KATCOFF, A. and UHLER, C. (2018). Characterizing and learning equivalence classes of causal DAGs under interventions. In *Proceedings of the* 35*th International Conference on Machine Learning* (J. Dy and A. Krause, eds.). *Proceedings of Machine Learning Research* **80** 5541–5550. PMLR.

[73] ZHANG, K., DUAN, X. and WU, J. (2016). Multigene disruption in undomesticated Bacillus subtilis ATCC 6051a using the CRISPR/Cas9 system. *Sci. Rep.* **6** 27943.

Statistical Science
2024, Vol. 39, No. 2, 322–345
https://doi.org/10.1214/23-STS910

Likelihood Asymptotics in Nonregular Settings: A Review with Emphasis on the Likelihood Ratio

Alessandra R. Brazzale and Valentina Mameli

Abstract. This paper reviews the most common situations in which the regularity conditions that underlie classical likelihood-based parametric inference fail, focusing on the large-sample properties of the likelihood ratio statistic. We identify three main classes of problems: boundary problems, indeterminate parameter problems—which include nonidentifiable parameters and singular information matrices—and change-point problems. We emphasise analytical solutions, consider software implementations where available, and summarise how the key results are derived.

Key words and phrases: Boundary point, change-point, finite mixture model, first order theory, identifiability, large-sample inference, singular information.

REFERENCES

ALGERI, S., AALBERS, J., MORA, K. D. and CONRAD, J. (2020). Searching for new phenomena with profile likelihood ratio tests. *Nat. Rev. Phys.* **2** 245–252.

ALGERI, S. and VAN DYK, D. A. (2020). Testing one hypothesis multiple times: The multidimensional case. *J. Comput. Graph. Statist.* **29** 358–371. MR4116048 https://doi.org/10.1080/10618600.2019.1677474

ANDREWS, D. W. K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica* **68** 399–405. MR1748009 https://doi.org/10.1111/1468-0262.00114

ANDREWS, D. W. K. (2001). Testing when a parameter is on the boundary of the maintained hypothesis. *Econometrica* **69** 683–734. MR1828540 https://doi.org/10.1111/1468-0262.00210

AUE, A. and HORVÁTH, L. (2013). Structural breaks in time series. *J. Time Series Anal.* **34** 1–16. MR3008012 https://doi.org/10.1111/j.1467-9892.2012.00819.x

AZZALINI, A. (1996). *Statistical Inference Based on the Likelihood. Monographs on Statistics and Applied Probability* **68**. CRC Press, London. MR1455371

BAEY, C. and KUHN, E. (2019). varTestnlme: Variance components testing in mixed-effect models. Available at https://github.com/baeyc/varTestnlme.

BANERJEE, M. (2005). Likelihood ratio tests under local alternatives in regular semiparametric models. *Statist. Sinica* **15** 635–644. MR2233903

BANERJEE, M. (2007). Likelihood based inference for monotone response models. *Ann. Statist.* **35** 931–956. MR2341693 https://doi.org/10.1214/009053606000001578

BANERJEE, M. and WELLNER, J. A. (2001). Likelihood ratio tests for monotone functions. *Ann. Statist.* **29** 1699–1731. MR1891743 https://doi.org/10.1214/aos/1015345959

BARNDORFF-NIELSEN, O. E. and COX, D. R. (1994). *Inference and Asymptotics. Monographs on Statistics and Applied Probability* **52**. CRC Press, London. MR1317097 https://doi.org/10.1007/978-1-4899-3210-5

BELLEC, P. C. (2018). Sharp oracle inequalities for least squares estimators in shape restricted regression. *Ann. Statist.* **46** 745–780. MR3782383 https://doi.org/10.1214/17-AOS1566

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

BHATTACHARYA, P. K. (1994). Some aspects of change-point analysis. In *Change-Point Problems* (*South Hadley, MA*, 1992). *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **23** 28–56. IMS, Hayward. MR1477912 https://doi.org/10.1214/lnms/1215463112

BLISCHKE, W. R., TRUELOVE, A. J. and MUNDLE, P. B. (1969). On non-regular estimation. I. Variance bounds for estimators of location parameters. *J. Amer. Statist. Assoc.* **64** 1056–1072.

BÖHNING, D., DIETZ, E., SCHAUB, R., SCHLATTMANN, P. and LINDSAY, B. G. (1994). The distribution of the likelihood ratio for mixture of densities from the one-parameter exponential family. *Ann. Inst. Statist. Math.* **46** 373–388.

BRAZZALE, A. R., DAVISON, A. C. and REID, N. (2007). *Applied Asymptotics: Case Studies in Small-Sample Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **23**. Cambridge Univ. Press, Cambridge. MR2342742 https://doi.org/10.1017/CBO9780511611131

Alessandra R. Brazzale is Associate Professor of Statistics, Department of Statistical Sciences, University of Padova, Via Cesare Battisti 241, 35121 Padova, Italy (e-mail: alessandra.brazzale@unipd.it; URL: https://homes.stat.unipd.it/alessandrarosalbabrazzale.html). Valentina Mameli is Associate Professor of Statistics, Department of Economics and Statistics, University of Udine, Via Tomadini 30/A, 33100 Udine, Italy (e-mail: valentina.mameli@uniud.it; URL: https://people.uniud.it/page/valentina.mameli).

BRAZZALE, A. R., KÜCHENHOFF, H., KRÜGEL, S., SCHIERGENS, T. S., TRENTZSCH, H. and HARTL, W. (2019). Nonparametric change point estimation for survival distributions with a partially constant hazard rate. *Lifetime Data Anal*. **25** 301–321. MR3924678 https://doi.org/10.1007/s10985-018-9431-x

BRAZZALE, A. R. and MAMELI, V. (2024). Supplement to "Likelihood asymptotics in nonregular settings: A review with emphasis on the likelihood ratio". https://doi.org/10.1214/23-STS910SUPP

CAVALIERE, G., NIELSEN, H. B., PEDERSEN, R. S. and RAHBEK, A. (2022). Bootstrap inference on the boundary of the parameter space, with application to conditional volatility models. *J. Econometrics* **227** 241–263. MR4377183 https://doi.org/10.1016/j.jeconom.2020.05.006

CHAUVEAU, D., GAREL, B. and MERCIER, S. (2018). Testing for univariate Gaussian mixture in practice. Available at https://hal.science/hal-01659771/, Version 2.

CHEN, H. and CHEN, J. (2001). Large sample distribution of the likelihood ratio test for normal mixtures. *Statist. Probab. Lett*. **52** 125–133. MR1841402 https://doi.org/10.1016/S0167-7152(00)00171-1

CHEN, H. and CHEN, J. (2003). Tests for homogeneity in normal mixtures in the presence of a structural parameter. *Statist. Sinica* **13** 351–365. MR1977730

CHEN, H., CHEN, J. and KALBFLEISCH, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **63** 19–29. MR1811988 https://doi.org/10.1111/1467-9868.00273

CHEN, J. (2017). On finite mixture models. *Stat. Theory Relat. Fields* **1** 15–27.

