Optimal test statistics for anytime-valid hypothesis tests

Lardy, T.D.

Bibliography

Adams, R. (2020). Safe hypothesis tests for the 2× 2 contingency table. Master's thesis, Delft University of Technology.

Agrawal, A. (2018). Lecture notes on Loewner order. `https://www.akshayagrawal.com/lecture-notes/html/loewner-order.html`.

Anderson, T. W. and Darling, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769.

Andersson, S. (1982). Distributions of maximal invariants using quotient measures. *The Annals of Statistics*, 10(3):955–961.

Anscombe, F. J. (1954). Fixed-sample-size analysis of sequential observations. *Biometrics*, 10(1):89–100.

Awad, Y., Bar-Lev, S. K., and Makov, U. (2022). A new class of counting distributions embedded in the Lee–Carter model for mortality projections: A Bayesian approach. *Risks*, 10(6):111.

Azadkia, M. and Chatterjee, S. (2021). A Simple Measure of Conditional Dependence. *The Annals of Statistics*, 49(6):3070 – 3102.

Balsubramani, A. and Ramdas, A. (2016). Sequential nonparametric testing with the law of the iterated logarithm. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, Uncertainty in Artificial Intelligence 2016, pages 42—-51, Arlington, Virginia, USA. Uncertainty in Artificial Intelligence Press.

Bar-Lev, S. K. (2020). Independent, though identical results: the class of Tweedie on power variance functions and the class of Bar-Lev and Enis on reproducible natural exponential families. *International Journal of Statistics and Probability*, 9(1):30–35.

Bar-Lev, S. K., Letac, G., and Ridder, A. (2024). A delineation of new classes of exponential dispersion models supported on the set of nonnegative integers. *Annals of the Institute of Statistical Mathematics*, 76(4):679–709.

Bar-Lev, S. K. and Ridder, A. (2021). New exponential dispersion models for count data – the ABM and LM classes. *ESAIM: Probability and Statistics*, 25:31–52.

Bar-Lev, S. K. and Ridder, A. (2023). Exponential dispersion models for overdispersed zero-inflated count data. *Communications in Statistics-Simulation and Computation*, 52(7):3286–3304.

Baringhaus, L. (1991). Testing for spherical symmetry of a multivariate distribution. *The Annals of Statistics*, pages 899–917.

Barnard, G. A. (1946). Sequential tests in industrial statistics. *Supplement to the Journal of the Royal Statistical Society*, 8(1):1–21.

Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory.* Wiley, Chichester, UK.

Barnett, A. (2018). It's a sign of how bad things have got that researchers think it's acceptable to write this in a nature journal: "we continuously increased the number of animals until statistical significance was reached to support our conclusions.". `https://x.com/aidybarnett/status/1036392482139865088`. X user @aidybarnett, id: 1036392482139865088, Accessed: 23-10-2024.

Barron, A. (1998). Information-Theoretic Characterization of Bayes Performance and the Choice of Priors in Parametric and Nonparametric Problems. In *Bayesian Statistics*, volume 6, pages 27–52. Oxford University Press, Oxford.

Berger, J. O., Pericchi, L. R., and Varshavsky, J. A. (1998). Bayes factors and marginal distributions in invariant situations. *Sankhyā: The Indian Journal of Statistics, Series A*, 60:307–321.

Berger, J. O. and Sun, D. (2008). Objective priors for the bivariate normal model. *The Annals of Statistics*, 36(2):963–982.

Berk, R. H. (1972). A note on sufficiency and invariance. *The Annals of Mathematical Statistics*, 43(2):647–650.

Berman, S. M. (1965). Sign-invariant random variables and stochastic processes with sign-invariant increments. *Transactions of the American Mathematical Society*, 119(2):216–243.

Bernardo, J. M. and Smith, A. F. (2009). *Bayesian theory*, volume 405. John Wiley & Sons.

Berrett, T. B., Wang, Y., Barber, R. F., and Samworth, R. J. (2020). The Conditional Permutation Test for Independence While Controlling for Confounders. *Journal of the Royal Statistical Society: Series B*, 82(1):175–197.

Bhowmik, J. L. and King, M. L. (2007). Maximal invariant likelihood based testing of semi-linear models. *Statistical Papers*, 48(3):357–383.

Bilodeau, B., Foster, D. J., and Roy, D. M. (2023). Minimax rates for conditional density estimation via empirical entropy. *The Annals of Statistics*, 51(2):762–790.

Bondar, J. V. (1976). Borel cross-sections and maximal invariants. *The Annals of Statistics*, 4(5):866–877.

Bondar, J. V. and Milnes, P. (1981). Amenability: A survey for statistical applications of Hunt-Stein and related conditions on groups. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57(1):103–128.

Bourbaki, N. (2004). *Integration II: Chapters 7–9*. Elements of Mathematics. Springer-Verlag, Berlin Heidelberg, 1st edition.

Brinda, W. D. (2018). *Adaptive estimation with Gaussian radial basis mixtures*. PhD thesis, Yale University.

Brown, L. D. (1966). On the admissibility of invariant estimators of one or more location parameters. *The Annals of Mathematical Statistics*, 37(5):1087–1136.

Brown, L. D. (1986). Fundamentals of statistical exponential families with applications in statistical decision theory. *Lecture Notes-Monograph Series*, 9:i–279.

Candès, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for Gold: 'Model-X' Knockoffs for High Dimensional Controlled Variable Selection. *Journal of the Royal Statistical Society: Series B*, 80(3):551–577.

Carney, D. R. (2016). My position on "power poses". `https://faculty.haas.berkeley.edu/dana_carney/pdf_my%20position%20on%20power%20poses.pdf`. Accessed: 23-10-2024.

Carney, D. R., Cuddy, A. J., and Yap, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological Science*, 21(10):1363–1368.

Casper, C., Cook, T., and on FORTRAN program ld98., O. A. P. B. (2022). *ldbounds: Lan-DeMets Method for Group Sequential Boundaries*. R package version 2.0.0.

Cesa-Bianchi, N. and Lugosi, G. (2001). Worst-case bounds for the logarithmic loss of predictors. *Machine Learning*, 43:247–264.

Chang, J. T. and Pollard, D. (1997). Conditioning as disintegration. *Statistica Neerlandica*, 51(3):287–317.

Chen, P., Chen, Y., and Rao, M. (2008). Metrics defined by Bregman divergences. *Communications in Mathematical Sciences*, 6(4):915–926.

Chiu, K. and Bloem-Reddy, B. (2023). Non-parametric hypothesis tests for distributional group symmetry. ArXiv preprint arXiv:2307.15834.

Cohen, T. and Welling, M. (2016). Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley Series in Telecommunications. John Wiley & Sons, Inc., New York.

Cox, D. R. (1952). Sequential tests for composite hypotheses. *Mathematical Proceedings of the Cambridge Philosophical Society*, 48(2):290–299.

Csiszár, I. (1963). Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der ergodizität von Markoffschen Ketten. *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei*, 8(1-2):85–108.

Csiszár, I. and Matúš, F. (2003). Information projections revisited. *IEEE Transactions on Information Theory*, 49(6):1474–1490.

Csiszár, I. and Tusnády, G. (1984). Information geometry and alternating minimization procedures. *Statistics and Decisions*, Supplementary Issue 1:205–237.

Cukier, K. and Mayer-Schoenberger, V. (2013). The rise of big data: How it's changing the way we think about the world. *Foreign Affairs*, 92(3):20–32.

Darling, D. and Robbins, H. (1967). Confidence Sequences for Mean, Variance, and Median. *Proceedings of the National Academy of Sciences*, 58(1):66–68.

Darling, D. A. and Robbins, H. (1968). Some nonparametric sequential tests with power one. *Proceedings of the National Academy of Sciences*, 61(3):804–809.

Dawid, A. P. (1979). Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society: Series B*, 41(1):1–15.

Dawid, A. P., Stone, M., and Zidek, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference. *Journal of the Royal Statistical Society, Series B (Methodological)*, 35(2):189–233.

Diaconis, P. (1988). Group representations in probability and statistics. *Lecture notes-monograph series*, 11:i–192.

Dodge, H. F. and Romig, H. G. (1929). A method of sampling inspection. *The Bell System Technical Journal*, 8(4):613–631.

Duan, B., Ramdas, A., and Wasserman, L. (2022). Interactive rank testing by betting. In *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 201–235.

Durrett, R. (2019). *Probability: Theory and examples.* Number 49 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 5th edition.

Eaton, M. L. (1989). Group invariance applications in statistics. *Regional Conference Series in Probability and Statistics*, 1:i–133.

Eaton, M. L. and Sudderth, W. D. (1999). Consistency and strong inconsistency of group-invariant predictive inferences. *Bernoulli*, 5(5):833–854. Publisher: International Statistical Institute (ISI) and Bernoulli Society for Mathematical Statistics and Probability.

Eaton, M. L. and Sudderth, W. D. (2002). Group invariant inference and right Haar measure. *Journal of Statistical Planning and Inference*, 103(1):87–99.

Efron, B. (2022). *Exponential Families in Theory and Practice.* Institute of Mathematical Statistics Textbooks. Cambridge University Press.

Fedorova, V., Gammerman, A., Nouretdinov, I., and Vovk, V. (2012). Plug-in martingales for testing exchangeability on-line. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1639–1646, New York, NY, USA. Omnipress.

Feller, W. K. (1940). Statistical aspects of esp. *The Journal of Parapsychology*, 4(2):271.

Fisher, R. A. (1936). Design of experiments. *British Medical Journal*, 1(3923):554.

Fontana, M., Zeni, G., and Vantini, S. (2023). Conformal prediction: a unified review of theory and new challenges. *Bernoulli*, 29(1):1–23.

Foster, D. J., Kale, S., Luo, H., Mohri, M., and Sridharan, K. (2018). Logistic regression: The Importance of Being Improper. In *Conference on learning theory*, pages 167–208.

Fraiman, R., Moreno, L., and Ransford, T. (2024). Application of the cramér–wold theorem to testing for invariance under group actions. *TEST*, 33(2):379–399.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2007). Kernel Measures of Conditional Dependence. In *Advances in Neural Information Processing Systems*, volume 20.

Giri, N., Kiefer, J., and Stein, C. (1963). Minimax character of Hotelling's T^2 test in the simplest case. *The Annals of Mathematical Statistics*, 34(4):1524 – 1535.

Greenwood, J. A. (1938). An empirical investigation of some sampling problems. *The Journal of Parapsychology*, 2(3):222.

Greenwood, J. A. and Greville, T. (1939). On the probability of attaining a given standard deviation ratio in an infinite series of trials. *The Annals of Mathematical Statistics*, 10(3):297–298.

Grünwald, P. (2007). *The minimum description length principle.* MIT press.

Grünwald, P. (2023). The E-posterior. *Philosophical Transactions of the Royal Society of London, Series A.*

Grünwald, P., de Heide, R., and Koolen, W. (2024). Safe testing. *Journal of the Royal Statistical Society, Series B.*

Grünwald, P. and Harremoës, P. (2009). Finiteness of redundancy, regret, Shtarkov sums, and Jeffreys integrals in exponential families. In *Proceedings for the International Symposium for Information Theory, Seoul, 2009*, pages 714–718. IEEE.

Grünwald, P., Henzi, A., and Lardy, T. (2023). Anytime valid tests of conditional independence under model-X. *Journal of the American Statistical Association*, 119(546):1554–1565.

Grünwald, P. D. (2024). Beyond neyman–pearson: E-values enable hypothesis testing with a data-driven alpha. *Proceedings of the National Academy of Sciences*, 121(39).

Grünwald, P. D. and Mehta, N. A. (2019). A tight excess risk bound via a unified pac-bayesian–rademacher–shtarkov–mdl complexity. In *Algorithmic Learning Theory*, pages 433–465. PMLR.

Grünwald, P. D. and Mehta, N. A. (2020). Fast rates for general unbounded loss functions: from ERM to generalized Bayes. *The Journal of Machine Learning Research*, 21(1):2040–2119.

Grünwald, P., Lardy, T., Hao, Y., Bar-Lev, S. K., and De Jong, M. (2024). Optimal e-values for exponential families: the simple case. ArXiv preprint, arXiv:2404.19465. A revised version of this manuscript has been accepted as a contribution to the Springer festschrift "Information Theory, Probability and Statistical Learning: A Festschrift in Honor of Andrew Barron.".

Hall, W. J., Wijsman, R. A., and Ghosh, J. K. (1965). The relationship between sufficiency and invariance with applications in sequential analysis. *The Annals of Mathematical Statistics*, 36(2):575–614.

Hall, W. J., Wijsman, R. A., and Ghosh, J. K. (1995). Correction: The relationship between sufficiency and invariance with applications in sequential analysis. *The Annals of Statistics*, 23(2):705–705.

Ham, D. W., Imai, K., and Janson, L. (2024). Using machine learning to test causal hypotheses in conjoint analysis. *Political Analysis*, 32(3):329–344.

Hao, Y., Grünwald, P., Lardy, T., Long, L., and Adams, R. (2024). E-values for k-sample tests with exponential families. *Sankhya A*, 86(1):596–636.

Haussler, D. and Opper, M. (1997). Mutual information, metric entropy and cumulative relative entropy risk. *The Annals of Statistics*, 25(6):2451–2492.

Henzi, A. and Law, M. (2024). A rank-based sequential test of independence. *Biometrika*.

Henzi, A., Puke, M., Dimitriadis, T., and Ziegel, J. (2023). A safe hosmer-lemeshow test. *The New England Journal of Statistics in Data Science*, 2(2):175–189.

Henzi, A. and Ziegel, J. F. (2022). Valid sequential inference on probability forecast performance. *Biometrika*, 109(3):647–663.

Howard, S. R., Ramdas, A., Mcauliffe, J., and Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080.

Jeffreys, H. (1961). *Theory of probability.* Oxford University Press, London, 3rd edition.

John, L. K., Loewenstein, G., and Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5):524–532.

Jørgensen, B. (1997). *The Theory of Exponential Dispersion Models*, volume 76 of *Monographs on Statistics and Probability.* Chapman and Hall, London.

Kariya, T. (1980). Locally robust tests for serial correlation in least squares regression. *The Annals of Statistics*, 8(5):1065–1070.

Katsevich, E. and Ramdas, A. (2022). On the Power of Conditional Independence Testing Under Model-X. *Electronic Journal of Statistics*, 16(2):6348 – 6394.

Kelly, J. L. (1956). A new interpretation of information rate. *The Bell System Technical Journal*, 35(4):917–926.

Koning, N. W. (2023). Online permutation tests: e-values and likelihood ratios for testing group invariance. ArXiv preprint arXiv:2310.01153.

Koolen, W. M. and Grünwald, P. (2022). Log-optimal anytime-valid e-values. *International Journal of Approximate Reasoning*, 141:69–82.

Kotłowski, W. and Grünwald, P. (2011). Maximum likelihood vs. sequential normalized maximum likelihood in on-line density estimation. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 457–476. JMLR Workshop and Conference Proceedings.

Kullback, S. and Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86.

Kumar, M. A. and Sason, I. (2016). Projection theorems for the Rényi divergence on α-convex sets. *IEEE Transactions on Information Theory*, 62(9):4924–4935.

Lai, T. L. (1976). On confidence sequences. *The Annals of Statistics*, 4(2):265–280.

Lai, T. L. (1977). Power-one tests based on sample sums. *The Annals of Statistics*, 5(5):866–880.

Lardy, T. (2021). E-values for hypothesis testing with covariates. Master's Thesis, Leiden University.

Lardy, T., Grünwald, P., and Harremoës, P. (2024). Reverse information projections and optimal e-statistics. *IEEE Transactions on Information Theory*, 70(11):7616–7631.

Lardy, T. and Pérez-Ortiz, M. F. (2024). Anytime-valid tests of group invariance through conformal prediction. ArXiv preprint arXiv:2401.15461.

Larsson, M., Ramdas, A., and Ruf, J. (2024). The numeraire e-variable and reverse information projection. ArXiv preprint arXiv:2402.18810.

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses.* Springer Texts in Statistics. Springer-Verlag, New York, 3rd edition.

Lehmann, E. L. and Stein, C. (1949). On the theory of some non-parametric hypotheses. *The Annals of Mathematical Statistics*, 20(1):28–45.

Levin, L. A. (1976). Uniform tests of randomness. In *Doklady Akademii Nauk*, volume 227, pages 33–35. Russian Academy of Sciences.

Lhéritier, A. and Cazals, F. (2018). A sequential non-parametric multivariate two-sample test. *IEEE Transactions on Information Theory*, 64(5):3361–3370.

Li, J. (1999). *Estimation of mixture models.* PhD thesis, Yale University, New Haven, CT.

Li, J. and Barron, A. (1999). Mixture density estimation. *Advances in neural information processing systems*, 12.

Li, S. (2010). Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics & Statistics*, 4(1):66–70.

Li, S. and Liu, M. (2023). Maxway crt: improving the robustness of the model-x inference. *Journal of the Royal Statistical Society, Series B*, 85(5):1441–1470.

Liang, F. and Barron, A. (2004). Exact minimax strategies for predictive density estimation, data compression, and model selection. *IEEE Transactions on Information Theory*, 50(11):2708–2726.

Liang, H. (2023). Stratified safe sequential testing for mean effect. Master's Thesis, University of Amsterdam.

Liese, F. and Vajda, I. (1987). *Convex Statistical Distances.* Teubner, Leipzig.

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318):399–402.

Lindon, M., Ham, D. W., Tingley, M., and Bojinov, I. (2024). Anytime-valid linear models and regression adjusted causal inference in randomized experiments. ArXiv preprint arXiv:2210.08589.

Liu, M., Katsevich, E., Janson, L., and Ramdas, A. (2021). Fast and Powerful Conditional Randomization Testing via Distillation. *Biometrika*, 109(2):277–293.

Malov, S. (1996). Sequential ranks and order statistics. *Journal of Mathematical Sciences*, 81:2434–2441.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models (2nd ed.).* CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall.

Meckes, E. S. (2019). *The random matrix theory of the classical compact groups*, volume 218. Cambridge University Press.

Morris, C. N. (1982). Natural exponential families with quadratic variance functions. *The Annals of Statistics*, 10(1):65–80.

Niu, Z., Chakraborty, A., Dukes, O., and Katsevich, E. (2024). Reconciling model-x and doubly robust approaches to conditional independence testing. *The Annals of Statistics*, 52(3):895–921.

Nogales, A. G. and Oyola, J. A. (1996). Some remarks on sufficiency, invariance and conditional independence. *The Annals of Statistics*, 24(2):906–909.

Novikov, A. (2024). Group sequential hypothesis tests with variable group sizes: Optimal design and performance evaluation. *Communications in Statistics-Theory and Methods*, 53(16):5744–5760.

O'Brien, P. C. and Fleming, T. R. (1979). A Multiple Testing Procedure for Clinical Trials. *Biometrics*, 35(3):549–556.

Pandeva, T., Bakker, T., Naesseth, C. A., and Forré, P. (2022). E-valuating classifier two-sample tests. *ArXiv preprint arXiv:2210.13027*.

Paterson, A. L. (1988). *Amenability.* Number 29 in Mathematical surveys and monographs. American Mathematical Soc., 1st edition.

Pérez-Ortiz, M. F., Lardy, T., De Heide, R., and Grünwald, P. D. (2024). E-statistics, group invariance and anytime-valid testing. *The Annals of Statistics*, 52(4):1410–1432.

Pitman, E. J. (1937). Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, 4(1):119–130.

Pocock, S. J. (1977). Group Sequential Methods in the Design and Analysis of Clinical Trials. *Biometrika*, 64(2):191–199.

Pocock, S. J. and Simon, R. (1975). Sequential Treatment Assignment with Balancing for Prognostic Factors in the Controlled Clinical Trial. *Biometrics*, 31(1):103–115.

Qian, G. and Field, C. (2002). Law of Iterated Logarithm and Consistent Model Selection Criterion in Logistic Regression. *Statistics & Probability Letters*, 56(1):101–112.

R Core Team (2022). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. (2023). Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601.

Ramdas, A., Ruf, J., Larsson, M., and Koolen, W. (2020). Admissible anytime-valid sequential inference must rely on nonnegative martingales. ArXiv preprint arXiv:2009.03167.

Ramdas, A., Ruf, J., Larsson, M., and Koolen, W. M. (2022). Testing Exchangeability: Fork-Convexity, Supermartingales and E-Processes. *International Journal of Approximate Reasoning*, 141:83–109.

Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S., and Weber, R. A. (2015). Assessing the robustness of power posing: No effect on hormones and risk tolerance in a large sample of men and women. *Psychological Science*, 26(5):653–656.

Reiter, H. and Stegeman, J. D. (2000). *Classical harmonic analysis and locally compact groups*. London Mathematical Society Monographs. Oxford University Press, Oxford, New York, 2nd edition.

Ren, Z. and Barber, R. F. (2024). Derandomised knockoffs: leveraging e-values for false discovery rate control. *Journal of the Royal Statistical Society, Series B*, 86(1):122–154.

Rényi, A. (1962). On the extreme elements of observations. *MTA III. Oszt. Közl*, 12:105–121.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58:527–535.

Robbins, H. (1970). Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409.

Robbins, H. and Siegmund, D. (1974). The expected sample size of some tests of power one. *The Annals of Statistics*, 2(3):415–436.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., and Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2):225–237.

Roy, S. N. and Bargmann, R. E. (1958). Tests of multiple independence and the associated confidence bounds. *The Annals of Mathematical Statistics*, 29(2):491–503.

Runge, J. (2018). Conditional Independence Testing Based on a Nearest-Neighbor Estimator of Conditional Mutual Information. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 938–947.

Rushton, S. (1950). On a sequential t-test. *Biometrika*, 37(3–4):326–333.

Sagiroglu, S. and Sinanc, D. (2013). Big data: A review. In *2013 international conference on collaboration technologies and systems (CTS)*, pages 42–47. IEEE.

Saha, A. and Ramdas, A. (2024). Testing exchangeability by pairwise betting. In *International Conference on Artificial Intelligence and Statistics*, pages 4915–4923. PMLR.

Savage, I. R. (1956). Contributions to the Theory of Rank Order Statistics-the Two-Sample Case. *The Annals of Mathematical Statistics*, 27(3):590–615. Publisher: Institute of Mathematical Statistics.

Schmitz, N. (1993). *Optimal sequentially planned decision procedures*, volume 79 of *Lecture Notes in Statistics.* Springer-Verlag New York.

Sen, P. K. and Ghosh, M. (1973a). A chernoff-savage representation of rank order statistics for stationary φ-mixing processes. *Sankhyā: The Indian Journal of Statistics, Series A*, 35(2):153–172. Publisher: Springer.

Sen, P. K. and Ghosh, M. (1973b). A Law of Iterated Logarithm for One-Sample Rank Order Statistics and an Application. *The Annals of Statistics*, 1(3):568–576. Publisher: Institute of Mathematical Statistics.

Sen, P. K. and Ghosh, M. (1974). Sequential Rank Tests for Location. *The Annals of Statistics*, 2(3):540–552. Publisher: Institute of Mathematical Statistics.

Shaer, S., Maman, G., and Romano, Y. (2023). Model-x sequential testing for conditional independence via testing by betting. In *International Conference on Artificial Intelligence and Statistics*, pages 2054–2086. PMLR.

Shafer, G. (2021). Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A*, 184(2):407–431.

Shah, R. D. and Peters, J. (2020). The Hardness of Conditional Independence Testing and the Generalized Covariance Measure. *The Annals of Statistics*, 48(3):1514–1538.

Shalaevskii, O. V. (1971). Minimax character of Hotelling's T2 test. I. In Kalinin, V. M. and Shalaevskii, O. V., editors, *Investigations in Classical Problems of Probability Theory and Mathematical Statistics: Part I*, pages 74–101. Springer US, Boston, MA, 1st edition.

Shiryaev, A. N. (2016). *Probability-1*, volume 95. Springer.

Sidak, Z., Sen, P. K., and Hajek, J. (1999). *Theory of Rank Tests.* Elsevier.

Simmons, J. P. and Simonsohn, U. (2017). Power posing: P-curving the evidence. *Psychological Science*, 28(5):687–693.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software*, 39(5):1–13.

Smith, A. F. (1981). On random sequences with centred spherical symmetry. *Journal of the Royal Statistical Society: Series B (Methodological)*, 43(2):208–209.

Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American statistical Association*, 69(347):730–737.

Subbaiah, P. and Mudholkar, G. S. (1978). A comparison of two tests for the significance of a mean vector. *Journal of the American Statistical Association*, 73(362):414–418.

Sun, D. and Berger, J. O. (2007). Objective Bayesian analysis for the multivariate normal model. *Bayesian Statistics*, 8:525–562.

Ter Schure, J. and Grünwald, P. (2019). Accumulation bias in meta-analysis: the need to consider time in error control. *F1000Research*, 8.

Ter Schure, J. and Grünwald, P. (2022). All-in meta-analysis: breathing life into living systematic reviews. *F1000Research*, 11.

ter Schure, J., Pérez-Ortiz, M. F., Ly, A., and Grünwald, P. D. (2024). The anytime-valid logrank test: Error control under continuous monitoring with unlimited horizon. *The New England Journal of Statistics in Data Science*, 2(2):190–214.

Topsøe, F. (2007). Information theory at the service of science. In Csiszár, I., Katona, G. O. H., and Tardos, G., editors, *Entropy, Search, Complexity*, volume 16 of *Bolyai Society Mathematical Studies*, pages 179–207. János Bolyai Mathematical Society and Springer-Verlag.

Turner, R. and Grünwald, P. (2022). Safe sequential testing and effect estimation in stratified count data. In *Proceedings of the Twenty-Sixth International Conference on Artificial Intelligence and Statistics (AISTATS) 2023*, volume 206 of *Proceedings of Machine Learning Research*.

Turner, R., Ly, A., and Grünwald, P. (2024). Generic e-variables for exact sequential k-sample tests that allow for optional stopping. *Statistical Planning and Inference*, 230.

Turner, R. J. and Grünwald, P. D. (2023). Exact anytime-valid confidence intervals for contingency tables and beyond. *Statistics & Probability Letters*, 198:109835.

van Erven, T., Grünwald, P. D., Mehta, N. A., Reid, M. D., and Williamson, R. C. (2015). Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16:1793–1861.

van Erven, T. and Harremoës, P. (2014). Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820.

De Jong, M. (2021). Tests of significance for linear regression using E-values. Master's Thesis, Leiden University.

Ville, J. (1939). *Etude critique de la notion de collectif.* Gauthier-Villars Paris.

Vovk, V. (2002). On-line confidence machines are well-calibrated. In *The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings.*, pages 187–196. IEEE.

Vovk, V. (2004). A universal well-calibrated algorithm for on-line classification. *The Journal of Machine Learning Research*, 5:575–604.

Vovk, V. (2023). The power of forgetting in statistical hypothesis testing. In *Conformal and Probabilistic Prediction with Applications*, pages 347–366. PMLR.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*, volume 29. Springer.

Vovk, V., Nouretdinov, I., and Gammerman, A. (2003). Testing exchangeability on-line. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 768–775.

Vovk, V. and Wang, R. (2021). E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754.

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186.

Wald, A. (1947). *Sequential analysis.* John Wiley & Sons, Inc.

Wallis, W. A. (1980). The statistical research group, 1942–1945. *Journal of the American Statistical Association*, 75(370):320–330.

Wang, Q., Wang, R., and Ziegel, J. (2024). E-backtesting. ArXiv preprint arXiv:200209.00991.

Wang, R. and Ramdas, A. (2022). False discovery rate control with e-values. *Journal of the Royal Statistical Society, Series B*, 84:822–852.

Wasserman, L., Ramdas, A., and Balakrishnan, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890.

Waudby-Smith, I. and Ramdas, A. (2024). Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society, Series B*, 86(1):1–27.

Weber, F., Hoang Do, J. P., Chung, S., Beier, K. T., Bikov, M., Saffari Doost, M., and Dan, Y. (2018). Regulation of rem and non-rem sleep by periaqueductal gabaergic neurons. *Nature Communications*, 9(1):354.

Wennerholm, U.-B., Saltvedt, S., Wessberg, A., Alkmark, M., Bergh, C., Wendel, S. B., Fadl, H., Jonsson, M., Ladfors, L., Sengpiel, V., et al. (2019). Induction of labour at 41 weeks versus expectant management and induction of labour at 42 weeks (SWEdish Post-term Induction Study, swepis): multicentre, open label, randomised, superiority trial. *British Medical Journal*, 367.

Williams, D. (1991). *Probability with martingales.* Cambridge university press.

Wong, W. H. and Shen, X. S. (1995). Probability Inequalities for Likelihood Ratios and Convergence Rates of Sieve MLES. *The Annals of Statistics*, 23(2):339—362.

Young, W. H. (1912). On classes of summable functions and their Fourier series. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 87(594):225–229.

Zhang, L.-X., Hu, F., Cheung, S. H., and Chan, W. S. (2007). Asymptotic Properties of Covariate-Adjusted Response-Adaptive Designs. *The Annals of Statistics*, 35(3):1166–1182.

Zhang, Z., Ramdas, A., and Wang, R. (2024). On the existence of powerful p-values and e-values for composite hypotheses. *The Annals of Statistics*, 52(5):2241–2267.

Appendices

A | Appendix to Chapter 3

A.1 Proofs

A.1.1 Proofs for Section 3.3

Before giving the intended results, we note that we introduced m_P as the averaged Bregman divergence associated with $\gamma(x) = x - 1 - \ln(x)$. For the proof, it will be useful to also define the Bregman divergence associated with $\gamma(x) = x - 1 - \ln(x)$ itself, which is the so-called Itakura-Saito divergence. For $f, g \in \mathcal{M}(\Omega, \mathbb{R}_{>0})$, it is given by

$$IS_P(f,g) = \int_\Omega \left(\frac{f}{g} - 1 - \ln \frac{f}{g} \right) \mathrm{d}P.$$

By definition, it holds that

$$m_P^2(f,g) = \frac{1}{2} IS\left(f, \frac{f+g}{2}\right) + \frac{1}{2} IS\left(g, \frac{f+g}{2}\right).$$

Furthermore, for $Q \in \mathcal{C}$, we have $IS_P(q,p) = D(P\|Q)$. We now state some auxiliary results before giving the proofs for Section 3.3.

Lemma A.1. *For $x, y \in \mathbb{R}_{>0}$, we have*

$$|\ln(x) - \ln(y)| = g(m_\gamma^2(x,y)),$$

where g denotes the function

$$g(t) = 2t + 2\ln\left(1 + (1 - \exp(-2t))^{1/2}\right).$$

The function g is concave and satisfies $g(t) \geq 2t$.

