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

THE ANNALS *of* STATISTICS

An Official Journal of the
Institute of Mathematical Statistics

Articles

Vol. 53, No. 6—December 2025

THE ANNALS OF STATISTICS Vol. 53, No. 6, pp. 2303–2756 December 2025

The Annals of Statistics **[ISSN 0090-5364 (print); ISSN 2168-8966 (online)],** Volume 53, Number 6, December 2025. 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
2025, Vol. 53, No. 6, 2303–2328
https://doi.org/10.1214/24-AOS2448

A COMMON-CAUSE PRINCIPLE FOR ELIMINATING SELECTION BIAS IN CAUSAL ESTIMANDS THROUGH COVARIATE ADJUSTMENT

BY MAYA B. MATHUR[1,a], TYLER J. VANDERWEELE[2,b] AND ILYA SHPITSER[3,c]

[1]*Quantitative Sciences Unit, Department of Biomedical Data Sciences, Stanford University,* [a]*mmathur@stanford.edu*
[2]*Department of Epidemiology, Harvard University,* [b]*tvanderw@hsph.harvard.edu*
[3]*Department of Computer Science, Johns Hopkins University,* [c]*ilyas@cs.jhu.edu*

Average treatment effects (ATEs) may be subject to selection bias when they are estimated among only a nonrepresentative subset of the target population. Selection bias can sometimes be eliminated by conditioning on a "sufficient adjustment set" of covariates, even for some forms of missingness not at random (MNAR). Without requiring full specification of the causal structure, we consider sufficient adjustment sets to allow nonparametric identification of conditional ATEs in the target population. Covariates in the sufficient set may be collected among only the selected sample. We establish that if a sufficient set exists, then the set consisting of common causes of the outcome and selection, excluding the exposure and its descendants, also suffices. We establish simple graphical criteria for when a sufficient set will not exist, which could help indicate whether this is plausible for a given study. Simulations considering selection due to missing data indicated that sufficiently-adjusted complete-case analysis (CCA) can considerably outperform multiple imputation under MNAR and, if the sample size is not large, sometimes even under missingness at random. Analogous to the common-cause principle for confounding, these sufficiency results clarify when and how selection bias can be eliminated through covariate adjustment.

REFERENCES

[1] ALLISON, P. D. (2000). Multiple imputation for missing data: A cautionary tale. *Sociol. Methods Res.* **28** 301–309.

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

[3] BANACK, H. R. and KAUFMAN, J. S. (2014). The obesity paradox: Understanding the effect of obesity on mortality among individuals with cardiovascular disease. *Prev. Med.* **62** 96–102.

[4] BAREINBOIM, E. and PEARL, J. (2012). Controlling selection bias in causal inference. In *Artificial Intelligence and Statistics* 100–108. PMLR.

[5] BAREINBOIM, E. and TIAN, J. (2015). Recovering causal effects from selection bias. In *Proceedings of the AAAI Conference on Artificial Intelligence* **29**.

[6] BAREINBOIM, E., TIAN, J. and PEARL, J. (2022). Recovering from selection bias in causal and statistical inference. In *Probabilistic and Causal Inference*: *The Works of Judea Pearl* 433–450.

[7] BARTLETT, J. W., CARPENTER, J. R., TILLING, K. and VANSTEELANDT, S. (2014). Improving upon the efficiency of complete case analysis when covariates are MNAR. *Biostatistics* **15** 719–730.

[8] BARTLETT, J. W., HAREL, O. and CARPENTER, J. R. (2015). Asymptotically unbiased estimation of exposure odds ratios in complete records logistic regression. *Amer. J. Epidemiol.* **182** 730–736.

[9] BUUREN, S. V. and GROOTHUIS-OUDSHOORN, K. (2010). mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* 1–68.

[10] CARPENTER, J. R. and SMUK, M. (2021). Missing data: A statistical framework for practice. *Biom. J.* **63** 915–947. MR4268947 https://doi.org/10.1002/bimj.202000196

[11] CORREA, J. and BAREINBOIM, E. (2017). Causal effect identification by adjustment under confounding and selection biases. In *Proceedings of the AAAI Conference on Artificial Intelligence* **31**.

MSC2020 subject classifications. 62D20.
Key words and phrases. Generalizability, missing data, complete-case analysis, imputation, collider bias.

[12] CORREA, J., TIAN, J. and BAREINBOIM, E. (2018). Generalized adjustment under confounding and selection biases. In *Proceedings of the AAAI Conference on Artificial Intelligence* **32**.

[13] CORREA, J. D., TIAN, J. and BAREINBOIM, E. (2019). Identification of causal effects in the presence of selection bias. In *Proceedings of the AAAI Conference on Artificial Intelligence* **33** 2744–2751.

[14] DANIEL, R. M., KENWARD, M. G., COUSENS, S. N. and DE STAVOLA, B. L. (2012). Using causal diagrams to guide analysis in missing data problems. *Stat. Methods Med. Res.* **21** 243–256. MR2906301 https://doi.org/10.1177/0962280210394469

[15] DIDELEZ, V., KREINER, S. and KEIDING, N. (2010). Graphical models for inference under outcome-dependent sampling. *Statist. Sci.* **25** 368–387. MR2791673 https://doi.org/10.1214/10-STS340

[16] GRAHAM, J. W., OLCHOWSKI, A. E. and GILREATH, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prev. Sci.* **8** 206–213.

[17] GREENLAND, S., PEARL, J. and ROBINS, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology* **10** 37–48.

[18] HENCKEL, L., PERKOVIĆ, E. and MAATHUIS, M. H. (2022). Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 579–599. MR4412998

[19] HERNÁN, M. A., HERNÁNDEZ-DÍAZ, S. and ROBINS, J. M. (2004). A structural approach to selection bias. *Epidemiology* **15** 615–625.

[20] HERNÁN, M. A. and ROBINS, J. M. (2020). *Causal Inference: What If.* CRC Press, CRC, Boca Raton, FL.

[21] HOWARD, B. V., ROMAN, M. J., DEVEREUX, R. B., FLEG, J. L., GALLOWAY, J. M., HENDERSON, J. A., HOWARD, W. J., LEE, E. T., METE, M. et al. (2008). Effect of lower targets for blood pressure and LDL cholesterol on atherosclerosis in diabetes: The SANDS randomized trial. *J. Amer. Med. Assoc.* **299** 1678–1689.

[22] HUGHES, R. A., HERON, J., STERNE, J. A. and TILLING, K. (2019). Accounting for missing data in statistical analyses: Multiple imputation is not always the answer. *Int. J. Epidemiol.* **48** 1294–1304.

[23] INTERNATIONAL COUNCIL FOR HARMONISATION OF TECHNICAL REQUIREMENTS FOR PHARMACEUTICALS IN HUMAN USE (2019). Addendum on estimands and sensitivity analysis in clinical trials: to the guideline on statistical principles for clinical trials. https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf. Accessed 2023-09-28.

[24] KENAH, E. (2023). A potential outcomes approach to selection bias. *Epidemiology* **34** 865–972.

[25] LAIRD, N. M. (1988). Missing data in longitudinal studies. *Stat. Med.* **7** 305–315.

[26] LASH, T. L. (2021). Chapter 29, Bias analysis. In *Modern Epidemiology* 711–748 4th ed. Wolters Kluwer, Philadelphia, PA.

[27] LEE, K. J. and CARLIN, J. B. (2012). Recovery of information from multiple imputation: A simulation study. *Emerg. Themes Epidemiol.* **9** 1–10.

[28] LI, P., STUART, E. A. and ALLISON, D. B. (2015). Multiple imputation: A flexible tool for handling missing data. *J. Amer. Med. Assoc.* **314** 1966–1967.

[29] LITTLE, R. J., D'AGOSTINO, R., COHEN, M. L., DICKERSIN, K., EMERSON, S. S., FARRAR, J. T., FRANGAKIS, C., HOGAN, J. W., MOLENBERGHS, G. et al. (2012). The prevention and treatment of missing data in clinical trials. *N. Engl. J. Med.* **367** 1355–1360.

[30] 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

[31] LU, H., COLE, S. R., HOWE, C. J. and WESTREICH, D. (2022). Toward a clearer definition of selection bias when estimating causal effects. *Epidemiology* **33** 699–706.

[32] MALINSKY, D., SHPITSER, I. and RICHARDSON, T. (2019). A potential outcomes calculus for identifying conditional path-specific effects. In *The 22nd International Conference on Artificial Intelligence and Statistics* 3080–3088. PMLR.

[33] MARSHALL, A., ALTMAN, D. G., ROYSTON, P. and HOLDER, R. L. (2010). Comparison of techniques for handling missing covariate data within prognostic modelling studies: A simulation study. *BMC Med. Res. Methodol.* **10** 1–16.

[34] MATHUR, M. B. (2023). The M-value: A simple sensitivity analysis for bias due to missing data in treatment effect estimates. *Amer. J. Epidemiol* **194** 612-620.

[35] MATHUR, M. B., PEACOCK, J. R., ROBINSON, T. N. and GARDNER, C. D. (2021). Effectiveness of a theory-informed documentary to reduce consumption of meat and animal products: Three randomized controlled experiments. *Nutrients* **13** 4555.

[36] MATHUR, M.B. and SHPITSER, I. (2025). Imputation without nightMARs: Graphical criteria for valid imputation of missing data. Preprint. Available at https://osf.io/preprints/osf/zqne9_v2.

[37] MATHUR, M. B. and SHPITSER, I. (2025). Simple graphical criteria for selection bias in general-population and selected-sample treatment effects. *Amer. J. Epidemiol.* **194** 267–277.

[38] MATHUR, M. B., SHPITSER, I., and VANDERWEELE, T. J., (2025). Resurrecting complete-case analysis: A defense. *Amer. J. Epidemiol.* To appear. Available at https://osf.io/preprints/osf/f9jvz_v2/.

[39] MATHUR, M. B., VANDERWEELE, T. J. and SHPITSER, I. (2025). Supplement to "A common-cause principle for eliminating selection bias in causal estimands through covariate adjustment." https://doi.org/10.1214/24-AOS2448SUPP

[40] MEALLI, F. and RUBIN, D. B. (2015). Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika* **102** 995–1000. MR3431570 https://doi.org/10.1093/biomet/asv035

[41] MOHAN, K. and PEARL, J. (2021). Graphical models for processing missing data. *J. Amer. Statist. Assoc.* **116** 1023–1037. MR4270041 https://doi.org/10.1080/01621459.2021.1874961

[42] MOLENBERGHS, G. and KENWARD, M. (2007). *Missing Data in Clinical Studies*. Wiley, New York.

[43] MOONS, K. G., DONDERS, R. A., STIJNEN, T. and HARRELL JR, F. E. (2006). Using the outcome for imputation of missing predictor values was preferred. *J. Clin. Epidemiol.* **59** 1092–1101.

[44] NABI, R., BHATTACHARYA, R., SHPITSER, I. and ROBINS, J. (2022). Causal and counterfactual views of missing data models. Preprint. Available at https://arxiv.org/abs/2210.05558.

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

[46] PIGOTT, T. D. (2001). A review of methods for missing data. *Educ. Res. Eval.* **7** 353–383.

[47] RICHARDSON, T. S. and ROBINS, J. M. (2013). Single world intervention graphs (SWIGs): a unification of the counterfactual and graphical approaches to causality. Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper 128.

[48] ROBINS, J. M., ROTNITZKY, A. and SCHARFSTEIN, D. O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials (Minneapolis, MN*, 1997). *IMA Vol. Math. Appl.* **116** 1–94. Springer, New York. MR1731681 https://doi.org/10.1007/978-1-4612-1284-3_1

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

[50] ROTNITZKY, A. and SMUCLER, E. (2020). Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *J. Mach. Learn. Res.* **21** Paper No. 188. MR4209474

[51] RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196 https://doi.org/10.1093/biomet/63.3.581

[52] RUBIN, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys. Wiley Classics Library*. Wiley, Hoboken, NJ. MR2117498

[53] SHPITSER, I. and PEARL, J. (2006). Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the National Conference on Artificial Intelligence* **21** 1219. MIT Press, Menlo Park, CA.

[54] SMITH, L. H. (2020). Selection mechanisms and their consequences: Understanding and addressing selection bias. *Curr. Epidemiol. Rep.* **7** 179–189.

[55] SMITH, L. H. and VANDERWEELE, T. J. (2019). Bounding bias due to selection. *Epidemiology* **30** 509.

[56] VANDERWEELE, T. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford Univ. Press, London.

[57] VANDERWEELE, T. J. (2012). Confounding and effect modification: Distribution and measure. *Epidemiol. Methods* **1** 55–82.

[58] VANDERWEELE, T. J., ROTHMAN, K. J. and LASH, T. L. (2021). Chapter 12, confounding and confounders. In *Modern Epidemiology* 263–286 4th ed. Wolters Kluwer, Philadelphia, PA.

[59] VANDERWEELE, T. J. and SHPITSER, I. (2011). A new criterion for confounder selection. *Biometrics* **67** 1406–1413. MR2872391 https://doi.org/10.1111/j.1541-0420.2011.01619.x

[60] VAN BUUREN, S. (2018). *Flexible Imputation of Missing Data*. CRC press, Boca Raton.

[61] WHITE, I. R. and CARLIN, J. B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat. Med.* **29** 2920–2931. MR2758387 https://doi.org/10.1002/sim.3944

[62] WHITTEMORE, A. S. (1978). Collapsibility of multidimensional contingency tables. *J. Roy. Statist. Soc. Ser. B, Methodol.* **40** 328–340. MR0522216

[63] WOOLDRIDGE, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *J. Econometrics* **141** 1281–1301. MR2413502 https://doi.org/10.1016/j.jeconom.2007.02.002

The Annals of Statistics
2025, Vol. 53, No. 6, 2329–2355
https://doi.org/10.1214/24-AOS2450

NEAR-OPTIMAL INFERENCE IN ADAPTIVE LINEAR REGRESSION

BY KOULIK KHAMARU[1,a], YASH DESHPANDE[2,b], TOR LATTIMORE[3,c], LESTER MACKEY[4,d] AND MARTIN J. WAINWRIGHT[5,e]

[1]*Department of Statistics, Rutgers University,* [a]*kk1241@stat.rutgers.edu*
[2]*Voleon Group,* [b]*ydeshpande@voleon.com*
[3]*DeepMind,* [c]*tor.lattimore@gmail.com*
[4]*Microsoft Research,* [d]*lmackey@microsoft.com*
[5]*Lab for Information and Decision Systems, Statistics and Data Science Center, Massachusetts Institute of Technology,* [e]*wainwrigwork@gmail.com*

When data is collected in an adaptive manner, even simple methods like ordinary least squares can exhibit nonnormal asymptotic behavior. As an undesirable consequence, hypothesis tests and confidence intervals based on asymptotic normality can lead to erroneous results. We propose a family of online debiasing estimators to correct these distributional anomalies in least squares estimation. Our proposed methods take advantage of the covariance structure present in the dataset and provide sharper estimates in directions for which more information has accrued. We establish an asymptotic normality property for our proposed online debiasing estimators under mild conditions on the data collection process and provide asymptotically exact confidence intervals. We additionally prove a minimax lower bound for the adaptive linear regression problem, thereby providing a baseline by which to compare estimators. There are various conditions under which our proposed estimators achieve the minimax lower bound. We demonstrate the usefulness of our theory via applications to multi-armed bandit, autoregressive time series estimation, and active learning with exploration.

REFERENCES

[1] ABBASI-YADKORI, Y., PÁL, D. and SZEPESVÁRI, C. (2011). Online least squares estimation with self-normalized processes: An application to bandit problems. Preprint. Available at arXiv:1102.2670.

[2] ÅSTRÖM, K. J. (2012). *Introduction to Stochastic Control Theory*. *Mathematics in Science and Engineering* **70**. Academic Press, New York. MR0270799

[3] BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities*: *A Nonasymptotic Theory of Independence*. Oxford Univ. Press, Oxford. MR3185193 https://doi.org/10.1093/acprof:oso/9780199535255.001.0001

[4] BOX, G. E. P., JENKINS, G. M., REINSEL, G. C. and LJUNG, G. M. (2016). *Time Series Analysis*: *Forecasting and Control*, 5th ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR3379415

[5] DESHPANDE, Y., JAVANMARD, A. and MEHRABI, M. (2019). Online debiasing for adaptively collected high-dimensional data. Preprint. Available at arXiv:1911.01040.

[6] DESHPANDE, Y., MACKEY, L., SYRGKANIS, V. and TADDY, M. (2018). Accurate inference for adaptive linear models. In *International Conference on Machine Learning* **35** 1194–1203.

[7] DICKEY, D. A. and FULLER, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *J. Amer. Statist. Assoc.* **74** 427–431. MR0548036

[8] DVORETZKY, A. (1972). Asymptotic normality for sums of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (*Univ. California*, *Berkeley*, *Calif.*, 1970/1971), *Vol. II*: *Probability Theory* 513–535. Univ. California Press, Berkeley, CA. MR0415728

[9] FONTAINE, X., PERRAULT, P., VALKO, M. and PERCHET, V. (2019). Online A-optimal design and active linear regression. Preprint. Available at arXiv:1906.08509.

MSC2020 subject classifications. Primary 62E20; secondary 6208.
Key words and phrases. Adaptive linear regression, online debiasing, minimax lower bound, asymptotic normality, multi-armed bandits, autoregressive time series.

[10] GOODWIN, G. C. and PAYNE, R. L. (1977). *Dynamic System Identification*: *Experiment Design and Data Analysis*. *Mathematics in Science and Engineering* **136**. Academic Press [Harcourt Brace Jovanovich, Publishers], New York–London. MR0499203

[11] HADAD, V., HIRSHBERG, D. A., ZHAN, R., WAGER, S. and ATHEY, S. (2021). Confidence intervals for policy evaluation in adaptive experiments. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **118** Paper No. e2014602118, 10 pp. MR4294058 https://doi.org/10.1073/pnas.2014602118

[12] HOWARD, S. R., RAMDAS, A., MCAULIFFE, J. and SEKHON, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *Ann*. *Statist*. **49** 1055–1080. MR4255119 https://doi.org/10.1214/20-aos1991

[13] JAMIESON, K., MALLOY, M., NOWAK, R. and BUBECK, S. (2014). lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory* **27** 423–439.

[14] KAUFMANN, E. and KOOLEN, W. M. (2021). Mixture martingales revisited with applications to sequential tests and confidence intervals. *J*. *Mach*. *Learn*. *Res*. **22** Paper No. 246, 44 pp. MR4353025

[15] KHAMARU, K., DESHPANDE, Y., LATTIMORE, T., MACKEY, L. and WAINWRIGHT, M. J. (2025). Supplement to "Near-optimal inference in adaptive linear regression." https://doi.org/10.1214/24-AOS2450SUPP

[16] LAI, T. L. (1994). Asymptotic properties of nonlinear least squares estimates in stochastic regression models. *Ann*. *Statist*. **22** 1917–1930. MR1329175 https://doi.org/10.1214/aos/1176325764

[17] LAI, T. L. and ROBBINS, H. (1979). Adaptive design and stochastic approximation. *Ann*. *Statist*. **7** 1196–1221. MR0550144

[18] LAI, T. L., ROBBINS, H. and WEI, C. Z. (1979). Strong consistency of least squares estimates in multiple regression. II. *J*. *Multivariate Anal*. **9** 343–361. MR0548786 https://doi.org/10.1016/0047-259X(79)90093-9

[19] LAI, T. L. and WEI, C. Z. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Ann*. *Statist*. **10** 154–166. MR0642726

[20] LATTIMORE, T. (2023). A lower bound for linear and kernel regression with adaptive covariates. ArXiv.

[21] LATTIMORE, T. and SZEPESVÁRI, C. (2020). *Bandit Algorithms*. Cambridge Univ. Press, Cambridge.

[22] NIE, X., TIAN, X., TAYLOR, J. and ZOU, J. (2018). Why adaptively collected data have negative bias and how to correct for it. In *International Conference on Artificial Intelligence and Statistics* **84** 1261–1269.

[23] SHIN, J., RAMDAS, A. and RINALDO, A. (2019). Are sample means in multi-armed bandits positively or negatively biased? In *Advances in Neural Information Processing Systems* **32** 7102–7111.

[24] SHIN, J., RAMDAS, A. and RINALDO, A. (2021). On the bias, risk, and consistency of sample means in multi-armed bandits. *SIAM J*. *Math*. *Data Sci*. **3** 1278–1300. MR4343253 https://doi.org/10.1137/20M1361249

[25] THOMPSON, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **25** 285–294.

[26] VILLAR, S. S., BOWDEN, J. and WASON, J. (2015). Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges. *Statist*. *Sci*. **30** 199–215. MR3353103 https://doi.org/10.1214/14-STS504

[27] 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

[28] WHITE, J. S. (1958). The limiting distribution of the serial correlation coefficient in the explosive case. *Ann*. *Math*. *Statist*. **29** 1188–1197. MR0100952 https://doi.org/10.1214/aoms/1177706450

[29] XU, M., QIN, T. and LIU, T.-Y. (2013). Estimation bias in multi-armed bandit algorithms for search advertising. In *Advances in Neural Information Processing Systems* **26** 2400–2408.

[30] ZHANG, K. W., JANSON, L. and MURPHY, S. A. Statistical inference with M-estimators on adaptively collected data.

[31] ZHANG, K. W., JANSON, L. and MURPHY, S. A. (2020). Inference for batched bandits. Preprint. Available at arXiv:2002.03217.

The Annals of Statistics
2025, Vol. 53, No. 6, 2356–2379
https://doi.org/10.1214/24-AOS2488

ASYMPTOTICALLY-EXACT SELECTIVE INFERENCE FOR QUANTILE REGRESSION

BY YUMENG WANG[1,a], SNIGDHA PANIGRAHI[1,b] AND XUMING HE[2,c]

[1]*Department of Statistics, University of Michigan,* [a]*yumengwa@umich.edu,* [b]*psnigdha@umich.edu*
[2]*Department of Statistics and Data Science, Washington University in St. Louis,* [c]*hex@wustl.edu*

In modern data analysis, it is common to select a model before performing statistical inference. Selective inference tools make adjustments for the model selection process in order to ensure reliable inference post selection. In this paper, we introduce an asymptotic pivot to infer about the effects of selected variables on conditional quantile functions. Utilizing estimators from smoothed quantile regression, our proposed pivot is easy to compute and yields asymptotically-exact selective inference without making strict distributional assumptions about the response variable.

At the core of our pivot is the use of external randomization variables, which allows us to utilize all available samples for both selection and inference, without partitioning the data into independent subsets or discarding samples at any step. From simulation studies, we find that: (i) the asymptotic confidence intervals based on our pivot achieve the desired coverage rates, even in cases where sample splitting fails due to insufficient sample size for inference; (ii) our intervals are consistently shorter than those produced by sample splitting across various models and signal settings. We report similar findings when we apply our approach to study risk factors for low birth weights in a publicly accessible dataset of US birth records from 2022.

REFERENCES

[1] ANGRIST, J., CHERNOZHUKOV, V. and FERNÁNDEZ-VAL, I. (2006). Quantile regression under misspecification, with an application to the U.S. wage structure. *Econometrica* **74** 539–563. MR2207400 https://doi.org/10.1111/j.1468-0262.2006.00671.x

[2] BACHOC, F., PREINERSTORFER, D. and STEINBERGER, L. (2020). Uniformly valid confidence intervals post-model-selection. *Ann. Statist.* **48** 440–463. MR4065169 https://doi.org/10.1214/19-AOS1815

[3] BEHL, P., CLAESKENS, G. and DETTE, H. (2014). Focussed model selection in quantile regression. *Statist. Sinica* **24** 601–624. MR3235391

[4] BELLONI, A., CHERNOZHUKOV, V. and KATO, K. (2019). Valid post-selection inference in high-dimensional approximately sparse quantile regression models. *J. Amer. Statist. Assoc.* **114** 749–758. MR3963177 https://doi.org/10.1080/01621459.2018.1442339

[5] BERK, R., BROWN, L., BUJA, A., ZHANG, K. and ZHAO, L. (2013). Valid post-selection inference. *Ann. Statist.* **41** 802–837. MR3099122 https://doi.org/10.1214/12-AOS1077

[6] BLENCOWE, H., KRASEVEC, J., DE ONIS, M., BLACK, R. E., AN, X., STEVENS, G. A., BORGHI, E., HAYASHI, C., ESTEVEZ, D. et al. (2019). National, regional, and worldwide estimates of low birthweight in 2015, with trends from 2000: A systematic analysis. *Lancet Glob. Health* **7** e849–e860.

[7] CADE, B. S. and NOON, B. R. (2003). A gentle introduction to quantile regression for ecologists. *Front. Ecol. Environ.* **1** 412–420.

[8] CANAY, I. A. (2011). A simple approach to quantile regression for panel data. *Econom. J.* **14** 368–386. MR2853232 https://doi.org/10.1111/j.1368-423X.2011.00349.x

[9] DAI, C., LIN, B., XING, X. and LIU, J. S. (2023). A scale-free approach for false discovery rate control in generalized linear models. *J. Amer. Statist. Assoc.* **118** 1551–1565. MR4646583 https://doi.org/10.1080/01621459.2023.2165930

MSC2020 subject classifications. 62G08, 62G20, 62E20.
Key words and phrases. Nonparametric statistics, post-selection inference, quantile regression, randomization, selective inference, smoothed quantile regression.

[10] DHARAMSHI, A., NEUFELD, A., MOTWANI, K., GAO, L. L., WITTEN, D. and BIEN, J. (2025). Generalized data thinning using sufficient statistics. *J. Amer. Statist. Assoc.* **120** 511–523. MR4893575 https://doi.org/10.1080/01621459.2024.2353948

[11] FERNANDES, M., GUERRE, E. and HORTA, E. (2021). Smoothing quantile regressions. *J. Bus. Econom. Statist.* **39** 338–357. MR4187194 https://doi.org/10.1080/07350015.2019.1660177

[12] GALVAO, A. F. and KATO, K. (2016). Smoothed quantile regression for panel data. *J. Econometrics* **193** 92–112. MR3500178 https://doi.org/10.1016/j.jeconom.2016.01.008

[13] GOLDENBERG, R. L. and CULHANE, J. F. (2007). Low birth weight in the United States. *Amer. J. Clin. Nutr.* **85** 584S–590S.

