ISSN 0090-5364 (print)
ISSN 2168-8966 (online)

THE ANNALS *of* STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles

Vol. 51, No. 5—October 2023

THE ANNALS OF STATISTICS Vol. 51, No. 5, pp. 1895–2297 October 2023

The Annals of Statistics **[ISSN 0090-5364 (print); ISSN 2168-8966 (online)],** Volume 51, Number 5, October 2023. Published bimonthly by the Institute of Mathematical Statistics, 9760 Smith Road, Waite Hill, OH 44094, USA. Periodicals postage paid at Cleveland, Ohio, and at additional mailing offices.

POSTMASTER: Send address changes to *The Annals of Statistics*, Institute of Mathematical Statistics, Dues and Subscriptions Office, PO Box 729, Middletown, MD 21769, USA.

The Annals of Statistics
2023, Vol. 51, No. 5, 1895–1920
https://doi.org/10.1214/23-AOS2308

THE IMPACTS OF UNOBSERVED COVARIATES ON COVARIATE-ADAPTIVE RANDOMIZED EXPERIMENTS

BY YANG LIU[1,a] AND FEIFANG HU[2,b]

[1]*Institute of Statistics and Big Data, Renmin University of China,* [a]*yangliu2022@ruc.edu.cn*
[2]*Department of Statistics, George Washington University,* [b]*feifang@gwu.edu*

Covariate-adaptive randomization (CAR) is commonly implemented in clinical trials to balance observed covariates. Recent studies have demonstrated the advantages of CAR procedures in balancing covariates and improving the subsequent statistical analysis. Covariate balance is crucial, but it is not a panacea for the valid statistical inferences. If the response to a treatment interacts with some unobserved covariates, the conclusion drawn from a CAR experiment may be affected, and thus, be inconsistent with other evidence. This paper aims to demonstrate the relationships between unobserved covariates and the analysis of treatment and covariate effects in CAR experiments. We first derive the asymptotic properties of the statistical methods based on a linear model framework with interactions between the treatment and an unobserved covariate. We also provide sufficient conditions for the identifiability of the treatment and covariate effects. Our results theoretically explain how inconsistent estimations are generated in CAR experiments when some important covariates are unobserved. Under these sufficient conditions, we show that the tests for the treatment and covariate effects can have reduced Type I errors under CAR procedures. A residual-based adjusted test is proposed to recover the Type I error when the effect can be correctly estimated. Numerical studies are conducted to evaluate the performance of our proposed procedure and theoretical findings.

REFERENCES

AUSTIN, P. C., MANCA, A., ZWARENSTEIN, M., JUURLINK, D. N. and STANBROOK, M. B. (2010). A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: A review of trials published in leading medical journals. *J. Clin. Epidemiol.* **63** 142–153. https://doi.org/10.1016/j.jclinepi.2009.06.002

BALDI ANTOGNINI, A. and ZAGORAIOU, M. (2011). The covariate-adaptive biased coin design for balancing clinical trials in the presence of prognostic factors. *Biometrika* **98** 519–535. MR2836404 https://doi.org/10.1093/biomet/asr021

BALDI ANTOGNINI, A. and ZAGORAIOU, M. (2015). On the almost sure convergence of adaptive allocation procedures. *Bernoulli* **21** 881–908. MR3338650 https://doi.org/10.3150/13-BEJ591

BALDI ANTOGNINI, A. and ZAGORAIOU, M. (2017). Estimation accuracy under covariate-adaptive randomization procedures. *Electron. J. Stat.* **11** 1180–1206. MR3634333 https://doi.org/10.1214/17-EJS1261

BLACKWELL, D. and HODGES, J. L. JR. (1957). Design for the control of selection bias. *Ann. Math. Stat.* **28** 449–460. MR0088849 https://doi.org/10.1214/aoms/1177706973

BUGNI, F. A., CANAY, I. A. and SHAIKH, A. M. (2018). Inference under covariate-adaptive randomization. *J. Amer. Statist. Assoc.* **113** 1784–1796. MR3902246 https://doi.org/10.1080/01621459.2017.1375934

BUGNI, F. A., CANAY, I. A. and SHAIKH, A. M. (2019). Inference under covariate-adaptive randomization with multiple treatments. *Quant. Econ.* **10** 1747–1785. MR4048253 https://doi.org/10.3982/qe1150

CIOLINO, J. D., PALAC, H. L., YANG, A., VACA, M. and BELLI, H. M. (2019). Ideal vs. real: A systematic review on handling covariates in randomized controlled trials. *BMC Med. Res. Methodol.* **19** 136.

EFRON, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika* **58** 403–417. MR0312660 https://doi.org/10.1093/biomet/58.3.403

MSC2020 subject classifications. Primary 62G99, 62L05; secondary 62F03, 60G42.
Key words and phrases. Treatment effect, covariate effect, identifiability, inconsistency, hypothesis tests, CAR procedures, conservativeness of test.

ELKASHEF, A., FUDALA, P. J., GORGON, L., LI, S.-H., KAHN, R., CHIANG, N., VOCCI, F., COLLINS, J., JONES, K. et al. (2006). Double-blind, placebo-controlled trial of selegiline transdermal system (STS) for the treatment of cocaine dependence. *Drug Alcohol Depend*. **85** 191–197.

EMA (2015). *Guideline on Adjustment for Baseline Covariates in Clinical Trials*. European Medicines Agency, Amsterdam.

FDA (2019). *Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biologics with Continuous Outcomes*. *Draft Guidance for Industry*. Center for Drug Evaluation and Research and Center for Biologics Evaluation and Research, Food and Drug Administration, U.S. Department of Health and Human Services.

FREEDMAN, D. A. (2008). On regression adjustments to experimental data. *Adv. in Appl. Math.* **40** 180–193. MR2388610 https://doi.org/10.1016/j.aam.2006.12.003

FRENCH, P. J., SWAGEMAKERS, S. M., NAGEL, J. H., KOUWENHOVEN, M. C., BROUWER, E., VAN DER SPEK, P., LUIDER, T. M., KROS, J. M., VAN DEN BENT, M. J. et al. (2005). Gene expression profiles associated with treatment response in oligodendrogliomas. *Cancer Res*. **65** 11335–11344.

GABRIELSEN, A. (1978). Consistency and identifiability. *J. Econometrics* **8** 261–263. MR0547145 https://doi.org/10.1016/0304-4076(78)90035-0

GAIL, M. H., WIEAND, S. and PIANTADOSI, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* **71** 431–444. MR0775390 https://doi.org/10.1093/biomet/71.3.431

GOLDER, W. A. (2017). Systematic errors in clinical studies: A comprehensive survey. *Ophthalmologe* **114** 215–223. https://doi.org/10.1007/s00347-017-0471-5

HALLSTROM, A. and DAVIS, K. (1988). Imbalance in treatment assignments in stratified blocked randomization. *Control. Clin. Trials* **9** 375–382. https://doi.org/10.1016/0197-2456(88)90050-5

HAMMER, S. M., SQUIRES, K. E., HUGHES, M. D., GRIMES, J. M., DEMETER, L. M., CURRIER, J. S., ERON JR., J. J., FEINBERG, J. E., BALFOUR JR., H. H. et al. (1997). A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less. AIDS Clinical Trials Group 320 Study Team. *N. Engl. J. Med.* **337** 725–733. https://doi.org/10.1056/NEJM199709113371101

HILGERS, R.-D., USCHNER, D., ROSENBERGER, W. F. and HEUSSEN, N. (2017). ERDO-a framework to select an appropriate randomization procedure for clinical trials. *BMC Med. Res. Methodol.* **17** 1–12.

HU, F. (2012). Statistical issues in trial design and personalized medicine. *Clin. Invest.* **2** 121–124.

HU, F., HU, Y., MA, Z. and ROSENBERGER, W. F. (2014). Adaptive randomization for balancing over covariates. *Wiley Interdiscip. Rev.: Comput. Stat.* **6** 288–303.

HU, F., YE, X. and ZHANG, L. X. (2023). Multi-arm covariate-adaptive randomization. *Sci. China Math.* **66** 163–190.

HU, F. and ZHANG, L. X. (2020). On the theory of covariate-adaptive designs. Preprint. Available at arXiv:2004.02994.

HU, Y. and HU, F. (2012). Asymptotic properties of covariate-adaptive randomization. *Ann. Statist.* **40** 1794–1815. MR3015044 https://doi.org/10.1214/12-AOS983

ICH (1998). ICH harmonised tripartite guideline E9: Statistical principles for clinical trials. In *International Conference on Harmonisation*.

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

LI, X., ZHOU, J. and HU, F. (2019). Testing hypotheses under adaptive randomization with continuous covariates in clinical trials. *Stat. Methods Med. Res.* **28** 1609–1621. MR3961953 https://doi.org/10.1177/0962280218770231

LI, Y., MA, W., QIN, Y. and HU, F. (2021). Testing for treatment effect in covariate-adaptive randomized trials with generalized linear models and omitted covariates. *Stat. Methods Med. Res.* **30** 2148–2164. MR4309241 https://doi.org/10.1177/09622802211008206

LIN, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Ann. Appl. Stat.* **7** 295–318. MR3086420 https://doi.org/10.1214/12-AOAS583

LIN, Y. and SU, Z. (2012). Balancing continuous and categorical baseline covariates in sequential clinical trials using the area between empirical cumulative distribution functions. *Stat. Med.* **31** 1961–1971. MR2956029 https://doi.org/10.1002/sim.5363

LIN, Y., ZHU, M. and SU, Z. (2015). The pursuit of balance: An overview of covariate-adaptive randomization techniques in clinical trials. *Contemp. Clin. Trials* **45** 21–25.

LIU, Y. and HU, F. (2022). Balancing unobserved covariates with covariate-adaptive randomized experiments. *J. Amer. Statist. Assoc.* **117** 875–886. MR4436319 https://doi.org/10.1080/01621459.2020.1825450

LIU, Y. and HU, F. (2023). Supplement to "The impacts of unobserved covariates on covariate-adaptive randomized experiments." https://doi.org/10.1214/23-AOS2308SUPP

MA, W., HU, F. and ZHANG, L. (2015). Testing hypotheses of covariate-adaptive randomized clinical trials. *J. Amer. Statist. Assoc.* **110** 669–680. MR3367256 https://doi.org/10.1080/01621459.2014.922469

MA, W., QIN, Y., LI, Y. and HU, F. (2020). Statistical inference for covariate-adaptive randomization procedures. *J. Amer. Statist. Assoc.* **115** 1488–1497. MR4143480 https://doi.org/10.1080/01621459.2019.1635483

MA, W., TU, F. and LIU, H. (2022). Regression analysis for covariate-adaptive randomization: A robust and efficient inference perspective. *Stat. Med.* **41** 5645–5661. MR4515034 https://doi.org/10.1002/sim.9585

MA, Z. and HU, F. (2013). Balancing continuous covariates based on kernel densities. *Contemp. Clin. Trials* **34** 262–269.

MATTS, J. P. and MCHUGH, R. B. (1978). Analysis of accrual randomized clinical trials with balanced groups in strata. *J. Chronic Dis.* **31** 725–740. https://doi.org/10.1016/0021-9681(78)90057-7

MEYN, S. and TWEEDIE, R. L. (2009). *Markov Chains and Stochastic Stability*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2509253 https://doi.org/10.1017/CBO9780511626630

MICKENAUTSCH, S. (2010). Systematic reviews, systematic error and the acquisition of clinical knowledge. *BMC Med. Res. Methodol.* **10** 53. https://doi.org/10.1186/1471-2288-10-53

NEUHAUS, J. M. (1998). Estimation efficiency with omitted covariates in generalized linear models. *J. Amer. Statist. Assoc.* **93** 1124–1129. MR1649206 https://doi.org/10.2307/2669855

NEUHAUS, J. M. and JEWELL, N. P. (1993). A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika* **80** 807–815. MR1282789 https://doi.org/10.1093/biomet/80.4.807

OTT, P. A., BANG, Y. J., PIHA-PAUL, S. A., RAZAK, A. R. A., BENNOUNA, J., SORIA, J. C., RUGO, H. S., COHEN, R. B., O'NEIL, B. H. et al. (2019). T-cell-inflamed gene-expression profile, programmed death ligand 1 expression, and tumor mutational burden predict efficacy in patients treated with pembrolizumab across 20 cancers: Keynote-028. *J. Clin. Oncol.* **37** 318–327.

PEARL, J. (1995). Causal diagrams for empirical research. *Biometrika* **82** 669–710. MR1380809 https://doi.org/10.1093/biomet/82.4.669

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

POCOCK, S. J., ASSMANN, S. E., ENOS, L. E. and KASTEN, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: Current practiceand problems. *Stat. Med.* **21** 2917–2930.

POCOCK, S. J. and SIMON, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* **31** 103–115.

QIN, Y., LI, Y., MA, W. and HU, F. (2016). Pairwise sequential randomization and its properties. Preprint. Available at arXiv:1611.02802.

RAO, P. (1971). Some notes on misspecification in multiple regressions. *Amer. Statist.* **25** 37–39.

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

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

ROSENBERGER, W. F. and LACHIN, J. M. (2016). *Randomization in Clinical Trials: Theory and Practice*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR3443072 https://doi.org/10.1002/9781118742112

ROSENBERGER, W. F. and SVERDLOV, O. (2008). Handling covariates in the design of clinical trials. *Statist. Sci.* **23** 404–419. MR2483911 https://doi.org/10.1214/08-STS269

SAVCI-HEIJINK, C. D., HALFWERK, H., KOSTER, J. and VAN DE VIJVER, M. J. (2017). Association between gene expression profile of the primary tumor and chemotherapy response of metastatic breast cancer. *BMC Cancer* **17** 1–8.

SHAO, J. (2021). Inference after covariate-adaptive randomisation: Aspects of methodology and theory. *Stat. Theory Relat. Fields* **5** 172–186. MR4311482 https://doi.org/10.1080/24754269.2021.1871873

SHAO, J. and YU, X. (2013). Validity of tests under covariate-adaptive biased coin randomization and generalized linear models. *Biometrics* **69** 960–969. MR3146791 https://doi.org/10.1111/biom.12062

SHAO, J., YU, X. and ZHONG, B. (2010). A theory for testing hypotheses under covariate-adaptive randomization. *Biometrika* **97** 347–360. MR2650743 https://doi.org/10.1093/biomet/asq014

SMITH, R. L. (1984). Properties of biased coin designs in sequential clinical trials. *Ann. Statist.* **12** 1018–1034. MR0751289 https://doi.org/10.1214/aos/1176346718

SØRLIE, T., PEROU, C. M., FAN, C., GEISLER, S., AAS, T., NOBEL, A., ANKER, G., AKSLEN, L. A., BOTSTEIN, D. et al. (2006). Gene expression profiles do not consistently predict the clinical treatment response in locally advanced breast cancer. *Mol. Cancer Ther.* **5** 2914–2918.

STRUTHERS, C. A. and KALBFLEISCH, J. D. (1986). Misspecified proportional hazard models. *Biometrika* **73** 363–369. MR0855896 https://doi.org/10.1093/biomet/73.2.363

TAVES, D. R. (1974). Minimization: A new method of assigning patients to treatment and control groups. *Clin. Pharmacol. Ther.* **15** 443–453.

TAVES, D. R. (2010). The use of minimization in clinical trials. *Contemp. Clin. Trials* **31** 180–184. https://doi.org/10.1016/j.cct.2009.12.005

TSIATIS, A. A., DAVIDIAN, M., ZHANG, M. and LU, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Stat. Med.* **27** 4658–4677. MR2528575 https://doi.org/10.1002/sim.3113

WANG, T. and MA, W. (2021). The impact of misclassification on covariate-adaptive randomized clinical trials. *Biometrics* **77** 451–464. MR4307647 https://doi.org/10.1111/biom.13308

WEI, L. J. (1978a). The adaptive biased coin design for sequential experiments. *Ann. Statist.* **6** 92–100. MR0471205

WEI, L. J. (1978b). An application of an urn model to the design of sequential controlled clinical trials. *J. Amer. Statist. Assoc.* **73** 559–563. MR0514157

YE, T. and SHAO, J. (2020). Robust tests for treatment effect in survival analysis under covariate-adaptive randomization. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 1301–1323. MR4176344

YU, L.-M., CHAN, A.-W., HOPEWELL, S., DEEKS, J. J. and ALTMAN, D. G. (2010). Reporting on covariate adjustment in randomised controlled trials before and after revision of the 2001 CONSORT statement: A literature review. *Trials* **11** 1–13.

ZELEN, M. (1974). The randomization and stratification of patients to clinical trials. *J. Chronic Dis.* **27** 365–375.

ZHANG, M., TSIATIS, A. A. and DAVIDIAN, M. (2008). Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics* **64** 707–715. MR2526620 https://doi.org/10.1111/j.1541-0420.2007.00976.x

ZHAO, Z., SONG, Y., JIANG, W. and TU, D. (2022). Consistent covariances estimation for stratum imbalances under marginal design for covariate adaptive randomization. Preprint. Available at arXiv:2209.13117.

ZHU, H., HU, F. and ZHAO, H. (2013). Adaptive clinical trial designs to detect interaction between treatment and a dichotomous biomarker. *Canad. J. Statist.* **41** 525–539. MR3101598 https://doi.org/10.1002/cjs.11184

The Annals of Statistics
2023, Vol. 51, No. 5, 1921–1945
https://doi.org/10.1214/23-AOS2310

SHARP OPTIMALITY FOR HIGH-DIMENSIONAL COVARIANCE TESTING UNDER SPARSE SIGNALS

BY SONG XI CHEN[1,a], YUMOU QIU[2,b] AND SHUYI ZHANG[3,c]

[1]*Guanghua School of Management and Center for Statistical Science, Peking University,* [a]*songxichen@pku.edu.cn*

[2]*School of Mathematical Sciences and Center for Statistical Science, Peking University,* [b]*qiuyumou@math.pku.edu.cn*

[3]*KLATASDS-MoE, School of Statistics, Academy of Statistics and Interdisciplinary Sciences, East China Normal University,* [c]*syzhang@fem.ecnu.edu.cn*

This paper considers one-sample testing of a high-dimensional covariance matrix by deriving the detection boundary as a function of the signal sparsity and signal strength under the sparse alternative hypotheses. It first shows that the optimal detection boundary for testing sparse means is the minimax detection lower boundary for testing the covariance matrix. A multilevel thresholding test is proposed and is shown to be able to attain the detection lower boundary over a substantial range of the sparsity parameter, implying that the multilevel thresholding test is sharp optimal in the minimax sense over the range. The asymptotic distribution of the multilevel thresholding statistic for covariance matrices is derived under both Gaussian and non-Gaussian distributions by developing a novel U-statistic decomposition in conjunction with the matrix blocking and the coupling techniques to handle the complex dependence among the elements of the sample covariance matrix. The superiority in the detection boundary of the multilevel thresholding test over the existing tests is also demonstrated.

REFERENCES

ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR1990662

ARIAS-CASTRO, E., BUBECK, S. and LUGOSI, G. (2012). Detection of correlations. *Ann. Statist.* **40** 412–435. MR3014312 https://doi.org/10.1214/11-AOS964

BERBEE, H. C. P. (1979). *Random Walks with Stationary Increments and Renewal Theory*. *Mathematical Centre Tracts* **112**. Mathematisch Centrum, Amsterdam. MR0547109

BERGER, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR0804611 https://doi.org/10.1007/978-1-4757-4286-2

BICKEL, P. J. and LEVINA, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969 https://doi.org/10.1214/009053607000000758

BRADLEY, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surv.* **2** 107–144. MR2178042 https://doi.org/10.1214/154957805100000104

CAI, T., LIU, W. and XIA, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *J. Amer. Statist. Assoc.* **108** 265–277. MR3174618 https://doi.org/10.1080/01621459.2012.758041

CAI, T. T. and JIANG, T. (2011). Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *Ann. Statist.* **39** 1496–1525. MR2850210 https://doi.org/10.1214/11-AOS879

CAI, T. T. and MA, Z. (2013). Optimal hypothesis testing for high dimensional covariance matrices. *Bernoulli* **19** 2359–2388. MR3160557 https://doi.org/10.3150/12-BEJ455

CHEN, S. X., QIU, Y. and ZHANG, S. (2023). Supplement to "Sharp optimality for high-dimensional covariance testing under sparse signals." https://doi.org/10.1214/23-AOS2310SUPPA, https://doi.org/10.1214/23-AOS2310SUPPB

CHEN, S. X., ZHANG, L.-X. and ZHONG, P.-S. (2010). Tests for high-dimensional covariance matrices. *J. Amer. Statist. Assoc.* **105** 810–819. MR2724863 https://doi.org/10.1198/jasa.2010.tm09560

MSC2020 subject classifications. 62C20, 62H15, 62H20.

Key words and phrases. Covariance matrix, detection boundary, high dimensionality, minimax, rare and faint signal, thresholding.

DELAIGLE, A., HALL, P. and JIN, J. (2011). Robustness and accuracy of methods for high dimensional data analysis based on Student's t-statistic. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 283–301. MR2815777 https://doi.org/10.1111/j.1467-9868.2010.00761.x

DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. MR2065195 https://doi.org/10.1214/009053604000000265

DONOHO, D. and JIN, J. (2015). Higher criticism for large-scale inference, especially for rare and weak effects. *Statist. Sci.* **30** 1–25. MR3317751 https://doi.org/10.1214/14-STS506

HALL, P. and JIN, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist.* **38** 1686–1732. MR2662357 https://doi.org/10.1214/09-AOS764

HOTELLING, H. (1953). New light on the correlation coefficient and its transforms. *J. Roy. Statist. Soc. Ser. B* **15** 193–232. MR0060794

INGSTER, Y. I. (1997). Some problems of hypothesis testing leading to infinitely divisible distributions. *Math. Methods Statist.* **6** 47–69. MR1456646

JIANG, T. (2004). The asymptotic distributions of the largest entries of sample correlation matrices. *Ann. Appl. Probab.* **14** 865–880. MR2052906 https://doi.org/10.1214/105051604000000143

LEDOIT, O. and WOLF, M. (2002). Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Ann. Statist.* **30** 1081–1102. MR1926169 https://doi.org/10.1214/aos/1031689018

LEHMANN, E. L. (1959). *Testing Statistical Hypotheses*. Wiley, New York. MR0107933

MOKKADEM, A. (1988). Mixing properties of ARMA processes. *Stochastic Process. Appl.* **29** 309–315. MR0958507 https://doi.org/10.1016/0304-4149(88)90045-2

NAGAO, H. (1973). On some test criteria for covariance matrix. *Ann. Statist.* **1** 700–709. MR0339405

QIU, Y. and CHEN, S. X. (2012). Test for bandedness of high-dimensional covariance matrices and bandwidth estimation. *Ann. Statist.* **40** 1285–1314. MR3015026 https://doi.org/10.1214/12-AOS1002

QIU, Y., CHEN, S. X. and NETTLETON, D. (2018). Detecting rare and faint signals via thresholding maximum likelihood estimators. *Ann. Statist.* **46** 895–923. MR3782388 https://doi.org/10.1214/17-AOS1574

SCHOTT, J. R. (2005). Testing for complete independence in high dimensions. *Biometrika* **92** 951–956. MR2234197 https://doi.org/10.1093/biomet/92.4.951

XUE, L., MA, S. and ZOU, H. (2012). Positive-definite ℓ_1-penalized estimation of large covariance matrices. *J. Amer. Statist. Assoc.* **107** 1480–1491. MR3036409 https://doi.org/10.1080/01621459.2012.725386

ZHONG, P.-S., CHEN, S. X. and XU, M. (2013). Tests alternative to higher criticism for high-dimensional means under sparsity and column-wise dependence. *Ann. Statist.* **41** 2820–2851. MR3161449 https://doi.org/10.1214/13-AOS1168

The Annals of Statistics
2023, Vol. 51, No. 5, 1946–1964
https://doi.org/10.1214/23-AOS2312

ESTIMATION OF MIXED FRACTIONAL STABLE PROCESSES USING HIGH-FREQUENCY DATA

BY FABIAN MIES[1,a] AND MARK PODOLSKIJ[2,b]

[1]*Department of Applied Mathematics, Delft University of Technology,* [a]*f.mies@tudelft.nl*
[2]*Department of Mathematics, University of Luxembourg,* [b]*mark.podolskij@uni.lu*

The linear fractional stable motion generalizes two prominent classes of stochastic processes, namely stable Lévy processes, and fractional Brownian motion. For this reason, it may be regarded as a basic building block for continuous time models. We study a stylized model consisting of a superposition of independent linear fractional stable motions and our focus is on parameter estimation of the model. Applying an estimating equations approach, we construct estimators for the whole set of parameters and derive their asymptotic normality in a high-frequency regime. The conditions for consistency turn out to be sharp for two prominent special cases: (i) for Lévy processes, that is, for the estimation of the successive Blumenthal–Getoor indices and (ii) for the mixed fractional Brownian motion introduced by Cheridito. In the remaining cases, our results reveal a delicate interplay between the Hurst parameters and the indices of stability. Our asymptotic theory is based on new limit theorems for multiscale moving average processes.

REFERENCES

ABRY, P., DELBEKE, L. and FLANDRIN, P. (1999). Wavelet based estimator for the self-similarity parameter of α-stable processes. In 1999 *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP*99 1729–1732.

