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

THE ANNALS *of* STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles

Vol. 50, No. 2—April 2022

THE ANNALS OF STATISTICS Vol. 50, No. 2, pp. 615–1245 April 2022

The Annals of Statistics **[ISSN 0090-5364 (print); ISSN 2168-8966 (online)],** Volume 50, Number 2, April 2022. Published bimonthly by the Institute of Mathematical Statistics, 9760 Smith Road, Waite Hill, Ohio 44094, USA. Periodicals postage paid at Cleveland, Ohio, and at additional mailing offices.

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

The Annals of Statistics
2022, Vol. 50, No. 2, 615–639
https://doi.org/10.1214/19-AOS1932

TESTABILITY OF HIGH-DIMENSIONAL LINEAR MODELS WITH NONSPARSE STRUCTURES

BY JELENA BRADIC[1,a], JIANQING FAN[2,b] AND YINCHU ZHU[3,c]

[1]*Department of Mathematics and Halicioğlu Data Science Institute, University of California, San Diego,* [a]*jbradic@ucsd.edu*
[2]*Department of Operations Research and Financial Engineering, Princeton University,* [b]*jqfan@princeton.edu*
[3]*Department of Economics and International Business School, Brandeis University,* [c]*yinchuzhu@brandeis.edu*

Understanding statistical inference under possibly nonsparse high-dimensional models has gained much interest recently. For a given component of the regression coefficient, we show that the difficulty of the problem depends on the sparsity of the corresponding row of the precision matrix of the covariates, not the sparsity of the regression coefficients. We develop new concepts of uniform and essentially uniform nontestability that allow the study of limitations of tests across a broad set of alternatives. Uniform nontestability identifies a collection of alternatives such that the power of any test, against any alternative in the group, is asymptotically at most equal to the nominal size. Implications of the new constructions include new minimax testability results that, in sharp contrast to the current results, do not depend on the sparsity of the regression parameters. We identify new tradeoffs between testability and feature correlation. In particular, we show that, in models with weak feature correlations, minimax lower bound can be attained by a test whose power has the $\sqrt{n}$ rate, regardless of the size of the model sparsity.

REFERENCES

BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.* **81** 608–650. MR3207983 https://doi.org/10.1093/restud/rdt044

BELLONI, A., CHERNOZHUKOV, V. and KATO, K. (2015). Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems. *Biometrika* **102** 77–94. MR3335097 https://doi.org/10.1093/biomet/asu056

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469 https://doi.org/10.1214/08-AOS620

BOYLE, E. A., LI, Y. I. and PRITCHARD, J. K. (2017). An expanded view of complex traits: From polygenic to omnigenic. *Cell* **169** 1177–1186. https://doi.org/10.1016/j.cell.2017.05.038

BRADIC, J., FAN, J. and ZHU, Y. (2022). Supplement to "Testability of high-dimensional linear models with nonsparse structures." https://doi.org/10.1214/19-AOS1932SUPP

BÜHLMANN, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* **19** 1212–1242. MR3102549 https://doi.org/10.3150/12-BEJSP11

CAI, T. T. and GUO, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Ann. Statist.* **45** 615–646. MR3650395 https://doi.org/10.1214/16-AOS1461

CAI, T. T. and GUO, Z. (2018). Accuracy assessment for high-dimensional linear regression. *Ann. Statist.* **46** 1807–1836. MR3819118 https://doi.org/10.1214/17-AOS1604

CAI, T. T. and LOW, M. G. (2004). An adaptation theory for nonparametric confidence intervals. *Ann. Statist.* **32** 1805–1840. MR2102494 https://doi.org/10.1214/009053604000000049

CAI, T. T. and LOW, M. G. (2006). Adaptive confidence balls. *Ann. Statist.* **34** 202–228. MR2275240 https://doi.org/10.1214/009053606000000146

CHAKRAVARTI, A. and TURNER, T. N. (2016). Revealing rate-limiting steps in complex disease biology: The crucial importance of studying rare, extreme-phenotype families. *BioEssays* **38** 578–586.