CHEN, J. and GUPTA, A. K. (2012). *Parametric Statistical Change Point Analysis*: *With Applications to Genetics*, *Medicine*, *and Finance*, 2nd ed. Birkhäuser/Springer, New York. MR3025631 https://doi.org/10.1007/978-0-8176-4801-5

CHEN, J. and LI, P. (2009). Hypothesis test for normal mixture models: The EM approach. *Ann. Statist*. **37** 2523–2542. MR2543701 https://doi.org/10.1214/08-AOS651

CHEN, R. and CABRERA, J. (2020). Bootstrap confidence intervals using the likelihood ratio test in changepoint detection. Available at https://arxiv.org/abs/2011.03718.

CHEN, S. X. and VAN KEILEGOM, I. (2009). A review on empirical likelihood methods for regression. *TEST* **18** 415–447. MR2566404 https://doi.org/10.1007/s11749-009-0159-5

CHENG, R. (2017). *Non-standard Parametric Statistical Inference*. Oxford Univ. Press, Oxford. MR3701991 https://doi.org/10.1093/oso/9780198505044.001.0001

CHENG, R. C. H. and TRAYLOR, L. (1995). Non-regular maximum likelihood problems. *J. Roy. Statist. Soc. Ser. B* **57** 3–44. With discussion and a reply by the authors. MR1325377

CHERNOFF, H. (1954). On the distribution of the likelihood ratio. *Ann. Math. Stat*. **25** 573–578. MR0065087 https://doi.org/10.1214/aoms/1177728725

CHERNOFF, H. and LANDER, E. (1995). Asymptotic distribution of the likelihood ratio test that a mixture of two binomials is a single binomial. *J. Statist. Plann. Inference* **43** 19–40. MR1314126 https://doi.org/10.1016/0378-3758(94)00006-H

CIUPERCA, G. (2002). Likelihood ratio statistic for exponential mixtures. *Ann. Inst. Statist. Math*. **54** 585–594. MR1932403 https://doi.org/10.1023/A:1022415228062

CONG, L. and YAO, W. (2021). A likelihood ratio test of a homoscedastic multivariate normal mixture against a heteroscedastic multivariate normal mixture. *Econom. Stat*. **18** 79–88. MR4238909 https://doi.org/10.1016/j.ecosta.2021.01.002

COX, D. R. (2006). *Principles of Statistical Inference*. Cambridge Univ. Press, Cambridge. MR2278763 https://doi.org/10.1017/CBO9780511813559

COX, D. R. and HINKLEY, D. V. (1974). *Theoretical Statistics*. CRC Press, London. MR0370837

CRAINICEANU, C. M. and RUPPERT, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **66** 165–185. MR2035765 https://doi.org/10.1111/j.1467-9868.2004.00438.x

CRAMÉR, H. (1946). *Mathematical Methods of Statistics*. *Princeton Mathematical Series*, *Vol*. 9. Princeton Univ. Press, Princeton. MR0016588

CSÖRGŐ, M. and HORVÁTH, L. (1997). *Limit Theorems in Change-Point Analysis*. *Wiley Series in Probability and Statistics*. Wiley, Chichester. With a foreword by David Kendall. MR2743035

DACUNHA-CASTELLE, D. and GASSIAT, É. (1997). Testing in locally conic models, and application to mixture models. *ESAIM Probab. Stat*. **1** 285–317. MR1468112 https://doi.org/10.1051/ps:1997111

DACUNHA-CASTELLE, D. and GASSIAT, E. (1999). Testing the order of a model using locally conic parametrization: Population mixtures and stationary ARMA processes. *Ann. Statist*. **27** 1178–1209. MR1740115 https://doi.org/10.1214/aos/1017938921

DARLING, D. A. and ERDŐS, P. (1956). A limit theorem for the maximum of normalized sums of independent random variables. *Duke Math. J*. **23** 143–155. MR0074712

DASGUPTA, A. (2008). *Asymptotic Theory of Statistics and Probability*. *Springer Texts in Statistics*. Springer, New York. MR2664452

DAVISON, A. C. (2003). *Statistical Models*. *Cambridge Series in Statistical and Probabilistic Mathematics* **11**. Cambridge Univ. Press, Cambridge. MR1998913 https://doi.org/10.1017/CBO9780511815850

DEL CASTILLO, J. and LÓPEZ-RATERA, A. (2006). Saddlepoint approximation in exponential models with boundary points. *Bernoulli* **12** 491–500. MR2232728 https://doi.org/10.3150/bj/1151525132

DETTE, H. and GÖSMANN, J. (2020). A likelihood ratio approach to sequential change point detection for a general class of parameters. *J. Amer. Statist. Assoc*. **115** 1361–1377. MR4143471 https://doi.org/10.1080/01621459.2019.1630562

DOSS, C. R. and WELLNER, J. A. (2019). Inference for the mode of a log-concave density. *Ann. Statist*. **47** 2950–2976. MR3988778 https://doi.org/10.1214/18-AOS1770

EFRON, B. and HASTIE, T. (2016). *Computer Age Statistical Inference*: *Algorithms*, *Evidence*, *and Data Science*. *Institute of Mathematical Statistics* (*IMS*) *Monographs* **5**. Cambridge Univ. Press, New York. MR3523956 https://doi.org/10.1017/CBO9781316576533

ELKANTASSI, S., BELLIO, R., BRAZZALE, A. R. and DAVISON, A. C. (2023). Improved inference for a boundary parameter. *Canad. J. Statist*. **51** 780–799. MR4635242 https://doi.org/10.1002/cjs.11791

ERDMAN, C. and EMERSON, J. W. (2007). bcp: An R package for performing a Bayesian analysis of change point problems. *J. Stat. Softw*. **23** 1–13.

FENG, C., WANG, H. and TU, X. M. (2012). The asymptotic distribution of a likelihood ratio test statistic for the homogeneity of Poisson distribution. *Sankhya A* **74** 263–268. MR3021560 https://doi.org/10.1007/s13171-012-0003-y

FENG, Z. and MCCULLOCH, C. E. (1992). Statistical inference using maximum likelihood estimation and the generalized likelihood ratio when the true parameter is on the boundary of the parameter space. *Statist. Probab. Lett*. **13** 325–332. MR1160755 https://doi.org/10.1016/0167-7152(92)90042-4

FENG, Z. and MCCULLOCH, C. E. (1996). Using bootstrap likelihood ratios in finite mixture models. *J. Roy. Statist. Soc. Ser. B* **58** 609–617.