ABRY, P., PESQUET-POPESCU, B. and TAQQU, M. S. (1999). Estimation Ondelette Des Paramètres de Stabilité et d'autosimilarité Des Processus α-Stables Autosimilaires. In 17^e *Colloque Sur Le Traitement Du Signal et Des Images, FRA*, 1999. *GRETSI, Groupe d'Etudes du Traitement du Signal et des Images*.

AÏT-SAHALIA, Y. and JACOD, J. (2008). Fisher's information for discretely sampled Lévy processes. *Econometrica* **76** 727–761. MR2433480 https://doi.org/10.1111/j.1468-0262.2008.00858.x

AÏT-SAHALIA, Y. and JACOD, J. (2009). Estimating the degree of activity of jumps in high frequency data. *Ann. Statist.* **37** 2202–2244. MR2543690 https://doi.org/10.1214/08-AOS640

AÏT-SAHALIA, Y. and JACOD, J. (2012). Identifying the successive Blumenthal–Getoor indices of a discretely observed process. *Ann. Statist.* **40** 1430–1464. MR3015031 https://doi.org/10.1214/12-AOS976

AÏT-SAHALIA, Y. and JACOD, J. (2014). *High-Frequency Financial Econometrics*. Princeton Univ. Press, Princeton, NJ.

ASTRAUSKAS, A. (1983). Limit theorems for sums of linearly generated random variables. *Lith. Math. J.* **23** 127–134. MR0706002

AYACHE, A. and HAMONIER, J. (2012). Linear fractional stable motion: A wavelet estimator of the α parameter. *Statist. Probab. Lett.* **82** 1569–1575. MR2930661 https://doi.org/10.1016/j.spl.2012.04.005

AZMOODEH, E., LJUNGDAHL, M. M. and THÄLE, C. (2022). Multi-dimensional normal approximation of heavy-tailed moving averages. *Stochastic Process. Appl.* **145** 308–334. MR4367889 https://doi.org/10.1016/j.spa.2021.11.011

BASSE-O'CONNOR, A., HEINRICH, C. and PODOLSKIJ, M. (2018). On limit theory for Lévy semi-stationary processes. *Bernoulli* **24** 3117–3146. MR3779712 https://doi.org/10.3150/17-BEJ956

BASSE-O'CONNOR, A., HEINRICH, C. and PODOLSKIJ, M. (2019). On limit theory for functionals of stationary increments Lévy driven moving averages. *Electron. J. Probab.* **24** Paper No. 79, 42. MR4003132 https://doi.org/10.1214/19-ejp336

MSC2020 subject classifications. Primary 60G22, 62F12; secondary 62E20, 62M09, 60F05.
Key words and phrases. High frequency data, linear fractional stable motion, Lévy processes, parametric estimation, self-similar processes.

BASSE-O'CONNOR, A., LACHIÈZE-REY, R. and PODOLSKIJ, M. (2017). Power variation for a class of stationary increments Lévy driven moving averages. *Ann. Probab.* **45** 4477–4528. MR3737916 https://doi.org/10.1214/16-AOP1170

BASSE-O'CONNOR, A. and PODOLSKIJ, M. (2017). On critical cases in limit theory for stationary increments Lévy driven moving averages. *Stochastics* **89** 360–383. MR3574707 https://doi.org/10.1080/17442508.2016.1191493

BULL, A. D. (2016). Near-optimal estimation of jump activity in semimartingales. *Ann. Statist.* **44** 58–86. MR3449762 https://doi.org/10.1214/15-AOS1349

CHERIDITO, P. (2001). Mixed fractional Brownian motion. *Bernoulli* **7** 913–934. MR1873835 https://doi.org/10.2307/3318626

CHONG, C., DELERUE, T. and LI, G. (2021). When frictions are fractional: Rough noise in high-frequency data. Available at arXiv:2106.16149.

CHONG, C., DELERUE, T. and MIES, F. (2022). Rate-optimal estimation of mixed semimartingales. Available at arXiv:2207.10464.

DANG, T. T. N. and ISTAS, J. (2017). Estimation of the Hurst and the stability indices of a H-self-similar stable process. *Electron. J. Stat.* **11** 4103–4150. MR3715823 https://doi.org/10.1214/17-EJS1357

GRAHOVAC, D., LEONENKO, N. N. and TAQQU, M. S. (2015). Scaling properties of the empirical structure function of linear fractional stable motion and estimation of its parameters. *J. Stat. Phys.* **158** 105–119. MR3296276 https://doi.org/10.1007/s10955-014-1126-4

JACOD, J. and SØRENSEN, M. (2018). A review of asymptotic theory of estimating functions. *Stat. Inference Stoch. Process.* **21** 415–434. MR3824976 https://doi.org/10.1007/s11203-018-9178-8

LJUNGDAHL, M. M. and PODOLSKIJ, M. (2020). A minimal contrast estimator for the linear fractional stable motion. *Stat. Inference Stoch. Process.* **23** 381–413. MR4123929 https://doi.org/10.1007/s11203-020-09216-2

LJUNGDAHL, M. M. and PODOLSKIJ, M. (2021). Multidimensional parameter estimation of heavy-tailed moving averages. *Scand. J. Stat.* **49** 593–624. MR4428498 https://doi.org/10.1111/sjos.12527

MAZUR, S., OTRYAKHIN, D. and PODOLSKIJ, M. (2020). Estimation of the linear fractional stable motion. *Bernoulli* **26** 226–252. MR4036033 https://doi.org/10.3150/19-BEJ1124

MIES, F. (2020). Rate-optimal estimation of the Blumenthal–Getoor index of a Lévy process. *Electron. J. Stat.* **14** 4165–4206. MR4175392 https://doi.org/10.1214/20-EJS1769

PIPIRAS, V. and TAQQU, M. S. (2003). Central limit theorems for partial sums of bounded functionals of infinite-variance moving averages. *Bernoulli* **9** 833–855. MR2047688 https://doi.org/10.3150/bj/1066418880

PIPIRAS, V., TAQQU, M. S. and ABRY, P. (2007). Bounds for the covariance of functions of infinite variance stable random variables with applications to central limit theorems and wavelet-based estimation. *Bernoulli* **13** 1091–1123. MR2364228 https://doi.org/10.3150/07-BEJ6143

REISS, M. (2013). Testing the characteristics of a Lévy process. *Stochastic Process. Appl.* **123** 2808–2828. MR3054546 https://doi.org/10.1016/j.spa.2013.03.016

STOEV, S., PIPIRAS, V. and TAQQU, M. S. (2002). Estimation of the self-similarity parameter in linear fractional stable motion. *Signal Process.* **82** 1873–1901.

STOEV, S. and TAQQU, M. S. (2005). Asymptotic self-similarity and wavelet estimation for long-range dependent fractional autoregressive integrated moving average time series with stable innovations. *J. Time Series Anal.* **26** 211–249. MR2122896 https://doi.org/10.1111/j.1467-9892.2005.00399.x

VAN ZANTEN, H. (2007). When is a linear combination of independent fBm's equivalent to a single fBm? *Stochastic Process. Appl.* **117** 57–70. MR2287103 https://doi.org/10.1016/j.spa.2006.05.013

XIAO, W.-L., ZHANG, W.-G. and ZHANG, X.-L. (2011). Maximum-likelihood estimators in the mixed fractional Brownian motion. *Statistics* **45** 73–85. MR2772157 https://doi.org/10.1080/02331888.2010.541254

MIES, F. and PODOLSKIJ, M. (2023). Supplement to "Estimation of mixed fractional stable processes using high-frequency data." https://doi.org/10.1214/23-AOS2312SUPP

The Annals of Statistics
2023, Vol. 51, No. 5, 1965–1988
https://doi.org/10.1214/23-AOS2313

EFFICIENT ESTIMATION OF THE MAXIMAL ASSOCIATION BETWEEN MULTIPLE PREDICTORS AND A SURVIVAL OUTCOME

BY TZU-JUNG HUANG[1,a], ALEX LUEDTKE[2,b] AND IAN W. MCKEAGUE[3,c]

[1]*Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center,* [a]*thuang2@fredhutch.org*
[2]*Department of Statistics, University of Washington,* [b]*aluedtke@uw.edu*
[3]*Department of Biostatistics, Columbia University,* [c]*im2131@columbia.edu*

This paper develops a new approach to post-selection inference for screening high-dimensional predictors of survival outcomes. Post-selection inference for right-censored outcome data has been investigated in the literature, but much remains to be done to make the methods both reliable and computationally-scalable in high dimensions. Machine learning tools are commonly used to provide *predictions* of survival outcomes, but the estimated effect of a selected predictor suffers from confirmation bias unless the selection is taken into account. The new approach involves the construction of semiparametrically efficient estimators of the linear association between the predictors and the survival outcome, which are used to build a test statistic for detecting the presence of an association between any of the predictors and the outcome. Further, a stabilization technique reminiscent of bagging allows a normal calibration for the resulting test statistic, which enables the construction of confidence intervals for the maximal association between predictors and the outcome and also greatly reduces computational cost. Theoretical results show that this testing procedure is valid even when the number of predictors grows superpolynomially with sample size, and our simulations support this asymptotic guarantee at moderate sample sizes. The new approach is applied to the problem of identifying patterns in viral gene expression associated with the potency of an antiviral drug.

REFERENCES

ANDERSEN, P. K. and GILL, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist.* **10** 1100–1120. MR0673646

ANTONIADIS, A., FRYZLEWICZ, P. and LETUÉ, F. (2010). The Dantzig selector in Cox's proportional hazards model. *Scand. J. Stat.* **37** 531–552. MR2779635 https://doi.org/10.1111/j.1467-9469.2009.00685.x

BINDER, H., PORZELIUS, C. and SCHUMACHER, M. (2011). An overview of techniques for linking high-dimensional molecular data to time-to-event endpoints by risk prediction models. *Biom. J.* **53** 170–189. MR2897395 https://doi.org/10.1002/bimj.201000152

BØVELSTAD, H. M., NYGÅRD, S. and BORGAN, Ø. (2009). Survival prediction from clinico-genomic models—a comparative study. *BMC Bioinform.* **10** Article 413.

BRADIC, J., FAN, J. and JIANG, J. (2011). Regularization for Cox's proportional hazards model with NP-dimensionality. *Ann. Statist.* **39** 3092–3120. MR3012402 https://doi.org/10.1214/11-AOS911

BUCKLEY, J. and JAMES, I. (1979). Linear regression with censored data. *Biometrika* **66** 429–436.

BUNEA, F. and MCKEAGUE, I. W. (2005). Covariate selection for semiparametric hazard function regression models. *J. Multivariate Anal.* **92** 186–204. MR2102251 https://doi.org/10.1016/j.jmva.2003.09.006

CAI, T., HUANG, J. and TIAN, L. (2009). Regularized estimation for the accelerated failure time model. *Biometrics* **65** 394–404. MR2751463 https://doi.org/10.1111/j.1541-0420.2008.01074.x

CHAI, H., ZHANG, Q., HUANG, J. and MA, S. (2019). Inference for low-dimensional covariates in a high-dimensional accelerated failure time model. *Statist. Sinica* **29** 877–894. MR3931392

DATTA, S., LE-RADEMACHER, J. and DATTA, S. (2007). Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and LASSO. *Biometrics* **63** 259–271. MR2345596 https://doi.org/10.1111/j.1541-0420.2006.00660.x

---

MSC2020 subject classifications. Primary 62N03, 62G10; secondary 62G20.
Key words and phrases. Marginal screening, post-selection inference, semiparametric efficiency.

DAVIDSON, R. and MACKINNON, J. G. (1987). Implicit alternatives and the local power of test statistics. *Econometrica* **55** 1305–1329. MR0923463 https://doi.org/10.2307/1913558

DEVLIN, S. J., GNANADESIKAN, R. and KETTENRING, J. R. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika* **62** 531–545.

ENGLER, D. and LI, Y. (2009). Survival analysis with high-dimensional covariates: An application in microarray studies. *Stat. Appl. Genet. Mol. Biol.* **8** Art. 14. MR2476392 https://doi.org/10.2202/1544-6115.1423

FAN, J., FENG, Y. and WU, Y. (2010). High-dimensional variable selection for Cox's proportional hazards model. In *Borrowing Strength: Theory Powering Applications—a Festschrift for Lawrence D. Brown. Inst. Math. Stat.* (*IMS*) *Collect.* **6** 70–86. IMS, Beachwood, OH. MR2798512

FAN, J. and LI, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30** 74–99. MR1892656 https://doi.org/10.1214/aos/1015362185

FANG, E. X., NING, Y. and LIU, H. (2017). Testing and confidence intervals for high dimensional proportional hazards models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1415–1437. MR3731669 https://doi.org/10.1111/rssb.12224

GAENSSLER, P., STROBEL, J. and STUTE, W. (1978). On central limit theorems for martingale triangular arrays. *Acta Math. Acad. Sci. Hung.* **31** 205–216. MR0471030 https://doi.org/10.1007/BF01901971

GILBERT, P. B., JURASKA, M., DECAMP, A. C., KARUNA, S., EDUPUGANTI, S., MGODI, N. et al. (2017). Basis and statistical design of the passive HIV-1 antibody mediated prevention (AMP) test-of-concept efficacy trials. *Stat. Commun. Infec. Dis.* **9** 20160001. MR3743441 https://doi.org/10.1515/scid-2016-0001

GORST-RASMUSSEN, A. and SCHEIKE, T. (2013). Independent screening for single-index hazard rate models with ultrahigh dimensional features. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 217–245. MR3021386 https://doi.org/10.1111/j.1467-9868.2012.01039.x

HE, X., WANG, L. and HONG, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Ann. Statist.* **41** 342–369. MR3059421 https://doi.org/10.1214/13-AOS1087

HONG, H. G., CHEN, X., CHRISTIANI, D. C. and LI, Y. (2018). Integrated powered density: Screening ultrahigh dimensional covariates with survival outcomes. *Biometrics* **74** 421–429. MR3825328 https://doi.org/10.1111/biom.12820

HONG, H. G., CHEN, X., KANG, J. and LI, Y. (2020). The L_q-norm learning for ultrahigh-dimensional survival data: An integrative framework. *Statist. Sinica* **30** 1213–1233. MR4257530 https://doi.org/10.5705/ss.202017.0537

HONG, H. G., KANG, J. and LI, Y. (2018). Conditional screening for ultra-high dimensional covariates with survival outcomes. *Lifetime Data Anal.* **24** 45–71. MR3742906 https://doi.org/10.1007/s10985-016-9387-7

HUANG, J. and MA, S. (2010). Variable selection in the accelerated failure time model via the bridge method. *Lifetime Data Anal.* **16** 176–195. MR2608284 https://doi.org/10.1007/s10985-009-9144-2

HUANG, J., MA, S. and XIE, H. (2006). Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics* **62** 813–820. MR2247210 https://doi.org/10.1111/j.1541-0420.2006.00562.x

HUANG, T.-J, LUEDTKE, A. and MCKEAGUE, I. W (2023). Supplement to "Efficient estimation of the maximal association between multiple predictors and a survival outcome." https://doi.org/10.1214/23-AOS2313SUPP

HUANG, T.-J., MCKEAGUE, I. W. and QIAN, M. (2019). Marginal screening for high-dimensional predictors of survival outcomes. *Statist. Sinica* **29** 2105–2139. MR3970349

JIN, Z., LIN, D. Y., WEI, L. J. and YING, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika* **90** 341–353. MR1986651 https://doi.org/10.1093/biomet/90.2.341

JOHNSON, B. A. (2008). Variable selection in semiparametric linear regression with censored data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 351–370. MR2424757 https://doi.org/10.1111/j.1467-9868.2008.00639.x

JOHNSON, B. A., LIN, D. Y. and ZENG, D. (2008). Penalized estimating functions and variable selection in semiparametric regression models. *J. Amer. Statist. Assoc.* **103** 672–680. MR2435469 https://doi.org/10.1198/016214508000000184

KOUL, H., SUSARLA, V. and VAN RYZIN, J. (1981). Regression analysis with randomly right-censored data. *Ann. Statist.* **9** 1276–1288. MR0630110

LAI, T. L. and YING, Z. (1991a). Large sample theory of a modified Buckley–James estimator for regression analysis with censored data. *Ann. Statist.* **19** 1370–1402. MR1126329 https://doi.org/10.1214/aos/1176348253

LAI, T. L. and YING, Z. (1991b). Rank regression methods for left-truncated and right-censored data. *Ann. Statist.* **19** 531–556. MR1105835 https://doi.org/10.1214/aos/1176348110

LI, J., ZHENG, Q., PENG, L. and HUANG, Z. (2016). Survival impact index and ultrahigh-dimensional model-free screening with survival outcomes. *Biometrics* **72** 1145–1154. MR3591599 https://doi.org/10.1111/biom.12499

LI, Y., DICKER, L. and ZHAO, S. D. (2014). The Dantzig selector for censored linear regression models. *Statist. Sinica* **24** 251–268. MR3183683

Liu, Y., Chen, X. and Li, G. (2020). A new joint screening method for right-censored time-to-event data with ultra-high dimensional covariates. *Stat. Methods Med. Res.* **29** 1499–1513. MR4106953 https://doi.org/10.1177/0962280219864710

Lockhart, R., Taylor, J., Tibshirani, R. J. and Tibshirani, R. (2014). A significance test for the lasso. *Ann. Statist.* **42** 413–468. MR3210970 https://doi.org/10.1214/13-AOS1175

Luedtke, A. R. and van der Laan, M. J. (2018). Parametric-rate inference for one-sided differentiable parameters. *J. Amer. Statist. Assoc.* **113** 780–788. MR3832226 https://doi.org/10.1080/01621459.2017.1285777

Ma, S. and Du, P. (2012). Variable selection in partly linear regression model with diverging dimensions for right censored data. *Statist. Sinica* **22** 1003–1020. MR2987481 https://doi.org/10.5705/ss.2010.267

Ma, S., Li, R. and Tsai, C.-L. (2017). Variable screening via quantile partial correlation. *J. Amer. Statist. Assoc.* **112** 650–663. MR3671759 https://doi.org/10.1080/01621459.2016.1156545

Magaret, C. A., Benkeser, D. C., Williamson, B. D., Borate, B. R., Carpp, L. N., Georgiev, I. S., Setliff, I., Dingens, A. S., Simon, N. et al. (2019). Prediction of VRC01 neutralization sensitivity by HIV-1 gp160 sequence features. *PLoS Comput. Biol.* **15** e1006952. https://doi.org/10.1371/journal.pcbi.1006952

Pan, W., Wang, X., Xiao, W. and Zhu, H. (2019). A generic sure independence screening procedure. *J. Amer. Statist. Assoc.* **114** 928–937. MR3963192 https://doi.org/10.1080/01621459.2018.1462709

Pfanzagl, J. (1982). *Contributions to a General Asymptotic Statistical Theory*. *Lecture Notes in Statistics* **13**. Springer, New York. With the assistance of W. Wefelmeyer. MR0675954

Pfanzagl, J. (1990). *Estimation in Semiparametric Models*: *Some Recent Developments*. *Lecture Notes in Statistics* **63**. Springer, New York. MR1048589 https://doi.org/10.1007/978-1-4612-3396-1

Ritov, Y. (1990). Estimation in a linear regression model with censored data. *Ann. Statist.* **18** 303–328. MR1041395 https://doi.org/10.1214/aos/1176347502

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B. et al. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N. Engl. J. Med.* **346** 1937–1947. https://doi.org/10.1056/NEJMoa012914

Sinnott, J. A. and Cai, T. (2016). Inference for survival prediction under the regularized Cox model. *Biostatistics* **17** 692–707. MR3604274 https://doi.org/10.1093/biostatistics/kxw016

Smola, A. J., Gretton, A. and Borgwardt, K. (2006). Maximum mean discrepancy. In 13*th International Conference*, *ICONIP* 2006, *Hong Kong*, *China*, *October* 3–6, 2006: *Proceedings*.

Song, R., Lu, W., Ma, S. and Jeng, X. J. (2014). Censored rank independence screening for high-dimensional survival data. *Biometrika* **101** 799–814. MR3286918 https://doi.org/10.1093/biomet/asu047

Stute, W. and Wang, J.-L. (1993). The strong law under random censorship. *Ann. Statist.* **21** 1591–1607. MR1241280 https://doi.org/10.1214/aos/1176349273

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35** 2769–2794. MR2382665 https://doi.org/10.1214/009053607000000505

Taylor, J. and Tibshirani, R. (2018). Post-selection inference for ℓ_1-penalized likelihood models. *Canad. J. Statist.* **46** 41–61. MR3767165 https://doi.org/10.1002/cjs.11313

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Stat. Med.* **16** 385–395. https://doi.org/10.1002/(sici)1097-0258(19970228)16:4<385::aid-sim380>3.0.co;2-3

Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *Ann. Statist.* **18** 354–372. MR1041397 https://doi.org/10.1214/aos/1176347504

van de Geer, S. (1995). Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *Ann. Statist.* **23** 1779–1801. MR1370307 https://doi.org/10.1214/aos/1176324323

van der Laan, M. J., Gill, R. D. and Robins, J. M. (2000). Locally efficient estimation in censored data models: Theory and examples Technical Report, Division of Biostatistics, Univ. California, Berkeley, CA.

van der Laan, M. J. and Hubbard, A. E. (1998). Locally efficient estimation of the survival distribution with right-censored data and covariates when collection of data is delayed. *Biometrika* **85** 771–783. MR1666754 https://doi.org/10.1093/biomet/85.4.771

van der Laan, M. J. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. *Springer Series in Statistics*. Springer, New York. MR1958123 https://doi.org/10.1007/978-0-387-21700-0

Whitney, D., Shojaie, A. and Carone, M. (2019). Comment: Models as (deliberate) approximations [MR4048582; MR4048583]. *Statist. Sci.* **34** 591–598. MR4048590 https://doi.org/10.1214/19-STS747

Wu, Y. (2012). Elastic net for Cox's proportional hazards model with a solution path algorithm. *Statist. Sinica* **22** 271–294. MR2933176 https://doi.org/10.5705/ss.2010.107

Xia, X. and Li, J. (2021). Copula-based partial correlation screening: A joint and robust approach. *Statist. Sinica* **31** 421–447. MR4270391 https://doi.org/10.5705/ss.20

Xia, X., Li, J. and Fu, B. (2019). Conditional quantile correlation learning for ultrahigh dimensional varying coefficient models and its application in survival analysis. *Statist. Sinica* **29** 645–669. MR3931382

YING, Z. (1993). A large sample study of rank estimation for censored regression data. *Ann. Statist.* **21** 76–99. MR1212167 https://doi.org/10.1214/aos/1176349016

YOON, H., MACKE, J., WEST, A. P. JR, FOLEY, B., BJORKMAN, P. J., KORBER, B. et al. (2015). CATNAP: A tool to compile, analyze and tally neutralizing antibody panels. *Nucleic Acids Res.* **43**.

YU, Y., BRADIC, J. and SAMWORTH, R. J. (2021). Confidence intervals for high-dimensional Cox models. *Statist. Sinica* **31** 243–267.

ZHANG, H. H. and LU, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika* **94** 691–703. MR2410017 https://doi.org/10.1093/biomet/asm037

ZHAO, S. D. and LI, Y. (2012). Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *J. Multivariate Anal.* **105** 397–411. MR2877525 https://doi.org/10.1016/j.jmva.2011.08.002

ZHAO, S. D. and LI, Y. (2014). Score test variable screening. *Biometrics* **70** 862–871. MR3295747 https://doi.org/10.1111/biom.12209

ZHONG, P.-S., HU, T. and LI, J. (2015). Tests for coefficients in high-dimensional additive hazard models. *Scand. J. Stat.* **42** 649–664. MR3391684 https://doi.org/10.1111/sjos.12127

The Annals of Statistics
2023, Vol. 51, No. 5, 1989–2014
https://doi.org/10.1214/23-AOS2316

ASSIGNING TOPICS TO DOCUMENTS BY SUCCESSIVE PROJECTIONS

BY OLGA KLOPP[1,a], MAXIM PANOV[2,b], SUZANNE SIGALLA[3,c] AND ALEXANDRE B. TSYBAKOV[3,d]

[1]*IDS Department, ESSEC Business School,* [a]*klopp@essec.edu*
[2]*ML Department, Mohamed bin Zayed University of Artificial Intelligence (MBZUAI),* [b]*panov.maxim@gmail.com*
[3]*Institut Polytechnique de Paris, CREST, ENSAE,* [c]*suzannesigalla@yahoo.fr,* [d]*alexandre.tsybakov@ensae.fr*

Topic models provide a useful tool to organize and understand the structure of large corpora of text documents, in particular, to discover hidden thematic structure. Clustering documents from big unstructured corpora into topics is an important task in various fields, such as image analysis, e-commerce, social networks and population genetics. Since the number of topics is typically substantially smaller than the size of the corpus and of the dictionary, the methods of topic modeling can lead to a dramatic dimension reduction. We study the problem of estimating the topic-document matrix, which gives the topics distribution for each document in a given corpus, that is, we focus on the clustering aspect of the problem. We introduce an algorithm that we call Successive Projection Overlapping Clustering (SPOC) inspired by the successive projection algorithm for separable matrix factorization. This algorithm is simple to implement and computationally fast. We establish upper bounds on the performance of the SPOC algorithm for estimation of the topic-document matrix, as well as near matching minimax lower bounds. We also propose a method that achieves analogous results when the number of topics is unknown and provides an estimate of the number of topics. Our theoretical results are complemented with a numerical study on synthetic and semisynthetic data.