MSC2020 subject classifications. Primary 62C20, 62F03; secondary 62F30, 62J07.
Key words and phrases. Minimax theory, ℓ_2-constraint, confidence intervals, uniform nontestability.

CHALKIDOU, A., O'DOHERTY, M. J. and MARSDEN, P. K. (2015). False discovery rates in PET and CT studies with texture features: A systematic review. *PLoS ONE* **10** e0124165. https://doi.org/10.1371/journal.pone.0124165

FURLONG, L. I. (2013). Human diseases through the lens of network biology. *Trends Genet*. **29** 150–159.

GANJGAHI, H., WINKLER, A. M., GLAHN, D. C., BLANGERO, J., DONOHUE, B., KOCHUNOV, P. and NICHOLS, T. E. (2018). Fast and powerful genome wide association of dense genetic data with high dimensional imaging phenotypes. *Nat. Commun*. **9** 3254. https://doi.org/10.1038/s41467-018-05444-6

GENOVESE, C. and WASSERMAN, L. (2008). Adaptive confidence bands. *Ann. Statist*. **36** 875–905. MR2396818 https://doi.org/10.1214/07-AOS500

GOEMAN, J. J., VAN DE GEER, S. A. and VAN HOUWELINGEN, H. C. (2006). Testing against a high dimensional alternative. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **68** 477–493. MR2278336 https://doi.org/10.1111/j.1467-9868.2006.00551.x

HOFFMANN, M. and NICKL, R. (2011). On adaptive inference and confidence bands. *Ann. Statist*. **39** 2383–2409. MR2906872 https://doi.org/10.1214/11-AOS903

HUMAN MICROBIOME PROJECT CONSORTIUM (2012). Structure, function and diversity of the healthy human microbiome. *Nature* **486** 207–214. https://doi.org/10.1038/nature11234

INGSTER, Y. I., TSYBAKOV, A. B. and VERZELEN, N. (2010). Detection boundary in sparse regression. *Electron. J. Stat*. **4** 1476–1526. MR2747131 https://doi.org/10.1214/10-EJS589

JANSON, L., FOYGEL BARBER, R. and CANDÈS, E. (2017). EigenPrism: Inference for high dimensional signal-to-noise ratios. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **79** 1037–1065. MR3689308 https://doi.org/10.1111/rssb.12203

JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res*. **15** 2869–2909. MR3277152

JAVANMARD, A. and MONTANARI, A. (2018). Debiasing the Lasso: Optimal sample size for Gaussian designs. *Ann. Statist*. **46** 2593–2622. MR3851749 https://doi.org/10.1214/17-AOS1630

KRISHNAN, A., ZHANG, R., YAO, V., THEESFELD, C. L., WONG, A. K., TADYCH, A., VOLFOVSKY, N., PACKER, A., LASH, A. et al. (2016). Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nat. Neurosci*. **19** 1454.

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

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist*. **34** 1436–1462. MR2278363 https://doi.org/10.1214/009053606000000281

NEGAHBAN, S. N., YU, B., WAINWRIGHT, M. J. and RAVIKUMAR, P. K. (2009). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems* 1348–1356.

NICKL, R. and VAN DE GEER, S. (2013). Confidence sets in sparse regression. *Ann. Statist*. **41** 2852–2876. MR3161450 https://doi.org/10.1214/13-AOS1170

RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inf. Theory* **57** 6976–6994. MR2882274 https://doi.org/10.1109/TIT.2011.2165799

ROBINS, J. and VAN DER VAART, A. (2006). Adaptive nonparametric confidence sets. *Ann. Statist*. **34** 229–253. MR2275241 https://doi.org/10.1214/009053605000000877

SHAH, R. D. and BÜHLMANN, P. (2018). Goodness-of-fit tests for high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **80** 113–135. MR3744714 https://doi.org/10.1111/rssb.12234

VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist*. **42** 1166–1202. MR3224285 https://doi.org/10.1214/14-AOS1221

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

VERHOEF, P. C., STEPHEN, A. T., KANNAN, P., LUO, X., ABHISHEK, V., ANDREWS, M., BART, Y., DATTA, H., FONG, N. et al. (2017). Consumer connectivity in a complex, technology-enabled, and mobile-oriented world with smart products. *J. Interact. Mark*. **40** 1–8.