[14] GU, Y., FAN, J., KONG, L., MA, S. and ZOU, H. (2018). ADMM for high-dimensional sparse penalized quantile regression. *Technometrics* **60** 319–331. MR3847169 https://doi.org/10.1080/00401706.2017.1345703

[15] HE, X., PAN, X., TAN, K. M. and ZHOU, W.-X. (2022). Scalable estimation and inference for censored quantile regression process. *Ann. Statist.* **50** 2899–2924. MR4500629 https://doi.org/10.1214/22-aos2214

[16] HE, X., PAN, X., TAN, K. M. and ZHOU, W.-X. (2023). Smoothed quantile regression with large-scale inference. *J. Econometrics* **232** 367–388. MR4539491 https://doi.org/10.1016/j.jeconom.2021.07.010

[17] HOROWITZ, J. L. (1998). Bootstrap methods for median regression models. *Econometrica* **66** 1327–1351. MR1654307 https://doi.org/10.2307/2999619

[18] HUANG, Y., PIRENNE, S., PANIGRAHI, S. and CLAESKENS, G. (2025). Selective inference using randomized group lasso estimators for general models. *Electron. J. Stat.* **19**. MR4947571 https://doi.org/10.1214/25-ejs2416

[19] KIVARANOVIC, D. and LEEB, H. (2024). A (tight) upper bound for the length of confidence intervals with conditional coverage. *Electron. J. Stat.* **18** 1677–1701. MR4729828 https://doi.org/10.1214/24-ejs2232

[20] KOENKER, R. and BASSETT, G. JR. (1978). Regression quantiles. *Econometrica* **46** 33–50. MR0474644 https://doi.org/10.2307/1913643

[21] LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927. MR3485948 https://doi.org/10.1214/15-AOS1371

[22] LEEB, H. and PÖTSCHER, B. M. (2003). The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. *Econometric Theory* **19** 100–142. MR1965844 https://doi.org/10.1017/S0266466603191050

[23] LEEB, H. and PÖTSCHER, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory* **21** 21–59. MR2153856 https://doi.org/10.1017/S0266466605050036

[24] LEINER, J., DUAN, B., WASSERMAN, L. and RAMDAS, A. (2025). Data fission: Splitting a single data point. *J. Amer. Statist. Assoc.* **120** 135–146. MR4893533 https://doi.org/10.1080/01621459.2023.2270748

[25] MAN, R., PAN, X., TAN, K. M. and ZHOU, W.-X. (2024). A unified algorithm for penalized convolution smoothed quantile regression. *J. Comput. Graph. Statist.* **33** 625–637. MR4754818 https://doi.org/10.1080/10618600.2023.2275999

[26] PANIGRAHI, S. (2023). Carving model-free inference. *Ann. Statist.* **51** 2318–2341. MR4682699 https://doi.org/10.1214/23-aos2318

[27] PANIGRAHI, S., FRY, K. and TAYLOR, J. (2024). Exact selective inference with randomization. *Biometrika* **111** 1109–1127. MR4830049 https://doi.org/10.1093/biomet/asae019

[28] PANIGRAHI, S., MACDONALD, P. W. and KESSLER, D. (2023). Approximate post-selective inference for regression with the group LASSO. *J. Mach. Learn. Res.* **24** Paper No. [79], 49. MR4582501

[29] PANIGRAHI, S., MARKOVIC, J. and TAYLOR, J. (2017). An MCMC-free approach to post-selective inference. Preprint. Available at arXiv:1703.06154.

[30] PANIGRAHI, S., MOHAMMED, S., RAO, A. and BALADANDAYUTHAPANI, V. (2023). Integrative Bayesian models using post-selective inference: A case study in radiogenomics. *Biometrics* **79** 1801–1813. MR4643957 https://doi.org/10.1111/biom.13740

[31] PANIGRAHI, S. and TAYLOR, J. (2023). Approximate selective inference via maximum likelihood. *J. Amer. Statist. Assoc.* **118** 2810–2820. MR4681622 https://doi.org/10.1080/01621459.2022.2081575

[32] PANIGRAHI, S., TAYLOR, J. and WEINSTEIN, A. (2021). Integrative methods for post-selection inference under convex constraints. *Ann. Statist.* **49** 2803–2824. MR4338384 https://doi.org/10.1214/21-aos2057

[33] RASINES, D. G. and YOUNG, G. A. (2023). Splitting strategies for post-selection inference. *Biometrika* **110** 597–614. MR4627773 https://doi.org/10.1093/biomet/asac070

[34] SONG, X., LI, G., ZHOU, Z., WANG, X., IONITA-LAZA, I. and WEI, Y. (2017). QRank: A novel quantile regression tool for eQTL discovery. *Bioinformatics* **33** 2123–2130.

[35] TAN, K. M., WANG, L. and ZHOU, W.-X. (2022). High-dimensional quantile regression: Convolution smoothing and concave regularization. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 205–233. MR4400395 https://doi.org/10.1111/rssb.12485

[36] TANG, Y., WANG, Y., WANG, H. J. and PAN, Q. (2022). Conditional marginal test for high dimensional quantile regression. *Statist. Sinica* **32** 869–892. MR4405101 https://doi.org/10.5705/ss.202019.0304

[37] TIAN, X. and TAYLOR, J. (2018). Selective inference with a randomized response. *Ann. Statist.* **46** 679–710. MR3782381 https://doi.org/10.1214/17-AOS1564

[38] TIBSHIRANI, R. J., RINALDO, A., TIBSHIRANI, R. and WASSERMAN, L. (2018). Uniform asymptotic inference and the bootstrap after model selection. *Ann. Statist.* **46** 1255–1287. MR3798003 https://doi.org/10.1214/17-AOS1584

[39] WANG, H. J., MCKEAGUE, I. W. and QIAN, M. (2018). Testing for marginal linear effects in quantile regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 433–452. MR3763699 https://doi.org/10.1111/rssb.12258

[40] WANG, Y., PANIGRAHI, S. and HE, X. (2025). Supplement to “Asymptotically-exact selective inference for quantile regression.” https://doi.org/10.1214/24-AOS2488SUPP

[41] WEI, Y., PERE, A., KOENKER, R. and HE, X. (2006). Quantile regression methods for reference growth charts. *Stat. Med.* **25** 1369–1382. MR2226792 https://doi.org/10.1002/sim.2271

[42] YAN, Y., WANG, X. and ZHANG, R. (2023). Confidence intervals and hypothesis testing for high-dimensional quantile regression: Convolution smoothing and debiasing. *J. Mach. Learn. Res.* **24** Paper No. [245], 49. MR4633972

[43] ZHANG, D., KHALILI, A. and ASGHARIAN, M. (2022). Post-model-selection inference in linear regression models: An integrated review. *Stat. Surv.* **16** 86–136. MR4390422 https://doi.org/10.1214/22-ss135

[44] ZRNIC, T. and JORDAN, M. I. (2023). Post-selection inference via algorithmic stability. *Ann. Statist.* **51** 1666–1691. MR4658572 https://doi.org/10.1214/23-aos2303

The Annals of Statistics
2025, Vol. 53, No. 6, 2380–2408
https://doi.org/10.1214/25-AOS2519

COUNTERFACTUAL INFERENCE IN SEQUENTIAL EXPERIMENTS

BY RAAZ DWIVEDI[1,a], KATHERINE TIAN[2,b], SABINA TOMKINS[3,d], PREDRAG KLASNJA[3,e], SUSAN MURPHY[2,c] AND DEVAVRAT SHAH[4,f]

[1]*Department of Operations Research and Information Engineering, Cornell Tech,* [a]*dwivedi@cornell.edu*
[2]*Department of Statistics, Harvard University,* [b]*kattian@alumni.harvard.edu,* [c]*samurphy@harvard.edu*
[3]*School of Information, University of Michigan,* [d]*stomkins@umich.edu,* [e]*klasnja@umich.edu*
[4]*Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology,* [f]*devavrat@mit.edu*

We consider after-study statistical inference for sequentially designed experiments wherein multiple units are assigned treatments for multiple time points using treatment policies that adapt over time. Our goal is to provide inference guarantees for the counterfactual mean at the smallest possible scale—mean outcome under different treatments *for each unit and each time*—with minimal assumptions on the adaptive treatment policy. Without any structural assumptions on the counterfactual means, this challenging task is infeasible due to more unknowns than observed data points. To make progress, we introduce a latent factor model over the counterfactual means that serves as a nonparametric generalization of the nonlinear mixed effects model and the bilinear latent factor model considered in prior works. For estimation, we use a nonparametric method, namely a variant of nearest neighbors, and establish a nonasymptotic high probability error bound for the counterfactual mean for each unit and each time. Under regularity conditions, this bound leads to asymptotically valid confidence intervals for the counterfactual mean as the number of units and time points grows to ∞ together at suitable rates. We illustrate our theory via several simulations and a case study involving data from a mobile health clinical trial HeartSteps.

REFERENCES

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

[2] AGARWAL, A., DAHLEH, M., SHAH, D. and SHEN, D. (2021). Causal matrix completion. arXiv preprint. Available at arXiv:2109.15154.

[3] AGARWAL, A., SHAH, D. and SHEN, D. (2021). Synthetic interventions. arXiv preprint. Available at arXiv:2006.07691.

[4] ALLEN, S. (2022). Stanford Computational Policy Lab Pretrial Nudges. https://policylab.stanford.edu/projects/nudge.html.

[5] ARKHANGELSKY, D., ATHEY, S., HIRSHBERG, D. A., IMBENS, G. W. and WAGER, S. (2021). Synthetic difference-in-differences. *Amer. Econ. Rev.* **111** 4088–4118. MR4348786 https://doi.org/10.1016/j.jeconom.2020.10.012

[6] ATHEY, S., BAYATI, M., DOUDCHENKO, N., IMBENS, G. and KHOSRAVI, K. (2021). Matrix completion methods for causal panel data models. *J. Amer. Statist. Assoc.* **116** 1716–1730. MR4353709 https://doi.org/10.1080/01621459.2021.1891924

[7] AVADHANULA, V., COLINI BALDESCHI, R., LEONARDI, S., SANKARARAMAN, K. A. and SCHRIJVERS, O. (2021). Stochastic bandits for multi-platform budget optimization in online advertising. In *Proceedings of the Web Conference* **2021** 2805–2817.

[8] BAI, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* **71** 135–171. MR1956857 https://doi.org/10.1111/1468-0262.00392

MSC2020 subject classifications. Primary 62L10, 62L05, 62K05; secondary 62G20.
Key words and phrases. Sequential experiments, counterfactual inference, adaptive randomization, nonlinear factor model, mixed effects model, nearest neighbors.

[9] BAI, J. (2024). Likelihood approach to dynamic panel models with interactive effects. *J. Econometrics* **240** Paper No. 105636, 31. MR4685892 https://doi.org/10.1016/j.jeconom.2023.105636

[10] BAI, J. and NG, S. (2021). Matrix completion, counterfactuals, and factor analysis of missing data. *J. Amer. Statist. Assoc.* **116** 1746–1763. MR4353711 https://doi.org/10.1080/01621459.2021.1967163

[11] BAKSHY, E. and ECKLES, D. (2013). Uncertainty in online experiments with dependent data: An evaluation of bootstrap methods. In *Proceedings of the* 19*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1303–1311.

[12] BERETTA, L. and SANTANIELLO, A. (2016). Nearest neighbor imputation algorithms: A critical evaluation. *BMC Med. Inform. Decis. Mak.* **16** 197–208.

[13] BERTSIMAS, D., KALLUS, N., WEINSTEIN, A. M. and ZHUO, Y. D. (2017). Personalized diabetes management using electronic medical records. *Diabetes Care* **40** 210–217.

[14] BERTSIMAS, D., PAWLOWSKI, C. and ZHUO, Y. D. (2017). From predictive methods to missing data imputation: An optimization approach. *J. Mach. Learn. Res.* **18** Paper No. 196, 39. MR3827084

[15] BIAN, Z., SHI, C., QI, Z. and WANG, L. (2025). Off-policy evaluation in doubly inhomogeneous environments. *J. Amer. Statist. Assoc.* **120** 1102–1114. MR4917281 https://doi.org/10.1080/01621459.2024.2395593

[16] BIBAUT, A., CHAMBAZ, A., DIMAKOPOULOU, M., KALLUS, N. and VAN DER LAAN, M. (2021). Post-contextual-bandit inference. In *NeurIPS*.

[17] BOJINOV, I., RAMBACHAN, A. and SHEPHARD, N. (2021). Panel experiments and dynamic causal effects: A finite population perspective. *Quant. Econ.* **12** 1171–1196. MR4399594 https://doi.org/10.3982/qe1744

[18] CAI, W., GROSSMAN, J., LIN, Z. J., SHENG, H., WEI, J. T.-Z., WILLIAMS, J. J. and GOEL, S. (2021). Bandit algorithms to personalize educational chatbots. *Mach. Learn.* **110** 2389–2418. MR4306054 https://doi.org/10.1007/s10994-021-05983-y

[19] CARIA, S., KASY, M., QUINN, S., SHAMI, S., TEYTELBOYM, A. et al. (2020). An adaptive targeted field experiment: Job search assistance for refugees in Jordan.

[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] DAVENPORT, M. A. and ROMBERG, J. (2016). An overview of low-rank matrix recovery from incomplete observations. *IEEE J. Sel. Top. Signal Process.* **10** 608–622. https://doi.org/10.1109/JSTSP.2016.2539100

[22] DAVIDIAN, M. and GILTINAN, D. M. (2017). *Nonlinear models for repeated measurement data*. Routledge, London.

[23] DWIVEDI, R., TIAN, K., TOMKINS, S., KLASNJA, P., MURPHY, S. and SHAH, D. (2025). Supplement to "Counterfactual inference in sequential experiments." https://doi.org/10.1214/25-AOS2519SUPP

[24] ERRAQABI, A., LAZARIC, A., VALKO, M., BRUNSKILL, E. and LIU, Y.-E. (2017). Trading off rewards and errors in multi-armed bandits. In *Artificial Intelligence and Statistics* 709–717. PMLR.

[25] FORMAN, E. M., KERRIGAN, S. G., BUTRYN, M. L., JUARASCIO, A. S., MANASSE, S. M., ONTAÑÓN, S., DALLAL, D. H., CROCHIERE, R. J. and MOSKOW, D. (2019). Can the artificial intelligence technique of reinforcement learning use continuously-monitored digital data to optimize treatment for weight loss? *Int. J. Behav. Med.* **42** 276–290.

[26] HADAD, V., HIRSHBERG, D. A., ZHAN, R., WAGER, S. and ATHEY, S. (2021). Confidence intervals for policy evaluation in adaptive experiments. *Proc. Natl. Acad. Sci. USA* **118** Paper No. e2014602118, 10. MR4294058 https://doi.org/10.1073/pnas.2014602118

[27] KALLUS, N. and UEHARA, M. (2020). Double reinforcement learning for efficient off-policy evaluation in Markov decision processes. *J. Mach. Learn. Res.* **21** Paper No. 167, 63. MR4209453

[28] KASY, M. and SAUTMANN, A. (2021). Adaptive treatment assignment in experiments for policy choice. *Econometrica* **89** 113–132. MR4220384 https://doi.org/10.3982/ecta17527

[29] LAIRD, N. M. and WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* 963–974.

[30] LEE, C. E., LI, Y., SHAH, D. and SONG, D. (2016). Blind regression: Nonparametric regression for latent variable models via collaborative filtering. In *Advances in Neural Information Processing Systems* **29** 2155–2163.

[31] LI, L., CHU, W., LANGFORD, J. and SCHAPIRE, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the* 19*th International Conference on World Wide Web* 661–670.

[32] LI, Y., SHAH, D., SONG, D. and YU, C. L. (2020). Nearest neighbors for matrix estimation interpreted as blind regression for latent variable model. *IEEE Trans. Inf. Theory* **66** 1760–1784. MR4077516 https://doi.org/10.1109/tit.2019.2950299

[33] LIAO, P., GREENEWALD, K., KLASNJA, P. and MURPHY, S. (2020). Personalized HeartSteps: A reinforcement learning algorithm for optimizing physical activity. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **4** 1–22.

[34] LIU, Y.-E., MANDEL, T., BRUNSKILL, E. and POPOVIC, Z. (2014). Trading off scientific knowledge and user learning with multi-armed bandits. In *EDM* 161–168.

[35] NEYMAN, J. (1923). Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes Master's Thesis.

[36] QI, Y., WU, Q., WANG, H., TANG, J. and SUN, M. (2018). Bandit learning with implicit feedback. *Adv. Neural Inf. Process. Syst.* **31**.

[37] RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196 https://doi.org/10.1093/biomet/63.3.581

[38] RUBIN, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys*. *Wiley Classics Library*. Wiley-Interscience, Hoboken, NJ. Reprint of the 1987 edition [John Wiley & Sons, Inc., New York; MR899519]. MR2117498

[39] RUSSO, D., ROY, B. V., KAZEROUNI, A., OSBAND, I. and WEN, Z. (2018). A tutorial on Thompson sampling. *Found. Trends Mach. Learn.* **11** 1–96. https://doi.org/10.1561/2200000070

[40] SAWANT, N., NAMBALLA, C. B., SADAGOPAN, N. and NASSIF, H. (2018). Contextual multi-armed bandits for causal marketing. arXiv preprint. Available at arXiv:1810.01859.

[41] SCHWARTZ, E. M., BRADLOW, E. T. and FADER, P. S. (2017). Customer acquisition via display advertising using multi-armed bandit experiments. *Mark. Sci.* **36** 500–522.

[42] SHAIKH, H., MODIRI, A., WILLIAMS, J. J. and RAFFERTY, A. N. (2019). Balancing student success and inferring personalized effects in dynamic experiments. In *EDM*.

[43] TOMKINS, S., LIAO, P., KLASNJA, P. and MURPHY, S. (2021). IntelligentPooling: Practical Thompson sampling for mHealth. *Mach. Learn.* **110** 2685–2727. MR4306062 https://doi.org/10.1007/s10994-021-05995-8

[44] UEHARA, M., SHI, C. and KALLUS, N. (2022). A review of off-policy evaluation in reinforcement learning. arXiv preprint. Available at arXiv:2212.06355.

[45] VONESH, E. F. and CARTER, R. L. (1992). Mixed-effects nonlinear regression for unbalanced repeated measures. *Biometrics* **48** 1–17. MR1165614 https://doi.org/10.2307/2532734

[46] WANG, S., HU, L., WANG, Y., CAO, L., SHENG, Q. Z. and ORGUN, M. (2019). Sequential recommender systems: Challenges, progress and prospects. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19. International Joint Conferences on Artificial Intelligence Organization* 6332–6338. https://doi.org/10.24963/ijcai.2019/883

[47] XU, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Polit. Anal.* **25** 57–76.

[48] YOM-TOV, E., FERARU, G., KOZDOBA, M., MANNOR, S., TENNENHOLTZ, M. and HOCHBERG, I. (2017). Encouraging physical activity in patients with diabetes: Treatment using a reinforcement learning system. *J. Med. Internet Res.* **19** e338.

[49] YOUNG, D. A., ZERBE, G. O. and HAY, W. W. JR (1997). Fieller's theorem, Scheffé simultaneous confidence intervals, and ratios of parameters of linear and nonlinear mixed-effects models. *Biometrics* 838–847.

[50] YU, C. L. (2022). Nonparametric matrix estimation with one-sided covariates. In 2022 *IEEE International Symposium on Information Theory* (*ISIT*) 892–897. IEEE.

[51] ZHAN, R., HADAD, V., HIRSHBERG, D. A. and ATHEY, S. (2021). Off-policy evaluation via adaptive weighting with data from contextual bandits. In *Proceedings of the* 27*th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

[52] ZHANG, K., JANSON, L. and MURPHY, S. (2020). Inference for batched bandits. In *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan and H. Lin, eds.) **33** 9818–9829. Curran Associates, Red Hook.

[53] ZHANG, K. W., JANSON, L. and MURPHY, S. A. (2020). Inference for batched bandits. In *Advances in Neural Information Processing Systems*.

[54] ZHANG, K. W. JANSON, L. and MURPHY, S. A. (2021). Statistical inference with M-estimators on adaptively collected data.

The Annals of Statistics
2025, Vol. 53, No. 6, 2409–2429
https://doi.org/10.1214/25-AOS2546

CLUSTERING RISK IN NONPARAMETRIC HIDDEN MARKOV AND I.I.D. MODELS

BY ÉLISABETH GASSIAT[1,a], IBRAHIM KADDOURI[1,b] AND ZACHARIE NAULET[2,c]

[1]*Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay,* [a]*elisabeth.gassiat@universite-paris-saclay.fr,* [b]*ibrahim.kaddouri@universite-paris-saclay.fr*

[2]*Université Paris-Saclay, INRAE, MaIAGE,* [c]*zacharie.naulet@inrae.fr*

We conduct an in-depth analysis of the Bayes risk of clustering in the context of hidden Markov and i.i.d. models. In both settings, we identify the situations where this risk is comparable to the Bayes risk of classification and those where its minimizer, the Bayes clusterer, can be derived from the Bayes classifier. While we demonstrate that clustering based on the Bayes classifier does not always match the optimal Bayes clusterer, we show that this difference is primarily theoretical and that the Bayes classifier remains nearly optimal for clustering. A key quantity emerges, capturing the fundamental difficulty of both classification and clustering tasks. Furthermore, by leveraging the identifiability of HMMs, we establish bounds on the clustering excess risk of a plug-in Bayes classifier in the general nonparametric setting, offering theoretical justification for its widespread use in practice. Simulations further illustrate our findings.

REFERENCES

[1] ABRAHAM, K., CASTILLO, I. and GASSIAT, E. (2022). Multiple testing in nonparametric hidden Markov models: An empirical Bayes approach. *J. Mach. Learn. Res.* **23** Paper No. [94]. MR4576679

[2] ABRAHAM, K., GASSIAT, E. and NAULET, Z. (2023). Frontiers to the learning of nonparametric hidden Markov models. arXiv preprint. Available at arXiv:2306.16293.

[3] ABRAHAM, K., GASSIAT, E. and NAULET, Z. (2023). Fundamental limits for learning hidden Markov model parameters. *IEEE Trans. Inf. Theory* **69** 1777–1794. MR4564681 https://doi.org/10.1109/tit.2022.3213429

[4] ALEXANDROVICH, G., HOLZMANN, H. and LEISTER, A. (2016). Nonparametric identification and maximum likelihood estimation for hidden Markov models. *Biometrika* **103** 423–434. MR3509896 https://doi.org/10.1093/biomet/asw001

[5] 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

[6] CAPPÉ, O., MOULINES, E. and RYDÉN, T. (2005). *Inference in Hidden Markov Models. Springer Series in Statistics*. Springer, New York. MR2159833

[7] DE CASTRO, Y., GASSIAT, É. and LACOUR, C. (2016). Minimax adaptive estimation of nonparametric hidden Markov models. *J. Mach. Learn. Res.* **17** Paper No. 111. MR3543517

[8] DE CASTRO, Y., GASSIAT, É. and LE CORFF, S. (2017). Consistent estimation of the filtering and marginal smoothing distributions in nonparametric hidden Markov models. *IEEE Trans. Inf. Theory* **63** 4758–4777. MR3683535 https://doi.org/10.1109/TIT.2017.2696959

[9] DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition. Applications of Mathematics* (*New York*) **31**. Springer, New York. MR1383093 https://doi.org/10.1007/978-1-4612-0711-5

[10] DE CASTRO, Y. (2016). Nonparametric HMM. Available at https://github.com/ydecastro/nonparametric-hmm.

[11] GAO, C., MA, Z., ZHANG, A. Y. and ZHOU, H. H. (2018). Community detection in degree-corrected block models. *Ann. Statist.* **46** 2153–2185. MR3845014 https://doi.org/10.1214/17-AOS1615

[12] GASSIAT, E., KADDOURI, I. and NAULET, Z. (2025). Supplement to "Clustering risk in nonparametric hidden Markov and i.i.d. models." https://doi.org/10.1214/25-AOS2546SUPP

MSC2020 subject classifications. Primary 62M05; secondary 62G99.
Key words and phrases. Clustering, nonparametric populations, hidden Markov models, mixture models.

[13] GHASSEMPOUR, S., GIROSI, F. and MAEDER, A. (2014). Clustering multivariate time series using hidden Markov models. *Int. J. Environ. Res. Public Health* **3**.

[14] GIRAUD, C. (2022). *Introduction to High-Dimensional Statistics*, 2nd ed. *Monographs on Statistics and Applied Probability* **168**. CRC Press, Boca Raton, FL. MR4436193 https://doi.org/10.1201/9781003158745

[15] GIRAUD, C. and VERZELEN, N. (2018). Partial recovery bounds for clustering with the relaxed K-means. *Math. Stat. Learn.* **1** 317–374. MR4059724

[16] GRÜN, B. (2019). Model-based clustering. In *Handbook of Mixture Analysis*. *Chapman & Hall/CRC Handb. Mod. Stat. Methods* 157–192. CRC Press, Boca Raton, FL. MR3889693

[17] KANNAN, R., SALMASIAN, H. and VEMPALA, S. (2005). The spectral method for general mixture models. In *Learning Theory*. *Lecture Notes in Computer Science* **3559** 444–457. Springer, Berlin. MR2203279 https://doi.org/10.1007/11503415_30

[18] KHIATANI, D. and GHOSE, U. (2017). Weather forecasting using hidden Markov model. In 2017 *International Conference on Computing and Communication Technologies for Smart Nation* (*IC3TSN*) 220–225.

[19] LEHÉRICY, L. (2018). State-by-state minimax adaptive estimation for nonparametric hidden Markov models. *J. Mach. Learn. Res.* **19** Paper No. 39. MR3862446

[20] LEHÉRICY, L. (2021). Nonasymptotic control of the MLE for misspecified nonparametric hidden Markov models. *Electron. J. Stat.* **15** 4916–4965. MR4347381 https://doi.org/10.1214/21-ejs1890

[21] LU, Y. and ZHOU, H. (2016). Statistical and Computational Guarantees of Lloyd's Algorithm and Its Variants. arXiv preprint. Available at arXiv:1612.02099.

[22] MARANDON, A., REBAFKA, T. and ROQUAIN, E. (2023). False clustering rate control in mixture models. arXiv preprint. Available at arXiv:2203.02597.

[23] MEILĂ, M. (2005). Comparing clusterings: An axiomatic view. In *Proceedings of the 22nd International Conference on Machine Learning* 577–584.

[24] MEILĂ, M. and HECKERMAN, D. (2001). An experimental comparison of model-based clustering methods. *Mach. Learn.* **42** 9–29.

[25] MIXON, D. G., VILLAR, S. and WARD, R. (2017). Clustering subgaussian mixtures by semidefinite programming. *Inf. Inference* **6** 389–415. MR3764529 https://doi.org/10.1093/imaiai/iax001

[26] NDAOUD, M. (2022). Sharp optimal recovery in the two component Gaussian mixture model. *Ann. Statist.* **50** 2096–2126. MR4474484 https://doi.org/10.1214/22-aos2178

[27] SCHLIEP, A., SCHÖNHUTH, A. and STEINHOFF, C. (2003). Using hidden Markov models to analyze gene expression time course data. *Bioinformatics*.

[28] SHI, T., BELKIN, M. and YU, B. (2009). Data spectroscopy: Eigenspaces of convolution operators and clustering. *Ann. Statist.* **37** 3960–3984. MR2572449 https://doi.org/10.1214/09-AOS700

[29] ZHANG, A. Y. and ZHOU, H. H. (2016). Minimax rates of community detection in stochastic block models. *Ann. Statist.* **44** 2252–2280. MR3546450 https://doi.org/10.1214/15-AOS1428

The Annals of Statistics
2025, Vol. 53, No. 6, 2430–2461
https://doi.org/10.1214/25-AOS2547

CAUSAL EFFECT ESTIMATION UNDER NETWORK INTERFERENCE WITH MEAN-FIELD METHODS

BY SOHOM BHATTACHARYA[1,a] AND SUBHABRATA SEN[2,b]

[1]*Department of Statistics, University of Florida,* [a]*bhattacharya.s@ufl.edu*
[2]*Department of Statistics, Harvard University,* [b]*subhabratasen@fas.harvard.edu*

We study causal effect estimation from observational data under interference. The interference pattern is captured by an observed network. We adopt the chain graph framework of (*J. Amer. Statist. Assoc.* **116** (2021) 833–844), which allows (i) interaction among the outcomes of distinct study units connected along the graph and (ii) long range interference, whereby the outcome of an unit may depend on the treatments assigned to distant units connected along the interference network. For "mean-field" interaction networks, we develop a new scalable iterative algorithm to estimate the causal effects. For Gaussian weighted networks, we introduce a novel causal effect estimation algorithm based on Approximate Message Passing (AMP). Our algorithms are provably consistent under a "high temperature" condition on the underlying model. We estimate the (unknown) parameters of the model from data using maximum pseudo-likelihood and establish $\sqrt{n}$-consistency of this estimator in all parameter regimes. Finally, we prove that the downstream estimators obtained by plugging in estimated parameters into the aforementioned algorithms are consistent at high temperature. Our methods can accommodate dense interactions among the study units—a setting beyond reach using existing techniques. Our algorithms originate from the study of variational inference approaches in high-dimensional statistics; overall, we demonstrate the usefulness of these ideas in the context of causal effect estimation under interference.