FU, Y., CHEN, J. and LI, P. (2008). Modified likelihood ratio test for homogeneity in a mixture of von Mises distributions. *J. Statist.*

Plann. Inference **138** 667–681. MR2382561 https://doi.org/10.1016/j.jspi.2007.01.003

GAREL, B. (2007). Recent asymptotic results in testing for mixtures. *Comput. Statist. Data Anal.* **51** 5295–5304. MR2370872 https://doi.org/10.1016/j.csda.2006.09.033

GEYER, C. J. (1994). On the asymptotics of constrained *M*-estimation. *Ann. Statist.* **22** 1993–2010. MR1329179 https://doi.org/10.1214/aos/1176325768

GHOSAL, S. and VAN DER VAART, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. *Cambridge Series in Statistical and Probabilistic Mathematics* **44**. Cambridge Univ. Press, Cambridge. MR3587782 https://doi.org/10.1017/9781139029834

GHOSH, J. K. and SEN, P. K. (1985). On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. II* (*Berkeley, Calif.*, 1983). *Wadsworth Statist./Probab. Ser.* 789–806. Wadsworth, Belmont. MR0822065

GODAMBE, V. P. (1991). *Estimating Functions*. *Oxford Statistical Science Series* **7**. Clarendon, Oxford. MR1163992

GOFFINET, B., LOISEL, P. and LAURENT, B. (1992). Testing in normal mixture models when the proportions are known. *Biometrika* **79** 842–846. MR1209483 https://doi.org/10.1093/biomet/79.4.842

GROENEBOOM, P. and JONGBLOED, G. (2018). Some developments in the theory of shape constrained inference. *Statist. Sci.* **33** 473–492. MR3881204 https://doi.org/10.1214/18-STS657

GROENEBOOM, P. and WELLNER, J. A. (2001). Computing Chernoff's distribution. *J. Comput. Graph. Statist.* **10** 388–400. MR1939706 https://doi.org/10.1198/10618600152627997

HAN, Q., SEN, B. and SHEN, Y. (2022). High-dimensional asymptotics of likelihood ratio tests in the Gaussian sequence model under convex constraints. *Ann. Statist.* **50** 376–406. MR4382021 https://doi.org/10.1214/21-aos2111

HARTIGAN, J. A. (1985). A failure of likelihood asymptotics for normal mixtures. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. II* (*Berkeley, Calif.*, 1983). *Wadsworth Statist./Probab. Ser.* 807–810. Wadsworth, Belmont. MR0822066

HATHAWAY, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Ann. Statist.* **13** 795–800. MR0790575 https://doi.org/10.1214/aos/1176349557

HAUGHTON, D. (1997). Packages for estimating finite mixtures: A review. *Amer. Statist.* **51** 194–205.

HAWKINS, D. M. (1977). Testing a sequence of observations for a shift in location. *J. Amer. Statist. Assoc.* **72** 180–186. MR0451496

HIRANO, K. and PORTER, J. R. (2003). Asymptotic efficiency in parametric structural models with parameter-dependent support. *Econometrica* **71** 1307–1338. MR2000249 https://doi.org/10.1111/1468-0262.00451

HOGG, R. V., MCKEAN, J. W. and CRAIG, A. T. (2019). *Introduction to Mathematical Statistics*, 8th ed. Pearson, Boston, MA.

HORVÁTH, L. and RICE, G. (2014a). Extensions of some classical methods in change point analysis. *TEST* **23** 219–255. MR3210268 https://doi.org/10.1007/s11749-014-0368-4

HORVÁTH, L. and RICE, G. (2014b). Rejoinder on: Extensions of some classical methods in change point analysis [MR3210269; MR3210270; MR3210271; MR3210272; MR3210273; MR3210274; MR3210275; MR3210276; MR3210268]. *TEST* **23** 287–290. MR3210277 https://doi.org/10.1007/s11749-014-0375-5

HUBER, P. J. and RONCHETTI, E. M. (2009). *Robust Statistics*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken. MR2488795 https://doi.org/10.1002/9780470434697

HUŠKOVÁ, M. and KIRCH, C. (2012). Bootstrapping sequential change-point tests for linear regression. *Metrika* **75** 673–708. MR2946654 https://doi.org/10.1007/s00184-011-0347-7

HUZURBAZAR, V. S. (1948). The likelihood equation, consistency and the maxima of the likelihood function. *Ann. Eugen.* **14** 185–200. MR0028000

IRVINE, J. M. (1986). *The Asymptotic Distribution of the Likelihood Ratio Test for a Change in the Mean*. *Statistical Research Division Report Series* **CENSUS/SRD/RR-86/10**. Bureau of the Census, Washington.

JARUŠKOVÁ, D. (1997). Some problems with application of change-point detection methods to environmental data. *Environmetrics* **8** 469–483.

KASAHARA, H. and SHIMOTSU, K. (2015). Testing the number of components in normal mixture regression models. *J. Amer. Statist. Assoc.* **110** 1632–1645. MR3449060 https://doi.org/10.1080/01621459.2014.986272

KHODADADI, A. and ASGHARIAN, M. (2008). Change-point problem and regression: An annotated bibliography. COBRA preprint series. Working paper 44. Available at biostats.bepress.com/cobra/art44.

KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Stat.* **27** 887–906. MR0086464 https://doi.org/10.1214/aoms/1177728066

KILLICK, R. and ECKLEY, I. A. (2014). Changepoint: An R package for changepoint analysis. *J. Stat. Softw.* **58** 1–19.

KIM, H.-J. and SIEGMUND, D. (1989). The likelihood ratio test for a change-point in simple linear regression. *Biometrika* **76** 409–423. MR1040636 https://doi.org/10.1093/biomet/76.3.409

KIRCH, C. (2008). Bootstrapping sequential change-point tests. *Sequential Anal.* **27** 330–349. MR2446906 https://doi.org/10.1080/07474940802241082

KIRCH, C. and STEINEBACH, J. (2006). Permutation principles for the change analysis of stochastic processes under strong invariance. *J. Comput. Appl. Math.* **186** 64–88. MR2190298 https://doi.org/10.1016/j.cam.2005.03.065

KOENKER, R., CHERNOZHUKOV, V., HE, X. and PENG, L., eds. (2018). *Handbook of Quantile Regression*. *Chapman & Hall/CRC Handbooks of Modern Statistical Methods*. CRC Press, Boca Raton. MR3728340

KOPYLEV, L. (2012). Constrained parameters in applications: Review of issues and approaches. *ISRN Biomathematics*. https://doi.org/10.5402/2012/872956

KOPYLEV, L. and SINHA, B. (2011). On the asymptotic distribution of likelihood ratio test when parameters lie on the boundary. *Sankhya B* **73** 20–41. MR2826318 https://doi.org/10.1007/s13571-011-0022-z