REFERENCES

[1] ANANDKUMAR, A., FOSTER, D. P., HSU, D. J., KAKADE, S. M. and LIU, Y.-K. (2012). A spectral algorithm for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems* 917–925.

[2] ANANDKUMAR, A., GE, R., HSU, D., KAKADE, S. M. and TELGARSKY, M. (2014). Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.* **15** 2773–2832. MR3270750

[3] ARAUJO, M. C. U., SALDANHA, T. C. B., GALVAO, R. K. H., YONEYAMA, T., CHAME, H. C. and VISANI, V. (2001). The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemom. Intell. Lab. Syst.* **57** 65–73. https://doi.org/10.1016/S0169-7439(01)00119-8

[4] ARORA, S., GE, R., HALPERN, Y., MIMNO, D., MOITRA, A., SONTAG, D., WU, Y. and ZHU, M. (2013). A practical algorithm for topic modeling with provable guarantees. In *International Conference on Machine Learning* 280–288.

[5] ARORA, S., GE, R., KOEHLER, F., MA, T. and MOITRA, A. (2016). Provable algorithms for inference in topic models. In *International Conference on Machine Learning* 2859–2867. PMLR, 48.

[6] ARORA, S., GE, R. and MOITRA, A. (2012). Learning topic models—going beyond SVD. In 2012 *IEEE 53rd Annual Symposium on Foundations of Computer Science—FOCS* 2012 1–10. IEEE Computer Soc., Los Alamitos, CA. MR3185945

[7] AZAR, Y., FIAT, A., KARLIN, A., MCSHERRY, F. and SAIA, J. (2001). Spectral analysis of data. In *Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing* 619–626.

[8] BANSAL, T., BHATTACHARYYA, C. and KANNAN, R. (2014). A provable SVD-based algorithm for learning topics in dominant admixture corpus. In *Advances in Neural Information Processing Systems* 1997–2005.

MSC2020 subject classifications. Primary 62G05; secondary 60C05.
Key words and phrases. Topic model, latent variable model, nonnegative matrix factorization, minimax rate of convergence, adaptive estimation.

[9] BEYGELZIMER, A., HAZAN, E., KALE, S. and LUO, H. (2015). Online gradient boosting. *Adv. Neural Inf. Process. Syst.* **28**.

[10] BICEGO, M., LOVATO, P., PERINA, A., FASOLI, M., DELLEDONNE, M., PEZZOTTI, M., POLVERARI, A. and MURINO, V. (2012). Investigating topic models' capabilities in expression microarray data classification. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **9** 1831–1836. https://doi.org/10.1109/TCBB.2012.121

[11] BING, X., BUNEA, F., STRIMAS-MACKEY, S. and WEGKAMP, M. (2022). Likelihood estimation of sparse topic distributions in topic models and its applications to Wasserstein document distance calculations. *Ann. Statist.* **50** 3307–3333. MR4524498 https://doi.org/10.1214/22-aos2229

[12] BING, X., BUNEA, F. and WEGKAMP, M. (2020). A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics. *Bernoulli* **26** 1765–1796. MR4091091 https://doi.org/10.3150/19-BEJ1166

[13] BING, X., BUNEA, F. and WEGKAMP, M. (2020). Optimal estimation of sparse topic models. *J. Mach. Learn. Res.* **21** Paper No. 177, 45. MR4209463

[14] BLEI, D. M. and LAFFERTY, J. D. (2006). Dynamic topic models. In *Proceedings of the* 23*rd International Conference on Machine Learning* 113–120.

[15] BLEI, D. M. and LAFFERTY, J. D. (2007). A correlated topic model of *Science*. *Ann. Appl. Stat.* **1** 17–35. MR2393839 https://doi.org/10.1214/07-AOAS114

[16] BLEI, D. M., NG, A. Y. and JORDAN, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3** 993–1022.

[17] CHIEN, J.-T. and CHUEH, C.-H. (2010). Dirichlet class language models for speech recognition. *IEEE Trans. Audio Speech Lang. Process.* **19** 482–495.

[18] CICHOCKI, A., ZDUNEK, R., PHAN, A. H. and AMARI, S.-I. (2009). *Nonnegative Matrix and Tensor Factorizations—Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation*. Wiley, New York.

[19] CURISKIS, S. A., DRAKE, B., OSBORN, T. R. and KENNEDY, P. J. (2020). An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit. *Inf. Process. Manag.* **57** 102034.

[20] DING, W., ROHBAN, M. H., ISHWAR, P. and SALIGRAMA, V. (2013). Topic discovery through data dependent and random projections. In *Proceedings of the* 30*th International Conference on Machine Learning* (*Sanjoy Dasgupta* (D. McAllester, ed.). *Proceedings of Machine Learning Research* **28** 1202–1210. PMLR, Atlanta, GA, USA.

[21] DONOHO, D. and STODDEN, V. (2004). When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems* (S. Thrun, L. Saul and B. Schölkopf, eds.) **16**. MIT Press, Cambridge.

[22] GILLIS, N. and VAVASIS, S. A. (2014). Fast and robust recursive algorithmsfor separable nonnegative matrix factorization. *IEEE Trans. Pattern Anal. Mach. Intell.* **36** 698–714. https://doi.org/10.1109/TPAMI.2013.226

[23] GILLIS, N. and VAVASIS, S. A. (2015). Semidefinite programming based preconditioning for more robust near-separable nonnegative matrix factorization. *SIAM J. Optim.* **25** 677–698. MR3325760 https://doi.org/10.1137/130940670

[24] GIRAUD, C. (2015). *Introduction to High-Dimensional Statistics*. *Monographs on Statistics and Applied Probability* **139**. CRC Press, Boca Raton, FL. MR3307991

[25] HARMAN, D. (1993). Overview of the first TREC conference. In *Proceedings of the* 16*th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. *SIGIR'*93 36–47. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/160688.160692

[26] HOFMANN, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the* 22*nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* 50–57.

[27] KE, Z. T. and WANG, M. (2022). Using SVD for topic modeling. *J. Amer. Statist. Assoc.* 1–16.

[28] KLOPP, O., PANOV, M., SIGALLA, S. and TSYBAKOV, A. B. (2023). Supplement to "Assigning topics to documents by successive projections." https://doi.org/10.1214/23-AOS2316SUPP

[29] LAFFERTY, J. D. and BLEI, D. M. (2006). Correlated topic models. In *Advances in Neural Information Processing Systems* 147–154.

[30] LEE, D. D. and SEUNG, H. S. (1999). Learning the parts of objects by nonnegative matrix factorization. *Nature* **401** 788–791.

[31] LEE, M., BINDEL, D. and MIMNO, D. (2015). Robust spectral inference for joint stochastic matrix factorization. *Adv. Neural Inf. Process. Syst.* **28**.

[32] LEE, M., BINDEL, D. and MIMNO, D. (2020). Prior-aware composition inference for spectral topic models. In *International Conference on Artificial Intelligence and Statistics* 4258–4268. PMLR, 108.

[33] LEE, M., CHO, S., BINDEL, D. and MIMNO, D. (2019). Practical correlated topic modeling and analysis via the rectified anchor word algorithm. In *Proceedings of the* 2019 *Conference on Empirical Methods in Natural Language Processing and the* 9*th International Joint Conference on Natural Language Processing* (*EMNLP-IJCNLP*) 4991–5001.
[34] LEI, J. and RINALDO, A. (2015). Consistency of spectral clustering in stochastic block models. *Ann*. *Statist*. **43** 215–237. MR3285605 https://doi.org/10.1214/14-AOS1274
[35] LI, J., RABANI, Y., SCHULMAN, L. J. and SWAMY, C. (2015). Learning arbitrary statistical mixtures of discrete distributions. In *STOC'*15*—Proceedings of the* 2015 *ACM Symposium on Theory of Computing* 743–752. ACM, New York. MR3388254
[36] LI, L.-J., WANG, C., LIM, Y., BLEI, D. M. and FEI-FEI, L. (2010). Building and using a semantivisual image hierarchy. In 2010 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 3336–3343. IEEE, Los Alamitos.
[37] LI, W. and MCCALLUM, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of the* 23*rd International Conference on Machine Learning* 577–584.
[38] LIU, K., TOKAR, R. and MCVEY, B. (1994). An integrated architecture of adaptive neural network control for dynamic systems. In *Advances in Neural Information Processing Systems* 7.
[39] MAO, X., SARKAR, P. and CHAKRABARTI, D. (2018). Overlapping clustering models, and one (class) svm to bind them all. In *Advances in Neural Information Processing Systems* 2126–2136.
[40] MAO, X., SARKAR, P. and CHAKRABARTI, D. (2021). Estimating mixed memberships with sharp eigenvector deviations. *J*. *Amer*. *Statist*. *Assoc*. **116** 1928–1940. MR4353723 https://doi.org/10.1080/01621459.2020.1751645
[41] MCCALLUM, A., CORRADA-EMMANUEL, A. and WANG, X. (2005). The author-recipient-topic model for topic and role discovery in social networks: Experiments with enron and academic email. *Comput*. *Sci*. *Dep*. *Fac*. *Publ*. *Ser*. **44**.
[42] MIZUTANI, T. (2014). Ellipsoidal rounding for nonnegative matrix factorization under noisy separability. *J*. *Mach*. *Learn*. *Res*. **15** 1011–1039. MR3195337
[43] MIZUTANI, T. (2016). Robustness analysis of preconditioned successive projection algorithm for general form of separable NMF problem. *Linear Algebra Appl*. **497** 1–22. MR3466631 https://doi.org/10.1016/j.laa.2016.02.016
[44] PALESE, B. and USAI, A. (2018). The relative importance of service quality dimensions in E-commerce experiences. *Internat*. *J*. *Inform*. *Management* **40** 132–140.
[45] PANOV, M., SLAVNOV, K. and USHAKOV, R. (2017). Consistent estimation of mixed memberships with successive projections. In *International Conference on Complex Networks and Their Applications* 53–64. Springer, Berlin.
[46] PARK, I. M., ARCHER, E. W., LATIMER, K. and PILLOW, J. W. (2013). Universal models for binary spike patterns using centered Dirichlet processes. *Adv*. *Neural Inf*. *Process*. *Syst*. **26**.
[47] PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A. et al. (2011). Scikit-learn: Machine learning in Python. *J*. *Mach*. *Learn*. *Res*. **12** 2825–2830. MR2854348
[48] PERRONE, V., JENKINS, P. A., SPANÒ, D. and TEH, Y. W. (2017). Poisson random fields for dynamic feature models. *J*. *Mach*. *Learn*. *Res*. **18** Paper No. 127, 45. MR3763761
[49] PORTEOUS, I., NEWMAN, D., IHLER, A., ASUNCION, A., SMYTH, P. and WELLING, M. (2008). Fast collapsed Gibbs sampling for latent Dirichlet allocation. In *Proceedings of the* 14*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 569–577.
[50] PRITCHARD, J. K., STEPHENS, M. and DONNELLY, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155** 945–959.
[51] RAMAGE, D., HALL, D., NALLAPATI, R. and MANNING, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the* 2009 *Conference on Empirical Methods in Natural Language Processing* 248–256.
[52] RECHT, B., RE, C., TROPP, J. and BITTORF, V. (2012). Factoring nonnegative matrices with linear programs. In *Advances in Neural Information Processing Systems* (F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger, eds.) **25**. Curran Associates, Inc., Red Hook.
[53] SILGE, J. and ROBINSON, D. (2020). Text Mining with R: A Tidy Approach.
[54] TEH, Y. W., JORDAN, M. I., BEAL, M. J. and BLEI, D. M. (2005). Sharing clusters among related groups: Hierarchical Dirichlet processes. In *Advances in Neural Information Processing Systems* 1385–1392.
[55] TROPP, J. A., (2015). An introduction to matrix concentration inequalities. *Found*. *Trends Mach*. *Learn*. **8** 1–230.
[56] WALLACH, H. M. (2006). Topic modeling: Beyond bag-of-words. In *Proceedings of the* 23*rd International Conference on Machine Learning* 977–984.
[57] YUAN, H., XU, W., LI, Q. and LAU, R. (2018). Topic sentiment mining for sales performance prediction in e-commerce. *Ann*. *Oper*. *Res*. **270** 553–576.

[58] ZHAI, K., BOYD-GRABER, J., ASADI, N. and ALKHOUJA, M. L. (2012). Mr. LDA: A flexible large scale topic modeling package using variational inference in mapreduce. In *Proceedings of the* 21*st International Conference on World Wide Web* 879–888.

[59] ZHU, Q., ZHONG, Y., ZHANG, L. and LI, D. (2017). Scene classification based on the fully sparse semantic topic model. *IEEE Trans. Geosci. Remote Sens.* **55** 5525–5538.

The Annals of Statistics
2023, Vol. 51, No. 5, 2015–2039
https://doi.org/10.1214/23-AOS2319

ADAPTIVE AND ROBUST MULTI-TASK LEARNING

BY YAQI DUAN[1,a] AND KAIZHENG WANG[2,b]

[1] *Leonard N. Stern School of Business, New York University,* [a] *yaqi.duan@stern.nyu.edu*
[2] *Department of IEOR and Data Science Institute, Columbia University,* [b] *kaizheng.wang@columbia.edu*

We study the multitask learning problem that aims to simultaneously analyze multiple data sets collected from different sources and learn one model for each of them. We propose a family of adaptive methods that automatically utilize possible similarities among those tasks while carefully handling their differences. We derive sharp statistical guarantees for the methods and prove their robustness against outlier tasks. Numerical experiments on synthetic and real data sets demonstrate the efficacy of our new methods.

REFERENCES

[1] ANDO, R. K. and ZHANG, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Mach. Learn. Res.* **6** 1817–1853. MR2249873

[2] ANGUITA, D., GHIO, A., ONETO, L., PARRA PEREZ, X. and REYES ORTIZ, J. L. (2013). A public domain dataset for human activity recognition using smartphones. In *Proceedings of the* 21*th International European Symposium on Artificial Neural Networks*, *Computational Intelligence and Machine Learning* 437–442.

[3] ANTONIADIS, A. (2007). Wavelet methods in statistics: Some recent developments and their applications. *Stat. Surv.* **1** 16–55. MR2520413 https://doi.org/10.1214/07-SS014

[4] ARGYRIOU, A., EVGENIOU, T. and PONTIL, M. (2008). Convex multi-task feature learning. *Mach. Learn.* **73** 243–272.

[5] ASIAEE, A., OYMAK, S., COOMBES, K. R. and BANERJEE, A. (2019). Data enrichment: Multi-task learning in high dimension with theoretical guarantees. In *Adaptive and Multitask Learning Workshop at the ICML*. IMLS, Long Beach, CA.

[6] BALCAN, M.-F., KHODAK, M. and TALWALKAR, A. (2019). Provable guarantees for gradient-based meta-learning. In *International Conference on Machine Learning* 424–433. PMLR.

[7] BAXTER, J. (2000). A model of inductive bias learning. *J. Artificial Intelligence Res.* **12** 149–198. MR1752410 https://doi.org/10.1613/jair.731

[8] BEN-DAVID, S. and SCHULLER, R. (2003). Exploiting task relatedness for multiple task learning. In *Learning Theory and Kernel Machines*: 16*th Annual Conference on Learning Theory and* 7*th Kernel Workshop*, *COLT/Kernel* 2003, *Washington*, *DC*, *USA*, *August* 24–27, 2003. *Proceedings* 567–580. Springer, Berlin.

[9] BICKEL, P. J. (1983). Minimax estimation of the mean of a normal distribution subject to doing well at a point. In *Recent Advances in Statistics* 511–528. Academic Press, New York. MR0736544

[10] BICKEL, P. J. (1984). Parametric robustness: Small biases can be worthwhile. *Ann. Statist.* **12** 864–879. MR0751278 https://doi.org/10.1214/aos/1176346707

[11] BREIMAN, L. and FRIEDMAN, J. H. (1997). Predicting multivariate responses in multiple linear regression. *J. Roy. Statist. Soc. Ser. B* **59** 3–54. MR1436554 https://doi.org/10.1111/1467-9868.00054

[12] CAI, T., LIU, M. and XIA, Y. (2022). Individual data protected integrative regression analysis of high-dimensional heterogeneous data. *J. Amer. Statist. Assoc.* **117** 2105–2119. MR4528492 https://doi.org/10.1080/01621459.2021.1904958

[13] CARUANA, R. (1997). Multitask learning. *Mach. Learn.* **28** 41–75.

[14] CHEN, A., OWEN, A. B. and SHI, M. (2015). Data enriched linear regression. *Electron. J. Stat.* **9** 1078–1112. MR3352068 https://doi.org/10.1214/15-EJS1027

[15] CHEN, J., ZHOU, J. and YE, J. (2011). Integrating low-rank and group-sparse structures for robust multi-task learning. In *Proceedings of the* 17*th ACM SIGKDD international conference on Knowledge discovery and data mining* 42–50.

MSC2020 subject classifications. Primary 62R07; secondary 62F10.
Key words and phrases. Multitask learning, adaptivity, robustness, model misspecification, clustering, low-rank model.

[16] CHEN, S., ZHANG, B. and YE, T. (2021). Minimax rates and adaptivity in combining experimental and observational data. arXiv preprint. Available at arXiv:2109.10522.

[17] CHEN, S., ZHENG, Q., LONG, Q. and SU, W. J. (2021). A theorem of the alternative for personalized federated learning. arXiv preprint. Available at arXiv:2103.01901.

[18] COLLIER, O. and DALALYAN, A. S. (2019). Multidimensional linear functional estimation in sparse Gaussian models and robust estimation of the mean. *Electron. J. Stat.* **13** 2830–2864. MR3998929 https://doi.org/10.1214/19-EJS1590

[19] DENEVI, G., CILIBERTO, C., GRAZZI, R. and PONTIL, M. (2019). Learning-to-learn stochastic gradient descent with biased regularization. In *International Conference on Machine Learning* 1566–1575. PMLR.

[20] DONOHO, D. and MONTANARI, A. (2016). High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *Probab. Theory Related Fields* **166** 935–969. MR3568043 https://doi.org/10.1007/s00440-015-0675-z

[21] DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455. MR1311089 https://doi.org/10.1093/biomet/81.3.425

[22] DONOHO, D. L., JOHNSTONE, I. M., HOCH, J. C. and STERN, A. S. (1992). Maximum entropy and the nearly black object. *J. Roy. Statist. Soc. Ser. B* **54** 41–81. MR1157714

[23] DU, S. S., HU, W., KAKADE, S. M., LEE, J. D. and LEI, Q. (2020). Few-Shot Learning via Learning the Representation, Provably. In *International Conference on Learning Representations*.

[24] DUAN, Y. and WANG, K. (2023). Supplement to "Adaptive and robust multi-task learning." https://doi.org/10.1214/23-AOS2319SUPP

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

[26] EFRON, B. and MORRIS, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators. II. The empirical Bayes case. *J. Amer. Statist. Assoc.* **67** 130–139. MR0323015

[27] EFRON, B. and MORRIS, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *J. Amer. Statist. Assoc.* **68** 117–130. MR0388597

[28] EVGENIOU, T., MICCHELLI, C. A. and PONTIL, M. (2005). Learning multiple tasks with kernel methods. *J. Mach. Learn. Res.* **6** 615–637. MR2249833

[29] EVGENIOU, T. and PONTIL, M. (2004). Regularized multi-task learning. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 109–117.

[30] GANNAZ, I. (2007). Robust estimation and wavelet thresholding in partially linear models. *Stat. Comput.* **17** 293–310. MR2409795 https://doi.org/10.1007/s11222-007-9019-x

[31] HANNEKE, S. and KPOTUFE, S. (2022). A no-free-lunch theorem for multitask learning. *Ann. Statist.* **50** 3119–3143. MR4524491 https://doi.org/10.1214/22-aos2189

[32] HIRIART-URRUTY, J.-B. and LEMARÉCHAL, C. (1993). *Convex Analysis and Minimization Algorithms. I: Fundamentals. Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **305**. Springer, Berlin. MR1261420

[33] HODGES, J. L. JR. and LEHMANN, E. L. (1952). The use of previous experience in reaching statistical decisions. *Ann. Math. Stat.* **23** 396–407. MR0050240 https://doi.org/10.1214/aoms/1177729384

[34] HSU, D., KAKADE, S. M. and ZHANG, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.* **17** no. 52, 6. MR2994877 https://doi.org/10.1214/ECP.v17-2079

[35] HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.* **35** 73–101. MR0161415 https://doi.org/10.1214/aoms/1177703732

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

[37] JACOB, L., BACH, F. and VERT, J.-P. (2008). Clustered multi-task learning: A convex formulation. In *Proceedings of the* 21*st International Conference on Neural Information Processing Systems* 745–752.

[38] JALALI, A., RAVIKUMAR, P. and SANGHAVI, S. (2013). A dirty model for multiple sparse regression. *IEEE Trans. Inf. Theory* **59** 7947–7968. MR3142275 https://doi.org/10.1109/TIT.2013.2280272

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

[40] KE, Z. T., FAN, J. and WU, Y. (2015). Homogeneity pursuit. *J. Amer. Statist. Assoc.* **110** 175–194. MR3338495 https://doi.org/10.1080/01621459.2014.892882

[41] KONSTANTINOV, N., FRANTAR, E., ALISTARH, D. and LAMPERT, C. (2020). On the sample complexity of adversarial multi-source PAC learning. In *International Conference on Machine Learning* 5416–5425. PMLR.

[42] KUMAR, A. and HAL, D. III (2012). Learning task grouping and overlap in multi-task learning. In *Proceedings of the* 29*th International Coference on International Conference on Machine Learning* 1723–1730.

[43] LENZERINI, M. (2002). Data integration: A theoretical perspective. In *Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* 233–246.

[44] LEPSKIĬ, O. V. (1991). On a problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35** 454–466.

[45] LIU, G., LIN, Z. and YU, Y. (2010). Robust subspace segmentation by low-rank representation. In *Proceedings of the* 27*th International Conference on International Conference on Machine Learning* 663–670.

[46] LOUNICI, K., PONTIL, M., VAN DE GEER, S. and TSYBAKOV, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* **39** 2164–2204. MR2893865 https://doi.org/10.1214/11-AOS896

[47] LOW, M. G. (1997). On nonparametric confidence intervals. *Ann. Statist.* **25** 2547–2554. MR1604412 https://doi.org/10.1214/aos/1030741084

[48] MAITY, S., SUN, Y. and BANERJEE, M. (2019). Meta-analysis of heterogeneous data: Integrative sparse regression in high-dimensions. arXiv preprint. Available at arXiv:1912.11928.

[49] MAURER, A., PONTIL, M. and ROMERA-PAREDES, B. (2016). The benefit of multitask representation learning. *J. Mach. Learn. Res.* **17** Paper No. 81, 32. MR3517104

[50] MCCOY, M. and TROPP, J. A. (2011). Two proposals for robust PCA using semidefinite programming. *Electron. J. Stat.* **5** 1123–1160. MR2836771 https://doi.org/10.1214/11-EJS636

[51] MCDONALD, A. M., PONTIL, M. and STAMOS, D. (2016). New perspectives on k-support and cluster norms. *J. Mach. Learn. Res.* **17** Paper No. 155, 38. MR3555046

[52] MEI, S., BAI, Y. and MONTANARI, A. (2018). The landscape of empirical risk for nonconvex losses. *Ann. Statist.* **46** 2747–2774. MR3851754 https://doi.org/10.1214/17-AOS1637

[53] MOUSAVI KALAN, M., FABIAN, Z., AVESTIMEHR, S. and SOLTANOLKOTABI, M. (2020). Minimax lower bounds for transfer learning with linear and one-hidden layer neural networks. *Adv. Neural Inf. Process. Syst.* **33** 1959–1969.

[54] NEGAHBAN, S. and WAINWRIGHT, M. J. (2008). Joint support recovery under high-dimensional scaling: Benefits and perils of $\ell_{1,\infty}$-regularization. *Adv. Neural Inf. Process. Syst.* **21** 1161–1168.

[55] OBOZINSKI, G., WAINWRIGHT, M. J. and JORDAN, M. I. (2011). Support union recovery in high-dimensional multivariate regression. *Ann. Statist.* **39** 1–47. MR2797839 https://doi.org/10.1214/09-AOS776

[56] PARIKH, N. and BOYD, S. (2014). Proximal algorithms. *Found. Trends Optim.* **1** 127–239.

[57] PONG, T. K., TSENG, P., JI, S. and YE, J. (2010). Trace norm regularization: Reformulations, algorithms, and multi-task learning. *SIAM J. Optim.* **20** 3465–3489. MR2763512 https://doi.org/10.1137/090763184

[58] SHE, Y. and OWEN, A. B. (2011). Outlier detection using nonconvex penalized regression. *J. Amer. Statist. Assoc.* **106** 626–639. MR2847975 https://doi.org/10.1198/jasa.2011.tm10390

[59] SHEN, X. and HUANG, H.-C. (2010). Grouping pursuit through a regularization solution surface. *J. Amer. Statist. Assoc.* **105** 727–739. MR2724856 https://doi.org/10.1198/jasa.2010.tm09380

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

[61] TANG, L. and SONG, P. X. K. (2016). Fused lasso approach in regression coefficients clustering—learning parameter heterogeneity in data integration. *J. Mach. Learn. Res.* **17** Paper No. 113, 23. MR3543519

[62] THRUN, S. and PRATT, L. (2012). *Learning to Learn*. Springer, Berlin.