Proof. Let $m = \frac{x+y}{2}$. Our goal is to determine the function g function such that

$$|\ln(x) - \ln(y)| = g(m_\gamma^2(x,y)).$$

We first rewrite the right-hand side

$$\begin{aligned} g(m_\gamma^2(x,y)) &= g\left(\ln(m) - \frac{1}{2}\ln(x) - \frac{1}{2}\ln(y)\right) \\ &= g\left(\frac{1}{2}\ln\left(\frac{m^2}{x \cdot y}\right)\right) \\ &= g\left(\frac{1}{2}\ln\left(\frac{\left(\frac{m}{y}\right)^2}{\frac{x}{y}}\right)\right) \\ &= g\left(\frac{1}{2}\ln\left(\frac{\left(\frac{1+\frac{x}{y}}{2}\right)^2}{\frac{x}{y}}\right)\right). \end{aligned}$$

Plugging this back in and replacing $\frac{x}{y}$ by w leads to

$$|\ln(w)| = g\left(\frac{1}{2}\ln\left(\frac{\left(\frac{1+w}{2}\right)^2}{w}\right)\right)$$

Then we solve the equation

$$\frac{1}{2}\ln\left(\frac{\left(\frac{1+w}{2}\right)^2}{w}\right) = t,$$

which gives

$$w = 2\exp(2t) - 1 + 2 \cdot (\exp(4t) - \exp(2t))^{1/2}$$

$$\begin{aligned} g(t) &= \ln\left(2\exp(2t) - 1 + 2 \cdot (\exp(4t) - \exp(2t))^{1/2}\right) \\ &= 2t + \ln\left(2 - \exp(-2t) + 2 \cdot (1 - \exp(-2t))^{1/2}\right) \\ &= 2t + 2\ln\left(1 + (1 - \exp(-2t))^{1/2}\right). \end{aligned}$$

The derivatives of g are

$$\begin{aligned}
g'(t) &= 2 + 2\frac{\left(1-\exp\left(-2t\right)\right)^{-1/2}\exp\left(-2t\right)}{1+\left(1-\exp\left(-2t\right)\right)^{1/2}} \\
&= \frac{2}{(1-\exp(-2t))^{1/2}} \\
g''(t) &= \frac{-\exp\left(-t/2\right)}{2^{1/2}\left(\sinh t\right)^{3/2}}.
\end{aligned}$$

We see that $g''(t) < 0$ and conclude that g is concave. Finally, we have

$$g\left(t\right) = 2t + 2\ln\left(1+\left(1-\exp\left(-2t\right)\right)^{1/2}\right) \geq 2t,$$

because $1-\exp\left(-2t\right) \geq 0$. □

Lemma A.2. *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of elements of $\mathcal{M}(\Omega, \mathbb{R}_{>0})$, then*

$$\limsup_{m,n\to\infty} m_P(f_m, f_n) = 0 \Leftrightarrow \limsup_{m,n\to\infty} \int_\Omega \left|\ln\left(\frac{f_m}{f_n}\right)\right| \mathrm{d}P = 0.$$

Proof. By Lemma A.1, we have for $m, n \in \mathbb{N}$,

$$\begin{aligned}
m_P^2(f_n, f_m) &= \int_\Omega m_\gamma^2(f_n, f_m)\,\mathrm{d}P \\
&\leq \frac{1}{2}\int_\Omega \left|\ln\left(\frac{f_m}{f_n}\right)\right| \mathrm{d}P,
\end{aligned}$$

as well as

$$\begin{aligned}
\int_\Omega \left|\ln\left(\frac{f_n}{f_m}\right)\right| \mathrm{d}P &= \int_\Omega g(m_\gamma^2(f_n, f_m))\,\mathrm{d}P \\
&\leq g\left(\int_\Omega m_\gamma^2(f_n, f_m)\,\mathrm{d}P\right) \\
&= g\left(m_P^2\left(f_n, f_m\right)\right).
\end{aligned}$$

The result then follows by continuity of g. □

Lemma A.3. *For $Q_1, Q_2 \in \mathcal{C}$ such that $P \ll Q_i$ for $i \in \{1, 2\}$, we have*

$$m_P^2(q_1, q_2) \leq \frac{D(P\|Q_1 \rightsquigarrow \mathcal{C}) + D(P\|Q_2 \rightsquigarrow \mathcal{C})}{2}.$$

Proof. Let $\bar{Q}$ denote the midpoint between Q_1 and Q_2. Then we have

$$\begin{aligned}
&\frac{D(P\|Q_1 \rightsquigarrow \mathcal{C}) + D(P\|Q_2 \rightsquigarrow \mathcal{C})}{2} \\
&= \frac{\sup_{Q\in\mathcal{C}} D(P\|Q_1 \rightsquigarrow Q) + \sup_{Q\in\mathcal{C}} D(P\|Q_2 \rightsquigarrow Q)}{2} \\
&\geq \frac{D(P\|Q_1 \rightsquigarrow \bar{Q}) + D(P\|Q_2 \rightsquigarrow \bar{Q})}{2} = m_P^2(q_1, q_2).
\end{aligned}$$

□

Proof of Proposition 3.4. This follows as a direct corollary of Lemma A.2. □

We now deviate slightly from the order of the results in Section 3.3 and first state the proof of Proposition 3.6, so that we can use its results in the proof of Theorem 3.5.

Proof of Proposition 3.6. The implications (3) → (2) → (1) are obvious, so we show here only the implication (1) → (3). Assume that P' is a measure such that $-\infty < D(P\|P' \rightsquigarrow \mathcal{C}) < \infty$. Then there exists a sequence of measures $Q_n \in \mathcal{C}$ such that

$$D(P\|P' \rightsquigarrow Q_n) \to D(P\|P' \rightsquigarrow \mathcal{C})$$

for $n \to \infty$. Without loss of generality we may assume that $-\infty < D(P\|P' \rightsquigarrow Q_n) < \infty$ for all n. The result follows because

$$D(P\|P' \rightsquigarrow \mathcal{C}) = D(P\|P' \rightsquigarrow Q_n) + D(P\|Q_n \rightsquigarrow \mathcal{C})$$

and all involved quantities are finite. □

Proof of Theorem 3.5 (1). Let $(Q_n)_{n\in\mathcal{C}}$ denote a sequence in $\mathcal{C}$ such that

$$\lim_{n\to\infty} D(P\|Q_n \rightsquigarrow \mathcal{C}) = \inf_{Q\in\mathcal{C}} D(P\|Q \rightsquigarrow \mathcal{C}) = 0,$$

where the last equality follows from Proposition 3.6. Without loss of generality, we may assume that $D(P\|Q_n \rightsquigarrow \mathcal{C}) < \infty$ for all n, so that $P \ll Q_n$ for all n. It then follows from Lemma A.3 that for $m, n \in \mathbb{N}$ we have

$$m_P^2(q_m, q_n) \leq \frac{D(P\|Q_m \rightsquigarrow \mathcal{C}) + D(P\|Q_n \rightsquigarrow \mathcal{C})}{2}.$$

It follows that $(q_n)_{n\in\mathbb{N}}$ is a Cauchy sequence with respect to m_P, so that $(q_n)_{n\in\mathbb{N}}$ converges to some function $\hat{q}$ in m_P. The latter follows from the completeness of

$(\mathcal{M}\left(\Omega,(0,\infty)\right),m_P)$, i.e. Proposition 3.4.

Furthermore, suppose that $(Q'_n)_{n\in\mathcal{C}}$ is another sequence in $\mathcal{C}$ such that

$$\lim_{n\to\infty} D(P\|Q'_n \rightsquigarrow \mathcal{C}) = 0.$$

Then, by the same reasoning as before, $Q_1, Q'_1, Q_2, Q'_2, Q_3, Q'_3, \dots$ is also a Cauchy sequence that converges and since a Cauchy sequence can only converge to a single element this implies the desired uniqueness. □

Proof of Theorem 3.5 (2). The equality

$$\int_\Omega \ln\frac{p'}{\hat{q}}\,\mathrm{d}P = \lim_{n\to\infty}\int_\Omega \ln\frac{p'}{q_n}\,\mathrm{d}P$$

follows from Theorem 3.5 (1) together with the fact that convergence of q_n in m_P implies convergence of the logarithms in $L_1(P)$. □

Proof of Theorem 3.5 (3). Let $(Q_n)_{n\in\mathcal{C}}$ denote a sequence in $\mathcal{C}$ such that

$$\lim_{n\to\infty} D(P\|Q_n \rightsquigarrow \mathcal{C}) = 0.$$

Without loss of generality, we may assume that $D(P\|Q_n \rightsquigarrow \mathcal{C}) < \infty$ for all n and that q_n converges to $\hat{q}$ P-almost surely. The latter is valid, because convergence in m_P implies convergence of the logarithms in $L_1(P)$ by Lemma A.2, which gives the existence of an almost surely converging sub-sequence.

Let $\tilde{Q} = (1-t)Q_1 + tQ$ for fixed $Q \in \mathcal{C}$ and fixed $0 < t < 1$. Let $Q_{n,s}$ denote the convex combination $Q_{n,s} = (1-s_n)Q_n + s_n\tilde{Q}$ and $s_n \in [0,1]$. By Theorem 3.5 (1), we know that there exists some $\hat{Q}$ such that $q_n \to \hat{q}$ in m_P.

Since $Q_{n,s} \in \mathcal{C}$ by convexity, we have that $D(P\|Q_n \rightsquigarrow Q_{n,s}) \le D(P\|Q_n \rightsquigarrow \mathcal{C})$. We also have

$$\begin{aligned} D(P\|Q_n \rightsquigarrow Q_{n,s}) &= s_n D(P\|Q_n \rightsquigarrow \tilde{Q}) + s_n IS_P\left(\tilde{q}, q_{n,s}\right) \\ &\quad + (1-s_n) IS_P\left(q_n, q_{n,s}\right) \\ &\ge s_n D(P\|Q_n \rightsquigarrow \tilde{Q}) + s_n IS_P(\tilde{q}, q_{n,s}). \end{aligned}$$

Hence

$$s_n D(P\|Q_n \rightsquigarrow \tilde{Q}) + s_n IS_P(\tilde{q}, q_{n,s}) \le D(P\|Q_n \rightsquigarrow \mathcal{C}).$$

Division by s_n gives

$$D(P\|Q_n \rightsquigarrow \tilde{Q}) + IS_P(\tilde{q}, q_{n,s}) \leq \frac{D(P\|Q_n \rightsquigarrow \mathcal{C})}{s_n}.$$

Choosing $s_n = D(P\|Q_n \rightsquigarrow \mathcal{C})^{1/2}$, this gives

$$D(P\|Q_n \rightsquigarrow \tilde{Q}) + IS_P(\tilde{q}, q_{n,s}) \leq s_n^{1/2}.$$

Then we get

$$IS_P(\tilde{q}, q_{n,s}) \leq D(P\|\tilde{Q} \rightsquigarrow Q_n) + s_n^{1/2}.$$
$$\int_\Omega \left(\frac{\tilde{q}}{q_{n,s}} + \ln \frac{q_{n,s}}{q_n} \right) \mathrm{d}P \leq P(\Omega) + \tilde{Q}(\Omega) - Q_n(\Omega) + s_n^{1/2}.$$

Writing q_n as $\frac{q_{n,s} - s_n \tilde{q}}{1 - s_n}$, we see

$$\begin{aligned}
\ln \frac{q_{n,s}}{q_n} &= \ln \frac{q_{n,s}}{\frac{q_{n,s} - s_n \tilde{q}}{1 - s_n}} \\
&= \ln(1 - s_n) - \ln \frac{q_{n,s} - s_n \tilde{q}}{q_{n,s}} \\
&= \ln(1 - s_n) - \ln \left(1 - s_n \frac{\tilde{q}}{q_{n,s}} \right) \\
&\geq \ln(1 - s_n) + s_n \frac{\tilde{q}}{q_{n,s}}.
\end{aligned}$$

Hence

$$\ln(1 - s_n) + (1 + s_n) \int_\Omega \frac{\tilde{q}}{q_{n,s}} \,\mathrm{d}P \leq P(\Omega) + \tilde{Q}(\Omega) - Q_n(\Omega) + s_n^{1/2}.$$

As $\lim_{n\to\infty} s_n = 0$, taking the limit inferior as $n \to \infty$ on both sides gives

$$\liminf_{n\to\infty} \int_\Omega \frac{\tilde{q}}{q_{n,s}} \,\mathrm{d}P \leq P(\Omega) + \tilde{Q}(\Omega) - \liminf_{n\to\infty} Q_n(\Omega).$$

An application of Fatou's lemma gives

$$\int_\Omega \frac{\mathrm{d}P}{\mathrm{d}\hat{Q}} \,\mathrm{d}\tilde{Q} \leq P(\Omega) + \tilde{Q}(\Omega) - \liminf_{n\to\infty} Q_n(\Omega).$$

Since $\tilde{Q} = (1-t)Q_1 + tQ$ we get the inequality

$$\begin{aligned}
&\int_\Omega \frac{\mathrm{d}P}{\mathrm{d}\hat{Q}} \,\mathrm{d}\left((1-t)\,Q_1 + tQ\right) \\
&\qquad \leq P(\Omega) + (1-t)Q_1(\Omega) + tQ(\Omega) - \liminf_{n\to\infty} Q_n(\Omega), \\
&(1-t)\int_\Omega \frac{\mathrm{d}P}{\mathrm{d}\hat{Q}} \,\mathrm{d}Q_1 + t\int_\Omega \frac{\mathrm{d}P}{\mathrm{d}\hat{Q}} \,\mathrm{d}Q \\
&\qquad \leq P(\Omega) + (1-t)Q_1(\Omega) + tQ(\Omega) - \liminf_{n\to\infty} Q_n(\Omega).
\end{aligned}$$

Finally we let t tend to one and obtain the desired result. □

Proof of Proposition 3.7. Let $Q \in \mathcal{C}$ arbitrarily. Then there exists a sequence $(w_i)_{i=1}^n$ in $[0,1]$ with $\sum_i w_i = 1$ such that $Q = \sum_{i=1}^n w_i Q_i$. It follows that

$$\begin{aligned}
D\left(P \| \frac{1}{n}\sum_i Q_i \rightsquigarrow Q\right) &= \int_\Omega \ln \frac{\sum_i w_i Q_i}{\frac{1}{n}\sum_i Q_i} \,\mathrm{d}P \\
&\leq \int_\Omega \ln \frac{\max_i w_i \sum_i Q_i}{\frac{1}{n}\sum_i Q_i} \,\mathrm{d}P \\
&= \ln(n) + \ln(\max_i w_i) \leq \ln(n).
\end{aligned}$$

The proposition follows by taking the supremum over Q on both sides. □

Proof of Proposition 3.8. Since Q^* is the normalized maximum likelihood distribution we have $\sup_Q \sup_\omega \ln \frac{\mathrm{d}Q}{\mathrm{d}Q^*} < \infty$. In particular

$$\begin{aligned}
\sup_{Q\in\mathcal{C}} D(P\|Q^* \rightsquigarrow Q) &= \sup_{Q\in\mathcal{C}} \int_\Omega \ln \frac{\mathrm{d}Q}{\mathrm{d}Q^*} \,\mathrm{d}P \\
&\leq \sup_{Q\in\mathcal{C}} \sup_\omega \ln \frac{\mathrm{d}Q}{\mathrm{d}Q^*}(\omega) < \infty.
\end{aligned}$$

□

Proof of Proposition 3.10. We can write

$$D(P\|Q_\theta \rightsquigarrow Q^*) = D(P\|Q_\theta \rightsquigarrow Q) + D(P\|Q \rightsquigarrow Q^*).$$

By assumption all terms are finite so that minimising $D(P\|Q_\theta \rightsquigarrow Q^*)$ over θ must be equivalent to minimising $D(P\|Q_\theta \rightsquigarrow Q)$ over θ. The same argument holds for step 5 in Algorithm 1. The result then follows from (Brinda, 2018, Theorem 3.0.13). Whereas the algorithm described there works by choosing θ_k to minimize $\int_\Omega \log((1-\alpha_k)q_{\theta_{k-1}} +$

$\alpha_k q_\theta)\,\mathrm{d}P$, the proof relies on (Li, 1999, Lemma 5.9), which indeed uses minimization of $D(P\|(1-\alpha_k)Q_{\theta_{k-1}}+\alpha_k Q_\theta \rightsquigarrow Q)$ as described here. □

Proof of Theorem 3.9. For any $a\in\mathbb{R}$ we have

$$f_0\left(i\right)+a\cdot f_1\left(i\right)=f_0\left(i\right)\cdot\left(1+a\cdot\frac{f_1\left(i\right)}{f_0\left(i\right)}\right). \tag{A.1}$$

Since $\frac{f_1(i)}{f_0(i)}\to 0$ for $i\to\infty$ we have that $f_0\left(i\right)+a\cdot f_1\left(i\right)\geq 0$ for i sufficiently large. Therefore, we can apply Fatou's lemma to the function and obtain

$$\begin{aligned}
&\sum f_0\left(i\right)\cdot q^*\left(i\right)+a\cdot\sum f_1\left(i\right)\cdot q^*\left(i\right)\\
&\quad=\sum\left(f_0\left(i\right)+a\cdot f_1\left(i\right)\right)\cdot q^*\left(i\right)\\
&\quad=\sum\liminf_{n\to\infty}\left(f_0\left(i\right)+a\cdot f_1\left(i\right)\right)\cdot q_n\left(i\right)\\
&\quad\leq\liminf_{n\to\infty}\sum_i\left(f_0\left(i\right)+a\cdot f_1\left(i\right)\right)\cdot q_n\left(i\right)\\
&\quad=\liminf_{n\to\infty}\left(\sum_i f_0\left(i\right)\cdot q_n\left(i\right)+a\cdot\sum_i f_1\left(i\right)\cdot q_n\left(i\right)\right)\\
&\quad=\liminf_{n\to\infty}\left(\lambda_0+a\cdot\lambda_1\right)=\lambda_0+a\cdot\lambda_1.
\end{aligned}$$

Hence

$$a\cdot\left(\sum f_1\left(i\right)\cdot q^*\left(i\right)-\lambda_1\right)\leq\lambda_0-\sum f_0\left(i\right)\cdot q^*\left(i\right). \tag{A.2}$$

This inequality should hold for all $a\in\mathbb{R}$, which is only possible if

$$\begin{aligned}
\sum f_1\left(i\right)\cdot q^*\left(i\right)-\lambda_1&=0.\\
\sum f_1\left(i\right)\cdot q^*\left(i\right)&=\lambda_1.
\end{aligned}$$

□

A.1.2 Proofs for Section 3.4

Proof of Proposition 3.12. Assume that $E_1,E_2,E_3,\ldots$ is a sequence of e-variables such that

$$\int_\Omega \ln\left(\frac{E_n}{E'}\right)\mathrm{d}P\to\sup_E\int_\Omega\ln\left(\frac{E}{E'}\right)\mathrm{d}P$$

for $n \to \infty$. Then $E_{n,m} = (E_m + E_n)/2$ are also e-variables and by convexity

$$\int_\Omega \ln\left(\frac{E_{m,n}}{E'}\right) \mathrm{d}P \to \sup_E \int_\Omega \ln\left(\frac{E}{E'}\right) \mathrm{d}P\,,$$

which implies that $m_\gamma^2(E_m, E_n) \to 0$ for $m, n \to \infty$. By completeness E_n converges to some e-variable E_∞. Using Lemma A.2 we see that $m_\gamma(E_n, E_\infty) \to 0$ implies that

$$\int_\Omega \ln\left(\frac{E_m}{E'}\right) \mathrm{d}P \to \int_\Omega \ln\left(\frac{E_\infty}{E'}\right) \mathrm{d}P$$

so that

$$\sup_E \int_\Omega \ln\left(\frac{E}{E'}\right) \mathrm{d}P = \int_\Omega \ln\left(\frac{E_\infty}{E'}\right) \mathrm{d}P\,.$$

Hence

$$\sup_E \int_\Omega \ln\left(\frac{E}{E_\infty}\right) \mathrm{d}P = 0$$

Therefore E_∞ is a strongest e-statistic.

Assume that both E_1 and E_2 are strongest e-variables. Then they are both stronger than the average $\bar{E} = (E_1 + E_2)/2$. Hence

$$0 \leq m_\gamma^2(E_1, E_2) = \frac{1}{2}\int \left(\ln\left(\frac{\bar{E}}{E_1}\right) + \ln\left(\frac{\bar{E}}{E_2}\right)\right) \mathrm{d}P \leq 0.$$

Therefore $E_1 = E_2$ P-almost surely. □

Proof of Theorem 3.13. Firstly, since $\hat{E} > 0$ holds P-almost surely, we have that $\hat{E}$ is stronger than any $E' \in \mathcal{E}_\mathcal{C}$ with $P(E' = 0) > 0$.

Secondly, let $E \in \mathcal{E}_\mathcal{C}$ be an e-statistic for which $E > 0$ holds P-almost surely. Furthermore, let Q_n be a sequence of measures in $\mathcal{C}$ such that $D(P\|Q_n \rightsquigarrow \mathcal{C}) \to 0$. We can define a sequence of sub-probability measures R_n by $R_n(F) = \int_F E \,\mathrm{d}Q_n$, which satisfies ${}^{\mathrm{d}R_n}/_{\mathrm{d}Q_n} = E$. We see

$$\begin{aligned}\int_\Omega \ln\left(\frac{\hat{E}}{E}\right) \mathrm{d}P &= \int_\Omega \ln\left(\frac{\mathrm{d}Q_n}{\mathrm{d}\hat{Q}}\right) \mathrm{d}P + D(P\|R_n)\\ &\quad + (P(\Omega) - R_n(\Omega))\\ &\geq \int_\Omega \ln\left(\frac{\mathrm{d}Q_n}{\mathrm{d}\hat{Q}}\right) \mathrm{d}P.\end{aligned}$$

The last expression goes to zero as $n \to \infty$, so we see that $\hat{E}$ is stronger than E. □

Proof of Proposition 3.14. Using the fact that $\ln(x) \le x - 1$ for $x > 0$, we see

$$\begin{aligned} D(P \| Q^* \rightsquigarrow Q) &= \int_\Omega \ln \frac{\mathrm{d}Q}{\mathrm{d}Q^*} \,\mathrm{d}P \\ &\le \int_\Omega \left(\frac{\mathrm{d}Q}{\mathrm{d}Q^*} - 1 \right) \mathrm{d}P \\ &= \int_\Omega \frac{\mathrm{d}P}{\mathrm{d}Q^*} \,\mathrm{d}Q - 1 \le 0, \end{aligned}$$

where the last inequality follows from the fact that ${}^{\mathrm{d}P}/_{\mathrm{d}Q^*}$ is an e-statistic. □

Proof of Theorem 3.16. Without loss of generality, assume that $\int_\Omega {}^{q'}/_{q} \,\mathrm{d}P = 1 + \epsilon$ for some $\epsilon > 0$. For the sake of brevity, we write $c_\beta := \|{}^{q'}/_{q}\|_{1+\beta}^{1+\beta}$. We now define a function $g : [0,1] \to \mathbb{R}_{\ge 0}$ as

$$g(\alpha) := D\left(P \| (1-\alpha)Q + \alpha Q' \rightsquigarrow \mathcal{C}\right).$$

Notice that $g(0) = \delta$ and $g(\alpha) \ge 0$, since $(1-\alpha)Q + \alpha Q' \in \mathcal{C}$. This function and its derivatives will guide the rest of the proofs, and we now list some properties that we will need:

$$g'(\alpha) := \frac{\mathrm{d}}{\mathrm{d}\alpha} g(\alpha) = \int_\Omega \frac{q - q'}{(1-\alpha)q + \alpha q'} \,\mathrm{d}P, \tag{A.3}$$

so that

$$g'(0) = \int_\Omega \left(1 - \frac{q'}{q}\right) \mathrm{d}P = -\epsilon, \tag{A.4}$$

$$g''(\alpha) := \frac{\mathrm{d}^2}{\mathrm{d}\alpha^2} g(\alpha) = \int_\Omega \left(\frac{q' - q}{(1-\alpha)q + \alpha q'} \right)^2 \mathrm{d}P, \tag{A.5}$$

so that

$$g''(0) = \int_\Omega \left(1 - \frac{q'}{q}\right)^2 \mathrm{d}P = 1 - 2(1+\epsilon) + c_1$$

and

$$0 \le g''(\alpha) \le \frac{1}{(1-\alpha)^2} g''(0). \tag{A.6}$$

We now prove (3.10). We start with the case $\beta = 1$ and will use the result for $\beta = 1$ to prove the case for $\beta < 1$. The proof for the case $\beta > 1$ comes later; it requires a completely different proof.

Case $\beta = 1$. The general idea is simple: at $\alpha = 0$ the function $g(\alpha)$ is equal to δ and has derivative $-\epsilon$. Its second derivative is positive and bounded by constant times $g''(0) \leq c_1$ for all $\alpha \leq {}^1\!/_2$. Thus, if ϵ is larger then a certain threshold, $g(\alpha)$ will become negative at some $\alpha \leq {}^1\!/_2$, but this is not possible since g is a description gain and we would arrive at a contradiction. The details to follow simply amount to calculating the threshold as a function of δ.

By Taylor's theorem, we have for any $\alpha \in [0, {}^1\!/_2]$ that

$$\begin{aligned} g(\alpha) &= g(0) + g'(0)\alpha + \max_{0 \leq \alpha^\circ \leq \alpha} \frac{g''(\alpha^\circ)}{2}\alpha^2 \\ &\leq g(0) + g'(0)\alpha + 2g''(0)\alpha^2 \\ &\leq \delta - \epsilon\alpha + 2\alpha^2 c_1, \end{aligned}$$

where we use the properties derived above. This final expression has a minimum in $\alpha^* = \min\{{}^\epsilon\!/_{4c_1}, {}^1\!/_2\}$. By nonnegativity of g, we know that $\delta - \epsilon\alpha^* + 2\alpha^{*2}c_1 \geq 0$. This gives $\epsilon \leq (8c_1\delta)^{1/2}$ in the case that $\alpha^* = {}^\epsilon\!/_{4c_1} < {}^1\!/_2$, and $\epsilon \leq 2\delta + c_1$ otherwise. In the latter case, it holds that $c_1 < {}^\epsilon\!/_2$, so the bound can be loosened slightly to find the simplification $\epsilon \leq 4\delta$. This concludes the proof for $\beta = 1$, which we now use to prove Case $\beta < 1$.

Case $\beta < 1$. For any $a > 0$, it holds that

$$\int_\Omega \frac{q'}{q}\,\mathrm{d}P = \int_\Omega \frac{q'}{q}\mathbf{1}_{\{q'/q \leq a\}}\,\mathrm{d}P + \int_\Omega \frac{q'}{q}\mathbf{1}_{\{q'/q > a\}}\,\mathrm{d}P. \tag{A.7}$$

We write $q'' := q'\mathbf{1}_{\{q'/q \leq a\}}$ and we will bound the first term on the right-hand side of (A.7) using the proof above with Q' replaced by Q''. Since Q'' is not necessarily an element of $\mathcal{C}$, we need to verify nonnegativity, which follows because for each $\alpha \in (0,1)$, we have that $D(P \| (1-\alpha)Q + \alpha Q'' \rightsquigarrow \mathcal{C}) \geq D(P \| (1-\alpha)Q + \alpha Q' \rightsquigarrow \mathcal{C}) \geq 0$.

Furthermore, it holds that

$$\begin{aligned}\left\|\frac{q''}{q}\right\|_2^2 &= \int_\Omega \left(\frac{q''}{q}\right)^2 \mathrm{d}P \\ &= \int_\Omega \left(\frac{q''}{q}\right)^{1+\beta} \left(\frac{q''}{q}\right)^{1-\beta} \mathrm{d}P \\ &\leq a^{1-\beta} c_\beta\end{aligned}$$

The results above therefore give

$$\int_\Omega \frac{q''}{q}\,\mathrm{d}P \leq 1 + \max\{(8a^{1-\beta}c_\beta\delta)^{1/2}, 2\delta\}.$$

For the second term on the right-hand side of (A.7), we use a Markov-type bound, i.e.

$$\begin{aligned}\int_\Omega \frac{q'}{q}\mathbf{1}_{\{q'/q>a\}}\,\mathrm{d}P &\leq \int_\Omega \frac{q'}{q}\left(\frac{q'/q}{a}\right)^\beta \mathbf{1}_{\{q'/q>a\}}\,\mathrm{d}P \\ &\leq a^{-\beta}c_\beta.\end{aligned}$$

Putting this together gives

$$\int_\Omega \frac{q'}{q}\,\mathrm{d}P \leq 1 + \max\{(8a^{1-\beta}c_\beta\delta)^{1/2}, 4\delta\} + a^{-\beta}c_\beta.$$

Since this holds for any a, we now pick it to minimize this bound. To this end, consider

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}a}(8a^{1-\beta}c_\beta\delta)^{1/2} + a^{-\beta}c_\beta \\ &\qquad = \frac{(1-\beta)(8c_\beta\delta)^{1/2}}{2}a^{-(1+\beta)/2} - \beta a^{-(1+\beta)}c_\beta.\end{aligned}$$

Setting this to zero, we find

$$a^* = \left(\frac{\beta c_\beta^{1/2}}{(1-\beta)(2\delta)^{1/2}}\right)^{\frac{2}{1+\beta}}.$$

The proof is concluded by noting that

$$\begin{aligned}(8a^{*1-\beta}c_\beta\delta)^{1/2} &= \left(8\left(\frac{\beta c_\beta^{1/2}}{(1-\beta)(2\delta)^{1/2}}\right)^{2\frac{1-\beta}{1+\beta}} c_\beta\delta\right)^{1/2}\\ &= 2c_\beta^{1/(\beta+1)}(2\delta)^{\beta/(\beta+1)}\left(\frac{\beta}{1-\beta}\right)^{\frac{1-\beta}{1+\beta}}\end{aligned}$$

and

$$\begin{aligned}a^{*-\beta}c_\beta &= \left(\frac{\beta c_\beta^{1/2}}{(1-\beta)(2\delta)^{1/2}}\right)^{\frac{-2\beta}{1+\beta}} c_\beta\\ &= c_\beta^{1/(\beta+1)}\left(\frac{\beta}{1-\beta}\right)^{\frac{-2\beta}{1+\beta}}(2\delta)^{\beta/(1+\beta)}.\end{aligned}$$

Case $\beta > 1$. We now prove the result for $\beta \in (1, \infty)$; the proof for $\beta = \infty$ follows by a minor modification of (A.9). If $\epsilon \leq 0$ there is nothing to prove, so without loss of generality we can write $\epsilon = \gamma\delta$ for some $\gamma > 0$; we will bound γ. Whereas the previous proof exploited the fact that the second derivative $g''(\alpha)$ was bounded above in terms of δ and hence 'not too large', the proof below uses the condition that c_β is finite to show first, (a), that $g''(\alpha)$ can also be bounded *below* in terms of (γ, δ). Therefore, if ϵ exceeds a certain threshold, as α moves away from the α^* at which $g(\alpha)$ achieves its minimum in the direction of the furthest boundary point (i.e. if $\alpha^* < 1/2$, we consider $\alpha \uparrow 1$, if $\alpha^* \geq 1/2$ we consider $\alpha \downarrow 0$), $g(\alpha)$ will become larger than $K\delta$ or δ respectively, and we arrive at a contradiction. (b) below gives the detailed calculation of this threshold.