KRISHNAIAH, P. R. and MIAO, B. Q., eds. (1988). Review about estimation of change points. In *Quality Control and Reliability*. *Handbook of Statistics* **7** 375–402. North-Holland, Amsterdam. MR0975506

KUDÔ, A. (1963). A multivariate analogue of the one-sided test. *Biometrika* **50** 403–418. MR0163386 https://doi.org/10.1093/biomet/50.3-4.403

LANCASTER, T. (2000). The incidental parameter problem since 1948. *J. Econometrics* **95** 391–413. MR1752336 https://doi.org/10.1016/S0304-4076(99)00044-5

LE CAM, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *Ann. Math. Stat.* **41** 802–828. MR0267676 https://doi.org/10.1214/aoms/1177696960

LE CAM, L. and YANG, G. L. (1990). *Asymptotics in Statistics*: *Some Basic Concepts*. *Springer Series in Statistics*. Springer, New York. MR1066869 https://doi.org/10.1007/978-1-4684-0377-0

LEE, T.-S. (2010). Change-point problems: Bibliography and review. *J. Stat. Theory Pract.* **4** 643–662. MR2758751 https://doi.org/10.1080/15598608.2010.10412010

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

LEMDANI, M. and PONS, O. (1997). Likelihood ratio tests for genetic linkage. *Statist. Probab. Lett.* **33** 15–22. MR1451126 https://doi.org/10.1016/S0167-7152(96)00105-8

LEMDANI, M. and PONS, O. (1999). Likelihood ratio tests in contamination models. *Bernoulli* **5** 705–719. MR1704563 https://doi.org/10.2307/3318698

LI, P., CHEN, J. and MARRIOTT, P. (2009). Non-finite Fisher information and homogeneity: An EM approach. *Biometrika* **96** 411–426. MR2507152 https://doi.org/10.1093/biomet/asp011

LI, S., CHEN, J. and LI, P. (2016). MixtureInf: Inference for finite mixture models. R package version 1.1. Available at https://CRAN.R-project.org/package=MixtureInf. MR2706578

LINDSAY, B. G. (1995). *Mixture models*: *Theory, geometry and applications*. *NSF-CBMS Regional Conf. Ser. Probab. Statist.* **5**. IMS, Hayward, CA. MR0994263 https://doi.org/10.1214/aos/1176347138

LIU, X. and SHAO, Y. (2003). Asymptotics for likelihood ratio tests under loss of identifiability. *Ann. Statist.* **31** 807–832. MR1994731 https://doi.org/10.1214/aos/1056562463

LO, Y. (2008). A likelihood ratio test of a homoscedastic normal mixture against a heteroscedastic normal mixture. *Stat. Comput.* **18** 233–240. MR2413381 https://doi.org/10.1007/s11222-008-9052-4

MCLACHLAN, G. and PEEL, D. (2000). *Finite Mixture Models*. *Wiley Series in Probability and Statistics*: *Applied Probability and Statistics*. Wiley-Interscience, New York. MR1789474 https://doi.org/10.1002/0471721182

MCLACHLAN, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *J. R. Stat. Soc., Ser. C* **36** 318–324.

MCLACHLAN, G. J., LEE, S. X. and RATHNAYAKE, S. I. (2019). Finite mixture models. *Annu. Rev. Stat. Appl.* **6** 355–378. MR3939525 https://doi.org/10.1146/annurev-statistics-031017-100325

MUGGEO, V. M. R. (2008a). Modeling temperature effects on mortality: Multiple segmented relationships with common break points. *Biostatistics* **9** 613–620.

MUGGEO, V. M. R. (2008b). segmented: An R package to fit regression models with broken-line relationships. *R News* **8** 20–25.

MUGGEO, V. M. R., ATKINS, D. C., GALLOP, R. J. and DIMIDJIAN, S. (2014). Segmented mixed models with random changepoints: A maximum likelihood approach with application to treatment for depression study. *Stat. Model.* **14** 293–313. MR3248010 https://doi.org/10.1177/1471082X13504721

MURPHY, S. A. and VAN DER VAART, A. W. (1997). Semiparametric likelihood ratio inference. *Ann. Statist.* **25** 1471–1509. MR1463562 https://doi.org/10.1214/aos/1031594729

MURPHY, S. A. and VAN DER VAART, A. W. (2000). On profile likelihood. *J. Amer. Statist. Assoc.* **95** 449–465. With comments and a rejoinder by the authors. MR1803168 https://doi.org/10.2307/2669386

NEYMAN, J. and SCOTT, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16** 1–32. MR0025113 https://doi.org/10.2307/1914288

NIU, Y. S., HAO, N. and ZHANG, H. (2016). Multiple change-point detection: A selective overview. *Statist. Sci.* **31** 611–623. MR3598742 https://doi.org/10.1214/16-STS587

OWEN, A. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.* **18** 90–120. MR1041387 https://doi.org/10.1214/aos/1176347494

OWEN, A. (1991). Empirical likelihood for linear models. *Ann. Statist.* **19** 1725–1747. MR1135146 https://doi.org/10.1214/aos/1176348368

PAGE, E. S. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika* **42** 523–527. MR0072412 https://doi.org/10.1093/biomet/42.3-4.523

PAGE, E. S. (1957). On problems in which a change in a parameter occurs at an unknown point. *Biometrika* **44** 248–252.

PAULINO, C. D. M. and PEREIRA, C. A. B. (1994). On identifiability of parametric statistical models. *J. Ital. Stat. Soc.* **1** 125–151.

PFANZAGL, J. (2017). *Mathematical Statistics*: *Essays on History and Methodology*. *Springer Series in Statistics*. Springer, Berlin. MR3822358

PIEGORSCH, W. W. and BAILER, A. J. (1997). *Statistics for Environmental Biology and Toxicology*. CRC Press, London.

PRAKASA RAO, B. L. S. (1992). *Identifiability in Stochastic Models*: *Characterization of Probability Distributions*. *Probability and Mathematical Statistics*. Academic Press, Boston, MA. MR1171013

QUANDT, R. E. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes. *J. Amer. Statist. Assoc.* **53** 873–880. MR0100314

QUANDT, R. E. (1960). Tests of the hypothesis that a linear regression system obeys two separate regimes. *J. Amer. Statist. Assoc.* **55** 324–330. MR0114269

R CORE TEAM (2020). *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

ROBERTSON, T., WRIGHT, F. T. and DYKSTRA, R. L. (1988). *Order Restricted Statistical Inference*. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. Wiley, Chichester. MR0961262

ROTHENBERG, T. J. (1971). Identification in parametric models. *Econometrica* **39** 577–591. MR0436944 https://doi.org/10.2307/1913267

ROTNITZKY, A., COX, D. R., BOTTAI, M. and ROBINS, J. (2000). Likelihood-based inference with singular information matrix. *Bernoulli* **6** 243–284. MR1748721 https://doi.org/10.2307/3318576

SAMWORTH, R. J. (2018). Recent progress in log-concave density estimation. *Statist. Sci.* **33** 493–509. MR3881205 https://doi.org/10.1214/18-STS666

SAMWORTH, R. J. and BODHISATTVA, B. (2018). Special issue on "Nonparametric inference under shape constraints". *Statist. Sci.* **33** 469–472. MR3881203 https://doi.org/10.1214/18-STS673

SCHEIPL, F., GREVEN, S. and KÜCHENHOFF, H. (2008). Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Comput. Statist. Data Anal.* **52** 3283–3299. MR2427346 https://doi.org/10.1016/j.csda.2007.10.022

SCRUCCA, L., FOP, M., MURPHY, T. B. and RAFTERY, A. E. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R J.* **8** 289–317.