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

[64] TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 91–108. MR2136641 https://doi.org/10.1111/j.1467-9868.2005.00490.x

[65] TRIPURANENI, N., JORDAN, M. and JIN, C. (2020). On the Theory of Transfer Learning: The Importance of Task Diversity. *Adv. Neural Inf. Process. Syst.* **33**.

[66] WU, S., ZHANG, H. R. and RÉ, C. (2020). Understanding and improving information transfer in multi-task learning. arXiv preprint. Available at arXiv:2005.00944.

[67] XU, H., CARAMANIS, C. and SANGHAVI, S. (2012). Robust PCA via outlier pursuit. *IEEE Trans. Inf. Theory* **58** 3047–3064. MR2952532 https://doi.org/10.1109/TIT.2011.2173156

[68] XU, K. and BASTANI, H. (2021). Learning across bandits in high dimension via robust statistics. arXiv preprint. Available at arXiv:2112.14233.

The Annals of Statistics
2023, Vol. 51, No. 5, 2040–2066
https://doi.org/10.1214/23-AOS2320

INFERENCE FOR EXTREMAL REGRESSION WITH DEPENDENT HEAVY-TAILED DATA

BY ABDELAATI DAOUIA[1,a], GILLES STUPFLER[2,b] AND ANTOINE USSEGLIO-CARLEVE[3,c]

[1] *Toulouse School of Economics, University of Toulouse Capitole,* [a] *abdelaati.daouia@tse-fr.eu*
[2] *CNRS, LAREMA, SFR MATHSTIC, Université d'Angers,* [b] *gilles.stupfler@univ-angers.fr*
[3] *LMA UPR 2151, Avignon Université,* [c] *antoine.usseglio-carleve@univ-avignon.fr*

Nonparametric inference on tail conditional quantiles and their least squares analogs, expectiles, remains limited to i.i.d. data. We develop a fully operational inferential theory for extreme conditional quantiles and expectiles in the challenging framework of α-mixing, conditional heavy-tailed data whose tail index may vary with covariate values. This requires a dedicated treatment to deal with data sparsity in the far tail of the response, in addition to handling difficulties inherent to mixing, smoothing and sparsity associated to covariate localization. We prove the pointwise asymptotic normality of our estimators and obtain optimal rates of convergence reminiscent of those found in the i.i.d. regression setting, but which had not been established in the conditional extreme value literature. Our assumptions hold in a wide range of models. We propose full bias and variance reduction procedures, and simple but effective data-based rules for selecting tuning hyperparameters. Our inference strategy is shown to perform well in finite samples and is showcased in applications to stock returns and tornado loss data.

REFERENCES

[1] ARTZNER, P., DELBAEN, F., EBER, J.-M. and HEATH, D. (1999). Coherent measures of risk. *Math. Finance* **9** 203–228. MR1850791 https://doi.org/10.1111/1467-9965.00068

[2] BEIRLANT, J., GOEGEBEUR, Y., TEUGELS, J. and SEGERS, J. (2004). *Statistics of Extremes: Theory and Applications. Wiley Series in Probability and Statistics*. Wiley, Chichester. MR2108013 https://doi.org/10.1002/0470012382

[3] BELLINI, F. and DI BERNARDINO, E. (2017). Risk management with expectiles. *Eur. J. Finance* **23** 487–506.

[4] BRADLEY, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surv.* **2** 107–144. MR2178042 https://doi.org/10.1214/154957805100000104

[5] CHAUDHURI, P. (1991). Global nonparametric estimation of conditional quantile functions and their derivatives. *J. Multivariate Anal.* **39** 246–269. MR1147121 https://doi.org/10.1016/0047-259X(91)90100-G

[6] CHERNOZHUKOV, V. (2005). Extremal quantile regression. *Ann. Statist.* **33** 806–839. MR2163160 https://doi.org/10.1214/009053604000001165

[7] CHERNOZHUKOV, V. and FERNÁNDEZ-VAL, I. (2011). Inference for extremal conditional quantile models, with an application to market and birthweight risks. *Rev. Econ. Stud.* **78** 559–589. MR2808129 https://doi.org/10.1093/restud/rdq020

[8] CHERNOZHUKOV, V., FERNÁNDEZ-VAL, I. and KAJI, T. (2018). Extremal quantile regression. In *Handbook of Quantile Regression* (R. Koenker, V. Chernozhukov, X. He and L. Peng, eds.). *Chapman & Hall/CRC Handb. Mod. Stat. Methods* 333–362. CRC Press, Boca Raton, FL. MR3821131

[9] DAOUIA, A., GARDES, L. and GIRARD, S. (2013). On kernel smoothing for extremal quantile regression. *Bernoulli* **19** 2557–2589. MR3160564 https://doi.org/10.3150/12-BEJ466

[10] DAOUIA, A., GARDES, L., GIRARD, S. and LEKINA, A. (2011). Kernel estimators of extreme level curves. *TEST* **20** 311–333. MR2834049 https://doi.org/10.1007/s11749-010-0196-0

MSC2020 subject classifications. Primary 62G32; secondary 62G05, 62G08, 62G15, 62G20, 62G30.
Key words and phrases. Conditional quantiles, conditional expectiles, extreme value analysis, heavy tails, inference, mixing, nonparametric regression.

[11] DAOUIA, A., GIRARD, S. and STUPFLER, G. (2018). Estimation of tail risk based on extreme expectiles. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 263–292. MR3763692 https://doi.org/10.1111/rssb.12254

[12] DAOUIA, A., STUPFLER, G. and USSEGLIO-CARLEVE, A. (2023). Supplement to "Inference for extremal regression with dependent heavy-tailed data." https://doi.org/10.1214/23-AOS2320SUPP

[13] DAVISON, A. C., PADOAN, S. A. and STUPFLER, G. (2023). Tail risk inference via expectiles in heavy-tailed time series. *J. Bus. Econom. Statist.* **41** 876–889. MR4600855 https://doi.org/10.1080/07350015.2022.2078332

[14] DE HAAN, L. and FERREIRA, A. (2006). *Extreme Value Theory: An Introduction. Springer Series in Operations Research and Financial Engineering*. Springer, New York. MR2234156 https://doi.org/10.1007/0-387-34471-3

[15] DREES, H. (2003). Extreme quantile estimation for dependent data, with applications to finance. *Bernoulli* **9** 617–657. MR1996273 https://doi.org/10.3150/bj/1066223272

[16] GIRARD, S., STUPFLER, G. and USSEGLIO-CARLEVE, A. (2021). Extreme conditional expectile estimation in heavy-tailed heteroscedastic regression models. *Ann. Statist.* **49** 3358–3382. MR4352533 https://doi.org/10.1214/21-aos2087

[17] GIRARD, S., STUPFLER, G. and USSEGLIO-CARLEVE, A. (2022). Nonparametric extreme conditional expectile estimation. *Scand. J. Stat.* **49** 78–115. MR4391048 https://doi.org/10.1111/sjos.12502

[18] GIRARD, S., STUPFLER, G. and USSEGLIO-CARLEVE, A. (2022). On automatic bias reduction for extreme expectile estimation. *Stat. Comput.* **32** Paper No. 64, 18 pp. MR4466976 https://doi.org/10.1007/s11222-022-10118-x

[19] GOMES, M. I. and PESTANA, D. (2007). A sturdy reduced-bias extreme quantile (VaR) estimator. *J. Amer. Statist. Assoc.* **102** 280–292. MR2345543 https://doi.org/10.1198/016214506000000799

[20] JONES, M. C. (1994). Expectiles and M-quantiles are quantiles. *Statist. Probab. Lett.* **20** 149–153. MR1293293 https://doi.org/10.1016/0167-7152(94)90031-0

[21] LINTON, O. and XIAO, Z. (2013). Estimation of and inference about the expected shortfall for time series with infinite variance. *Econometric Theory* **29** 771–807. MR3092463 https://doi.org/10.1017/S0266466612000692

[22] NEWEY, W. K. and POWELL, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica* **55** 819–847. MR0906565 https://doi.org/10.2307/1911031

[23] SMITH, R. L. and WEISSMAN, I. (1996). Characterization and estimation of the multivariate extremal index. Technical report, Univ. North Carolina.

[24] USSEGLIO-CARLEVE, A. (2018). Estimation of conditional extreme risk measures from heavy-tailed elliptical random vectors. *Electron. J. Stat.* **12** 4057–4093. MR3887178 https://doi.org/10.1214/18-EJS1499

[25] WANG, H. J., LI, D. and HE, X. (2012). Estimation of high conditional quantiles for heavy-tailed distributions. *J. Amer. Statist. Assoc.* **107** 1453–1464. MR3036407 https://doi.org/10.1080/01621459.2012.716382

[26] WASSERMAN, L. (2006). *All of Nonparametric Statistics. Springer Texts in Statistics*. Springer, New York. MR2172729

[27] WEISSMAN, I. (1978). Estimation of parameters and large quantiles based on the k largest observations. *J. Amer. Statist. Assoc.* **73** 812–815. MR0521329

[28] YEE, T. W. and WILD, C. J. (1996). Vector generalized additive models. *J. Roy. Statist. Soc. Ser. B* **58** 481–493. MR1394361

[29] ZIEGEL, J. F. (2016). Coherence and elicitability. *Math. Finance* **26** 901–918. MR3551510 https://doi.org/10.1111/mafi.12080

The Annals of Statistics
2023, Vol. 51, No. 5, 2067–2092
https://doi.org/10.1214/23-AOS2321

DIFFERENTIALLY PRIVATE INFERENCE VIA NOISY OPTIMIZATION

BY MARCO AVELLA-MEDINA[1,a], CASEY BRADSHAW[1,b] AND PO-LING LOH[2,c]

[1]*Department of Statistics, Columbia University,* [a]*marco.avella@columbia.edu,* [b]*cb3431@columbia.edu*
[2]*Department of Pure Mathematics and Mathematical Statistics, University of Cambridge,* [c]*pll28@cam.ac.uk*

We propose a general optimization-based framework for computing differentially private M-estimators and a new method for constructing differentially private confidence regions. First, we show that robust statistics can be used in conjunction with noisy gradient descent or noisy Newton methods in order to obtain optimal private estimators with global linear or quadratic convergence, respectively. We establish local and global convergence guarantees, under both local strong convexity and self-concordance, showing that our private estimators converge with high probability to a small neighborhood of the nonprivate M-estimators. Second, we tackle the problem of parametric inference by constructing differentially private estimators of the asymptotic variance of our private M-estimators. This naturally leads to approximate pivotal statistics for constructing confidence regions and conducting hypothesis testing. We demonstrate the effectiveness of a bias correction that leads to enhanced small-sample empirical performance in simulations. We illustrate the benefits of our methods in several numerical examples.

REFERENCES

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

[2] ACHARYA, J., SUN, Z. and ZHANG, H. (2018). Differentially private testing of identity and closeness of discrete distributions. *Adv. Neural Inf. Process. Syst.* **31** 6878–6891.

[3] AGARWAL, A., BARTLETT, P. L., RAVIKUMAR, P. and WAINWRIGHT, M. J. (2012). Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Trans. Inf. Theory* **58** 3235–3249. MR2952543 https://doi.org/10.1109/TIT.2011.2182178

[4] AVELLA-MEDINA, M. (2021). Privacy-preserving parametric inference: A case for robust statistics. *J. Amer. Statist. Assoc.* **116** 969–983. MR4270037 https://doi.org/10.1080/01621459.2019.1700130

[5] AVELLA-MEDINA, M., BRADSHAW, C. and LOH, P.-L. (2023). Supplement to "Differentially private inference via noisy optimization." https://doi.org/10.1214/23-AOS2321SUPPA, https://doi.org/10.1214/23-AOS2321SUPPB

[6] AWAN, J. and SLAVKOVIĆ, A. (2018). Differentially private uniformly most powerful tests for binomial data. *Adv. Neural Inf. Process. Syst.* **2018** 4208–4218.

[7] BALLE, B., KAIROUZ, P., MCMAHAN, B., THAKKAR, O. D. and THAKURTA, A. (2020). Privacy amplification via random check-ins. *Adv. Neural Inf. Process. Syst.* **33**.

[8] BARBER, R. F. and DUCHI, J. (2014). Privacy: A few definitional aspects and consequences for minimax mean-squared error. In 53*rd IEEE Conference on Decision and Control* 1365–1369. IEEE, New York.

[9] BARRIENTOS, A. F., REITER, J. P., MACHANAVAJJHALA, A. and CHEN, Y. (2019). Differentially private significance tests for regression coefficients. *J. Comput. Graph. Statist.* **28** 440–453. MR3974892 https://doi.org/10.1080/10618600.2018.1538881

[10] BASSILY, R., FELDMAN, V., TALWAR, K. and THAKURTA, A. (2019). Private stochastic convex optimization with optimal rates. Preprint. Available at arXiv:1908.09970.

[11] BASSILY, R., SMITH, A. and THAKURTA, A. (2014). Private empirical risk minimization: Efficient algorithms and tight error bounds. In 55*th Annual IEEE Symposium on Foundations of Computer Science—FOCS* 2014 464–473. IEEE Computer Soc., Los Alamitos, CA. MR3344896 https://doi.org/10.1109/FOCS.2014.56

MSC2020 subject classifications. Primary 62F10; secondary 68P27.
Key words and phrases. Differential privacy, M-estimation, statistical inference, gradient descent, Newton's method.

[12] BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge Univ. Press, Cambridge. MR2061575 https://doi.org/10.1017/CBO9780511804441

[13] BU, Z., DONG, J., LONG, Q. and SU, W. J. (2020). Deep learning with Gaussian differential privacy. *Harv. Data Sci. Rev.* **2020**.

[14] BUBECK, S. (2015). Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.* **8** 231–357.

[15] CAI, T. T., WANG, Y. and ZHANG, L. (2020). The cost of privacy in generalized linear models: Algorithms and minimax lower bounds. Preprint. Available at arXiv:2011.03900.

[16] CAI, T. T., WANG, Y. and ZHANG, L. (2021). The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *Ann. Statist.* **49** 2825–2850. MR4338894 https://doi.org/10.1214/21-aos2058

[17] CHADHA, K., DUCHI, J. and KUDITIPUDI, R. (2021). Private confidence sets. In *NeurIPS* 2021 *Workshop Privacy in Machine Learning*.

[18] CHAUDHURI, K. and MONTELEONI, C. (2008). Privacy-preserving logistic regression. In *Advances in Neural Information Processing Systems* **22** 289–296. Citeseer.

[19] CHAUDHURI, K., MONTELEONI, C. and SARWATE, A. D. (2011). Differentially private empirical risk minimization. *J. Mach. Learn. Res.* **12** 1069–1109. MR2786918

[20] COVINGTON, C., HE, X., HONAKER, J. and KAMATH, G. (2021). Unbiased statistical estimation and valid confidence intervals under differential privacy. Preprint. Available at arXiv:2110.14465.

[21] D'ASPREMONT, A. (2008). Smooth optimization with approximate gradient. *SIAM J. Optim.* **19** 1171–1183. MR2460737 https://doi.org/10.1137/060676386

[22] DEVOLDER, O., GLINEUR, F. and NESTEROV, Y. (2014). First-order methods of smooth convex optimization with inexact oracle. *Math. Program.* **146** 37–75. MR3232608 https://doi.org/10.1007/s10107-013-0677-5

[23] DING, B., KULKARNI, J. and YEKHANIN, S. (2017). Collecting telemetry data privately. In *Proceedings of the* 31*st International Conference on Neural Information Processing Systems* 3574–3583.

[24] DONG, J., ROTH, A. and SU, W. J. (2022). Gaussian differential privacy. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 3–54. MR4400389

[25] DUCHI, J. C., JORDAN, M. I. and WAINWRIGHT, M. J. (2018). Minimax optimal procedures for locally private estimation. *J. Amer. Statist. Assoc.* **113** 182–201. MR3803452 https://doi.org/10.1080/01621459.2017.1389735

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

[27] DWORK, C., TALWAR, K., THAKURTA, A. and ZHANG, L. (2014). Analyze Gauss: Optimal bounds for privacy-preserving principal component analysis. In *STOC'*14*—Proceedings of the* 2014 *ACM Symposium on Theory of Computing* 11–20. ACM, New York. MR3238926 https://doi.org/10.1145/2591796.2591883

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

[29] EVFIMIEVSKI, A., GEHRKE, J. and SRIKANT, R. (2003). Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* 211–222.

[30] FELDMAN, V., KOREN, T. and TALWAR, K. (2020). Private stochastic convex optimization: Optimal rates in linear time. In *STOC '*20*—Proceedings of the* 52*nd Annual ACM SIGACT Symposium on Theory of Computing* 439–449. ACM, New York. MR4141772

[31] GABOARDI, M., LIM, H., ROGERS, R. and VADHAN, S. (2016). Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *International Conference on Machine Learning* 2111–2120. PMLR.

[32] GANESH, A., HAGHIFAM, M., STEINKE, T. and THAKURTA, A. (2023). Faster differentially private convex optimization via second-order methods. Preprint. Available at arXiv:2305.13209.

[33] GARFINKEL, S., ABOWD, J. M. and MARTINDALE, C. (2019). Understanding database reconstruction attacks on public data. *Commun. ACM* **62** 46–53.

[34] GHADIMI, S. and LAN, G. (2012). Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework. *SIAM J. Optim.* **22** 1469–1492. MR3023780 https://doi.org/10.1137/110848864

[35] HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. and STAHEL, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. *Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics*. Wiley, New York. MR0829458

[36] HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.* **35** 73–101. MR0161415 https://doi.org/10.1214/aoms/1177703732

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

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

[39] IYENGAR, R., NEAR, J. P., SONG, D., THAKKAR, O., THAKURTA, A. and WANG, L. (2019). Towards practical differentially private convex optimization. In 2019 *IEEE Symposium on Security and Privacy* (*SP*) 299–316. IEEE, New York.

[40] JAIN, P., KOTHARI, P. and THAKURTA, A. (2012). Differentially private online learning. In *Conference on Learning Theory* 24–1. JMLR Workshop and Conference Proceedings.

[41] JAIN, P. and THAKURTA, A. G. (2014). (Near) dimension independent risk bounds for differentially private learning. In *International Conference on Machine Learning* 476–484. PMLR.

[42] KARIMIREDDY, S. P., STICH, S. U. and JAGGI, M. (2018). Global linear convergence of Newton's method without strong-convexity or Lipschitz gradients. Preprint. Available at arXiv:1806.00413.

[43] KARWA, V. and VADHAN, S. (2017). Finite sample differentially private confidence intervals. Preprint. Available at arXiv:1711.03908.

[44] KASIVISWANATHAN, S. P., LEE, H. K., NISSIM, K., RASKHODNIKOVA, S. and SMITH, A. (2011). What can we learn privately? *SIAM J. Comput.* **40** 793–826. MR2823508 https://doi.org/10.1137/090756090

[45] KIFER, D., SMITH, A. and THAKURTA, A. (2012). Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory* 25–1. JMLR Workshop and Conference Proceedings.

[46] KULKARNI, T., JÄLKÖ, J., KOSKELA, A., KASKI, S. and HONKELA, A. (2021). Differentially private Bayesian inference for generalized linear models. In *International Conference on Machine Learning* 5838–5849. PMLR.

[47] LEE, J. and KIFER, D. (2018). Concentrated differentially private gradient descent with adaptive per-iteration privacy budget. In *Proceedings of the* 24*th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 1656–1665.

[48] LOH, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *Ann. Statist.* **45** 866–896. MR3650403 https://doi.org/10.1214/16-AOS1471

[49] LOH, P.-L. and WAINWRIGHT, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *J. Mach. Learn. Res.* **16** 559–616. MR3335800

[50] NESTEROV, Y. (2018). *Lectures on Convex Optimization*, 2nd ed. *Springer Optimization and Its Applications* **137**. Springer, Cham. MR3839649 https://doi.org/10.1007/978-3-319-91578-4

[51] PEÑA, V. and BARRIENTOS, A. F. (2021). Differentially private methods for managing model uncertainty in linear regression models. Preprint. Available at arXiv:2109.03949.

[52] RAJKUMAR, A. and AGARWAL, S. (2012). A differentially private stochastic gradient descent algorithm for multiparty classification. In *Artificial Intelligence and Statistics* 933–941. PMLR.

[53] ROGERS, R. and KIFER, D. (2017). A new class of private chi-square hypothesis tests. In *Artificial Intelligence and Statistics* 991–1000. PMLR.

[54] ROOSTA-KHORASANI, F. and MAHONEY, M. W. (2019). Sub-sampled Newton methods. *Math. Program.* **174** 293–326. MR3935082 https://doi.org/10.1007/s10107-018-1346-5

[55] SAVITSKY, T. D., WILLIAMS, M. R. and HU, J. (2022). Bayesian pseudo posterior mechanism under asymptotic differential privacy. *J. Mach. Learn. Res.* **23** Paper No. [55], 37 pp. MR4420780 https://doi.org/10.4995/agt.2022.16126

[56] SHEFFET, O. (2017). Differentially private ordinary least squares. In *International Conference on Machine Learning* 3105–3114. PMLR.

[57] SLAVKOVIC, A. and MOLINARI, R. (2021). Perturbed M-estimation: A further investigation of robust statistics for differential privacy. Preprint. Available at arXiv:2108.08266.

[58] SONG, S., CHAUDHURI, K. and SARWATE, A. D. (2013). Stochastic gradient descent with differentially private updates. In 2013 *IEEE Global Conference on Signal and Information Processing* 245–248. IEEE, New York.

[59] SONG, S., STEINKE, T., THAKKAR, O. and THAKURTA, A. (2021). Evading the curse of dimensionality in unconstrained private GLMs. In *International Conference on Artificial Intelligence and Statistics* 2638–2646. PMLR.

[60] SUN, T., NECOARA, I. and TRAN-DINH, Q. (2020). Composite convex optimization with global and local inexact oracles. *Comput. Optim. Appl.* **76** 69–124. MR4081182 https://doi.org/10.1007/s10589-020-00174-2

[61] SUN, T. and TRAN-DINH, Q. (2019). Generalized self-concordant functions: A recipe for Newton-type methods. *Math. Program.* **178** 145–213. MR4019949 https://doi.org/10.1007/s10107-018-1282-4

[62] TALWAR, K., THAKURTA, A. and ZHANG, L. (2015). Nearly-optimal private LASSO. In *Proceedings of the* 28*th International Conference on Neural Information Processing Systems—Volume* 2 3025–3033.

[63] TANG, J., KOROLOVA, A., BAI, X., WANG, X. and WANG, X. (2017). Privacy loss in Apple's implementation of differential privacy on MacOS 10.12. Preprint. Available at arXiv:1709.02753.

[64] UHLER, C., SLAVKOVIĆ, A. and FIENBERG, S. E. (2013). Privacy-preserving data sharing for genome-wide association studies. *J. Priv. Confid.* **5** 137.

[65] VU, D. and SLAVKOVIC, A. (2009). Differential privacy for clinical trial data: Preliminary evaluations. In 2009 *IEEE International Conference on Data Mining Workshops* 138–143. IEEE, New York.

[66] WANG, D., YE, M. and XU, J. (2017). Differentially private empirical risk minimization revisited: Faster and more general. In *Proceedings of the* 31*st International Conference on Neural Information Processing Systems* 2719–2728.

[67] WANG, X., MA, S., GOLDFARB, D. and LIU, W. (2017). Stochastic quasi-Newton methods for nonconvex stochastic optimization. *SIAM J. Optim.* **27** 927–956. MR3651489 https://doi.org/10.1137/15M1053141

[68] WANG, Y., FIENBERG, S. and SMOLA, A. (2015). Privacy for free: Posterior sampling and stochastic gradient Monte Carlo. In *International Conference on Machine Learning* 2493–2502. PMLR.

[69] WANG, Y., KIFER, D. and LEE, J. (2019). Differentially private confidence intervals for empirical risk minimization. *J. Priv. Confid.* **9**.

[70] WANG, Y. X. (2018). Revisiting differentially private linear regression: Optimal and adaptive prediction & estimation in unbounded domain. Preprint. Available at arXiv:1803.02596v2.

[71] WARNER, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *J. Amer. Statist. Assoc.* **60** 63–69.

[72] WASSERMAN, L. and ZHOU, S. (2010). A statistical framework for differential privacy. *J. Amer. Statist. Assoc.* **105** 375–389. MR2656057 https://doi.org/10.1198/jasa.2009.tm08651

[73] YU, F., RYBAR, M., UHLER, C. and FIENBERG, S. E. (2014). Differentially-private logistic regression for detecting multiple-SNP association in GWAS databases. In *International Conference on Privacy in Statistical Databases* 170–184. Springer, Berlin.