WAINWRIGHT, M. J., LAFFERTY, J. D. and RAVIKUMAR, P. K. (2007). High-dimensional graphical model selection using l_1-regularized logistic regression. In *Advances in Neural Information Processing Systems* 1465–1472.

ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **76** 217–242. MR3153940 https://doi.org/10.1111/rssb.12026

ZHU, Y. and BRADIC, J. (2018a). Significance testing in non-sparse high-dimensional linear models. *Electron. J. Stat*. **12** 3312–3364. MR3861831 https://doi.org/10.1214/18-EJS1443

ZHU, Y. and BRADIC, J. (2018b). Linear hypothesis testing in dense high-dimensional linear models. *J. Amer. Statist. Assoc.* **113** 1583–1600. MR3902231 https://doi.org/10.1080/01621459.2017.1356319

The Annals of Statistics
2022, Vol. 50, No. 2, 640–672
https://doi.org/10.1214/21-AOS2059

ADAPTIVE ESTIMATION IN MULTIVARIATE RESPONSE REGRESSION WITH HIDDEN VARIABLES

BY XIN BING[a], YANG NING[b] AND YAOSHENG XU[c]

Department of Statistics and Data Science, Cornell University, [a]*xb43@cornell.edu,* [b]*yn265@cornell.edu,* [c]*yx433@cornell.edu*

A prominent concern of scientific investigators is the presence of unobserved hidden variables in association analysis. Ignoring hidden variables often yields biased statistical results and misleading scientific conclusions. Motivated by this practical issue, this paper studies the multivariate response regression with hidden variables, $Y = (\Psi^*)^T X + (B^*)^T Z + E$, where $Y \in \mathbb{R}^m$ is the response vector, $X \in \mathbb{R}^p$ is the observable feature, $Z \in \mathbb{R}^K$ represents the vector of unobserved hidden variables, possibly correlated with X, and E is an independent error. The number of hidden variables K is unknown and both m and p are allowed, but not required, to grow with the sample size n.

Though Ψ^* is shown to be nonidentifiable due to the presence of hidden variables, we propose to identify the projection of Ψ^* onto the orthogonal complement of the row space of B^*, denoted by Θ^*. The quantity $(\Theta^*)^T X$ measures the effect of X on Y that cannot be explained through the hidden variables, and thus Θ^* is treated as the parameter of interest. Motivated by the identifiability proof, we propose a novel estimation algorithm for Θ^*, called HIVE, under homoscedastic errors. The first step of the algorithm estimates the best linear prediction of Y given X, in which the unknown coefficient matrix exhibits an additive decomposition of Ψ^* and a dense matrix due to the correlation between X and Z. Under the sparsity assumption on Ψ^*, we propose to minimize a penalized least squares loss by regularizing Ψ^* and the dense matrix via group-lasso and multivariate ridge, respectively. Nonasymptotic deviation bounds of the in-sample prediction error are established. Our second step estimates the row space of B^* by leveraging the covariance structure of the residual vector from the first step. In the last step, we estimate Θ^* via projecting Y onto the orthogonal complement of the estimated row space of B^* to remove the effect of hidden variables. Nonasymptotic error bounds of our final estimator of Θ^*, which are valid for any m, p, K and n, are established. We further show that, under mild assumptions, the rate of our estimator matches the best possible rate with known B^* and is adaptive to the unknown sparsity of Θ^* induced by the sparsity of Ψ^*. The model identifiability, estimation algorithm and statistical guarantees are further extended to the setting with heteroscedastic errors. Thorough numerical simulations and two real data examples are provided to back up our theoretical results.