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

SEN, P. K. and SILVAPULLE, M. J. (2002). An appraisal of some aspects of statistical inference under inequality constraints. *J. Statist. Plann. Inference* **107** 3–43. MR1927753 https://doi.org/10.1016/S0378-3758(02)00242-2

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

SEVERINI, T. A. (2000). *Likelihood Methods in Statistics*. *Oxford Statistical Science Series* **22**. Oxford Univ. Press, Oxford. MR1854870

SEVERINI, T. A. (2004). A modified likelihood ratio statistic for some nonregular models. *Biometrika* **91** 603–612. MR2090625 https://doi.org/10.1093/biomet/91.3.603

SHABAN, S. A. (1980). Change point problem and two-phase regression: An annotated bibliography. *Int. Stat. Rev.* **48** 83–93. MR0576777

SHAPIRO, A. (1985). Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. *Biometrika* **72** 133–144. MR0790208 https://doi.org/10.1093/biomet/72.1.133

SHAPIRO, A. (1988). Towards a unified theory of inequality constrained testing in multivariate analysis. *Int. Stat. Rev.* **56** 49–62. MR0963140 https://doi.org/10.2307/1403361

SILVAPULLE, M. J. and SEN, P. K. (2005). *Constrained Statistical Inference*: *Inequality, Order, and Shape Restrictions*. *Wiley Series in Probability and Statistics*. Wiley-Interscience, New York. MR2099529

SILVEY, S. D. (1959). The Lagrangian multiplier test. *Ann. Math. Stat.* **30** 389–407. MR0104307 https://doi.org/10.1214/aoms/1177706259

SINHA, B. K., KOPYLEV, L. and FOX, J. (2012). Some new aspects of statistical inference for multistage dose-response models with applications. *Pak. J. Stat. Oper. Res.* **8** 441–478. MR2975211 https://doi.org/10.18187/pjsor.v8i3.519

SMITH, A. F. M. and COOK, D. G. (1980). Straight lines with a change-point: A Bayesian analysis of some renal transplant data. *J. R. Stat. Soc., Ser. C* **29** 180–189. MR0585607 https://doi.org/10.2307/2986304

SMITH, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika* **72** 67–90. MR0790201 https://doi.org/10.1093/biomet/72.1.67

SMITH, R. L. (1989). A survey of nonregular problems. *Bull. Inst. Int. Stat.* **53** 353–372. MR1093694

SOFRONOV, G., WENDLER, M. and LIEBSCHER, V. (eds.) (2020). Part 1: Special Issue on Change Point Detection (first 10 articles). *Statist. Papers* **61** 1347–1588. MR4127477 https://doi.org/10.1007/s00362-020-01199-9

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

SUN, H.-J. (1988). A FORTRAN subroutine for computing normal orthant probabilities of dimensions up to nine. *Comm. Statist. Simulation Comput.* **17** 1097–1111.

THODE, H. C. JR., FINCH, S. J. and MENDELL, N. R. (1988). Simulated percentage points for the null distribution of the likelihood ratio test for a mixture of two normals. *Biometrics* **44** 1195–1201. MR0981003 https://doi.org/10.2307/2531747

TITTERINGTON, D. M. (1990). Some recent research in the analysis of mixture distributions. *Statistics* **21** 619–641. MR1087291 https://doi.org/10.1080/02331889008802274

ULM, K. W. (1991). A statistical method for assessing a threshold in epidemiological studies. *Stat. Med.* **10** 341–349.

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

WALD, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Stat.* **20** 595–601. MR0032169 https://doi.org/10.1214/aoms/1177729952

WICHITCHAN, S., YAO, W. and YANG, G. (2019). Hypothesis testing for finite mixture models. *Comput. Statist. Data Anal.* **132** 180–189. MR3913143 https://doi.org/10.1016/j.csda.2018.05.005

WILKS, S. S. (1938). The large sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.* **1** 60–62.

WOLAK, F. A. (1987). An exact test for multiple inequality and equality constraints in the linear regression model. *J. Amer. Statist. Assoc.* **82** 782–793. MR0909983

WORSLEY, K. J. (1979). On the likelihood ratio test for a shift in location of normal populations. *J. Amer. Statist. Assoc.* **74** 365–367. MR0548027

YAO, Y.-C. and DAVIS, R. A. (1986). The asymptotic behavior of the likelihood ratio statistic for testing a shift in mean in a sequence of independent normal variates. *Sankhyā Ser. A* **48** 339–353. MR0905446

YAU, C. Y. and ZHAO, Z. (2016). Inference for multiple change points in time series via likelihood ratio scan statistics. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 895–916. MR3534355 https://doi.org/10.1111/rssb.12139

YU, M. and CHEN, X. (2022). A robust bootstrap change point test for high-dimensional location parameter. *Electron. J. Stat.* **16** 1096–1152. MR4377134 https://doi.org/10.1214/21-ejs1915

YU, Y. (2018). mixR: Finite mixture modeling for raw and binned data. R package version 0.1.1. Available at https://CRAN.R-project.org/package=mixR.

ZEILEIS, A. (2006). Implementing a class of structural change tests: An econometric computing approach. *Comput. Statist. Data Anal.* **50** 2987–3008. MR2239654 https://doi.org/10.1016/j.csda.2005.07.001

ZEILEIS, A., LEISCH, F., HORNIK, K. and KLEIBER, C. (2002). Strucchange: An R package for testing for structural change in linear regression models. *J. Stat. Softw.* **7** 1–38.

ZHANG, Y. (2018). lmeVarComp: Testing for a subset of variance components in linear mixed models. R package version 1.1. Available at https://CRAN.R-project.org/package=lmeVarComp.