REFERENCES

[1] AKIYAMA, S., KURAMASHI, Y., YAMASHITA, T. and YOSHIMURA, Y. (2019). Phase transition of four-dimensional Ising model with higher-order tensor renormalization group. *Phys. Rev. D* **100** 054510. MR4025367 https://doi.org/10.1103/physrevd.100.054510

[2] ANGRIST, J. D. (2014). The perils of peer effects. *Labour Econ.* **30** 98–108.

[3] ARONOW, P. M. and SAMII, C. (2017). Estimating average causal effects under general interference, with application to a social network experiment. *Ann. Appl. Stat.* **11** 1912–1947. MR3743283 https://doi.org/10.1214/16-AOAS1005

[4] ATHEY, S., ECKLES, D. and IMBENS, G. W. (2018). Exact p-values for network interference. *J. Amer. Statist. Assoc.* **113** 230–240. MR3803460 https://doi.org/10.1080/01621459.2016.1241178

[5] BAI, Z. and SILVERSTEIN, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2567175 https://doi.org/10.1007/978-1-4419-0661-8

[6] BASAK, A. and MUKHERJEE, S. (2017). Universality of the mean-field for the Potts model. *Probab. Theory Related Fields* **168** 557–600. MR3663625 https://doi.org/10.1007/s00440-016-0718-0

[7] BASSE, G. W., FELLER, A. and TOULIS, P. (2019). Randomization tests of causal effects under interference. *Biometrika* **106** 487–494. MR3949317 https://doi.org/10.1093/biomet/asy072

[8] BAUERSCHMIDT, R. and BODINEAU, T. (2019). A very simple proof of the LSI for high temperature spin systems. *J. Funct. Anal.* **276** 2582–2588. MR3926125 https://doi.org/10.1016/j.jfa.2019.01.007

[9] 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

MSC2020 subject classifications. Primary 62H22, 62F12; secondary 82B44.
Key words and phrases. Causal inference, network interference, Variational Inference, Approximate Message Passing.

[10] BENAYCH-GEORGES, F., BORDENAVE, C. and KNOWLES, A. (2019). Largest eigenvalues of sparse inhomogeneous Erdős-Rényi graphs. *Ann. Probab.* **47** 1653–1676. MR3945756 https://doi.org/10.1214/18-AOP1293

[11] BESAG, J. (1975). Statistical analysis of non-lattice data. *J. R. Stat. Soc., Ser. D, Stat.* **24** 179–195.

[12] BHATTACHARYA, R., MALINSKY, D. and SHPITSER, I. (2020). Causal inference under interference and network uncertainty. In *Uncertainty in Artificial Intelligence* 1028–1038. PMLR.

[13] BHATTACHARYA, S., DEB, N. and MUKHERJEE, S. (2023). Gibbs measures with multilinear forms. arXiv preprint. Available at arXiv:2307.14600.

[14] BHATTACHARYA, S., DEB, N. and MUKHERJEE, S. (2024). LDP for inhomogeneous U-statistics. *Ann. Appl. Probab.* **34** 5769–5808. MR4839638 https://doi.org/10.1214/24-aap2107

[15] BHATTACHARYA, S., MUKHERJEE, R. and RAY, G. (2025). Sharp signal detection under ferromagnetic Ising models. *IEEE Trans. Inf. Theory* **71** 3925–3949. MR4897833 https://doi.org/10.1109/tit.2025.3546976

[16] BHATTACHARYA, S. and SEN, S. (2025). Supplement to "Causal effect estimation under network interference with mean-field methods." https://doi.org/10.1214/25-AOS2547SUPP

[17] BLEI, D. M., KUCUKELBIR, A. and MCAULIFFE, J. D. (2017). Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* **112** 859–877. MR3671776 https://doi.org/10.1080/01621459.2017.1285773

[18] BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford Univ. Press, Oxford. MR3185193 https://doi.org/10.1093/acprof:oso/9780199535255.001.0001

[19] BRAMOULLÉ, Y., DJEBBARI, H. and FORTIN, B. (2009). Identification of peer effects through social networks. *J. Econometrics* **150** 41–55. MR2525993 https://doi.org/10.1016/j.jeconom.2008.12.021

[20] CHATTERJEE, S. and DEMBO, A. (2016). Nonlinear large deviations. *Adv. Math.* **299** 396–450. MR3519474 https://doi.org/10.1016/j.aim.2016.05.017

[21] CHEN, W.-K. and LAM, W.-K. (2021). Universality of approximate message passing algorithms. *Electron. J. Probab.* **26** 36. MR4235487 https://doi.org/10.1214/21-EJP604

[22] CHEN, W.-K. and TANG, S. (2021). On convergence of the cavity and Bolthausen's TAP iterations to the local magnetization. *Comm. Math. Phys.* **386** 1209–1242. MR4294289 https://doi.org/10.1007/s00220-021-04103-0

[23] CHIN, A. (2019). Regression adjustments for estimating the global treatment effect in experiments with interference. *J. Causal Inference* **7** 20180026. MR4350071 https://doi.org/10.1515/jci-2018-0026

[24] CORTEZ-RODRIGUEZ, M., EICHHORN, M. and YU, C. L. (2023). Exploiting neighborhood interference with low-order interactions under unit randomized design. *J. Causal Inference* **11** 20220051. MR4622849 https://doi.org/10.1515/jci-2022-0051

[25] DASKALAKIS, C., DIKKALA, N. and PANAGEAS, I. (2019). Regression from dependent observations. In *STOC'19—Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing* 881–889. ACM, New York. MR4003392

[26] DUDEJA, R., LU, Y. M. and SEN, S. (2023). Universality of approximate message passing with semirandom matrices. *Ann. Probab.* **51** 1616–1683. MR4642220 https://doi.org/10.1214/23-aop1628

[27] ECKLES, D., KARRER, B. and UGANDER, J. (2017). Design and analysis of experiments in networks: Reducing bias from interference. *J. Causal Inference* **5** 20150021. MR4323809 https://doi.org/10.1515/jci-2015-0021

[28] EICHHORN, M., KHAN, S., UGANDER, J. and YU, C. L. (2024). Low-order outcomes and clustered designs: Combining design and analysis for causal inference under network interference. arXiv preprint. Available at arXiv:2405.07979.

[29] ELDAN, R. (2018). Gaussian-width gradient complexity, reverse log-Sobolev inequalities and nonlinear large deviations. *Geom. Funct. Anal.* **28** 1548–1596. MR3881829 https://doi.org/10.1007/s00039-018-0461-z

[30] ELDAN, R. (2020). Taming correlations through entropy-efficient measure decompositions with applications to mean-field approximation. *Probab. Theory Related Fields* **176** 737–755. MR4087482 https://doi.org/10.1007/s00440-019-00924-2

[31] ELDAN, R., KOEHLER, F. and ZEITOUNI, O. (2022). A spectral condition for spectral gap: Fast mixing in high-temperature Ising models. *Probab. Theory Related Fields* **182** 1035–1051. MR4408509 https://doi.org/10.1007/s00440-021-01085-x

[32] EMMENEGGER, C., SPOHN, M.-L., ELMER, T. and BÜHLMANN, P. (2025). Treatment effect estimation with observational network data using machine learning. *J. Causal Inference* **13** 20230082. MR4886357 https://doi.org/10.1515/jci-2023-0082

[33] FERRACCI, M., JOLIVET, G. and VAN DEN BERG, G. J. (2014). Evidence of treatment spillovers within markets. *Rev. Econom. Statist.* **96** 812–823.

[34] FORASTIERE, L., AIROLDI, E. M. and MEALLI, F. (2021). Identification and estimation of treatment and interference effects in observational studies on networks. *J. Amer. Statist. Assoc.* **116** 901–918. MR4270033 https://doi.org/10.1080/01621459.2020.1768100

[35] FRIEDLI, S. and VELENIK, Y. (2018). *Statistical Mechanics of Lattice Systems: A Concrete Mathematical Introduction*. Cambridge Univ. Press, Cambridge. MR3752129

[36] GALANIS, A., KALAVASIS, A. and KANDIROS, A. V. (2024). On sampling from Ising models with spectral constraints. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques. LIPIcs. Leibniz Int. Proc. Inform.* **317** 70. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern. MR4802580 https://doi.org/10.4230/lipics.approx/random.2024.70

[37] GAMARNIK, D. (2013). Correlation decay method for decision, optimization, and inference in large-scale networks. In *Theory Driven by Influential Applications* 108–121. INFORMS.

[38] GHANG, W., MARTIN, Z. and WARUHIU, S. (2014). The sharp log-Sobolev inequality on a compact interval. *Involve* **7** 181–186. MR3133718 https://doi.org/10.2140/involve.2014.7.181

[39] GHOSAL, P. and MUKHERJEE, S. (2020). Joint estimation of parameters in Ising model. *Ann. Statist.* **48** 785–810. MR4102676 https://doi.org/10.1214/19-AOS1822

[40] GOLDSMITH-PINKHAM, P. and IMBENS, G. W. (2013). Social networks and the identification of peer effects. *J. Bus. Econom. Statist.* **31** 253–264. MR3173674 https://doi.org/10.1080/07350015.2013.801251

[41] GRAHAM, B. S. (2008). Identifying social interactions through conditional variance restrictions. *Econometrica* **76** 643–660. MR2406869 https://doi.org/10.1111/j.1468-0262.2008.00850.x

[42] HAYES, R. J. and MOULTON, L. H. (2017). *Cluster Randomised Trials*. CRC Press, Boca Raton.

[43] HONG, G. and RAUDENBUSH, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *J. Amer. Statist. Assoc.* **101** 901–910. MR2324091 https://doi.org/10.1198/016214506000000447

[44] HUDGENS, M. G. and HALLORAN, M. E. (2008). Toward causal inference with interference. *J. Amer. Statist. Assoc.* **103** 832–842. MR2435472 https://doi.org/10.1198/016214508000000292

[45] JAGADEESAN, R., PILLAI, N. S. and VOLFOVSKY, A. (2020). Designs for estimating the treatment effect in networks with interference. *Ann. Statist.* **48** 679–712. MR4102672 https://doi.org/10.1214/18-AOS1807

[46] JAIN, V., KOEHLER, F. and MOSSEL, E. (2018). The mean-field approximation: Information inequalities, algorithms, and complexity. In *Conference on Learning Theory* 1326–1347. PMLR.

[47] JAIN, V., RISTESKI, A. and KOEHLER, F. (2019). Mean-field approximation, convex hierarchies, and the optimality of correlation rounding: A unified perspective. In *STOC'19—Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing* 1226–1236. ACM, New York. MR4003424 https://doi.org/10.1145/3313276.3316299

[48] JIANG, K., MUKHERJEE, R., SEN, S. and SUR, P. (2025). A new central limit theorem for the augmented IPW estimator: Variance inflation, cross-fit covariance and beyond. *Ann. Statist.* **53** 647–675. MR4900161 https://doi.org/10.1214/24-aos2476

[49] KANG, H. and IMBENS, G. (2016). Peer encouragement designs in causal inference with partial interference and identification of local average network effects. arXiv preprint. Available at arXiv:1609.04464.

[50] KRIVELEVICH, M. and SUDAKOV, B. (2003). The largest eigenvalue of sparse random graphs. *Combin. Probab. Comput.* **12** 61–72. MR1967486 https://doi.org/10.1017/S0963548302005424

[51] KUNISKY, D. (2024). Optimality of Glauber dynamics for general-purpose Ising model sampling and free energy approximation. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)* 5013–5028. SIAM, Philadelphia, PA. MR4699425 https://doi.org/10.1137/1.9781611977912.180

[52] LATAŁA, R. and OLESZKIEWICZ, K. (2000). Between Sobolev and Poincaré. In *Geometric Aspects of Functional Analysis. Lecture Notes in Math.* **1745** 147–168. Springer, Berlin. MR1796718 https://doi.org/10.1007/BFb0107213

[53] LAURITZEN, S. L. and RICHARDSON, T. S. (2002). Chain graph models and their causal interpretations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 321–361. MR1924296 https://doi.org/10.1111/1467-9868.00340

[54] LEE, L. (2007). Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. *J. Econometrics* **140** 333–374. MR2408910 https://doi.org/10.1016/j.jeconom.2006.07.001

[55] LEUNG, M. P. (2022). Rate-optimal cluster-randomized designs for spatial interference. *Ann. Statist.* **50** 3064–3087. MR4500634 https://doi.org/10.1214/22-aos2224

[56] LI, S. and WAGER, S. (2022). Random graph asymptotics for treatment effect estimation under network interference. *Ann. Statist.* **50** 2334–2358. MR4474493 https://doi.org/10.1214/22-aos2191

[57] LIU, L. and HUDGENS, M. G. (2014). Large sample randomization inference of causal effects in the presence of interference. *J. Amer. Statist. Assoc.* **109** 288–301. MR3180564 https://doi.org/10.1080/01621459.2013.844698

[58] LIU, T., MUKHERJEE, S. and BISWAS, R. (2024). Tensor recovery in high-dimensional Ising models. *J. Multivariate Anal.* **203** 105335. MR4752227 https://doi.org/10.1016/j.jmva.2024.105335

[59] LUNDIN, M. and KARLSSON, M. (2014). Estimation of causal effects in observational studies with interference between units. *Stat. Methods Appl.* **23** 417–433. MR3233970 https://doi.org/10.1007/s10260-014-0257-8

[60] MANSKI, C. F. (1993). Identification of endogenous social effects: The reflection problem. *Rev. Econ. Stud.* **60** 531–542. MR1236836 https://doi.org/10.2307/2298123

[61] MANSKI, C. F. (2013). Identification of treatment response with social interactions. *Econom. J.* **16** S1–S23. MR3030060 https://doi.org/10.1111/j.1368-423X.2012.00368.x

[62] MÉZARD, M. and MONTANARI, A. (2009). *Information, Physics, and Computation. Oxford Graduate Texts*. Oxford Univ. Press, Oxford. MR2518205 https://doi.org/10.1093/acprof:oso/9780198570837.001.0001

[63] MUKHERJEE, S., NIU, Z., HALDER, S., BHATTACHARYA, B. B. and MICHAILIDIS, G. (2024). Logistic regression under network dependence. *J. Mach. Learn. Res.* **25** 1–62.

[64] MUKHERJEE, S. and SEN, S. (2022). Variational inference in high-dimensional linear regression. *J. Mach. Learn. Res.* **23** 304. MR4577743

[65] MUKHERJEE, S., SON, J. and BHATTACHARYA, B. B. (2022). Estimation in tensor Ising models. *Inf. Inference* **11** 1457–1500. MR4526327 https://doi.org/10.1093/imaiai/iaac007

[66] OGBURN, E. L., SHPITSER, I. and LEE, Y. (2020). Causal inference, social networks and chain graphs. *J. Roy. Statist. Soc. Ser. A* **183** 1659–1676. MR4157830 https://doi.org/10.1111/rssa.12594

[67] OGBURN, E. L., SOFRYGIN, O., DÍAZ, I. and VAN DER LAAN, M. J. (2024). Causal inference for social network data. *J. Amer. Statist. Assoc.* **119** 597–611. MR4713916 https://doi.org/10.1080/01621459.2022.2131557

[68] PARK, C. and KANG, H. (2023). Assumption-lean analysis of cluster randomized trials in infectious diseases for intent-to-treat effects and network effects. *J. Amer. Statist. Assoc.* **118** 1195–1206. MR4595487 https://doi.org/10.1080/01621459.2021.1983437

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

[70] ROBINS, J. M. and ROTNITZKY, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology: Methodological Issues* 297–331. Springer, Berlin.

[71] ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866. MR1294730

[72] ROSENBAUM, P. R. (2007). Interference between units in randomized experiments. *J. Amer. Statist. Assoc.* **102** 191–200. MR2345537 https://doi.org/10.1198/016214506000001112

[73] RUBIN, D. B. (1990). Formal mode of statistical inference for causal effects. *J. Statist. Plann. Inference* **25** 279–292.

[74] SASAKURA, N. and SATO, Y. (2014). Ising model on random networks and the canonical tensor model. *Prog. Theor. Exp. Phys.* **2014** 053B03.

[75] SÄVJE, F., ARONOW, P. M. and HUDGENS, M. G. (2021). Average treatment effects in the presence of unknown interference. *Ann. Statist.* **49** 673–701. MR4255103 https://doi.org/10.1214/20-aos1973

[76] SHERMAN, E. and SHPITSER, I. (2018). Identification and estimation of causal effects from dependent data. *Adv. Neural Inf. Process. Syst.* **31**.

[77] SHIRANI, S. and BAYATI, M. (2024). Causal message-passing for experiments with unknown and general network interference. *Proc. Natl. Acad. Sci. USA* **121** e2322232121. MR4826978 https://doi.org/10.1073/pnas.2322232121

[78] SHPITSER, I., TCHETGEN, E. T. and ANDREWS, R. (2017). Modeling interference via symmetric treatment decomposition. arXiv preprint. Available at arXiv:1709.01050.

[79] SOBEL, M. E. (2006). What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. *J. Amer. Statist. Assoc.* **101** 1398–1407. MR2307573 https://doi.org/10.1198/016214506000000636

[80] SOFRYGIN, O. and VAN DER LAAN, M. J. (2017). Semi-parametric estimation and inference for the mean outcome of the single time-point intervention in a causally connected population. *J. Causal Inference* **5** 20160003. MR4323811 https://doi.org/10.1515/jci-2016-0003

[81] TALAGRAND, M. (2010). *Mean Field Models for Spin Glasses: Volume I: Basic Examples* **54**. Springer, Berlin.

[82] TCHETGEN TCHETGEN, E. J., FULCHER, I. R. and SHPITSER, I. (2021). Auto-g-computation of causal effects on a network. *J. Amer. Statist. Assoc.* **116** 833–844. MR4270028 https://doi.org/10.1080/01621459.2020.1811098

[83] TCHETGEN TCHETGEN, E. J. and VANDERWEELE, T. J. (2012). On causal inference in the presence of interference. *Stat. Methods Med. Res.* **21** 55–75. MR2867538 https://doi.org/10.1177/0962280210386779

[84] TOULIS, P. and KAO, E. (2013). Estimation of causal peer influence effects. In *International Conference on Machine Learning* 1489–1497. PMLR.

[85] UGANDER, J., KARRER, B., BACKSTROM, L. and KLEINBERG, J. (2013). Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the* 19*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 329–337.

[86] VANDERWEELE, T. J. (2010). Direct and indirect effects for neighborhood-based clustered and longitudinal data. *Sociol. Methods Res.* **38** 515–544. MR2758165 https://doi.org/10.1177/0049124110366236

[87] VAN DER LAAN, M. J. (2014). Causal inference for a population of causally connected units. *J. Causal Inference* **2** 13–74. MR4289412 https://doi.org/10.1515/jci-2013-0002

[88] VAN DER LAAN, M. J. and RUBIN, D. (2006). Targeted maximum likelihood learning. *Int. J. Biostat.* **2** 11. MR2306500 https://doi.org/10.2202/1557-4679.1043

[89] VIVIANO, D., LEI, L., IMBENS, G., KARRER, B., SCHRIJVERS, O. and SHI, L. (2023). Causal clustering: Design of cluster experiments under network interference. arXiv preprint. Available at arXiv:2310.14983.

[90] WAINWRIGHT, M. J., JORDAN, M. I. et al. (2008). Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* **1** 1–305.

[91] WANG, T., ZHONG, X. and FAN, Z. (2024). Universality of approximate message passing algorithms and tensor networks. *Ann. Appl. Probab.* **34** 3943–3994. MR4783034 https://doi.org/10.1214/24-aap2056

[92] YAN, J. (2020). Nonlinear large deviations: Beyond the hypercube. *Ann. Appl. Probab.* **30** 812–846. MR4108123 https://doi.org/10.1214/19-AAP1516

[93] YU, C. L., AIROLDI, E. M., BORGS, C. and CHAYES, J. T. (2022). Estimating the total treatment effect in randomized experiments with unknown network structure. *Proc. Natl. Acad. Sci. USA* **119** e2208975119. MR4542829 https://doi.org/10.1073/pnas.2208975119

The Annals of Statistics
2025, Vol. 53, No. 6, 2462–2487
https://doi.org/10.1214/25-AOS2548

THE EMPIRICAL COPULA PROCESS IN HIGH DIMENSIONS: STUTE'S REPRESENTATION AND APPLICATIONS

BY AXEL BÜCHER[1,a] AND CAMBYSE PAKZAD[2,b]

[1]*Fakultät für Mathematik, Ruhr-Universität Bochum,* [a]*axel.buecher@rub.de*
[2]*Université Paris Nanterre, Modal'X,* [b]*cambyse.pakzad@parisnanterre.fr*

The empirical copula process, a fundamental tool for copula inference, is studied in the high dimensional regime where the dimension is allowed to grow to infinity exponentially in the sample size. Under natural, weak smoothness assumptions on the underlying copula, it is shown that Stute's representation is valid in the following sense: all low-dimensional margins of fixed dimension of the empirical copula process can be approximated by a functional of the low-dimensional margins of the standard empirical process, with the almost sure error term being uniform in the margins. The result has numerous potential applications, and is exemplary applied to the problem of testing pairwise stochastic independence in high dimensions, leading to various extensions of recent results in the literature: for certain test statistics based on pairwise association measures, type-I error control is obtained for models beyond mutual independence. Moreover, bootstrap-based critical values are shown to yield strong control of the familywise error rate for a large class of data generating processes.

REFERENCES

[1] ANATOLYEV, S. and PYRLIK, V. (2022). Copula shrinkage and portfolio allocation in ultra-high dimensions. *J. Econom. Dynam. Control* **143** Paper No. 104508, 21. MR4470729 https://doi.org/10.1016/j.jedc.2022.104508

[2] BASTIAN, P., DETTE, H. and HEINY, J. (2024). Testing for practically significant dependencies in high dimensions via bootstrapping maxima of U-statistics. *Ann. Statist.* **52** 628–653. MR4744190 https://doi.org/10.1214/24-aos2361

[3] BEARE, B. K. and SEO, J. (2020). Randomization tests of copula symmetry. *Econometric Theory* **36** 1025–1063. MR4190702 https://doi.org/10.1017/S0266466619000410

[4] BÜCHER, A., GENEST, C., LOCKHART, R. A. and NEŠLEHOVÁ, J. G. (2023). Asymptotic behavior of an intrinsic rank-based estimator of the Pickands dependence function constructed from B-splines. *Extremes* **26** 101–138. MR4544217 https://doi.org/10.1007/s10687-022-00451-9

[5] BÜCHER, A., KOJADINOVIC, I., ROHMER, T. and SEGERS, J. (2014). Detecting changes in cross-sectional dependence in multivariate time series. *J. Multivariate Anal.* **132** 111–128. MR3266263 https://doi.org/10.1016/j.jmva.2014.07.012

[6] BÜCHER, A. and PAKZAD, C. (2024). Testing for independence in high dimensions based on empirical copulas. *Ann. Statist.* **52** 311–334. MR4718417 https://doi.org/10.1214/23-aos2348

[7] BÜCHER, A. and PAKZAD, C. (2025). Supplement to "The empirical copula process in high dimensions: Stute's representation and applications." https://doi.org/10.1214/25-AOS2548SUPP

[8] CHATELAIN, S., FOUGÈRES, A.-L. and NEŠLEHOVÁ, J. G. (2020). Inference for Archimax copulas. *Ann. Statist.* **48** 1025–1051. MR4102686 https://doi.org/10.1214/19-AOS1836

[9] 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

[10] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41** 2786–2819. MR3161448 https://doi.org/10.1214/13-AOS1161

MSC2020 subject classifications. Primary 62G20, 62H05; secondary 62G09.
Key words and phrases. High dimensional statistics, Kendall's tau, multiplier bootstrap, rank-based inference, Spearman's rho, testing independence.

[11] 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

[12] DEHEUVELS, P. (1979). La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance. *Acad. Roy. Belg. Bull. Cl. Sci.* (5) **65** 274–292. MR0573609

[13] DEHEUVELS, P. (1981). An asymptotic decomposition for multivariate distribution-free tests of independence. *J. Multivariate Anal.* **11** 102–113. MR0612295 https://doi.org/10.1016/0047-259X(81)90136-6

[14] DEO, C. M. (1972). Some limit theorems for maxima of absolute values of Gaussian sequences. *Sankhyā Ser. A* **34** 289–292. MR0334319

[15] DRTON, M., HAN, F. and SHI, H. (2020). High-dimensional consistent independence testing with maxima of rank correlations. *Ann. Statist.* **48** 3206–3227. MR4185806 https://doi.org/10.1214/19-AOS1926

[16] ENGELKE, S., MALINOWSKI, A., KABLUCHKO, Z. and SCHLATHER, M. (2015). Estimation of Hüsler-Reiss distributions and Brown-Resnick processes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 239–265. MR3299407 https://doi.org/10.1111/rssb.12074

[17] ENGELKE, S., LALANCETTE, M. and VOLGUSHEV, S. (2024). Learning extremal graphical structures in high dimensions. Available at arXiv:2111.00840.