Proof of (a). Fix some $0 \leq \tilde{\alpha} < 1$ (we will derive a bound for any such $\tilde{\alpha}$ and later optimize for $\tilde{\alpha}$; for a sub-optimal yet easier derivation take $\tilde{\alpha} = 1/2$). By Taylor's theorem, we have $0 \leq g(\tilde{\alpha}) = \delta - \tilde{\alpha}\epsilon + (1/2)\tilde{\alpha}^2 g''(\alpha^\circ)$ for some $0 \leq \alpha^\circ \leq \tilde{\alpha}$. Plugging in $\epsilon = \gamma\delta$ we find that

$$g''(\alpha^\circ) \geq \frac{2}{\tilde{\alpha}^2}(\tilde{\alpha}\gamma - 1)\delta.$$

This gives a lower bound on $g''(\alpha^\circ)$ for *some* α° in terms of (γ, δ). We now turn this into a weaker lower bound on *all* α. First, using (A.6) and then $\alpha^\circ \leq \tilde{\alpha}$ and then the above lower bound, we find

$$\begin{aligned} g''(0) &\geq \max_{\alpha \in [0, \tilde{\alpha}]} (1-\alpha)^2 g''(\alpha) \geq (1-\alpha^\circ)^2 g''(\alpha^\circ) \\ &\geq (1-\tilde{\alpha})^2 g''(\alpha^\circ) \geq 2 f_{\tilde{\alpha}}(\gamma, \delta), \end{aligned} \tag{A.8}$$

where $f_{\tilde{\alpha}}(\gamma, \delta) := ((1-\tilde{\alpha})/\tilde{\alpha})^2 (\tilde{\alpha}\gamma - 1)\delta$ is a function that is linear in γ and δ. We have now lower bounded $g''(0)$ in terms of γ, δ. We next show that, under our condition that $c_\beta < \infty$, this implies a (weaker) lower bound on $g''(\alpha)$ for all α. For this, fix any $C > 1$. We have for all $0 < \alpha \leq 1$:

$$\begin{aligned} g''(\alpha) &\geq \int_\Omega \mathbf{1}_{q' \leq Cq} \cdot \left(\frac{q' - q}{(1-\alpha)q + \alpha q'} \right)^2 \mathrm{d}P \\ &\geq \int_\Omega \mathbf{1}_{q' \leq Cq} \cdot \left(\frac{q' - q}{(1-\alpha)q + \alpha C q} \right)^2 \mathrm{d}P \\ &= \int_\Omega \mathbf{1}_{q' \leq Cq} \cdot \left(\frac{q' - q}{q} \right)^2 \mathrm{d}P \cdot \frac{1}{(1 + \alpha(C-1))^2} \\ &\geq \frac{1}{(1 + (C-1))^2} \left(g''(0) - \int_\Omega \mathbf{1}_{q' > Cq} \left(\frac{q'}{q} - 1 \right)^2 \mathrm{d}P \right) \\ &\geq \frac{1}{C^2} \left(2 f_{\tilde{\alpha}}(\gamma, \delta) - C^{1-\beta} c_\beta \right), \end{aligned} \tag{A.9}$$

where in the fourth line we used the definition of $g''(0)$, and in the fifth line we used (A.8) and a Markov-type bound on the integral, i.e. we used that $\int_\Omega \mathbf{1}_{q' > Cq} \cdot (q'/q - 1)^2 \, \mathrm{d}P$ is bounded by

$$\begin{aligned} \int_\Omega \mathbf{1}_{q' > Cq} \cdot \left(\frac{q'}{q} \right)^2 \mathrm{d}P &\leq \int_\Omega \left(\frac{q'/q}{C} \right)^{\beta - 1} \cdot \left(\frac{q'}{q} \right)^2 \mathrm{d}P \\ &= C^{1-\beta} c_\beta. \end{aligned}$$

By differentiation we can determine the C that maximizes the bound (A.9). This gives $C^{1-\beta} = f_{\tilde{\alpha}}(\gamma, \delta)(4/c_\beta(1+\beta))$. and with this choice of C, (A.9) becomes

$$g''(\alpha) \geq f_{\tilde{\alpha}}(\gamma, \delta)^{(\beta+1)/(\beta-1)} c_\beta^{2/(1-\beta)} h(\beta) \tag{A.10}$$

where $h(\beta) = (4/(1+\beta))^{2/(\beta-1)} \cdot 2(\beta - 1)/(1+\beta)$. We are now ready to continue to:

Proof of (b). Let $\alpha^* \in [0, 1]$ be the point at which $g(\alpha)$ achieves its minimum. If

$\alpha^* \leq {}^1\!/\!_2$, a second-order Taylor approximation of $g(1)$ around α^* gives that

$$\begin{aligned} K\delta \geq g(1) &\geq \frac{1}{2}(1-\alpha^*)^2 \min_{\alpha \in [\alpha^*, 1]} g''(\alpha) \\ &\geq \frac{1}{8} f_{\tilde{\alpha}}(\gamma, \delta)^{(\beta+1)/(\beta-1)} c_\beta^{2/(1-\beta)} h(\beta), \end{aligned}$$

so that after some manipulations

$$f_{\tilde{\alpha}}(\gamma, \delta)^{(1+\beta)/(\beta-1)} \leq 8K' c_\beta^{2/(\beta-1)} \cdot h(\beta)^{-1} \delta, \tag{A.11}$$

with $K' = K$. If $\alpha^* > {}^1\!/\!_2$, we perform a completely analogous second-order Taylor approximation of $g(0)$ around α^*, which will then give (A.11) again but with K' replaced by 1. We thus always have (A.11) with $K' = \max\{K, 1\}$. Unpacking $f_{\tilde{\alpha}}$ in (A.11) and rearranging gives:

$$\gamma \leq \frac{\tilde{\alpha}}{(1-\tilde{\alpha})^2} \cdot V + \frac{1}{\tilde{\alpha}}$$

with

$$V = c_\beta^{2/(1+\beta)} \cdot \left(\frac{8K'}{h(\beta)}\right)^{\frac{\beta-1}{1+\beta}} \delta^{\frac{-2}{1+\beta}}.$$

We now pick the $\tilde{\alpha}$ that makes both terms on the right equal, so that the right-hand side becomes equal to ${}^2\!/\!_{\tilde{\alpha}}$. This is the solution to the equation $({}^{\tilde{\alpha}}\!/\!_{(1-\tilde{\alpha})})^2 V = 1$ which must clearly be obtained for some $0 < \tilde{\alpha} < 1$, so this $\tilde{\alpha}$ satisfies our assumptions. Basic calculation gives

$$\gamma \leq \frac{2}{\tilde{\alpha}} = 2 \cdot \left(V^{1/2} + 1\right)$$

and unpacking V we obtain

$$\epsilon = \gamma\delta \leq c^* \cdot \delta^{\frac{\beta}{1+\beta}} + 2\delta.$$

where

$$c^* = c_\beta^{1/(1+\beta)} \cdot \left(\frac{8K'}{h(\beta)}\right)^{\frac{\beta-1}{2(1+\beta)}}.$$

Unpacking $h(\beta)$ gives the desired result.

□

A.2 RIPr Strict Sub-Probability Measure

In this appendix, we discuss a general way to construct a measure P and convex set of distributions $\mathcal{C}$ such that the reverse information projection of P on $\mathcal{C}$ is a strict sub-probability measure. For simplicity, we take $\Omega = \mathbb{N}$ and $\mathcal{F} = 2^{\mathbb{N}}$, though the idea should easily translate to more general settings.

Proposition A.4. *Let $g : \mathbb{N} \to \mathbb{R}_{>0}$ be a function, and let $\mathcal{C}$ denote the set of measures $\{Q : \sum_i g\,(i)\, q(i) \leq \nu\}$ for some $\nu > 0$. Then for any P that is not in $\mathcal{C}$ we have that $E\,(i) = {}^{g(i)}/_{\nu}$ is the optimal e-statistic.*

Proof. The extreme points in $\mathcal{C}$ are the measure with total mass 0 and measures of the form $\frac{\nu}{g(i)}\delta_i$, i.e. measures concentrated in single points. An e-statistic E must satisfy

$$\sum_j E\,(j)\,\frac{\nu}{g\,(i)}\delta_i\,(j) \leq 1$$

or, equivalently, $E\,(i)\,\frac{\nu}{g(i)} \leq 1$. Hence $E \leq {}^{g}/_{\nu}$ so the optimal e-statistic is ${}^{g}/_{\nu}$. □

Let $g : \mathbb{N} \to \mathbb{R}_{>0}$ be any function that satisfies

$$\lim_{n\to\infty} g(n) = 0.$$

Furthermore, let P denote a probability measure on the natural numbers such that

$$\sum_i \frac{p(i)}{g(i)} = c$$

for some $c \in \mathbb{R}_{>0}$. Fix $\nu^* \in (0, 1/c)$ and let $\mathcal{C}_{\nu^*}$ denote the set of measures $\{Q : \sum_i g\,(i)\,q(i) \leq \nu^*\}$. Note that we do not yet require all measures in $\mathcal{C}_{\nu^*}$ to be probability measures so that the set $\mathcal{C}_{\nu^*}$ is compact. It follows that there exists a unique element of $\mathcal{C}_{\nu^*}$ that minimizes $\sum_i p(i) \ln\left(p(i)/q(i)\right)$.

The optimal e-statistic is $E_{\nu^*} = {}^{g}/_{\nu^*}$, and we may define the measure Q_{ν^*} by

$$q_{\nu^*}(i) = \frac{p(i)}{E_{\nu^*}(i)} = \nu^* p(i)/g(i),$$

and we can check that $Q_{\nu^*} \in \mathcal{C}_{\nu^*}$. Hence Q_{ν^*} minimizes $\sum_i p(i) \ln\left(p(i)/q(i)\right)$.

This is a strict sub-probability measure:

$$\begin{aligned}\sum_i q_{\nu^*}(i) &= \nu^* \sum_i \frac{p(i)}{g(i)} \\ &= \nu^* c \\ &< 1,\end{aligned}$$

where we use that $\nu^* < {}^1\!/_c$.

The next step is to prove that the information projection does not change if we restrict to the set of probability measures in $\mathcal{C}_{\nu^*}$, which we denote by $\tilde{\mathcal{C}}_{\nu^*}$. To this end, note first that for $\nu < \nu^*$, we have that $\sum g\,(i)\, q_\nu(i) < \nu^*$, so that for all $\nu < \nu^*$ there exists $n_\nu \in \mathbb{N}$ such that the probability measure defined by

$$q_\nu(i) + \left(1 - \sum_j q_\nu(j)\right) \delta_{n_\nu}\,(i)$$

is an element of $\tilde{\mathcal{C}}_{\nu^*}$. Hence

$$\begin{aligned}D(P\|\tilde{C}_{\nu^*}) \leq D &\left(P \middle\| Q_\nu + \left(1 - \sum_{j\in\mathbb{N}} q_\nu(j)\right) \delta_{n_\nu}\right) \\ &= \sum_{i\in\mathbb{N}} p(i) \ln \left(\frac{p(i)}{Q_\nu(i) + \left(1 - \sum_{j\in\mathbb{N}} q_\nu(j)\right) \delta_{n_\nu}\,(i)}\right) \\ &= -p(n_\nu) \ln \left(\frac{p(n_\nu)}{q_\nu(n_\nu)}\right) \\ &\quad + p(n_\nu) \ln \left(\frac{p(n_\nu)}{q_\nu(n_\nu) + 1 - \sum_{j\in\mathbb{N}} q_\nu(j)}\right) \\ &\quad + \sum_{i\in\mathbb{N}} p(i) \ln \left(\frac{p(i)}{q_\nu(i)}\right).\end{aligned}$$

The first term can be written as

$$\begin{aligned}p(n_\nu) \ln \left(\frac{p(n_\nu)}{q_\nu(n_\nu)}\right) &= q_\nu(n_\nu) \frac{p(n_\nu)}{q_\nu(n_\nu)} \ln \left(\frac{p(n_\nu)}{q_\nu(n_\nu)}\right) \\ &= q_\nu(n_\nu) \frac{g\,(n_\nu)}{\nu} \ln \left(\frac{g\,(n_\nu)}{\nu}\right)\end{aligned}$$

Then notice that for $\nu \to \nu^*$, we must have that $n_\nu \to \infty$. Using that $c \ln (c) \to 0$ for

$c \to 0$ we see the first term tends to 0 for $\nu \to \nu^*$. Similarly, the second term can be written as

$$\begin{aligned}
&p(n_\nu) \ln \left(\frac{p(n_\nu)}{q_\nu(n_\nu) + 1 - \sum_{j \in \mathbb{N}} q_\nu(j)} \right) \\
&\quad = \left(q_\nu(n_\nu) + 1 - \sum_{j \in \mathbb{N}} q_\nu(j) \right) \frac{p(n_\nu)}{q_\nu(n_\nu) + 1 - \sum_{j \in \mathbb{N}} q_\nu(j)} \\
&\qquad \cdot \ln \left(\frac{p(n_\nu)}{q_\nu(n_\nu) + 1 - \sum_{j \in \mathbb{N}} q_\nu(j)} \right).
\end{aligned}$$

We also have

$$\frac{p(n_\nu)}{q_\nu(n_\nu) + 1 - \sum_i q_\nu(i)} \to 0$$

for $\nu \to \nu^*$ and using that $c \ln(c) \to 0$ for $c \to 0$ we get the second term tends to 0 for $\nu \to \nu^*$. Therefore we see

$$\begin{aligned}
D(P \| \tilde{\mathcal{C}}_{\nu^*}) &\leq \lim_{\nu \to \nu^*} D \left(P \middle\| Q_\nu + \left(1 - \sum_i q_\nu(i) \right) \delta_{n_\nu} \right) \\
&\leq \sum_i p(i) \ln \left(\frac{p(i)}{q_{\nu^*}(i)} \right) \\
&= \inf_{Q \in \mathcal{C}_{\nu^*}} \sum_i p(i) \ln \left(\frac{p(i)}{q(i)} \right).
\end{aligned}$$

The inequality trivially also holds the other way around, so we find that

$$D(P \| \tilde{\mathcal{C}}_{\nu^*}) = \inf_{Q \in \mathcal{C}_{\nu^*}} \sum_i p(i) \ln \left(\frac{p(i)}{q(i)} \right).$$

It follows that Q_{ν^*} is a strict sub-probability measure, and at the same time it is the reverse information projection of P onto $\tilde{\mathcal{C}}_{\nu^*}$.

A.3 Convexity

One of the main assumptions made throughout the main text is that the set of measures $\mathcal{C}$ is convex, i.e. closed under finite mixtures. However, one can also consider stronger notions of convexity, such as σ-convexity and Choquet-convexity. In this appendix, we investigate whether considering different levels of convexity can change the reverse information projection.

Definition A.5. A set $\mathcal{C}'$ of measures is said to be *σ-convex* if $Q_1\,, Q_2\,, Q_3 \cdots \in \mathcal{C}'$ implies that $\sum_{i=1}^{\infty} w_i Q_i \in \mathcal{C}'$ when $w_i \geq 0$ and $\sum_{i=1}^{\infty} w_i = 1$. The σ-convex hull of a set of measures $\mathcal{C}$, denoted by σ-conv$(\mathcal{C})$, is the smallest σ-convex set containing $\mathcal{C}$.

In order to avoid topological complications we will restrict the discussion of Choquet-convexity to Polish spaces, i.e. spaces for which there exists a complete metric that generates the topology. That is, assume that Ω is a Polish space equipped with the Borel σ-algebra. Let Θ be another Polish space and let $\{Q_\theta : \theta \in \Theta\}$ denote a parameterized set of probability measures on Ω such that $\theta \to \int_\Omega f \, \mathrm{d}Q_\theta$ is Borel measurable for any measurable function $f : \Omega \to \mathbb{R}$. Then for any probability measure ν on Θ the *Choquet-convex mixture* μ_ν can be defined by

$$\int_\Omega f \, \mathrm{d}\mu_\nu = \int_\Theta \left(\int_\Omega f \, \mathrm{d}\mu_\theta \right) \mathrm{d}\nu,$$

for any measurable function $f : \Omega \to \mathbb{R}$.

Definition A.6. A set $\mathcal{C}'$ of measures is said to be *Choquet-convex* if it is closed under Choquet convex mixtures. The Choquet-convex hull of a set of measures $\mathcal{C}$ is the smallest Choquet-convex set that contains $\mathcal{C}$.

So far, we have assumed that all of the measures in $\mathcal{C}$ are finite. However, a countable or Choquet convex mixture of finite measures may not be finite. It follows that our results on the existence of the RIPr might not be applicable to the σ-convex and Choquet-convex hull of $\mathcal{C}$. We therefore assume for the remainder of this section that all involved measures are sub-probability measures, in which case this problem does not arise. With all of this in place, it is relatively straightforward to construct examples where the RIPr of P on a convex set does not exist, whereas the RIPr of P on its σ-convex hull does exist.

Example A.1. Let P denote a geometric distribution on $\mathbb{N}_0$ and let $\mathcal{C}$ denote the set of probability measures on $\mathbb{N}_0$ with finite support. Then $D(P \| Q \rightsquigarrow \mathcal{C}) = -\infty$ for any $Q \in \mathcal{C}$. Therefore the reverse information projection of P on $\mathcal{C}$ is not defined according to the definitions given in Chapter 3. However, the σ-convex hull of $\mathcal{C}$ consists of all probability measures on $\mathbb{N}_0$, which implies that the reverse information projection on the σ-convex hull is well-defined and equals P.

However, as the following results show, if the RIPr of P on $\mathcal{C}$ does exist, then it must coincide with the RIPr of P on σ-conv$(\mathcal{C})$.

Lemma A.7. *Let P and Q be sub-probability measures and let $Q_1, Q_2, \ldots$ be a sequence of sub-probability measures such that $D(P\|Q \rightsquigarrow Q_1) > -\infty$, and let $w_1, w_2, \ldots$ be a sequence of positive numbers with sum 1. Then*

$$D\left(P \middle\| Q \rightsquigarrow \frac{\sum_{i=1}^{n} w_i \cdot Q_i}{\sum_{i=1}^{n} w_i}\right) \to D\left(P \middle\| Q \rightsquigarrow \sum_{i=1}^{\infty} w_i \cdot Q_i\right)$$

for $n \to \infty$.

Proof. Firstly, note that

$$\ln \frac{\mathrm{d}\sum_{i=1}^{n+1} w_i Q_i}{\mathrm{d}Q} \geq \ln \frac{\mathrm{d}\sum_{i=1}^{n} w_i Q_i}{\mathrm{d}Q}$$

and

$$\begin{aligned}\int_\Omega \ln \frac{\mathrm{d}\sum_{i=1}^{n} w_i Q_i}{\mathrm{d}Q}\,\mathrm{d}P &\geq \int_\Omega \ln \frac{\mathrm{d}w_1 Q_1}{\mathrm{d}Q}\,\mathrm{d}P \\ &= D(P\|Q \rightsquigarrow Q_1) + \ln w_1 \\ &\quad + (Q_1(\Omega) - Q(\Omega)) \\ &> -\infty.\end{aligned}$$

Since $\sum_{i=1}^{n} w_i q_i \to \sum_{i=1}^{\infty} w_i q_i$ pointwise, applying the monotone convergence theorem to the sequence

$$\left(\ln \frac{\mathrm{d}\sum_{i=1}^{n} w_i Q_i}{\mathrm{d}Q} - \ln \frac{\mathrm{d}w_1 Q_1}{\mathrm{d}Q}\right)_{n\in\mathbb{N}}$$

gives that

$$\begin{aligned}&\int_\Omega \ln \frac{\mathrm{d}\sum_{i=1}^{n} w_i Q_i}{\mathrm{d}Q} - \ln \frac{\mathrm{d}w_1 Q_1}{\mathrm{d}Q}\,\mathrm{d}P \\ &\qquad \to \int_\Omega \ln \frac{\mathrm{d}\sum_{i=1}^{\infty} w_i Q_i}{\mathrm{d}Q} - \ln \frac{\mathrm{d}w_1 Q_1}{\mathrm{d}Q}\,\mathrm{d}P.\end{aligned}$$

We get

$$\int_\Omega \ln \frac{\mathrm{d}\sum_{i=1}^{n} w_i Q_i}{\mathrm{d}Q}\,\mathrm{d}P \to \int_\Omega \ln \frac{\mathrm{d}\sum_{i=1}^{\infty} w_i Q_i}{\mathrm{d}Q}\,\mathrm{d}P$$

for $n \to \infty$. Finally, we see that

$$\begin{aligned}
D&\left(P \middle\| Q \rightsquigarrow \frac{\sum_{i=1}^n w_i \cdot Q_i}{\sum_{i=1}^n w_i}\right) \\
&= \int_\Omega \ln \frac{\mathrm{d}\sum_{i=1}^n w_i Q_i}{\mathrm{d}Q} \,\mathrm{d}P - (Q_n(\Omega) - Q(\Omega)) - \ln \sum_{i=1}^n w_i \\
&\to \int_\Omega \ln \frac{\mathrm{d}\sum_{i=1}^\infty w_i Q_i}{\mathrm{d}Q} \,\mathrm{d}P - (Q_\infty(\Omega) - Q(\Omega)) \\
&= D(P\|Q \rightsquigarrow Q_\infty),
\end{aligned}$$

where $Q_\infty := \sum_{i=1}^\infty w_i Q_i$ and we use that $\ln \sum_{i=1}^n w_i \to 0$ and $Q_n(\Omega) \to Q_\infty(\Omega)$. To see the latter, note that

$$Q_n(\Omega) = \int_\Omega \frac{\sum_{i=1}^n q_i(\omega) w_i}{\sum_{i=1}^n w_i} \,\mathrm{d}\mu(\omega),$$

and $0 \leq \sum_{i=1}^n q_i(\omega)w_i / \sum_{i=1}^n w_i \leq q_\infty(\omega)/w_1$, where the RHS integrates, so that the desired convergence follows from the dominated convergence theorem. □

Theorem A.8. *Let P be a finite measure and $\mathcal{C}$ a convex set of sub-probability measures such that $D(P\|Q \rightsquigarrow \mathcal{C}) = 0$. If $Q_1, Q_2, \dots$ is a sequence of measures in $\mathcal{C}$ such that $D(P\|Q_n \rightsquigarrow \mathcal{C}) \to 0$, then $D(P\|Q_n \rightsquigarrow \sigma\text{-conv}(\mathcal{C})) \to 0$.*

Proof. Fix $Q^* \in \mathcal{C}$ such that $D(P\|Q^* \rightsquigarrow \mathcal{C}) \leq \varepsilon$ and let $\bar{Q} = \sum_{i=1}^\infty w_i Q_i \in \sigma\text{-conv}(\mathcal{C})$ arbitrarily. Let $s \in (0,1)$ and consider $\tilde{Q} := s \cdot Q^* + (1-s) \cdot \bar{Q} = \sum_{i=0}^\infty \tilde{w}_i Q_i$, where $Q_0 := Q^*$, $\tilde{w}_0 = s$ and $\tilde{w}_i = (1-s) \cdot w_i$ for $i = 1, 2, \dots$. Note that $D(P\|Q^* \rightsquigarrow Q_0) = 0$, so it follows from Lemma A.7 that

$$\lim_{n\to\infty} D\left(P \middle\| Q^* \rightsquigarrow \frac{\sum_{i=0}^n \tilde{w}_i Q_i}{\sum_{i=0}^n \tilde{w}_i}\right) = D(P\|Q^* \rightsquigarrow \tilde{Q}).$$

The left hand side is, by definition of Q^*, bounded by ε since $\sum_{i=0}^n \tilde{w}_i Q_i / \sum_{i=0}^n \tilde{w}_i \in \mathcal{C}$, so that we find $D(P\|Q^* \rightsquigarrow \tilde{Q}) \leq \varepsilon$. Furthermore, by concavity of the log,

$$\begin{aligned}
\varepsilon &\geq D(P\|Q^* \rightsquigarrow \tilde{Q}) \\
&\geq s \cdot D(P\|Q^* \rightsquigarrow Q_0) + (1-s) \cdot D(P\|Q^* \rightsquigarrow \bar{Q}) \\
&= (1-s) \cdot D(P\|Q^* \rightsquigarrow \bar{Q}).
\end{aligned}$$

Taking the limit of $s \to 0$, we see $D(P\|Q^* \rightsquigarrow \bar{Q}) \leq \varepsilon$. Finally, the result follows by taking the supremum over $\bar{Q}$. □

We conjecture that if $\mathcal{C}$ is a σ-convex set of sub-probability measures and $\mathcal{C}'$ is the Choquet-convex hull of $\mathcal{C}$ then $D(P\|Q \rightsquigarrow \mathcal{C}) = D(P\|Q \rightsquigarrow \mathcal{C}')$ for any sub-probability measures P and Q such that P, Q, and the sub-probability measures in $\mathcal{C}$ all have densities with respect to a common σ-finite measure.

B | Appendix to Chapter 4

B.1 Application in Practice: k Separate I.I.D. Data Streams

In the simplest practical applications, we observe one block at a time, i.e. at time n, we have observed $\boldsymbol{X}_{(1)}, \ldots, \boldsymbol{X}_{(n)}$, where each $\boldsymbol{X}_{(i)} = (X_{i,1}, \ldots, X_{i,k})$ is a block, i.e. a vector with one outcome for each of the k groups. This is a rather restrictive setup, but we can easily extend it to blocks of data in which each group has a different number of outcomes. For example, if data comes in blocks with m_j outcomes in group j, for $j = 1 \ldots k$, $X_{(i)} = (X_{i,1,1}, \ldots, X_{i,1,m_1}, X_{i,2,1}, \ldots, X_{i,2,m_2}, \ldots, X_{i,k,1}, \ldots, X_{i,k,m_k})$, we can re-organize this having $k' = \sum_{j=1}^{k} m_j$ groups, having 1 outcome in each group, and having an alternative in which the first m_1 entries of the outcome vector share the same mean $\mu'_1 = \ldots = \mu'_{m_1} = \mu_1$; the next m_2 entries share the same mean $\mu'_{m_1+1} = \ldots = \mu'_{m_1+m_2} = \mu_2$, and so on.

Even more generally though, we will be confronted with k separate i.i.d streams and data in each stream may arrive at a different rate. We can still handle this case by pre-determining a multiplicity $m_1, \ldots, m_k$ for each stream. As data comes in, we fill virtual 'blocks' with m_j outcomes for group j, $j = 1 \ldots k$. Once a (number of) virtual block(s) has been filled entirely, the analysis can be performed as usual, restricted to the filled blocks. That is, if for some integer B we have observed Bm_j outcomes in stream j, for all $j = 1 \ldots k$, but for some j, we have not yet observed $(B+1)m_j$ outcomes, and we decide to stop the analysis and calculate the evidence against the null, then we output the product of e-variables for the first B blocks and ignore any additional data for the time being. Importantly, if we find out, while analyzing the streams, that some streams are providing data at a much faster rate than others, we may adapt $m_1, \ldots, m_k$ dynamically: whenever a virtual block has been finished, we may decide on alternative multiplicities for the next block; see Turner et al. (2024) for a detailed description for the case that $k = 2$.

B.2 Proofs for Section 4.2

In the proofs we freely use, without specific mention, basic facts about derivatives of (log-) densities of exponential families. These can all be found in, for example, Barndorff-Nielsen (1978).

B.2.1 Proof of Proposition 4.6

Proof. Since $S_{\text{GRO}(\mathcal{M})}$ was already shown to be an E-variable in Lemma 4.4, the 'if' part of the statement holds. The 'only-if' part follows directly from Corollary 2 to Theorem 1 in (Grünwald et al., 2024), which states that there can be at most one E-variable of the form $p_{\boldsymbol{\mu}}(X^k)/r(X^k)$ where r is a probability density for X^k. □

B.2.2 Proof of Proposition 4.7

Proof. Define $g(\mu_0) := \mathbb{E}_{p_{\langle \mu_0 \rangle}}\left[S_{\text{PSEUDO}(\mathcal{M})}\right]$ and $B(\mu_i) := A\left(\lambda(\mu_i) + \lambda(\mu_0) - \lambda(\mu_0^*)\right)$.

$$\begin{aligned}
g(\mu_0) &= \mathbb{E}_{p_{\langle \mu_0 \rangle}}\left[\prod_{i=1}^k \frac{p_{\mu_i}(X_i)}{p_{\mu_0^*}(X_i)}\right] = \prod_{i=1}^k \mathbb{E}_{Y \sim p_{\mu_0}}\left[\frac{p_{\mu_i}(Y)}{p_{\mu_0^*}(Y)}\right] \\
&= \prod_{i=1}^k \int \exp\left(\lambda(\mu_0)y - A\left(\lambda(\mu_0)\right)\right) \cdot \frac{\exp\left(\lambda(\mu_i)y - A\left(\lambda(\mu_i)\right)\right)}{\exp\left(\lambda(\mu_0^*)y - A\left(\lambda(\mu_0^*)\right)\right)} d\rho(y) \\
&= \prod_{i=1}^k \int \exp\left(\left(\lambda(\mu_i) + \lambda(\mu_0) - \lambda(\mu_0^*)\right)y - A\left(\lambda(\mu_i)\right) - A\left(\lambda(\mu_0)\right) + A\left(\lambda(\mu_0^*)\right)\right) d\rho(y) \\
&= \prod_{i=1}^k \exp\left(A\left(\lambda(\mu_0^*)\right) - A\left(\lambda(\mu_i)\right) - A\left(\lambda(\mu_0)\right)\right) \exp\left(B(\mu_i)\right) \\
&\qquad \cdot \int \exp\left(\left(\lambda(\mu_i) + \lambda(\mu_0) - \lambda(\mu_0^*)\right)y - B(\mu_i)\right) d\rho(y) \\
&= \prod_{i=1}^k \exp\left(A\left(\lambda(\mu_0^*)\right) - A\left(\lambda(\mu_i)\right) - A\left(\lambda(\mu_0)\right)\right) \exp\left(B(\mu_i)\right) \cdot 1 \\
&= \exp\left(kA\left(\lambda(\mu_0^*)\right) - \sum_{i=1}^k A\left(\lambda(\mu_i)\right) - kA\left(\lambda(\mu_0)\right) + \sum_{i=1}^k B(\mu_i)\right). \qquad \text{(B.1)}
\end{aligned}$$

Taking first and second derivatives with respect to μ_0, we find

$$\frac{d}{d\mu_0} g(\mu_0) = g(\mu_0) \cdot \frac{d}{d\mu_0} \left(\sum_{i=1}^{k} B(\mu_i) - kA\left(\lambda(\mu_0)\right) \right) \tag{B.2}$$

and

$$\begin{aligned} \frac{d^2}{d\mu_0^2} g(\mu_0) = & \left(\frac{d}{d\mu_0} g(\mu_0) \right) \cdot \frac{d}{d\mu_0} \left(\sum_{i=1}^{k} B(\mu_i) - kA\left(\lambda(\mu_0)\right) \right) \\ & + g(\mu_0) \cdot \frac{d^2}{d\mu_0^2} \left(\sum_{i=1}^{k} B(\mu_i) - kA\left(\lambda(\mu_0)\right) \right) \\ = & g(\mu_0) \left(\sum_{i=1}^{k} (\mu_i + \mu_0 - \mu_0^*) - k\mu_0 \right)^2 \\ & + g(\mu_0) \left(\sum_{i=1}^{k} \text{VAR}_{P_{\mu_i + \mu_0 - \mu_0^*}}[X] - k\text{VAR}_{P_{\mu_0}}[X] \right) \\ = & g(\mu_0) \left(\sum_{i=1}^{k} \text{VAR}_{P_{\mu_i + \mu_0 - \mu_0^*}}[X] - k\text{VAR}_{P_{\mu_0}}[X] \right) = g(\mu_0) \cdot f(\mu_0). \end{aligned} \tag{B.3}$$

where the second equality holds because of (B.2), $(d/d\lambda(\mu))A(\lambda(\mu)) = \mathbb{E}_{P_\mu}[X]$ and $(d^2/d\lambda(\mu)^2)A(\lambda(\mu)) = \text{VAR}_{P_\mu}[X]$. (B.3) is continuous with respect to μ_0. Therefore, if $f(\mu_0^*) > 0$ holds, it means that there exists an interval $\mathtt{M}^* \subset \mathtt{M}$ with μ_0^* in the interior of $\mathtt{M}^*$ on which (B.1) is strictly convex. Then there must exist a point $\mu_0' \in \mathtt{M}^*$ satisfying $\mathbb{E}_{P_{\langle \mu_0' \rangle}}\left[S_{\text{PSEUDO}(\mathcal{M})}\right] > \mathbb{E}_{P_{\langle \mu_0^* \rangle}}\left[S_{\text{PSEUDO}(\mathcal{M})}\right] = 1$, i.e. $S_{\text{PSEUDO}(\mathcal{M})}$ is not an E-variable. Conversely, $f(\mu_0^*) < 0$ means that there exists an interval $\mathtt{M}^* \subset \mathtt{M}$ with μ_0^* in the interior of $\mathtt{M}^*$, on which (B.1) is strictly concave. The result follows. □

B.2.3 Proof of Theorem 4.8

To prepare for the proof of Theorem 4.8, let us first recall Young's [1912] inequality:

Lemma B.1. [Young's inequality] *Let p, q be positive real numbers satisfying $\frac{1}{p} + \frac{1}{q} = 1$. Then if a, b are nonnegative real numbers, $ab \leq \frac{a^p}{p} + \frac{b^q}{q}$.*

The proof of Theorem 4.8 follows exactly the same argument as the one used by Turner et al. (2024) to prove this statement in the special case that $\mathcal{M}$ is the Bernoulli model.

Proof. We first show that $S_{\text{GRO(IID)}}$ as defined in the theorem statement is an E-variable.