The Annals of Statistics
2023, Vol. 51, No. 5, 2093–2115
https://doi.org/10.1214/23-AOS2322

ROBUST HIGH-DIMENSIONAL TUNING FREE MULTIPLE TESTING

BY JIANQING FAN[1,a], ZHIPENG LOU[2,b] AND MENGXIN YU[3,c]

[1]*Department of Operations Research and Financial Engineering, Princeton University,* [a]*jqfan@princeton.edu*
[2]*Department of Statistics, University of Pittsburgh,* [b]*ZHL318@pitt.edu*
[3]*Department of Statistics and Data Science, University of Pennsylvania,* [c]*mengxiny@wharton.upenn.edu*

A stylized feature of high-dimensional data is that many variables have heavy tails, and robust statistical inference is critical for valid large-scale statistical inference. Yet, the existing developments such as Winsorization, Huberization and median of means require the bounded second moments and involve variable-dependent tuning parameters, which hamper their fidelity in applications to large-scale problems. To liberate these constraints, this paper revisits the celebrated Hodges–Lehmann (HL) estimator for estimating location parameters in both the one- and two-sample problems, from a nonasymptotic perspective. Our study develops Berry–Esseen inequality and Cramér-type moderate deviation for the HL estimator based on newly developed nonasymptotic Bahadur representation and builds data-driven confidence intervals via a weighted bootstrap approach. These results allow us to extend the HL estimator to large-scale studies and propose *tuning-free* and *moment-free* high-dimensional inference procedures for testing global null and for large-scale multiple testing with false discovery proportion control. It is convincingly shown that the resulting tuning-free and moment-free methods control false discovery proportion at a prescribed level. The simulation studies lend further support to our developed theory.

REFERENCES

[1] ARCONES, M. A. (1995). The asymptotic accuracy of the bootstrap of U-quantiles. *Ann. Statist.* **23** 1802–1822. MR1370308 https://doi.org/10.1214/aos/1176324324

[2] ARCONES, M. A. (1996). The Bahadur-Kiefer representation for U-quantiles. *Ann. Statist.* **24** 1400–1422. MR1401857 https://doi.org/10.1214/aos/1032526976

[3] BAI, Z. and SARANADASA, H. (1996). Effect of high dimension: By an example of a two sample problem. *Statist. Sinica* **6** 311–329. MR1399305

[4] BAUER, D. F. (1972). Constructing confidence sets using rank statistics. *J. Amer. Statist. Assoc.* **67** 687–690.

[5] BELLONI, A. and CHERNOZHUKOV, V. (2011). ℓ_1-penalized quantile regression in high-dimensional sparse models. *Ann. Statist.* **39** 82–130. MR2797841 https://doi.org/10.1214/10-AOS827

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

[7] BLANCHARD, G. and ROQUAIN, É. (2009). Adaptive false discovery rate control under independence and dependence. *J. Mach. Learn. Res.* **10** 2837–2871. MR2579914

[8] BROWN, B. and KILDEA, D. (1978). Reduced U-statistics and the Hodges–Lehmann estimator. *Ann. Statist.* **6** 828–835. MR0491556 https://doi.org/10.1214/aos/1176344256

[9] BROWNLEES, C., JOLY, E. and LUGOSI, G. (2015). Empirical risk minimization for heavy-tailed losses. *Ann. Statist.* **43** 2507–2536. MR3405602 https://doi.org/10.1214/15-AOS1350

[10] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer Series in Statistics.* Springer, Heidelberg. MR2807761 https://doi.org/10.1007/978-3-642-20192-9

[11] CATONI, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 1148–1185. MR3052407 https://doi.org/10.1214/11-AIHP454

[12] CHANG, J., CHEN, X. and WU, M. (in press). Central limit theorems for high dimensional dependent data. *Bernoulli.*

MSC2020 subject classifications. 62F35, 62F40, 62J15.
Key words and phrases. Robust statistical inference, heavy-tailed data, tuning free, weighted bootstrap, large-scale multiple testing.

[13] CHANG, J., ZHENG, C., ZHOU, W.-X. and ZHOU, W. (2017). Simulation-based hypothesis testing of high dimensional means under covariance heterogeneity. *Biometrics* **73** 1300–1310. MR3744543 https://doi.org/10.1111/biom.12695

[14] CHEN, S. X. and QIN, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.* **38** 808–835. MR2604697 https://doi.org/10.1214/09-AOS716

[15] CHEN, X. (2018). Gaussian and bootstrap approximations for high-dimensional U-statistics and their applications. *Ann. Statist.* **46** 642–678. MR3782380 https://doi.org/10.1214/17-AOS1563

[16] CHEN, X. and ZHOU, W.-X. (2020). Robust inference via multiplier bootstrap. *Ann. Statist.* **48** 1665–1691. MR4124339 https://doi.org/10.1214/19-AOS1863

[17] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2017). Central limit theorems and bootstrap in high dimensions. *Ann. Probab.* **45** 2309–2352. MR3693963 https://doi.org/10.1214/16-AOP1113

[18] CHERNOZHUOKOV, V., CHETVERIKOV, D., KATO, K. and KOIKE, Y. (2022). Improved central limit theorem and bootstrap approximations in high dimensions. *Ann. Statist.* **50** 2562–2586. MR4500619 https://doi.org/10.1214/22-aos2193

[19] CHI, Z. (2007). On the performance of FDR control: Constraints and a partial solution. *Ann. Statist.* **35** 1409–1431. MR2351091 https://doi.org/10.1214/009053607000000037

[20] CONT, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quant. Finance* **1** 223.

[21] DEHLING, H. and MIKOSCH, T. (1994). Random quadratic forms and the bootstrap for U-statistics. *J. Multivariate Anal.* **51** 392–413. MR1321305 https://doi.org/10.1006/jmva.1994.1069

[22] EKLUND, A., NICHOLS, T. E. and KNUTSSON, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proc. Natl. Acad. Sci. USA* **113** 7900–7905.

[23] FAN, J., FAN, Y. and BARUT, E. (2014). Adaptive robust variable selection. *Ann. Statist.* **42** 324–351. MR3189488 https://doi.org/10.1214/13-AOS1191

[24] FAN, J., GU, Y. and ZHOU, W.-X. (2022). How do noise tails impact on deep ReLU networks? Preprint. Available at arXiv:2203.10418.

[25] FAN, J., HALL, P. and YAO, Q. (2007). To how many simultaneous hypothesis tests can normal, Student's t or bootstrap calibration be applied? *J. Amer. Statist. Assoc.* **102** 1282–1288. MR2372536 https://doi.org/10.1198/016214507000000969

[26] FAN, J., KE, Y., SUN, Q. and ZHOU, W.-X. (2019). FarmTest: Factor-adjusted robust multiple testing with approximate false discovery control. *J. Amer. Statist. Assoc.* **114** 1880–1893. MR4047307 https://doi.org/10.1080/01621459.2018.1527700

[27] FAN, J., LI, Q. and WANG, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 247–265. MR3597972 https://doi.org/10.1111/rssb.12166

[28] FAN, J., LI, R., ZHANG, C.-H. and ZOU, H. (2020). *Statistical Foundations of Data Science*. CRC Press, Boca Raton.

[29] FAN, J., LIAO, Y. and YAO, J. (2015). Power enhancement in high-dimensional cross-sectional tests. *Econometrica* **83** 1497–1541. MR3384226 https://doi.org/10.3982/ECTA12749

[30] FAN, J., LOU, Z. and YU, M. (2023). Supplement to "Robust high-dimensional tuning free multiple testing." https://doi.org/10.1214/23-AOS2322SUPP

[31] FAN, J., LOU, Z. and YU, M. (2023). Are latent factor regression and sparse regression adequate? *J. Amer. Statist. Assoc.* 1–13.

[32] FAN, J., MA, C. and WANG, K. (2020). Comment on "A tuning-free robust and efficient approach to high-dimensional regression" [MR4189748]. *J. Amer. Statist. Assoc.* **115** 1720–1725. MR4189751 https://doi.org/10.1080/01621459.2020.1837138

[33] FAN, J., WANG, W. and ZHU, Z. (2021). A shrinkage principle for heavy-tailed data: High-dimensional robust low-rank matrix recovery. *Ann. Statist.* **49** 1239–1266. MR4298863 https://doi.org/10.1214/20-aos1980

[34] FAN, J., YANG, Z. and YU, M. (2022). Understanding implicit regularization in over-parameterized single index model. *J. Amer. Statist. Assoc.* 1–14.

[35] FAN, J. and YAO, Q. (2017). *The Elements of Financial Econometrics*. Cambridge Univ. Press, Cambridge.

[36] FANG, X., LUO, L. and SHAO, Q.-M. (2020). A refined Cramér-type moderate deviation for sums of local statistics. *Bernoulli* **26** 2319–2352. MR4091111 https://doi.org/10.3150/20-BEJ1195

[37] FERREIRA, J. A. and ZWINDERMAN, A. H. (2006). On the Benjamini-Hochberg method. *Ann. Statist.* **34** 1827–1849. MR2283719 https://doi.org/10.1214/009053606000000425

[38] FINOTELLO, F. and CAMILLO, B. D. (2015). Measuring differential gene expression with RNA-seq: Challenges and strategies for data analysis. *Brief. Funct. Genomics* **14** 130–142. https://doi.org/10.1093/bfgp/elu035

[39] GENOVESE, C. and WASSERMAN, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.* **32** 1035–1061. MR2065197 https://doi.org/10.1214/009053604000000283

[40] GOLDSTEIN, L., MINSKER, S. and WEI, X. (2018). Structured signal recovery from non-linear and heavy-tailed measurements. *IEEE Trans. Inf. Theory* **64** 5513–5530. MR3832320 https://doi.org/10.1109/TIT.2018.2842216

[41] GUPTA, S., ELLIS, S. E., ASHAR, F. N., MOES, A., BADER, J. S., ZHAN, J., WEST, A. B. and ARKING, D. E. (2014). Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nat. Commun.* **5** 1–8.

[42] HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2722294 https://doi.org/10.1007/978-0-387-84858-7

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

[44] HE, X. and SHAO, Q.-M. (1996). A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. *Ann. Statist.* **24** 2608–2630. MR1425971 https://doi.org/10.1214/aos/1032181172

[45] HE, X. and SHAO, Q.-M. (2000). On parameters of increasing dimensions. *J. Multivariate Anal.* **73** 120–135. MR1766124 https://doi.org/10.1006/jmva.1999.1873

[46] HODGES, J. L. JR. and LEHMANN, E. L. (1963). Estimates of location based on rank tests. *Ann. Math. Stat.* **34** 598–611. MR0152070 https://doi.org/10.1214/aoms/1177704172

[47] HØYLAND, A. (1965). Robustness of the Hodges-Lehmann estimates for shift. *Ann. Math. Stat.* **36** 174–197. MR0175252 https://doi.org/10.1214/aoms/1177700281

[48] HSU, D. and SABATO, S. (2016). Loss minimization and parameter estimation with heavy tails. *J. Mach. Learn. Res.* **17** Paper No. 18, 40. MR3491112

[49] HUBER, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1** 799–821. MR0356373

[50] JANSSEN, P. (1994). Weighted bootstrapping of U-statistics. *J. Statist. Plann. Inference* **38** 31–41. MR1256846 https://doi.org/10.1016/0378-3758(92)00156-X

[51] KOENKER, R. and HALLOCK, K. F. (2001). Quantile regression. *J. Econ. Perspect.* **15** 143–156.

[52] LEHMANN, E. L. (1963). Nonparametric confidence intervals for a shift parameter. *Ann. Math. Stat.* **34** 1507–1512. MR0164412 https://doi.org/10.1214/aoms/1177703882

[53] LI, J., WITTEN, D. M., JOHNSTONE, I. M. and TIBSHIRANI, R. (2012). Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics* **13** 523–538.

[54] LI, J. and CHEN, S. X. (2012). Two sample tests for high-dimensional covariance matrices. *Ann. Statist.* **40** 908–940. MR2985938 https://doi.org/10.1214/12-AOS993

[55] LI, J. and TIBSHIRANI, R. (2013). Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. *Stat. Methods Med. Res.* **22** 519–536. MR3190673 https://doi.org/10.1177/0962280211428386

[56] LIU, W. and SHAO, Q.-M. (2014). Phase transition and regularized bootstrap in large-scale t-tests with false discovery rate control. *Ann. Statist.* **42** 2003–2025. MR3262475 https://doi.org/10.1214/14-AOS1249

[57] LOH, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *Ann. Statist.* **45** 866–896. MR3650403 https://doi.org/10.1214/16-AOS1471

[58] MAMMEN, E. (1989). Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *Ann. Statist.* **17** 382–400. MR0981457 https://doi.org/10.1214/aos/1176347023

[59] MINSKER, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli* **21** 2308–2335. MR3378468 https://doi.org/10.3150/14-BEJ645

[60] MINSKER, S. (2018). Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *Ann. Statist.* **46** 2871–2903. MR3851758 https://doi.org/10.1214/17-AOS1642

[61] NAGALAKSHMI, U., WANG, Z., WAERN, K., SHOU, C., RAHA, D., GERSTEIN, M. and SNYDER, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320** 1344–1349.

[62] NEMIROVSKY, A. S. and YUDIN, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization*. *Wiley-Interscience Series in Discrete Mathematics*. Wiley, New York. MR0702836

[63] PETROV, V. V. (1975). *Sums of Independent Random Variables*. *Ergebnisse der Mathematik und Ihrer Grenzgebiete* [*Results in Mathematics and Related Areas*], *Band* 82. Springer, New York. MR0388499

[64] ROSENKRANZ, G. K. (2010). A note on the Hodges–Lehmann estimator. *Pharm. Stat.* **9** 162–167. https://doi.org/10.1002/pst.387

[65] SHENDURE, J. and JI, H. (2008). Next-generation DNA sequencing. *Nat. Biotechnol.* **26** 1135–1145. https://doi.org/10.1038/nbt1486

[66] STOCK, J. H. and WATSON, M. W. (2002). Macroeconomic forecasting using diffusion indexes. *J. Bus. Econom. Statist.* **20** 147–162. MR1963257 https://doi.org/10.1198/073500102317351921

[67] STOREY, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 479–498. MR1924302 https://doi.org/10.1111/1467-9868.00346

[68] STOREY, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann. Statist.* **31** 2013–2035. MR2036398 https://doi.org/10.1214/aos/1074290335

[69] SUN, Q., ZHOU, W.-X. and FAN, J. (2020). Adaptive Huber regression. *J. Amer. Statist. Assoc.* **115** 254–265. MR4078461 https://doi.org/10.1080/01621459.2018.1543124

[70] WAINWRIGHT, M. J. (2019). *High-Dimensional Statistics: A Non-asymptotic Viewpoint. Cambridge Series in Statistical and Probabilistic Mathematics* **48**. Cambridge Univ. Press, Cambridge. MR3967104 https://doi.org/10.1017/9781108627771

[71] WANG, B. and FAN, J. (2022). Robust matrix completion with heavy-tailed noise. Preprint. Available at arXiv:2206.04276.

[72] WANG, L., PENG, B., BRADIC, J., LI, R. and WU, Y. (2020). A tuning-free robust and efficient approach to high-dimensional regression. *J. Amer. Statist. Assoc.* **115** 1700–1714. MR4189748 https://doi.org/10.1080/01621459.2020.1840989

[73] WANG, L., PENG, B. and LI, R. (2015). A high-dimensional nonparametric multivariate test for mean vector. *J. Amer. Statist. Assoc.* **110** 1658–1669. MR3449062 https://doi.org/10.1080/01621459.2014.988215

[74] WANG, Q. and JING, B.-Y. (2004). Weighted bootstrap for U-statistics. *J. Multivariate Anal.* **91** 177–198. MR2087842 https://doi.org/10.1016/j.jmva.2004.01.002

[75] CHEN, Y., CHI, Y., FAN, J., MA, C. et al. (2021). Spectral methods for data science: A statistical perspective. *Found. Trends Mach. Learn.* **14** 566–806.

[76] WANG, Z., GERSTEIN, M. and SNYDER, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10** 57–63. https://doi.org/10.1038/nrg2484

[77] XIA, Y., CAI, T. T. and LI, H. (2018). Joint testing and false discovery rate control in high-dimensional multivariate regression. *Biometrika* **105** 249–269. MR3804401 https://doi.org/10.1093/biomet/asx085

[78] XU, G., LIN, L., WEI, P. and PAN, W. (2016). An adaptive two-sample test for high-dimensional means. *Biometrika* **103** 609–624. MR3551787 https://doi.org/10.1093/biomet/asw029

[79] YANG, Z., BALASUBRAMANIAN, K. and LIU, H. (2017). High-dimensional non-Gaussian single index models via thresholded score function estimation. In *International Conference on Machine Learning* 3851–3860. PMLR.

[80] YOHAI, V. J. and MARONNA, R. A. (1979). Asymptotic behavior of M-estimators for the linear model. *Ann. Statist.* **7** 258–268. MR0520237

[81] ZHANG, J.-T., GUO, J., ZHOU, B. and CHENG, M.-Y. (2020). A simple two-sample test in high dimensions based on L^2-norm. *J. Amer. Statist. Assoc.* **115** 1011–1027. MR4107696 https://doi.org/10.1080/01621459.2019.1604366

[82] ZHANG, X. (2015). Testing high dimensional mean under sparsity. Preprint. Available at arXiv:1509.08444.

[83] ZHANG, Y., WANG, R. and SHAO, X. (2023). Adaptive testing for high-dimensional data. Preprint. Available at arXiv:2303.08197.

[84] ZHENG, Q., PENG, L. and HE, X. (2015). Globally adaptive quantile regression with ultra-high dimensional data. *Ann. Statist.* **43** 2225–2258. MR3396984 https://doi.org/10.1214/15-AOS1340

[85] ZHOU, W.-X., BOSE, K., FAN, J. and LIU, H. (2018). A new perspective on robust M-estimation: Finite sample theory and applications to dependence-adjusted multiple testing. *Ann. Statist.* **46** 1904–1931. MR3845005 https://doi.org/10.1214/17-AOS1606

The Annals of Statistics
2023, Vol. 51, No. 5, 2116–2144
https://doi.org/10.1214/23-AOS2323

NONPARAMETRIC CONDITIONAL LOCAL INDEPENDENCE TESTING

BY ALEXANDER MANGULAD CHRISTGAU[a], LASSE PETERSEN[b] AND NIELS RICHARD HANSEN[c]

Department of Mathematical Sciences, University of Copenhagen, [a]*amc@math.ku.dk,* [b]*lassepetersen@protonmail.com,* [c]*Niels.R.Hansen@math.ku.dk*

Conditional local independence is an asymmetric independence relation among continuous time stochastic processes. It describes whether the evolution of one process is directly influenced by another process given the histories of additional processes, and it is important for the description and learning of causal relations among processes. We develop a model-free framework for testing the hypothesis that a counting process is conditionally locally independent of another process. To this end, we introduce a new functional parameter called the Local Covariance Measure (LCM), which quantifies deviations from the hypothesis. Following the principles of double machine learning, we propose an estimator of the LCM and a test of the hypothesis using nonparametric estimators and sample splitting or cross-fitting. We call this test the (cross-fitted) Local Covariance Test ((X)-LCT), and we show that its level and power can be controlled uniformly, provided that the nonparametric estimators are consistent with modest rates. We illustrate the theory by an example based on a marginalized Cox model with time-dependent covariates, and we show in simulations that when double machine learning is used in combination with cross-fitting, then the test works well without restrictive parametric assumptions.

REFERENCES

AALEN, O. O. (1987). Dynamic modelling and causality. *Scand. Actuar. J.* **3–4** 177–190. MR0943579 https://doi.org/10.1016/j.rser.2011.04.029

AALEN, O. O., RØYSLAND, K., GRAN, J. M. and LEDERGERBER, B. (2012). Causality, mediation and time: A dynamic viewpoint. *J. Roy. Statist. Soc. Ser. A* **175** 831–861. MR2993496 https://doi.org/10.1111/j.1467-985X.2011.01030.x

ACHAB, M., BACRY, E., GAÏFFAS, S., MASTROMATTEO, I. and MUZY, J. F. (2017). Uncovering causality from multivariate Hawkes integrated cumulants. In *Proceedings of the* 34*th International Conference on Machine Learning* **70** 1–10. PMLR.

ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. and KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. *Springer Series in Statistics*. Springer, New York. MR1198884 https://doi.org/10.1007/978-1-4612-4348-9

BACRY, E., BOMPAIRE, M., DEEGAN, P., GAÏFFAS, S. and POULSEN, S. V. (2017). `tick`: A Python library for statistical learning, with an emphasis on Hawkes processes and time-dependent models. *J. Mach. Learn. Res.* **18** Paper No. 214. MR3827102

BRÉMAUD, P. (1981). *Point Processes and Queues*: *Martingale Dynamics*. *Springer Series in Statistics*. Springer, New York-Berlin. MR0636252

CAI, R., WU, S., QIAO, J., HAO, Z., ZHANG, K. and ZHANG, X. (2022). THPs: Topological Hawkes processes for learning causal structure on event sequences. *IEEE Trans. Neural Netw. Learn. Syst.* 1–15.

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

CHRISTGAU, A. M., PETERSEN, L. and HANSEN, N. R. (2023). Supplement to "Nonparametric conditional local independence testing." https://doi.org/10.1214/23-AOS2323SUPPA, https://doi.org/10.1214/23-AOS2323SUPPB

MSC2020 subject classifications. Primary 62G10; secondary 62G20.
Key words and phrases. Nonparametric inference, local independence, double machine learning, functional CLT, stochastic processes.

COMMENGES, D. and GÉGOUT-PETIT, A. (2009). A general dynamical statistical model with causal interpretation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 719–736. MR2749916 https://doi.org/10.1111/j.1467-9868.2009.00703.x

DIDELEZ, V. (2007). Graphical models for composable finite Markov processes. *Scand. J. Stat.* **34** 169–185. MR2325249 https://doi.org/10.1111/j.1467-9469.2006.00528.x

DIDELEZ, V. (2008). Graphical models for marked point processes based on local independence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 245–264. MR2412641 https://doi.org/10.1111/j.1467-9868.2007.00634.x

DIDELEZ, V. (2015). Causal reasoning for events in continuous time: A decision-theoretic approach. In *Proceedings of the UAI* 2015 *Workshop on Advances in Causal Inference*.

GRANGER, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37** 424–438.

HAREZLAK, J., COULL, B. A., LAIRD, N. M., MAGARI, S. R. and CHRISTIANI, D. C. (2007). Penalized solutions to functional regression problems. *Comput. Statist. Data Anal.* **51** 4911–4925. MR2364549 https://doi.org/10.1016/j.csda.2006.09.034

LOK, J. J. (2008). Statistical modeling of causal effects in continuous time. *Ann. Statist.* **36** 1464–1507. MR2418664 https://doi.org/10.1214/009053607000000820

LUNDBORG, A. R., KIM, I., SHAH, R. D. and SAMWORTH, R. J. (2022). The projected covariance measure for assumption-lean variable significance testing. ArXiv preprint. Available at arXiv:2211.02039.

LUNDBORG, A. R., SHAH, R. D. and PETERS, J. (2022). Conditional independence testing in Hilbert spaces with applications to functional data analysis. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 1821–1850. MR4515559 https://doi.org/10.1111/rssb.12544

MALFAIT, N. and RAMSAY, J. O. (2003). The historical functional linear model. *Canad. J. Statist.* **31** 115–128. MR2016223 https://doi.org/10.2307/3316063

MOGENSEN, S. W. and HANSEN, N. R. (2020). Markov equivalence of marginalized local independence graphs. *Ann. Statist.* **48** 539–559. MR4065173 https://doi.org/10.1214/19-AOS1821

MOGENSEN, S. W. and HANSEN, N. R. (2022). Graphical modeling of stochastic processes driven by correlated noise. *Bernoulli* **28** 3023–3050. MR4474571 https://doi.org/10.3150/21-bej1446

MOGENSEN, S. W., MALINSKY, D. and HANSEN, N. R. (2018). Causal learning for partially observed stochastic dynamical systems. In *Proceedings of the* 34*th Conference on Uncertainty in Artificial Intelligence* 350–360.

NEWEY, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica* **62** 1349–1382. MR1303237 https://doi.org/10.2307/2951752

NEYKOV, M., BALAKRISHNAN, S. and WASSERMAN, L. (2021). Minimax optimal conditional independence testing. *Ann. Statist.* **49** 2151–2177. MR4319245 https://doi.org/10.1214/20-aos2030

PETERSEN, L. and HANSEN, N. R. (2021). Testing conditional independence via quantile regression based partial copulas. *J. Mach. Learn. Res.* **22** Paper No. 70. MR4253763

ROGERS, L. C. G. and WILLIAMS, D. (2000). *Diffusions, Markov Processes, and Martingales. Vol.* 2. *Cambridge Mathematical Library*. Cambridge Univ. Press, Cambridge. MR1780932 https://doi.org/10.1017/CBO9781107590120

RØYSLAND, K., RYALEN, P., NYGÅRD, M. and DIDELEZ, V. (2022). Graphical criteria for the identification of marginal causal effects in continuous-time survival and event-history analyses. ArXiv preprint. Available at arXiv:2202.02311.

SCHEIDEGGER, C., HÖRRMANN, J. and BÜHLMANN, P. (2022). The weighted generalised covariance measure. *J. Mach. Learn. Res.* **23** 1–68.

SCHILLING, R. L. and PARTZSCH, L. (2014). *Brownian Motion: An Introduction to Stochastic Processes*, 2nd ed. *De Gruyter Graduate*. de Gruyter, Berlin. MR3234570 https://doi.org/10.1515/9783110307306

SCHWEDER, T. (1970). Composable Markov processes. *J. Appl. Probab.* **7** 400–410. MR0264755 https://doi.org/10.2307/3211973

SHAH, R. D. and PETERS, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *Ann. Statist.* **48** 1514–1538. MR4124333 https://doi.org/10.1214/19-AOS1857

XU, H., FARAJTABAR, M. and ZHA, H. (2016). Learning granger causality for Hawkes processes. In *Proceedings of the* 33*rd International Conference on Machine Learning* **48** 1717–1726.