[18] FERMANIAN, J.-D., RADULOVIĆ, D. and WEGKAMP, M. (2004). Weak convergence of empirical copula processes. *Bernoulli* **10** 847–860. MR2093613 https://doi.org/10.3150/bj/1099579158

[19] GENEST, C., NEŠLEHOVÁ, J. and QUESSY, J.-F. (2012). Tests of symmetry for bivariate copulas. *Ann. Inst. Statist. Math.* **64** 811–834. MR2927772 https://doi.org/10.1007/s10463-011-0337-6

[20] GENEST, C., NEŠLEHOVÁ, J. G., RÉMILLARD, B. and MURPHY, O. A. (2019). Testing for independence in arbitrary distributions. *Biometrika* **106** 47–68. MR3912383 https://doi.org/10.1093/biomet/asy059

[21] GENEST, C., QUESSY, J.-F. and REMILLARD, B. (2007). Asymptotic local efficiency of Cramér-von Mises tests for multivariate independence. *Ann. Statist.* **35** 166–191. MR2332273 https://doi.org/10.1214/009053606000000984

[22] GENEST, C. and RÉMILLARD, B. (2004). Tests of independence and randomness based on the empirical copula process. *TEST* **13** 335–370. MR2154005 https://doi.org/10.1007/BF02595777

[23] GENEST, C., RÉMILLARD, B. and BEAUDOIN, D. (2009). Goodness-of-fit tests for copulas: A review and a power study. *Insurance Math. Econom.* **44** 199–213. MR2517885 https://doi.org/10.1016/j.insmatheco.2007.10.005

[24] GENEST, C. and SEGERS, J. (2009). Rank-based inference for bivariate extreme-value copulas. *Ann. Statist.* **37** 2990–3022. MR2541453 https://doi.org/10.1214/08-AOS672

[25] GHOUDI, K., KULPERGER, R. J. and RÉMILLARD, B. (2001). A nonparametric test of serial independence for time series and residuals. *J. Multivariate Anal.* **79** 191–218. MR1868288 https://doi.org/10.1006/jmva.2000.1967

[26] HAN, F., CHEN, S. and LIU, H. (2017). Distribution-free tests of independence in high dimensions. *Biometrika* **104** 813–828. MR3737306 https://doi.org/10.1093/biomet/asx050

[27] KASA, S. R., BHATTACHARYA, S. and RAJAN, V. (2019). Gaussian mixture copulas for high-dimensional clustering and dependency-based subtyping. *Bioinformatics* **36** 621–628. https://doi.org/10.1093/bioinformatics/btz599

[28] KOJADINOVIC, I. and HOLMES, M. (2009). Tests of independence among continuous random vectors based on Cramér-von Mises functionals of the empirical copula process. *J. Multivariate Anal.* **100** 1137–1154. MR2508377 https://doi.org/10.1016/j.jmva.2008.10.013

[29] KOJADINOVIC, I. and YAN, J. (2011). A goodness-of-fit test for multivariate multiparameter copulas based on multiplier central limit theorems. *Stat. Comput.* **21** 17–30. MR2746600 https://doi.org/10.1007/s11222-009-9142-y

[30] LANCASTER, H. O. (1969). *The Chi-Squared Distribution*. Wiley, New York. MR0253452

[31] LEUNG, D. and DRTON, M. (2018). Testing independence in high dimensions with sums of rank correlations. *Ann. Statist.* **46** 280–307. MR3766953 https://doi.org/10.1214/17-AOS1550

[32] LI, W., WANG, Q. and YAO, J. (2024). Distance correlation test for high-dimensional independence. *Bernoulli* **30** 3165–3192. MR4779430 https://doi.org/10.3150/23-bej1710

[33] LIU, H., HAN, F., YUAN, M., LAFFERTY, J. and WASSERMAN, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.* **40** 2293–2326. MR3059084 https://doi.org/10.1214/12-AOS1037

[34] LIU, H., LAFFERTY, J. and WASSERMAN, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.* **10** 2295–2328. MR2563983

[35] MÜLLER, D. and CZADO, C. (2019). Dependence modelling in ultra high dimensions with vine copulas and the graphical lasso. *Comput. Statist. Data Anal.* **137** 211–232. MR3924530 https://doi.org/10.1016/j.csda.2019.02.007

[36] NAAMAN, M. (2021). On the tight constant in the multivariate Dvoretzky-Kiefer-Wolfowitz inequality. *Statist. Probab. Lett.* **173** Paper No. 109088, 8. MR4233456 https://doi.org/10.1016/j.spl.2021.109088

[37] RÉMILLARD, B. (2017). Goodness-of-fit tests for copulas of multivariate time series. *Econometrics* **5**. https://doi.org/10.3390/econometrics5010013

[38] ROMANO, J. P. and WOLF, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *J. Amer. Statist. Assoc.* **100** 94–108. MR2156821 https://doi.org/10.1198/016214504000000539

[39] RÜSCHENDORF, L. (1976). Asymptotic distributions of multivariate rank order statistics. *Ann. Statist.* **4** 912–923. MR0420794

[40] SAHIN, Ö. and CZADO, C. (2024). High-dimensional sparse vine copula regression with application to genomic prediction. *Biometrics* **80** Paper No. ujad042, 9. MR4880435 https://doi.org/10.1093/biomtc/ujad042

[41] SCHMID, F., SCHMIDT, R., BLUMENTRITT, T., GAISSER, S. and RUPPERT, M. (2010). Copula-based measures of multivariate association. In *Copula Theory and Its Applications*. *Lect. Notes Stat. Proc.* **198** 209–236. Springer, Heidelberg. MR3051270 https://doi.org/10.1007/978-3-642-12465-5_10

[42] SEGERS, J. (2012). Asymptotics of empirical copula processes under non-restrictive smoothness assumptions. *Bernoulli* **18** 764–782. MR2948900 https://doi.org/10.3150/11-BEJ387

[43] SHORACK, G. R. and WELLNER, J. A. (2009). *Empirical Processes with Applications to Statistics*. *Classics in Applied Mathematics* **59**. SIAM, Philadelphia, PA. MR3396731 https://doi.org/10.1137/1.9780898719017.ch1

[44] STUTE, W. (1984). The oscillation behavior of empirical processes: The multivariate case. *Ann. Probab.* **12** 361–379. MR0735843

[45] TSUKAHARA, H. (2005). Semiparametric estimation in copula models. *Canad. J. Statist.* **33** 357–375. MR2193980 https://doi.org/10.1002/cjs.5540330304

[46] 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

[47] YAO, S., ZHANG, X. and SHAO, X. (2018). Testing mutual independence in high dimension via distance covariance. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 455–480. MR3798874 https://doi.org/10.1111/rssb.12259

The Annals of Statistics
2025, Vol. 53, No. 6, 2488–2515
https://doi.org/10.1214/25-AOS2550

SEMI-SUPERVISED U-STATISTICS

BY ILMUN KIM[1,a], LARRY WASSERMAN[2,b], SIVARAMAN BALAKRISHNAN[2,c] AND MATEY NEYKOV[3,d]

[1]*Department of Mathematical Sciences, KAIST,* [a]*ilmunk@kaist.ac.kr*
[2]*Department of Statistics and Data Science, Carnegie Mellon University,* [b]*larry@stat.cmu.edu,* [c]*siva@stat.cmu.edu*
[3]*Department of Statistics and Data Science, Northwestern University,* [d]*mneykov@northwestern.edu*

Semi-supervised datasets are ubiquitous across diverse domains where obtaining fully labeled data is costly or time-consuming. The prevalence of such datasets has consistently driven the demand for new tools and methods that exploit the potential of unlabeled data. Responding to this demand, we introduce semi-supervised U-statistics enhanced by the abundance of unlabeled data, and investigate their statistical properties. We show that the proposed approach is asymptotically Normal and exhibits notable efficiency gains over classical U-statistics by effectively integrating various powerful prediction tools into the framework. To understand the fundamental difficulty of the problem, we derive minimax lower bounds in semi-supervised settings and showcase that our procedure is semi-parametrically efficient under regularity conditions. Moreover, tailored to bivariate kernels, we propose a refined approach that outperforms the classical U-statistic across all degeneracy regimes, and demonstrate its optimality properties. Simulation studies are conducted to corroborate our findings and to further demonstrate our framework.

REFERENCES

ANGELOPOULOS, A. N., BATES, S., FANNJIANG, C., JORDAN, M. I. and ZRNIC, T. (2023). Prediction-powered inference. *Science* **382** 669–674. MR4672910

ARVESEN, J. N. (1969). Jackknifing U-statistics. *Ann. Math. Statist.* **40** 2076–2100. MR0264805 https://doi.org/10.1214/aoms/1177697287

AZRIEL, D., BROWN, L. D., SKLAR, M., BERK, R., BUJA, A. and ZHAO, L. (2022). Semi-supervised linear regression. *J. Amer. Statist. Assoc.* **117** 2238–2251. MR4528501 https://doi.org/10.1080/01621459.2021.1915320

BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–972. MR2216189 https://doi.org/10.1111/j.1541-0420.2005.00377.x

BHATTACHARYYA, A. (1946). On some analogues of the amount of information and their use in statistical estimation. *Sankhyā* **8** 1–14. MR0020242

BIAU, G. and SCORNET, E. (2016). A random forest guided tour. *TEST* **25** 197–227. MR3493512 https://doi.org/10.1007/s11749-016-0481-7

BISHOP, C. M. (1994). Mixture density networks. Technical report, Aston Univ.

BLOM, G. (1976). Some properties of incomplete U-statistics. *Biometrika* **63** 573–580. MR0474582 https://doi.org/10.1093/biomet/63.3.573

BOUSQUET, O. and ELISSEEFF, A. (2002). Stability and generalization. *J. Mach. Learn. Res.* **2** 499–526. MR1929416 https://doi.org/10.1162/153244302760200704

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

CAI, T. T. and GUO, Z. (2020). Semisupervised inference for explained variance in high dimensional linear regression and its applications. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 391–419. MR4084169

CANNINGS, T. I. and FAN, Y. (2022). The correlation-assisted missing data estimator. *J. Mach. Learn. Res.* **23** Paper No. [41], 49. MR4420766

CHAKRABORTTY, A. and CAI, T. (2018). Efficient and adaptive linear regression in semi-supervised settings. *Ann. Statist.* **46** 1541–1572. MR3819109 https://doi.org/10.1214/17-AOS1594

MSC2020 subject classifications. 62G10, 62E20, 62C20.
Key words and phrases. Adaptivity, minimax risk, semi-supervised inference, U-statistics, van Trees inequality.

CHAKRABORTTY, A., DAI, G. and CARROLL, R. J. (2022). Semi-supervised quantile estimation: Robust and efficient inference in high dimensional settings. arXiv preprint, arXiv:2201.10208.

CHAKRABORTTY, A., DAI, G. and TCHETGEN, E. T. (2022). A general framework for treatment effect estimation in semi-supervised and high dimensional settings. arXiv preprint, arXiv:2201.00468.

CHAKRABORTTY, A., LU, J., CAI, T. T. and LI, H. (2019). High dimensional M-estimation with missing outcomes: a semi-parametric framework. arXiv preprint. Available at arXiv:1911.11345.

CHAN, S. F., HEJBLUM, B. P., CHAKRABORTTY, A. and CAI, T. (2020). Semi-supervised estimation of covariance with application to phenome-wide association studies with electronic medical records data. *Stat. Methods Med. Res.* **29** 455–465. MR4876744 https://doi.org/10.1177/0962280219837676

CHAPELLE, O., SCHÖLKOPF, B. and ZIEN, A. (2006). *Semi-Supervised Learning*. MIT Press, Cambridge.

CHEN, L. H. Y., GOLDSTEIN, L. and SHAO, Q.-M. (2011). *Normal Approximation by Stein's Method*. *Probability and Its Applications* (*New York*). Springer, Heidelberg. MR2732624 https://doi.org/10.1007/978-3-642-15007-4

CHEN, Q., SYRGKANIS, V. and AUSTERN, M. (2022). Debiased machine learning without sample-splitting for stable estimators. *Adv. Neural Inf. Process. Syst.* **35** 3096–3109.

CHEN, S., WANG, Y., LIN, C.-T., DING, W. and CAO, Z. (2019). Semi-supervised feature learning for improving writer identification. *Inform. Sci.* **482** 156–170. MR3898730 https://doi.org/10.1016/j.ins.2019.01.024

CHEN, T. and GUESTRIN, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794. ACM, New York.

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

CHERNOZHUKOV, V., NEWEY, W., SINGH, R. and SYRGKANIS, V. (2020). Adversarial estimation of Riesz representers. arXiv preprint, arXiv:2101.00009.

DENG, S., NING, Y., ZHAO, J. and ZHANG, H. (2024). Optimal and safe estimation for high-dimensional semi-supervised learning. *J. Amer. Statist. Assoc.* **119** 2748–2759. MR4833912 https://doi.org/10.1080/01621459.2023.2277409

DICICCIO, C. and ROMANO, J. (2022). CLT for U-statistics with growing dimension. *Statist. Sinica* **32** 323–344. MR4359635 https://doi.org/10.5705/ss.202020.0048

ELISSEEFF, A. (2000). A study about algorithmic stability and their relation to generalization performances. Technical report, Université Lyon 2.

ELISSEEFF, A. and PONTIL, M. (2003). Leave-one-out error and stability of learning algorithms with applications. *NATO Sci. Se. Sub Ser. III Comput. Syst. Sci.* **190** 111–130.

FAN, J., YAO, Q. and TONG, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika* **83** 189–206. MR1399164 https://doi.org/10.1093/biomet/83.1.189

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

GAN, F., LIANG, W. and ZOU, C. (2024). Prediction de-correlated inference: A safe approach for post-prediction inference. *Aust. N. Z. J. Stat.* **66** 417–440. MR4838008 https://doi.org/10.1111/anzs.12429

GILL, R. D. and LEVIT, B. Y. (1995). Applications of the Van Trees inequality: A Bayesian Cramér–Rao bound. *Bernoulli* **1** 59–79. MR1354456 https://doi.org/10.2307/3318681

GOODFELLOW, I., BENGIO, Y. and COURVILLE, A. (2016). *Deep Learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR3617773

GRONSBELL, J. L. and CAI, T. (2018). Semi-supervised approaches to efficient evaluation of model prediction performance. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 579–594. MR3798879 https://doi.org/10.1111/rssb.12264

GYÖRFI, L., KOHLER, M., KRZYŻAK, A. and WALK, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York.

HAN, J., LIANG, X., XU, H., CHEN, K., HONG, L., MAO, J., YE, C., ZHANG, W., LI, Z. et al. (2021). SODA10m: A large-scale 2d self/semi-supervised object detection dataset for autonomous driving. In *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (*Round* 2).

HARDT, M., RECHT, B. and SINGER, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning* 1225–1234. PMLR.

HINTON, G. E., OSINDERO, S. and TEH, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Comput.* **18** 1527–1554. MR2224485 https://doi.org/10.1162/neco.2006.18.7.1527

HIRSHBERG, D. A. and WAGER, S. (2021). Augmented minimax linear estimation. *Ann. Statist.* **49** 3206–3227. MR4352528 https://doi.org/10.1214/21-aos2080

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

IZBICKI, R. and LEE, A. B. (2017). Converting high-dimensional regression to high-dimensional conditional density estimation. *Electron*. *J*. *Stat*. **11** 2800–2831. MR3679910 https://doi.org/10.1214/17-EJS1302

JIAO, R., ZHANG, Y., DING, L. XUE, B. ZHANG, J. CAI, R. JIN, C. (2024). Learning with limited annotations: A survey on deep semi-supervised learning for medical image segmentation. *Comput*. *Biol*. *Med*. **169** 107840.

KALE, S., KUMAR, R. and VASSILVITSKII, S. (2011). Cross-validation and mean-square stability. In *ICS* 487–495.

KENNEDY, E. H. (2022). Semiparametric doubly robust targeted double machine learning: a review. arXiv preprint. Available at arXiv:2203.06469.

KENNEDY, E. H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects. *Electron*. *J*. *Stat*. **17** 3008–3049. MR4667730 https://doi.org/10.1214/23-ejs2157

KIM, I., BALAKRISHNAN, S. and WASSERMAN, L. (2020). Robust multivariate nonparametric tests via projection averaging. *Ann*. *Statist*. **48** 3417–3441. MR4185814 https://doi.org/10.1214/19-AOS1936

KIM, I., BALAKRISHNAN, S. and WASSERMAN, L. (2022). Minimax optimality of permutation tests. *Ann*. *Statist*. **50** 225–251. MR4382015 https://doi.org/10.1214/21-aos2103

KIM, I., WASSERMAN, L., BALAKRISHNAN, S. and NEYKOV, M. (2025). Supplement to “Semi-Supervised U-statistics.” https://doi.org/10.1214/25-AOS2550SUPPA, https://doi.org/10.1214/25-AOS2550SUPPB

KPOTUFE, S. (2011). k-NN regression adapts to local intrinsic dimension. In *Advances in Neural Information Processing Systems* 729–737.

KPOTUFE, S. and GARG, V. (2013). Adaptivity to local smoothness and dimension in kernel regression. In *Advances in Neural Information Processing Systems* 3075–3083.

LEE, A. J. (1990). *U-Statistics*: *Theory and Practice*. *Statistics*: *Textbooks and Monographs* **110**. Dekker, New York. MR1075417

LI, M., NEYKOV, M. and BALAKRISHNAN, S. (2022). Minimax optimal conditional density estimation under total variation smoothness. *Electron*. *J*. *Stat*. **16** 3937–3972. MR4453593 https://doi.org/10.1214/22-ejs2037

LINCHENG, Z. and ZHIJUN, L. (1985). Strong consistency of the kernel estimators of conditional density function. *Acta Math*. *Sin*. *New Ser.* **1** 314–318. MR0867902 https://doi.org/10.1007/BF02564838

LUEDTKE, A. R. and VAN DER LAAN, M. J. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Ann*. *Statist*. **44** 713–742. MR3476615 https://doi.org/10.1214/15-AOS1384

POLYANSKIY, Y. and WU, Y. (2023). *Information Theory*: *From Coding to Learning*. Cambridge Univ. Press, Cambridge.

RIGOLLET, P. and TSYBAKOV, A. B. (2007). Linear and convex aggregation of density estimators. *Math*. *Methods Statist*. **16** 260–280. MR2356821 https://doi.org/10.3103/S1066530707030052

ROBERT, C. P. and CASELLA, G. (2004). *Monte Carlo Statistical Methods*, 2nd ed. *Springer Texts in Statistics*. Springer, New York. MR2080278 https://doi.org/10.1007/978-1-4757-4145-2

ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J*. *Amer. Statist*. *Assoc*. **89** 846–866. MR1294730

ROSENBLATT, M. (1969). Conditional probability density and regression estimators. In *Multivariate Analysis*, *II* (*Proc*. *Second Internat*. *Sympos*., *Dayton*, *Ohio*, 1968) 25–31. Academic Press, New York. MR0254987

ROTNITZKY, A., LEI, Q., SUED, M. and ROBINS, J. M. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika* **99** 439–456. MR2931264 https://doi.org/10.1093/biomet/ass013

SCHMUTZ, H., HUMBERT, O. and MATTEI, P.-A. (2022). Don’t fear the unlabelled: Safe semi-supervised learning via debiasing. In *The Eleventh International Conference on Learning Representations*.

SCHRAB, A., KIM, I., GUEDJ, B. and GRETTON, A. (2022). Efficient aggregated kernel tests using incomplete U-statistics. *Adv*. *Neural Inf. Process*. *Syst*. **35** 18793–18807.

SONG, S., LIN, Y. and ZHOU, Y. (2024). A general M-estimation theory in semi-supervised framework. *J*. *Amer. Statist*. *Assoc*. **119** 1065–1075. MR4765970 https://doi.org/10.1080/01621459.2023.2169699

STEINWART, I. and SCOVEL, C. (2012). Mercer’s theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constr. Approx*. **35** 363–417. MR2914365 https://doi.org/10.1007/s00365-012-9153-3

SUGIYAMA, M., TAKEUCHI, I., SUZUKI, T., KANAMORI, T., HACHIYA, H. and OKANOHARA, D. (2010). Conditional density estimation via least-squares density ratio estimation. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. *JMLR Workshop and Conference Proceedings*. 781–788.

TSIATIS, A. A. (2006). *Semiparametric Theory and Missing Data*. *Springer Series in Statistics*. Springer, New York. MR2233926

TSYBAKOV, A. B. (2003). Optimal rates of aggregation. In *Learning Theory and Kernel Machines*: 16*th Annual Conference on Learning Theory and* 7*th Kernel Workshop*, *COLT/Kernel* 2003. *Proceedings*. 303–313. Springer, Berlin.

TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. MR2724359 https://doi.org/10.1007/b13794

TUKEY, J. W. (1947). Non-parametric estimation. II. Statistically equivalent blocks and tolerance regions–the continuous case. *Ann*. *Math*. *Stat*. **18** 529–539. MR0023033 https://doi.org/10.1214/aoms/1177730343

TUKEY, J. W. (1960). Curves as parameters, and touch estimation. In *Proc*. 4*th Berkeley Sympos*. *Math*. *Statist*. *and Prob*., *Vol*. *I* 681–694. Univ. California Press, Berkeley, CA. MR0132677

VAN ENGELEN, J. E. and HOOS, H. H. (2020). A survey on semi-supervised learning. *Mach*. *Learn*. **109** 373–440. MR4070203 https://doi.org/10.1007/s10994-019-05855-6

VAN TREES, H. L. (1968). *Detection*, *Estimation*, *and Modulation Theory*, *Part I*. Wiley, New York.

VAN DER LAAN, M. J., POLLEY, E. C. and HUBBARD, A. E. (2007). Super learner. *Stat*. *Appl*. *Genet*. *Mol*. *Biol*. **6** Art. 25, 23. MR2349918 https://doi.org/10.2202/1544-6115.1309

VAN DER LAAN, M. J. and RUBIN, D. (2006). Targeted maximum likelihood learning. *Int*. *J*. *Biostat*. **2** Art. 11, 40. MR2306500 https://doi.org/10.2202/1557-4679.1043

VAN DER VAART, A. W. (2000). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

WANG, D., QI, Y., LIN, J., CUI, P., JIA, Q., WANG, Z., FANG, Y., YU, Q., ZHOU, J. et al. (2019). A semi-supervised graph attentive network for financial fraud detection. In 2019 *IEEE International Conference on Data Mining* (*ICDM*) IEEE.

WASSERMAN, L. (2004). *All of Statistics*. *Springer Texts in Statistics*. Springer, New York. A concise course in statistical inference. MR2055670 https://doi.org/10.1007/978-0-387-21736-9

WASSERMAN, L., RAMDAS, A. and BALAKRISHNAN, S. (2020). Universal inference. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **117** 16880–16890. MR4242731 https://doi.org/10.1073/pnas.1922664117

WILLIAMSON, B. D., GILBERT, P. B., SIMON, N. R. and CARONE, M. (2023). A general framework for inference on algorithm-agnostic variable importance. *J*. *Amer*. *Statist*. *Assoc*. **118** 1645–1658. MR4646595 https://doi.org/10.1080/01621459.2021.2003200

ZHANG, A., BROWN, L. D. and CAI, T. T. (2019). Semi-supervised inference: General theory and estimation of means. *Ann*. *Statist*. **47** 2538–2566. MR3988765 https://doi.org/10.1214/18-AOS1756

ZHANG, Y. and BRADIC, J. (2022). High-dimensional semi-supervised learning: In search of optimal inference of the mean. *Biometrika* **109** 387–403. MR4430964 https://doi.org/10.1093/biomet/asab042

ZHANG, Y., CHAKRABORTTY, A. and BRADIC, J. (2023a). Double robust semi-supervised inference for the mean: Selection bias under MAR labeling with decaying overlap. *Inf*. *Inference* **12** Paper No. iaad021, 94. MR4616868 https://doi.org/10.1093/imaiai/iaad021

ZHANG, Y., CHAKRABORTTY, A. and BRADIC, J. (2023b). Semi-Supervised Causal Inference: Generalizable and Double Robust Inference for Average Treatment Effects under Selection Bias with Decaying Overlap. arXiv preprint. Available at arXiv:2305.12789.

ZHENG, W. and VAN DER LAAN, M. J. (2010). *Asymptotic Theory for Cross-Validated Targeted Maximum Likelihood Estimation*. *UC Berkeley Division of Biostatistics Working Paper Series* **273**. 1–58.

ZHU, B., DING, M., JACOBSON, P., WU, M., ZHAN, W., JORDAN, M. I. and JIAO, J. (2023). Doubly-robust self-training. In *Advances in Neural Information Processing Systems* 41413–41431.

ZHU, X. J. (2008). Semi-supervised learning literature survey. Technical Report.

ZRNIC, T. and CANDÈS, E. J. (2024). Cross-prediction-powered inference. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **121** Paper No. e2322083121, 12. MR4740351

The Annals of Statistics
2025, Vol. 53, No. 6, 2516–2535
https://doi.org/10.1214/25-AOS2553

A FLEXIBLE DEFENSE AGAINST THE WINNER'S CURSE

BY TIJANA ZRNIC[1,a] AND WILLIAM FITHIAN[2,b]

[1]*Stanford Data Science & Department of Statistics, Stanford University,* [a]*tijana.zrnic@stanford.edu*
[2]*Department of Statistics, University of California, Berkeley,* [b]*wfithian@berkeley.edu*

Across science and policy, decision-makers often need to draw conclusions about the best candidate among competing alternatives. For instance, researchers may seek to infer the effectiveness of the most successful treatment or determine which demographic group benefits most from a specific treatment. Similarly, in machine learning, practitioners are often interested in the population performance of the model that performs best empirically. However, cherry-picking the best candidate leads to the winner's curse: the observed performance for the winner is biased upwards, rendering conclusions based on standard measures of uncertainty invalid. We introduce the *zoom correction*, a novel approach for valid inference on the winner. Our method is flexible: it can be employed in both parametric and nonparametric settings, can handle arbitrary dependencies between candidates, and automatically adapts to the level of selection bias. The method easily extends to important related problems, such as inference on the top k winners, inference on the value and identity of the population winner, and inference on "near-winners."

REFERENCES

[1] ANDREWS, I., BOWEN, D., KITAGAWA, T. and MCCLOSKEY, A. (2022). Inference for losers. In *AEA Papers and Proceedings* **112** 635–640. American Economic Association.

[2] ANDREWS, I., KITAGAWA, T. and MCCLOSKEY, A. (2024). Inference on winners. *Q. J. Econ.* **139** 305–358.

[3] BENJAMINI, Y., HECHTLINGER, Y. and STARK, P. B. (2019). Confidence intervals for selected parameters. Preprint. Available at arXiv:1906.00505.

[4] BRADLEY, R. A. and TERRY, M. E. (1952). Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika* **39** 324–345. MR0070925 https://doi.org/10.2307/2334029

[5] CHIANG, W.-L., ZHENG, L., SHENG, Y., ANGELOPOULOS, A. N., LI, T., LI, D., ZHU, B., ZHANG, H., JORDAN, M. et al. (2024). Chatbot arena: An open platform for evaluating LLMs by human preference. In *Forty-First International Conference on Machine Learning*.

[6] DANESCU-NICULESCU-MIZIL, C., SUDHOF, M., JURAFSKY, D., LESKOVEC, J. and POTTS, C. (2013). A computational approach to politeness with application to social factors. In 51*st Annual Meeting of the Association for Computational Linguistics* 250–259. ACL.

[7] DENG, J., DONG, W., SOCHER, R., LI, L.-J., LI, K. and FEI-FEI, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 *IEEE Conference on Computer Vision and Pattern Recognition* 248–255. IEEE.

[8] DICKHAUS, T. (2014). *Simultaneous Statistical Inference*: *With Applications in the Life Sciences*. Springer, Heidelberg. MR3184277 https://doi.org/10.1007/978-3-642-45182-9

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

[10] FUENTES, C., CASELLA, G. and WELLS, M. T. (2018). Confidence intervals for the means of the selected populations. *Electron. J. Stat.* **12** 58–79. MR3743737 https://doi.org/10.1214/17-EJS1374

[11] HUNG, K. and FITHIAN, W. (2019). Rank verification for exponential families. *Ann. Statist.* **47** 758–782. MR3909949 https://doi.org/10.1214/17-AOS1634

[12] KIVARANOVIC, D. and LEEB, H. (2021). On the length of post-model-selection confidence intervals conditional on polyhedral constraints. *J. Amer. Statist. Assoc.* **116** 845–857. MR4270029 https://doi.org/10.1080/01621459.2020.1732989

MSC2020 subject classifications. Primary 62J15; secondary 62F25, 62G15.
Key words and phrases. Simultaneous inference, winner's curse, selective inference, selection bias.

[13] LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927. MR3485948 https://doi.org/10.1214/15-AOS1371

[14] LEINER, J., DUAN, B., WASSERMAN, L. and RAMDAS, A. (2025). Data fission: Splitting a single data point. *J. Amer. Statist. Assoc.* **120** 135–146. MR4893533 https://doi.org/10.1080/01621459.2023.2270748

[15] MILLER, R. G. JR. (1981). *Simultaneous Statistical Inference*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR0612319

[16] MOGSTAD, M., ROMANO, J. P., SHAIKH, A. M. and WILHELM, D. (2024). Inference for ranks with applications to mobility across neighbourhoods and academic achievement across countries. *Rev. Econ. Stud.* **91** 476–518.

[17] NEUFELD, A., DHARAMSHI, A., GAO, L. L. and WITTEN, D. (2024). Data thinning for convolution-closed distributions. *J. Mach. Learn. Res.* **25** Paper No. [57], 35. MR4749093

[18] PERRY, R., XU, Z., MCGOUGH, O. and WITTEN, D. (2024). Infer-and-widen versus split-and-condition: Two tales of selective inference. Preprint. Available at arXiv:2408.06323.

[19] RASINES, D. G. and YOUNG, G. A. (2023). Splitting strategies for post-selection inference. *Biometrika* **110** 597–614. MR4627773 https://doi.org/10.1093/biomet/asac070

[20] RASP, S., DUEBEN, P. D., SCHER, S., WEYN, J. A., MOUATADID, S. and THUEREY, N. (2020). WeatherBench: A benchmark data set for data-driven weather forecasting. *J. Adv. Model. Earth Syst.* **12** e2020MS002203.

[21] TIAN, X. and TAYLOR, J. (2018). Selective inference with a randomized response. *Ann. Statist.* **46** 679–710. MR3782381 https://doi.org/10.1214/17-AOS1564

[22] TORCHVISION MAINTAINERS AND CONTRIBUTORS (2016). TorchVision: PyTorch's Computer Vision library. https://github.com/pytorch/vision. GitHub repository.

[23] VENTER, J. (1988). Confidence bounds based on the largest treatment mean. *South Afr. J. Sci.* **84** 340.