ZHANG, Y., STAICU, A.-M. and MAITY, A. (2016). Testing for additivity in non-parametric regression. *Canad. J. Statist.* **44** 445–462. MR3574131 https://doi.org/10.1002/cjs.11295

Statistical Science
2024, Vol. 39, No. 2, 346–354
https://doi.org/10.1214/23-STS912

J. B. S. Haldane's Rule of Succession

Eric-Jan Wagenmakers, Sandy Zabell and Quentin F. Gronau

Abstract. After Bayes, the oldest Bayesian account of enumerative induction is given by Laplace's so-called *rule of succession*: if all n observed instances of a phenomenon to date exhibit a given character, the probability that the next instance of that phenomenon will also exhibit the character is $\frac{n+1}{n+2}$. Laplace's rule however has the apparently counterintuitive mathematical consequence that the corresponding "universal generalization" (every future observation of this type will also exhibit that character) has zero probability. In 1932, the British scientist J. B. S. Haldane proposed an alternative rule giving a universal generalization the positive probability $\frac{n+1}{n+2} \times \frac{n+3}{n+2}$. A year later, Harold Jeffreys proposed essentially the same rule in the case of a finite population. A related variant rule results in a predictive probability of $\frac{n+1}{n+2} \times \frac{n+4}{n+3}$. These arguably elegant adjustments of the original Laplacean form have the advantage that they give predictions better aligned with intuition and common sense. In this paper, we discuss J. B. S. Haldane's rule and its variants, placing them in their historical context, and relating them to subsequent philosophical discussions.

Key words and phrases: Universal generalization, statistical evidence.

REFERENCES

[1] BAYES, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philos. Trans. R. Soc. Lond.* **53** 370–418.
[2] BROAD, C. D. (1968). On the relation between induction and probability (part I.). *Mind* **27** 389–404.
[3] CARNAP, R. (1950). *Logical Foundations of Probability*. Univ. Chicago Press, Chicago, IL. MR0040253
[4] CARNAP, R. (1952). *The Continuum of Inductive Methods*. Univ. Chicago Press, Chicago, IL. MR0048379
[5] DE MORGAN, A. (1847). *Formal Logic*: *The Calculus of Inference, Necessary and Probable*. Taylor and Walton, London.
[6] DIACONIS, P. and FREEDMAN, D. (1980). de Finetti's theorem for Markov chains. *Ann. Probab.* **8** 115–130. MR0556418
[7] DIACONIS, P. and SKYRMS, B. (2018). *Ten Great Ideas About Chance*. Princeton Univ. Press, Princeton, NJ. MR3702017
[8] ETZ, A. and WAGENMAKERS, E.-J. (2017). J. B. S. Haldane's contribution to the Bayes factor hypothesis test. *Statist. Sci.* **32** 313–329. MR3648962 https://doi.org/10.1214/16-STS599
[9] HALDANE, J. B. S. (1932). A note on inverse probability. *Math. Proc. Cambridge Philos. Soc.* **28** 55–61.
[10] HINTIKKA, J. (1966). A two-dimensional continuum of inductive methods. In *Aspects of Inductive Logic* (J. Hintikka and P. Suppes, eds.). 113–132. North-Holland, Amsterdam. MR0201269
[11] HINTIKKA, J. and NIINILUOTO, I. (1976). An axiomatic foundation for the logic of inductive generalization. In *Formal Methods in the Methodology of Empirical Sciences* (*Proc. Conf., Warsaw*, 1974) (K. Szaniawski and R. Wójcicki, eds.) *Synthese Lib.* **103** 57–81. Reidel, Dordrecht. MR0538458
[12] JEFFREY, R. (1992). *Probability and the Art of Judgment. Cambridge Studies in Probability, Induction, and Decision Theory*. Cambridge Univ. Press, Cambridge. MR1160559 https://doi.org/10.1017/CBO9781139172394
[13] JEFFREYS, H. (1931). *Scientific Inference*, 1st ed. Cambridge Univ. Press, Cambridge, UK.
[14] JEFFREYS, H. (1933). On the prior probability in the theory of sampling. *Proc. Camb. Philos. Soc.* **29** 83–87.
[15] JEFFREYS, H. (1939). *Theory of Probability*. Oxford Univ. Press, Oxford. MR0000924
[16] JEFFREYS, H. (1948). *Theory of Probability*, 2nd ed. Oxford Univ. Press, Oxford.
[17] JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Clarendon, Oxford. MR0187257
[18] JEFFREYS, H. (1973). *Scientific Inference*, 3rd ed. Cambridge Univ. Press, Cambridge, UK.
[19] JOHNSON, W. E. (1932). Probability: The deductive and inductive problems. *Mind* **41** 409–423.
[20] KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795. MR3363402 https://doi.org/10.1080/01621459.1995.10476572
[21] KUIPERS, T. A. F. (1978). *Studies in Inductive Probability and Rational Expectation. Synthese Library* **123**. Reidel, Dordrecht-Boston, Mass. MR0490761
[22] LAPLACE, P.-S. (1986). Memoir on the probability of the causes of events. *Statist. Sci.* **1** 364–378.

Eric-Jan Wagenmakers is Professor of Bayesian Methodology at the Psychological Methods Unit, University of Amsterdam, Amsterdam, The Netherlands (e-mail: EJ.Wagenmakers@gmail.com). Sandy Zabell is Professor of Mathematics and Statistics and Data Science at the Departments of Mathematics, and Statistics and Data Science, Northwestern University, Evanston, Illinois, USA (e-mail: zabell@math.northwestern.edu). Quentin F. Gronau is a postdoctoral fellow at the School of Psychological Sciences, University of Newcastle, Newcastle, Australia (e-mail: quentin.f.gronau@gmail.com).

[23] LAPLACE, P.-S. (1995). *Philosophical Essay on Probabilities*. *Sources in the History of Mathematics and Physical Sciences* **13**. Springer, New York. Translated from the fifth (1825) French edition, and with notes and a preface by Andrew I. Dale. MR1325241 https://doi.org/10.1007/978-1-4612-4184-3

[24] NIINILUOTO, I. (1981). Analogy and Inductive Logic. *Erkenntnis* **16** 1–34.

[25] NIINILUOTO, I. (2011). The development of the Hintikka program. In *Inductive Logic* (D. M. Gabbay, S. Hartmann and J. Woods eds.). *Handbook of the History of Logic* **10** 311–356. North-Holland, Amsterdam.

[26] OLIVEIRA E SILVA, T., HERZOG, S. and PARDI, S. (2014). Empirical verification of the even Goldbach conjecture and computation of prime gaps up to $4 \cdot 10^{18}$. *Math. Comp.* **83** 2033–2060. MR3194140 https://doi.org/10.1090/S0025-5718-2013-02787-1

[27] PASEAU, A. C. (2021). Arithmetic, enumerative induction and size bias. *Synthese* **199** 9161–9184. MR4351389 https://doi.org/10.1007/s11229-021-03198-1

[28] TUYL, F. (2019). A method to handle zero counts in the multinomial model. *Amer. Statist.* **73** 151–158. MR3953627 https://doi.org/10.1080/00031305.2018.1444673

[29] WALZER, R. (1944). *Galen on Medical Experience*: *First Edition of the Arabic Version with English Translation and Notes*. Oxford Univ. Press, London.