ZHOU, K., ZHA, H. and SONG, L. (2013). Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. In *Proceedings of the* 16*th International Conference on Artificial Intelligence and Statistics*.

The Annals of Statistics
2023, Vol. 51, No. 5, 2145–2169
https://doi.org/10.1214/23-AOS2324

ON BACKWARD SMOOTHING ALGORITHMS

BY HAI-DANG DAU[a] AND NICOLAS CHOPIN[b]

CREST-ENSAE, Institut Polytechnique de Paris, [a]*hai-dang.dau@stats.ox.ac.uk,* [b]*nicolas.chopin@ensae.fr*

In the context of state-space models, skeleton-based smoothing algorithms rely on a backward sampling step, which by default, has a $\mathcal{O}(N^2)$ complexity (where N is the number of particles). Existing improvements in the literature are unsatisfactory: a popular rejection sampling-based approach, as we shall show, might lead to badly behaved execution time; another rejection sampler with stopping lacks complexity analysis; yet another MCMC-inspired algorithm comes with no stability guarantee. We provide several results that close these gaps. In particular, we prove a novel nonasymptotic stability theorem, thus enabling smoothing with truly linear complexity and adequate theoretical justification. We propose a general framework, which unites most skeleton-based smoothing algorithms in the literature and allows to simultaneously prove their convergence and stability, both in online and offline contexts. Furthermore, we derive, as a special case of that framework, a new coupling-based smoothing algorithm applicable to models with intractable transition densities. We elaborate practical recommendations and confirm those with numerical experiments.

REFERENCES

ANDRIEU, C., DOUCET, A. and HOLENSTEIN, R. (2010). Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 269–342. MR2758115 https://doi.org/10.1111/j.1467-9868.2009.00736.x

BESKOS, A., PAPASPILIOPOULOS, O., ROBERTS, G. O. and FEARNHEAD, P. (2006). Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 333–382. MR2278331 https://doi.org/10.1111/j.1467-9868.2006.00552.x

BISHOP, C. M. (2006). *Pattern Recognition and Machine Learning. Information Science and Statistics*. Springer, New York. MR2247587 https://doi.org/10.1007/978-0-387-45528-0

BUNCH, P. and GODSILL, S. (2013). Improved particle approximations to the joint smoothing distribution using Markov chain Monte Carlo. *IEEE Trans. Signal Process.* **61** 956–963.

CHOPIN, N. and PAPASPILIOPOULOS, O. (2020). *An Introduction to Sequential Monte Carlo. Springer Series in Statistics*. Springer, Cham. MR4215639 https://doi.org/10.1007/978-3-030-47845-2

DAU, H.-D and CHOPIN, N. (2023). Supplement to "On backward smoothing algorithms." https://doi.org/10.1214/23-AOS2324SUPP

DEL MORAL, P. (2004). *Feynman–Kac Formulae: Genealogical and Interacting Particle Systems with Applications. Probability and Its Applications (New York)*. Springer, New York. MR2044973 https://doi.org/10.1007/978-1-4684-9393-1

DEL MORAL, P. (2013). *Mean Field Simulation for Monte Carlo Integration. Monographs on Statistics and Applied Probability* **126**. CRC Press, Boca Raton, FL. MR3060209

DEL MORAL, P., DOUCET, A. and SINGH, S. S. (2010). A backward particle interpretation of Feynman–Kac formulae. *M2AN Math. Model. Numer. Anal.* **44** 947–975. MR2731399 https://doi.org/10.1051/m2an/2010048

DEL MORAL, P. and MICLO, L. (2001). Genealogies and increasing propagation of chaos for Feynman–Kac and genetic models. *Ann. Appl. Probab.* **11** 1166–1198. MR1878294

DOUC, R., GARIVIER, A., MOULINES, E. and OLSSON, J. (2011). Sequential Monte Carlo smoothing for general state space hidden Markov models. *Ann. Appl. Probab.* **21** 2109–2145. MR2895411 https://doi.org/10.1214/10-AAP735

DUBARRY, C. and LE CORFF, S. (2013). Non-asymptotic deviation inequalities for smoothed additive functionals in nonlinear state-space models. *Bernoulli* **19** 2222–2249. MR3160552 https://doi.org/10.3150/12-BEJ450

DUFFIELD, S. and SINGH, S. S. (2022). Online particle smoothing with application to map-matching. *IEEE Trans. Signal Process.* **70** 497–508. MR4372360 https://doi.org/10.1109/TSP.2022.3141259

MSC2020 subject classifications. Primary 62M05, 65C05; secondary 65Y20.
Key words and phrases. State-space model, smoothing, sequential Monte Carlo.

FEARNHEAD, P., PAPASPILIOPOULOS, O. and ROBERTS, G. O. (2008). Particle filters for partially observed diffusions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 755–777. MR2523903 https://doi.org/10.1111/j.1467-9868.2008.00661.x

FEARNHEAD, P., WYNCOLL, D. and TAWN, J. (2010). A sequential smoothing algorithm with linear computational cost. *Biometrika* **97** 447–464. MR2650750 https://doi.org/10.1093/biomet/asq013

GERBER, M. and CHOPIN, N. (2015). Sequential quasi Monte Carlo. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 509–579. MR3351446 https://doi.org/10.1111/rssb.12104

GLOAGUEN, P., LE CORFF, S. and OLSSON, J. (2022). A pseudo-marginal sequential Monte Carlo online smoothing algorithm. *Bernoulli* **28** 2606–2633. MR4474556 https://doi.org/10.3150/21-bej1431

GODSILL, S. J., DOUCET, A. and WEST, M. (2004). Monte Carlo smoothing for nonlinear times series. *J. Amer. Statist. Assoc.* **99** 156–168. MR2054295 https://doi.org/10.1198/016214504000000151

GORDON, N. J., SALMOND, D. J. and SMITH, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc. F, Commun. Radar Signal Process.* **140** 107–113.

GUARNIERO, P., JOHANSEN, A. M. and LEE, A. (2017). The iterated auxiliary particle filter. *J. Amer. Statist. Assoc.* **112** 1636–1647. MR3750887 https://doi.org/10.1080/01621459.2016.1222291

HENING, A. and NGUYEN, D. H. (2018). Stochastic Lotka–Volterra food chains. *J. Math. Biol.* **77** 135–163. MR3800804 https://doi.org/10.1007/s00285-017-1192-8

JACOB, P. E., LINDSTEN, F. and SCHÖN, T. B. (2020). Smoothing with couplings of conditional particle filters. *J. Amer. Statist. Assoc.* **115** 721–729. MR4107675 https://doi.org/10.1080/01621459.2018.1548856

JASRA, A., KAMATANI, K., LAW, K. J. H. and ZHOU, Y. (2017). Multilevel particle filters. *SIAM J. Numer. Anal.* **55** 3068–3096. MR3735293 https://doi.org/10.1137/17M1111553

KITAGAWA, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *J. Comput. Graph. Statist.* **5** 1–25. MR1380850 https://doi.org/10.2307/1390750

LOTKA, A. J. (1926). *Elements of Physical Biology*. Williams & Wilkins.

MASTROTOTARO, A., OLSSON, J. and ALENLÖV, J. (2021). Fast and numerically stable particle-based online additive smoothing: The AdaSmooth algorithm. ArXiv preprint. Available at arXiv:2108.00432.

NORDH, J. and ANTONSSON, J. (2015). A quantitative evaluation of Monte Carlo smoothers. Technical report.

OLSSON, J. and WESTERBORN, J. (2017). Efficient particle-based online smoothing in general hidden Markov models: The PaRIS algorithm. *Bernoulli* **23** 1951–1996. MR3624883 https://doi.org/10.3150/16-BEJ801

PITT, M. K. and SHEPHARD, N. (1999). Filtering via simulation: Auxiliary particle filters. *J. Amer. Statist. Assoc.* **94** 590–599. MR1702328 https://doi.org/10.2307/2670179

SAMET, H. (2006). *Foundations of Multidimensional and Metric Data Structures*. Morgan Kaufmann, San Mateo.

SEN, D., THIERY, A. H. and JASRA, A. (2018). On coupling particle filter trajectories. *Stat. Comput.* **28** 461–475. MR3747574 https://doi.org/10.1007/s11222-017-9740-z

TAGHAVI, E., LINDSTEN, F., SVENSSON, L. and SCH ÖN, T. B. (2013). Adaptive stopping for fast particle smoothing. In 2013 *IEEE International Conference on Acoustics, Speech and Signal Processing* 6293–6297.

VOLTERRA, V. (1928). Variations and fluctuations of the number of individuals in animal species living together. *ICES J. Mar. Sci.* **3** 3–51.

YONEKURA, S. and BESKOS, A. (2022). Online smoothing for diffusion processes observed with noise. *J. Comput. Graph. Statist.* **31** 1344–1360. MR4513392 https://doi.org/10.1080/10618600.2022.2027243

The Annals of Statistics
2023, Vol. 51, No. 5, 2170–2193
https://doi.org/10.1214/23-AOS2326
This research was funded, in whole or in part, by European Research Council, 101019489 and EPSRC, EP/W016117/1. A CC BY 4.0 license is applied to this article arising from this submission, in accordance with the grant's open access conditions

OPTIMAL NONPARAMETRIC TESTING OF MISSING COMPLETELY AT RANDOM AND ITS CONNECTIONS TO COMPATIBILITY

BY THOMAS B. BERRETT[1,a] AND RICHARD J. SAMWORTH[2,b]

[1]*Department of Statistics, University of Warwick,* [a]*tom.berrett@warwick.ac.uk*
[2]*Statistical Laboratory, Centre for Mathematical Sciences,* [b]*r.samworth@statslab.cam.ac.uk*

Given a set of incomplete observations, we study the nonparametric problem of testing whether data are Missing Completely At Random (MCAR). Our first contribution is to characterise precisely the set of alternatives that can be distinguished from the MCAR null hypothesis. This reveals interesting and novel links to the theory of Fréchet classes (in particular, compatible distributions) and linear programming, that allow us to propose MCAR tests that are consistent against all detectable alternatives. We define an incompatibility index as a natural measure of ease of detectability, establish its key properties and show how it can be computed exactly in some cases and bounded in others. Moreover, we prove that our tests can attain the minimax separation rate according to this measure, up to logarithmic factors. Our methodology does not require any complete cases to be effective, and is available in the R package MCARtest.

REFERENCES

ABRAMSKY, S. (2017). Contextuality: At the borders of paradox. In *Categories for the Working Philosopher* 262–285. Oxford Univ. Press, Oxford. MR3822113

ABRAMSKY, S. and BRANDENBURGER, A. (2011). The sheaf-theoretic structure of non-locality and contextuality. *New J. Phys.* **13** 113036.

AHUJA, R. K., MAGNANTI, T. L. and ORLIN, J. B. (1989). Network flows. In *Optimization. Handbooks Oper. Res. Management Sci.* **1** 211–369. North-Holland, Amsterdam. MR1105103 https://doi.org/10.1016/S0927-0507(89)01005-4

ALEXANDROFF, P. (1924). Über die Metrisation der im Kleinen kompakten topologischen Räume. *Math. Ann.* **92** 294–301. MR1512216 https://doi.org/10.1007/BF01448011

BELL, J. S. (1966). On the problem of hidden variables in quantum mechanics. *Rev. Modern Phys.* **38** 447–452. MR0208927 https://doi.org/10.1103/RevModPhys.38.447

BERRETT, T. B. and SAMWORTH, R. J. (2022). MCARtest: Optimal nonparametric testing of Missing Completely At Random. R package version 1.1. Available at https://cran.r-project.org/web/packages/MCARtest/index.html.

BERRETT, T. B. and SAMWORTH, R. J. (2023). Supplement to "Optimal nonparametric testing of Missing Completely At Random and its connections to compatibility." https://doi.org/10.1214/23-AOS2326SUPP

BLANCHARD, G., CARPENTIER, A. and GUTZEIT, M. (2018). Minimax Euclidean separation rates for testing convex hypotheses in $\mathbb{R}^d$. *Electron. J. Stat.* **12** 3713–3735. MR3873534 https://doi.org/10.1214/18-ejs1472

CAI, T. T. and ZHANG, L. (2019). High dimensional linear discriminant analysis: Optimality, adaptive algorithm and missing data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 675–705. MR3997097

CHEN, H. Y. and LITTLE, R. (1999). A test of Missing Completely At Random for generalised estimating equations with missing data. *Biometrika* **86** 1–13. MR1688067 https://doi.org/10.1093/biomet/86.1.1

CLAUSER, J. F. and SHIMONY, A. (1978). Bell's theorem. Experimental tests and implications. *Rep. Progr. Phys.* **41** 1881.

COONS, J. I., CUMMINGS, J., HOLLERING, B. and MARAJ, A. (2020). Generalized cut polytopes for binary hierarchical models. *Algeb. Stat.* To appear.

DALL'AGLIO, G., KOTZ, S. and SALINETTI, G. (2012). *Advances in Probability Distributions with Given Marginals: Beyond the Copulas*. Springer, Berlin.

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

MSC2020 subject classifications. 62G10.
Key words and phrases. Missing data, Missing Completely At Random, minimax testing, linear programming.

DE LOERA, J. A. and KIM, E. D. (2014). Combinatorics and geometry of transportation polytopes: An update. In *Discrete Geometry and Algebraic Combinatorics*. *Contemp*. *Math*. **625** 37–76. Amer. Math. Soc., Providence, RI. MR3289405 https://doi.org/10.1090/conm/625/12491

DEZA, M. M. and LAURENT, M. (2010). *Geometry of Cuts and Metrics*. *Algorithms and Combinatorics* **15**. Springer, Heidelberg. MR2841334 https://doi.org/10.1007/978-3-642-04295-9

DUDLEY, R. M. (2002). *Real Analysis and Probability*. *Cambridge Studies in Advanced Mathematics* **74**. Cambridge Univ. Press, Cambridge. MR1932358 https://doi.org/10.1017/CBO9780511755347

ELSENER, A. and VAN DE GEER, S. (2019). Sparse spectral estimation with missing and corrupted measurements. *Stat* **8** e229. MR3978409 https://doi.org/10.1002/sta4.229

EMBRECHTS, P. and PUCCETTI, G. (2010). Bounds for the sum of dependent risks having overlapping marginals. *J*. *Multivariate Anal*. **101** 177–190. MR2557627 https://doi.org/10.1016/j.jmva.2009.07.004

ERIKSSON, N., FIENBERG, S. E., RINALDO, A. and SULLIVANT, S. (2006). Polyhedral conditions for the nonexistence of the MLE for hierarchical log-linear models. *J*. *Symbolic Comput*. **41** 222–233. MR2197157 https://doi.org/10.1016/j.jsc.2005.04.003

FARKAS, J. (1902). Theorie der einfachen Ungleichungen. *J*. *Reine Angew*. *Math*. **124** 1–27. MR1580578 https://doi.org/10.1515/crll.1902.124.1

FIENBERG, S. E. (1968). The geometry of an $r \times c$ contingency table. *Ann*. *Math*. *Stat*. **39** 1186–1190. MR0232525 https://doi.org/10.1214/aoms/1177698242

FOLLAIN, B., WANG, T. and SAMWORTH, R. J. (2022). High-dimensional changepoint estimation with heterogeneous missingness. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **84** 1023–1055. MR4460584

FUCHS, C. (1982). Maximum likelihood estimation and model selection in contingency tables with missing data. *J*. *Amer*. *Statist*. *Assoc*. **77** 270–278.

GALE, D. (1957). A theorem on flows in networks. *Pacific J*. *Math*. **7** 1073–1082. MR0091855

GEYER, C. J. and MEEDEN, G. D. (2021). rcdd: Computational Geometryats. R package version 1.5. Available at https://cran.r-project.org/web/packages/rcdd/index.html.

GUROBI OPTIMIZATION, LLC (2021). Gurobi Optimizer Reference Manual.

HOŞTEN, S. and SULLIVANT, S. (2002). Gröbner bases and polyhedral geometry of reducible and cyclic models. *J*. *Combin*. *Theory Ser*. *A* **100** 277–301. MR1940337 https://doi.org/10.1006/jcta.2002.3301

ISII, K. (1964). Inequalities of the types of Chebyshev and Cramér-Rao and mathematical programming. *Ann*. *Inst*. *Statist*. *Math*. **16** 277–293. MR0176836 https://doi.org/10.1007/BF02868576

JAMSHIDIAN, M. and JALAL, S. (2010). Tests of homoscedasticity, normality, and Missing Completely At Random for incomplete multivariate data. *Psychometrika* **75** 649–674. MR2741492 https://doi.org/10.1007/s11336-010-9175-3

JIAO, J., HAN, Y. and WEISSMAN, T. (2018). Minimax estimation of the L_1 distance. *IEEE Trans*. *Inf*. *Theory* **64** 6672–6706. MR3860754 https://doi.org/10.1109/TIT.2018.2846245

JOE, H. (1997). *Multivariate Models and Dependence Concepts*. *Monographs on Statistics and Applied Probability* **73**. CRC Press, London. MR1462613 https://doi.org/10.1201/b13150

KANTOROVICH, L. V. (2004). On mass transportation. *Zap*. *Nauchn*. *Sem*. *S*.*-Peterburg*. *Otdel*. *Mat*. *Inst*. *Steklov*. (*POMI*) **312** 11–14. MR2117876 https://doi.org/10.1007/s10958-006-0049-2

KANTOROVITCH, L. (1942). On the translocation of masses. *C*. *R*. (*Dokl*.) *Acad*. *Sci*. *URSS* **37** 199–201. MR0009619

KELLERER, H. G. (1984). Duality theorems for marginal problems. *Z*. *Wahrsch*. *Verw*. *Gebiete* **67** 399–432. MR0761565 https://doi.org/10.1007/BF00532047

KIM, K. H. and BENTLER, P. M. (2002). Tests of homogeneity of means and covariance matrices for multivariate incomplete data. *Psychometrika* **67** 609–623. MR2227891 https://doi.org/10.1007/BF02295134

LAURITZEN, S. L., SPEED, T. P. and VIJAYAN, K. (1984). Decomposable graphs and hypergraphs. *J*. *Aust*. *Math*. *Soc*. *A* **36** 12–29. MR0719998

LAURITZEN, S. L. and SPIEGELHALTER, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *J*. *Roy*. *Statist*. *Soc*. *Ser*. *B* **50** 157–224. MR0964177

LEIGHTON, T. and RAO, S. (1999). Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms. *J*. *ACM* **46** 787–832. MR1753034 https://doi.org/10.1145/331524.331526

LI, J. and YU, Y. (2015). A nonparametric test of Missing Completely At Random for incomplete multivariate data. *Psychometrika* **80** 707–726. MR3392026 https://doi.org/10.1007/s11336-014-9410-4

LITTLE, R. J. A. (1988). A test of Missing Completely At Random for multivariate data with missing values. *J*. *Amer*. *Statist*. *Assoc*. **83** 1198–1202. MR0997603

LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR1925014 https://doi.org/10.1002/9781119013563

LOH, P.-L. and TAN, X. L. (2018). High-dimensional robust precision matrix estimation: Cellwise corruption under ϵ-contamination. *Electron*. *J*. *Stat*. **12** 1429–1467. MR3804842 https://doi.org/10.1214/18-EJS1427

LOH, P.-L. and WAINWRIGHT, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann*. *Statist*. **40** 1637–1664. MR3015038 https://doi.org/10.1214/12-AOS1018

MAIER, D. (1983). *The Theory of Relational Databases*. *Computer Software Engineering Series*. Computer Science Press, Rockville, MD. MR0691493

MCMULLEN, P. (1970). The maximum numbers of faces of a convex polytope. *Mathematika* **17** 179–184. MR0283691 https://doi.org/10.1112/S0025579300002850

NELSEN, R. B. (2006). *An Introduction to Copulas*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2197664 https://doi.org/10.1007/s11229-005-3715-x

QU, A. and SONG, P. X.-K. (2002). Testing ignorable missingness in estimating equation approaches for longitudinal data. *Biometrika* **89** 841–850. MR1946514 https://doi.org/10.1093/biomet/89.4.841

REEVE, H. W., CANNINGS, T. I. and SAMWORTH, R. J. (2021). Optimal subgroup selection. *Ann*. *Statist*. To appear.

ROCKAFELLAR, R. T. (1997). *Convex Analysis*. *Princeton Landmarks in Mathematics*. Princeton Univ. Press, Princeton, NJ. MR1451876

RÜSCHENDORF, L. (2013). *Mathematical Risk Analysis*: *Dependence*, *risk bounds*, *optimal allocations and portfolios*. *Springer Series in Operations Research and Financial Engineering*. Springer, Heidelberg. MR3051756 https://doi.org/10.1007/978-3-642-33590-7

SPOHN, M.-L., NÄF, J., MICHEL, L. and MEINSHAUSEN, N. (2021). PKLM: A flexible MCAR test using Classification. ArXiv preprint. Available at arXiv:2109.10150.

VLACH, M. (1986). Conditions for the existence of solutions of the three-dimensional planar transportation problem. *Discrete Appl*. *Math*. **13** 61–78. MR0829339 https://doi.org/10.1016/0166-218X(86)90069-7

VOROBEV, N. N. (1962). Consistent families of measures and their extensions. *Theory Probab. Appl*. **7** 147–163.

WAINWRIGHT, M. J. and JORDAN, M. I. (2003). Variational inference in graphical models: The view from the marginal polytope. In *Proceedings of the Annual Allerton Conference on Communication Control and Computing* **41** 961–971.

WAINWRIGHT, M. J. and JORDAN, M. I. (2008). *Graphical Models*, *Exponential Families*, *and Variational Inference*. Now Publishers Inc., Hanover, MA.

WEI, Y., WAINWRIGHT, M. J. and GUNTUBOYINA, A. (2019). The geometry of hypothesis testing over convex cones: Generalized likelihood ratio tests and minimax radii. *Ann*. *Statist*. **47** 994–1024. MR3909958 https://doi.org/10.1214/18-AOS1701

WU, Y. and YANG, P. (2016). Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Trans*. *Inf*. *Theory* **62** 3702–3720. MR3506758 https://doi.org/10.1109/TIT.2016.2548468

ZHU, Z., WANG, T. and SAMWORTH, R. J. (2022). High-dimensional principal component analysis with heterogeneous missingness. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **84** 2000–2031. MR4515564

The Annals of Statistics
2023, Vol. 51, No. 5, 2194–2220
https://doi.org/10.1214/23-AOS2327

THE LASSO WITH GENERAL GAUSSIAN DESIGNS WITH APPLICATIONS TO HYPOTHESIS TESTING

BY MICHAEL CELENTANO[1,a], ANDREA MONTANARI[2,b], AND YUTING WEI[3,c]

[1]*Department of Statistics, University of California at Berkeley,* [a]*mcelentano@berkeley.edu*
[2]*Department of Statistics, Stanford University,* [b]*montanar@stanford.edu*
[3]*Department of Statistics and Data Science, University of Pennsylvania,* [c]*ytwei@wharton.upenn.edu*

The Lasso is a method for high-dimensional regression, which is now commonly used when the number of covariates p is of the same order or larger than the number of observations n. Classical asymptotic normality theory does not apply to this model due to two fundamental reasons: (1) The regularized risk is nonsmooth; (2) The distance between the estimator $\widehat{\boldsymbol{\theta}}$ and the true parameters vector $\boldsymbol{\theta}^*$ cannot be neglected. As a consequence, standard perturbative arguments that are the traditional basis for asymptotic normality fail.

On the other hand, the Lasso estimator can be precisely characterized in the regime in which both n and p are large and n/p is of order one. This characterization was first obtained in the case of Gaussian designs with i.i.d. covariates: here we generalize it to Gaussian correlated designs with non-singular covariance structure. This is expressed in terms of a simpler "fixed-design" model. We establish nonasymptotic bounds on the distance between the distribution of various quantities in the two models, which hold uniformly over signals $\boldsymbol{\theta}^*$ in a suitable sparsity class and over values of the regularization parameter.

As an application, we study the distribution of the debiased Lasso and show that a degrees-of-freedom correction is necessary for computing valid confidence intervals.

REFERENCES

[1] AMELUNXEN, D., LOTZ, M., MCCOY, M. B. and TROPP, J. A. (2014). Living on the edge: Phase transitions in convex programs with random data. *Inf. Inference* **3** 224–294. MR3311453 https://doi.org/10.1093/imaiai/iau005

[2] BAYATI, M., ERDOGDU, M. A. and MONTANARI, A. (2013). Estimating lasso risk and noise level. In *Advances in Neural Information Processing Systems* 944–952.