REFERENCES

[1] AHN, S. C. and HORENSTEIN, A. R. (2013). Eigenvalue ratio test for the number of factors. *Econometrica* **81** 1203–1227. MR3064065 https://doi.org/10.3982/ECTA8968

[2] ALDRIN, M. (1996). Moderate projection pursuit regression for multivariate response data. *Comput. Statist. Data Anal.* **21** 501–531. MR1395234 https://doi.org/10.1016/0167-9473(94)00029-8

[3] ANDERSON, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd ed. *Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. Wiley, New York. MR0771294

[4] 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. 62H12, 62J07.

Key words and phrases. High-dimensional models, multivariate response regression, shrinkage, nonsparse estimation, hidden variables, confounding, surrogate variable analysis.

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

[6] Bai, Z. D. and Yin, Y. Q. (1993). Limit of the smallest eigenvalue of a large-dimensional sample covariance matrix. *Ann. Probab.* **21** 1275–1294. MR1235416

[7] Belloni, A., Chernozhukov, V. and Wang, L. (2014). Pivotal estimation via square-root Lasso in nonparametric regression. *Ann. Statist.* **42** 757–788. MR3210986 https://doi.org/10.1214/14-AOS1204

[8] Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469 https://doi.org/10.1214/08-AOS620

[9] Bing, X., Ning, Y. and Xu, Y. (2022). Supplement to "Adaptive estimation in multivariate response regression with hidden variables." https://doi.org/10.1214/21-AOS2059SUPP

[10] Bing, X. and Wegkamp, M. H. (2019). Adaptive estimation of the rank of the coefficient matrix in high-dimensional multivariate response regression models. *Ann. Statist.* **47** 3157–3184. MR4025738 https://doi.org/10.1214/18-AOS1774

[11] Bloom, J. S., Ehrenreich, I. M., Loo, W. T., Lite, T.-L. V. and Kruglyak, L. (2013). Finding the sources of missing heritability in a yeast cross. *Nature* **494** 234–237.

[12] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*. *Springer Series in Statistics*. Springer, Heidelberg. MR2807761 https://doi.org/10.1007/978-3-642-20192-9

[13] Buja, A. and Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivar. Behav. Res.* **27** 509–540. https://doi.org/10.1207/s15327906mbr2704_2

[14] Bunea, F., Lederer, J. and She, Y. (2014). The group square-root lasso: Theoretical properties and fast algorithms. *IEEE Trans. Inf. Theory* **60** 1313–1325. MR3164977 https://doi.org/10.1109/TIT.2013.2290040

[15] Bunea, F., She, Y. and Wegkamp, M. H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann. Statist.* **39** 1282–1309. MR2816355 https://doi.org/10.1214/11-AOS876

[16] Bunea, F., She, Y. and Wegkamp, M. H. (2012). Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *Ann. Statist.* **40** 2359–2388. MR3097606 https://doi.org/10.1214/12-AOS1039

[17] Bunea, F., Strimas-Mackey, S. and Wegkamp, M. (2020). Interpolation under latent factor regression models.

[18] Candès, E. J., Li, X., Ma, Y. and Wright, J. (2011). Robust principal component analysis? *J. ACM* **58** 11:1–11:37. MR2811000 https://doi.org/10.1145/1970392.1970395

[19] Candès, E. J. and Tao, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inf. Theory* **56** 2053–2080. MR2723472 https://doi.org/10.1109/TIT.2010.2044061

[20] Ćevid, D., Bühlmann, P. and Meinshausen, N. (2020). Spectral deconfounding via perturbed sparse linear models. *J. Mach. Learn. Res.* **21** Paper No. 232. MR4209518 https://doi.org/10.22405/2226-8383-2020-21-1-221-232

[21] Chakraborty, S., Datta, S. and Datta, S. (2012). Surrogate variable analysis using partial least squares (SVA-PLS) in gene expression studies. *Bioinformatics* **28** 799–806. https://doi.org/10.1093/bioinformatics/bts022

[22] Chandrasekaran, V., Parrilo, P. A. and Willsky, A. S. (2012). Latent variable graphical model selection via convex optimization. *Ann. Statist.* **40** 1935–1967. MR3059067 https://doi.org/10.1214/11-AOS949

[23] Chernozhukov, V., Hansen, C. and Liao, Y. (2017). A lava attack on the recovery of sums of dense and sparse signals. *Ann. Statist.* **45** 39–76. MR3611486 https://doi.org/10.1214/16-AOS1434

[24] Diaz, E. (2017). Causality and surrogate variable analysis. Available at arXiv:1704.00588.