[24] ZRNIC, T. and FITHIAN, W. (2024). Locally simultaneous inference. *Ann. Statist.* **52** 1227–1253. MR4784076 https://doi.org/10.1214/24-aos2391

[25] ZRNIC, T. and FITHIAN, W. (2025). Supplement to "A flexible defense against the winner's curse." https://doi.org/10.1214/25-AOS2553SUPP

[26] ZRNIC, T. and FITHIAN, W. (2025). A flexible defense against the winner's curse: Code.

[27] ZRNIC, T. and JORDAN, M. I. (2023). Post-selection inference via algorithmic stability. *Ann. Statist.* **51** 1666–1691. MR4658572 https://doi.org/10.1214/23-aos2303

The Annals of Statistics
2025, Vol. 53, No. 6, 2536–2562
https://doi.org/10.1214/25-AOS2554

CLUSTERING BY HILL-CLIMBING: CONSISTENCY RESULTS

BY ERY ARIAS-CASTRO[1,a] AND WANLI QIAO[2,b]

[1]*Department of Mathematics and Halıcıoğlu Data Science Institute, University of California,* [a]*eariascastro@ucsd.edu*
[2]*Department of Statistics, George Mason University,* [b]*wqiao@gmu.edu*

We consider several hill-climbing approaches to clustering as formulated by Fukunaga and Hostetler (*IEEE Trans. Inf. Theory* **IT-21** (1975) 32–40) in the 1970s. We study both continuous-space and discrete-space (i.e., medoid) variants and establish their consistency.

REFERENCES

[1] ALIYARI GHASSABEH, Y. (2015). A sufficient condition for the convergence of the mean shift algorithm with Gaussian kernel. *J. Multivariate Anal.* **135** 1–10. MR3306422 https://doi.org/10.1016/j.jmva.2014.11.009

[2] ARIAS-CASTRO, E., MASON, D. and PELLETIER, B. (2016). On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *J. Mach. Learn. Res.* **17** Paper No. 43, 28. MR3491137

[3] ARIAS-CASTRO, E. and QIAO, W. (2025). Supplement to "Clustering by hill-climbing: Consistency results." https://doi.org/10.1214/25-AOS2554SUPP

[4] ARIAS-CASTRO, E. and QIAO, W. (2023). A unifying view of modal clustering. *Inf. Inference* **12** 897–920. MR4565754 https://doi.org/10.1093/imaiai/iaac030

[5] BANYAGA, A. and HURTUBISE, D. (2004). *Lectures on Morse Homology*. *Kluwer Texts in the Mathematical Sciences* **29**. Kluwer Academic, Dordrecht. MR2145196 https://doi.org/10.1007/978-1-4020-2696-6

[6] CARREIRA-PERPIÑÁN, M. A. (2000). Mode-finding for mixtures of Gaussian distributions. *IEEE Trans. Pattern Anal. Mach. Intell.* **22** 1318–1323.

[7] CARREIRA-PERPINÁN, M. A. (2006). Fast nonparametric clustering with Gaussian blurring mean-shift. In *International Conference on Machine Learning* 153–160.

[8] CARREIRA-PERPINÁN, M. A. (2008). Generalised blurring mean-shift algorithms for nonparametric clustering. In 2008 *IEEE Conference on Computer Vision and Pattern Recognition* 1–8.

[9] CARREIRA-PERPIÑÁN, M. Á. (2016). Clustering methods based on kernel density estimators: Mean-shift algorithms. In *Handbook of Cluster Analysis* (C. Hennig, M. Meila, F. Murtagh and R. Rocci, eds.). *Chapman & Hall/CRC Handb. Mod. Stat. Methods* 383–417. CRC Press, Boca Raton, FL. MR3644721

[10] CHACÓN, J. E. (2015). A population background for nonparametric density-based clustering. *Statist. Sci.* **30** 518–532. MR3432839 https://doi.org/10.1214/15-STS526

[11] CHACÓN, J. E. and DUONG, T. (2013). Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting. *Electron. J. Stat.* **7** 499–532. MR3035264 https://doi.org/10.1214/13-EJS781

[12] CHAZAL, F., GUIBAS, L. J., OUDOT, S. Y. and SKRABA, P. (2013). Persistence-based clustering in Riemannian manifolds. *J. ACM* **60** Art. 41, 38. MR3144911 https://doi.org/10.1145/2535927

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

[14] CHEN, Y.-C., GENOVESE, C. R. and WASSERMAN, L. (2015). Asymptotic theory for density ridges. *Ann. Statist.* **43** 1896–1928. MR3375871 https://doi.org/10.1214/15-AOS1329

[15] CHEN, Y.-C., GENOVESE, C. R. and WASSERMAN, L. (2017). Statistical inference using the Morse-Smale complex. *Electron. J. Stat.* **11** 1390–1433. MR3635917 https://doi.org/10.1214/17-EJS1271

[16] CHEN, Y.-C., GENOVESE, C. R. and WASSERMAN, L. (2017). Density level sets: Asymptotics, inference, and visualization. *J. Amer. Statist. Assoc.* **112** 1684–1696. MR3750891 https://doi.org/10.1080/01621459.2016.1228536

MSC2020 subject classifications. Primary 62H30; secondary 62G20.
Key words and phrases. Clustering, gradient lines, gradient flow, dynamical systems, ordinary differential equations, Euler scheme, Morse theory, Mean Shift, Max Shift, Max Slope Shift, hill-climbing methods for clustering.

[17] CHENG, M.-Y., HALL, P. and HARTIGAN, J. A. (2004). Estimating gradient trees. In *A Festschrift for Herman Rubin*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **45** 237–249. IMS, Beachwood, OH. MR2126901 https://doi.org/10.1214/lnms/1196285394

[18] CHENG, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE Trans*. *Pattern Anal*. *Mach*. *Intell*. **17** 790–799.

[19] COMANICIU, D. and MEER, P. (1999). Mean shift analysis and applications. In *Proceedings of the Seventh IEEE International Conference on Computer Vision* **2** 1197–1203. IEEE.

[20] COMANICIU, D. and MEER, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Trans*. *Pattern Anal*. *Mach*. *Intell*. **24** 603–619.

[21] CUEVAS, A., GONZÁLEZ-MANTEIGA, W. and RODRÍGUEZ-CASAL, A. (2006). Plug-in estimation of general level sets. *Aust*. *N*. *Z*. *J*. *Stat*. **48** 7–19. MR2234775 https://doi.org/10.1111/j.1467-842X.2006.00421.x

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

[23] EINMAHL, U. and MASON, D. M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *Ann*. *Statist*. **33** 1380–1403. MR2195639 https://doi.org/10.1214/009053605000000129

[24] FUKUNAGA, K. and HOSTETLER, L. D. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Trans*. *Inf*. *Theory* **IT-21** 32–40. MR0388638 https://doi.org/10.1109/tit.1975.1055330

[25] GENOVESE, C. R., PERONE-PACIFICO, M., VERDINELLI, I. and WASSERMAN, L. (2014). Nonparametric ridge estimation. *Ann*. *Statist*. **42** 1511–1545. MR3262459 https://doi.org/10.1214/14-AOS1218

[26] GENOVESE, C. R., PERONE-PACIFICO, M., VERDINELLI, I. and WASSERMAN, L. (2016). Non-parametric inference for density modes. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **78** 99–126. MR3453648 https://doi.org/10.1111/rssb.12111

[27] HARAUX, A. (2012). Some applications of the Łojasiewicz gradient inequality. *Commun*. *Pure Appl*. *Anal*. **11** 2417–2427. MR2912754 https://doi.org/10.3934/cpaa.2012.11.2417

[28] HARTIGAN, J. A. (1975). *Clustering Algorithms*. *Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. MR0405726

[29] HARTIGAN, J. A. and WONG, M. (1979). Algorithm AS 136: A K-means clustering algorithm. *J*. *R*. *Stat*. *Soc*., *Ser*. *C* **28** 100–108.

[30] HIRSCH, M. W., SMALE, S. and DEVANEY, R. L. (2013). *Differential Equations*, *Dynamical Systems*, *and an Introduction to Chaos*, 3rd ed. Elsevier/Academic Press, Amsterdam. MR3293130 https://doi.org/10.1016/B978-0-12-382010-5.00001-4

[31] ISERLES, A. (2009). *A First Course in the Numerical Analysis of Differential Equations*, 2nd ed. *Cambridge Texts in Applied Mathematics*. Cambridge Univ. Press, Cambridge. MR2478556

[32] JIANG, H. (2017). On the consistency of quick shift. In *Advances in Neural Information Processing Systems*.

[33] JIANG, H., JANG, J. and KPOTUFE, S. (2018). Quickshift++: Provably good initializations for sample-based mean shift. In *International Conference on Machine Learning* 2294–2303.

[34] KERNIGHAN, B. W. and LIN, S. (1970). An efficient heuristic procedure for partitioning graphs. *Bell Syst*. *Tech*. *J*. **49** 291–307.

[35] KOONTZ, W. L. G., NARENDRA, P. M. and FUKUNAGA, K. (1976). A graph-theoretic approach to nonparametric cluster analysis. *IEEE Trans*. *Comput*. **C-25** 936–944. MR0489093 https://doi.org/10.1109/tc.1976.1674719

[36] LI, X., HU, Z. and WU, F. (2007). A note on the convergence of the mean shift. *Pattern Recognit*. **40** 1756–1762.

[37] LLOYD, S. P. (1982). Least squares quantization in PCM. *IEEE Trans*. *Inf*. *Theory* **28** 129–137. MR0651807 https://doi.org/10.1109/TIT.1982.1056489

[38] MA, C., WANG, K., CHI, Y. and CHEN, Y. (2020). Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Found*. *Comput*. *Math*. **20** 451–632. MR4099988 https://doi.org/10.1007/s10208-019-09429-9

[39] MASON, D. M. (2012). Proving consistency of non-standard kernel estimators. *Stat*. *Inference Stoch*. *Process*. **15** 151–176. MR2928244 https://doi.org/10.1007/s11203-012-9068-4

[40] MASON, D. M. and SWANEPOEL, J. W. H. (2011). A general result on the uniform in bandwidth consistency of kernel-type function estimators. *TEST* **20** 72–94. MR2806311 https://doi.org/10.1007/s11749-010-0188-0

[41] MILNOR, J. (1963). *Morse Theory*. *Annals of Mathematics Studies*, *No*. 51. Princeton Univ. Press, Princeton, NJ. MR0163331

[42] QIAO, W. (2021). Nonparametric estimation of surface integrals on level sets. *Bernoulli* **27** 155–191. MR4177365 https://doi.org/10.3150/20-BEJ1232

[43] QIAO, W. and POLONIK, W. (2016). Theoretical analysis of nonparametric filament estimation. *Ann. Statist.* **44** 1269–1297. MR3485960 https://doi.org/10.1214/15-AOS1405

[44] QIAO, W. and SHEHU, A. (2022). Space partitioning and regression maxima seeking via a mean-shift-inspired algorithm. *Electron. J. Stat.* **16** 5623–5658. MR4497867 https://doi.org/10.1214/22-ejs2073

[45] RAND, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. Amer. Statist. Assoc.* **66** 846–850.

[46] STÖGER, D. and SOLTANOLKOTABI, M. (2021). Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. *Adv. Neural Inf. Process. Syst.* **34**.

[47] SUN, R. and LUO, Z.-Q. (2016). Guaranteed matrix completion via non-convex factorization. *IEEE Trans. Inf. Theory* **62** 6535–6579. MR3565131 https://doi.org/10.1109/TIT.2016.2598574

[48] VEDALDI, A. and SOATTO, S. (2008). Quick shift and kernel methods for mode seeking. In *European Conference on Computer Vision* 705–718.

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

[50] YAMASAKI, R. and TANAKA, T. (2024). Convergence analysis of mean shift. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

The Annals of Statistics
2025, Vol. 53, No. 6, 2563–2591
https://doi.org/10.1214/25-AOS2555

KURTOSIS-BASED PROJECTION PURSUIT FOR MATRIX-VALUED DATA

BY UNA RADOJIČIĆ[1,a], KLAUS NORDHAUSEN[2,b] AND JONI VIRTA[3,c]

[1]*Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology,* [a]*una.radojicic@tuwien.ac.at*

[2]*Department of Mathematics and Statistics, University of Helsinki,* [b]*klaus.nordhausen@helsinki.fi*

[3]*Department of Mathematics and Statistics, University of Turku,* [c]*joni.virta@utu.fi*

We develop projection pursuit for data that admit a natural representation in matrix form. For projection indices, we propose extensions of the classical kurtosis and Mardia's multivariate kurtosis. The first index estimates projections for both sides of the matrices simultaneously, while the second index finds the two projections separately. Both indices are shown to recover the optimally separating projection for two-group Gaussian mixtures in the absence of any label information. We further establish the strong consistency of the corresponding sample estimators, as well as the asymptotic normality and high-dimensional consistency for the first estimator. Simulations and real data examples on hand-written postal codes and video data are used to demonstrate the method.

REFERENCES

[1] BECKMANN, C. F. and SMITH, S. M. (2005). Tensorial extensions of independent component analysis for multisubject fMRI analysis. *NeuroImage* **25** 294–311.

[2] BERRO, A., LARABI MARIE-SAINTE, S. and RUIZ-GAZEN, A. (2010). Genetic algorithms and particle swarm optimization for exploratory projection pursuit. *Ann. Math. Artif. Intell.* **60** 153–178. MR2817813 https://doi.org/10.1007/s10472-010-9211-0

[3] BICKEL, P. J., KUR, G. and NADLER, B. (2018). Projection pursuit in high dimensions. *Proc. Natl. Acad. Sci. USA* **115** 9151–9156. MR3856112 https://doi.org/10.1073/pnas.1801177115

[4] BOLTON, R. J. and KRZANOWSKI, W. J. (2003). Projection pursuit clustering for exploratory data analysis. *J. Comput. Graph. Statist.* **12** 121–142. MR1965212 https://doi.org/10.1198/1061860031374

[5] CAUSSINUS, H. and RUIZ, A. (1990). Interesting projections of multidimensional data by means of generalized principal component analyses. In *Compstat* (K. Momirović and V. Mildner, eds.) 121–126. Physica-Verlag HD, Heidelberg.

[6] CHEN, Q., ZHONG, D., ZHANG, Y., LIU, Z., PU, Y. and ZHANG, L. (2023). Research on artificial intelligence based on projection pursuit and real coded genetic algorithm. In 2023 4*th International Seminar on Artificial Intelligence*, *Networking and Information Technology* (*AINIT*) 575–578.

[7] DAVIES, P. L. (1987). Asymptotic behaviour of *S*-estimates of multivariate location parameters and dispersion matrices. *Ann. Statist.* **15** 1269–1292. MR0902258 https://doi.org/10.1214/aos/1176350505

[8] DIACONIS, P. and FREEDMAN, D. (1984). Asymptotics of graphical projection pursuit. *Ann. Statist.* **12** 793–815. MR0751274 https://doi.org/10.1214/aos/1176346703

[9] DUTILLEUL, P. (1999). The MLE algorithm for the matrix normal distribution. *J. Stat. Comput. Simul.* **64** 105–123.

[10] ECKART, C. and YOUNG, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika* **1** 211–218.

[11] ESPEZUA, S., VILLANUEVA, E. and MACIEL, C. D. (2014). Towards an efficient genetic algorithm optimizer for sequential projection pursuit. *Neurocomputing* **123** 40–48.

[12] FISCHER, D., BERRO, A., NORDHAUSEN, K. and RUIZ-GAZEN, A. (2021). REPPlab: An R package for detecting clusters and outliers using exploratory projection pursuit. *Comm. Statist. Simulation Comput.* **50** 3397–3419. MR4336345 https://doi.org/10.1080/03610918.2019.1626880

[13] FRIEDMAN, J. H. and TUKEY, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.* **100** 881–890.

MSC2020 subject classifications. 62F12, 62H25.
Key words and phrases. Discriminant analysis, matrix-variate Gaussian mixture, rank-1 projection.

[14] GUPTA, A. K. and NAGAR, D. K. (2000). *Matrix Variate Distributions*. *Chapman & Hall/CRC Monographs and Surveys in Pure and Applied Mathematics* **104**. CRC Press/CRC, Boca Raton, FL. MR1738933

[15] HU, W., SHEN, W., ZHOU, H. and KONG, D. (2020). Matrix linear discriminant analysis. *Technometrics* **62** 196–205. MR4095745 https://doi.org/10.1080/00401706.2019.1610069

[16] HUA, G., VIOLA, P. A. and DRUCKER, S. M. (2007). Face recognition using discriminatively trained orthogonal rank one tensor projections. In 2007 *IEEE Conference on Computer Vision and Pattern Recognition* 1–8. IEEE.

[17] HUBER, P. J. (1985). Projection pursuit. *Ann*. *Statist*. **13** 435–525. MR0790553 https://doi.org/10.1214/aos/1176349519

[18] HUNG, H., WU, P., TU, I. and HUANG, S. (2012). On multilinear principal component analysis of order-two tensors. *Biometrika* **99** 569–583. MR2966770 https://doi.org/10.1093/biomet/ass019

[19] KINGMA, D. P. and BA, J. (2014). Adam: A method for stochastic optimization. *CoRR*. Available at arXiv:1412.6980.

[20] LI, L., HUANG, W., GU, I. Y.-H. and TIAN, Q. (2004). Statistical modeling of complex backgrounds for foreground object detection. *IEEE Trans*. *Image Process*. **13** 1459–1472.

[21] LINDQUIST, M. A. (2008). The statistical analysis of fMRI data. *Statist*. *Sci*. **23** 439–464. MR2530545 https://doi.org/10.1214/09-STS282

[22] LIU, C., HE, K., ZHOU, J-L. and GAO, C.-B. (2011). Discriminant orthogonal rank-one tensor projections for face recognition. In *Asian Conference on Intelligent Information and Database Systems* 203–211. Springer, Berlin.

[23] LOPERFIDO, N. (2015). Vector-valued skewness for model-based clustering. *Statist*. *Probab*. *Lett*. **99** 230–237. MR3321519 https://doi.org/10.1016/j.spl.2015.01.018

[24] LOPUHAÄ, H. P. (1991). Multivariate τ-estimators for location and scatter. *Canad*. *J*. *Statist*. **19** 307–321. MR1144148 https://doi.org/10.2307/3315396

[25] LU, H., PLATANIOTIS, K. N. and VENETSANOPOULOS, A. N. (2011). A survey of multilinear subspace learning for tensor data. *Pattern Recognit*. **44** 1540–1551.

[26] MANNING, C. D., RAGHAVAN, P. and SCHÜTZE, H. (2008). *Introduction to Information Retrieval*. Cambridge Univ. Press, Cambridge.

[27] MARDIA, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika* **57** 519–530. MR0397994 https://doi.org/10.1093/biomet/57.3.519

[28] MARONNA, R. A. (1976). Robust M-estimators of multivariate location and scatter. *Ann*. *Statist*. **4** 51–67. MR0388656

[29] MIETTINEN, J., NORDHAUSEN, K., OJA, H., TASKINEN, S. and VIRTA, J. (2017). The squared symmetric FastICA estimator. *Signal Process*. **131** 402–411.

[30] MOHAMMADI, F. G., AMINI, M. H. and ARABNIA, H. (2020). Evolutionary Computation, Optimization, and Learning Algorithms for Data Science. In *Optimization*, *Learning*, *and Control for Interdependent Complex Networks* 37–65. Springer, Berlin.

[31] MOLSTAD, A. J. (2018). MatrixLDA: Penalized matrix-normal linear discriminant analysis. R package version 0.2.

[32] MURPHY, K. P. (2012). *Machine Learning*: *A Probabilistic Perspective*. *Adaptive Computation and Machine Learning Series*. MIT Press, Cambridge.

[33] NORDHAUSEN, K., OJA, H., TYLER, D. E. and VIRTA, J. (2021). ICtest: Estimating and testing the number of interesting components in linear dimension reduction. R package version 0.3-3.

[34] PEÑA, D. and PRIETO, F. J. (2001). Cluster identification using projections. *J*. *Amer*. *Statist*. *Assoc*. **96** 1433–1445. MR1946588 https://doi.org/10.1198/016214501753382345

[35] PEÑA, D., PRIETO, F. J. and VILADOMAT, J. (2010). Eigenvectors of a kurtosis matrix as interesting directions to reveal cluster structure. *J*. *Multivariate Anal*. **101** 1995–2007. MR2671197 https://doi.org/10.1016/j.jmva.2010.04.014

[36] PFEIFFER, R. M., FORZANI, L. and BURA, E. (2012). Sufficient dimension reduction for longitudinally measured predictors. *Stat*. *Med*. **31** 2414–2427. MR2972256 https://doi.org/10.1002/sim.4437

[37] PIRES, A. and BRANCO, J. (2019). High dimensionality: The latest challenge to data analysis. arXiv preprint. Available at arXiv:1902.04679.

[38] R CORE TEAM (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

[39] RADOJIČIĆ, U., NORDHAUSEN, K. and OJA, H. (2020). Notion of information and independent component analysis. *Appl*. *Math*. **65** 311–330. MR4114255 https://doi.org/10.21136/AM.2020.0326-19

[40] RADOJIČIĆ, U., NORDHAUSEN, K. and VIRTA, J. (2021). Large-sample properties of unsupervised estimation of the linear discriminant using projection pursuit. *Electron*. *J*. *Stat*. **15** 6677–6739. MR4357275 https://doi.org/10.1214/21-ejs1956

[41] RADOJIČIĆ, U., NORDHAUSEN, K. and VIRTA, J. (2025). Supplement to "Kurtosis-based projection pursuit for matrix-valued data." https://doi.org/10.1214/25-AOS2555SUPP

[42] ROUSSEEUW, P. (1985). Multivariate estimation with high breakdown point. In *Mathematical Statistics and Applications*, *Vol*. *B* (*Bad Tatzmannsdorf*, 1983) 283–297. Reidel, Dordrecht. MR0851060

[43] ROUSSEEUW, P. J., RAYMAEKERS, J. and HUBERT, M. (2018). A measure of directional outlyingness with applications to image data and video. *J*. *Comput*. *Graph*. *Statist*. **27** 345–359. MR3816270 https://doi.org/10.1080/10618600.2017.1366912

[44] RUDER, S. (2017). An overview of gradient descent optimization algorithms. arXiv preprint. Available at arXiv:1609.04747.

[45] SALOMON, R. (2004). The curse of high-dimensional search spaces: Observing premature convergence in unimodal functions. In *Proceedings of the* 2004 *Congress on Evolutionary Computation* **1** 918–923.

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

[47] SCRUCCA, L. and SERAFINI, A. (2019). Projection pursuit based on Gaussian mixtures and evolutionary algorithms. *J*. *Comput*. *Graph*. *Statist*. **28** 847–860. MR4045853 https://doi.org/10.1080/10618600.2019.1598871

[48] TYLER, D. E., CRITCHLEY, F., DÜMBGEN, L. and OJA, H. (2009). Invariant co-ordinate selection. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **71** 549–592. MR2749907 https://doi.org/10.1111/j.1467-9868.2009.00706.x

[49] UNGUREANU, R.-M. and BREABAN, M. E. (2020). An evolutionary distributed framework for projection pursuit. *Proc*. *Comput*. *Sci*. **176** 1023–1032.

[50] VIRTA, J., KOESNER, C. L., LI, B., NORDHAUSEN, K., OJA, H. and RADOJICIC, U. (2021). tensorBSS: Blind source separation methods for tensor-valued observations. R package version 0.3.8.

[51] VIRTA, J., LI, B., NORDHAUSEN, K. and OJA, H. (2017). Independent component analysis for tensor-valued data. *J*. *Multivariate Anal*. **162** 172–192. MR3719342 https://doi.org/10.1016/j.jmva.2017.09.008

[52] VIRTA, J., LI, B., NORDHAUSEN, K. and OJA, H. (2018). JADE for tensor-valued observations. *J*. *Comput*. *Graph*. *Statist*. **27** 628–637. MR3863763 https://doi.org/10.1080/10618600.2017.1407324

[53] VIRTA, J., LIETZÉN, N., ILMONEN, P. and NORDHAUSEN, K. (2021). Fast tensorial JADE. *Scand*. *J*. *Stat*. **48** 164–187. MR4233169 https://doi.org/10.1111/sjos.12445

[54] VIRTA, J. and NORDHAUSEN, K. (2017). Blind source separation for nonstationary tensor-valued time series. In 2017 *IEEE* 27*th International Workshop on Machine Learning for Signal Processing* (*MLSP*) 1–6.

[55] VIRTA, J. and NORDHAUSEN, K. (2017). Blind source separation of tensor-valued time series. *Signal Process*. **141** 204–216.

[56] VIRTA, J., NORDHAUSEN, K. and OJA, H. (2016). Projection pursuit for non-Gaussian independent components. arXiv preprint. Available at arXiv:1612.05445.

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

[58] WU, S., LI, W., WEI, Z. and YANG, J. (2011). Local discriminative orthogonal rank-one tensor projection for image feature extraction. In *The First Asian Conference on Pattern Recognition* 367–371. IEEE.

[59] WU, X., LAI, J. and CHEN, X. (2011). Rank-1 tensor projection via regularized regression for action classification. *Int*. *J*. *Wavelets Multiresolut*. *Inf*. *Process*. **9** 1025–1041. MR2875516 https://doi.org/10.1142/S0219691311004420

[60] YE, J. (2005). Generalized low rank approximations of matrices. *Mach*. *Learn*. **61** 167–191.

[61] ZHANG, D. and ZHOU, Z.-H. (2005). $(2D)^2$PCA: Two-directional two-dimensional PCA for efficient face representation and recognition. *Neurocomputing* **69** 224–231.

[62] ZHONG, W., XING, X. and SUSLICK, K. (2015). Tensor sufficient dimension reduction. *Wiley Interdiscip*. *Rev*.: *Comput*. *Stat*. **7** 178–184. MR3349297 https://doi.org/10.1002/wics.1350

The Annals of Statistics
2025, Vol. 53, No. 6, 2592–2616
https://doi.org/10.1214/25-AOS2556

A GEOMETRICAL ANALYSIS OF KERNEL RIDGE REGRESSION AND ITS APPLICATIONS

BY GEORGIOS GAVRILOPOULOS[1,a], GUILLAUME LECUÉ[2,b] AND ZONG SHANG[3,c]

[1]*ETH Zurich, Seminar for Statistics,* [a]*georgios.gavrilopoulos@stat.math.ethz.ch*
[2]*ESSEC, Business School,* [b]*lecue@essec.edu*
[3]*CREST, ENSAE, IPParis,* [c]*zong.shang@ensae.fr*

We obtain upper bounds for the estimation error of Kernel Ridge Regression (KRR) for all nonnegative regularization parameters, offering a geometric perspective on various phenomena in KRR. As applications: 1. We address the Multiple Descents problem, unifying the proofs of (In *Proceedings of Thirty Third Conference on Learning Theory* (2020) 2683–2711 PMLR) and (*Ann. Statist.* **49** (2021) 1029–1054) for polynomial kernels in nonasymptotic regime and we establish Multiple Descents for the generalization error of KRR for polynomial kernel under sub-Gaussian design in asymptotic regimes. 2. In the nonasymptotic regime, we have established a one-sided isomorphic version of the Gaussian Equivalence Conjecture for sub-Gaussian design vectors. 3. We offer a novel perspective on the linearization of kernel matrices of nonlinear kernels, extending it to the power regime for polynomial kernels. 4. Our theory extends the results in (*J. Mach. Learn. Res.* **24** (2023) Paper No. [123]) under weak moment assumption.

Our proof is based on three mathematical tools developed in this paper that can be of independent interest: 1. Dvoretzky–Milman theorem for ellipsoids under (very) weak moment assumptions. 2. Restricted isomorphic property in reproducing kernel Hilbert spaces with embedding index conditions. 3. A concentration inequality for finite-degree polynomial kernel functions.