[30] WRINCH, D. and JEFFREYS, H. (1921). On certain fundamental principles of scientific inquiry. *Philos. Mag.* **42** 369–390.

[31] ZABELL, S. L. (1982). W. E. Johnson's "sufficientness" postulate. *Ann. Statist.* **10** 1090–1099. MR0673645

[32] ZABELL, S. L. (1988). Symmetry and its discontents. In *Causation*, *Chance*, *and Credence*: *Volume I* (B. Skyrms and W. L. Harper, eds.). 155–190. Kluwer Academic, Dordrecht.

[33] ZABELL, S. L. (1989). The rule of succession. *Erkenntnis* **31** 283–321.

[34] ZABELL, S. L. (1992). Predicting the unpredictable. *Synthese* **90** 205–232. MR1148566 https://doi.org/10.1007/BF00485351

[35] ZABELL, S. L. (1996). Confirming universal generalizations. *Erkenntnis* **45** 267–283.

[36] ZABELL, S. L. (1997). The continuum of inductive methods revisited. In *The Cosmos of Science*: *Essays of Exploration* (J. Earman and J. D. Norton, eds.). 351–385. Univ. Pittsburgh Press, Pittsburgh, PA.

[37] ZABELL, S. L. (2011). Carnap and the logic of inductive inference. In *Inductive Logic* (D. M. Gabbay, S. Hartmann and J. Woods eds.). *Handbook of the History of Logic* **10** 265–309. North-Holland, Amsterdam.

Statistical Science
2024, Vol. 39, No. 2, 355–356
https://doi.org/10.1214/23-STS913

On the Certainty of an Inductive Inference: The Binomial Case

Frank Tuyl, Richard Gerlach and Kerrie Mengersen

Abstract. In the context of the binomial distribution, the potential need for a prior point mass on $\theta = 0\,(1)$, given $x = 0\,(n)$, was identified more than 100 years ago by Jeffreys. Given previous proposals to implement such a point mass, a slightly different approach is proposed, followed by the corresponding posterior probability of "homogeneity" and posterior predictive distribution.

Key words and phrases: Bayesian inference, Laplace's rule of succession, point mass.

REFERENCES

BERNARDO, J.-M. and SMITH, A. F. M. (1994). *Bayesian Theory*. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. Wiley, Chichester. MR1274699 https://doi.org/10.1002/9780470316870

BROAD, C. D. (1918). On the relation between induction and probability (Part I.). *Mind*. **27** 389–404.

HALDANE, J. B. S. (1932). A note on inverse probability. *Math. Proc. Cambridge Philos. Soc*. **28** 55–61. https://doi.org/10.1017/S0305004100010495

HUZURBAZAR, V. S. (1955). On the certainty of an inductive inference. *Math. Proc. Cambridge Philos. Soc*. **51** 761–762. MR0071388 https://doi.org/10.1017/s0305004100030826

JAYNES, E. T. (2003). *Probability Theory*: *The Logic of Science*. Cambridge Univ. Press, Cambridge. Edited and with a foreword by G. Larry Bretthorst. MR1992316 https://doi.org/10.1017/CBO9780511790423

JEFFREYS, H. (1939). *Theory of Probability*. Oxford Univ. Press, Oxford. MR0000924

JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Clarendon, Oxford. MR0187257

TUYL, F. (2019). A method to handle zero counts in the multinomial model. *Amer. Statist*. **73** 151–158. MR3953627 https://doi.org/10.1080/00031305.2018.1444673

WRINCH, D. H. and JEFFREYS, H. (1919). On some aspects of the theory of probability. *Philos. Mag. Ser.* 6 **38** 715–731.

Frank Tuyl is Honorary Lecturer, School of Information and Physical Sciences, The University of Newcastle, Newcastle, Australia (e-mail: frank.tuyl@newcastle.edu.au). Richard Gerlach is Professor, Business School, University of Sydney, Sydney, Australia (e-mail: richard.gerlach@sydney.edu.au). Kerrie Mengersen is Distinguished Professor, School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia (e-mail: k.mengersen@qut.edu.au).

Statistical Science
2024, Vol. 39, No. 2, 357–373
https://doi.org/10.1214/23-STS906

A Conversation with Guido W. Imbens

Fabrizia Mealli and Julie Holland Mortimer

Abstract. Guido Wilhelmus Imbens is the Applied Econometrics Professor and Professor of Economics with a joint appointment at the Graduate School of Business and the Department of Economics at Stanford University. He has made fundamental contributions to econometric and statistical methods for drawing causal inferences in experimental and observational studies, and applications to a wide range of disciplines beyond economics, including psychology, education, policy, law, epidemiology, public health and other social and biomedical sciences. Together with his longtime collaborator, Joshua Angrist, Guido was awarded half the 2021 Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel for their methodological contributions to the analysis of causal relationships, with the other half going to David Card.

REFERENCES

ABADIE, A., DIAMOND, A. and HAINMUELLER, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *J. Amer. Statist. Assoc.* **105** 493–505. MR2759929 https://doi.org/10.1198/jasa.2009.ap08746

ABADIE, A., DIAMOND, A. and HAINMUELLER, J. (2015). Comparative politics and the synthetic control method. *Amer. J. Polit. Sci.* **59** 495–510.

ABADIE, A. and IMBENS, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica* **74** 235–267. MR2194325 https://doi.org/10.1111/j.1468-0262.2006.00655.x

ANGRIST, J. and PISCHKE, J.-S. (2009). *Mostly Harmless Econometrics*: *An Empiricist's Companion*. Princeton Univ. Press, USA.

ANGRIST, J. D. (1990). Lifetime earnings and the Vietnam era draft lottery: Evidence from social security administrative records. *Amer. Econ. Rev.* **80** 313–336.

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

ANGRIST, J. D. and KRUEGER, A. B. (1999). Empirical strategies in labor economics. In *Handbook of Labor Economics* **3** 1277–1366. Elsevier, Amsterdam.

ANGRIST, J. D. and PISCHKE, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *J. Econ. Perspect.* **24** 3–30.

ATHEY, S., CHETTY, R., IMBENS, G. W. and KANG, H. (2019). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. *Working Paper* **3895**.

BAJARI, P., BURDICK, B., IMBENS, G. W., MASOERO, L., MCQUEEN, J., RICHARDSON, T. S. and ROSEN, I. M. (2023). Experimental design in marketplaces. *Statist. Sci.* **38** 458–476. MR4630378 https://doi.org/10.1214/23-sts883

BERTRAND, M. and MULLAINATHAN, S. (2004). Are emily and greg more employable than lakisha and jamal? A field experiment on labor market discrimination. *Amer. Econ. Rev.* **94** 991–1013.