For this, we set $p_0^*(X) = \frac{1}{k}\sum\limits_{i=1}^{k} p_{\mu_i}(X)$. We have:

$$\mathbb{E}_{X^k \sim P_{\langle \mu_0 \rangle}}\left[S_{\text{GRO(IID)}}\right] = \mathbb{E}_{X_1 \sim P_{\mu_0}}\left[\frac{p_{\mu_1}(X_1)}{p_0^*(X_1)}\right] \cdot \ldots \cdot \mathbb{E}_{X_k \sim P_{\mu_0}}\left[\frac{p_{\mu_k}(X_k)}{p_0^*(X_k)}\right]. \tag{B.4}$$

We also have

$$\begin{aligned} &\frac{1}{k}\mathbb{E}_{X_1 \sim P_{\mu_0}}\left[\frac{p_{\mu_1}(X_1)}{p_0^*(X_1)}\right] + \cdots + \frac{1}{k}\mathbb{E}_{X_k \sim P_{\mu_0}}\left[\frac{p_{\mu_k}(X_k)}{p_0^*(X_k)}\right] \\ =&\frac{1}{k}\mathbb{E}_{X \sim P_{\mu_0}}\left[\frac{p_{\mu_1}(X)}{\frac{1}{k}\sum\limits_{i=1}^{k} p_{\mu_i}(X)} + \cdots + \frac{p_{\mu_k}(X)}{\frac{1}{k}\sum\limits_{i=1}^{k} p_{\mu_i}(X)}\right] = 1. \end{aligned} \tag{B.5}$$

We need to show that (B.4) ≤ 1, for which we can use (B.5). Stated more simply, it is sufficient to prove $\prod\limits_{i=1}^{k} r_i \leq 1$ with $\frac{1}{k}\sum\limits_{i=1}^{k} r_i \leq 1$, $r_i \in \mathbb{R}^+$. But this is easily established:

$$\begin{aligned} \frac{1}{k}\sum_{i=1}^{k} r_i = \frac{k-1}{k} \cdot \frac{\sum_{i=1}^{k-1} r_i}{k-1} + \frac{r_k}{k} &\geq \left(\frac{\sum_{i=1}^{k-1} r_i}{k-1}\right)^{\frac{k-1}{k}} r_k^{\frac{1}{k}} \\ &= \left(\frac{k-2}{k-1} \cdot \frac{\sum_{i=1}^{k-2} r_i}{k-2} + \frac{r_{k-1}}{k-1}\right)^{\frac{k-1}{k}} r_k^{\frac{1}{k}} \\ &\geq \left(\frac{\sum_{i=1}^{k-2} r_i}{k-2}\right)^{\frac{k-2}{k}} r_{k-1}^{\frac{1}{k}} r_k^{\frac{1}{k}} \\ &\vdots \\ &\geq \left(\frac{r_1 + r_2}{2}\right)^{\frac{2}{k}} \prod_{i=3}^{k} r_i^{\frac{1}{k}} \geq \prod_{i=1}^{k} r_i^{\frac{1}{k}} \end{aligned} \tag{B.6}$$

where the first inequality holds because of Young's inequality, by setting $\frac{1}{p} := \frac{k-1}{k}$, $\frac{1}{q} := \frac{1}{k}$, $a^p := \frac{\sum_{i=1}^{k-1} r_i}{k-1}$, $b^q := r_k$ in Lemma B.1. The other inequalities are established in the same way. It follows that $\prod\limits_{i=1}^{k} r_i^{\frac{1}{k}} \leq 1$ and further $\prod\limits_{i=1}^{k} r_i \leq 1$.

This shows that $S_{\text{GRO(IID)}}$ is a e-variable. It remains to show that $S_{\text{GRO(IID)}}$ is indeed the GRO e-variable relative to $\mathcal{H}_0(\text{IID})$; once we have shown this, it follows by Lemma 2 that it is the unique such e-variable and therefore by Lemma 1 that P_0^* achieves the minimum in Lemma 1. Since we already know that $S_{\text{GRO(IID)}}$ is an e-variable, the fact that it is the GRO e-variable relative to $\mathcal{H}_0(\textsc{iid})$ follows immediately from Corollary 2 of Theorem 1 in Grünwald et al. (2024), which states that there can be at most one e-variable of form $p_{\boldsymbol{\mu}}(X^k)/r(X^k)$ where r is a probability density. Since $S_{\text{GRO(IID)}}$ is such an e-variable, Lemma 1 gives that it must be the GRO e-variable. □

B.2.4 Proof of Proposition 4.11

Proof. The observed values of $X_1, X_2, \ldots, X_k$ are denoted as x^k $(:= x_1, \ldots, x_k)$. With $X_k(x^{k-1}, z) := z - \sum_{i=1}^{k-1} x_i$ and $\mathcal{C}(z)$ as in (4.12) and $p_{\boldsymbol{\mu};[Z]}(z)$ and $\rho(x^{k-1})$ as in (4.11), we get:

$$
\begin{aligned}
p_{\boldsymbol{\mu}}\left(x^{k-1} \middle| Z = z\right) &= \frac{p_{\boldsymbol{\mu}}\left(x^k\right)}{p_{\boldsymbol{\mu};[Z]}\left(z\right)} \\
&= \frac{\exp\left(\sum\limits_{i=1}^{k} (\lambda(\mu_i)x_i - A(\lambda(\mu_i)))\right)}{\int_{\mathcal{C}(z)} \exp\left(\sum\limits_{i=1}^{k-1} (\lambda(\mu_i)y_i - A(\lambda(\mu_i)) + \lambda(\mu_k)X_k(y^{k-1}, z)) - A(\lambda(\mu_k)))\right) d\rho(y^{k-1})} \\
&= \frac{\exp\left(\lambda(\mu_k)z + \sum\limits_{i=1}^{k-1} (\lambda(\mu_i) - \lambda(\mu_k))x_i\right)}{\int_{\mathcal{C}(z)} \exp\left(\lambda(\mu_k)z + \sum\limits_{i=1}^{k-1} (\lambda(\mu_i) - \lambda(\mu_k))y_i\right) d\rho(y^{k-1})} \\
&= \frac{\exp\left(\sum\limits_{i=1}^{k-1} (\lambda(\mu_i) - \lambda(\mu_k))x_i\right)}{\int_{\mathcal{C}(z)} \exp\left(\sum\limits_{i=1}^{k-1} (\lambda(\mu_i) - \lambda(\mu_k))y_i\right) d\rho(y^{k-1})}.
\end{aligned}
$$

□

B.3 Proofs for Section 4.3

B.3.1 Proof of Theorem 4.12

Proof. We prove the theorem using an elaborate Taylor expansion of $F(\delta)$, defined below, around $\delta = 0$. We first calculate the first four derivatives of $F(\delta)$. Thus we define and derive, with $\mu_i = \mu_0 + \alpha_i\delta$ and $f_y(\delta) = \sum\limits_{i=1}^{k} p_{\mu_i}(y)$ defined as in the theorem statement,

$$
\begin{aligned}
F(\delta) :=& \mathbb{E}_{P_{\langle \mu_0 \rangle + \boldsymbol{\alpha}\delta}} \left[\log S_{\text{PSEUDO}(\mathcal{M})} - \log S_{\text{GRO(IID)}}\right] \\
=& \mathbb{E}_{P_{\boldsymbol{\mu}}} \left[\log \prod_{j=1}^{k} \left(\frac{1}{k} \sum_{i=1}^{k} p_{\mu_i}(X_j) \right) - \log p_{\langle \mu_0 \rangle}(X^k) \right] \\
=& \mathbb{E}_{P_{\boldsymbol{\mu}}} \left[\sum_{j=1}^{k} \log f_{X_j}(\delta) - \sum_{j=1}^{k} \log p_{\mu_0}(X_j) \right] - k \log k \\
\overset{(a)}{=}& \sum_{j=1}^{k} \mathbb{E}_{X \sim P_{\mu_j}} \left[\log f_X(\delta) - \log p_{\mu_0}(X) \right] - k \log k \\
\overset{(b)}{=}& \overbrace{\int_{y \in \mathcal{X}} f_y(\delta) \log f_y(\delta) d\rho(y)}^{F_1(\delta)} + \overbrace{\left(- \int_{y \in \mathcal{X}} f_y(\delta) \log p_{\mu_0}(y) d\rho(y) \right)}^{F_2(\delta)} - k \log k, \quad \text{(B.7)}
\end{aligned}
$$

where we define $F_1(\delta)$ to be equal to the leftmost term in (B.7) and $F_2(\delta)$ to be equal to the second, and (a) and (b) both hold provided that

$$
\text{for all } j \in \{1, \ldots, k\}\text{: } \mathbb{E}_{X_j \sim P_{\mu_j}} \left[|\log f_{X_j}(\delta) - \log p_{\mu_0}(X_j)\ | \right] < \infty \qquad \text{(B.8)}
$$

is finite. In Appendix B.6 we verify that this condition, as well as a plethora of related finiteness-of-expectation-of-absolute-value conditions hold for all δ sufficiently close to 0. Together these not just imply (a) and (b), but also (c) that we can freely exchange integration over y and differentiation over δ for all such δ when computing the first k derivatives of $F_1(\delta)$ and $F_2(\delta)$, for any finite k and (d) that all these derivatives are finite for δ in a compact interval including 0 (since the details are straightforward but quite tedious and long-winded we deferred these to Appendix B.6). Thus, using (c), we will freely differentiate under the integral sign in the remainder of the proof below, and using (d), we will be able to conclude that the final result is finite.

For each derivative, we first compute the derivative of $F_1(\delta)$ and then that of $F_2(\delta)$.

$$
\begin{aligned}
F_1'(\delta) &= \int f_y'(\delta) d\rho(y) + \int f_y'(\delta) \log f_y(\delta) d\rho(y) = 0, \\
F_2'(\delta) &= - \int f_y'(\delta) \log p_{\mu_0}(y) d\rho(y) = 0, \text{ so } F'(0) = F_1'(0) + F_2'(0) = 0, \qquad \text{(B.9)}
\end{aligned}
$$

where the above formulas hold since $f_x'(0) = 0$ for all $x \in \mathcal{X}$, which can be obtained by

$$\begin{aligned}
f'_x(\delta^\circ) =& \sum_{j=1}^{k} \frac{dp_{\mu_j}(x)}{d\mu_j}\frac{d\mu_j}{d\delta}(\delta^\circ), \\
f'_x(0) =& \frac{dp_{\mu_0}(x)}{d\mu_0}\sum_{j=1}^{k}\frac{d\mu_j}{d\delta}(0) = \frac{dp_{\mu_0}(x)}{d\mu_0}\sum_{j=1}^{k}\alpha_j = 0, \qquad \text{(B.10)}
\end{aligned}$$

where we used that all μ_j are equal to μ_0 at $\delta = 0$. We turn to the second derivatives:

$$\begin{aligned}
F''_1(\delta) &= \int f''_y(\delta)d\rho(y) + \int \left(f''_y(\delta)\log f_y(\delta) + \frac{\left(f'_y(\delta)\right)^2}{f_y(\delta)} \right) d\rho(y) \\
&= \int \left(f''_y(\delta)\log f_y(\delta) + \frac{\left(f'_y(\delta)\right)^2}{f_y(\delta)} \right) d\rho(y) \\
F''_1(0) &= \int \left(f''_y(0)\log f_y(0) + \frac{\left(f'_y(0)\right)^2}{f_y(0)} \right) d\rho(y); \\
&= \int f''_y(0)\log p_{\mu_0}(y)d\rho(y) + \int_{y\in\mathcal{X}} \left(f''_y(0)\log k\right) d\rho(y) \qquad \text{(B.11)} \\
&= \int \left(f''_y(0)\log p_{\mu_0}(y)\right) d\rho(y),
\end{aligned}$$

where $\int f''_y(\delta)d\rho(y) = 0$ because $\int f_y(\delta)d\rho(y) = k$, in which k is a constant that does not depend on δ. Then $F''_2(\delta)$ is given by

$$\begin{aligned}
F''_2(\delta) = &-\int f''_y(\delta)\log p_{\mu_0}(y)d\rho(y) \ ; \ F''_2(0) = -\int f''_y(0)\log p_{\mu_0}(y)d\rho(y), \text{ so} \\
F''(0) =& F''_1(0) + F''_2(0) = 0. \qquad \text{(B.12)}
\end{aligned}$$

Now we compute the third derivative of $F(\delta)$, denoted as $F^{(3)}(\delta)$.

$$\begin{aligned}
F_1^{(3)}(\delta) &= \int \left(f_y^{(3)}(\delta)\log f_y(\delta) + \frac{f''_y(\delta)f'_y(\delta)}{f_y(\delta)} + \frac{2f''_y(\delta)f'_y(\delta)f_y(\delta) - (f'_y(\delta))^3}{(f_y(\delta))^2} \right) d\rho(y) \\
F_1^{(3)}(0) &= \int f_y^{(3)}(0)\log f_y(0)d\rho(y) = \int f_y^{(3)}(0)\log p_{\mu_0}(y)d\rho(y) + \int f_y^{(3)}(0)\log k d\rho(y) \\
&= \int f_y^{(3)}(0)\log p_{\mu_0}(y)d\rho(y) \\
F_2^{(3)}(\delta) &= -\int f_y^{(3)}(\delta)\log p_{\mu_0}(y)d\rho(y)
\end{aligned}$$

$$F_2^{(3)}(0) = -\int f_y^{(3)}(0) \log p_{\mu_0}(y) d\rho(y), \text{ so } F^{(3)}(0) = F_1^{(3)}(0) + F_2^{(3)}(0) = 0,$$

which holds since $f_y'(0) = 0$ and $\int f_y(0) d\rho(y) = k$.

The fourth derivative of $F(\delta)$ can be computed as follows:

$$\begin{aligned} F_1^{(4)}(\delta) =& \int \left(f_y^{(4)}(\delta) \log f_y(\delta) + \frac{f_y^{(3)}(\delta) f_y'(\delta)}{f_y(\delta)} \right) d\rho(y) \\ &+ \int 3 \cdot \frac{\left(f_y^{(3)}(\delta) f_y'(\delta) + (f_y''(\delta))^2 \right) f_y(\delta) - f_y''(\delta) \left(f_y'(\delta) \right)^2}{\left(f_y(\delta) \right)^2} d\rho(y) \\ &- \int \frac{3 \left(f_y(\delta) f_y'(\delta) \right)^2 \cdot f_y''(\delta) - 2 \left(f_y'(\delta) \right)^4 \cdot f_y(\delta)}{\left(f_y(\delta) \right)^4} d\rho(y) \ ; \end{aligned} \tag{B.13}$$

$$\begin{aligned} F_1^{(4)}(0) =& \int \left(f_y^{(4)}(0) \log f_y(0) + \frac{3 \left(f_y''(0) \right)^2}{f_y(0)} \right) d\rho(y) \\ =& \int f_y^{(4)}(0) \log p_{\mu_0}(y) d\rho(y) + \log k \int_{y \in \mathcal{X}} f_y^{(4)}(0) d\rho(y) + \int_{y \in \mathcal{X}} \frac{3 \left(f_y''(0) \right)^2}{f_y(0)} d\rho(y) \\ =& \int f_y^{(4)}(0) \log p_{\mu_0}(y) d\rho(y) + \int_{y \in \mathcal{X}} \frac{3 \left(f_y''(0) \right)^2}{f_y(0)} d\rho(y), \end{aligned}$$

and $F_2^{(4)}(\delta)$ can be computed by

$$\begin{aligned} &F_2^{(4)}(\delta) = -\int f_y^{(4)}(\delta) \log p_{\mu_0}(y) d\rho(y), \ F_2^{(4)}(0) = -\int f_y^{(4)}(0) \log p_{\mu_0}(y) d\rho(y), \text{ so} \\ &F^{(4)}(0) = F_1^{(4)}(0) + F_2^{(4)}(0) = \int \frac{3 \left(f_y''(0) \right)^2}{f_y(0)} d\rho(y) > 0. \end{aligned}$$

Based on the above derivatives, we can now do a fourth-order Taylor expansion of $F(\delta)$ around $\delta = 0$, which gives:

$$\begin{aligned} \mathbb{E}_{P_\mu} \left[\log S_{\text{PSEUDO}(\mathcal{M})} - \log S_{\text{GRO(IID)}} \right] =& \frac{1}{4!} F^{(4)}(0) \delta^4 + o(\delta^4) \\ =& \frac{1}{8} \int_{y \in \mathcal{X}} \frac{\left(f_y''(0) \right)^2}{f_y(0)} d\rho(y) \cdot \delta^4 + o\left(\delta^4 \right), \end{aligned}$$

where $f_y(0) = \sum_{i=1}^{k} p_{\mu_0}(y) = kp_{\mu_0}(y)$ and $f_y''(0) = \left(\sum_{i=1}^{k} \alpha_i^2\right) \cdot \frac{d^2}{d\mu^2} p_\mu(y) \mid_{\mu=\mu_0} = \frac{d^2}{d\mu^2} p_\mu(y) \mid_{\mu=\mu_0}$. □

B.3.2 Proof of Theorem 4.13

Proof. We obtain the result using an even more involved Taylor expansion than in the previous theorem. As in that theorem, we will freely differentiate (with respect to δ) under the integral sign — that this is allowed is again verified in Appendix B.6.

Let $\boldsymbol{\mu}$, $\boldsymbol{\alpha}, \mathcal{C}(z), \rho(x^{k-1}), P_{\boldsymbol{\mu}}$ etc. be as in the theorem statement. We have:

$$\begin{aligned} f(\delta) :=& \mathbb{E}_{P_{\boldsymbol{\mu}}} \left[\log S_{\text{PSEUDO}(\mathcal{M})} - \log S_{\text{COND}}\right] \\ =& \mathbb{E}_{P_{\boldsymbol{\mu}}} \left[\log \frac{p_{\boldsymbol{\mu}}\left(X^k\right)}{p_{\langle\mu_0\rangle}\left(X^k\right)} - \log \frac{p_{\boldsymbol{\mu}}\left(X^{k-1} \mid Z\right)}{p_{\langle\mu_0\rangle}\left(X^{k-1} \mid Z\right)}\right] \\ =& \mathbb{E}_{P_{\boldsymbol{\mu}}} \left[\log \frac{p_{\boldsymbol{\mu}}\left(X^k\right)}{p_{\langle\mu_0\rangle}\left(X^k\right)} - \log \frac{p_{\boldsymbol{\mu}}\left(X^k\right)}{p_{\langle\mu_0\rangle}\left(X^k\right)} + \log \frac{\int_{\mathcal{C}(z)} p_{\boldsymbol{\mu}}\left(x^k\right) d\rho(x^{k-1})}{\int_{\mathcal{C}(z)} p_{\langle\mu_0\rangle}\left(x^k\right) d\rho(x^{k-1})}\right] \\ =& D\left(P_{\langle\mu_0\rangle+\boldsymbol{\alpha}\delta;[Z]} \| P_{\langle\mu_0\rangle;[Z]}\right). \end{aligned}$$

We will prove the result by doing a Taylor expansion for $f(\delta)$ around $\delta = 0$. It is obvious that $f(0) = 0$ and the first derivative $f'(0) = 0$ since $f(0)$ is the minimum of $f(\delta)$ over an open set, and $f(\delta)$ is differentiable. We proceed to compute the second derivative of $f(\delta)$, using the notation $g_z(\delta) = p_{\langle\mu_0\rangle+\boldsymbol{\alpha}\delta;[Z]}(z)$ as in the theorem statement, with g_z' and g_z'' denoting first and second derivatives.

$$\begin{aligned} f'(\delta) &= \int g_z'(\delta) \log \frac{g_z(\delta)}{g_z(0)} d\rho_{[Z]}(z) + \int g_z'(\delta) d\rho_{[Z]}(z) = \int g_z'(\delta) \log \frac{g_z(\delta)}{g_z(0)} d\rho_{[Z]}(z). \\ f''(\delta) &= \int g_z''(\delta) \log \frac{g_z(\delta)}{g_z(0)} d\rho_{[Z]}(z) + \int \frac{\left(g_z'(\delta)\right)^2}{g_z(\delta)} d\rho_{[Z]}(z), \end{aligned}$$

where in the first line, the second equality follows since the second term does not change if we interchanging differentiation and integration and the fact that $\int g_z(\delta) dz = 1$ is constant in δ. We obtain

$$f''(0) = \int \frac{\left(g_z'(0)\right)^2}{g_z(0)} d\rho_{[Z]}(z), \tag{B.14}$$

and, with x_k set to $X_k(x^{k-1}, z)$ and recalling that $\boldsymbol{\mu} = \langle \mu_0 \rangle + \boldsymbol{\alpha}\delta$ and $\mu_j = \mu_0 + \alpha_j \delta$,

$$\begin{aligned}
g'_z(\delta) &= \int_{\mathcal{C}(z)} \frac{d}{d\delta} p_{\langle \mu_0 \rangle + \boldsymbol{\alpha}\delta}(x^k) d\rho(x^{k-1}) \\
&= \int_{\mathcal{C}(z)} \sum_{j=1}^{k} \prod_{i \in \{1,\ldots,k\} \setminus j} p_{\mu_i}(x_i) \frac{dp_{\mu_j}(x_j)}{d\delta} d\rho(x^{k-1}) \\
&= \int_{\mathcal{C}(z)} \sum_{j=1}^{k} p_{\mu_1,\ldots,\mu_{j-1},\mu_{j+1},\ldots,\mu_k}(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k) \frac{dp_{\mu_j}(x_j)}{d\mu_j} \frac{d\mu_j}{d\delta} d\rho(x^{k-1}) \\
&= \int_{\mathcal{C}(z)} \sum_{j=1}^{k} p_{\boldsymbol{\mu}}(x^k) \frac{d \log p_{\mu_j}(x_j)}{d\mu_j} \alpha_j d\rho(x^{k-1}) \\
&= \int_{\mathcal{C}(z)} \sum_{j=1}^{k} p_{\boldsymbol{\mu}}(x^k) \left(I(\mu_j) x_j - \mu_j I(\mu_j) \right) \alpha_j d\rho(x^{k-1})
\end{aligned}$$

where $I(\mu_j)$ is the Fisher information. The final equality follows because, with $\lambda(\mu_j)$ denoting the canonical parameter corresponding to μ_j, we have $d\lambda(\mu_j)/d\mu_j = I(\mu_j)$ and $dA(\beta)/d\beta) \mid_{\beta=\lambda(\mu_j)} = \mu_j$; see e.g. (Grünwald, 2007, Chapter 18). Now

$$\begin{aligned}
g'_z(0) &= \int_{\mathcal{C}(z)} \sum_{j=1}^{k} p_{\langle \mu_0 \rangle}(x^k) \left(I(\mu_0) x_j - \mu_0 I(\mu_0) \right) \alpha_j d\rho(x^{k-1}) \\
&= \int_{\mathcal{C}(z)} p_{\langle \mu_0 \rangle}(x^k) I(\mu_0) \sum_{j=1}^{k} x_j \alpha_j d\rho(x^{k-1}) && \text{(B.15)} \\
&= I(\mu_0) \cdot \int_{\mathcal{C}(z)} p_{\langle \mu_0 \rangle}(x^k) \sum_{j=1}^{k} x_j \alpha_j d\rho(x^{k-1}) && \text{(B.16)}
\end{aligned}$$

where the second equality follows from $\sum_{j=1}^{k} \alpha_j = 0$. Because X^k i.i.d. $\sim P_{\mu_0}$ under $P_{\langle \mu_0 \rangle}$ and the integral in (B.15) is over a set of exchangeable sequences, (For understanding the statement, we can consider the simple case $k = 2$, X_1 and X_2 can be exchangeable because they are 'symmetric' for given $\mathcal{C}(z)$.) we must have that (B.15) remains valid if we re-order the α_j's in round-robin fashion, i.e. for all $i = 1..k$, we have, with $\alpha_{j,i} = \alpha_{(j+i-1) \mod k}$,

$$g'_z(0) = I(\mu_0) \cdot \int_{\mathcal{C}(z)} p_{\langle \mu_0 \rangle}(x^k) \sum_{j=1}^{k} x_j \alpha_{j,i} d\rho(x^{k-1}).$$

Summing these k equations we get, using that $\sum_{i=1}^{k} \alpha_i = 0$, that $kg'_z(0) = 0$ so that $g'_z(0) = 0$. From (B.14) we now see that

$$f''(0) = 0.$$

Now we compute the third derivative of $f(\delta)$, denoted as $f^{(3)}(\delta)$:

$$\begin{aligned} f^{(3)}(\delta) = &\int \left(g_z^{(3)}(\delta) \log \frac{g_z(\delta)}{g_z(0)} + \frac{g''_z(\delta) g'_z(\delta)}{g_z(\delta)} \right) d\rho_{[Z]}(z) \\ &+ \int \left(\frac{2g''_z(\delta) g'_z(\delta) g_z(\delta) - (g'_z(\delta))^3}{(g_z(\delta))^2} \right) d\rho_{[Z]}(z) \end{aligned}$$

So since $g'_z(0) = 0$ we must also have

$$f^{(3)}(0) = 0.$$

The fourth derivative of $f(\delta)$ is now computed as follows:

$$\begin{aligned} f^{(4)}(\delta) = &\int \left(g_z^{(4)}(\delta) \log \frac{g_z(\delta)}{g_z(0)} + \frac{g_z^{(3)}(\delta) \cdot g'_z(\delta)}{g_z(\delta)} \right) d\rho_{[Z]}(z) \\ &+ \int 3 \cdot \frac{\left(g_z^{(3)}(\delta) \cdot g'_z(\delta) + (g''_z(\delta))^2 \right) g_z(\delta) - g''_z(\delta) \cdot (g'_z(\delta))^2}{(g_z(\delta))^2} d\rho_{[Z]}(z). \end{aligned}$$

Then

$$f^{(4)}(0) = \int \frac{3 \left(g''_z(0) \right)^2}{g_z(0)} d\rho_{[Z]}(z) > 0.$$

We now have all ingredients for a fourth-order Taylor expansion of $f(\delta)$ around $\delta = 0$, which gives:

$$\mathbb{E}_{P_\mu} \left[\log S_{\text{PSEUDO}(\mathcal{M})} - \log S_{\text{COND}} \right] = \frac{1}{8} \int \frac{\left(g''_z(0) \right)^2}{g_z(0)} d\rho_{[Z]}(z) \cdot \delta^4 + o\left(\delta^4 \right)$$

which is what we had to prove. □

B.4 Proofs for Section 4.4

In this section, we prove all the statements in Table 4.1.

B.4.1 Bernoulli Family

We prove that for $\mathcal{M}$ equal to the Bernoulli family, we have $S_{\text{PSEUDO}(\mathcal{M})} = S_{\text{GRO}(\mathcal{M})} = S_{\text{GRO(IID)}} \succ S_{\text{COND}}$.

Proof. We set $\mu_0^* = \frac{1}{k} \sum_{i=1}^{k} \mu_i$.

$$S_{\text{GRO(IID)}} := \frac{p_{\boldsymbol{\mu}}(X^k)}{\prod_{j=1}^{k} \left(\frac{1}{k} \sum_{i=1}^{k} p_{\mu_i}(X_j) \right)} = \frac{p_{\boldsymbol{\mu}}(X^k)}{\prod_{j=1}^{k} \left(\frac{1}{k} \sum_{i=1}^{k} \left(\mu_i^{X_j} (1-\mu_i)^{1-X_j} \right) \right)} \tag{B.17}$$

$$= \frac{p_{\boldsymbol{\mu}}(X^k)}{\prod_{j=1}^{k} \left((\mu_0^*)^{X_j} (1-\mu_0^*)^{1-X_j} \right)}$$

$$= \frac{p_{\boldsymbol{\mu}}(X^k)}{\prod_{j=1}^{k} p_{\mu_0^*}(X_j)} = S_{\text{PSEUDO}(\mathcal{M})} \tag{B.18}$$

where the third equality holds since $X_i \in \{0,1\}$. So $S_{\text{PSEUDO}(\mathcal{M})}$ is an E-variable and $S_{\text{PSEUDO}(\mathcal{M})} = S_{\text{GRO}(\mathcal{M})}$ according to Theorem 4.6. Then the claim follows using (4.9) together with the fact that when $Z = 0$ or $Z = 2$, we have $S_{\text{COND}} = 1$, while this is not true for the other *e*-variables, so that $S_{\text{COND}} \neq S_{\text{GRO}(\mathcal{M})} = S_{\text{PSEUDO}(\mathcal{M})} = S_{\text{GRO(IID)}}$. The result then follows from (4.9). □

B.4.2 Poisson and Gaussian Family With Free Mean and Fixed Variance

We prove that for $\mathcal{M}$ equal to the family of Gaussian distributions with free mean and fixed variance σ^2, we have $S_{\text{PSEUDO}(\mathcal{M})} = S_{\text{GRO}(\mathcal{M})} = S_{\text{COND}} \succ S_{\text{GRO(IID)}}$. The proof that the same holds for $\mathcal{M}$ equal to the family of Poisson distributions is omitted, as it is completely analogous.

Proof. Note that if we let $Z := \sum_{i=1}^{k} X_i$, then we have that $Z \sim \mathcal{N}(\sum_{i=1}^{k} \mu_i, k\sigma^2)$ if $X^k \sim P_{\boldsymbol{\mu}}$. Let μ_0^* be given by (4.8) relative to fixed alternative $P_{\boldsymbol{\mu}}$ as in the definition of $S_{\text{PSEUDO}(\mathcal{M})}$ underneath (4.8). Since $k\mu_0^* = \sum_{i=1}^{k} \mu_i$, we have that Z has the same distribution for $X^k \sim P_{\langle \mu_0^* \rangle}$. This can be used to write

$$S_{\text{COND}} = \frac{p_{\boldsymbol{\mu}}\left(X^k \mid Z\right)}{p_{\langle \mu_0^* \rangle}\left(X^k \mid Z\right)} = \frac{p_{\boldsymbol{\mu}}\left(X^k\right)}{p_{\langle \mu_0^* \rangle}\left(X^k\right)} \frac{p_{\langle \mu_0^* \rangle}(Z)}{p_{\boldsymbol{\mu}}(Z)} = \frac{p_{\boldsymbol{\mu}}\left(X^k\right)}{p_{\langle \mu_0^* \rangle}\left(X^k\right)} = S_{\text{PSEUDO}(\mathcal{M})}.$$

Therefore, $S_{\text{PSEUDO}(\mathcal{M})}$ is also an e-variable, so we derive that $S_{\text{PSEUDO}(\mathcal{M})} = S_{\text{GRO}(\mathcal{M})}$ by Theorem 4.6. Furthermore, we have that the denominator of $S_{\text{GRO(IID)}}$ is given by a different distribution than $p_{\langle \mu_0^* \rangle}$, so that $S_{\text{GRO(IID)}} \neq S_{\text{GRO}(\mathcal{M})} = S_{\text{PSEUDO}(\mathcal{M})} = S_{\text{COND}}$. The result then follows from (4.9).

□

B.4.3 The Families for Which $S_{\text{pseudo}(\mathcal{M})}$ Is Not an E-variable

Here, we prove that $S_{\text{PSEUDO}(\mathcal{M})}$ is not an e-variable for $\mathcal{M}$ equal to the family of beta distributions with free β and fixed α. It then follows from (4.9) that $S_{\text{PSEUDO}(\mathcal{M})} \succ S_{\text{GRO}(\mathcal{M})}$. (4.9) also gives $S_{\text{GRO}(\mathcal{M})} \succeq S_{\text{GRO(IID)}}$ and $S_{\text{GRO}(\mathcal{M})} \succeq S_{\text{COND}}$. The same is true for $\mathcal{M}$ equal to the family of geometric distributions and the family of Gaussian distributions with free variance and fixed mean, as the proof that $S_{\text{PSEUDO}(\mathcal{M})}$ is not an e-variable is entirely analogous to the proof for the beta distributions given below. In all of these cases, one easily shows by simulation that in general, $S_{\text{GRO}(\mathcal{M})} \neq S_{\text{GRO(IID)}}$ and $S_{\text{GRO}(\mathcal{M})} \neq S_{\text{COND}}$, so then $S_{\text{GRO}(\mathcal{M})} \succ S_{\text{GRO(IID)}}$ and $S_{\text{GRO}(\mathcal{M})} \succ S_{\text{COND}}$ follow.