REFERENCES

[1] AUBIN, B., KRZAKALA, F., LU, Y. and ZDEBOROVÁ, L. (2020). Generalization error in high-dimensional perceptrons: Approaching Bayes error with convex optimization. In *Advances in Neural Information Processing Systems* **33** 12199–12210. Curran Associates, Red Hook.

[2] BA, J., ERDOGDU, M. A., SUZUKI, T., WANG, Z., WU, D. and YANG, G. (2022). High-dimensional asymptotics of feature learning: How one gradient step improves the representation. Available at arXiv:2205.01445 [cs, math, stat].

[3] BACH, F. (2024). *Learning Theory from First Principles*. MIT Press, Cambridge.

[4] BARTLETT, P. L., LONG, P. M., LUGOSI, G. and TSIGLER, A. (2020). Benign overfitting in linear regression. *Proc. Natl. Acad. Sci. USA* **117** 30063–30070. MR4263288 https://doi.org/10.1073/pnas.1907378117

[5] BARTLETT, P. L., MONTANARI, A. and RAKHLIN, A. (2021). Deep learning: A statistical viewpoint. *Acta Numer.* **30** 87–201. MR4295218 https://doi.org/10.1017/S0962492921000027

[6] BARZILAI, D. and SHAMIR, O. (2024). Generalization in kernel regression under realistic assumptions. Available at arXiv:2312.15995 [cs, stat].

[7] BAUER, F., PEREVERZEV, S. and ROSASCO, L. (2007). On regularization algorithms in learning theory. *J. Complexity* **23** 52–72. MR2297015 https://doi.org/10.1016/j.jco.2006.07.001

[8] BELKIN, M. (2021). Fit without fear: Remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numer.* **30** 203–248. MR4298218 https://doi.org/10.1017/S0962492921000039

[9] BELKIN, M., MA, S. and MANDAL, S. (2018). To understand deep learning we need to understand kernel learning. In *Proceedings of the* 35*th International Conference on Machine Learning* 541–549. PMLR. ISSN: 2640-3498.

MSC2020 subject classifications. 62F35.
Key words and phrases. Kernel ridge regression, deep learning models, benign overfitting, Dvoretsky–Milman theorem, restricted isomorphy property.

[10] BIETTI, A. and BACH, F. R. (2021). Deep equals shallow for ReLU networks in kernel regimes. In *9th International Conference on Learning Representations*, *ICLR* 2021, *Virtual Event*, *Austria*, *May* 3–7, 2021 OpenReview.net.

[11] BIETTI, A., BRUNA, J., SANFORD, C. and SONG, M. J. (2022). Learning single-index models with shallow neural networks. Available at arXiv:2210.15651 [cs, math, stat].

[12] BIETTI, A. and MAIRAL, J. (2019). On the inductive bias of neural tangent kernels. In *Advances in Neural Information Processing Systems* **32**. Curran Associates, Red Hook.

[13] BLANCHARD, G., MATHÉ, P. and MÜCKE, N. (2019). Lepskii principle in supervised learning. Available at arXiv:1905.10764 [math, stat].

[14] BLANCHARD, G. and MÜCKE, N. (2016). Kernel regression, minimax rates and effective dimensionality: Beyond the regular case. Available at arXiv:1611.03979 [stat].

[15] BOYER, C. (2022). Living la vida loca: Learning in interpolation regimes.

[16] BUCHHOLZ, S. (2022). Kernel interpolation in Sobolev spaces is not consistent in low dimensions. In *Proceedings of Thirty Fifth Conference on Learning Theory* 3410–3440. PMLR. ISSN: 2640-3498.

[17] CHENG, X. and SINGER, A. (2013). The spectrum of random inner-product kernel matrices. *Random Matrices Theory Appl*. **2** 1350010. MR3149440 https://doi.org/10.1142/S201032631350010X

[18] CHINOT, G., LÖFFLER, M. and VAN DE GEER, S. (2022). On the robustness of minimum norm interpolators and regularized empirical risk minimizers. *Ann*. *Statist*. **50** 2306–2333. MR4474492 https://doi.org/10.1214/22-aos2190

[19] CHIZAT, L., COLOMBO, M., FERNÁNDEZ-REAL, X. and FIGALLI, A. (2024). Infinite-width limit of deep linear neural networks. *Comm*. *Pure Appl*. *Math*. **77** 3958–4007. MR4814914 https://doi.org/10.1002/cpa.22200

[20] CHIZAT, L., OYALLON, E. and BACH, F. (2019). On lazy training in differentiable programming. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems* **264** 2937–2947. Curran Associates, Red Hook, NY, USA.

[21] DAMIAN, A., LEE, J. D. and SOLTANOLKOTABI, M. (2022). Neural networks can learn representations with gradient descent. Available at arXiv:2206.15144 [cs, math, stat].

[22] DANDI, Y., KRZAKALA, F., LOUREIRO, B., PESCE, L. and STEPHAN, L. (2024). How two-layer neural networks learn, one (giant) step at a time. *J*. *Mach*. *Learn*. *Res*. **25** Paper No. [349]. MR4829193

[23] DEVORE, R., HANIN, B. and PETROVA, G. (2021). Neural network approximation. *Acta Numer.* **30** 327–444. MR4298220 https://doi.org/10.1017/S0962492921000052

[24] DHIFALLAH, O. and LU, Y. M. (2020). A precise performance analysis of learning with random features. Available at arXiv:2008.11904 [cs, math].

[25] DO, Y. and VU, V. (2013). The spectrum of random kernel matrices: Universality results for rough and varying kernels. *Random Matrices Theory Appl*. **2** 1350005. MR3109422 https://doi.org/10.1142/S2010326313500056

[26] DONHAUSER, K., RUGGERI, N., STOJANOVIC, S. and YANG, F. (2022). Fast rates for noisy interpolation require rethinking the effect of inductive bias. In *Proceedings of the 39th International Conference on Machine Learning* 5397–5428. PMLR. ISSN: 2640-3498.

[27] DONHAUSER, K., WU, M. and YANG, F. (2021). How rotational invariance of common kernels prevents generalization in high dimensions. In *Proceedings of the 38th International Conference on Machine Learning* 2804–2814. PMLR. ISSN: 2640-3498.

[28] EL KAROUI, N. (2010). The spectrum of kernel random matrices. *Ann*. *Statist*. **38** 1–50. MR2589315 https://doi.org/10.1214/08-AOS648

[29] FAN, Z. and MONTANARI, A. (2019). The spectral norm of random inner-product kernel matrices. *Probab. Theory Related Fields* **173** 27–85. MR3916104 https://doi.org/10.1007/s00440-018-0830-4

[30] FISCHER, S. and STEINWART, I. (2020). Sobolev norm learning rates for regularized least-squares algorithms. *J*. *Mach*. *Learn*. *Res*. **21** Paper No. 205. MR4209491

[31] FOUCART, S. and RAUHUT, H. (2013). *A Mathematical Introduction to Compressive Sensing*. *Applied and Numerical Harmonic Analysis*. Birkhäuser/Springer, New York. MR3100033 https://doi.org/10.1007/978-0-8176-4948-7

[32] GAVRILOPOULOS, G., LECUÉ, G. and SHANG, Z. (2025). Supplement to "A geometrical analysis of kernel ridge regression and its applications." https://doi.org/10.1214/25-AOS2556SUPP

[33] GHORBANI, B., MEI, S., MISIAKIEWICZ, T. and MONTANARI, A. (2021). Linearized two-layers neural networks in high dimension. *Ann*. *Statist*. **49** 1029–1054. MR4255118 https://doi.org/10.1214/20-aos1990

[34] GOLDT, S., MÉZARD, M., KRZAKALA, F. and ZDEBOROVÁ, L. (2020). Modelling the influence of data structure on learning in neural networks: The hidden manifold model. *Phys*. *Rev. X* **10** 041044. Available at arXiv:1909.11500 [cond-mat, stat].

[35] GUÉDON, O., LITVAK, A. E., PAJOR, A. and TOMCZAK-JAEGERMANN, N. (2017). On the interval of fluctuation of the singular values of random matrices. *J. Eur. Math. Soc.* (*JEMS*) **19** 1469–1505. MR3635358 https://doi.org/10.4171/JEMS/697

[36] HAAS, M., HOLZMÜLLER, D., VON LUXBURG, U. and STEINWART, I. (2023). Mind the spikes: Benign overfitting of kernels and neural networks in fixed dimension. Available at arXiv:2305.14077 [cs, math, stat].

[37] HASTIE, T., MONTANARI, A., ROSSET, S. and TIBSHIRANI, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *Ann. Statist.* **50** 949–986. MR4404925 https://doi.org/10.1214/21-aos2133

[38] 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

[39] JACOT, A., GABRIEL, F. and HONGLER, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Proceedings of the* 32*nd International Conference on Neural Information Processing Systems*. *NIPS'*18 8580–8589. Curran Associates, Red Hook, NY, USA.

[40] KINGMA, D. P. and BA, J. (2017). Adam: A method for stochastic optimization. Available at arXiv:1412.6980 [cs].

[41] KOEHLER, F., ZHOU, L., SUTHERLAND, D. J. and SREBRO, N. (2021). Uniform convergence of interpolators: Gaussian width, norm bounds and benign overfitting.

[42] LAX, P. D. (2002). *Funct. Anal.*. Wiley–Blackwell, New York.

[43] LECUÉ, G. and SHANG, Z. (2025). A geometrical viewpoint on the benign overfitting property of the minimum ℓ_2-norm interpolant estimator and its universality. *Probab. Theory Related Fields* **191** 1401–1484. MR4898104 https://doi.org/10.1007/s00440-024-01336-7

[44] LECUÉ, G. and SHANG, Z. (2025). A geometric viewpoint on the benign overfitting property of the minimum norm interpolant estimator in regression and classification. In preparation.

[45] LI, Y., ZHANG, H. and LIN, Q. (2023). On the asymptotic learning curves of kernel ridge regression under power-law decay. Available at arXiv:2309.13337 [cs, math, stat].

[46] LI, Z. Personal communication.

[47] LIANG, T., RAKHLIN, A. and ZHAI, X. (2020). On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Proceedings of Thirty Third Conference on Learning Theory* 2683–2711. PMLR. ISSN: 2640-3498.

[48] LONG, P. M. (2021). Properties of the after kernel. Available at arXiv:2105.10585 [cs].

[49] MCRAE, A. D., KARNIK, S., DAVENPORT, M. and MUTHUKUMAR, V. K. (2022). Harmless interpolation in regression and classification with structured features. In *Proceedings of the* 25*th International Conference on Artificial Intelligence and Statistics* 5853–5875. PMLR ISSN: 2640-3498.

[50] MEI, S., MISIAKIEWICZ, T. and MONTANARI, A. (2022). Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Appl. Comput. Harmon. Anal.* **59** 3–84. MR4412180 https://doi.org/10.1016/j.acha.2021.12.003

[51] MEI, S. and MONTANARI, A. (2022). The generalization error of random features regression: Precise asymptotics and the double descent curve. *Comm. Pure Appl. Math.* **75** 667–766. MR4400901 https://doi.org/10.1002/cpa.22008

[52] MENDELSON, S. and NEEMAN, J. (2010). Regularization in kernel learning. *Ann. Statist.* **38** 526–565. MR2590050 https://doi.org/10.1214/09-AOS728

[53] MILMAN, V. (2023). Asymptotic geometric analysis: Achievements and perspective. In *Geometric Aspects of Functional Analysis* (R. Eldan, B. Klartag, A. Litvak and E. Milman, eds.). *Lecture Notes in Math.* **2327** 1–55. Springer, Cham. MR4651215 https://doi.org/10.1007/978-3-031-26300-2_1

[54] MISIAKIEWICZ, T. (2022). Spectrum of inner-product kernel matrices in the polynomial regime and multiple descent phenomenon in kernel ridge regression. Available at arXiv:2204.10425 [math, stat].

[55] MONIRI, B., LEE, D., HASSANI, H. and DOBRIBAN, E. (2023). A theory of non-linear feature learning with one gradient step in two-layer neural networks. Available at arXiv:2310.07891 [cs, stat].

[56] MOURTADA, J. and ROSASCO, L. (2022). An elementary analysis of ridge regression with random design. *C. R. Math. Acad. Sci. Paris* **360** 1055–1063. MR4491836 https://doi.org/10.5802/crmath.367

[57] PILLAUD-VIVIEN, L. and BACH, F. (2023). Kernelized diffusion maps. In *Proceedings of Thirty Sixth Conference on Learning Theory* 5236–5259. PMLR. ISSN: 2640-3498.

[58] PISIER, G. (1989). *The Volume of Convex Bodies and Banach Space Geometry*. *Cambridge Tracts in Mathematics* **94**. Cambridge Univ. Press, Cambridge. MR1036275 https://doi.org/10.1017/CBO9780511662454

[59] RADHAKRISHNAN, A., BEAGLEHOLE, D., PANDIT, P. and BELKIN, M. (2022). Feature learning in neural networks and kernel machines that recursively learn features. Available at arXiv:2212.13881 [cs].

[60] RAKHLIN, A. and ZHAI, X. (2019). Consistency of interpolation with Laplace kernels is a high-dimensional phenomenon. In *Conference on Learning Theory*, *COLT* 2019, 25-28 *June* 2019, *Phoenix*, *AZ*, *USA* (A. Beygelzimer and D. Hsu, eds.). *Proceedings of Machine Learning Research* **99** 2595–2623. PMLR.

[61] RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR2514435

[62] SAITOH, S. and SAWANO, Y. (2016). *Theory of Reproducing Kernels and Applications*. *Developments in Mathematics* **44**. Springer, Singapore. MR3560890 https://doi.org/10.1007/978-981-10-0530-5

[63] SAXE, A. M., MCCLELLAND, J. L. and GANGULI, S. (2014). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. Available at arXiv:1312.6120 [cond-mat, q-bio, stat].

[64] SCHMIDT-HIEBER, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *Ann*. *Statist*. **48** 1875–1897. MR4134774 https://doi.org/10.1214/19-AOS1875

[65] SEDDIK, M. E. A., LOUART, C., TAMAAZOUSTI, M. and COUILLET, R. (2020). Random matrix theory proves that deep learning representations of GAN-data behave as Gaussian mixtures. Available at arXiv:2001.08370 [cs, stat].

[66] SHAMIR, O. (2022). The implicit bias of benign overfitting. In *Proceedings of Thirty Fifth Conference on Learning Theory* 448–478. PMLR ISSN: 2640-3498.

[67] SHAWE-TAYLOR, J. and CRISTIANINI, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge Univ. Press, Cambridge.

[68] STEINWART, I. and CHRISTMANN, A. (2008). *Support Vector Machines*. *Information Science and Statistics*. Springer, New York. MR2450103

[69] STEINWART, I., HUSH, D. R. and SCOVEL, C. (2009). Optimal rates for regularized least squares regression. In *COLT* 79–93.

[70] SUZUKI, T. (2018). Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: Optimal rate and curse of dimensionality.

[71] TIKHOMIROV, K. (2018). Sample covariance matrices of heavy-tailed distributions. *Int*. *Math*. *Res*. *Not*. *IMRN* **20** 6254–6289. MR3872323 https://doi.org/10.1093/imrn/rnx067

[72] TSIGLER, A. and BARTLETT, P. L. (2023). Benign overfitting in ridge regression. *J*. *Mach*. *Learn*. *Res*. **24** Paper No. [123]. MR4583284

[73] VERSHYNIN, R. (2018). *High-Dimensional Probability*: *An Introduction with Applications in Data Science*, *with a Foreword by Sara van de Geer*. *Cambridge Series in Statistical and Probabilistic Mathematics* **47**. Cambridge Univ. Press, Cambridge. MR3837109 https://doi.org/10.1017/9781108231596

[74] 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

[75] WANG, G., DONHAUSER, K. and YANG, F. (2022). Tight bounds for minimum ℓ_1-norm interpolation of noisy data. In *Proceedings of the* 25*th International Conference on Artificial Intelligence and Statistics* 10572–10602. PMLR. ISSN: 2640-3498.

[76] WOODWORTH, B., GUNASEKAR, S., LEE, J. D., MOROSHKO, E., SAVARESE, P., GOLAN, I., SOUDRY, D. and SREBRO, N. (2020). Kernel and rich regimes in overparametrized models. In *Proceedings of Thirty Third Conference on Learning Theory* 3635–3673. PMLR. ISSN: 2640-3498.

[77] YANG, G. (2020). Tensor programs II: Neural tangent kernel for any architecture. *CoRR*. Available at arXiv:2006.14548.

[78] YAROTSKY, D. (2017). Error bounds for approximations with deep ReLU networks. *Neural Netw*. **94** 103–114.

[79] ZHANG, C., BENGIO, S., HARDT, M., RECHT, B. and VINYALS, O. (2017). Understanding deep learning requires rethinking generalization. In 5*th International Conference on Learning Representations*, *ICLR* 2017, *Toulon*, *France*, *April* 24-26, 2017, *Conference Track Proceedings* OpenReview.net.

[80] ZHANG, H., LI, Y. and LIN, Q. (2024). On the optimality of misspecified spectral algorithms. *J*. *Mach*. *Learn*. *Res*. **25** Paper No. [188]. MR4777430

The Annals of Statistics
2025, Vol. 53, No. 6, 2617–2641
https://doi.org/10.1214/25-AOS2557

A TWO-WAY HETEROGENEITY MODEL FOR DYNAMIC NETWORKS

BY BINYAN JIANG[1,a], CHENLEI LENG[2,b], TING YAN[3,c], QIWEI YAO[4,d] AND XINYANG YU[4,e]

[1]*Department of Data Science and AI, Hong Kong Polytechnic University,* [a]*by.jiang@polyu.edu.hk*

[2]*Department of Applied Mathematics, Hong Kong Polytechnic University,* [b]*chenlei.leng@polyu.edu.hk*

[3]*Department of Statistics, Hubei Key Laboratory of Mathematical Sciences, Key Laboratory of Nonlinear Analysis & Applications, Central China Normal University,* [c]*tingyanty@mail.ccnu.edu.cn*

[4]*Department of Statistics, London School of Economics,* [d]*q.yao@lse.ac.uk,* [e]*X.Yu59@lse.ac.uk*

Analysis of networks that evolve dynamically requires the joint modelling of individual snapshots and time dynamics. This paper proposes a new flexible two-way heterogeneity model towards this goal. The new model equips each node of the network with two heterogeneity parameters, one to characterize the propensity to form ties with other nodes statically and the other to differentiate the tendency to retain existing ties over time. With n observed networks each having p nodes, we develop a new asymptotic theory for the maximum likelihood estimation of $2p$ parameters when $np \to \infty$. We overcome the global nonconvexity of the negative log-likelihood function by the virtue of its local convexity, and propose a novel method of moment estimator as the initial value for a simple algorithm that leads to the consistent local maximum likelihood estimator (MLE). To establish the upper bounds for the estimation error of the MLE, we derive a new uniform deviation bound, which is of independent interest. The theory of the model and its usefulness are further supported by extensive simulation and the analysis of some real network data sets.

REFERENCES

[1] BHATTACHARJEE, M., BANERJEE, M. and MICHAILIDIS, G. (2020). Change point estimation in a dynamic stochastic block model. *J. Mach. Learn. Res.* **21** Paper No. 107, 59. MR4119175

[2] CHATTERJEE, S., DIACONIS, P. and SLY, A. (2011). Random graphs with a given degree sequence. *Ann. Appl. Probab.* **21** 1400–1435. MR2857452 https://doi.org/10.1214/10-AAP728

[3] CHEN, M., KATO, K. and LENG, C. (2021). Analysis of networks via the sparse β-model. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **83** 887–910. MR4349121 https://doi.org/10.1111/rssb.12444

[4] DURANTE, D. and DUNSON, D. B. (2016). Locally adaptive dynamic networks. *Ann. Appl. Stat.* **10** 2203–2232. MR3592054 https://doi.org/10.1214/16-AOAS971

[5] GRAGG, W. B. and TAPIA, R. A. (1974). Optimal error bounds for the Newton–Kantorovich theorem. *SIAM J. Numer. Anal.* **11** 10–13. MR0343594 https://doi.org/10.1137/0711002

[6] GRAHAM, B. S. (2017). An econometric model of network formation with degree heterogeneity. *Econometrica* **85** 1033–1063. MR3681763 https://doi.org/10.3982/ECTA12679

[7] HAN, R., CHEN, K. and TAN, C. (2020). Bivariate gamma model. *J. Multivariate Anal.* **180** 104666, 12. MR4137711 https://doi.org/10.1016/j.jmva.2020.104666

[8] HAN, R., XU, Y. and CHEN, K. (2023). A general pairwise comparison model for extremely sparse networks. *J. Amer. Statist. Assoc.* **118** 2422–2432. MR4681593 https://doi.org/10.1080/01621459.2022.2053137

[9] HANNEKE, S., FU, W. and XING, E. P. (2010). Discrete temporal models of social networks. *Electron. J. Stat.* **4** 585–605. MR2660534 https://doi.org/10.1214/09-EJS548

[10] HANNEKE, S. and XING, E. P. (2007). Discrete temporal models of social networks. In *Statistical Network Analysis: Models, Issues, and New Directions: ICML* 2006 *Workshop on Statistical Network Analysis, Pittsburgh, PA, USA, June* 29, 2006, *Revised Selected Papers* 115–125. Springer, Berlin.

[11] HILLAR, C. J., LIN, S. and WIBISONO, A. (2012). Inverses of symmetric, diagonally dominant positive matrices and applications. arXiv preprint. Available at arXiv:1203.6812.

MSC2020 subject classifications. 62F12, 62J99, 62R99.

Key words and phrases. Degree heterogeneity, dynamic networks, maximum likelihood estimation, uniform deviation bound.

[12] HOLLAND, P. W. and LEINHARDT, S. (1981). An exponential family of probability distributions for directed graphs. *J. Amer. Statist. Assoc.* **76** 33–50. MR0608176

[13] JIANG, B., LENG, C., YAN, T., YAO, Q. and YU, X. (2025). Supplement to "A two-way heterogeneity model for dynamic networks." https://doi.org/10.1214/25-AOS2557SUPP

[14] JIANG, B., LI, J. and YAO, Q. (2023). Autoregressive networks. *J. Mach. Learn. Res.* **24** Paper No. [227], 69. MR4633954

[15] JIN, J. (2015). Fast community detection by SCORE. *Ann. Statist.* **43** 57–89. MR3285600 https://doi.org/10.1214/14-AOS1265

[16] JIN, J., KE, Z. T., LUO, S. and WANG, M. (2023). Optimal estimation of the number of network communities. *J. Amer. Statist. Assoc.* **118** 2101–2116. MR4646629 https://doi.org/10.1080/01621459.2022.2035736

[17] KARRER, B. and NEWMAN, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **83** 016107, 10. MR2788206 https://doi.org/10.1103/PhysRevE.83.016107

[18] KARWA, V. and SLAVKOVIĆ, A. (2016). Inference using noisy degrees: Differentially private β-model and synthetic graphs. *Ann. Statist.* **44** 87–112. MR3449763 https://doi.org/10.1214/15-AOS1358

[19] KE, Z. T. and JIN, J. (2023). The SCORE normalization, especially for highly heterogeneous network and text data. arXiv preprint. Available at arXiv:2204.11097.

[20] KOLACZYK, E. D. and CSÁRDI, G. (2020). *Statistical Analysis of Network Data with R*, 2nd ed. *Use R!* Springer, Cham. MR4174392 https://doi.org/10.1007/978-3-030-44129-6_1

[21] KRIVITSKY, P. N. and HANDCOCK, M. S. (2014). A separable model for dynamic networks. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 29–46. MR3153932 https://doi.org/10.1111/rssb.12014

[22] MATIAS, C. and MIELE, V. (2017). Statistical clustering of temporal networks through a dynamic stochastic block model. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1119–1141. MR3689311 https://doi.org/10.1111/rssb.12200

[23] MERSCH, D. P., CRESPI, A. and KELLER, L. (2013). Tracking individuals shows spatial fidelity is a key regulator of ant social organization. *Science* **340** 1090–1093.

[24] MINAI, A. A. and WILLIAMS, R. D. (1993). On the derivatives of the sigmoid. *Neural Netw.* **6** 845–853.

[25] NEWMAN, M. (2018). *Networks*, 2nd ed. Oxford Univ. Press, Oxford. MR3838417 https://doi.org/10.1093/oso/9780198805090.001.0001

[26] PENSKY, M. (2019). Dynamic network models and graphon estimation. *Ann. Statist.* **47** 2378–2403. MR3953455 https://doi.org/10.1214/18-AOS1751

[27] SENGUPTA, S. and CHEN, Y. (2018). A block model for node popularity in networks with community structure. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 365–386. MR3763696 https://doi.org/10.1111/rssb.12245

[28] SHAO, M., ZHANG, Y., WANG, Q., ZHANG, Y., LUO, J. and YAN, T. (2021). L-2 Regularized maximum likelihood for β-model in large and sparse networks. arXiv preprint. Available at arXiv:2110.11856.

[29] STEIN, S., FENG, R. and LENG, C. (2024). A sparse beta regression model for network analysis. *J. Amer. Statist. Assoc.* **120** 1281–1293. MR4917295 https://doi.org/10.1080/01621459.2024.2411073

[30] VAN DER VAART, A. W. (2000). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

[31] VAN DER VAART, A. W. and WELLNER, J. A. (1996). Weak convergence. In *Weak Convergence and Empirical Processes*: *With Applications to Statistics Springer Series in Statistics* 16–28 Springer, New York. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2

[32] 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

[33] YAN, T., JIANG, B., FIENBERG, S. E. and LENG, C. (2019). Statistical inference in a directed network model with covariates. *J. Amer. Statist. Assoc.* **114** 857–868. MR3963186 https://doi.org/10.1080/01621459.2018.1448829

[34] YAN, T., LENG, C. and ZHU, J. (2015). Supplement to "Asymptotics in directed exponential random graph models with an increasing bi-degree sequence". *Ann. Statist.*

[35] YAN, T., LENG, C. and ZHU, J. (2016). Asymptotics in directed exponential random graph models with an increasing bi-degree sequence. *Ann. Statist.* **44** 31–57. MR3449761 https://doi.org/10.1214/15-AOS1343

[36] YAN, T. and XU, J. (2012). Approximating the inverse of a balanced symmetric matrix with positive elements. arXiv preprint. Available at arXiv:1202.1058.

[37] YAN, T. and XU, J. (2013). A central limit theorem in the β-model for undirected random graphs with a diverging number of vertices. *Biometrika* **100** 519–524. MR3068452 https://doi.org/10.1093/biomet/ass084

The Annals of Statistics
2025, Vol. 53, No. 6, 2642–2672
https://doi.org/10.1214/25-AOS2558

OPTIMAL SEQUENCING DEPTH FOR SINGLE-CELL RNA-SEQUENCING IN WASSERSTEIN SPACE

BY JAKWANG KIM[1,a], SHARVAJ KUBAL[2,b] AND GEOFFREY SCHIEBINGER[2,c]

[1]*School of Data Science, The Chinese University of Hong Kong,* [a]*jakwangkim@cuhk.edu.cn*
[2]*Department of Mathematics, University of British Columbia,* [b]*sharvaj@math.ubc.ca,* [c]*geoff@math.ubc.ca*

How many samples should one collect for an empirical distribution to be as close as possible to the true population? This question is not trivial in the context of single-cell RNA-sequencing. With limited sequencing depth, profiling more cells comes at the cost of fewer reads per cell. Therefore, one must strike a balance between the number of cells sampled and the accuracy of each measured gene expression profile. In this paper, we analyze an empirical distribution of cells and obtain upper and lower bounds on the Wasserstein distance to the true population. Our analysis holds for general, nonparametric distributions of cells, and is validated by simulation experiments on a real single-cell dataset.