[3] BAYATI, M., LELARGE, M. and MONTANARI, A. (2015). Universality in polytope phase transitions and message passing algorithms. *Ann. Appl. Probab.* **25** 753–822. MR3313755 https://doi.org/10.1214/14-AAP1010

[4] BAYATI, M. and MONTANARI, A. (2012). The LASSO risk for Gaussian matrices. *IEEE Trans. Inf. Theory* **58** 1997–2017. MR2951312 https://doi.org/10.1109/TIT.2011.2174612

[5] BELLEC, P. C. (2023). Out-of-sample error estimation for M-estimators with convex penalty. *Inf. Inference* **12** 2782–2817. MR4660702 https://doi.org/10.1093/imaiai/iaad031

[6] BELLEC, P. C., LECUÉ, G. and TSYBAKOV, A. B. (2018). Slope meets Lasso: Improved oracle bounds and optimality. *Ann. Statist.* **46** 3603–3642. MR3852663 https://doi.org/10.1214/17-AOS1670

[7] BELLEC, P. C. and SHEN, Y. (2022). Derivatives and residual distribution of regularized m-estimators with application to adaptive tuning. In *Proceedings of Thirty Fifth Conference on Learning Theory* (P.-L. Loh and M. Raginsky, eds.) *Proceedings of Machine Learning Research* **178** 1912–1947. PMLR.

[8] BELLEC, P. C. and ZHANG, C.-H. (2018). Second order stein: Sure for sure and other applications in high-dimensional inference.

[9] BELLEC, P. C. and ZHANG, C.-H. (2022). De-biasing the lasso with degrees-of-freedom adjustment. *Bernoulli* **28** 713–743. MR4389062 https://doi.org/10.3150/21-BEJ1348

MSC2020 subject classifications. Primary 62J07, 62E17; secondary 62F05, 62F12.
Key words and phrases. Lasso, debiased Lasso, exact asymptotics, convex Gaussian min–max theorem, Gaussian designs, Gaussian width.

[10] BELLEC, P. C. and ZHANG, C.-H. (2023). Debiasing convex regularized estimators and interval estimation in linear models. *Ann*. *Statist*. **51** 391–436. MR4600987 https://doi.org/10.1214/22-aos2243
[11] BICKEL, P. J. and LEVINA, E. (2008). Covariance regularization by thresholding. *Ann*. *Statist*. **36** 2577–2604. MR2485008 https://doi.org/10.1214/08-AOS600
[12] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann*. *Statist*. **37** 1705–1732. MR2533469 https://doi.org/10.1214/08-AOS620
[13] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. *Springer Series in Statistics*. Springer, Heidelberg. MR2807761 https://doi.org/10.1007/978-3-642-20192-9
[14] CAI, T. T., ZHANG, C.-H. and ZHOU, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann*. *Statist*. **38** 2118–2144. MR2676885 https://doi.org/10.1214/09-AOS752
[15] CANDÈS, E., FAN, Y., JANSON, L. and LV, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **80** 551–577. MR3798878 https://doi.org/10.1111/rssb.12265
[16] CELENTANO, M. (2021). Approximate separability of symmetrically penalized least squares in high dimensions: Characterization and consequences. *Inf*. *Inference* **10** 1105–1165. MR4312091 https://doi.org/10.1093/imaiai/iaaa037
[17] CELENTANO, M. and MONTANARI, A. (2021). Cad: Debiasing the lasso with inaccurate covariate model.
[18] CELENTANO, M., MONTANARI, A. and WEI, Y. (2023). Supplement to "The Lasso with general Gaussian designs with applications to hypothesis testing." https://doi.org/10.1214/23-AOS2327SUPP
[19] CHANDRASEKARAN, V., RECHT, B., PARRILO, P. A. and WILLSKY, A. S. (2012). The convex geometry of linear inverse problems. *Found*. *Comput*. *Math*. **12** 805–849. MR2989474 https://doi.org/10.1007/s10208-012-9135-7
[20] CHEN, Y., FAN, J., MA, C. and YAN, Y. (2019). Inference and uncertainty quantification for noisy matrix completion. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **116** 22931–22937. MR4036123 https://doi.org/10.1073/pnas.1910053116
[21] CHETVERIKOV, D., LIAO, Z. and CHERNOZHUKOV, V. (2016). On cross-validated lasso. Available at arXiv:1605.02214.
[22] DONOHO, D. and MONTANARI, A. (2016). High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *Probab*. *Theory Related Fields* **166** 935–969. MR3568043 https://doi.org/10.1007/s00440-015-0675-z
[23] DONOHO, D. and TANNER, J. (2009). Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philos*. *Trans*. *R*. *Soc*. *Lond*. *Ser*. *A Math*. *Phys*. *Eng*. *Sci*. **367** 4273–4293. MR2546388 https://doi.org/10.1098/rsta.2009.0152
[24] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2011). The noise-sensitivity phase transition in compressed sensing. *IEEE Trans*. *Inf*. *Theory* **57** 6920–6941. MR2882271 https://doi.org/10.1109/TIT.2011.2165823
[25] DONOHO, D. L. and TANNER, J. (2005). Neighborliness of randomly projected simplices in high dimensions. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **102** 9452–9457. MR2168716 https://doi.org/10.1073/pnas.0502258102
[26] DONOHO, D. L. and TANNER, J. (2009). Counting faces of randomly projected polytopes when the projection radically lowers dimension. *J*. *Amer*. *Math*. *Soc*. **22** 1–53. MR2449053 https://doi.org/10.1090/S0894-0347-08-00600-0
[27] EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann*. *Statist*. **32** 407–499. MR2060166 https://doi.org/10.1214/009053604000000067
[28] EFRON, B. and TIBSHIRANI, R. (1997). Improvements on cross-validation: The 632+ bootstrap method. *J*. *Amer*. *Statist*. *Assoc*. **92** 548–560. MR1467848 https://doi.org/10.2307/2965703
[29] EL KAROUI, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann*. *Statist*. **36** 2717–2756. MR2485011 https://doi.org/10.1214/07-AOS559
[30] EL KAROUI, N. and PURDOM, E. (2018). Can we trust the bootstrap in high-dimensions? The case of linear models. *J*. *Mach*. *Learn*. *Res*. **19** Paper No. 5. MR3862412
[31] FISHER, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos*. *Trans*. *R*. *Soc*. *Lond*., *Ser*. *A*, *Contain*. *Pap*. *Math*. *Phys*. *Character* **222** 309–368.
[32] GEER, S. A. and VAN DE GEER, S. (2000). *Empirical Processes in M-Estimation* **6**. Cambridge University Press, Cambridge.
[33] HAN, Q. and SHEN, Y. (2023). Universality of regularized regression estimators in high dimensions. *Ann*. *Statist*. **51** 1799–1823. MR4658577 https://doi.org/10.1214/23-aos2309
[34] HASTIE, T. J. (2017). *Generalized Additive Models*. Routledge, London.
[35] HU, H. and LU, Y. M. (2023). Universality laws for high-dimensional learning with random features. *IEEE Trans*. *Inf*. *Theory* **69** 1932–1964. MR4564688

[36] JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909. MR3277152
[37] JAVANMARD, A. and MONTANARI, A. (2014). Hypothesis testing in high-dimensional regression under the Gaussian random design model: Asymptotic theory. *IEEE Trans. Inf. Theory* **60** 6522–6554. MR3265038 https://doi.org/10.1109/TIT.2014.2343629
[38] JAVANMARD, A. and MONTANARI, A. (2018). Debiasing the Lasso: Optimal sample size for Gaussian designs. *Ann. Statist.* **46** 2593–2622. MR3851749 https://doi.org/10.1214/17-AOS1630
[39] KATSEVICH, E. and RAMDAS, A. (2022). On the power of conditional independence testing under model-X. *Electron. J. Stat.* **16** 6348–6394. MR4517344 https://doi.org/10.1214/22-ejs2085
[40] LE CAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory. Springer Series in Statistics*. Springer, New York. MR0856411 https://doi.org/10.1007/978-1-4612-4946-7
[41] LI, G., FAN, W. and WEI, Y. (2023). Approximate message passing from random initialization with applications to $\mathbb{Z}_2$ synchronization. *Proc. Natl. Acad. Sci. USA* **120** Paper No. e2302930120. MR4637851
[42] LI, G. and WEI, Y. (2022). A non-asymptotic framework for approximate message passing in spiked models. ArXiv preprint. Available at arXiv:2208.03313.
[43] LI, Y. and WEI, Y. (2021). Minimum ℓ_1-norm interpolators: Precise asymptotics and multiple descent. ArXiv preprint. Available at arXiv:2110.09502.
[44] LIU, M., KATSEVICH, E., JANSON, L. and RAMDAS, A. (2022). Fast and powerful conditional randomization testing via distillation. *Biometrika* **109** 277–293. MR4430958 https://doi.org/10.1093/biomet/asab039
[45] MIOLANE, L. and MONTANARI, A. (2021). The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning. *Ann. Statist.* **49** 2313–2335. MR4319252 https://doi.org/10.1214/20-aos2038
[46] MONTANARI, A. and NGUYEN, P.-M. (2017). Universality of the elastic net error. In 2017 *IEEE International Symposium on Information Theory (ISIT)* 2338–2342. IEEE Press, New York.
[47] MONTANARI, A. and SAEED, B. N. (2022). Universality of empirical risk minimization. In *Conference on Learning Theory* 4310–4312. PMLR.
[48] NEGAHBAN, S. N., RAVIKUMAR, P., WAINWRIGHT, M. J. and YU, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.* **27** 538–557. MR3025133 https://doi.org/10.1214/12-STS400
[49] OYMAK, S. and TROPP, J. A. (2018). Universality laws for randomized dimension reduction, with applications. *Inf. Inference* **7** 337–446. MR3858331 https://doi.org/10.1093/imaiai/iax011
[50] REN, Z., SUN, T., ZHANG, C.-H. and ZHOU, H. H. (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *Ann. Statist.* **43** 991–1026. MR3346695 https://doi.org/10.1214/14-AOS1286
[51] SU, W., BOGDAN, M. and CANDÈS, E. (2017). False discoveries occur early on the Lasso path. *Ann. Statist.* **45** 2133–2150. MR3718164 https://doi.org/10.1214/16-AOS1521
[52] SUN, T. and ZHANG, C.-H. (2012). Comment: "Minimax estimation of large covariance matrices under ℓ_1-norm" [MR3027084]. *Statist. Sinica* **22** 1354–1358. MR3027086
[53] SUR, P. and CANDÈS, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proc. Natl. Acad. Sci. USA* **116** 14516–14525. MR3984492 https://doi.org/10.1073/pnas.1810420116
[54] THRAMPOULIDIS, C., ABBASI, E. and HASSIBI, B. (2018). Precise error analysis of regularized M-estimators in high dimensions. *IEEE Trans. Inf. Theory* **64** 5592–5628. MR3832326 https://doi.org/10.1109/TIT.2018.2840720
[55] THRAMPOULIDIS, C., OYMAK, S. and HASSIBI, B. (2015). Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory* 1683–1709.
[56] TROPP, J. A. (2015). Convex recovery of a structured signal from independent random linear measurements. In *Sampling Theory, a Renaissance. Appl. Numer. Harmon. Anal.* 67–101. Birkhäuser/Springer, Cham. MR3467419
[57] VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285 https://doi.org/10.1214/14-AOS1221
[58] WANG, H., YANG, Y., BU, Z. and SU, W. (2020). The complete lasso tradeoff diagram. In *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan and H. Lin, eds.) **33** 20051–20060. Curran Associates, Red Hook.
[59] ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940 https://doi.org/10.1111/rssb.12026
[60] ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2007). On the "degrees of freedom" of the lasso. *Ann. Statist.* **35** 2173–2192. MR2363967 https://doi.org/10.1214/009053607000000127

The Annals of Statistics
2023, Vol. 51, No. 5, 2221–2245
https://doi.org/10.1214/23-AOS2332

SPATIAL QUANTILES ON THE HYPERSPHERE

BY DIMITRI KONEN[a] AND DAVY PAINDAVEINE[b]

ECARES and Department of Mathematics, Université libre de Bruxelles, [a]*Dimitri.Konen@ulb.be,* [b]*Davy.Paindaveine@ulb.be*

We propose a concept of quantiles for probability measures on the unit hypersphere $\mathcal{S}^{d-1}$ of $\mathbb{R}^d$. The innermost quantile is the Fréchet median, that is, the L_1-analog of the Fréchet mean. The proposed quantiles $\mu^m_{\alpha,u}$ are directional in nature: they are indexed by a scalar order $\alpha \in [0, 1]$ and a unit vector u in the tangent space $T_m\mathcal{S}^{d-1}$ to $\mathcal{S}^{d-1}$ at m. To ensure computability in any dimension d, our quantiles are essentially obtained by considering the Euclidean (Chaudhuri (*J. Amer. Statist. Assoc.* **91** (1996) 862–872)) spatial quantiles in a suitable stereographic projection of $\mathcal{S}^{d-1}$ onto $T_m\mathcal{S}^{d-1}$. Despite this link with Euclidean spatial quantiles, studying the proposed spherical quantiles requires understanding the nature of the (Chaudhuri (1996)) quantiles in a version of the projective space where all points at infinity are identified. We thoroughly investigate the structural properties of our quantiles and we further study the asymptotic behavior of their sample versions, which requires controlling the impact of estimating m. Our spherical quantile concept also allows for companion concepts of ranks and depth on the hypersphere. We illustrate the relevance of our construction by considering two inferential applications, related to supervised classification and to testing for rotational symmetry.

REFERENCES

AGOSTINELLI, C. and ROMANAZZI, M. (2013). Nonparametric analysis of directional data based on data depth. *Environ. Ecol. Stat.* **20** 253–270. MR3068658 https://doi.org/10.1007/s10651-012-0218-z

BHATTACHARYA, R. and LIN, L. (2017). Omnibus CLTs for Fréchet means and nonparametric inference on non-Euclidean spaces. *Proc. Amer. Math. Soc.* **145** 413–428. MR3565392 https://doi.org/10.1090/proc/13216

BHATTACHARYA, R. and PATRANGENARU, V. (2003). Large sample theory of intrinsic and extrinsic sample means on manifolds. I. *Ann. Statist.* **31** 1–29. MR1962498 https://doi.org/10.1214/aos/1046294456

BROWN, B. M. (1983). Statistical uses of the spatial median. *J. Roy. Statist. Soc. Ser. B* **45** 25–30. MR0701072

CARDOT, H., CÉNAC, P. and GODICHON-BAGGIONI, A. (2017). Online estimation of the geometric median in Hilbert spaces: Nonasymptotic confidence balls. *Ann. Statist.* **45** 591–614. MR3650394 https://doi.org/10.1214/16-AOS1460

CARDOT, H., CÉNAC, P. and ZITT, P.-A. (2013). Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli* **19** 18–43. MR3019484 https://doi.org/10.3150/11-BEJ390

CHAKRABORTY, A. and CHAUDHURI, P. (2014). The spatial distribution in infinite dimensional spaces and related quantiles and depths. *Ann. Statist.* **42** 1203–1231. MR3224286 https://doi.org/10.1214/14-AOS1226

CHAUDHURI, P. (1996). On a geometric notion of quantiles for multivariate data. *J. Amer. Statist. Assoc.* **91** 862–872. MR1395753 https://doi.org/10.2307/2291681

CHENG, Y. and DE GOOIJER, J. G. (2007). On the uth geometric conditional quantile. *J. Statist. Plann. Inference* **137** 1914–1930. MR2323873 https://doi.org/10.1016/j.jspi.2006.02.014

CHERNOZHUKOV, V., GALICHON, A., HALLIN, M. and HENRY, M. (2017). Monge-Kantorovich depth, quantiles, ranks and signs. *Ann. Statist.* **45** 223–256. MR3611491 https://doi.org/10.1214/16-AOS1450

CHOWDHURY, J. and CHAUDHURI, P. (2019). Nonparametric depth and quantile regression for functional data. *Bernoulli* **25** 395–423. MR3892324 https://doi.org/10.3150/17-bej991

DAI, X. and LIN, Z. (2022). manifold: Operations for Riemannian Manifolds. R package version 0.1.1.

DAI, X. and LOPEZ-PINTADO, S. (2023). Tukey's depth for object data. *J. Amer. Statist. Assoc.* **118** 1760–1772. MR4646604 https://doi.org/10.1080/01621459.2021.2011298

MSC2020 subject classifications. Primary 62R30, 62G99; secondary 62E10, 62G20.

Key words and phrases. Centrality regions, directional statistics, Fréchet medians, multivariate quantiles, spatial quantiles, statistical depth.

DAI, X. and MÜLLER, H.-G. (2018). Principal component analysis for functional data on Riemannian manifolds and spheres. *Ann. Statist.* **46** 3334–3361. MR3852654 https://doi.org/10.1214/17-AOS1660

DYCKERHOFF, R. and NAGY, S. (2023). Exact computation of angular halfspace depth. *Manuscript in Preparation.*

ELTZNER, B. and HUCKEMANN, S. F. (2019). A smeary central limit theorem for manifolds with application to high-dimensional spheres. *Ann. Statist.* **47** 3360–3381. MR4025745 https://doi.org/10.1214/18-AOS1781

FRANCISCI, G., NIETO-REYES, A. and AGOSTINELLI, C. (2020). Generalization of the simplicial depth: No vanishment outside the convex hull of the distribution support. Available at arXiv:1909.02739v2.

FRÉCHET, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. Henri Poincaré* **10** 215–310. MR0027464

GARCÍA-PORTUGUÉS, E., PAINDAVEINE, D. and VERDEBOUT, T. (2020). On optimal tests for rotational symmetry against new classes of hyperspherical distributions. *J. Amer. Statist. Assoc.* **115** 1873–1887. MR4189764 https://doi.org/10.1080/01621459.2019.1665527

GHOSH, A. K. and CHAUDHURI, P. (2005). On maximum depth and related classifiers. *Scand. J. Stat.* **32** 327–350. MR2188677 https://doi.org/10.1111/j.1467-9469.2005.00423.x

GIRARD, S. and STUPFLER, G. (2015). Extreme geometric quantiles in a multivariate regular variation framework. *Extremes* **18** 629–663. MR3418771 https://doi.org/10.1007/s10687-015-0226-0

GIRARD, S. and STUPFLER, G. (2017). Intriguing properties of extreme geometric quantiles. *REVSTAT* **15** 107–139. MR3614901 https://doi.org/10.4310/jsg.2017.v15.n1.a4

HALLIN, M., DEL BARRIO, E., CUESTA-ALBERTOS, J. and MATRÁN, C. (2021). Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *Ann. Statist.* **49** 1139–1165. MR4255122 https://doi.org/10.1214/20-aos1996

HALLIN, M., PAINDAVEINE, D. and ŠIMAN, M. (2010). Multivariate quantiles and multiple-output regression quantiles: From L_1 optimization to halfspace depth. *Ann. Statist.* **38** 635–669. MR2604670 https://doi.org/10.1214/09-AOS723

JUPP, P. E. and KUME, A. (2020). Measures of goodness of fit obtained by almost-canonical transformations on Riemannian manifolds. *J. Multivariate Anal.* **176** 104579. MR4045270 https://doi.org/10.1016/j.jmva.2019.104579

KENDALL, W. S. and LE, H. (2011). Limit theorems for empirical Fréchet means of independent and non-identically distributed manifold-valued random variables. *Braz. J. Probab. Stat.* **25** 323–352. MR2832889 https://doi.org/10.1214/11-BJPS141

KOLTCHINSKII, V. I. (1997). M-estimation, convexity and quantiles. *Ann. Statist.* **25** 435–477. MR1439309 https://doi.org/10.1214/aos/1031833659

KONEN, D. and PAINDAVEINE, D. (2023). Supplement to "Spatial quantiles on the hypersphere." https://doi.org/10.1214/23-AOS2332SUPP

LEY, C., SABBAH, C. and VERDEBOUT, T. (2014). A new concept of quantiles for directional data and the angular Mahalanobis depth. *Electron. J. Stat.* **8** 795–816. MR3217789 https://doi.org/10.1214/14-EJS904

LEY, C. and VERDEBOUT, T. (2017a). *Modern Directional Statistics*. *Chapman & Hall/CRC Interdisciplinary Statistics Series*. CRC Press, Boca Raton, FL. MR3752655

LEY, C. and VERDEBOUT, T. (2017b). Skew-rotationally-symmetric distributions and related efficient inferential procedures. *J. Multivariate Anal.* **159** 67–81. MR3668548 https://doi.org/10.1016/j.jmva.2017.02.010

LI, J., CUESTA-ALBERTOS, J. A. and LIU, R. Y. (2012). DD-classifier: Nonparametric classification procedure based on DD-plot. *J. Amer. Statist. Assoc.* **107** 737–753. MR2980081 https://doi.org/10.1080/01621459.2012.688462

LIU, R. Y. (1990). On a notion of data depth based on random simplices. *Ann. Statist.* **18** 405–414. MR1041400 https://doi.org/10.1214/aos/1176347507

LIU, R. Y. and SINGH, K. (1992). Ordering directional data: Concepts of data depth on circles and spheres. *Ann. Statist.* **20** 1468–1484. MR1186260 https://doi.org/10.1214/aos/1176348779

MARDIA, K. V. and JUPP, P. E. (2000). *Directional Statistics*. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR1828667

MCKEAGUE, I. W., LÓPEZ-PINTADO, S., HALLIN, M. and ŠIMAN, M. (2011). Analyzing growth trajectories. *J. Dev. Orig. Health Dis.* **2** 322–329.

NAGY, S. (2021). Halfspace depth does not characterize probability distributions. *Statist. Papers* **62** 1135–1139. MR4262188 https://doi.org/10.1007/s00362-019-01130-x

PAINDAVEINE, D. and VIRTA, J. (2021). On the behavior of extreme d-dimensional spatial quantiles under minimal assumptions. In *Advances in Contemporary Statistics and Econometrics—Festschrift in Honor of Christine Thomas-Agnan* (A. Daouia and A. Ruiz-Gazen, eds.) 243–259. Springer, Cham. MR4299283 https://doi.org/10.1007/978-3-030-73249-3_13

PANDOLFO, G., PAINDAVEINE, D. and PORZIO, G. C. (2018). Distance-based depths for directional data. *Canad. J. Statist.* **46** 593–609. MR3902616 https://doi.org/10.1002/cjs.11479

SERFLING, R. J. (2002). Quantile functions for multivariate analysis: Approaches and applications. *Stat. Neerl.* **56** 214–232.

TUKEY, J. W. (1975). Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians* (*Vancouver, B.C.*, 1974), *Vol.* 2 523–531. Canad. Math. Congr., Montreal, QC. MR0426989

TYLER, D. E. (1987). Statistical analysis for the angular central Gaussian distribution on the sphere. *Biometrika* **74** 579–589. MR0909362 https://doi.org/10.1093/biomet/74.3.579

WANG, X., ZHU, J., PAN, W., ZHU, J. and ZHANG, H. (2021). Nonparametric statistical inference via metric distribution function in metric spaces. Available at arXiv:2107.07317.

WEI, Y. (2008). An approach to multivariate covariate-dependent quantile contours with application to bivariate conditional growth charts. *J. Amer. Statist. Assoc.* **103** 397–409. MR2420242 https://doi.org/10.1198/016214507000001472

YANG, L. (2011). Some properties of Fréchet medians in Riemannian manifolds. ArXiv preprint. Available at arXiv:1110.3899v2.

ZUO, Y. and SERFLING, R. (2000). General notions of statistical depth function. *Ann. Statist.* **28** 461–482. MR1790005 https://doi.org/10.1214/aos/1016218226

The Annals of Statistics
2023, Vol. 51, No. 5, 2246–2271
https://doi.org/10.1214/23-AOS2333

A CLT FOR THE LSS OF LARGE-DIMENSIONAL SAMPLE COVARIANCE MATRICES WITH DIVERGING SPIKES

BY ZHIJUN LIU[a], JIANG HU[b], ZHIDONG BAI[c] AND HAIYAN SONG[d]

School of Mathematics and Statistics, Northeast Normal University, [a]*liuzj037@nenu.edu.cn,* [b]*huj156@nenu.edu.cn,* [c]*baizd@nenu.edu.cn,* [d]*songhy716@nenu.edu.cn*

In this paper, we establish the central limit theorem (CLT) for linear spectral statistics (LSSs) of a large-dimensional sample covariance matrix when the population covariance matrices are involved with diverging spikes. This constitutes a nontrivial extension of the Bai–Silverstein theorem (BST) (*Ann. Probab.* **32** (2004) 553–605), a theorem that has strongly influenced the development of high-dimensional statistics, especially in the applications of random matrix theory to statistics. Recently, there has been a growing realization that the assumption of uniform boundedness of the population covariance matrices in the BST is not satisfied in some fields, such as economics, where the variances of principal components may diverge as the dimension tends to infinity. Therefore, in this paper, we aim to eliminate this obstacle to applications of the BST. Our new CLT accommodates spiked eigenvalues, which may either be bounded or tend to infinity. A distinguishing feature of our result is that the variance in the new CLT is related to both spiked eigenvalues and bulk eigenvalues, with dominance being determined by the divergence rate of the largest spiked eigenvalues. The new CLT for LSS is then applied to test the hypothesis that the population covariance matrix is the identity matrix or a generalized spiked model. The asymptotic distributions of the corrected likelihood ratio test statistic and the corrected Nagao's trace test statistic are derived under the alternative hypothesis. Moreover, we present power comparisons between these two LSSs and Roy's largest root test. In particular, we demonstrate that except for the case in which there is only one spike, the LSSs could exhibit higher asymptotic power than Roy's largest root test.