Proof. First, let $Q_{\alpha,\beta}$ represent a beta distribution in its standard parameterization, so that its density is given by

$$q_{\alpha,\beta}(u) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} u^{\alpha-1}(1-u)^{\beta-1}, \qquad \alpha,\beta > 0; u \in [0,1].$$

To simplify the proof, we assume $\alpha = 1$ here. Then

$$q_{1,\beta}(u) = \frac{\Gamma(1+\beta)}{\Gamma(\beta)}(1-u)^{\beta-1} = \frac{1}{1-u}\exp\left(\beta \log(1-u) - \log\frac{1}{\beta}\right)$$

where the first equality holds since $\Gamma(1+\beta) = \beta\Gamma(\beta)$. Comparing this to (4.1), we see that β is the canonical parameter corresponding to the family $\{Q_{1,\beta} : \beta > 0\}$, and we have

$$\lambda(\mu) = \beta, \quad t(u) = \log(1-u), \quad A(\beta) = \log\frac{1}{\beta}.$$

To prove the statement, according to Proposition 4.7, we just need to show, for any $\mu_1, \ldots, \mu_k$ that are not all equal to each other, that, with $X = t(U) = \log(1-U)$ and $\mu_0^* = \frac{1}{k}\sum_{i=1}^{k} \mu_i$ defined as in (4.8), we have

$$\sum_{i=1}^{k} \text{VAR}_{P_{\mu_i}}[X] - k\text{VAR}_{P_{\mu_0^*}}[X] > 0. \tag{B.19}$$

Straightforward calculation gives

$$\text{VAR}_{P_{\mu_i}}[X] = \text{VAR}_{Q_{1,\beta_i}}[X] = \frac{d^2}{d^2\beta_i}(\log \frac{1}{\beta_i}) = \frac{1}{\beta_i^2} \text{ in particular } \text{VAR}_{P_{\mu_0^*}}[X] = \frac{1}{(\beta_0^*)^2} \tag{B.20}$$

where β_i corresponds to μ_i, i.e. $\mathbb{E}_{Q_{1,\beta_i}}[(X)] = \mu_i$. We also have:

$$\mathbb{E}_{P_{\beta_0^*}}[(X)] = \mu_0^* = \frac{1}{k}\sum_{i=1}^{k} \mu_i = \frac{1}{k}\sum_{i=1}^{k} \mathbb{E}_{P_{\beta_i}}[(X)]\,. \tag{B.21}$$

While $\mathbb{E}_{P_{\beta_i}}[(X)] = \frac{d}{d\beta_i}(\log \frac{1}{\beta_i}) = -\frac{1}{\beta_i}$, therefore $\frac{1}{\beta_0^*} = \frac{1}{k}\sum_{i=1}^{k} \frac{1}{\beta_i}$. We obtain, together with (B.20) and (B.21), that

$$\sum_{i=1}^{k} \text{VAR}_{P_{\mu_i}}[(X)] - k\text{VAR}_{P_{\mu_0^*}}[(X)] = \sum_{i=1}^{k} \frac{1}{(\beta_i)^2} - k\left(\frac{1}{k}\sum_{i=1}^{k}\frac{1}{\beta_i}\right)^2. \tag{B.22}$$

Jensen's inequality now gives that (B.22) is strictly positive, whenever at least one of the μ_i is not equal to μ_0^*, which is what we had to show. □

B.5 Graphical Depiction of RIPr-Approximation

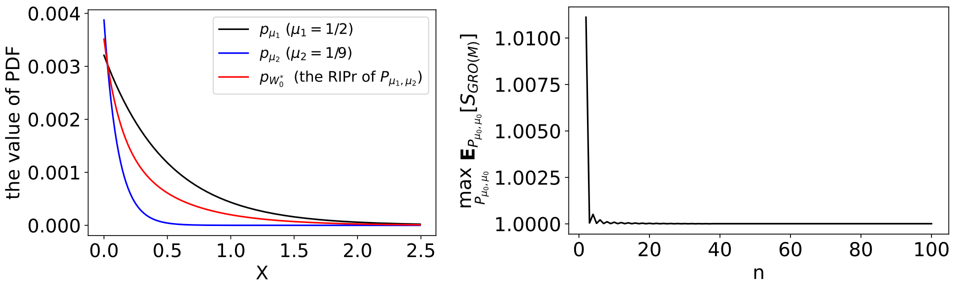

Figure B.1: Exponential distribution. On the right, n represents number of iterations with Li's algorithm, starting at iteration 2

We illustrate RIPr-approximation and convergence of Li's algorithm with four distributions: exponential, beta with free β and fixed α, geometric and Gaussian with free variance and fixed mean, each with one particular (randomly chosen) setting of the parameters. The pictures on the left in Figure B.1– B.4 give the probability density functions (for geometric distributions, discrete probability mass functions) after

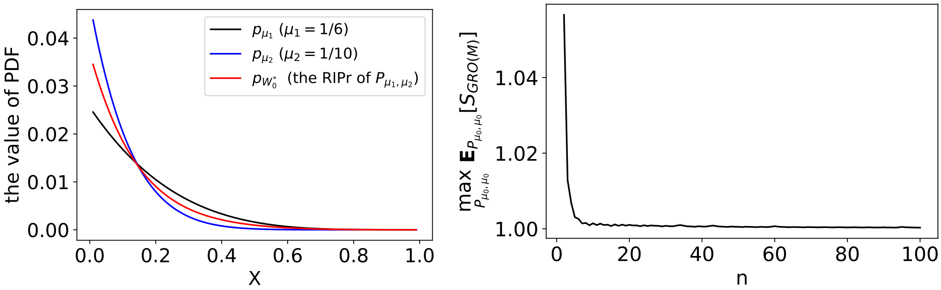

Figure B.2: beta with free β and fixed α. On the right, n represents number of iterations with Li's algorithm, starting at iteration 2

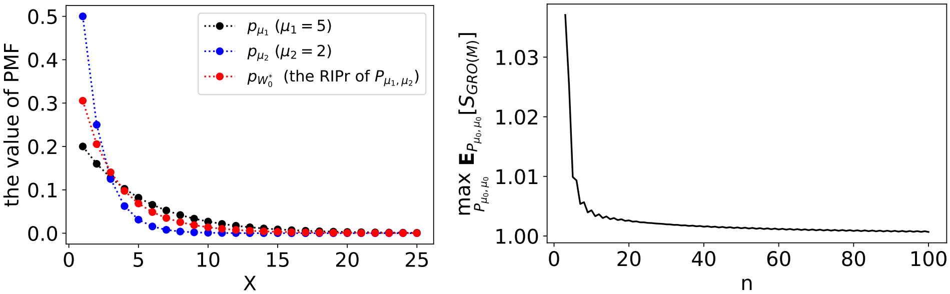

Figure B.3: geometric distribution. On the right, n represents number of iterations with Li's algorithm, starting at iteration 3

$n = 100$ iterations of Li's algorithm. The pictures on the right illustrate the speed of convergence of Li's algorithm. The pictures on the right do not show the first (or the first two, for geometric and Gaussian with free variance) iteration(s), since the worst-case expectation $\sup_{\mu_0 \in \mathtt{M}}[S_{\text{GRO}(\mathcal{M})}]$ is invariably incomparably larger in these initial steps. We empirically find that Li's algorithm converges quite fast for computing the true $S_{\text{GRO}(\mathcal{M})}$. In each step of Li's algorithm, we searched for the best mixture weight α in $P_{(m)}$ over a uniformly spaced grid of 100 points in $[0, 1]$, and for the novel component $P' = P_{\mu',\mu'}$ by searching for μ' in a grid of 100 equally spaced points inside the parameter space $\mathtt{M}$ where the left- and right- endpoints of the grid were determined by trial and error. While with this ad-hoc discretization strategy we obviously cannot guarantee any formal approximation results, in practice it invariably worked well: in all cases, we found that $\max_{\mu_0 \in \mathtt{M}} \mathbb{E}_{P_{\mu_0,\mu_0}}[S_{\text{GRO}(\mathcal{M})}] \leq 1.005$ after 15 iterations. For comparison, we show the best approximation that can be obtained by brute-force combining of just two components, for the same parameter values, in Table B.1.

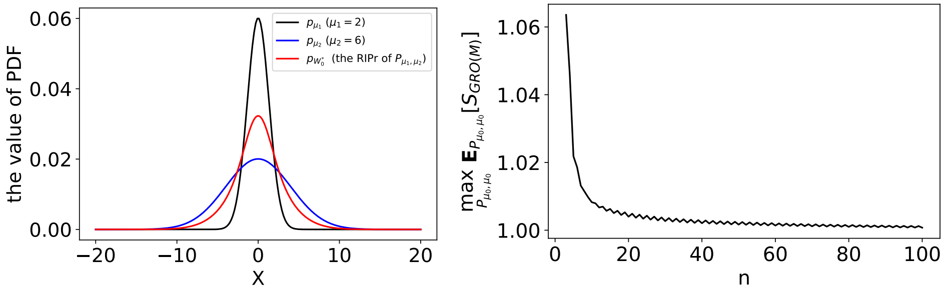

Figure B.4: Gaussian with free variance and fixed mean. On the right, n represents number of iterations with Li's algorithm, starting at iteration 3

Distributions	(μ_1, μ_2)	α	(μ_{01}, μ_{02})	$\sup_{\mu_0 \in \mathtt{M}} \mathbb{E}_{X_1,X_2 \sim P_{\mu_0,\mu_0}}[S]$
beta	$(\frac{1}{6}, \frac{1}{10})$	0.57	(0.12, 0.16)	1.00071
geometric	(5, 2)	0.39	(2.52, 4.21)	1.00035
Exponential	$(\frac{1}{2}, \frac{1}{9})$	0.53	(0.13, 0.51)	1.00083
Gaussian with free variance and fixed mean	(2, 6)	0.41	(5.82, 3.36)	1.00035

Table B.1: Analogue of Table 4.2 for μ_1, μ_2 corresponding to the parameters used in Figures B.1–B.4

B.6 Further Details

In this section, we verify that all conditions are met for the implicit use of Fubini's theorem and differentiation under the integral sign in the proofs of Theorem 2 and 3, and that all derivatives of interest are bounded.

B.6.1 Theorem 2

In the chapter, notation is as follows:

$$\begin{aligned}
\mu_j &= \mu_0 + \delta\alpha_j \\
\lambda(\mu_j) &= \text{nat. param. } \lambda \text{ corresponding to mean } \mu = \mu_j \\
p_\mu(y) &= e^{\lambda(\mu)y - A(\lambda(\mu))} \\
f_y(\delta) &= \sum_{j=1}^{k} p_{\mu_i}(y).
\end{aligned}$$

As this will simplify the notation for the derivatives, we write $g_y(\lambda) = e^{\lambda y - A(\lambda)}$, so that

$$f_y(\delta) = \sum_{j=1}^{k} g_y(\lambda(\mu_j)) \text{ and } p_{\mu_0}(y) = g_y(\lambda(\mu_0)). \tag{B.23}$$

To stress dependence on δ, we write $\mu_j(\delta)$ instead of μ_j in the following.

Step 1 We first establish the finiteness condition (B.8). We note that

$$\begin{aligned}
\log \sum_{j=1}^{k} g_y(\lambda(\mu_j(\delta))) &\leq \log(\max_j g_y(\lambda(\mu_j(\delta)))k) \\
&= \max_j \log(g_y(\lambda(\mu_j(\delta)))) + \log k \\
&\leq \max_j \log(\max\{g_y(\lambda(\mu_j(\delta))), 1\}) + \log k \\
&\leq \sum_j \log(\max\{g_y(\lambda(\mu_j(\delta))), 1\}) + \log k \\
&\leq \sum_j |\lambda(\mu_j(\delta))y - \log A(\lambda(\mu_j(\delta)))| + \log k.
\end{aligned}$$

and

$$\begin{aligned}
\log \sum_{j=1}^{k} g_y(\lambda(\mu_j(\delta))) &= \log \frac{1}{k} \sum_{j=1}^{k} g_y(\lambda(\mu_j(\delta))) + \log k \\
&\geq \frac{1}{k} \sum_{j=1}^{k} \log g_y(\lambda(\mu_j(\delta))) + \log k \\
&= \frac{1}{k} \sum_{j=1}^{k} \lambda(\mu_j(\delta))y - A(\lambda(\mu_j(\delta))) + \log k.
\end{aligned}$$

Putting these together, we see that

$$\begin{aligned}
&|\log f_y(\delta)| \leq \\
&\max\left\{ \sum_j |\lambda(\mu_j(\delta))y - A(\lambda(\mu_j(\delta)))| + \log k, \left| \frac{1}{k} \sum_{j=1}^{k} (\lambda(\mu_j(\delta))y - A(\lambda(\mu_j(\delta)))) + \log k \right| \right\} \\
&\leq \sum_j |\lambda(\mu_j(\delta))y - A(\lambda(\mu_j(\delta)))| + \log k,
\end{aligned} \tag{B.24}$$

and, more trivially,

$$|\log g_y(\lambda(\mu_0))| \leq |\lambda(\mu_0)y - A(\lambda(\mu_0)|\,. \tag{B.25}$$

We know that $\lambda(\mu_j(\delta))$ and $A(\lambda(\mu_j(\delta)))$ are smooth, hence finite functions for $\mu_j(\delta)$ in the interior of the mean-value parameter space $\mathtt{M}$ (see (Barndorff-Nielsen, 1978, Chapter 9, Theorem 9.1 and Eq. (2))). Since $\mathtt{M}$ is open and for all $j = 1..k$, $\mu_j(0) = \mu_0 \in \mathtt{M}$, it follows that $|\log f(y)(\delta) - \log g_y(\lambda(\mu_0))|$ can be written as a smooth, in particular finite function of $|y|$ for all δ in a compact subset of $\mathbb{R}$ with 0 in its interior. Since $|y| \leq 1 + y^2$ has finite expectation under all P_μ with $\mu \in \mathtt{M}$, finiteness of (B.8) follows by (B.23).

Step 2 We now proceed to establish that we can differentiate with respect to δ for δ in a compact subset of $\mathbb{R}$ with 0 in its interior. The proof will make use of (B.24) and (B.25). We denote derivatives of functions f_y and g_y as

$$g_y^s(\lambda) = \frac{\mathrm{d}^s}{\mathrm{d}\lambda^s} g_y(\lambda) \quad \text{and} \quad f_y^s(\delta) = \frac{\mathrm{d}^s}{\mathrm{d}\delta^s} f_y(\delta).$$

We will argue that, for any $s \in \mathbb{N}$, the family $\{\frac{\mathrm{d}^s}{\mathrm{d}\delta^s} f_y(\delta) \log f_y(\delta) - f_y(\delta) \log g_y(\lambda(\mu_0)) : \delta \in \Delta\}$ is uniformly integrable for any compact $\Delta \subset \mathbb{R}$, so that we are allowed to interchange differentiation and integration (see e.g. Williams, 1991, Chapter A16).

Using standard results for exponential families, we have, for λ in the interior of the canonical parameter space,

$$\begin{aligned} g_y^{(1)}(\lambda) &= (y - \mu(\lambda))g_y(\lambda) \\ g_y^{(2)}(\lambda) &= -I(\lambda)g_y(\lambda) + (y - \mu(\lambda))^2 g_y(\lambda), \end{aligned}$$

where $\mu(\lambda)$ denotes the mean-value parameter corresponding to λ and $I(\lambda)$ the corresponding Fisher information.

Continuing this using the fact that $(d^s/d\lambda^s)A(\lambda)$ is continuous for all s, gives

$$g_y^{(s)}(\lambda) = g_y(\lambda) \cdot h_{y,s}(\lambda) \text{ with } h_{y,s}(\lambda) = \sum_{t=1}^{s} h_{[t,s]}(\lambda)(y - \mu(\lambda))^t \tag{B.26}$$

for some smooth functions $h_{[1,s]}, h_{[2,s]}, \ldots, h_{[s,s]}$ of λ (we do not need to know precise definitions of these functions). Similarly

$$f_y^{(1)}(\delta) = \sum_j g_y^{(1)}(\lambda_{\mu_j(\delta)}) \cdot (\lambda(\mu_j(\delta)))'$$

where $\lambda(\mu_j(\delta))' = \frac{\mathrm{d}}{\mathrm{d}\delta}\lambda(\mu_j(\delta))$. We know that $\lambda'(\mu_j(\delta))$ and further derivatives are smooth functions for $\mu_j(\delta)$ in the interior of the mean-value parameter space $\mathtt{M}$ (see (Barndorff-Nielsen, 1978, Chapter 9, Theorem 9.1 and Eq. (2))). Since this space is open and for all $j = 1..k$, $\mu_j(0) = \mu_0 \in \mathtt{M}$, it follows that $\lambda'(\mu_j(\delta))$ are smooth functions of δ for δ in a compact subset of $\mathbb{R}$ with 0 in its interior. Thus, analogously to what we did above with $g^{(s)}$, we get that

$$f_y^{(s)}(\delta) = \sum_j \sum_{t=1}^{s} g_y^{(t)}(\lambda(\mu_j(\delta))) \cdot r_{t,s}(\mu_j) \tag{B.27}$$

for some smooth functions $r_{t,s}$, the details of which we do not need to know. In particular this gives, with

$$b_y^{(s)} := \frac{f_y^{(s)}(\delta)}{f_y(\delta)}$$

that

$$\begin{aligned}\left|b_y^{(s)}\right| &\leq \frac{\sum_j g_y(\lambda(\mu_j(\delta))) \cdot (\sum_{t=1}^{s} |h_{y,t}(\lambda(\mu_j(\delta))) \cdot r_{t,s}(\mu_j(\delta))|)}{\sum_j g_y(\lambda(\mu_j(\delta)))} \\ &\leq \sum_j \sum_{t=1}^{s} |h_{y,t}(\lambda(\mu_j(\delta))) \cdot r_{t,s}(\mu_j(\delta))|.\end{aligned}$$

Inspecting the proof in the main text, we informally note that all terms without logarithms in the first four derivatives of $F_0(\delta)$ and $F_1(\delta)$ can be written as products $f_y(\delta) \cdot b_y^{(s_1)}(\delta) \cdot \ldots \cdot b_y^{(s_u)}(\delta)$ for the $b_y^{(s)}$ we just bounded in terms of polynomials in $|y|$; similarly, the terms involving logarithms can be bounded in terms of such polynomials as well using (B.24) and (B.25), suggesting that all terms inside all integrals can be such bounded. This is indeed the case: formalizing the reasoning, we see that

$$\int \left(\frac{\mathrm{d}^s}{\mathrm{d}\delta^s} f_y(\delta) \log f_y(\delta) - f_y(\delta) \log g_y(\lambda(\mu_0)) \right)^2 d\rho(y) =$$

$$\int \left(f_y^{(s)} (\log f_y(\delta) - \log g_y(\lambda(\mu_0))) + f_y(\delta) \sum_u c_u \cdot b_y^{(s_2)}(\delta) \cdot \ldots \cdot b_y^{(s_u)}(\delta) \right)^2 d\rho(y)$$

$$= \int (f_y^{(s)} (\log f_y(\delta) - \log g_y(\lambda(\mu_0))))^2 + \left(f_y(\delta) \sum_u c_u \cdot b_y^{(s_1)}(\delta) \cdot \ldots \cdot b_y^{(s_u)}(\delta) \right)^2$$
$$+ f_y(\delta) f_y^{(s)} (\log f_y(\delta) - \log g_y(\lambda(\mu_0))) \sum_u c_u \cdot b_y^{(s_1)}(\delta) \cdot \ldots \cdot b_y^{(s_u)}(\delta) d\rho(y).$$

By (B.24) and (B.25) and the bound on $|b_y^{(s)}|$ given above, all the terms within the integral can be bounded by polynomials in y (or $|y|$), so the integral is given by linear functions of moments of ρ and P_μ. Therefore, using also that ρ is itself a probability measure and a member of the exponential family under consideration (equal to P_μ with $\lambda(\mu) = 0$), the integral can be uniformly bounded over δ in a compact subset of the mean-value parameter space. It follows that the family $\{\frac{\mathrm{d}^s}{\mathrm{d}\delta^s} f_y(\delta) \log f_y(\delta) - f_y(\delta) \log g_y(\lambda(\mu_0)) : \delta \in \Delta\}$ is uniformly integrable (see e.g. Williams, 1991, Chapter 13.3), so integration and differentiation may be interchanged freely (see e.g. Williams, 1991, Chapter A16). It also follows that the quantity on the right-hand side in the theorem statement is bounded.

B.6.2 Theorem 3

As in the proof of Theorem 3, let $f(\delta) = \mathbb{E}_{P_{\boldsymbol{\mu}}} \left[\log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle \mu_0 \rangle}(X^k)} - \log \frac{p_{\boldsymbol{\mu}}(X^k|Z)}{p_{\langle \mu_0 \rangle}(X^k|Z)} \right]$.

To validate the proof in the main text we merely need to show that $f(\delta)$ is finite, and that we can interchange differentiation and expectation with respect to δ in a compact interval containing $\delta = 0$. Thus, we want to show that, for any $s \in \mathbb{N}$, we have that

$$\frac{\mathrm{d}^s}{\mathrm{d}\delta^s} f(\delta) = \mathbb{E} \left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s} \left(\log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle \mu_0 \rangle}(X^k)} - \log \frac{p_{\boldsymbol{\mu}}(X^k \mid Z)}{p_{\langle \mu_0 \rangle}(X^k \mid Z)} \right) \right].$$

To show this, first note that both $\mathbb{E}_{P_{\boldsymbol{\mu}}} \left[\log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle \mu_0 \rangle}(X^k)} \right]$ and $\mathbb{E}_{P_{\boldsymbol{\mu}}} \left[\log \frac{p_{\boldsymbol{\mu}}(X^k|Z)}{p_{\langle \mu_0 \rangle}(X^k|Z)} \mid Z \right]$ are KL divergences between members of exponential families (the fact that conditioning on a sum of sufficient statistics results in a new, derived full exponential family is shown by, for example, Brown (1986)), which are finite as long as δ is in a sufficiently small interval containing 0 in its interior (since then $\boldsymbol{\mu}$ is in the interior of the mean-value parameter space). This already shows that $f(\delta)$ is finite, and it also allows us to rewrite

$$f(\delta) = \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle \mu_0 \rangle}(X^k)}\right] - \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log \frac{p_{\boldsymbol{\mu}}(X^k \mid Z)}{p_{\langle \mu_0 \rangle}(X^k \mid Z)}\right].$$

Furthermore, (Brown, 1986, Theorem 2.2) in combination with Theorem 9.1. and Chapter 9, Eq.2. of Barndorff-Nielsen (1978) shows that for any full exponential family, for any finite $k > 0$, the k-th derivative of the KL divergence with respect to its first argument, given in the mean-value parameterization, exists, is finite, and can be obtained by differentiating under the integral sign, at any μ in the interior of the mean-value parameter space. We are therefore allowed to interchange expectation and differentiation for such terms separately for all δ in any compact interval containing 0. Thus, starting with the previous display, we can write

$$\begin{aligned}
&\frac{\mathrm{d}^s}{\mathrm{d}\delta^s} f(\delta) = \frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle \mu_0 \rangle}(X^k)}\right] - \frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log \frac{p_{\boldsymbol{\mu}}(X^k \mid Z)}{p_{\langle \mu_0 \rangle}(X^k \mid Z)}\right] \\
&= \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle \mu_0 \rangle}(X^k)}\right] - \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\log \frac{p_{\boldsymbol{\mu}}(X^k \mid Z)}{p_{\langle \mu_0 \rangle}(X^k \mid Z)}\right] \\
&= \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle \mu_0 \rangle}(X^k)}\right] - \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle \mu_0 \rangle}(X^k)} + \log \frac{p_{\boldsymbol{\mu};[Z]}(Z)}{p_{\langle \mu_0 \rangle;[Z]}(Z)}\right] = \\
&\mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle \mu_0 \rangle}(X^k)}\right] - \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle \mu_0 \rangle}(X^k)}\right] + \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\log \frac{p_{\boldsymbol{\mu};[Z]}(Z)}{p_{\langle \mu_0 \rangle;[Z]}(Z)}\right] \\
&= \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\log \frac{p_{\boldsymbol{\mu};[Z]}(Z)}{p_{\langle \mu_0 \rangle;[Z]}(Z)}\right],
\end{aligned}$$

where in the last line we use that all involved terms are finite. This is what we had to show.

C | Appendix to Chapter 5

C.1 Details for Section 5.4.4

We need to establish that $\Sigma_p(\boldsymbol{\mu}) - \Sigma_q^{(\theta^\circ)}(\boldsymbol{\mu}) = \Sigma_q^{(0)}(\boldsymbol{\mu}) - \Sigma_q^{(\theta^\circ)}(\boldsymbol{\mu})$ is positive semidefinite for all $\boldsymbol{\mu} \in \mathbb{R}^d$.

Thus, take any $\boldsymbol{\mu}^* \in \mathbb{R}^d$. By (5.26), we have that $q_{\boldsymbol{\mu}^*}^{(\theta^\circ)} = f_{\lambda^\circ,\boldsymbol{\beta}^\circ}^{(\theta^\circ)}$ and $p_{\boldsymbol{\mu}^*} = q_{\boldsymbol{\mu}^*}^{(0)} = f_{\lambda^*,\boldsymbol{\beta}^*}^{(0)}$ for some $\lambda^\circ, \boldsymbol{\beta}^\circ$ and $\lambda^*, \boldsymbol{\beta}^*$ that are related to each other via the normal equations (5.27). Based on the sufficient statistics (5.25), we can thus write, for $\theta \in \{0, \theta^\circ\}$, that

$$\Sigma_q^{(\theta)}(\boldsymbol{\mu}^*) = \begin{pmatrix} A^{(\theta)} & B^{(\theta)} \\ (B^{(\theta)})^T & C^{(\theta)} \end{pmatrix}$$

where $A^{(\theta^\circ)}$ is the variance of $\sum Y_i^2$ according to distribution $F_{\lambda^\circ,\boldsymbol{\beta}^\circ}^{(\theta^\circ)}$ and $C^{(\theta^\circ)}$ is the $d \times d$ covariance matrix of the $t_j(Y^n)$ according to this distribution and

$$B^{(\theta^\circ)} = \left(\operatorname{cov}\left(\sum Y_i^2, t_1(Y^n)\right), \ldots, \operatorname{cov}\left(\sum Y_i^2, t_d(Y^n)\right)\right)$$

where the covariances are again under this distribution. Similarly, $A^{(0)}$ is the variance of $\sum Y_i^2$ according to distribution $F_{\lambda^*,\boldsymbol{\beta}^*}^{(0)}$ and $B^{(0)}$, $C^{(0)}$ are defined accordingly.

Positive semidefiniteness of $\Sigma_q^{(0)}(\boldsymbol{\mu}^*) - \Sigma_q^{(\theta^\circ)}(\boldsymbol{\mu}^*)$ is easily seen to be implied[1] if we can show that $C^{(0)} - C^{(\theta^\circ)}$ is positive definite and that

$$(A^{(0)} - A^{(\theta^\circ)}) - (B^{(0)} - B^{(\theta^\circ)})^T (C^{(0)} - C^{(\theta^\circ)})^{-1} (B^{(0)} - B^{(\theta^\circ)}) \geq 0. \tag{C.1}$$

To show that $C^{(0)} - C^{(\theta^\circ)}$ is positive definite, note that $C^{(\theta^\circ)}$ (as is readily established, for example, by twice differentiating $\log Z_q^{(\theta^\circ)}(\lambda, \boldsymbol{\beta}; \boldsymbol{\mu}^*)$ at $\lambda = 0, \boldsymbol{\beta} = 0$) is simply

[1] For an explicit derivation see `https://math.stackexchange.com/questions/2280671/definiteness-of-a-general-partitioned-matrix-mathbf-m-left-beginmatrix-bf`.

the standard covariance matrix in linear regression scaled by $1/\sigma^{\circ 2}$, i.e. $C^{(\theta^\circ)} = \sigma^{\circ 2} \sum \mathbf{x}_i \mathbf{x}_i^T$ which by the maximal rank assumption is positive definite. Similarly $C^{(0)} = \sigma^{*2} \sum \mathbf{x}_i \mathbf{x}_i^T$ so that, since by assumption $\theta^\circ \neq 0$ and using the normal equations (5.27), we have that $C^{(0)} - C^{(\theta)} = cC^{(\theta)}$ for $c = \sigma^{*2} - \sigma^{\circ 2} > 0$ is also positive definite.

It only remains to show (C.1). As again easily established (for example, by twice differentiating $\log Z_q^{(\theta^\circ)}(\lambda, \boldsymbol{\beta}; \boldsymbol{\mu}^*)$ at $\lambda = 0, \boldsymbol{\beta} = 0$), we have that $A^{(\theta^\circ)} = 2\sigma^{\circ 2} \left(2(\sum \nu_i^{\circ 2}) + n\sigma^{\circ 2}\right)$ and similarly we find $A^{(0)} = 2\sigma^{*2} \left(2(\sum \nu_i^{*2}) + n\sigma^{*2}\right)$ and $B_j^{(\theta^\circ)} = -2\sigma^{\circ 2} \left(\sum \nu_i^\circ x_{i,j}\right)$ and similarly $B_j^{(0)} = -2\sigma^{*2} \left(\sum \nu_i^* x_{i,j}\right)$. By the normal equations (5.27) we find that $B_j^{(0)} - B_j^{(\theta^\circ)} = -2(\sigma^{*2} - \sigma^{\circ 2}) \sum \nu_i^* x_{i,j}$. After some matrix multiplications (where we may use the cyclic property of the trace of a matrix product) we get that (C.1) is equivalent to

$$(A^{(0)} - A^{(\theta)^\circ}) - 4(\sigma^{*2} - \sigma^{\circ 2}) \sum \nu_i^{*2} \geq 0.$$

But this is easily verified: it is equivalent to

$$2\sigma^{*2} \left(2\left(\sum \nu_i^{*2}\right) + n\sigma^{*2} - 2\left(\sum \nu_i^{*2}\right)\right) - 2\sigma^{\circ 2} \left(2\left(\sum \nu_i^{\circ 2}\right) + n\sigma^{\circ 2} - 2\left(\sum \nu_i^{*2}\right)\right) \geq 0$$

which in turn is equivalent to

$$2n\sigma^{*4} - 2n\sigma^{\circ 4} + 4(\sum \nu_i^{*2} - \sum \nu_i^{\circ 2})\sigma^{\circ 2} \geq 0$$

which by the normal equations is equivalent to

$$\sigma^{*4} - \sigma^{\circ 4} + 2(\sigma^{*2} - \sigma^{\circ 2})\sigma^{\circ 2} \geq 0$$

but this must be the case since by the normal equations, $\sigma^{*2} > \sigma^{\circ 2}$.

D | Appendix to Chapter 6

D.1 Additional Simulations

D.1.1 Effect of Truncation on Power

Figure D.1 shows the same plot as Figure 1 in Chapter 6 but without truncation for the probabilities in the e-statistics, i.e. $\varepsilon = 0$.

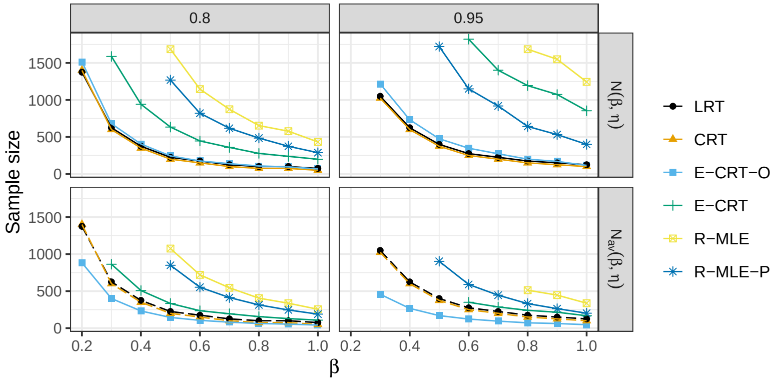

Figure D.1: Sample sizes for different methods as in Figure 1 in Chapter 6 but with $\varepsilon = 0$ for the e-statistics.

D.1.2 Robustness With Respect to Misspecification

We test the robustness of the randomization based e-statistics with respect to misspecification of the conditional distribution of X in the same way as in the simulation study of Berrett et al. (2020). All simulations in this section are under the null hypothesis, i.e. $\beta = 0$. Rejection rates of the e-statistics are again computed with a maximal sample size of 2000 and with optional stopping, i.e. rejection if the level $1/\alpha$ is exceeded at least once, and with truncation level $\varepsilon = 0.05$. For comparison, the conditional randomization test is applied to a sample of fixed size, for sizes 200, 1000, and 2000, and additionally with the unconditional absolute correlation $|\mathrm{cor}(X, Y)|$ as test statistic, as in Berrett et al. (2020), for sample sizes 200 and 2000.