REFERENCES

[1] AHLMANN-ELTZE, C. and HUBER, W. (2023). Comparison of transformations for single-cell RNA-seq data. *Nat. Methods* **20** 665–672. https://doi.org/10.1038/s41592-023-01814-1
[2] BACHER, R. and KENDZIORSKI, C. (2016). Design and computational analysis of single-cell RNA-sequencing experiments. *Genome Biol.* **17** 1–14. https://doi.org/10.1186/s13059-016-0927-y
[3] BEIGLBÖCK, M. and JUILLET, N. (2016). On a problem of optimal transport under marginal martingale constraints. *Ann. Probab.* **44** 42–106. MR3456332 https://doi.org/10.1214/14-AOP966
[4] BENAMOU, J.-D. and BRENIER, Y. (2000). A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numer. Math.* **84** 375–393. MR1738163 https://doi.org/10.1007/s002110050002
[5] BLOCK, A., JIA, Z., POLYANSKIY, Y. and RAKHLIN, A. (2022). Intrinsic dimension estimation using Wasserstein distance. *J. Mach. Learn. Res.* **23** Paper No. [313]. MR4577752
[6] BOISSARD, E. (2011). Simple bounds for convergence of empirical and occupation measures in 1-Wasserstein distance. *Electron. J. Probab.* **16** no. 83, 2296–2333. MR2861675 https://doi.org/10.1214/EJP.v16-958
[7] BOISSARD, E. and LE GOUIC, T. (2014). On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. *Ann. Inst. Henri Poincaré Probab. Stat.* **50** 539–563. MR3189084 https://doi.org/10.1214/12-AIHP517
[8] BOLLEY, F., GUILLIN, A. and VILLANI, C. (2007). Quantitative concentration inequalities for empirical measures on non-compact spaces. *Probab. Theory Related Fields* **137** 541–593. MR2280433 https://doi.org/10.1007/s00440-006-0004-7
[9] BRENIER, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.* **44** 375–417. MR1100809 https://doi.org/10.1002/cpa.3160440402
[10] CAFFARELLI, L. and FIGALLI, A. (2013). Regularity of solutions to the parabolic fractional obstacle problem. *J. Reine Angew. Math.* **680** 191–233. MR3100955 https://doi.org/10.1515/crelle.2012.036
[11] CAFFARELLI, L. A. (2000). Monotonicity properties of optimal transportation and the FKG and related inequalities. *Comm. Math. Phys.* **214** 547–563. MR1800860 https://doi.org/10.1007/s002200000257
[12] CARRILLO, J. A., DIFRANCESCO, M., FIGALLI, A., LAURENT, T. and SLEPČEV, D. (2011). Global-in-time weak measure solutions and finite-time aggregation for nonlocal interaction equations. *Duke Math. J.* **156** 229–271. MR2769217 https://doi.org/10.1215/00127094-2010-211
[13] CHIZAT, L. and BACH, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport. *Adv. Neural Inf. Process. Syst.* **31**.

MSC2020 subject classifications. Primary 62G05, 49Q22; secondary 62D99.
Key words and phrases. RNA sequencing, optimal transport, Wasserstein distance, optimal depth, empirical distribution, nonparametric inference.

[14] CUTURI, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems* (C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger, eds.) **26**. Curran Associates, Red Hook.

[15] DAL MOLIN, A. and DI CAMILLO, B. (2019). How to design a single-cell RNA-sequencing experiment: Pitfalls, challenges and perspectives. *Brief. Bioinform.* **20** 1384–1394.

[16] DE BORTOLI, V., THORNTON, J., HENG, J. and DOUCET, A. (2021). Diffusion Schrödinger bridge with applications to score-based generative modeling. *Adv. Neural Inf. Process. Syst.* **34** 17695–17709.

[17] DIAO, M. Z., BALASUBRAMANIAN, K., CHEWI, S. and SALIM, A. (2023). Forward-backward Gaussian variational inference via JKO in the Bures-Wasserstein space. In *International Conference on Machine Learning* 7960–7991. PMLR.

[18] DUDLEY, R. M. (1968). The speed of mean Glivenko-Cantelli convergence. *Ann. Math. Statist.* **40** 40–50. MR0236977 https://doi.org/10.1214/aoms/1177697802

[19] ECKER, J. R., GESCHWIND, D. H., KRIEGSTEIN, A. R., NGAI, J., OSTEN, P., POLIOUDAKIS, D., REGEV, A., SESTAN, N., WICKERSHAM, I. R. et al. (2017). The BRAIN initiative cell census consortium: Lessons learned toward generating a comprehensive brain cell atlas. *Neuron* **96** 542–557.

[20] EFRON, B. (2016). Empirical Bayes deconvolution estimates. *Biometrika* **103** 1–20. MR3465818 https://doi.org/10.1093/biomet/asv068

[21] FIGALLI, A., FUSCO, N., MAGGI, F., MILLOT, V. and MORINI, M. (2015). Isoperimetry and stability properties of balls with respect to nonlocal energies. *Comm. Math. Phys.* **336** 441–507. MR3322379 https://doi.org/10.1007/s00220-014-2244-1

[22] FIGALLI, A., MAGGI, F. and PRATELLI, A. (2010). A mass transportation approach to quantitative isoperimetric inequalities. *Invent. Math.* **182** 167–211. MR2672283 https://doi.org/10.1007/s00222-010-0261-z

[23] FLAMARY, R., COURTY, N., GRAMFORT, A., ALAYA, M. Z., BOISBUNON, A., CHAMBON, S., CHAPEL, L., CORENFLOS, A., FATRAS, K. et al. (2021). POT: Python optimal transport. *J. Mach. Learn. Res.* **22** 1–8.

[24] FOURNIER, N. and GUILLIN, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Related Fields* **162** 707–738. MR3383341 https://doi.org/10.1007/s00440-014-0583-7

[25] GARCÍA TRILLOS, N., KIM, J. and JACOBS, M. (2023). The multimarginal optimal transport formulation of adversarial multiclass classification. *J. Mach. Learn. Res.* **24** Paper No. [45]. MR4582467

[26] GARCÍA TRILLOS, N. and MURRAY, R. (2022). Adversarial classification: Necessary conditions and geometric flows. *J. Mach. Learn. Res.* **23** Paper No. [187]. MR4577140

[27] GOZLAN, N., ROBERTO, C., SAMSON, P.-M. and TETALI, P. (2017). Kantorovich duality for general transport costs and applications. *J. Funct. Anal.* **273** 3327–3405. MR3706606 https://doi.org/10.1016/j.jfa.2017.08.015

[28] GRABSKI, I. N., STREET, K. and IRIZARRY, R. A. (2023). Significance analysis for clustering with single-cell RNA-sequencing data. *Nat. Methods* **20** 1196–1202.

[29] HAFEMEISTER, C. and SATIJA, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20** 296.

[30] HAMDOUCHE, M., HENRY-LABORDERE, P. and PHAM, H. (2023). Generative modeling for time series via Schrödinger bridge. arXiv Preprint. Available at arXiv:2304.05093.

[31] HAQUE, A., ENGEL, J., TEICHMANN, S. A. and LÖNNBERG, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Gen. Med.* **9** 1–12.

[32] HEIMBERG, G., BHATNAGAR, R., EL-SAMAD, H. and THOMSON, M. (2016). Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing. *Cell Syst.* **2** 239–250.

[33] HOBSON, D. and NEUBERGER, A. (2012). Robust bounds for forward start options. *Math. Finance* **22** 31–56. MR2881879 https://doi.org/10.1111/j.1467-9965.2010.00473.x

[34] HUANG, M., WANG, J., TORRE, E., DUECK, H., SHAFFER, S., BONASIO, R., MURRAY, J. I., RAJ, A., LI, M. et al. (2018). SAVER: Gene expression recovery for single-cell RNA sequencing. *Nat. Methods* **15** 539–542.

[35] HÜTTER, J.-C. and RIGOLLET, P. (2021). Minimax estimation of smooth optimal transport maps. *Ann. Statist.* **49** 1166–1194. MR4255123 https://doi.org/10.1214/20-aos1997

[36] JAITIN, D. A., KENIGSBERG, E., KEREN-SHAUL, H., ELEFANT, N., PAUL, F., ZARETSKY, I., MILDNER, A., COHEN, N., JUNG, S. et al. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343** 776–779.

[37] KANTOROVICH, L. V. (1948). On a problem of Monge. In *CR* (*Doklady*) *Acad. Sci. URSS* (*NS*) **3** 225–226.

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

[39] KIM, J., KIM, Y.-H., RUAN, Y. and WARREN, A. (2024). Statistical inference of convex order by Wasserstein projection.

[40] KIM, Y.-H. and MCCANN, R. J. (2010). Continuity, curvature, and the general covariance of optimal transportation. *J. Eur. Math. Soc.* (*JEMS*) **12** 1009–1040. MR2654086 https://doi.org/10.4171/JEMS/221

[41] KIVIOJA, T., VÄHÄRAUTIO, A., KARLSSON, K., BONKE, M., ENGE, M., LINNARSSON, S. and TAIPALE, J. (2012). Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9** 72–74.

[42] KLEIN, A., MAZUTIS, L., AKARTUNA, I., TALLAPRAGADA, N., VERES, A., LI, V., PESHKIN, L., WEITZ, D. and KIRSCHNER, M. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161** 1187–1201.

[43] LAMBERT, M., CHEWI, S., BACH, F., BONNABEL, S. and RIGOLLET, P. (2022). Variational inference via Wasserstein gradient flows. *Adv. Neural Inf. Process. Syst.* **35** 14434–14447.

[44] LAVENANT, H., ZHANG, S., KIM, Y.-H. and SCHIEBINGER, G. (2024). Toward a mathematical theory of trajectory inference. *Ann. Appl. Probab.* **34** 428–500. MR4696282 https://doi.org/10.1214/23-aap1969

[45] LEDOUX, M. and ZHU, J.-X. (2021). On optimal matching of Gaussian samples III. *Probab. Math. Statist.* **41** 237–265. MR4359822 https://doi.org/10.37190/0208-4147.41.2.3

[46] LOTT, J. and VILLANI, C. (2009). Ricci curvature for metric-measure spaces via optimal transport. *Ann. of Math.* (2) **169** 903–991. MR2480619 https://doi.org/10.4007/annals.2009.169.903

[47] MACOSKO, E. Z., BASU, A., SATIJA, R., NEMESH, J., SHEKHAR, K., GOLDMAN, M., TIROSH, I., BIALAS, A. R., KAMITAKI, N. et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161** 1202–1214.

[48] MASOERO, L., CAMERLENGHI, F., FAVARO, S. and BRODERICK, T. (2022). More for less: Predicting and maximizing genomic variant discovery via Bayesian nonparametrics. *Biometrika* **109** 17–32. MR4374638 https://doi.org/10.1093/biomet/asab012

[49] MASSRI, A. J., BERRIO, A., AFANASSIEV, A., GREENSTREET, L., PIPHO, K., BYRNE, M., SCHIEBINGER, G., MCCLAY, D. R. and WRAY, G. A. (2025). Single-cell transcriptomics reveals evolutionary reconfiguration of embryonic cell fate specification in the sea urchin heliocidaris erythrogramma. *Gen. Biol. Evol.* **17** evae258.

[50] MCCANN, R. J. (1994). *A Convexity Theory for Interacting Gases and Equilibrium Crystals*. Princeton University.

[51] MCCANN, R. J. (1997). A convexity principle for interacting gases. *Adv. Math.* **128** 153–179. MR1451422 https://doi.org/10.1006/aima.1997.1634

[52] MEI, S., MONTANARI, A. and NGUYEN, P.-M. (2018). A mean field view of the landscape of two-layer neural networks. *Proc. Natl. Acad. Sci. USA* **115** E7665–E7671. MR3845070 https://doi.org/10.1073/pnas.1806579115

[53] MÉRIGOT, Q., SANTAMBROGIO, F. and SARRAZIN, C. (2021). Non-asymptotic convergence bounds for Wasserstein approximation using point clouds. *Adv. Neural Inf. Process. Syst.* **34** 12810–12821.

[54] MONGE, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.* 666–704.

[55] NILES-WEED, J. and BACH, F. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli* **25** 2620–2648. MR4003560 https://doi.org/10.3150/18-BEJ1065

[56] NILES-WEED, J. and BERTHET, Q. (2022). Minimax estimation of smooth densities in Wasserstein distance. *Ann. Statist.* **50** 1519–1540. MR4441130 https://doi.org/10.1214/21-aos2161

[57] OTTO, F. (2001). The geometry of dissipative evolution equations: The porous medium equation. *Comm. Partial Differential Equations* **26** 101–174. MR1842429 https://doi.org/10.1081/PDE-100002243

[58] OTTO, F. and VILLANI, C. (2000). Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *J. Funct. Anal.* **173** 361–400. MR1760620 https://doi.org/10.1006/jfan.1999.3557

[59] PEYRÉ, G. and CUTURI, M. (2019). Computational optimal transport: With applications to data science. *Found. Trends Mach. Learn.* **11** 355–607.

[60] POLLEN, A. A., NOWAKOWSKI, T. J., SHUGA, J., WANG, X., LEYRAT, A. A., LUI, J. H., LI, N., SZPANKOWSKI, L., FOWLER, B. et al. (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* **32** 1053–1058.

[61] PYDI, M. S. and JOG, V. (2021). The many faces of adversarial risk. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

[62] QIAN, J., FRUIT, R., PIROTTA, M. and LAZARIC, A. (2020). Concentration inequalities for multinoulli random variables. arXiv Preprint. Available at arXiv:2001.11595.

[63] QIU, C., CAO, J., MARTIN, B. K., LI, T., WELSH, I. C., SRIVATSAN, S., HUANG, X., CALDERON, D., NOBLE, W. S. et al. (2022). Systematic reconstruction of cellular trajectories across mouse embryogenesis. *Nat. Genet.* **54** 328–341.

[64] ROCH, S. (2024). *Modern Discrete Probability: An Essential Toolkit*. Cambridge Univ. Press, Cambridge.

[65] ROTSKOFF, G. M. and VANDEN-EIJNDEN, E. (2022). Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Comm. Pure Appl. Math.* **75** 1889–1935. MR4465905 https://doi.org/10.1002/cpa.22074

[66] SANTAMBROGIO, F. (2015). *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. *Progress in Nonlinear Differential Equations and Their Applications* **87**. Birkhäuser/Springer, Cham. MR3409718 https://doi.org/10.1007/978-3-319-20828-2

[67] SCHIEBINGER, G., SHU, J., TABAKA, M., CLEARY, B., SUBRAMANIAN, V., SOLOMON, A., GOULD, J., LIU, S., LIN, S. et al. (2019). Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* **176** 928–943.

[68] SHAHAN, R., HSU, C.-W., NOLAN, T. M., COLE, B. J., TAYLOR, I. W., GREENSTREET, L., ZHANG, S., AFANASSIEV, A., VLOT, A. H. C. et al. (2022). A single-cell Arabidopsis root atlas reveals developmental trajectories in wild-type and cell identity mutants. *Dev. Cell* **57** 543–560.

[69] SHALEK, A. K., SATIJA, R., SHUGA, J., TROMBETTA, J. J., GENNERT, D., LU, D., CHEN, P., GERTNER, R. S., GAUBLOMME, J. T. et al. (2014). Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510** 363–369.

[70] SIRIGNANO, J. and SPILIOPOULOS, K. (2020). Mean field analysis of neural networks: A law of large numbers. *SIAM J. Appl. Math.* **80** 725–752. MR4074020 https://doi.org/10.1137/18M1192184

[71] STREETS, A. M. and HUANG, Y. (2014). How deep is enough in single-cell RNA-seq? *Nat. Biotechnol.* **32** 1005–1006.

[72] TOWNES, F. W., HICKS, S. C., ARYEE, M. J. and IRIZARRY, R. A. (2019). Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol.* **20** 1–16.

[73] TRAAG, V. A., WALTMAN, L. and VAN ECK, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Sci. Rep.* **9** 1–12.

[74] VAN DIJK, D., SHARMA, R., NAINYS, J., YIM, K., KATHAIL, P., CARR, A. J., BURDZIAK, C., MOON, K. R., CHAFFER, C. L. et al. (2018). Recovering gene interactions from single-cell data using data diffusion. *Cell* **174** 716–729.

[75] VILLANI, C. (2003). *Topics in Optimal Transportation*. *Graduate Studies in Mathematics* **58**. Amer. Math. Soc., Providence, RI. MR1964483 https://doi.org/10.1090/gsm/058

[76] VILLANI, C. (2009). *Optimal Transport: Old and New*. *Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **338**. Springer, Berlin. MR2459454 https://doi.org/10.1007/978-3-540-71050-9

[77] VIRSHUP, I., RYBAKOV, S., THEIS, F. J., ANGERER, P. and WOLF, F. A. (2021). anndata: Annotated data. *bioRxiv*.

[78] WANG, G., JIAO, Y., XU, Q., WANG, Y. and YANG, C. (2021). Deep generative learning via Schrödinger bridge. In *International Conference on Machine Learning* 10794–10804. PMLR.

[79] WANG, J., HUANG, M., TORRE, E., DUECK, H., SHAFFER, S., MURRAY, J., RAJ, A., LI, M. and ZHANG, N. R. (2018). Gene expression distribution deconvolution in single-cell RNA sequencing. *Proc. Natl. Acad. Sci. USA* **115** E6437–E6446. MR3831463 https://doi.org/10.1073/pnas.1721085115

[80] WEISSMAN, T., ORDENTLICH, E., SEROUSSI, G., VERDU, S. and WEINBERGER, M. J. (2003). Inequalities for the L1 deviation of the empirical distribution.

[81] WOLF, F. A., ANGERER, P. and THEIS, F. J. (2018). SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* **19** 1–5.

[82] ZHANG, M. J., NTRANOS, V. and TSE, D. (2020). Determining sequencing depth in a single-cell RNA-seq experiment. *Nat. Commun.* **11** 774.

The Annals of Statistics
2025, Vol. 53, No. 6, 2673–2701
https://doi.org/10.1214/25-AOS2559

HIGH-DIMENSIONAL HILBERT–SCHMIDT LINEAR REGRESSION WITH HILBERT MANIFOLD VARIABLES

BY CHANGWON CHOI[a] AND BYEONG U. PARK[b]

Department of Statistics, Seoul National University, [a]*cwchoi94@gmail.com,* [b]*bupark@snu.ac.kr*

In this paper, we propose a novel high-dimensional linear regression technique for variables taking values in Hilbert manifolds. Our approach offers a flexible framework where the response and covariates originate from distinct spaces and are interconnected by Hilbert–Schmidt operators. Our methodology is designed for situations where the number of the covariates grows exponentially fast as the sample size increases and some of the covariates together with the response take values in infinite dimensional spaces. It is formulated under a general penalization scheme that includes various nonconvex penalty functions. Leveraging modern statistical theory for data residing on Hilbert manifolds, we establish the oracle property and derive error bounds for the proposed estimators. We also provide an efficient computational algorithm to solve the associated constrained optimization problem. The practical performance of the proposed method is demonstrated via numerical simulation and real data applications.

REFERENCES

ARNAUDON, M. and MICLO, L. (2014). Means in complete manifolds: Uniqueness and approximation. *ESAIM Probab. Stat.* **18** 185–206. MR3230874 https://doi.org/10.1051/ps/2013033

ARSIGNY, V., FILLARD, P., PENNEC, X. and AYACHE, N. (2007). Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM J. Matrix Anal. Appl.* **29** 328–347. MR2288028 https://doi.org/10.1137/050637996

BOYD, S., PARIKH, N., CHU, E., PELEATO, B. and ECKSTEIN, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **3** 1–122. https://doi.org/10.1561/2200000016

CAI, X., XUE, L. and CAO, J. (2022). Variable selection for multiple function-on-function linear regression. *Statist. Sinica* **32** 1435–1465. MR4449901 https://doi.org/10.5705/ss.202020.0473

CHEN, Y., LIN, Z. and MÜLLER, H.-G. (2023). Wasserstein regression. *J. Amer. Statist. Assoc.* **118** 869–882. MR4595462 https://doi.org/10.1080/01621459.2021.1956937

CHO, S., JEON, J. M., KIM, D., YU, K. and PARK, B. U. (2024). Partially linear additive regression with a general Hilbertian response. *J. Amer. Statist. Assoc.* **119** 942–956. MR4765960 https://doi.org/10.1080/01621459.2022.2149407

CHOI, C., LIN, Z. and PARK, B. U. (2025). High-dimensional partially linear additive models on Riemannian manifolds. *Bernoulli* **31** 3285–3308. MR4931371 https://doi.org/10.3150/24-BEJ1847

CHOI, C. and PARK, B. U. (2025). Supplement to "High-dimensional Hilbert–Schmidt Linear regression with Hilbert manifold variables." https://doi.org/10.1214/25-AOS2559SUPP

DAI, X. (2022). Statistical inference on the Hilbert sphere with application to random densities. *Electron. J. Stat.* **16** 700–736. MR4366819 https://doi.org/10.1214/21-ejs1942

DAI, X., LIN, Z. and MÜLLER, H.-G. (2021). Modeling sparse longitudinal data on Riemannian manifolds. *Biometrics* **77** 1328–1341. MR4357841 https://doi.org/10.1111/biom.13385

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581 https://doi.org/10.1198/016214501753382273

GHODRATI, L. and PANARETOS, V. M. (2022). Distribution-on-distribution regression via optimal transport maps. *Biometrika* **109** 957–974. MR4519110 https://doi.org/10.1093/biomet/asac005

HALL, P. and HOROWITZ, J. L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.* **35** 70–91. MR2332269 https://doi.org/10.1214/009053606000000957

MSC2020 subject classifications. Primary 62R30; secondary 62J07.

Key words and phrases. Non-Euclidean data, high-dimensional regression, Hilbert–Schmidt operators, spectral decomposition, penalization, Hilbert spaces, Riemannian manifolds, Wasserstein spaces.

HAN, K., MÜLLER, H.-G. and PARK, B. U. (2018). Smooth backfitting for additive modeling with small errors-in-variables, with an application to additive functional regression for multiple predictor functions. *Bernoulli* **24** 1233–1265. MR3706793 https://doi.org/10.3150/16-BEJ898

HAN, K., MÜLLER, H.-G. and PARK, B. U. (2020). Additive functional regression for densities as responses. *J. Amer. Statist. Assoc.* **115** 997–1010. MR4107695 https://doi.org/10.1080/01621459.2019.1604365

HSING, T. and EUBANK, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators. Wiley Series in Probability and Statistics*. Wiley, Chichester. MR3379106 https://doi.org/10.1002/9781118762547

JEON, J. M. and PARK, B. U. (2020). Additive regression with Hilbertian responses. *Ann. Statist.* **48** 2671–2697. MR4152117 https://doi.org/10.1214/19-AOS1902

JEON, J. M., PARK, B. U. and VAN KEILEGOM, I. (2021). Additive regression for non-Euclidean responses and predictors. *Ann. Statist.* **49** 2611–2641. MR4338377 https://doi.org/10.1214/21-aos2048

JEON, J. M., PARK, B. U. and VAN KEILEGOM, I. (2022). Nonparametric regression on Lie groups with measurement errors. *Ann. Statist.* **50** 2973–3008. MR4500632 https://doi.org/10.1214/22-aos2218

JEON, J. M. and VAN BEVER, G. (2022). Additive regression with general imperfect variables. arXiv preprint. Available at arXiv:2212.05745.

KADISON, R. V. and RINGROSE, J. R. (1997). *Fundamentals of the Theory of Operator Algebras. Vol. I: Elementary Theory. Graduate Studies in Mathematics* **15**. Amer. Math. Soc., Providence, RI. Reprint of the 1983 original. MR1468229 https://doi.org/10.1090/gsm/015

KLINGENBERG, W. P. A. (1995). *Riemannian Geometry*, 2nd ed. *De Gruyter Studies in Mathematics* **1**. de Gruyter, Berlin. MR1330918 https://doi.org/10.1515/9783110905120

KONG, D., XUE, K., YAO, F. and ZHANG, H. H. (2016). Partially functional linear regression in high dimensions. *Biometrika* **103** 147–159. MR3465827 https://doi.org/10.1093/biomet/asv062

LANG, S. (1999). *Fundamentals of Differential Geometry. Graduate Texts in Mathematics* **191**. Springer, New York. MR1666820 https://doi.org/10.1007/978-1-4612-0541-8

LIN, Z., KONG, D. and WANG, L. (2023). Causal inference on distribution functions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **85** 378–398. MR4725988 https://doi.org/10.1093/jrsssb/qkad008

LIN, Z. and MÜLLER, H.-G. (2021). Total variation regularized Fréchet regression for metric-space valued data. *Ann. Statist.* **49** 3510–3533. MR4352539 https://doi.org/10.1214/21-aos2095

LIN, Z., MÜLLER, H.-G. and PARK, B. U. (2023). Additive models for symmetric positive-definite matrices and Lie groups. *Biometrika* **110** 361–379. MR4588358 https://doi.org/10.1093/biomet/asac055

LIN, Z. and YAO, F. (2019). Intrinsic Riemannian functional data analysis. *Ann. Statist.* **47** 3533–3577. MR4025751 https://doi.org/10.1214/18-AOS1787

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

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

LOH, P.-L. and WAINWRIGHT, M. J. (2017). Support recovery without incoherence: A case for nonconvex regularization. *Ann. Statist.* **45** 2455–2482. MR3737898 https://doi.org/10.1214/16-AOS1530

MATABUENA, M. and PETERSEN, A. (2023). Distributional data analysis of accelerometer data from the NHANES database using nonparametric survey regression models. *J. R. Stat. Soc., Ser. C, Appl. Stat.* **72** 294–313. MR4719277 https://doi.org/10.1093/jrsssc/qlad007

PANARETOS, V. M. and ZEMEL, Y. (2020). *An Invitation to Statistics in Wasserstein Space. SpringerBriefs in Probability and Mathematical Statistics*. Springer, Cham. MR4350694 https://doi.org/10.1007/978-3-030-38438-8

PENNEC, X. (2018). Barycentric subspace analysis on manifolds. *Ann. Statist.* **46** 2711–2746. MR3851753 https://doi.org/10.1214/17-AOS1636

PENNEC, X., SOMMER, S. and FLETCHER, T., eds. (2020). *Riemannian Geometric Statistics in Medical Image Analysis* Elsevier, London. MR4254211

PETERSEN, A. and MÜLLER, H.-G. (2019). Fréchet regression for random objects with Euclidean predictors. *Ann. Statist.* **47** 691–719. MR3909947 https://doi.org/10.1214/17-AOS1624

PETERSEN, A., ZHANG, C. and KOKOSZKA, P. (2022). Modeling probability density functions as data objects. *Econom. Stat.* **21** 159–178. MR4366852 https://doi.org/10.1016/j.ecosta.2021.04.004

RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *J. Mach. Learn. Res.* **11** 2241–2259. MR2719855

SCHÖTZ, C. (2019). Convergence rates for the generalized Fréchet mean via the quadruple inequality. *Electron. J. Stat.* **13** 4280–4345. MR4023955 https://doi.org/10.1214/19-EJS1618

SHAO, L., LIN, Z. and YAO, F. (2022). Intrinsic Riemannian functional data analysis for sparse longitudinal observations. *Ann. Statist.* **50** 1696–1721. MR4441137 https://doi.org/10.1214/22-aos2172

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

TUCKER, D. C., WU, Y. and MÜLLER, H.-G. (2023). Variable selection for global Fréchet regression. *J. Amer. Statist. Assoc.* **118** 1023–1037. MR4595474 https://doi.org/10.1080/01621459.2021.1969240

VAN DE GEER, S. A. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **3** 1360–1392. MR2576316 https://doi.org/10.1214/09-EJS506

WAGHMARE, K. G., MASAK, T. and PANARETOS, V. M. (2023). The functional graphical Lasso. arXiv preprint. Available at arXiv:2306.02347.