BLACK, S. E. (1999). Do better schools matter? Parental valuation of elementary education. *Q. J. Econ.* **114** 577–599.

BLOOM, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Eval. Rev.* **8** 225–246.

DEHEJIA, R. H. and WAHBA, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *J. Amer. Statist. Assoc.* **94** 1053–1062.

DEKKER, E. (2021). *Jan Tinbergen* (1903–1994) *and the Rise of Economic Expertise*. Cambridge Univ. Press, Cambridge.

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

GELMAN, A. and IMBENS, G. (2013). Why ask why? Forward causal inference and reverse causal questions. *NBER Working Papers 19614, National Bureau of Economic Research, Inc.*

GOLDIN, C. and ROUSE, C. (2000). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *Amer. Econ. Rev.* **90** 715–741.

GRILICHES, Z. and MASON, W. M. (1972). Education, income, and ability. *J. Polit. Econ.* **80** S74–S103.

GUPTA, S., KOHAVI, R., TANG, D., XU, Y., ANDERSEN, R., BAKSHY, E., CARDIN, N., CHANDRAN, S., CHEN, N. et al. (2019). Top challenges from the first practical online controlled experiments summit. *ACM SIGKDD Explor. Newsl.* **21** 20–35.

HECKMAN, J. (1990). Varieties of selection bias. *Amer. Econ. Rev.* **80** 313–318.

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 https://doi.org/10.1111/1468-0262.00442

IMBENS, G. (2021). Statistical significance, p-values, and the reporting of uncertainty. *J. Econ. Perspect.* **35** 157–174. https://doi.org/10.1257/jep.35.3.157

Fabrizia Mealli is Professor of Econometrics, European University Institute, Florence, Italy, and Professor of Statistics, University of Florence, Florence, Italy (e-mail: fabrizia.mealli@eui.eu). Julie Holland Mortimer is the Kenneth G. Elzinga Professor of Economics and the Law, University of Virginia, Charlottesville, Virginia 22904-4182, USA (e-mail: juliemortimer@virginia.edu).

IMBENS, G. W. (1992). An efficient method of moments estimator for discrete choice models with choice-based sampling. *Econometrica* **60** 1187–1214. MR1180239 https://doi.org/10.2307/2951544

IMBENS, G. W. (1994). Transition models in a non-stationary environment. *Rev. Econ. Stat.* 703–720.

IMBENS, G. W. (2020). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *J. Econ. Lit.* **58** 1129–79.

IMBENS, G. W. and ANGRIST, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62** 467–775.

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

IMBENS, G. W., RUBIN, D. B. and SACERDOTE, B. I. (2001). Estimating the effect of unearned income on labor earnings, savings, and consumption: Evidence from a survey of lottery players. *Amer. Econ. Rev.* **91** 778–794.

KOLESÁR, M., CHETTY, R., FRIEDMAN, J., GLAESER, E. and IMBENS, G. W. (2015). Identification and inference with many invalid instruments. *J. Bus. Econom. Statist.* **33** 474–484. MR3416595 https://doi.org/10.1080/07350015.2014.978175

KRAMER, A. D., GUILLORY, J. E. and HANCOCK, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proc. Natl. Acad. Sci. USA* **111** 8788–8790.

LALONDE, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *Amer. Econ. Rev.* 604–620.

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

LEAMER, E. E. (1983). Let's take the con out of econometrics. *Amer. Econ. Rev.* **73** 31–43.

LEE, D. S., MORETTI, E. and BUTLER, M. J. (2004). Do voters affect or elect policies? Evidence from the US house. *Q. J. Econ.* **119** 807–859.

MANSKI, C. (1990). Non-parametric bounds on treatment effects. *Am. Econ. Rev. Pap. Proc.* **80** 319–323.

NEYMAN, J., IWASZKIEWICZ, K. and KOŁODZIEJCZYK, S. (1935). Statistical problems in agricultural experimentation. *Suppl. J. R. Stat. Soc.* **2** 107–154.

PAPADOGEORGOU, G., MEALLI, F. and ZIGLER, C. M. (2019). Causal inference with interfering units for cluster and population level treatment allocation programs. *Biometrics* **75** 778–787. MR4012083 https://doi.org/10.1111/biom.13049

PEARL, J. (2009). *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2548166 https://doi.org/10.1017/CBO9780511803161

PHILLIPS, A. W. (1958). The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1861–1957. *Economica* **25** 283–299.

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.

ROMER, C. D. and ROMER, D. H. (1994). Monetary policy matters. *J. Monet. Econ.* **34** 75–88.

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 importance of randomization. In *Proceedings of the Social Statistics Section of the American Statistical Association* 233-239.

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

THISTLEWAITE, D. and CAMPBELL, D. (1960). Regression-discontinuity analysis: An alternative to the ex-post facto experiment. *J. Educ. Psychol.* **51** 309–317.

TINBERGEN, J. (1940). Econometric business cycle research. *Rev. Econ. Stud.* **7** 73–90.

VAN DER KLAAUW, W. (2002). Estimating the effect of financial aid offers on college enrollment: A regression–discontinuity approach. *Internat. Econom. Rev.* **43** 1249–1287.

WRIGHT, S. (1934). The method of path coefficients. *Ann. Math. Stat.* **5** 161–215.

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: Michael Kosorok, Department of Biostatistics and Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, Chapel Hill, NC 27599, USA

President-Elect: Tony Cai, Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA 19104-6304, USA

Past President: Peter Bühlmann, Seminar für Statistik, ETH Zürich, 8092 Zürich, Switzerland

Executive Secretary: Peter Hoff, Department of Statistical Science, Duke University, Durham, NC 27708-0251, USA

Treasurer: Jiashun Jin, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA

Program Secretary: Annie Qu, Department of Statistics, University of California, Irvine, Irvine, CA 92697-3425, USA

IMS EDITORS

The Annals of Statistics. *Editors*: Enno Mammen, Institute for Mathematics, Heidelberg University, 69120 Heidelberg, Germany. Lan Wang, Miami Herbert Business School, University of Miami, Coral Gables, FL 33124, USA

The Annals of Applied Statistics. *Editor-in-Chief*: Ji Zhu, Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA

The Annals of Probability. *Editors*: Paul Bourgade, Courant Institute of Mathematical Sciences, New York University, New York, NY 10012-1185, USA. Julien Dubedat, Department of Mathematics, Columbia University, New York, NY 10027, USA

The Annals of Applied Probability. *Editors*: Kavita Ramanan, Division of Applied Mathematics, Brown University, Providence, RI 02912, USA. Qi-Man Shao, Department of Statistics and Data Science, Southern University of Science and Technology, Shenzhen, Guangdong 518055, P.R. China

Statistical Science. *Editor*: Moulinath Banerjee, Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA

The IMS Bulletin. *Editor*: Tati Howell, bulletin@imstat.org