First, instead of sampling X with conditional mean μ_Z as defined in (6.14), we set the mean to

$$\begin{aligned} &\mu_Z - \xi\mu_Z^3 && \text{(cubic misspecification)}, \\ &\mu_Z + \xi\mu_Z^2 && \text{(quadratic misspecification)}, \\ &\tanh(\xi\mu_Z)/\xi && \text{(hyperbolic tangent)}, \end{aligned}$$

which are the same misspecifications as in Berrett et al. (2020, Section 6.1.1). They are illustrated in Figure D.2 for different values of ξ, the range of which has been selected for each misspecification type in such a way that the relative misspecification compared to the true mean approximately matches the one in the simulations by Berrett et al. (2020). When the parameter ξ equals zero, understood as limit $\xi \to 0$ for the hyperbolic tangent, the model is correctly specified. Panel (a) of Figure D.3 shows that both the CRT and the e-statistics are robust with respect to slight misspecifications of the conditional mean. The CRT based on the likelihood is much more robust than the other two tests, due to the fact that re-estimating the logistic regression model with simulated X is invariant under affine transformations of X and Z and hence able to correct much of the misspecification. The e-statistic based test is less robust than this variant of the CRT, since it does not re-estimate the logistic model with simulated X, but still substantially more robust than the CRT based on unconditional correlation, which already with $n = 200$, as compared to $n = 2000$ for the e-values, has rejection rates strongly exceeding the nominal level as ξ increases.

In panel (b) of Figure D.3, the rejection rates of the tests are shown when the distribution of X_p is estimated on an independent unlabeled data set, for different sizes of this data set. The estimation of the conditional distribution is by linear regression, with the maximum likelihood estimator for the conditional variance. Here the e-statistics have rejection rates below the nominal level, even for unlabeled sample size as small as 50. Also the CRT with logistic likelihood as test statistic has rejection rate close to the nominal level.

Finally, in panel (c) of Figure D.3 the rejection rates are depicted for the case when the same data is used both for estimating the distribution of X and for testing. The

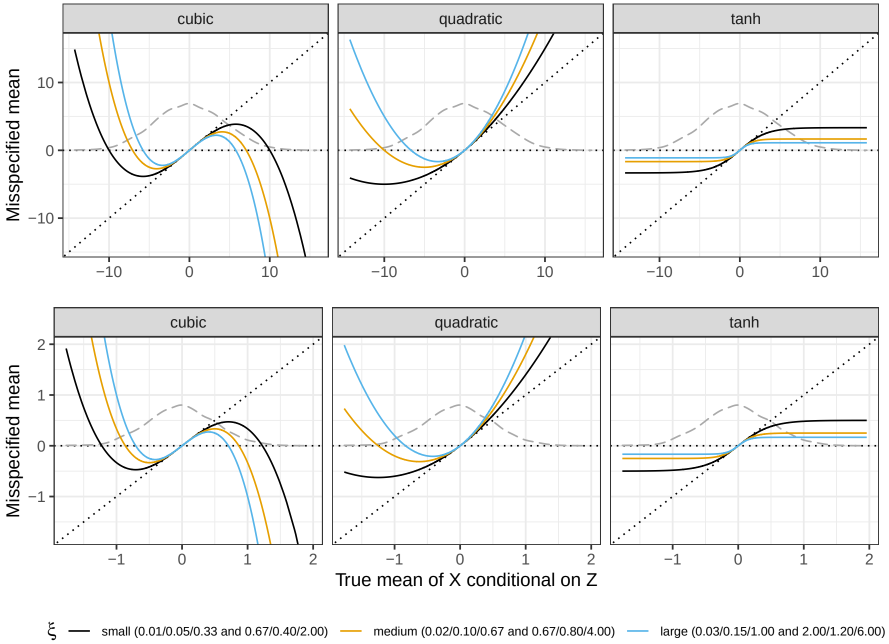

Figure D.2: Misspecification in the conditional mean of X given Z for the three different functions from Section D.1.2. Upper row of plots: $X \mid Z$ generated as in Berrett et al. (2020, Section 6.1.1). Lower row: $X \mid Z$ generated as in Section 6.4 with $q = 4$. The dashed line shows the (height adjusted) density of the conditional expectation of X given Z. The values for ξ given in the legend refer to the misspecifications in the same order as the panel colums (cubic/quadratic/tanh), with the first triple giving ξ for $X \mid Z$ as simulated by Berrett et al. (2020) (upper three figures), and the second triple the values of ξ applied when $X \mid Z$ is generated as in Section 6.4 (lower three figures).

estimation is as described in the previous paragraph. For the CRT, the distribution of X is estimated on the same data to which the test is applied, like in the simulation study of Berrett et al. (2020). For the e-statistics, a slightly different approach is taken, tailored to sequential settings. We start with a potentially small unlabeled sample, and each time a new instance is observed, the estimate of the distribution of X is updated with all the data available so far. Again, all tests except for the correlation based CRT with sample size 2000 have rejection rate close to the nominal level.

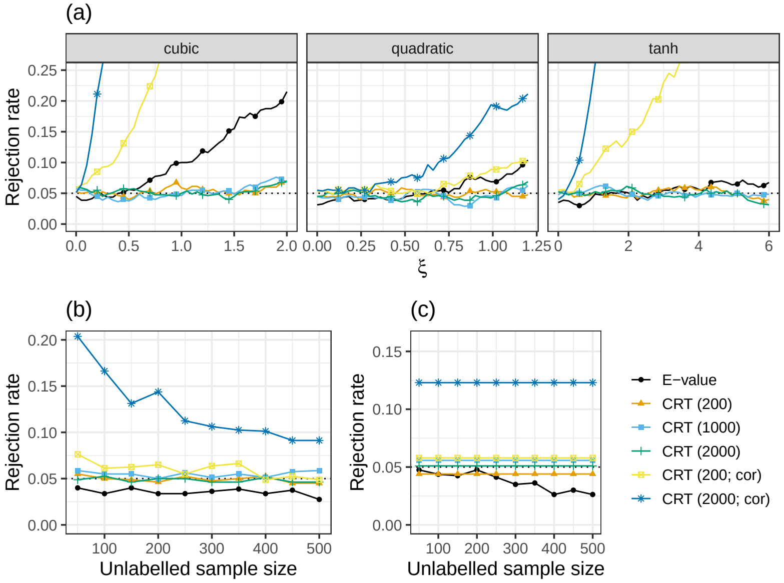

Figure D.3: (a) Rejection rates of e-values and conditional randomization test (with sample sizes $n = 200$, 1000, 2000 and likelihood as test statistic, and $n = 200$, 2000 and correlation as test statistic) at the level $\alpha = 0.05$, under different misspecifications for the conditional mean of X. (b) Rejection rates when the distribution of X is estimated on a separate sample, for varying sample sizes. (c) Rejection rates when the same data is used both for estimating the conditional distribution of X and applying the test, as described in the text.

D.2 Proofs of Main Results

D.2.1 Proof of Theorem 6.1

Proof. Let $P \in \mathcal{H}_0$ arbitrarily. The proof relies on the simple insight that we can separate the expectation with respect to (Y_n, Z_n) from that with X_n,

$$\begin{aligned}
&\mathbb{E}_P[E^{\mathrm{CI}}_{h_n}(X_n, Y_n, Z_n) \mid D^{n-1}] \\
&\quad = \mathbb{E}_P\left[\mathbb{E}_P\left[\frac{h_n(X_n, Y_n, Z_n)}{\int_{\mathcal{X}} h_n(x, Y_n, Z_n)\,\mathrm{d}Q_{Z_n}(x)}\middle| Y_n, Z_n, D^{n-1}\right]\middle| D^{n-1}\right] \\
&\quad = \mathbb{E}_P\left[\frac{\int_{\mathcal{X}} h_n(x', Y_n, Z_n)\,\mathrm{d}Q_{Z_n}(x')}{\int_{\mathcal{X}} h_n(x, Y_n, Z_n)\,\mathrm{d}Q_{Z_n}(x)}\middle| D^{n-1}\right] = 1,
\end{aligned}$$

where in the last step we use that $P_{X_n|Y_n,Z_n} = P_{X_n|Z_n} = Q_{Z_n}$. □

D.2.2 Proof of Proposition 6.2

Proof. Define $\tilde{X}_0 = X_n$. The random variables $\tilde{X}_0, \dots, \tilde{X}_M$ are exchangeable, so

$$\check{E}^{\mathrm{CI}}_{h_n;j}(D_n) := \frac{h_n(\tilde{X}_j, Y_n, Z_n)}{\sum_{i=0}^M h_n(\tilde{X}_i, Y_n, Z_n)/(M+1)},\ j = 0, \dots, M,$$

have the same expected value as $\check{E}^{\mathrm{CI}}_{h_n}(D_n) = \check{E}^{\mathrm{CI}}_{h_n;0}(D_n)$. Since $\sum_{i=0}^M \check{E}^{\mathrm{CI}}_{h_n;i}(D_n) \equiv M+1$, this implies $\mathbb{E}_P[\check{E}^{\mathrm{CI}}_{h_n}(D_n) \mid D^{n-1}] = 1$. □

D.2.3 Proof of Theorem 6.3

Proof. Let $f = f_{X,Y,Z}(x,y,z)$ be the density of (X,Y,Z) with respect to a measure $\sigma \times \mu \times \nu$ on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. Then the conditional density $f_{Y|X,Z}$ equals

$$f_{Y|X,Z}(y \mid x,z) = \frac{f(x,y,z)}{\int_{\mathcal{Y}} f(x,s,z)\,\mathrm{d}\sigma(s)}.$$

The density of Q_Z must equal the conditional density $f_{X|Z}$, which is given by

$$f_{X|Z}(x \mid z) = \frac{\int_{\mathcal{Y}} f(x,s,z)\,\mathrm{d}\sigma(s)}{\int_{\mathcal{X}} \int_{\mathcal{Y}} f(r,s,z)\,\mathrm{d}\sigma(s)\,\mathrm{d}\mu(r)},$$

so that, with $h(x,y,z) = f_{Y|X,Z}(y|x,z)$,

$$\int_{\mathcal{X}} h(x,y,z)\,\mathrm{d}Q_z(x) = \int_{\mathcal{X}} \frac{f(r,y,z)}{\int_{\mathcal{X}} \int_{\mathcal{Y}} f(r',s,z)\,\mathrm{d}\sigma(s)\,\mathrm{d}\mu(r')}\,\mathrm{d}\mu(r) = f_{Y|Z}(y \mid z).$$

Hence the e-statistic with this choice of h is equal to

$$E^{\mathrm{CI}}_{f_{Y|X,Z}}(X_i, Y_i, Z_i) = \frac{f_{Y|X,Z}(Y_i \mid X_i, Z_i)}{f_{Y|Z}(Y_i \mid Z_i)} = \frac{f_{X,Y,Z}(X_i, Y_i, Z_i)}{f_{Y|Z}(Y_i \mid Z_i) f_{X|Z}(X_i \mid Z_i) f_Z(Z_i)}.$$

The denominator is the density of an element of $\mathcal{H}_0$ as in (6.1). Theorem 1 by Grünwald et al. (2024) states that this e-statistic must therefore be the GRO e-statistic for a single data point (X_i, Y_i, Z_i) and the same argument can be applied to the product of these e-statistics. Finally, a slight rewriting shows that the e-statistic corresponds to the ratio of the joint conditional density of (X,Y) given Z divided by the product of its marginals. For all i, the expected value of $\log E^{\mathrm{CI}}_{f_{Y|X,Z}}(X_i, Y_i, Z_i)$ conditional on Z is therefore equal to the conditional mutual information of X and Y given Z. □

D.2.4 Proof of Proposition 6.4

Proof. Since the distribution Q_Z is well-specified, we denote $g_{Y|Z}$ for the density $\int g_{Y|X,Z}\,\mathrm{d}Q_Z$. Then the quantity of interest is given by

$$\begin{aligned}
\mathbb{E}_f\left[\log E^{\mathrm{CI}}_{g_{Y|X,Z}}(x,y,z)\right] &= \mathbb{E}_f\left[\log\frac{g_{Y|X,Z}(y\mid x,z)}{g_{Y|Z}(y\mid z)}\right]\\
&= I_f(X;Y\mid Z) + \mathbb{E}_f\left[\log\frac{g_{Y|X,Z}(y\mid x,z)}{g_{Y|Z}(y\mid z)} - \log\frac{f(x,y,z)}{f_{X|Z}(x\mid z)f_{Y|Z}(y\mid z)f_Z(z)}\right]\\
&= I_f(X;Y\mid Z) + \mathbb{E}_f\left[\log\frac{g_{Y|X,Z}(y\mid x,z)}{g_{Y|Z}(y\mid z)} - \log\frac{f_{Y|X,Z}(y\mid x,z)}{f_{Y|Z}(y\mid z)}\right]\\
&= I_f(X;Y\mid Z) + \mathbb{E}_f[\mathrm{KL}(f_{Y|Z}\|g_{Y|Z})] - \mathbb{E}_f[\mathrm{KL}(f_{Y|X,Z}\|g_{Y|X,Z})].
\end{aligned}$$

The desired result follows from the nonnegativity of KL divergence. □

D.2.5 Proof of Theorem 6.6

Proof. Fix $N\in\mathbb{N}$ and $\alpha\in(0,1)$. Conditional on Y_i, Z_i, $i=1,\dots,N$, the randomness of the process $S_n = S_n(X^n) = \prod_{i=1}^n \tilde{E}^{\mathrm{CI}}_{h_n}$, $n=1,\dots,N$, solely stems from $X_1,\dots,X_N$, and we will write Y_i, Z_i with lower case letters y_i, z_i to reflect that all statements are conditional on their values. So the e-value at time n writes as

$$\tilde{E}^{\mathrm{CI}}_{h_n} = \frac{h_n(X_n,y_n,z_n\mid X^{n-1},y^{n-1},z^{n-1})}{\int_{\mathcal{X}} h_n(x,y_n,z_n\mid X^{n-1},y^{n-1},z^{n-1})\,\mathrm{d}\hat{Q}_{z_n}(x)}.$$

The condition $h_n>0$ ensures that this e-value is well-defined. For $n>N$, set $h_n\equiv 1$, so that $S_n = S_N$ for $n>N$. If X^N has distribution $\hat{Q}^N_{Z^N}$, then the process $(S_n)_{n\in\mathbb{N}}$ is a nonnegative martingale with respect to the filtration $\mathcal{F}_n = \sigma(X_1,\dots,X_n)$, because

$$\mathbb{E}\left[S_n|X^{n-1}\right] = \int_{\mathcal{X}}\frac{h_n(x,y_n,z_n\mid X^{n-1},y^{n-1},z^{n-1})}{\int_{\mathcal{X}} h_n(x,y_n,z_n\mid X^{n-1},y^{n-1},z^{n-1})\,\mathrm{d}\hat{Q}_{Z_n}(x)}\,\mathrm{d}\hat{Q}_{Z_n}(x) = 1$$

almost surely. Hence by Ville's inequality, $P\left(\exists\, n\le N\colon S_n\ge 1/\alpha\right)\le\alpha$. Let

$$A = \left\{x^n\in\mathcal{X}^n\colon \exists\, n\le N \text{ s.t. } S_n(x^n)\ge 1/\alpha\right\}.$$

Then, since $Q^N_{Z^N}(A) = P(\exists\, n\le N\colon S_n\ge 1/\alpha\mid Y^N=y^N, Z^N=z^N)$,

$$P(\exists\, n\le N\colon S_n\ge 1/\alpha\mid Y^N,Z^N)\le \hat{Q}^N_{z^N}(A)+d_{\mathrm{TV}}(Q^N_{Z^N},\hat{Q}^N_{Z^N})\le\alpha+d_{\mathrm{TV}}(Q^N_{Z^N},\hat{Q}^N_{Z^N}).$$

□

D.2.6 Proof of Proposition 6.7

Proof. The subgaussianity assumption (i) implies that

$$P\left(|u^\top((X,Z)-\mathbb{E}[(X,Z)])|\geq\eta\right)\leq 2\exp(-\eta^2/(2\|u\|^2\sigma^2))\right),\ \eta>0,\ u\in\mathbb{R}^{p+q}, \quad \text{(D.1)}$$

and that $\mathbb{E}[\|(X,Z)\|^k]<\infty$ for all $k\in\mathbb{N}$. As a consequence of the latter and of assumption (i)(a), Theorem 1 of Qian and Field (2002) implies that the MLE $\hat{\theta}_n$ exists with asymptotic probability one and satisfies $\|\hat{\theta}_n-\theta\|=\mathcal{O}(n^{-1/2}\log(\log(n))^{1/2})$ almost surely.

We now study the properties of the function $\theta\mapsto\log(p_\theta(y\mid x,z))$ for $\theta\in\mathbb{R}^{p+q}$. The derivative of $\log(p_\theta(y\mid x,z))$ with respect to θ_j equals

$$\frac{d}{d\theta_j}\log(p_\theta(y\mid x,z))=\begin{cases}yx_j-x_jp_\theta(1\mid x,z) & \text{if}\, j\leq p\\ yz_{j-p},-z_{j-p}p_\theta(1\mid x,z) & \text{else.}\end{cases}$$

Consequently, for any $\theta,\theta'\in\mathbb{R}^{p+q}$,

$$|\log(p_\theta(y\mid x,z))-\log(p_{\theta'}(y\mid x,z))|\leq\|(x,z)\|\|\theta-\theta'\|. \quad \text{(D.2)}$$

This implies that

$$\begin{aligned}\frac{1}{n}\left|\sum_{i=1}^n\log(p_{\hat{\theta}_{i-1}}(Y_i\mid X_i,Z_i))-\log(p_\theta(Y_i\mid X_i,Z_i))\right|&\leq\frac{1}{n}\sum_{i=1}^n\|\hat{\theta}_{i-1}-\theta\|\|(X_i,Z_i)\|\\&\leq\left(\frac{1}{n}\sum_{i=1}^n\|\hat{\theta}_{i-1}-\theta\|^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^n\|(X_i,Z_i)\|^2\right)^{1/2}.\end{aligned}$$

Since $\|(X_i,Z_i)\|^2$, $i\in\mathbb{N}$, are independent with expectation $\mathbb{E}[\|(X,Z)\|^2]<\infty$, the law of large number implies that $\sum_{i=1}^n\|(X_i,Z_i)\|^2/n\to\mathbb{E}[\|(X,Z)\|^2]<\infty$ almost surely, and $\sum_{i=1}^n\|\hat{\theta}_{i-1}-\theta\|^2/n\to 0$ since $\|\hat{\theta}_n-\theta\|\to 0$ almost surely as $n\to\infty$. It remains to show an analogous convergence result for the denominator in S_n^{CI}. Define

$$r_n=\frac{\int p_\theta(Y_n\mid x,Z_n)\,\mathrm{d}Q_{Z_n}(x)}{\int p_{\hat{\theta}_{n-1}}(Y_n\mid x,Z_n)\,\mathrm{d}Q_{Z_n}(x)}.$$

We want to show that $\liminf_{n\to\infty}\sum_{i=1}^n\log(r_i)/n\geq 0$ almost surely. To this end, write

$$
\begin{aligned}
r_n &= \frac{\int p_\theta(Y_n \mid x, Z_n)\,\mathrm{d}Q_{Z_n}(x)}{\int p_\theta(Y_n \mid x, Z_n)\,\mathrm{d}Q_{Z_n}(x) + \int (p_{\hat{\theta}_{n-1}}(Y_n \mid x, Z_n) - p_\theta(Y_n \mid x, Z_n))\,\mathrm{d}Q_{Z_n}(x)} \\
&\geq \frac{\int p_\theta(Y_n \mid x, Z_n)\,\mathrm{d}Q_{Z_n}(x)}{\int p_\theta(Y_n \mid x, Z_n)\,\mathrm{d}Q_{Z_n}(x) + \int |p_{\hat{\theta}_{n-1}}(Y_n \mid x, Z_n) - p_\theta(Y_n \mid x, Z_n)|\,\mathrm{d}Q_{Z_n}(x)}.
\end{aligned}
$$

Since $\log(1+x) \leq x$, we have $\log(1/(1+x)) = -\log(1+x) \geq -x$, for $x > -1$. So

$$
\log(r_n) \geq -\frac{\int |p_{\hat{\theta}_{n-1}}(Y_n \mid x, Z_n) - \int p_\theta(Y_n \mid x, Z_n)|\,\mathrm{d}Q_{Z_n}(x)}{\int p_\theta(Y_n \mid x, Z_n)\,\mathrm{d}Q_{Z_n}(x)}
$$

The function $\theta \mapsto p_\theta(y \mid x, z)$ is Lipschitz continuous, because for $k = 1, \ldots, p+q$,

$$
\left|\frac{d}{d\theta_k} p_\theta(y \mid x, z)\right| = \begin{cases} |x_k| p_\theta(1 \mid x, z)(1 - p_\theta(1 \mid x, z)) \leq |x_k| & \text{if } k = 1, \ldots, p \\ |z_{k-p}| p_\theta(1 \mid x, z)(1 - p_\theta(1 \mid x, z)) \leq |z_{k-p}| & \text{else.} \end{cases}
$$

This implies that

$$
\log(r_n) \geq -\frac{\|\hat{\theta}_{n-1} - \theta\| \int \|(x, Z_n)\|\,\mathrm{d}Q_{Z_n}(x)}{\int p_\theta(Y_n \mid x, Z_n)\,\mathrm{d}Q_{Z_n}(x)}.
$$

To bound this from below, we now show that the denominator $\int p_\theta(Y_n \mid x, Z_n)\,\mathrm{d}Q_{Z_n}(x)$ is small only with a small probability. Let $\kappa_n = n^{-\delta}/2$ for $\delta > 0$. Define the events

$$
A_n = \left\{ \min_{y=0,1} p_\theta(y \mid X_n, Z_n) \leq \kappa_n \right\}.
$$

Let $\operatorname{logit}(p) = \log(p/(1-p))$. Then,

$$
\min_{y=0,1} p_\theta(y \mid x, z) \leq \kappa_n \iff |\theta^\top (x, z)| \geq |\operatorname{logit}(\kappa_n)|,
$$

and therefore, since $|\operatorname{logit}(p)| \geq |\log(2p)|$ for $p \in (0, 1/2]$,

$$
A_n \subseteq \{|\theta^\top (X_n, Z_n)| \geq |\log(2\kappa_n)|\} = \{|\theta^\top (X_n, Z_n)| \geq \delta \log(n)\},
$$

The above derivations yield $P(A_n) \leq P(|\theta^\top (X_n, Z_n)| \geq \delta \log(n))$, and (D.1) implies, with $B = \|\theta\|$,

$$\begin{aligned} P(|\theta^\top(X,Z)| \geq \delta\log(n)) &\leq P\left(|\theta^\top((X,Z) - \mathbb{E}[(X,Z)])| \geq \delta\log(n)) - |\theta^\top\mathbb{E}[(X,Z)]|\right) \\ &\leq 2\exp(-\delta^2\log(n)^2/(8B^2\sigma^2)), \end{aligned}$$

for n large enough such that $\delta\log(n)/2 \geq |\theta^\top\mathbb{E}[(X,Z)]|$. In a next step, we use this to bound $\min_{y=0,1}\int p_\theta(y \mid x, Z_n)\,\mathrm{d}Q_{Z_n}(x)$. First, note that for $y \in \{0,1\}$,

$$\begin{aligned} \int p_\theta(y \mid x, Z_n)\,\mathrm{d}Q_{Z_n}(x) &= \int p_\theta(y \mid x, Z_n)1\{p_\theta(y \mid x, Z_n) \geq 1-\kappa_n\}\,\mathrm{d}Q_{Z_n}(x) \\ &\quad+ \int p_\theta(y \mid x, Z_n)1\{p_\theta(y \mid x, Z_n) < 1-\kappa_n\}\,\mathrm{d}Q_{Z_n}(x) \\ &\leq Q_{Z_n}(p_\theta(y \mid X_n, Z_n) \geq 1-\kappa_n) + 1 - \kappa_n. \end{aligned}$$

It follows that for $\eta > 0$, if $\int p_\theta(y \mid x, Z_n)\,\mathrm{d}Q_{Z_n}(x) \geq 1 - n^{-\eta}$, then $Q_{Z_n}(p_\theta(y \mid X_n, Z_n) \geq 1-\kappa_n) \geq \kappa_n - n^{-\eta}$. Recall that $\kappa_n = n^{-\delta}/2$ with $\delta > 0$ unspecified so far. For n large enough such that $n^{-\eta/2} \leq 1/4$, choosing $\delta = \eta/2$ implies $\kappa_n - n^{-\eta} = n^{-\eta/2}(1/2 - n^{-\eta/2}) \geq n^{-\eta/2}/4$. Consequently, for large n, by Markov's inequality,

$$\begin{aligned} &P\left(\int p_\theta(y \mid x, Z_n)\,\mathrm{d}Q_{Z_n}(x) \geq 1 - n^{-\eta}\right) \\ &\leq P\left(Q_{Z_n}(p_\theta(y \mid X_n, Z_n) \geq 1-\kappa_n) \geq n^{-\eta/2}/4\right) \\ &\leq 4n^{\eta/2}\mathbb{E}[Q_{Z_n}(p_\theta(y \mid X_n, Z_n) \geq 1-\kappa_n)] \\ &= 4n^{\eta/2}P(p_\theta(y \mid X_n, Z_n) \geq 1-\kappa_n). \end{aligned} \tag{D.3}$$

But it has already been shown that

$$P(p_\theta(y \mid X_n, Z_n) \geq 1-\kappa_n) = P(p_\theta(1-y \mid X_n, Z_n) \leq \kappa_n) \leq 2\exp(-\delta^2\log(n)^2/(8B^2\sigma^2))$$

for large n, which in (D.3) gives an upper bound of

$$8\exp\left(-\log(n)(\eta^2\log(n)/(32B^2\sigma^2) - \eta/2)\right).$$

Since $\eta^2\log(n)/(32B^2\sigma^2) - \eta/2 \to \infty$ as $n \to \infty$, it holds that $\eta^2\log(n)/(32B^2\sigma^2) - \eta/2 > 1$ for n large enough, and we can conclude

$$\sum_{n=1}^\infty P\left(\min_{y=0,1}\int p_\theta(y \mid x, Z_n)\,\mathrm{d}Q_{Z_n}(x) \leq n^{-\eta}\right) < \infty.$$

Thus the Borel-Cantelli Lemma implies that $\min_{y=0,1} \int p_\theta(y \mid x, Z_n)\, \mathrm{d}Q_{Z_n}(x) \leq n^{-\eta}$ holds for only finitely many n with probability one. Now

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \log(r_i) &\geq -\frac{1}{n}\sum_{i=1}^{n} \frac{\|\hat{\theta}_{i-1} - \theta\| \int \|(x, Z_i)\|\, \mathrm{d}Q_{Z_i}(x)}{\int p_\theta(Y_i \mid x, Z_i)\, \mathrm{d}Q_{Z_i}(x)} \\
&\geq -\frac{M}{n} - \sum_{i=1}^{n} i^{\eta} \|\hat{\theta}_{i-1} - \theta\| \int \|(x, Z_i)\|\, \mathrm{d}Q_{Z_i}(x) \\
&= -\frac{M}{n} - \frac{1}{n}\sum_{i=1}^{n} i^{\eta} \|\hat{\theta}_{i-1} - \theta\| \mathbb{E}[\|(X_i, Z_i)\| \mid Z_i] \\
&\geq -\frac{M}{n} - \left(\frac{1}{n}\sum_{i=1}^{n} i^{2\eta} \|\hat{\theta}_{i-1} - \theta\|^2\right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\|(X_i, Z_i)\| \mid Z_i]^2\right)^{1/2},
\end{aligned} \tag{D.4}$$

where

$$M = \sum_{i=1}^{\infty} 1\left\{\int p_\theta(Y_i \mid x, Z_i)\, \mathrm{d}Q_{Z_i}(x) \leq i^{-\eta}\right\} \frac{\|\hat{\theta}_{i-1} - \theta\| \int \|(x, Z_i)\|\, \mathrm{d}Q_{Z_i}(x)}{\int p_\theta(Y_i \mid x, Z_i)\, \mathrm{d}Q_{Z_i}(x)}$$

is the sum of $\log(r_i)$ over all almost surely finitely many i s.t. $\int p_\theta(Y_i \mid x, Z_i)\, \mathrm{d}Q_{Z_i}(x) \leq i^{-\eta}$. Since (X_i, Z_i), $i \in \mathbb{N}$, are independent and identically distributed with

$$\mathbb{E}[\mathbb{E}[\|(X,Z)\||Z]^2] \leq \mathbb{E}[\|(X,Z)\|^2] < \infty,$$

the law of large numbers implies

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\|(X_i, Z_i)\| \mid Z_i]^2 \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\|(X_i, Z_i)\|^2 \mid Z_i] \to \mathbb{E}[\|(X,Z)\|^2] < \infty$$

almost surely as $n \to \infty$. On the other hand, $n^{2\eta}\|\hat{\theta}_{n-1} - \theta\|^2 = \mathcal{O}(n^{2\eta-1}\log(\log(n)))$ almost surely, so that for $\eta < 1/2$, we have $n^{2\eta}\|\hat{\theta}_{n-1} - \theta\|^2 \to 0$ almost surely as $n \to \infty$. Finally, since M only takes finite values, also $M/n \to 0$ for $n \to \infty$. Hence (D.4) converges to 0 almost surely. It follows that

$$\liminf_{n\to\infty} \frac{1}{n}\left(\log(S_n^{CI}) - \log\left(\prod_{i=1}^{n} \frac{p_\theta(Y_i \mid X_i, Z_i)}{\int p_\theta(Y_i \mid x, Z_i)\, \mathrm{d}Q_{Z_i}(x)}\right)\right) \geq 0$$

almost surely. Since

$$\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{p_\theta(Y_i \mid X_i, Z_i)}{\int p_\theta(Y_i \mid x, Z_i)\,\mathrm{d}Q_{Z_i}(x)}\right) \to I(X;Y \mid Z) > 0,\ n \to \infty,$$

almost surely, by the law of large numbers, this proves the theorem. □

D.3 Anytime-Valid E-Statistics

In this section, we discuss an alternative way to define anytime-valid tests using e-statistics and show that, in the setting of Chapter 6, this method coincides with the method discussed in Section 6.2.2. In Section 6.2.2, we mentioned that a sequence of conditional e-statistics gives rise to a test martingale $(S_n(D^n))_{n\in\mathbb{N}}$, which satisfies $\mathbb{E}_P[S_\tau(D^\tau)] \leq 1$ for any stopping time τ and $P \in \mathcal{H}_0$. Rather than taking the latter as a consequence, Koolen and Grünwald (2022) take this as the definition of what they call *anytime-valid* e-statistics. That is, they call a nonnegative process $(E_n(D^n))_{n\in\mathbb{N}}$ an anytime-valid e-statistic if $\mathbb{E}_P[E_\tau(D^\tau)] \leq 1$ for any stopping time τ and $P \in \mathcal{H}_0$. The same object is referred to as e-process in Ramdas et al. (2022), and it can be shown that the class of anytime-valid e-statistics (or e-processes) is strictly larger than the class of test martingales. A priori it is not obvious whether the GRO criterion, which maximizes the expected growth rate without referring to any particular stopping time, also yields powerful e-statistics when specific stopping rules τ are applied. Therefore, Koolen and Grünwald (2022) propose, for fixed alternative distribution $\mathcal{H}_1 = \{P^*\}$ and stopping time τ, to look for the anytime-valid e-statistic that maximizes

$$(E_n)_{n\in\mathbb{N}} \mapsto \mathbb{E}_{P^*}[\log E_\tau(D^\tau)]. \tag{D.5}$$

It turns out that there are settings in which the optimal anytime-valid e-statistic is actually equal to the GRO test martingale. One of the settings in which this happens is given in their Theorem 12. We present a slightly rephrased version of this theorem here.