WANG, J.-L., CHIOU, J.-M. and MÜLLER, H.-G. (2016). Functional data analysis. *Annu. Rev. Stat. Appl.* **3** 257–295. https://doi.org/10.1146/annurev-statistics-041715-033624

YUN, H. and PARK, B. U. (2023). Exponential concentration for geometric-median-of-means in non-positive curvature spaces. *Bernoulli* **29** 2927–2960. MR4632126 https://doi.org/10.3150/22-bej1569

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701 https://doi.org/10.1214/09-AOS729

ZHANG, X., PARK, B. U. and WANG, J.-L. (2013). Time-varying additive models for longitudinal data. *J. Amer. Statist. Assoc.* **108** 983–998. MR3174678 https://doi.org/10.1080/01621459.2013.778776

The Annals of Statistics
2025, Vol. 53, No. 6, 2702–2727
https://doi.org/10.1214/25-AOS2561

ONLINE ESTIMATION WITH ROLLING VALIDATION: ADAPTIVE NONPARAMETRIC ESTIMATION WITH STREAMING DATA

BY TIANYU ZHANG[1,a] AND JING LEI[2,b]

[1]*Department of Statistics and Applied Probability, University of California, Santa Barbara,* [a]*tyz@ucsb.edu*
[2]*Department of Statistics & Data Science, Carnegie Mellon University,* [b]*jinglei@andrew.cmu.edu*

Online nonparametric estimators are gaining popularity due to their efficient computation and competitive generalization abilities. An important example includes variants of stochastic gradient descent. These algorithms often take one sample point at a time and incrementally update the parameter estimate of interest. In this work, we consider model selection/hyperparameter tuning for such online algorithms. We propose a weighted rolling validation procedure, an online variant of leave-one-out cross-validation, that costs minimal extra computation for many typical stochastic gradient descent estimators and maintains their online nature. Similar to batch cross-validation, it can boost base estimators to achieve better heuristic performance and adaptive convergence rate. Our analysis is straightforward, relying mainly on some general statistical stability assumptions. The simulation study underscores the significance of diverging weights in practice and demonstrates its favorable sensitivity even when there is only a slim difference between candidate estimators.

REFERENCES

[1] ARLOT, S. and CELISSE, A. (2010). A survey of cross-validation procedures for model selection. *Stat. Surv.* **4** 40–79. MR2602303 https://doi.org/10.1214/09-SS054

[2] ARMSTRONG, J. S. and GROHMAN, M. C. (1972). A comparative study of methods for long-range market forecasting. *Manag. Sci.* **19** 211–221.

[3] BACH, F. and MOULINES, E. (2013). Non-strongly-convex smooth stochastic approximation with convergence rate O (1/n). *Adv. Neural Inf. Process. Syst.* **26**.

[4] BATES, S., HASTIE, T. and TIBSHIRANI, R. (2024). Cross-validation: What does it estimate and how well does it do it? *J. Amer. Statist. Assoc.* **119** 1434–1445. MR4766000 https://doi.org/10.1080/01621459.2023.2197686

[5] BENKESER, D., JU, C., LENDLE, S. and VAN DER LAAN, M. (2018). Online cross-validation-based ensemble learning. *Stat. Med.* **37** 249–260. MR3738747 https://doi.org/10.1002/sim.7320

[6] BOUSQUET, O. and ELISSEEFF, A. (2002). Stability and generalization. *J. Mach. Learn. Res.* **2** 499–526. MR1929416 https://doi.org/10.1162/153244302760200704

[7] BROWNE, M. W. (2000). Cross-validation methods. *J. Math. Psych.* **44** 108–132. MR1770004 https://doi.org/10.1006/jmps.1999.1279

[8] CALANDRIELLO, D., LAZARIC, A. and VALKO, M. (2017). Efficient second-order online kernel learning with adaptive embedding. *Adv. Neural Inf. Process. Syst.* **30**.

[9] CHEN, X. and KLUSOWSKI, J. M. (2024). Stochastic Gradient Descent for Additive Nonparametric Regression. arXiv Preprint. Available at arXiv:2401.00691.

[10] DIEULEVEUT, A. and BACH, F. (2016). Nonparametric stochastic approximation with large step-sizes. *Ann. Statist.* **44** 1363–1399. MR3519927 https://doi.org/10.1214/15-AOS1391

[11] DING, J., TAROKH, V. and YANG, Y. (2018). Model selection techniques: An overview. *IEEE Signal Process. Mag.* **35** 16–34.

[12] 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

[13] GYÖRFI, L., KOHLER, M., KRZYŻAK, A. and WALK, H. (2002). *A Distribution-Free Theory of Nonparametric Regression. Springer Series in Statistics*. Springer, New York. MR1920390 https://doi.org/10.1007/b97848

MSC2020 subject classifications. 62L12, 62G05.
Key words and phrases. Online learning, model selection, stochastic gradient descent.

[14] HARDT, M., RECHT, B. and SINGER, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning* 1225–1234. PMLR.

[15] HOI, S. C., SAHOO, D., LU, J. and ZHAO, P. (2021). Online learning: A comprehensive survey. *Neurocomputing* **459** 249–289.

[16] HOMRIGHAUSEN, D. and MCDONALD, D. J. (2017). Risk consistency of cross-validation with lasso-type procedures. *Statist. Sinica* **27** 1017–1036. MR3699692

[17] KISSEL, N. and LEI, J. (2022). On High-Dimensional Gaussian Comparisons for Cross-Validation. arXiv Preprint. Available at arXiv:2211.04958.

[18] LANGFORD, J., LI, L. and ZHANG, T. (2009). Sparse online learning via truncated gradient. *J. Mach. Learn. Res.* **10** 777–801. MR2491758

[19] MARTEAU-FEREY, U., BACH, F. and RUDI, A. (2019). Globally convergent Newton methods for ill-conditioned generalized self-concordant losses. *Adv. Neural Inf. Process. Syst.* **32**.

[20] PICARD, R. R. and COOK, R. D. (1984). Cross-validation of regression models. *J. Amer. Statist. Assoc.* **79** 575–583. MR0763576

[21] RAO, R. B., FUNG, G. and ROSALES, R. (2008). On the dangers of cross-validation. An experimental evaluation. In *Proceedings of the* 2008 *SIAM International Conference on Data Mining* 588–596. SIAM, Philadelphia.

[22] ROBERTS, D. R., BAHN, V., CIUTI, S., BOYCE, M. S., ELITH, J., GUILLERA-ARROITA, G., HAUENSTEIN, S., LAHOZ-MONFORT, J. J., SCHRÖDER, B. et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **40** 913–929.

[23] SHAO, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* **88** 486–494. MR1224373

[24] STONE, M. (1978). Cross-validation: A review. *Math. Operforsch. Stat., Ser. Stat.* **9** 127–139. MR0494581 https://doi.org/10.1080/02331887808801414

[25] TARRÈS, P. and YAO, Y. (2014). Online learning as stochastic approximation of regularization paths: Optimality and almost-sure convergence. *IEEE Trans. Inf. Theory* **60** 5716–5735. MR3252416 https://doi.org/10.1109/TIT.2014.2332531

[26] TASHMAN, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *Int. J. Forecast.* **16** 437–450.

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

[28] TIBSHIRANI, R. J. and TIBSHIRANI, R. (2009). A bias correction for the minimum error rate in cross-validation. *Ann. Appl. Stat.* **3** 822–829. MR2750683 https://doi.org/10.1214/08-AOAS224

[29] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. Springer, New York. MR2724359 https://doi.org/10.1007/b13794

[30] VAN DER VAART, A. W., DUDOIT, S. and VAN DER LAAN, M. J. (2006). Oracle inequalities for multi-fold cross validation. *Statist. Decisions* **24** 351–371. MR2305112 https://doi.org/10.1524/stnd.2006.24.3.351

[31] VIEN, B. S., WONG, L., KUEN, T., ROSE, L. and CHIU, W. K. (2021). A machine learning approach for anaerobic reactor performance prediction using long short-term memory recurrent neural network. *Struct. Health Monit.* 8*apwshm* **18** 61.

[32] VLADIMIROVA, M., GIRARD, S., NGUYEN, H. and ARBEL, J. (2020). Sub-Weibull distributions: Generalizing sub-Gaussian and sub-exponential properties to heavier tailed distributions. *Stat* **9** e318. MR4193421 https://doi.org/10.1002/sta4.318

[33] WEI, C. Z. (1992). On predictive least squares principles. *Ann. Statist.* **20** 1–42. MR1150333 https://doi.org/10.1214/aos/1176348511

[34] YANG, Y. (2007). Consistency of cross validation for comparing regression procedures. *Ann. Statist.* **35** 2450–2473. MR2382654 https://doi.org/10.1214/009053607000000514

[35] YANG, Y. and YAO, F. (2023). Online estimation for functional data. *J. Amer. Statist. Assoc.* **118** 1630–1644. MR4646594 https://doi.org/10.1080/01621459.2021.2002158

[36] YING, Y. and PONTIL, M. (2008). Online gradient descent learning algorithms. *Found. Comput. Math.* **8** 561–596. MR2443089 https://doi.org/10.1007/s10208-006-0237-y

[37] ZHANG, P. (1993). Model selection via multifold cross validation. *Ann. Statist.* **21** 299–313. MR1212178 https://doi.org/10.1214/aos/1176349027

[38] ZHANG, T. and LEI, J. (2025). Supplement to "Online estimation with rolling validation: Adaptive nonparametric estimation with streaming data." https://doi.org/10.1214/25-AOS2561SUPP

[39] ZHANG, T. and SIMON, N. (2022). A sieve stochastic gradient descent estimator for online nonparametric regression in Sobolev ellipsoids. *Ann. Statist.* **50** 2848–2871. MR4500627 https://doi.org/10.1214/22-aos2212

[40] ZHANG, T. and SIMON, N. (2023). An online projection estimator for nonparametric regression in reproducing kernel Hilbert spaces. *Statist*. *Sinica* **33** 127–148. MR4527765

[41] ZHANG, T. and SIMON, N. (2023). Regression in tensor product spaces by the method of sieves. *Electron*. *J*. *Stat*. **17** 3660–3727. MR4675048 https://doi.org/10.1214/23-ejs2188

[42] ZOU, D., WU, J., BRAVERMAN, V., GU, Q. and KAKADE, S. (2021). Benign overfitting of constant-stepsize sgd for linear regression. In *Conference on Learning Theory* 4633–4635. PMLR.

The Annals of Statistics
2025, Vol. 53, No. 6, 2728–2756
https://doi.org/10.1214/25-AOS2562

FUNDAMENTAL LIMITS OF COMMUNITY DETECTION FROM MULTI-VIEW DATA: MULTI-LAYER, DYNAMIC AND PARTIALLY LABELED BLOCK MODELS

BY XIAODONG YANG[a], BUYU LIN[b] AND SUBHABRATA SEN[c]

Department of Statistics, Harvard University, [a]*xyang@g.harvard.edu,* [b]*blin@g.harvard.edu,* [c]*subhabratasen@fas.harvard.edu*

Multi-view data arises frequently in modern network analysis for example, relations of multiple types among individuals in social network analysis, longitudinal measurements of interactions among observational units, annotated networks with noisy partial labeling of vertices etc. We study community detection in these disparate settings via a unified theoretical framework, and investigate the fundamental thresholds for community recovery. We characterize the mutual information between the data and the latent parameters, provided the degrees are sufficiently large. Based on this general result, (i) we derive a sharp threshold for community detection in an inhomogeneous multi-layer block model (*Ann. Statist.* **50** (2022) 2664–2693), (ii) characterize a sharp threshold for weak recovery in a dynamic stochastic block model (*J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** (2017) 1119–1141), and (iii) identify the limiting mutual information in an unbalanced partially labeled block model. Finally, we introduce iterative algorithms based on Approximate Message Passing for community detection in these problems.

REFERENCES

[1] ABBE, E. (2017). Community detection and stochastic block models: Recent developments. *J. Mach. Learn. Res.* **18** 6446–6531. MR3827065

[2] ALLAHVERDYAN, A. E., VER STEEG, G. and GALSTYAN, A. (2010). Community detection with and without prior information. *Europhys. Lett.* **90** 18002.

[3] AUBIN, B., MAILLARD, A., KRZAKALA, F., MACRIS, N. and ZDEBOROVÁ, L. (2018). The committee machine: Computational to statistical gaps in learning a two-layers neural network. *Adv. Neural Inf. Process. Syst.* **31**.

[4] BAKKEN, T. E., MILLER, J. A., DING, S.-L., SUNKIN, S. M., SMITH, K. A., NG, L., SZAFER, A., DALLEY, R. A., ROYALL, J. J. et al. (2016). A comprehensive transcriptional map of primate brain development. *Nature* **535** 367–375.

[5] BAYATI, M. and MONTANARI, A. (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. Inf. Theory* **57** 764–785. MR2810285 https://doi.org/10.1109/TIT.2010.2094817

[6] BEHNE, J. K. and REEVES, G. (2022). Fundamental limits for rank-one matrix estimation with groupwise heteroskedasticity. In *International Conference on Artificial Intelligence and Statistics* 8650–8672. PMLR.

[7] BERTHIER, R., MONTANARI, A. and NGUYEN, P.-M. (2020). State evolution for approximate message passing with non-separable functions. *Inf. Inference* **9** 33–79. MR4079177 https://doi.org/10.1093/imaiai/iay021

[8] BHATTACHARJEE, M., BANERJEE, M. and MICHAILIDIS, G. (2020). Change point estimation in a dynamic stochastic block model. *J. Mach. Learn. Res.* **21** 4330–4388. MR4119175

[9] BHATTACHARYYA, S. and CHATTERJEE, S. (2020). General community detection with optimal recovery conditions for multi-relational sparse networks with dependent layers. arXiv preprint. Available at arXiv:2004.03480.

[10] BOBKOV, S. G., CHISTYAKOV, G. P. and GÖTZE, F. (2024). Strictly subgaussian probability distributions. *Electron. J. Probab.* **29** 1–28. MR4736269 https://doi.org/10.1214/24-ejp1120

MSC2020 subject classifications. Primary 62F15, 94A16; secondary 60C05.
Key words and phrases. Stochastic block model, multi-layer networks, dynamical networks, weak recovery.

[11] BORDENAVE, C., LELARGE, M. and MASSOULIÉ, L. (2015). Non-backtracking spectrum of random graphs: Community detection and non-regular Ramanujan graphs. In 2015 *IEEE 56th Annual Symposium on Foundations of Computer Science—FOCS* 2015 1347–1357. IEEE Comput. Soc., Los Alamitos, CA. MR3473374 https://doi.org/10.1109/FOCS.2015.86

[12] CAI, T. T., LIANG, T. and RAKHLIN, A. (2020). Weighted message passing and minimum energy flow for heterogeneous stochastic block models with side information. *J. Mach. Learn. Res.* **21** 346–379. MR4071194

[13] CHEN, H., MOURRAT, J.-C. and XIA, J. (2022). Statistical inference of finite-rank tensors. *Ann. Henri Lebesgue* **5** 1161–1189. MR4526249 https://doi.org/10.5802/ahl.146

[14] CHEN, S., LIU, S. and MA, Z. (2022). Global and individualized community detection in inhomogeneous multilayer networks. *Ann. Statist.* **50** 2664–2693. MR4500621 https://doi.org/10.1214/22-aos2202

[15] COVER, T. M. and THOMAS, J. A. (2005). *Elements of Information Theory*, Wiley, New York.

[16] DESHPANDE, Y., ABBE, E. and MONTANARI, A. (2017). Asymptotic mutual information for the balanced binary stochastic block model. *Inf. Inference* **6** 125–170. MR3671474 https://doi.org/10.1093/imaiai/iaw017

[17] DESHPANDE, Y., SEN, S., MONTANARI, A. and MOSSEL, E. (2018). Contextual stochastic block models. *Adv. Neural Inf. Process. Syst.* **31**.

[18] DING, D. Y., LI, S., NARASIMHAN, B. and TIBSHIRANI, R. (2022). Cooperative learning for multiview analysis. *Proc. Natl. Acad. Sci. USA* **119** e2202113119.

[19] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2009). Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci. USA* **106** 18914–18919.

[20] FENG, O. Y., VENKATARAMANAN, R., RUSH, C. and SAMWORTH, R. J. (2022). A unifying tutorial on approximate message passing. *Found. Trends Mach. Learn.* **15** 335–536.

[21] GERBELOT, C. and BERTHIER, R. (2023). Graph-based approximate message passing iterations. *Inf. Inference* **12** Paper No. iaad020, 67. MR4644961 https://doi.org/10.1093/imaiai/iaad020

[22] GRENANDER, U. and SZEGÖ, G. (1958). *Toeplitz Forms and Their Applications*. *California Monographs in Mathematical Sciences*. Univ. California Press, Berkeley, CA. MR0094840

[23] GUO, D., SHAMAI, S. and VERDÚ, S. (2005). Mutual information and minimum mean-square error in Gaussian channels. *IEEE Trans. Inf. Theory* **51** 1261–1282. MR2241490 https://doi.org/10.1109/TIT.2005.844072

[24] JI, P., JIN, J., KE, Z. T. and LI, W. (2022). Co-citation and co-authorship networks of statisticians. *J. Bus. Econom. Statist.* **40** 469–485. MR4410871 https://doi.org/10.1080/07350015.2021.1978469

[25] JIANG, Y. and KE, T. (2022). Semi-supervised community detection via structural similarity metrics. In *The Eleventh International Conference on Learning Representations*.

[26] KANADE, V., MOSSEL, E. and SCHRAMM, T. (2016). Global and local information in clustering labeled block models. *IEEE Trans. Inf. Theory* **62** 5906–5917. MR3552430 https://doi.org/10.1109/TIT.2016.2516564

[27] KARRER, B. and NEWMAN, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **83** 016107, 10. MR2788206 https://doi.org/10.1103/PhysRevE.83.016107

[28] KERIVEN, N. and VAITER, S. (2022). Sparse and smooth: Improved guarantees for spectral clustering in the dynamic stochastic block model. *Electron. J. Stat.* **16** 1330–1366. MR4387844 https://doi.org/10.1214/22-ejs1986

[29] KIVELÄ, M., ARENAS, A., BARTHELEMY, M., GLEESON, J. P., MORENO, Y. and PORTER, M. A. (2014). Multilayer networks. *J. Complex Netw.* **2** 203–271.

[30] KRZAKALA, F., MOORE, C., MOSSEL, E., NEEMAN, J., SLY, A., ZDEBOROVÁ, L. and ZHANG, P. (2013). Spectral redemption in clustering sparse networks. *Proc. Natl. Acad. Sci. USA* **110** 20935–20940. MR3174850 https://doi.org/10.1073/pnas.1312486110

[31] LEI, J., CHEN, K. and LYNCH, B. (2020). Consistent community detection in multi-layer network data. *Biometrika* **107** 61–73. MR4064140 https://doi.org/10.1093/biomet/asz068

[32] LEI, J., ZHANG, A. R. and ZHU, Z. (2024). Computational and statistical thresholds in multi-layer stochastic block models. *Ann. Statist.* **52** 2431–2455. MR4829494 https://doi.org/10.1214/24-aos2441

[33] LELARGE, M. and MIOLANE, L. (2019). Fundamental limits of symmetric low-rank matrix estimation. *Probab. Theory Related Fields* **173** 859–929. MR3936148 https://doi.org/10.1007/s00440-018-0845-x

[34] LENG, M. and MA, T. (2019). Semi-supervised community detection: A survey. In *Proceedings of the* 2019 *7th International Conference on Information Technology: IoT and Smart City* 137–140.

[35] LESIEUR, T., KRZAKALA, F. and ZDEBOROVÁ, L. (2017). Constrained low-rank matrix estimation: Phase transitions, approximate message passing and applications. *J. Stat. Mech. Theory Exp.* **7** 073403, 86. MR3683819 https://doi.org/10.1088/1742-5468/aa7284

[36] LI, A., CORNELIUS, S. P., LIU, Y.-Y., WANG, L. and BARABÁSI, A.-L. (2017). The fundamental advantages of temporal networks. *Science* **358** 1042–1046.

[37] 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, 7. MR4637851

[38] LIN, K. Z. and LEI, J. (2024). Dynamic clustering for heterophilic stochastic block models with time-varying node memberships. arXiv preprint. Available at arXiv:2403.05654.

[39] LONGEPIERRE, L. and MATIAS, C. (2019). Consistency of the maximum likelihood and variational estimators in a dynamic stochastic block model. *Electron. J. Stat.* **13** 4157–4223. MR4021264 https://doi.org/10.1214/19-EJS1624

[40] LU, C. and SEN, S. (2023). Contextual stochastic block model: Sharp thresholds and contiguity. *J. Mach. Learn. Res.* **24** 1–34. MR4582476

[41] MA, Z. and NANDY, S. (2023). Community detection with contextual multilayer networks. *IEEE Trans. Inf. Theory* **69** 3203–3239. MR4581286 https://doi.org/10.1109/tit.2023.3238352

[42] MANOEL, A., KRZAKALA, F., MÉZARD, M. and ZDEBOROVÁ, L. (2017). Multi-layer generalized linear estimation. In 2017 *IEEE International Symposium on Information Theory* (*ISIT*) 2098–2102. IEEE Press, New York.

[43] MASUDA, N. and HOLME, P. (2017). *Introduction to Temporal Network Epidemiology*. Springer, Berlin.

[44] MATIAS, C. and MIELE, V. (2017). Statistical clustering of temporal networks through a dynamic stochastic block model. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1119–1141. MR3689311 https://doi.org/10.1111/rssb.12200

[45] MAYYA, V. and REEVES, G. (2019). Mutual information in community detection with covariate information and correlated networks. In 2019 57*th Annual Allerton Conference on Communication*, *Control*, *and Computing* (*Allerton*) 602–607. IEEE Press, New York.

[46] MONTANARI, A. and SEN, S. (2022). A short tutorial on mean-field spin glass techniques for non-physicists. arXiv preprint. Available at arXiv:2204.02909.

[47] MONTANARI, A. and VENKATARAMANAN, R. (2021). Estimation of low-rank matrices via approximate message passing. *Ann. Statist.* **49** 321–345. MR4206680 https://doi.org/10.1214/20-AOS1958

[48] MOSSEL, E., NEEMAN, J. and SLY, A. (2012). Stochastic block models and reconstruction. arXiv preprint. Available at arXiv:1202.1499.

[49] MOSSEL, E., NEEMAN, J. and SLY, A. (2015). Reconstruction and estimation in the planted partition model. *Probab. Theory Related Fields* **162** 431–461. MR3383334 https://doi.org/10.1007/s00440-014-0576-6

[50] MOSSEL, E., NEEMAN, J. and SLY, A. (2018). A proof of the block model threshold conjecture. *Combinatorica* **38** 665–708. MR3876880 https://doi.org/10.1007/s00493-016-3238-8

[51] NEWMAN, C. M. (1975). Inequalities for Ising models and field theories which obey the Lee–Yang theorem. *Comm. Math. Phys.* **41** 1–9. MR0376061

[52] PENSKY, M. and ZHANG, T. (2019). Spectral clustering in the dynamic stochastic block model. *Electron. J. Stat.* **13** 678–709. MR3914178 https://doi.org/10.1214/19-ejs1533

[53] 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

[54] REEVES, G. (2020). Information-theoretic limits for the matrix tensor product. *IEEE J. Sel. Areas Inf. Theory* **1** 777–798.

[55] RICHARD, E. and MONTANARI, A. (2014). A statistical model for tensor PCA. *Adv. Neural Inf. Process. Syst.* **27**.

[56] ROSSETTI, R. and REEVES, G. (2023). Approximate message passing for the matrix tensor product model. arXiv preprint. Available at arXiv:2306.15580.

[57] SAADE, A., KRZAKALA, F., LELARGE, M. and ZDEBOROVÁ, L. (2018). Fast randomized semi-supervised clustering. *J. Phys.*, *Conf. Ser.* **1036** 012015. IOP Publishing.

[58] SHAPIRA, B., ROKACH, L. and FREILIKHMAN, S. (2013). Facebook single and cross domain data for recommendation systems. *User Model. User-Adapt. Interact.* **23** 211–247.

[59] TRENCH, W. F. (2010). Spectral decomposition of Kac–Murdock–Szego matrices. The Selected Works of William F. Trench. Available at http://works.bepress.com/william_trench/133.

[60] WANG, T., ZHONG, X. and FAN, Z. (2024). Universality of approximate message passing algorithms and tensor networks. *Ann. Appl. Probab.* **34** 3943–3994. MR4783034 https://doi.org/10.1214/24-aap2056

[61] WELCH, G. and BISHOP, G. (1995). An introduction to the Kalman filter.

[62] WESTON, J., ZHOU, D., ELISSEEFF, A., NOBLE, W. and LESLIE, C. (2003). Semi-supervised protein classification using cluster kernels. *Adv. Neural Inf. Process. Syst.* **16**.

[63] WU, Y. and XU, J. (2018). Statistical problems with planted structures: Information-theoretical and computational limits. arXiv preprint. Available at arXiv:1806.00118.

[64] WU, Y. and XU, J. (2021). Statistical problems with planted structures: Information-theoretical and computational limits. *Inf.-Theor. Methods Data Sci.* **383** 13.

[65] YANG, T., CHI, Y., ZHU, S., GONG, Y. and JIN, R. (2011). Detecting communities and their evolutions in dynamic social networks—a Bayesian approach. *Mach. Learn.* **82** 157–189. MR3108191 https://doi.org/10.1007/s10994-010-5214-7

[66] YANG, X., LIN, B. and SEN, S. (2025). Supplement to "Fundamental limits of community detection from multi-view data: multi-layer, dynamic and partially labeled block models." https://doi.org/10.1214/25-AOS2562SUPP

[67] ZHANG, J., WANG, J. and WANG, X. (2024). Consistent community detection in inter-layer dependent multi-layer networks. *J. Amer. Statist. Assoc.* **119** 3141–3151. MR4833943 https://doi.org/10.1080/01621459.2024.2308848

[68] ZHANG, P., MOORE, C. and ZDEBOROVÁ, L. (2014). Phase transitions in semisupervised clustering of sparse networks. *Phys. Rev. E* **90** 052802.

VOLUME 53

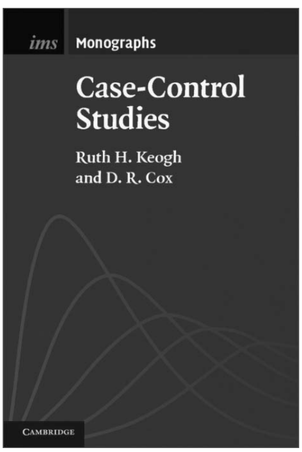
2025

CONTENTS OF VOLUME 53

Articles

The Institute of Mathematical Statistics presents

IMS MONOGRAPHS



Case-Control Studies

Ruth H. Keogh
and D. R. Cox

The case-control approach is a powerful method for investigating factors that may explain a particular event. It is extensively used in epidemiology to study disease incidence, one of the best-known examples being Bradford Hill and Doll's investigation of the possible connection between cigarette smoking and lung cancer. More recently, case-control studies have been increasingly used in other fields, including sociology and econometrics.

With a particular focus on statistical analysis, this book is ideal for applied and theoretical statisticians wanting an up-to-date introduction to the field. It covers the fundamentals of case-control study design and analysis as well as more recent developments, including two-stage studies, case-only studies and methods for case-control sampling in time. The latter have important applications in large prospective cohorts which require case-control sampling designs to make efficient use of resources. More theoretical background is provided in an appendix for those new to the field.

IMS member? Claim your 40% discount:
www.cambridge.org/ims

Hardback price US$48.00 (non-member price $80.00)

Cambridge University Press, in conjunction with the Institute of Mathematical Statistics, established the IMS Monographs and IMS Textbooks series of high-quality books. The Series Editors are Xiao-Li Meng, Susan Holmes, Ben Hambly, D. R. Cox and Alan Agresti.