REFERENCES

ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR1990662

BAI, J. and NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221. MR1926259 https://doi.org/10.1111/1468-0262.00273

BAI, Z., HU, J., PAN, G. and ZHOU, W. (2015). Convergence of the empirical spectral distribution function of Beta matrices. *Bernoulli* **21** 1538–1574. MR3352053 https://doi.org/10.3150/14-BEJ613

BAI, Z., JIANG, D., YAO, J.-F. and ZHENG, S. (2009). Corrections to LRT on large-dimensional covariance matrix by RMT. *Ann. Statist.* **37** 3822–3840. MR2572444 https://doi.org/10.1214/09-AOS694

BAI, Z., LI, H. and PAN, G. (2019). Central limit theorem for linear spectral statistics of large dimensional separable sample covariance matrices. *Bernoulli* **25** 1838–1869. MR3961233 https://doi.org/10.3150/18-BEJ1038

BAI, Z. and YAO, J. (2008). Central limit theorems for eigenvalues in a spiked population model. *Ann. Inst. Henri Poincaré Probab. Stat.* **44** 447–474. MR2451053 https://doi.org/10.1214/07-AIHP118

BAI, Z. and YAO, J. (2012). On sample eigenvalues in a generalized spiked population model. *J. Multivariate Anal.* **106** 167–177. MR2887686 https://doi.org/10.1016/j.jmva.2011.10.009

BAI, Z. D. (1999). Methodologies in spectral analysis of large-dimensional random matrices, a review. *Statist. Sinica* **9** 611–677. MR1711663

BAI, Z. D., MIAO, B. Q. and PAN, G. M. (2007). On asymptotics of eigenvectors of large sample covariance matrix. *Ann. Probab.* **35** 1532–1572. MR2330979 https://doi.org/10.1214/009117906000001079

MSC2020 subject classifications. Primary 60B20; secondary 60F05.

Key words and phrases. Empirical spectral distribution, linear spectral statistic, random matrix, Stieltjes transform.

BAI, Z. D. and SILVERSTEIN, J. W. (2004). CLT for linear spectral statistics of large-dimensional sample covariance matrices. *Ann. Probab.* **32** 553–605. MR2040792 https://doi.org/10.1214/aop/1078415845

BAIK, J., LEE, J. O. and WU, H. (2018). Ferromagnetic to paramagnetic transition in spherical spin glass. *J. Stat. Phys.* **173** 1484–1522. MR3878351 https://doi.org/10.1007/s10955-018-2150-6

BAIK, J. and SILVERSTEIN, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *J. Multivariate Anal.* **97** 1382–1408. MR2279680 https://doi.org/10.1016/j.jmva.2005.08.003

BANNA, M., NAJIM, J. and YAO, J. (2020). A CLT for linear spectral statistics of large random information-plus-noise matrices. *Stochastic Process. Appl.* **130** 2250–2281. MR4074699 https://doi.org/10.1016/j.spa.2019.06.017

BAO, Z., HU, J., XU, X. and ZHANG, X. (2022). Spectral statistics of sample block correlation matrices. Preprint. Available at arXiv:2207.06107.

BLOEMENDAL, A., KNOWLES, A., YAU, H.-T. and YIN, J. (2016). On the principal components of sample covariance matrices. *Probab. Theory Related Fields* **164** 459–552. MR3449395 https://doi.org/10.1007/s00440-015-0616-x

CAI, T. T., HAN, X. and PAN, G. (2020). Limiting laws for divergent spiked eigenvalues and largest nonspiked eigenvalue of sample covariance matrices. *Ann. Statist.* **48** 1255–1280. MR4124322 https://doi.org/10.1214/18-AOS1798

CHEN, B. and PAN, G. (2015). CLT for linear spectral statistics of normalized sample covariance matrices with the dimension much larger than the sample size. *Bernoulli* **21** 1089–1133. MR3338658 https://doi.org/10.3150/14-BEJ599

DING, X. and YANG, F. (2018). A necessary and sufficient condition for edge universality at the largest singular values of covariance matrices. *Ann. Appl. Probab.* **28** 1679–1738. MR3809475 https://doi.org/10.1214/17-AAP1341

DOBRIBAN, E. (2020). Permutation methods for factor analysis and PCA. *Ann. Statist.* **48** 2824–2847. MR4152122 https://doi.org/10.1214/19-AOS1907

DONOHO, D., GAVISH, M. and JOHNSTONE, I. (2018). Optimal shrinkage of eigenvalues in the spiked covariance model. *Ann. Statist.* **46** 1742–1778. MR3819116 https://doi.org/10.1214/17-AOS1601

GAO, J., HAN, X., PAN, G. and YANG, Y. (2017). High dimensional correlation matrices: The central limit theorem and its applications. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 677–693. MR3641402 https://doi.org/10.1111/rssb.12189

HU, J., LI, W., LIU, Z. and ZHOU, W. (2019). High-dimensional covariance matrices in elliptical distributions with application to spherical test. *Ann. Statist.* **47** 527–555. MR3909941 https://doi.org/10.1214/18-AOS1699

JIANG, D. and BAI, Z. (2021). Generalized four moment theorem and an application to CLT for spiked eigenvalues of high-dimensional covariance matrices. *Bernoulli* **27** 274–294. MR4177370 https://doi.org/10.3150/20-BEJ1237

JIANG, T. and YANG, F. (2013). Central limit theorems for classical likelihood ratio tests for high-dimensional normal distributions. *Ann. Statist.* **41** 2029–2074. MR3127857 https://doi.org/10.1214/13-AOS1134

JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. MR1863961 https://doi.org/10.1214/aos/1009210544

JOHNSTONE, I. M. (2008). Multivariate analysis and Jacobi ensembles: Largest eigenvalue, Tracy–Widom limits and rates of convergence. *Ann. Statist.* **36** 2638–2716. MR2485010 https://doi.org/10.1214/08-AOS605

JOHNSTONE, I. M. and NADLER, B. (2017). Roy's largest root test under rank-one alternatives. *Biometrika* **104** 181–193. MR3626473 https://doi.org/10.1093/biomet/asw060

JOHNSTONE, I. M. and ONATSKI, A. (2020). Testing in high-dimensional spiked models. *Ann. Statist.* **48** 1231–1254. MR4124321 https://doi.org/10.1214/18-AOS1697

JOHNSTONE, I. M. and PAUL, D. (2018). PCA in high dimensions: An orientation. *Proc. IEEE* **106** 1277–1292.

JUNG, S. and MARRON, J. S. (2009). PCA consistency in high dimension, low sample size context. *Ann. Statist.* **37** 4104–4130. MR2572454 https://doi.org/10.1214/09-AOS709

KRITCHMAN, S. and NADLER, B. (2009). Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory. *IEEE Trans. Signal Process.* **57** 3930–3941. MR2683143 https://doi.org/10.1109/TSP.2009.2022897

LEDOIT, O. and WOLF, M. (2002). Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Ann. Statist.* **30** 1081–1102. MR1926169 https://doi.org/10.1214/aos/1031689018

LI, H. and BAI, Z. (2015). Extreme eigenvalues of large dimensional quaternion sample covariance matrices. *J. Statist. Plann. Inference* **159** 1–14. MR3299085 https://doi.org/10.1016/j.jspi.2014.10.005

LI, Z., HAN, F. and YAO, J. (2020). Asymptotic joint distribution of extreme eigenvalues and trace of large sample covariance matrix in a generalized spiked population model. *Ann. Statist.* **48** 3138–3160. MR4185803 https://doi.org/10.1214/19-AOS1882

LI, Z., WANG, Q. and LI, R. (2021). Central limit theorem for linear spectral statistics of large dimensional Kendall's rank correlation matrices and its applications. *Ann. Statist.* **49** 1569–1593. MR4298873 https://doi.org/10.1214/20-aos2013

LIU, Z., HU, J., BAI, Z. and SONG, H. (2023). Supplement to "A CLT for the LSS of large-dimensional sample covariance matrices with diverging spikes." https://doi.org/10.1214/23-AOS2333SUPP

NADLER, B. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach. *Ann. Statist.* **36** 2791–2817. MR2485013 https://doi.org/10.1214/08-AOS618

NAGAO, H. (1973). On some test criteria for covariance matrix. *Ann. Statist.* **1** 700–709. MR0339405

NAJIM, J. and YAO, J. (2016). Gaussian fluctuations for linear spectral statistics of large random covariance matrices. *Ann. Appl. Probab.* **26** 1837–1887. MR3513608 https://doi.org/10.1214/15-AAP1135

OLSON, C. L. (1974). Comparative robustness of six tests in multivariate analysis of variance. *J. Amer. Statist. Assoc.* **69** 894–908.

ONATSKI, A., MOREIRA, M. J. and HALLIN, M. (2013). Asymptotic power of sphericity tests for high-dimensional data. *Ann. Statist.* **41** 1204–1231. MR3113808 https://doi.org/10.1214/13-AOS1100

ONATSKI, A., MOREIRA, M. J. and HALLIN, M. (2014). Signal detection in high dimension: The multispiked case. *Ann. Statist.* **42** 225–254. MR3189485 https://doi.org/10.1214/13-AOS1181

PAN, G. (2014). Comparison between two types of large sample covariance matrices. *Ann. Inst. Henri Poincaré Probab. Stat.* **50** 655–677. MR3189088 https://doi.org/10.1214/12-AIHP506

PAN, G. M. and ZHOU, W. (2008). Central limit theorem for signal-to-interference ratio of reduced rank linear receiver. *Ann. Appl. Probab.* **18** 1232–1270. MR2418244 https://doi.org/10.1214/07-AAP477

PAUL, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17** 1617–1642. MR2399865

PERRY, A., WEIN, A. S., BANDEIRA, A. S. and MOITRA, A. (2018). Optimality and sub-optimality of PCA I: Spiked random matrix models. *Ann. Statist.* **46** 2416–2451. MR3845022 https://doi.org/10.1214/17-AOS1625

SILVERSTEIN, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. *J. Multivariate Anal.* **55** 331–339. MR1370408 https://doi.org/10.1006/jmva.1995.1083

WANG, Q. and YAO, J. (2013). On the sphericity test with large-dimensional observations. *Electron. J. Stat.* **7** 2164–2192. MR3104916 https://doi.org/10.1214/13-EJS842

WANG, Q. and YAO, J. (2017). Extreme eigenvalues of large-dimensional spiked Fisher matrices with application. *Ann. Statist.* **45** 415–460. MR3611497 https://doi.org/10.1214/16-AOS1463

WANG, W. and FAN, J. (2017). Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *Ann. Statist.* **45** 1342–1374. MR3662457 https://doi.org/10.1214/16-AOS1487

WILKS, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.* **9** 60–62.

YANG, J. and JOHNSTONE, I. M. (2018). Edgeworth correction for the largest eigenvalue in a spiked PCA model. *Statist. Sinica* **28** 2541–2564. MR3839873

YANG, Y. and PAN, G. (2015). Independence test for high dimensional data based on regularized canonical correlation coefficients. *Ann. Statist.* **43** 467–500. MR3316187 https://doi.org/10.1214/14-AOS1284

YAO, J., ZHENG, S. and BAI, Z. (2015). *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. *Cambridge Series in Statistical and Probabilistic Mathematics* **39**. Cambridge Univ. Press, New York. MR3468554 https://doi.org/10.1017/CBO9781107588080

YAO, Z., ZHANG, Y., BAI, Z. and EDDY, W. F. (2018). Estimating the number of sources in magnetoencephalography using spiked population eigenvalues. *J. Amer. Statist. Assoc.* **113** 505–518. MR3832204 https://doi.org/10.1080/01621459.2017.1341411

YIN, Y. (2022). Spectral statistics of high dimensional sample covariance matrix with unbounded population spectral norm. *Bernoulli* **28** 1729–1756. MR4411509 https://doi.org/10.3150/21-bej1391

ZHANG, Z., ZHENG, S., PAN, G. and ZHONG, P.-S. (2022). Asymptotic independence of spiked eigenvalues and linear spectral statistics for large sample covariance matrices. *Ann. Statist.* **50** 2205–2230. MR4474488 https://doi.org/10.1214/22-aos2183

ZHENG, S. (2012). Central limit theorems for linear spectral statistics of large dimensional F-matrices. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 444–476. MR2954263 https://doi.org/10.1214/11-AIHP414

ZHENG, S., BAI, Z. and YAO, J. (2015). Substitution principle for CLT of linear spectral statistics of high-dimensional sample covariance matrices with applications to hypothesis testing. *Ann. Statist.* **43** 546–591. MR3316190 https://doi.org/10.1214/14-AOS1292

ZHENG, S., CHENG, G., GUO, J. and ZHU, H. (2019). Test for high-dimensional correlation matrices. *Ann. Statist.* **47** 2887–2921. MR3988776 https://doi.org/10.1214/18-AOS1768

ZHOU, Y.-H. and MARRON, J. S. (2015). High dimension low sample size asymptotics of robust PCA. *Electron. J. Stat.* **9** 204–218. MR3312407 https://doi.org/10.1214/15-EJS992

The Annals of Statistics
2023, Vol. 51, No. 5, 2272–2297
https://doi.org/10.1214/23-AOS2337

ESTIMATION OF EXPECTED EULER CHARACTERISTIC CURVES OF NONSTATIONARY SMOOTH RANDOM FIELDS

BY FABIAN J. E. TELSCHOW[1,a], DAN CHENG[2,b], PRATYUSH PRANAV[3,c] AND ARMIN SCHWARTZMAN[4,d]

[1]*Department of Mathematics, Humboldt-Universität zu Berlin,* [a]*fabian.telschow@hu-berlin.de*
[2]*School of Mathematical and Statistical Sciences, Arizona State University,* [b]*chengdan@asu.edu*
[3]*Centre de Recherche Astrophysique de Lyon,* [c]*pratyuze@gmail.com*
[4]*Halıcıoğlu Data Science Institute, University of California, San Diego,* [d]*a7schwartzman@ucsd.edu*

The expected Euler characteristic (EEC) of excursion sets of a smooth Gaussian-related random field over a compact manifold approximates the distribution of its supremum for high thresholds. Viewed as a function of the excursion threshold, the EEC of a Gaussian-related field is expressed by the Gaussian kinematic formula (GKF) as a finite sum of known functions multiplied by the Lipschitz–Killing curvatures (LKCs) of the generating Gaussian field. This paper proposes consistent estimators of the LKCs as linear projections of "pinned" Euler characteristic (EC) curves obtained from realizations of zero-mean, unit variance Gaussian processes. As observed, data seldom is Gaussian and the exact mean and variance is unknown, yet the statistic of interest often satisfies a CLT with a Gaussian limit process; we adapt our LKC estimators to this scenario using a Gaussian multiplier bootstrap approach. This yields consistent estimates of the LKCs of the possibly nonstationary Gaussian limiting field that have low variance and are computationally efficient for complex underlying manifolds. For the EEC of the limiting field, a parametric plug-in estimator is presented, which is more efficient than the nonparametric average of EC curves. The proposed methods are evaluated using simulations of 2D fields, and illustrated on cosmological observations and simulations on the 2-sphere and 3D fMRI volumes.

REFERENCES

[1] ADE, P. A., AGHANIM, N., ARMITAGE-CAPLAN, C., ARNAUD, M., ASHDOWN, M., ATRIO-BARANDELA, F., AUMONT, J., BACCIGALUPI, C., BANDAY, A. J. et al. (2014). Planck 2013 results. XXIII. Isotropy and statistics of the CMB. *Astron. Astrophys.* **571** A23.

[2] ADE, P. A., AGHANIM, N., ARNAUD, M., ASHDOWN, M., AUMONT, J., BACCIGALUPI, C., BANDAY, A., BARREIRO, R., BARTLETT, J. et al. (2016). Planck 2015 results-XII. Full focal plane simulations. *Astron. Astrophys.* **594** A12.

[3] ADLER, R. J. (1977). A spectral moment estimation problem in two dimensions. *Biometrika* **64** 367–373. MR0466102 https://doi.org/10.1093/biomet/64.2.367

[4] ADLER, R. J., BARTZ, K., KOU, S. C. and MONOD, A. (2017). Estimating thresholding levels for random fields via Euler characteristics. ArXiv preprint. Available at arXiv:1704.08562.

[5] ADLER, R. J. and TAYLOR, J. E. (2007). *Random Fields and Geometry*. *Springer Monographs in Mathematics*. Springer, New York. MR2319516

[6] BARTLETT, J., BUCHER, M., CARDOSO, J., CASTEX, G., DELABROUILLE, J., GANGA, K., GIRAUD-HÉRAUD, Y., LE JEUNE, M., PATANCHON, G. et al. (2016). Planck 2015 results: IX. Diffuse component separation: CMB maps. *Astron. Astrophys.* **594** A9–A9.

[7] BIERMÉ, H., DI BERNARDINO, E., DUVAL, C. and ESTRADE, A. (2019). Lipschitz–Killing curvatures of excursion sets for two-dimensional random fields. *Electron. J. Stat.* **13** 536–581. MR3911693 https://doi.org/10.1214/19-EJS1530

MSC2020 subject classifications. Primary 60G15; secondary 60G60, 62M40.
Key words and phrases. Random field theory, Lipschitz–Killing curvatures, Gaussian related fields, Euler characteristic.

[8] Bullmore, E. T., Suckling, J., Overmeyer, S., Rabe-Hesketh, S., Taylor, E. and Brammer, M. J. (1999). Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. *IEEE Trans. Med. Imag.* **18** 32–42. https://doi.org/10.1109/42.750253

[9] Cabaña, E. M. (1985). Estimation of the spectral moment by means of the extrema. *Trabajos de Estadística e Investigación Operativa* **36** 71–80.

[10] Cheng, D., Cammarota, V., Fantaye, Y., Marinucci, D. and Schwartzman, A. (2020). Multiple testing of local maxima for detection of peaks on the (celestial) sphere. *Bernoulli* **26** 31–60. MR4036027 https://doi.org/10.3150/18-BEJ1068

[11] Collaboration, P., Ade, P. A. R., Aghanim, N., Armitage-Caplan, C., Arnaud, M., Ashdown, M., Atrio-Barandela, F., Aumont, J., Baccigalupi, C. et al. (2014). Planck 2013 results. XXIII. Isotropy and statistics of the CMB. *Astron. Astrophys.* **571** A23.

[12] Davenport, S. and Telschow, F. J. (2022). On the finiteness of second moments of the number of critical points of Gaussian random fields. ArXiv preprint. Available at arXiv:2201.01591.

[13] Degras, D. A. (2011). Simultaneous confidence bands for nonparametric regression with functional data. *Statist. Sinica* **21** 1735–1765. MR2895997 https://doi.org/10.5705/ss.2009.207

[14] Di Bernardino, E., Estrade, A. and León, J. R. (2017). A test of Gaussianity based on the Euler characteristic of excursion sets. *Electron. J. Stat.* **11** 843–890. MR3629017 https://doi.org/10.1214/17-EJS1248

[15] Eklund, A., Nichols, T. E. and Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proc. Natl. Acad. Sci. USA* **113** 7900–7905.

[16] Eriksen, H. K., Hansen, F. K., Banday, A. J., Górski, K. M. and Lilje, P. B. (2004). Asymmetries in the Cosmic Microwave Background Anisotropy Field. *Astrophys. J.* **605** 14–20.

[17] Gass, L. and Stecconi, M. (2023). The number of critical points of a Gaussian field: Finiteness of moments. ArXiv preprint. Available at arXiv:2305.17586.

[18] Goresky, M. and MacPherson, R. (1988). *Stratified Morse Theory*. *Ergebnisse der Mathematik und Ihrer Grenzgebiete* (3) [*Results in Mathematics and Related Areas* (3)] **14**. Springer, Berlin. MR0932724 https://doi.org/10.1007/978-3-642-71714-7

[19] Gorski, K. M., Hivon, E., Banday, A. J., Wandelt, B. D., Hansen, F. K., Reinecke, M. and Bartelmann, M. (2005). HEALPix: A framework for high-resolution discretization and fast analysis of data distributed on the sphere. *Astrophys. J.* **622** 759.

[20] Heiss, T. and Wagner, H. (2017). Streaming algorithm for Euler characteristic curves of multidimensional images. In *Computer Analysis of Images and Patterns. Part I. Lecture Notes in Computer Science* **10424** 397–409. Springer, Cham. MR3695725 https://doi.org/10.1007/978-3-319-64689-3

[21] Hikage, C., Suto, Y., Kayo, I., Taruya, A., Matsubara, T., Vogeley, M. S., Hoyle, F., Gott III, J. R., Brinkmann, J. et al. (2002). Three-dimensional genus statistics of galaxies in the SDSS early data release. *Publ. Astron. Soc. Jpn.* **54** 707–717.

[22] Kiebel, S. J., Poline, J. B., Friston, K. J., Holmes, A. P. and Worsley, K. J. (1999). Robust smoothness estimation in statistical parametric maps using standardized residuals from the general linear model. *NeuroImage* **10** 756–766. https://doi.org/10.1006/nimg.1999.0508

[23] Land, K. and Magueijo, J. (2005). Is the universe odd? *Phys. Rev. D* **72** 101302. MR2188178 https://doi.org/10.1103/PhysRevD.72.101302

[24] Liebl, D. and Reimherr, M. (2023). Fast and fair simultaneous confidence bands for functional parameters. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **85** 842–868. https://doi.org/10.1093/jrsssb/qkad026

[25] Milnor, J. (1963). *Morse Theory*. *Annals of Mathematics Studies*, *No*. 51. Princeton Univ. Press, Princeton, NJ. MR0163331

[26] Moran, J. M., Jolly, E. and Mitchell, J. P. (2012). Social-cognitive deficits in normal aging. *J. Neurosci.* **32** 5553–5561.

[27] Nichols, T. E. (2012). Multiple testing corrections, nonparametric methods, and random field theory. *NeuroImage* **62** 811–815. https://doi.org/10.1016/j.neuroimage.2012.04.014

[28] Piterbarg, V. (1996). Rice's method for large excursions of Gaussian random fields. Technical Report 478, Center for Stochastic Processes, Univ. North Carolina.

[29] Pranav, P. (2022). Anomalies in the topology of the temperature fluctuations in the cosmic microwave background: An analysis of the NPIPE and FFP10 data releases. *Astron. Astrophys.* **659** A115.

[30] Pranav, P., Adler, R. J., Buchert, T., Edelsbrunner, H., Jones, B. J., Schwartzman, A., Wagner, H. and van de Weygaert, R. (2019). Unexpected topology of the temperature fluctuations in the cosmic microwave background. *Astron. Astrophys.* **627** A163.

[31] Schwarz, D. J., Copi, C. J., Huterer, D. and Starkman, G. D. (2016). CMB anomalies after Planck. *Classical Quantum Gravity* **33** 184001.

[32] SOMMERFELD, M., SAIN, S. and SCHWARTZMAN, A. (2018). Confidence regions for spatial excursion sets from repeated random field observations, with an application to climate. *J. Amer. Statist. Assoc.* **113** 1327–1340. MR3862360 https://doi.org/10.1080/01621459.2017.1341838

[33] TAYLOR, J., TAKEMURA, A. and ADLER, R. J. (2005). Validity of the expected Euler characteristic heuristic. *Ann. Probab.* **33** 1362–1396. MR2150192 https://doi.org/10.1214/009117905000000099

[34] TAYLOR, J. E. (2006). A Gaussian kinematic formula. *Ann. Probab.* **34** 122–158. MR2206344 https://doi.org/10.1214/009117905000000594

[35] TAYLOR, J. E. and WORSLEY, K. J. (2007). Detecting sparse signals in random fields, with an application to brain mapping. *J. Amer. Statist. Assoc.* **102** 913–928. MR2354405 https://doi.org/10.1198/016214507000000815

[36] TELSCHOW, F. J, CHENG, D., PRANAV, P. and SCHWARTZMAN, A. (2023). Supplement to "Estimation of Expected Euler Characteristic Curves of Nonstationary Smooth Random Fields." https://doi.org/10.1214/23-AOS2337SUPP

[37] TELSCHOW, F. J. E., DAVENPORT, S. and SCHWARTZMAN, A. (2022). Functional delta residuals and applications to simultaneous confidence bands of moment based statistics. *J. Multivariate Anal.* **192** Paper No. 105085. MR4463046 https://doi.org/10.1016/j.jmva.2022.105085

[38] TELSCHOW, F. J. E., PIERRYNOWSKI, M. R. and HUCKEMANN, S. F. (2023). Confidence tubes for curves on SO(3) and identification of subject-specific gait change after kneeling. *J. R. Stat. Soc. Ser. C. Appl. Stat.* qlad060.

[39] TELSCHOW, F. J. E. and SCHWARTZMAN, A. (2022). Simultaneous confidence bands for functional data using the Gaussian kinematic formula. *J. Statist. Plann. Inference* **216** 70–94. MR4273839 https://doi.org/10.1016/j.jspi.2021.05.008

[40] WORSLEY, K. J., EVANS, A. C., MARRETT, S. and NEELIN, P. (1992). A three-dimensional statistical analysis for CBF activation studies in human brain. *J. Cereb. Blood Flow Metab.* **12** 900–918.

[41] WORSLEY, K. J., MARRETT, S., NEELIN, P., VANDAL, A. C., FRISTON, K. J. and EVANS, A. C. (1996). A unified statistical approach for determining significant signals in images of cerebral activation. *Hum. Brain Mapp.* **4** 58–73.

[42] WORSLEY, K. J., TAYLOR, J. E., TOMAIUOLO, F. and LERCH, J. (2004). Unified univariate and multivariate random field theory. *NeuroImage* **23** S189–S195.