Theorem D.1 (Koolen and Grünwald (2022)). *Assume that the data is given by an i.i.d. stream $(D_i)_{i\in\mathbb{N}}$ and that the alternative is given by $\mathcal{H}_1 = \{P^*\}$, where P^* admits a density p^*. Suppose further that the GRO e-statistic is given by the likelihood ratio p^*/q, where q is the density of an element of $\mathcal{H}_0$. Then the process $(p^*(D_i)/q(D_i))_{i\in\mathbb{N}}$ also maximizes* (D.5) *for any stopping time τ.*

In the proof of our Theorem 6.3 (see Section D.2.3), we show that the GRO e-variable is exactly of the form described in Theorem D.1. It therefore follows that the test martingale that we give in (6.7) is actually also the optimal anytime-valid e-statistic. We therefore chose to focus on the GRO property in Chapter 6.

E | Appendix to Chapter 7

E.1 Proofs

E.1.1 Proof of Proposition 7.4

Proposition 7.4. For $X^{n-1} \in \mathcal{X}^{n-1}$ and $X_n \in \mathcal{X}$, define $F_n(\gamma_{n-1}(X^{n-1}), X_n) = \gamma_n((\gamma_{n-1}(X^{n-1}), X_n))$, where, with a slight abuse of notation, we use $(\gamma_{n-1}(X^{n-1}), X_n)$ to refer to the concatenation of $\gamma_{n-1}(X^{n-1})$ and X_n. We will show that F_n has the claimed properties. First, we will show that the vectors $(\gamma_{n-1}(X^{n-1}), X_n)$ and X^n are in the same orbit, so that also $\gamma_n((\gamma_{n-1}(X^{n-1}), X_n)) = \gamma_n(X^n)$. To this end, let $g' \in G_{n-1}$ denote the group element such that $g'X^{n-1} = \gamma_{n-1}(X^{n-1})$. Then it holds that

$$\begin{aligned}\{g(\gamma_{n-1}(X^{n-1}), X_n) : g \in G_n\} &= \{g(g'X^{n-1}, X_n) : g \in G_n\} \\ &= \{g\imath_n(g')X^n : g \in G_n\} \\ &= \{gX^n : g \in G_n\},\end{aligned}$$

where we used (iii) of Definition 7.2 for the second equality and called X^n the concatenation of X^{n-1} and X_n. This shows the first claim. For the second claim, that $F_n(\,\cdot\,, X_n)$ is one-to-one for each fixed X_n, we show that we can reconstruct $\gamma_{n-1}(X^{n-1})$ from X_n and $\gamma_n(X^n)$.

Pick any $g_{X_n} \in G_n$ such that $(g_{X_n}\gamma_n(X^n))_n = X_n$. We furthermore know that there exists some $g \in G_n$ such that $gX^n = \gamma_n(X^n)$. Note that $g_{X_n}g$ does nothing to the final coordinate of X^n, so by item (iii) of Definition 7.2 there is a $g^*_{n-1} \in G_{n-1}$

such that $g_{X_n} g X^n = \imath(g^*_{n-1}) X^n$. Then we see

$$\begin{aligned}\{\imath(g_{n-1}) g_{X^n} \gamma_n(X^n) : g_{n-1} \in G_{n-1}\} &= \{\imath(g_{n-1}) g_{X^n} g X^n : g_{n-1} \in G_{n-1}\} \\ &= \{\imath(g_{n-1}) \imath(g^*_{n-1}) X^n : g_{n-1} \in G_{n-1}\} \\ &= \{\imath(g_{n-1}) X^n : g_{n-1} \in G_{n-1}\}.\end{aligned}$$

It follows from item (ii) of Definition 7.2 that $G_{n-1}\mathrm{proj}_{n-1}(g_{X_n}\gamma_n(X^n)) = G_{n-1}X^{n-1}$. It therefore follows that $\gamma_{n-1}(\mathrm{proj}_{n-1}(g_{X_n}\gamma_n(X^n))) = \gamma_{n-1}(X^{n-1})$. □

E.1.2 Proof of Theorem 7.8

Theorem 7.8. The proof can be divided in two main steps: (1) to show that, conditionally on $\gamma_n(X^n)$, R_n is uniformly distributed for each n and (2) to show that $R_1, R_2, \ldots$ are also independent. The second step is completely analogous to the proof of Theorem 3 by Vovk (2002). For each n, define the σ-algebra $\mathcal{G}_n = \sigma(\gamma_n(X^n), X_{n+1}, X_{n+2}, \ldots)$. Notice that $\mathcal{G}_n$ contains—among others—all G_n-invariant functions of X^n because γ_n is a maximally invariant function of X^n—any other G_n-invariant function of X^n is a function of $\gamma_n(X^n)$. Let $g' \in G_n$ such that $\gamma_n(X^n) = g'X^n$, then we have that $\{g \in G_n : A((gX^n)_n, \gamma_n(X^n))_n < \alpha_n\} = \{g \in G_n : A((g\gamma_n(X^n)_n, \gamma_n(X^n))_n < \alpha_n\}g'$. Here, we define $Bg = \{bg : b \in B\}$ for a subset $B \subseteq G_n$. By the invariance of μ_n—it is the Haar probability measure—, it follows that

$$\begin{aligned}&\mu_n(\{g \in G_n : A((gX^n)_n, \gamma_n(X^n))_n < \alpha_n\}) \\ &= \mu_n(\{g \in G_n : A((g\gamma_n(X^n))_n, \gamma_n(X^n))_n < \alpha_n\}).\end{aligned}$$

An analogous identity can be derived for the second term in (7.3). We have $\alpha_n \mid \mathcal{G}_n \overset{\mathcal{D}}{=} A((U\gamma_n(X^n))_n, \gamma_n(X^n))_n \mid \mathcal{G}_n$.

We will denote $F(b) := \mu(\{g \in G_n : A((g\gamma_n(X^n))_n, \gamma_n(X^n))_n < b\})$ and define $G(\delta) = \sup\{b \in \mathbb{R} : F(b) \leq \delta\}$. If $\alpha_n \mid \mathcal{G}_n$ is continuous, then F is the CDF of that distribution, otherwise it is the CDF minus the probability of equality. In any case, F is is increasing and right-continuous. For any $\delta \in (0,1)$, we have that $F(G(\delta)) = \delta'$ for some $\delta' \leq \delta$, with equality if F is continuous in $G(\delta)$. Then we can write

$$\mathbb{P}(R_n \leq \delta \mid \mathcal{G}_n) = \mathbb{P}(R_n \leq \delta' \mid \mathcal{G}_n) + \mathbb{P}(\delta' < R_n \leq \delta \mid \mathcal{G}_n). \tag{E.1}$$

For any $\theta \in (0,1]$, we have that $R_n = F(\alpha_n) + \theta(F(\alpha_n^+) - F(\alpha_n)) \leq \delta'$ if and only if either $F(\alpha_n) < \delta'$ or $F(\alpha_n^+) - F(\alpha_n) = 0$, which happens precisely when $\alpha_n < G(\delta)$.

We therefore see

$$\mathbb{P}(R_n \leq \delta' \mid \mathcal{G}_n) = \mathbb{P}(\alpha_n < G(\delta') \mid \mathcal{G}_n) = F(G(\delta')) = \delta'.$$

If F is continuous in $G(\delta)$, then this shows that $\mathbb{P}(R_n \leq \delta \mid \mathcal{G}_n) = \delta$, since $\delta' = \delta$ in that case. If F is not continuous in $G(\delta)$, then we have that

$$\mathbb{P}(\delta' < R_n \leq \delta \mid \mathcal{G}_n) = \mathbb{P}(\delta' < F(\alpha_n) + \theta(F(\alpha_n^+) - F(\alpha_n)) \leq \delta \mid \mathcal{G}_n).$$

Notice that $\delta' < F(\alpha_n) + \theta(F(\alpha_n^+) - F(\alpha_n)) \leq \delta$ if and only if $\alpha_n = G(\delta)$ and $\theta < (\delta - \delta')/(F(\alpha_n^+) - F(\alpha_n))$, so that we can write

$$\begin{aligned}\mathbb{P}(\delta' < R_n \leq \delta \mid \mathcal{G}_n) &= \mathbb{P}(\alpha_n = G(\delta) \mid \mathcal{G}_n)\mathbb{P}\left(\theta \leq \frac{\delta - \delta'}{F(G(\delta')^+) - F(G(\delta'))} \mid \mathcal{G}_n\right) \\ &= (F(G(\delta')^+) - F(G(\delta')))\frac{\delta - \delta'}{(F(G(\delta')^+) - F(G(\delta')))} \\ &= \delta - \delta'.\end{aligned}$$

Putting everything together, we see that $\mathbb{P}(R_n \leq \delta \mid \mathcal{G}_n) = \delta$. This shows the first part, that R_n has a conditional uniform distribution on $[0, 1]$.

For the second part of the proof, we show that the sequence $R_1, R_2, \ldots$ is also an independent sequence. We have that R_n is $\mathcal{G}_{n-1}$-measurable because it is invariant under transformations of the form $X^n \mapsto (gX^{n-1}, X_n)$ for $g \in G_{n-1}$ (see also Vovk, 2004, Lemma 2). We proceed (implicitly) by induction:

$$\begin{aligned}\mathbb{P}(R_n \leq \delta_n, \ldots, R_1 \leq \delta_1 \mid \mathcal{G}_n) &= \mathbf{E}\left[\mathbf{1}\left\{R_n \leq \delta_n, \ldots, R_1 \leq \delta_1\right\} \mid \mathcal{G}_n\right] \\ &= \mathbf{E}\left[\mathbf{E}\left[\mathbf{1}\left\{R_n \leq \delta_n, \ldots, R_1 \leq \delta_1\right\} \mid \mathcal{G}_{n-1}\right] \mid \mathcal{G}_n\right] \\ &= \mathbf{E}\left[\mathbf{1}\left\{R_n \leq \delta_n\right\}\mathbf{E}\left[\mathbf{1}\left\{p_{n-1} \leq \delta_{n-1}, \ldots, R_1 \leq \delta_1\right\} \mid \mathcal{G}_{n-1}\right] \mid \mathcal{G}_n\right] \\ &= \mathbf{E}\left[\mathbf{1}\left\{R_n \leq \delta_n\right\}\right]\delta_{n-1} \cdots \delta_1 \\ &= \delta_n \cdots \delta_1.\end{aligned}$$

It follows by the law of total expectation that

$$\mathbb{P}(R_n \leq \delta_n, \ldots, R_1 \leq \delta_1) = \delta_n \cdots \delta_1,$$

which shows that $R_1, R_2, \ldots, R_n$ are independent and uniformly distributed on $[0, 1]$ for any $n \in \mathbb{N}$. This implies that the distribution of $R_1, R_2, \ldots$ coincides with U^∞

by Kolmogorov's extension theorem (see e.g. Shiryaev, 2016, Theorem II.3.3). This shows the claim of the theorem. □

E.1.3 Proof of Proposition 7.9

The proof of Proposition 7.9 follows directly from Lemma E.1. It states that, with probability one, enough of the original data can be recovered using the smoothed ranks and the orbit representative. We state Lemma E.1, prove Proposition 7.9 and then prove Lemma E.1.

Lemma E.1. *Suppose, for each $n \in \mathbb{N}$, that $A(\,\cdot\,, \gamma_n(X^n))$ is a one-to-one function of X_n, then there exists a map $D_n : [0,1]^n \times \mathcal{X}^n \to [0,1]^n \times \mathcal{X}^n$ s.t. for any $Q \in \mathcal{H}_0$, $\widetilde{Q}(D_n(R^n, \gamma_n(X^n)) = (\widetilde{\theta}^n, X^n)) = 1$. Here, $\widetilde{\theta}^n = (\widetilde{\theta}_n)_{n\in\mathbb{N}}$ is the sequence given by $\widetilde{\theta}_n = \theta_n \mathbf{1}\{\mu_n(\{g \in G_n : A((gX^n)_n, \gamma_n(X^n))_n = \alpha_n\}) \neq 0\}$.*

Proposition 7.9. Consider, without loss of generality, the case that $A(X^n, \gamma_n(X^n)) = X_n$. Because of the independence of R_n and γ_n under P and the assumption that the marginal distribution of γ_n under Q^* and under P are equal, $M_n = \frac{\mathrm{d}\widetilde{P}(R^n,\gamma_n(X^n))}{\mathrm{d}\widetilde{Q}^*(R^n,\gamma_n(X^n))}$. Using the sequence of functions $(D_n)_{n\in\mathbb{N}}$ from Lemma E.1 and that the external randomization is independent of X^n, the claim follows. □

Lemma E.1. As in the proof of Theorem 7.8, we will denote $F(b) = \mu_n(\{g \in G_n : A((gX^n)_n, \gamma_n(X^n))_n < b\})$ and define $G(\delta) = \sup\{b \in \mathbb{R} : F(b) \leq \delta\}$. Furthermore, we will write $\mathbb{P}_{\alpha_n|\gamma_n(X^n)}$ for the distribution of α_n given $\gamma_n(X^n)$ and denote its support by

$$\operatorname{supp}(\mathbb{P}_{\alpha_n|\gamma_n(X^n)}) := \{x \in \mathbb{R} \mid \text{for all } I \text{ open, if } x \in I \text{ then } \mathbb{P}_{\alpha_n|\gamma_n(X^n)}(I) > 0\},$$

If $b \in \operatorname{int}(\operatorname{supp}(\mathbb{P}_{\alpha_n|\gamma_n(X^n)}))$, then there exists an open interval B with $b \in B$ and $B \subseteq \operatorname{supp}(\mathbb{P}_{\alpha_n|\gamma_n(X^n)}))$. For all $c \in B$ with $c > b$, we have that $F(c) - F(b) = \mathbb{P}_{\alpha_n|\gamma_n(X^n)}([b,c)) > 0$, since $[b,c)$ contains an open neighborhood of an interior point of the support. It follows that $F(c) > F(b)$. In words, there are no points c to the right of b such that $F(c) > F(b)$. Consequently, we have

$$G(F(b)) = \sup\{a \in \mathbb{R} : F(a) \leq F(b)\} = b.$$

In a similar fashion, we can conclude that the same identity holds whenever $b \in \operatorname{supp}(\mathbb{P}_{\alpha_n|\gamma_n(X^n)}) \backslash \operatorname{int}(\operatorname{supp}(\mathbb{P}_{\alpha_n|\gamma_n(X^n)}))$. Notice furthermore that $G(R_n) = G(F(\alpha_n) +$

$\theta_n(F(\alpha_n^+) - F(\alpha_n))) = G(F(\alpha_n))$ whenever $\theta_n < 1$, which happens with probability one. Together with the fact that $\mathbb{P}_{\alpha_n|\gamma_n(X^n)}(\mathrm{supp}(\mathbb{P}_{\alpha_n|\gamma_n(X^n)})) = 1$, this gives $\mathbb{P}_{\alpha_n|\gamma_n(X^n)}(G(R_n) = \alpha_n) = 1$, so also $\mathbb{P}(G(R_n) = \alpha_n) = 1$. If $(F(G(R_n)^+) - F(G(R_n))) = \mu_n(\{g \in G_n : (g\alpha^n)_n = \alpha_n\}) = 0$, set $\widetilde{\theta}_n = 0$. If $\mu_n(\{g \in G_n : (g\alpha^n)_n = \alpha_n\}) > 0$, then it follows that $\mathbb{P}(\theta_n = (R_n - F(G(R_n)))/(F(G(R_n)^+) - F(G(R_n)))) = 1$, so set $\widetilde{\theta}_n = (R_n - F(G(R_n)))/(F(G(R_n)^+) - F(G(R_n)))$. Since $A(\cdot, \gamma_n(X^n))$ is one-to-one by assumption, its inverse maps α_n to X_n. By Proposition 7.4, there also exists a map from X_n and $\gamma_n(X^n)$ to $\gamma_{n-1}(X^{n-1})$. At this point, we can repeat the procedure above to recover X_{n-1} from $(R_{n-1}, \gamma_{n-1}(X^{n-1}))$, from which we can then recover $\gamma_{n-2}(X^{n-2})$, etc. Together, all of the maps involved give the function as in the statement of the proposition. □

E.1.4 Proof of Theorem 7.10

Theorem 7.10. We first show (7.6). Assume that $\widetilde{P}$ is such that $R^n \perp \gamma_n(X^n)$ for all n. Let Q^* denote the distribution under which the marginal of $\gamma_n(X^n)$ coincides with that under P, and such that $X^n \mid \gamma_n(X^n) \overset{\mathcal{D}}{=} U\gamma_n(X^n) \mid \gamma_n(X^n)$, where $U \sim \mu_n$ is uniform on G_n and independent from $\gamma_n(X^n)$. First note that

$$\begin{aligned}
\widetilde{Q}^*\left(\prod_{i=1}^{\tau} f_i(R_i) = \frac{\mathrm{d}P}{\mathrm{d}Q^*}(X^\tau)\right) &\geq \widetilde{Q}^*\left(\forall t : \prod_{i=1}^{t} f_i(R_i) = \frac{\mathrm{d}P}{\mathrm{d}Q^*}(X^t)\right) \\
&= 1 - \widetilde{Q}^*\left(\exists t : \prod_{i=1}^{t} f_i(R_i) \neq \frac{\mathrm{d}P}{\mathrm{d}Q^*}(X^t)\right) \\
&= 1 - \widetilde{Q}^*\left(\bigcup_{t=1}^{\infty}\left\{\prod_{i=1}^{t} f_i(R_i) \neq \frac{\mathrm{d}P}{\mathrm{d}Q^*}(X^t)\right\}\right) \\
&\geq 1 - \sum_{t=1}^{\infty}\widetilde{Q}^*\left(\left\{\prod_{i=1}^{t} f_i(R_i) \neq \frac{\mathrm{d}P}{\mathrm{d}Q^*}(X^t)\right\}\right) = 1.
\end{aligned}$$

In the last inequality, we used Lemma E.1. By assumption, we have $\widetilde{P} \ll \widetilde{Q}^*$, so we also have $\widetilde{P}\left(\prod_{i=1}^{\tau} f_i(R_i) = \frac{\mathrm{d}P}{\mathrm{d}Q^*}(X^\tau)\right) = 1$. We have shown that M_τ is a modification of the likelihood ratio evaluated at X^τ. We now show that the latter is optimal.

Denote $\ell_n = \frac{\mathrm{d}P}{\mathrm{d}Q^*}(X^n)$ and let $f(\alpha) = \mathbf{E}_{\widetilde{P}}[\ln((1-\alpha)\ell_\tau + \alpha E'_\tau)]$; a concave function. We will show that the derivative of f in 0 is negative, which implies that f

attains its maximum in $\alpha = 0$. This in turn implies our claim. Indeed,

$$\begin{aligned} f'(0) &= \mathbf{E}_{\widetilde{P}}\left[\frac{E'_\tau - \ell_\tau}{\ell_\tau}\right] \\ &= \sum_{i=1}^{\infty} \mathbf{E}_{\widetilde{P}}\left[\frac{E'_i}{\ell_i}\mathbf{1}\{\tau = i\}\right] - 1 \\ &= \sum_{i=1}^{\infty} \mathbf{E}_{\widetilde{Q}^*}\left[E'_i \mathbf{1}\{\tau = i\}\right] - 1 \\ &= \mathbf{E}_{\widetilde{Q}^*}\left[E'_\tau\right] - 1 \leq 0, \end{aligned}$$

where we use that differentiation and integration can be interchanged, because

$$|f'(\alpha)| = \left|\frac{E'_\tau - \ell_\tau}{(1-\alpha)\ell_\tau + \alpha E'_\tau}\right| \leq \max\left\{\frac{1}{1-\alpha}, \frac{1}{\alpha}\right\},$$

so that the dominated convergence theorem is applicable. Finally, this gives that $\mathbf{E}_{\widetilde{P}}[\ln \prod_{i=1}^{\tau} f(R_i)] = \mathbf{E}_{\widetilde{P}}[\ln E'_\tau] \geq \mathbf{E}_{\widetilde{P}}[\ln E'_\tau]$. The proof of (7.5) follows from the same argument, but using $\ell'_n = \frac{\mathrm{d}P}{\mathrm{d}Q^*}(R^n)$. □

E.2 Linear Models and Isotropy Groups

The rotational symmetry described in Section 7.5.2 is that of symmetry around the origin, which we argued is equivalent to testing whether $X_i \sim \mathcal{N}(0, \sigma)$ for some $\sigma \in \mathbb{R}^+$. Of course, there are many applications where it is not reasonable to assume that the data is zero-mean and it is more interesting to test whether the data is spherically symmetric around some point other than the origin. One particular instance of such noncentered sphericity is to test whether, for each n, the data can be written as $X^n = \mu \mathbf{1}_n + \epsilon^n$, where $\mu \in \mathbb{R}$, the error ϵ^n is spherically symmetric and $\mathbf{1}_n$ is the n-vector of all ones. If μ is known, we can test for spherical symmetry of $X^n - \mu\mathbf{1}_n$ under $\mathrm{O}(n)$ and the problem reduces to that of the previous section. It is still possible treat the more realistic case where μ is unknown because the null model is still symmetric under a family of rotations. Notice the following: for any $O_n \in \mathrm{O}(n)$ it holds that $O_n X^n = \mu O_n \mathbf{1}_n + O_n \epsilon^n$. Unless $\mu = 0$, it follows that $X^n \overset{\mathcal{D}}{=} O_n X^n$ every time that $O_n \mathbf{1}_n = \mathbf{1}_n$. That is, the null distribution of X^n is invariant under the isotropy group of $\mathbf{1}_n$, i.e. $G_n = \{O_n \in \mathrm{O}(n) : O_n \mathbf{1}_n = \mathbf{1}_n\}$. Invariance under the action of G_n has previously appeared in the literature as centered spherical symmetry (Smith, 1981). Through the lens of test martingales, testing sequentially for centered spherical symmetry is equivalent to testing whether the data was generated by any Gaussian. This holds because any probability distribution on $\mathbb{R}^\infty$ for which the marginal of the first n coordinates is centered spherically symmetric for any n can be written as a mixture of Gaussians (Smith, 1981; Eaton, 1989, Theorem 8.13).

Using some geometry, a test is readily obtained. Note that we can write $X^n = X^n_{\mathbf{1}_n} + X^n_{\perp\mathbf{1}_n}$, where $X^n_{\mathbf{1}_n} = \frac{\langle X^n, \mathbf{1} \rangle}{n}\mathbf{1}_n$ is the projection of X^n onto the span of $\mathbf{1}_n$, and $X^n_{\perp\mathbf{1}_n}$ the projection onto its orthogonal complement. We have that $gX^n = X^n_{\mathbf{1}_n} + gX^n_{\perp\mathbf{1}_n}$ for any $g \in G_n$. Consequently, the orbit of X^n under G_n is given by the intersection of $S^{n-1}(\|X^n\|)$ and the hyperplane $H_n(X^n)$ defined by $H_n(X^n) = \{x'^n \in \mathbb{R}^n : \langle x'^n, \mathbf{1}_n \rangle = \langle X^n, \mathbf{1}_n \rangle\}$. There is a unique line that is perpendicular to $H_n(X^n)$ and passes through the origin $0_n = (0, \dots, 0)$; it intersects $H_n(X^n)$ in the point $0_{H_n} := \frac{\langle X^n, \mathbf{1}_n \rangle}{n}\mathbf{1}_n$. For any $x'^n \in S^{n-1}(\|X^n\|) \cap H_n(X^n)$, Pythagoras' theorem gives that $\|x'^n - 0_{H_n}\|^2 = \|X^n\|^2 - \|0_{H_n} - 0_n\|^2$. In other words, $S^{n-1}(\|X^n\|) \cap H_n(X^n)$ forms an $(n-2)$-dimensional sphere of radius $(\|X^n\|^2 - \|0_{H_n} - 0_n\|^2)^{1/2}$ around 0_{H_n}. If one considers the projection of this sphere on the n-th coordinate, then the highest possible value is given by $\|X^n\|$, and the lowest value therefore by $\frac{\langle X^n, \mathbf{1}_n \rangle}{n} - \frac{1}{2}(\|X^n\| - \frac{\langle X^n, \mathbf{1}_n \rangle}{n})$. The relative value of X_n is therefore given by $\widetilde{X}_n := X_n - \frac{\langle X^n, \mathbf{1}_n \rangle}{n} + \frac{1}{2}(\|X^n\| - \frac{\langle X^n, \mathbf{1}_n \rangle}{n})$. As a result, R_n is the relative surface area of the $(n-2)$-dimensional hyper-spherical cap with co-latitude angle $\varphi = \pi - \cos^{-1}(\widetilde{X}_n / (\|X^n\|^2 - \|0_{H_n} - 0_n\|^2)^{1/2})$, so that equation (7.9) can again be used to determine R_n. With this construction, we recover what Vovk (2023) refers to as the "full Gaussian model", which is an online compression model that is defined in terms of the summary statistic $\sigma_n = (\langle X^n, \mathbf{1}_n \rangle, \|X^n\|)$.

This model can be extended to the case in which there are covariates, i.e. $X_n = (Y_n, Z_n^d)$ for some $Y_n \in \mathbb{R}$ and $Z_n^d \in \mathbb{R}^d$. Denote Z_n for the matrix with row-vectors Z_n^d and, as is a standard assumption in regression, assume that Z_n is full rank for every n. The model of interest is $Y^n = Z_n\beta + \epsilon^n$ where $\beta \in \mathbb{R}^d$ and ϵ^n is spherically symmetric for each n. Similar to the reasoning above, this model is invariant under the intersection of the isotropy groups of the column vectors of Z_n, i.e. $G_n = \{O_n \in \mathrm{O}(n) : O_n Z_n = Z_n\}$. The orbit of X^n under G_n is given by the intersection of $S^{n-1}(\|X^n\|)$ with the intersection of the d hyperplanes defined by the columns of Z_n, so that for $\alpha^n(Y^n, Z_n) = Y^n$, computing R_n is analogous. Interestingly, however, it does not always hold that testing for invariance under G_n is equivalent to testing for normality with mean $Z_n\beta^d$. A sufficient condition for the equivalence to hold is that $\lim_{n\to\infty}(Z_n' Z_n)^{-1} = 0$, which is essentially the condition that the parameter vector β can be consistently estimated by means of least squares (Eaton, 1989, Section 9.3).

F | Appendix to Chapter 8

F.1 Invariance and Sufficiency

The relationship between invariance and sufficiency has been thoroughly investigated (Hall et al., 1965, 1995; Berk, 1972; Nogales and Oyola, 1996). Consider a G-invariant hypothesis testing problem such that a sufficient statistic is available. If the action of G on the original data space induces a free action on the sufficient statistic—that is, if the sufficient statistic is equivariant—, there must be a maximally invariant function of the sufficient statistic. With this structure in mind, the results presented thus far suggest two approaches for solving the hypothesis testing problem. The first is to reduce the data using the sufficient statistic, and to test the problem using the maximally invariant function of the sufficient statistic. The second approach is to use the maximally invariant function of the original data. These two approaches yield two potentially different growth-optimal e-statistics, and one question arises naturally: are both approaches equivalent? In this section we show that this is indeed the case, under certain conditions.

We now introduce the setup formally. At the end of this section we revisit our guiding example, the t-test, and show how the results of this section apply to it. Let Θ be the parameter space, and let $\delta = \delta(\theta)$ be a maximally invariant function of θ for the action of G on Θ. Let $s_n : \mathcal{X}^n \to \mathcal{S}_n$ be a sufficient statistic for $\theta \in \Theta$. Consider again the hypothesis testing problem in the form presented in (8.1). Assume further that G acts freely and continuously on the image space $\mathcal{S}_n$ of the sufficient statistic $S_n = s_n(X^n)$. Denote by $(g, s) \mapsto gs$ the action of G on $\mathcal{S}_n$. We assume that s_n is equivariant, that is, s_n is compatible with the action of G in the sense that, for any $X^n \in \mathcal{X}^n$ and any $g \in G$, the identity $gs_n(X^n) = s_n(gX^n)$ holds. Let $M_{\mathcal{X},n} = m_{\mathcal{X},n}(X^n)$ and $M_{\mathcal{S},n} = m_{\mathcal{S},n}(S_n)$ be two maximally invariant functions for the actions of G on $\mathcal{X}^n$ and $\mathcal{S}_n$, respectively. Because of their invariance, the distributions of $M_{\mathcal{X},n}$ and $M_{\mathcal{S},n}$ depend only on the maximally invariant parameter δ. Hall et al. (1965, Section II.3) proved that, under regularity conditions, if $S_{\mathcal{X},n} = s_{\mathcal{X},n}(X^n)$ is sufficient for $\theta \in \Theta$, then the statistic $M_{\mathcal{S},n} = m_{\mathcal{S},n}(s_n(X^n))$ is sufficient for δ. In that case, we call $M_{\mathcal{S},n}$ invariantly sufficient. Here we state the version of their result, attributed by Hall et al. (1965) to C. Stein, that suits best our purposes (see Remark F.1).

Theorem F.1 (C. Stein). *If there exists a right Haar measure on the group G and G is σ-finite, the statistic $M_{\mathcal{S},n} = m_{\mathcal{S},n}(s_n(X^n))$ is invariantly sufficient, that is, it is sufficient for the maximally invariant parameter δ.*

With this theorem at hand, and the fact that the KL divergence does not decrease by the application of sufficient transformations, we obtain the following proposition.

Proposition F.2. *Let $s_n : \mathcal{X}^n \to \mathcal{S}_n$ be sufficient statistic for $\theta \in \Theta$. Assume that G acts freely on $\mathcal{S}_n$ and that $s_n(gX^n) = gs_n(x^n)$ for all $X^n \in \mathcal{X}^n$ and $g \in G$. Let $m_{\mathcal{S},n}$ be a maximal invariant for the action of G on $\mathcal{S}_n$, and let $M_{\mathcal{S},n} = m_{\mathcal{S},n}(s_n(X^n))$. Then,*

$$\mathrm{KL}\left(\mathbf{P}_{\delta_1}^{M_{\mathcal{X},n}}, \mathbf{P}_{\delta_0}^{M_{\mathcal{X},n}}\right) = \mathrm{KL}\left(\mathbf{P}_{\delta_1}^{M_{\mathcal{S},n}}, \mathbf{P}_{\delta_0}^{M_{\mathcal{S},n}}\right).$$

Proof. The function $M_{\mathcal{S},n} = m_{\mathcal{S},n}(s_n(X^n))$ is invariant, and consequently its distribution only depends on the maximally invariant parameter δ. Since $M_{\mathcal{X},n}$ is maximally invariant for the action of G on $\mathcal{X}^n$, there is a function f such that $M_{\mathcal{S},n} = f(M_{\mathcal{X},n})$. By Stein's theorem, Theorem F.1, $M_{\mathcal{S},n}$ is sufficient for δ. Consequently, f is a sufficient transformation. Hence, from the invariance of the KL divergence under sufficient transformations, the result follows. □

Via the factorization theorem of Fisher and Neyman, the likelihood ratio for the maximal invariant $M_{\mathcal{X},n}$ coincides with that of the invariantly sufficient $M_{\mathcal{S},n}$. As a consequence, we obtain the answer to the motivating question of this section: performing an invariance reduction on the original data and on the sufficient statistic are equivalent.

Corollary F.3. *Under the assumptions of Proposition F.2, if $S_n = s_n(X^n)$,*

$$\frac{q^{M_{\mathcal{X},n}}(m_{\mathcal{X},n}(X^n))}{p^{M_{\mathcal{X},n}}(m_{\mathcal{X},n}(X^n))} = \frac{q^{M_{\mathcal{S},n}}(m_{\mathcal{S},n}(S_n))}{p^{M_{\mathcal{S},n}}(m_{\mathcal{S},n}(S_n))}.$$

Hence, if assumptions of Corollary 8.3 also hold, the likelihood ratio for the invariantly sufficient statistic $M_{\mathcal{S},n}$ is (relatively) GROW.

Example F.1 (continues=ex:t-test)**.** We have seen that a maximally invariant function of the data is $M_{\mathcal{X},n} = m_{\mathcal{X},n}(X^n) = (X_1/\left|X_1\right|, \ldots, X_n/\left|X_1\right|)$ while the t-statistic $M_{\mathcal{S},n} = m_{\mathcal{S},n}(X^n) \propto \hat{\mu}_n/\hat{\sigma}_n$ is a maximally invariant function of the sufficient statistic $s_n(X^n) = (\hat{\mu}_n, \hat{\sigma}_n)$. Stein's theorem (Theorem F.1) shows that the t-statistic $M_{\mathcal{S},n}$ is sufficient for the maximally invariant parameter $\delta = \mu/\sigma$. Corollary F.3 shows that the likelihood ratio for the t-statistic is relatively GROW.

Remark. In the present form, the assumptions in Theorem F.1 avoid issues that may arise with almost-invariant functions (see Lehmann and Romano, 2005, Section 6.5). Almost-invariant functions are functions that are invariant under the action of a group almost surely up to a null set that may depend on the group element in question. Under the assumptions in Theorem F.1, every almost invariant function is equivalent to an invariant one (Lehmann and Romano, 2005, Theorem 6.5.1). In turn, the assumptions in Theorem F.1 are implied by Assumption 8.1, so that the same is true in the general setting of Chapter 8. See also Hall et al. (1965, discussion in p. 581).

F.2 Detailed Comparison to Sun and Berger (2007) and Liang and Barron (2004)

As the example in Section 8.5.1 illustrates, it is sometimes possible to represent the same $\mathcal{H}_0$ and $\mathcal{H}_1$ via (at least) two different groups, say G_a and G_b. Group G_a is combined with parameter of interest in some space Δ_a and priors $\mathbf{\Pi}_j^{*\delta_a}$ on Δ_a achieving (8.18) relative to group G_a, for $j = 0, 1$; group G_b has parameter of interest in Δ_b and priors $\mathbf{\Pi}_j^{*\delta_b}$ achieving (8.18) relative to group G_b; yet the tuples $\mathcal{T}_a = (G_a, \Delta_a, \{\mathbf{\Pi}_j^{*\delta_a}\}_{j=0,1})$ and $\mathcal{T}_b = (G_b, \Delta_b, \{\mathbf{\Pi}_j^{*\delta_b}\}_{j=0,1})$ define the same hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$. That is, the set of distributions $\{\mathbf{P}_g^*\}_{g \in G_a}$ obtained by applying Proposition 8.7 with group G_a (representing $\mathcal{H}_0$ defined relative to group G_a) coincides with the set of distributions $\{\mathbf{P}_g^*\}_{g \in G_b}$ obtained by applying Proposition 8.7 with group G_b (representing $\mathcal{H}_0$ defined relative to group G_b); and analogously for the set of distributions $\{\mathbf{P}_g^*\}_{g \in G_a}$ and the set of distributions $\{\mathbf{P}_g^*\}_{g \in G_b}$. In the example, G_a was GL(d) and the priors $\mathbf{\Pi}_0^{*\delta_a}, \mathbf{\Pi}_1^{*\delta_a}$ were degenerate priors on 0 and γ as in (8.23), respectively; G_b was the lower triangular group with a specific prior as indicated in the example. In such a case with multiple representations of the same $\mathcal{H}_0$ and $\mathcal{H}_1$, using the fact that the notion of "GROW" does not refer to the underlying group, Corollary 8.8 can be used to identify the GROW e-statistic as soon as the assumptions of Proposition 8.7 hold for at least one of the tuples $\mathcal{T}_a$ or $\mathcal{T}_b$. Namely, if the assumptions hold for just one of the two tuples, we use Corollary 8.8 with that tuple; then T^* as defined in the corollary must be GROW, irrespective of whether T^* based on the other tuple is the same (as it was in the example above) or different. If the assumptions hold for *both* groups, then, using the fact that the GROW e-statistic is essentially unique (see Theorem 1 of GHK for definition and proof), it follows that $T^*(X^n)$ as defined in Corollary 8.8 must coincide for both tuples.

Superficially, this may seem to contradict Sun and Berger (2007) who point out that in some settings, the right Haar prior is not uniquely defined, and different choices for right Haar prior give different posteriors. To resolve the paradox, note that, whereas we always formulate two models $\mathcal{H}_0$ and $\mathcal{H}_1$, Sun and Berger (2007) start with a single probabilistic model, say $\mathcal{P}$, that can be written as in (8.3) for some group G. Their example shows that the same $\mathcal{P}$ can sometimes arise from two different groups, and then it is not clear what group, and hence what Haar prior to pick, and their quantity of interest, the Bayesian posterior, can depend on the choice.

In contrast, our quantity of interest, the GROW e-statistic T_n^*, is uniquely defined as soon as there exists one group G with $\mathcal{H}_0$ and $\mathcal{H}_1$ as in (8.1) for which the assumptions of Theorem 8.2 hold; or more generally, as soon as there exists one tuple $\mathcal{T} = (G, \Delta, \{\mathbf{\Pi}_j^{*\delta}\}_{j=0,1})$ for which the assumptions of Proposition 8.7 hold, even if there exist other such tuples.

To reconcile uniqueness of the GROW e-statistic T_n^* with nonuniqueness of the Bayes posterior, note that the former is a ratio between Bayes marginals for different models $\mathcal{H}_0$ and $\mathcal{H}_1$ at the same sample size n. In contrast, the Bayes predictive distribution based on a single model $\mathcal{P}$ is a ratio between Bayes marginals for the same $\mathcal{P}$ at different sample sizes n and $n-1$. The role of 'same' and 'different' being interchanged, it turns out that this Bayes predictive distribution *can* depend on the group on which the right Haar prior for $\mathcal{P}$ is based. Since the Bayes predictive distribution can be rewritten as a marginal over the Bayes posterior for $\mathcal{P}$, it is then not surprising that this Bayes posterior may also change if the underlying group is changed.

The consideration of two families $\mathcal{H}_0$ and $\mathcal{H}_1$ vs. a single $\mathcal{P}$ is also one of the main differences between our setting and the one of Liang and Barron (2004), who provide exact min-max procedures for predictive density estimation for general location and scale families under Kullback-Leibler loss. Their results apply to any invariant probabilistic model $\mathcal{P}$ as in (8.3) where the invariance is with respect to location or scale (and more generally, with respect to some other groups including the subset of the affine group that we consider in Section 8.4.2). Consider then such a $\mathcal{P}$ and let $p^{M_n}(m_n(X^n))$ be as in (8.5). As is well-known, provided that n' is larger than some minimum value, for all $n > n'$, $r(X_{n'+1}, \ldots, X_n \mid X_1, \ldots, X_{n'}) := p^{M_n}(m_n(X^n))/p^{M_{n'}}(m_{n'}(X^{n'}))$ defines a conditional probability density for $X_{n'+1}, \ldots, X_n$; this is a consequence of the formal-Bayes posterior corresponding to the right Haar prior becoming proper after n' observations, a.s. under all $\mathbf{P} \in \mathcal{P}$. For example, in the t-test setting, $n' = 1$. Liang and Barron (2004) show that the distribution corresponding to r minimizes the $\mathbf{P}^{n'}$-expected KL divergence to the conditional distribution $\mathbf{P}^n \mid X^{n'}$, in the worst case over all $\mathbf{P} \in \mathcal{P}$. Even though their optimal density r is defined in terms of the same quantities as our optimal statistic T^*_n, it is, just as Berger and Sun (2008), considered above, a ratio between likelihoods for the same model at different sample sizes, rather than, as in our setting, between likelihoods for different models, both composite, at the same sample sizes. Our setting requires a joint KL minimization over two families, and therefore our proof techniques turn out quite different from their information- and decision-theoretic ones.

F.3 Anytime-Valid Testing Under Optional Stopping and Optional Continuation

Consider the setting of Section 8.2.2. Let $X = (X_n)_{n\in\mathbb{N}}$ be a random process, where each X_n is an observation that takes values on a space $\mathcal{X}$. Let $(M_n)_{n\in\mathbb{N}}$ be a sequence where, for each n, $M_n = m_n(X^n)$ is a maximally invariant function for the action of G on $\mathcal{X}^n$.

Suppose that data $X_1, X_2, \ldots$ are gathered one by one. Here, a sequential test is a sequence of zero-one-valued statistics $\xi = (\xi_n)_{n\in\mathbb{N}}$ adapted to the natural filtration generated by $X_1, X_2, \ldots$. We consider the test defined by $\xi_n = \mathbf{1}\left\{T^{M_n} \geq 1/\alpha\right\}$ for some value α. We note that Wald-style—Sequential Probability Ratio Tests—tests are different because they would output "no decision" until a particular sample size n. Afterwards, they would output 1 ("reject the null") or 0 ("there is no evidence to reject the null") forever. In contrast, in the present setting $\xi_n = 1$ means "if you stop now, for whatever reason, it is safe to reject the null". Below we prove the anytime validity of ξ. Additionally, we show that, for certain stopping times $\tau \leq \infty$, the optionally stopped e-statistic T^{M_τ} remains an e-statistic. This fact validates the use of the stopped T^{M_τ} for optional continuation because we can multiply the e-statistics T^{M_τ} across studies while retaining type-I error control. This result is not new and we add it merely for completeness; it follows by standard arguments as Ramdas et al. (2023) or GHK.

Proposition F.4. *Let $T^* = (T^{M_n})_{n \in \mathbb{N}}$, where, for each n, T^{M_n} is the likelihood ratio for the maximally invariant function $M_n = m_n(X^n)$ for the action of G on $\mathcal{X}^n$. Let $\xi = (\xi_n)_{n \in \mathbb{N}}$ be the sequential test given by $\xi_n = \mathbf{1}\left\{T^{M_n} \geq 1/\alpha\right\}$. Then, the following two properties hold:*

1. *The sequential test ξ is anytime valid at level α, that is,*

$$\text{for any random time } N, \ \sup_{\theta_0 \in \Theta_0} \mathbf{P}_{\theta_0}\left\{\xi_N = 1\right\} \leq \alpha.$$

2. *Suppose that $\tau \leq \infty$ is a stopping time with respect to the filtration induced by $M = (M_n)_{n \in \mathbb{N}}$. Then the optionally stopped e-statistic T^{M_τ} is also an e-statistic, that is,*

$$\sup_{\theta_0 \in \Theta_0} \mathbf{E}^{\mathbf{P}}_{\theta_0}[T^{M_\tau}] \leq 1. \tag{F.1}$$

It is natural to ask whether (F.1) also holds for stopping times that are adapted to the full data $(X^n)_{n \in \mathbb{N}}$ but not to the reduced $(M_n)_{n \in \mathbb{N}}$. In our t-test example, this could be a stopping time τ^* such as "$\tau^* := 1$ if $|X_1| \notin [a, b]$; $\tau^* = 2$ otherwise" for some $0 < a < b$. The answer is negative: after proving Proposition F.4, we show that, for appropriate choice of a and b, this τ^* is a counterexample. This means that such nonadapted τ^* cannot be safely used under optional continuation. However, using such a stopping time has no repercussions for optional stopping, since the time N in part 1 of the proposition above is not even required to be a stopping time—N is not restricted by the filtration induced by M and it is even allowed to depend on future observations.

Proof of Proposition F.4. From Proposition 8.6, we know that $T^* = (T^{M_n})_{n \in \mathbb{N}}$ is a nonnegative martingale with expected value equal to one. Let $\xi = (\xi_n)_n$ be the sequential test given by $\xi_n = \mathbf{1}\left\{T^{M_n} \geq 1/\alpha\right\}$. The anytime-validity at level α of ξ, is a consequence of Ville's inequality, and the fact that the distribution of each T^{M_n} does not depend on g. Indeed, these two, together, imply that

$$\sup_{g \in G} \mathbf{P}_g\{T^{M_n} \geq 1/\alpha \text{ for some } n \in \mathbb{N}\} \leq \alpha.$$

This implies the first statement. Now, let $\tau \leq \infty$ be a stopping time with respect to the filtration induced by M. If the stopping time τ is almost surely bounded, T^{M_τ} is an e-statistic by virtue of the optional stopping theorem. However, since T^* is a nonnegative martingale, Doob's martingale convergence theorem implies the existence of an almost sure limit T^*_∞. Even when τ might be infinite with positive probability, Theorem 4.8.4 of Durrett (2019) implies that T^{M_τ} is still an e-statistic. □

F.3.1 Importance of the Filtration for Randomly Stopped E-Statistics

Consider the t-test as in Example 8.1. Fix some $0 < a < b$, and define the stopping time $\tau^* := 1$ if $|X_1| \notin [a, b]$. $\tau^* = 2$ otherwise. Then τ^* is not adapted to (hence not a stopping time relative to) $(M_n)_n$ as defined in that example, since $M_1 \in \{-1, 1\}$ coarsens out all information in X_1 except its sign. Now let $\delta_0 := 0$ (so that $\mathcal{H}_0$ represents the normal distributions with mean $\mu = 0$ and arbitrary variance). Let $T_n^{*,\delta_1}(X^n)$ be equal to the GROW e-statistic $T^{M_n}(X^n)$ as in (8.6); here we make explicit its dependence on δ_1. For $\mathcal{H}_1$, to simplify computations, we put a prior $\tilde{\mathbf{\Pi}}_1^\delta$ on $\Delta_1 := \mathbb{R}$. We take $\tilde{\mathbf{\Pi}}_1^\delta$ to be a normal distribution with mean 0 and variance κ. We can now apply Corollary 8.9 (with prior $\tilde{\mathbf{\Pi}}_0^\delta$ putting mass 1 on $\delta = \delta_0 = 0$), which gives that $\tilde{T}_n = \tilde{t}_n(X^n)$ is an e-statistic, where

$$\tilde{t}_n(x^n) = \int \frac{1}{\sqrt{2\pi\kappa^2}} \exp\left(-\frac{\delta_1^2}{2\kappa^2}\right) \cdot T_n^{*,\delta_1}(x^n) \mathrm{d}\delta_1$$

coincides with a standard type of Bayes factor used in Bayesian statistics. By exchanging the integrals in the numerator, this expression can be calculated analytically. The Bayes factor $\tilde{T}_1$ for $x^1 = x_1$ is found to be equal to 1 for all $x_1 \neq 0$, and the Bayes factor for (x_1, x_2) is given by:

$$\tilde{T}_2 = \frac{\sqrt{2\kappa^2 + 1} \cdot (x_1^2 + x_2^2)}{\kappa^2(x_1 - x_2)^2 + (x_1^2 + x_2^2)}.$$

Now we consider the function

$$f(x) := \mathbf{E}_{X_2 \sim N(0,1)}[\tilde{t}_2(x, X_2)].$$

$f(x)$ is continuous and even. We want to show that, with τ^* as above, $\tilde{T}_{\tau^*}$ is not an E-variable for some specific choices of a, b and κ. Since, for any $\sigma > 0$, the null contains the distribution under which the X_i are i.i.d. $N(0,\sigma)$, the data may, under the null, in particular be sampled from $N(0,1)$. It thus suffices to show that

$$\mathbf{E}_{X_1,X_2\sim N(0,1)}[\tilde{T}_{\tau^*}] = \mathbf{P}_{X_1\sim N(0,1)}\{|X_1| \notin [a,b]\} + \mathbf{E}_{X_1\sim N(0,1)}[\mathbf{1}_{|X_1|\in[a,b]} f(X_1)] > 1.$$

From numerical integration we find that $f(x) > 1$ on $[a,b]$ and $[-b,-a]$ if we take $\kappa = 200$, $a \approx 0.44$ and $b \approx 1.70$. The above expectation is then approximately equal to 1.19, which shows that, even though $\tilde{T}_n$ is an e-statistic at each n by Corollary 8.9 (it is even a GROW one), $\tilde{T}_{\tau^*}$ is not an e-statistic (its expectation is 0.19 too large), providing the claimed counterexample.

F.4 Further Derivations, Computations and Proofs

In this appendix, we prove the technical lemmas whose proof was omitted from the main text. In Section F.4.1, we prove the lemmas used in the proof of Theorem 8.2. In Section F.4.2, we show the computations omitted from Section 8.4.1.

F.4.1 Proof of Technical Lemmas 8.11, 8.12, and 8.13 for Theorem 8.2

Proof of Lemma 8.11. Let $\{\varepsilon_i\}_i$ be a sequence of positive numbers decreasing to zero. Let $\{K_i\}_{i\in\mathbb{N}}$ and $\{L_i\}_{i\in\mathbb{N}}$ be two arbitrary sequences of compact symmetric subsets that increase to cover G. Fix $i \in \mathbb{N}$. The set K_iL_i is compact and by our assumption there exists a sequence $\{J_l\}_{l\in\mathbb{N}}$ and such that $\rho\{J_l\}/\rho\{J_lK_iL_i\} \to 1$ as $l \to \infty$. Pick $l(i)$ to be such that $\rho\{J_{l(i)}\}/\rho\{J_{l(i)}K_iL_i\} \geq 1-\varepsilon_i$. The claim follows from a relabeling of the sequences. □

Proof of Lemma 8.12. Let $h \in N$. Then we can write

$$\begin{aligned}
&\int \mathbf{1}\{g \in NL\}\ q_g(h|m)\mathrm{d}\rho(g) = \int \mathbf{1}\{g \in NL\}\ q_1(g^{-1}h|m)\mathrm{d}\rho(g)\\
=&\int \mathbf{1}\left\{g \in (NL)^{-1}\right\}\ q_1(gh|m)\mathrm{d}\lambda(g) = \Delta(h^{-1})\int \mathbf{1}\left\{g \in (NL)^{-1}h\right\}\ q_1(g|m)\mathrm{d}\lambda(g)\\
=&\Delta(h^{-1})\mathbf{Q}_1\{H \in (NL)^{-1}h \mid M = m\}
\end{aligned}$$

The same computation can be carried out for p. Consequently

$$\begin{aligned}\ln\frac{\int \mathbf{1}\{g\in NL\}\;\, q_g(h|m)\mathrm{d}\rho(g)}{\int \mathbf{1}\{g\in NL\}\;\, p_g(h|m)\mathrm{d}\rho(g)} &= \ln\frac{\mathbf{Q}_1\{H\in (NL)^{-1}h \mid M=m\}}{\mathbf{P}_1\{H\in (NL)^{-1}h \mid M=m\}}\\ &\le -\ln \mathbf{P}_1\{H\in (NL)^{-1}h \mid M=m\}.\end{aligned}$$

By our assumption that $h\in N$, we have that $(NL)^{-1}h = L^{-1}N^{-1}h \supseteq L^{-1} = L$. This implies that the last quantity of the previous display is smaller than $-\ln\mathbf{P}_1\{H\in L \mid M=m\}$. The result follows. □

Proof of Lemma 8.13. The result follows from a rewriting and an application of Jensen's inequality. Indeed,

$$\begin{aligned}-\ln\frac{\int p_g(h|m)\mathrm{d}\mathbf{\Pi}(g)}{\int q_g(h|m)\mathrm{d}\mathbf{\Pi}(g)} &= -\ln\frac{\int q_g(h|m)\frac{p_g(h|m)}{q_g(h|m)}\mathrm{d}\mathbf{\Pi}(g)}{\int q_g(h|m)\mathrm{d}\mathbf{\Pi}(g)} = -\ln\int\frac{p_g(h|m)}{q_g(h|m)}\mathrm{d}\mathbf{\Pi}(g|h,m)\\ &\le -\int \ln\frac{p_g(h|m)}{q_g(h|m)}\mathrm{d}\mathbf{\Pi}(g|h,m) = \int\ln\frac{q_g(h|m)}{p_g(h|m)}\mathrm{d}\mathbf{\Pi}(g|h,m),\end{aligned}$$

as it was to be shown. □

F.4.2 Derivation and Computation for Section 8.4.1

We now provide Proposition F.5, giving the derivation underlying Lemma 8.10 in the main text about the likelihood ratio $T^*_{\mathcal{S},n}$ for $\delta_0=0$, followed by details about numerical computation.

Proposition F.5. *Let $X\sim N(\gamma,I)$, and let $mS\sim W(m,I)$ be independent random variables. Let $LL'=S$ be the Cholesky decomposition of S, and let $M=\frac{1}{\sqrt{m}}L^{-1}X$. If $\mathbf{P}_{0,n}$ is the probability distribution under which $X\sim N(0,I)$, then, the likelihood $p^M_{\gamma,m}/p^M_{0,m}$ ratio is given by*

$$\frac{p^M_{\gamma,m}(M)}{p^M_{0,m}(M)} = \mathrm{e}^{-\frac{1}{2}\|\gamma\|^2}\int \mathrm{e}^{\langle\gamma,TA^{-1}M\rangle}\mathrm{d}\mathbf{P}_{m+1,I}(T)$$

where $A\in\mathcal{L}^+$ is the Cholesky factor $AA'=I+MM'$, and $\mathbf{P}^T_{m+1,I}$ is the probability distribution on $\mathcal{L}^+$ such that $TT'\sim W(m+1,I)$.

Proof. Let $\Sigma=\Lambda\Lambda'$ be the Cholesky decomposition of Σ. The density $p^X_{\gamma,\Lambda}$ of X with respect to the Lebesgue measure on $\mathbb{R}^d$ is

$$p^X_{\gamma,\Lambda}(X) = \frac{1}{(2\pi)^{d/2}\det(\Lambda)}\mathrm{etr}\left(-\frac{1}{2}(\Lambda^{-1}X-\gamma)(\Lambda^{-1}X-\gamma)'\right),$$

where, for a square matrix A, we define $\mathrm{etr}(A)$ to be the exponential of the trace of A. Let $W = mS$. Then, the density $p^W_{\gamma,\Lambda}$ of W with respect to the Lebesgue measure on $\mathbb{R}^{d(d-1)/2}$ is

$$p^W_{\gamma,\Lambda}(W) = \frac{1}{2^{md/2}\Gamma_d(n/2)\det(\Lambda)^m}\det(S)^{(m-d-1)/2}\mathrm{etr}\left(-\frac{1}{2}(\Lambda\Lambda')^{-1}W\right).$$

Now, let $W = TT'$ be the Cholesky decomposition of W. We seek to compute the distribution of the random lower lower triangular matrix T. To this end, the change of variables $W \mapsto T$ is one-to-one, and has Jacobian determinant equal to $2^d\prod_{i=1}^d t_{ii}^{d-i+1}$. Consequently, the density $p^T_{\gamma,\Lambda}(T)$ of T with respect to the Lebesgue measure is

$$p^T_{\gamma,\Lambda}(T) = \frac{2^d}{2^{md/2}\Gamma_d(m/2)}\det(\Lambda^{-1}T)^m\mathrm{etr}\left(-\frac{1}{2}(\Lambda^{-1}T)(\Lambda^{-1}T)'\right)\prod_{i=1}^d t_{ii}^{-i}.$$

We recognize $\mathrm{d}\nu(T) = \prod_{i=1}^d t_{ii}^{-i}\mathrm{d}T$ to be a left Haar measure on $\mathcal{L}_+$, and consequently

$$\tilde{p}^T_{\gamma,\Lambda}(T) = \frac{2^d}{2^{md/2}\Gamma_d(m/2)}\det(\Lambda^{-1}T)^m\mathrm{etr}\left(-\frac{1}{2}(\Lambda^{-1}T)(\Lambda^{-1}T)'\right)$$

is the density of T with respect to $\mathrm{d}\nu(T)$. After these rewritings, The density $\tilde{p}^{X,T}_{\gamma,\Lambda}(X,T)$ of the pair (X,T) with respect to $\mathrm{d}X \times \mathrm{d}\nu(T)$ is given by

$$\tilde{p}^{X,T}_{\gamma,\Lambda}(X,T) = \frac{2^d}{K}\frac{\det(\Lambda^{-1}T)^m}{\det(\Lambda)}\mathrm{etr}\left(-\frac{1}{2}(\Lambda^{-1}T)(\Lambda^{-1}T)' - \frac{1}{2}(\Lambda^{-1}X-\gamma)(\Lambda^{-1}X-\gamma)'\right)$$

with $K = (2\pi)^{d/2}2^{md/2}\Gamma_d(n/2)$. The change of variables $(X,T) \mapsto (T^{-1}X, T)$ has Jacobian determinant equal to $\det(T)$. If $M = T^{-1}X$, then, the density $\tilde{p}^{M,T}_{\gamma,\Lambda}$ of (M,T) with respect to $\mathrm{d}M \times \mathrm{d}\nu(T)$ is given by

$$\frac{\det(\Lambda^{-1}T)^{m+1}}{K''}\mathrm{etr}\left(-\frac{1}{2}(\Lambda^{-1}T)(\Lambda^{-1}T)' - \frac{1}{2}(\Lambda^{-1}TM-\gamma)(\Lambda^{-1}TM-\gamma)'\right).$$

We now marginalize T to obtain the distribution of the maximal invariant M. Since the integral is with respect to the left Haar measure $\mathrm{d}\nu(T)$, we have that

$$\int_{T\in\mathcal{L}^+} \tilde{p}^{M,T}_{\gamma,\Lambda}(M,T)\mathrm{d}\nu(T) = \int_{T\in\mathcal{L}^+} \tilde{p}^{M,T}_{\gamma,I}(M,\Lambda^{-1}T)\mathrm{d}\nu(T) = \int_{T\in\mathcal{L}^+} \tilde{p}^{M,T}_{\gamma,I}(M,T)\mathrm{d}\nu(T),$$

and consequently,

$$\begin{aligned} p^M_{\gamma,\Lambda}(M) &= \frac{2^d}{K}\int_{T\in\mathcal{L}^+} \det(T)^{m+1}\mathrm{etr}\left(-\frac{1}{2}TT' - \frac{1}{2}(TM-\gamma)(TM-\gamma)'\right)\mathrm{d}\nu(T) \\ &= \frac{2^d}{K}\mathrm{e}^{-\frac{1}{2}\|\gamma\|^2}\int_{T\in\mathcal{L}^+} \det(T)^{m+1}\mathrm{etr}\left(-\frac{1}{2}T(I+MM')T' + \gamma(TM)'\right)\mathrm{d}\nu(T). \end{aligned}$$

The matrix $I+MM'$ is positive definite and symmetric. It is then possible to perform its Cholesky decomposition $(I+MM') = AA'$. With this at hand, the previous display can be written as

$$p^M_{\gamma,\Lambda}(M) = \frac{\mathrm{e}^{-\frac{1}{2}\|\gamma\|^2}}{K}\int_{T\in\mathcal{L}^+} \det(T)^{m+1}\mathrm{etr}\left(-\frac{1}{2}(TA)(TA)' + \gamma(TM)'\right)\mathrm{d}\nu(T).$$

We now perform the change of variable $T \mapsto TA^{-1}$. To this end, notice that $\mathrm{d}\nu(A^{-1}) = \mathrm{d}\nu(T)\prod_{i=1}^d a_{ii}^{-(d-2i+1)}$, and consequently

$$\begin{aligned} p^M_{\gamma,\Lambda}(M) &= \frac{2^d}{K}\frac{\mathrm{e}^{-\frac{1}{2}\|\gamma\|^2}\prod_{i=1}^d a_{ii}^{2i}}{\det(A)^{m+d+2}}\int_{T\in\mathcal{L}^+} \det(T)^{m+1}\mathrm{etr}\left(-\frac{1}{2}TT' + \gamma(TA^{-1}M)'\right)\mathrm{d}\nu(T) \\ &= \frac{\Gamma_d\left(\frac{m+1}{2}\right)}{\pi^{d/2}\Gamma_d\left(\frac{m}{2}\right)}\frac{\prod_{i=1}^d a_{ii}^{2i}}{\det(A)^{m+d+2}}\mathrm{e}^{-\frac{1}{2}\|\gamma\|^2}\mathbf{P}^T_{m+1}\left[\mathrm{e}^{\langle\gamma,TA^{-1}M\rangle}\right], \end{aligned}$$

so that that at $\gamma = 0$ the density $p^M_{0,\Lambda}(M)$ takes the form

$$p^M_{0,\Lambda}(M) = \frac{\Gamma_d\left(\frac{m+1}{2}\right)}{\pi^{d/2}\Gamma_d\left(\frac{m}{2}\right)}\frac{\prod_{i=1}^d a_{ii}^{2i}}{\det(A)^{m+d+2}},$$

and consequently the likelihood ratio is

$$\frac{p^M_{\gamma,\Lambda}(M)}{p^M_{0,\Lambda}(M)} = \mathrm{e}^{-\frac{1}{2}\|\gamma\|^2}\int \mathrm{e}^{\langle\gamma,TA^{-1}M\rangle}\mathrm{d}\mathbf{P}_{m+1}(T).$$

□

Remark (Numerical computation)**.** Computing the optimal *e*-statistic is feasible numerically. We are interested in computing

$$\int \mathrm{e}^{\langle x,Ty\rangle} \mathrm{d}\mathbf{P}_{m+1}(T),$$

where T is a $\mathcal{L}^{+}$-valued random lower triangular matrix such that $TT' \sim W(m+1, I)$, and $x, y \in \mathbb{R}^d$. Define, for $i \geq j$, the numbers $a_{ij} = x_i y_j$. Then $\langle x, Ty\rangle = \sum_{i\geq j} a_{ij} T_{ij}$. By Bartlett's decomposition, the entries of the matrix T are independent and $T_{ii}^2 \sim \chi^2((m+1)-i+1)$, and $T_{ij} \sim N(0,1)$ for $i > j$. Hence, our target quantity satisfies

$$\int [\mathrm{e}^{\langle x,Ty\rangle}]\mathbf{P}_{m+1}(T) = \int \mathrm{e}^{\sum_{i\geq j} a_{ij}T_{ij}} \mathrm{d}\mathbf{P}_{m+1}(T) = \int \prod_{i\geq j} \mathrm{e}^{a_{ij}T_{ij}} \mathrm{d}\mathbf{P}_{m+1}(T).$$

On the one hand, for the off-diagonal elements satisfy, using the expression for the moment generating function of a standard normal random variable,

$$\mathbf{E}_{m+1}^{\mathbf{P}}[\mathrm{e}^{a_{ij}T_{ij}}] = \exp\left(\frac{1}{2}a_{ij}^2\right).$$

For the diagonal elements the situation is not as simple, but a numerical solution is possible. Indeed, for $a_{ii} \geq 0$, and $k_i = (m+1) - i + 1$

$$\begin{aligned}\mathbf{E}_{m}^{\mathbf{P}}[\mathrm{e}^{a_{ii}T_{ii}}] &= \frac{1}{2^{\frac{k_i}{2}}\Gamma\left(\frac{k_i}{2}\right)} \int_0^\infty x^{\frac{k_i}{2}-1} \exp\left(-\frac{1}{2}x + a_{ii}\sqrt{x}\right) \mathrm{d}x \\ &= {}_1F_1\left(\frac{k_i}{2}, \frac{1}{2}, \frac{a_{ii}^2}{2}\right) + \frac{\sqrt{2}a_{ii}\Gamma\left(\frac{k_i+1}{2}\right)}{\Gamma\left(\frac{k_i}{2}\right)} {}_1F_1\left(\frac{k_i+1}{2}, \frac{3}{2}, \frac{a_{ii}^2}{2}\right),\end{aligned}$$

where ${}_1F_1(a,b,z)$ is the Kummer confluent hypergeometric function. For $a_{ii} < 0$,

$$\frac{1}{2^{k_i/2}\Gamma\left(\frac{k_i}{2}\right)} \int_0^\infty x^{k_i/2-1} \exp\left(-\frac{1}{2}x + a_{ii}\sqrt{x}\right) \mathrm{d}x = \frac{\Gamma(k_i)}{2^{k_i-1}\Gamma\left(\frac{k_i}{2}\right)} U\left(\frac{k_i}{2}, \frac{1}{2}, \frac{a_{ii}^2}{2}\right),$$

and U is Kummer's U function.

Curriculum Vitae

Tyron Darnell Lardy was born in Haarlem on December 6, 1996. He completed his bilingual high school education at the Mendelcollege in 2014, after which he pursued a bachelor's degree in both mathematics and physics at Leiden University. Ultimately, Tyron decided to focus on mathematics, earning a master's degree in the field (cum laude) while working part-time as a software developer to gain practical experience. Alongside his academic activities, Tyron competed internationally in karate as a member of the Dutch national team, achieving third place at the World U21 Championships in 2015, first place at the European U21 Championships in 2016, and third place at the European Championships in 2018. He retired as an athlete to pursue a PhD at the Mathematical Institute of Leiden University under the supervision of Professor Peter Grünwald and Professor Wouter Koolen, completing his doctoral studies in 2025.