[25] Dicker, L. H. (2016). Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli* **22** 1–37. MR3449775 https://doi.org/10.3150/14-BEJ609

[26] Fan, J., Liao, Y. and Mincheva, M. (2011). High-dimensional covariance matrix estimation in approximate factor models. *Ann. Statist.* **39** 3320–3356. MR3012410 https://doi.org/10.1214/11-AOS944

[27] Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 603–680. MR3091653 https://doi.org/10.1111/rssb.12016

[28] Fan, J., Xue, L. and Yao, J. (2017). Sufficient forecasting using factor models. *J. Econometrics* **201** 292–306. MR3717565 https://doi.org/10.1016/j.jeconom.2017.08.009

[29] Gagnon-Bartsch, J. A. and Speed, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics* **13** 539–552.

[30] Giraud, C. (2011). Low rank multivariate regression. *Electron. J. Stat.* **5** 775–799. MR2824816 https://doi.org/10.1214/11-EJS625

[31] Guo, Z., Ćevid, D. and Bühlmann, P. (2020). Doubly debiased lasso: High-dimensional inference under hidden confounding.

[32] HOUSEMAN, E. A., ACCOMANDO, W. P., KOESTLER, D. C., CHRISTENSEN, B. C., MARSIT, C. J., NELSON, H. H., WIENCKE, J. K. and KELSEY, K. T. (2012). Dna methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinform*. **13** 86. https://doi.org/10.1186/1471-2105-13-86

[33] HOX, J. J. and BECHGER, T. M. (1998). An introduction to structural equation modeling.

[34] HSU, D., KAKADE, S. M. and ZHANG, T. (2011). Robust matrix decomposition with sparse corruptions. *IEEE Trans*. *Inf*. *Theory* **57** 7221–7234. MR2883652 https://doi.org/10.1109/TIT.2011.2158250

[35] HSU, D., KAKADE, S. M. and ZHANG, T. (2014). Random design analysis of ridge regression. *Found*. *Comput*. *Math*. **14** 569–600. MR3201956 https://doi.org/10.1007/s10208-014-9192-1

[36] IZENMAN, A. J. (2008). *Modern Multivariate Statistical Techniques*: *Regression*, *Classification*, *and Manifold Learning*. *Springer Texts in Statistics*. Springer, New York. MR2445017 https://doi.org/10.1007/978-0-387-78189-1

[37] LAM, C. and YAO, Q. (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. *Ann*. *Statist*. **40** 694–726. MR2933663 https://doi.org/10.1214/12-AOS970

[38] LEE, S., SUN, W., WRIGHT, F. A. and ZOU, F. (2017). An improved and explicit surrogate variable analysis procedure by coefficient adjustment. *Biometrika* **104** 303–316. MR3698255 https://doi.org/10.1093/biomet/asx018

[39] LEEK, J. T. and STOREY, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*. **3** 1724.

[40] LEEK, J. T. and STOREY, J. D. (2008). A general framework for multiple testing dependence. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **105** 18718–18723. https://doi.org/10.1073/pnas.0808709105

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

[42] MCKENNAN, C. and NICOLAE, D. (2019). Accounting for unobserved covariates with varying degrees of estimability in high-dimensional biological data. *Biometrika* **106** 823–840. MR4031201 https://doi.org/10.1093/biomet/asz037

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

[44] RUDELSON, M. and ZHOU, S. (2013). Reconstruction from anisotropic random measurements. *IEEE Trans*. *Inf. Theory* **59** 3434–3447. MR3061256 https://doi.org/10.1109/TIT.2013.2243201

[45] SUN, Y., ZHANG, N. R. and OWEN, A. B. (2012). Multiple hypothesis testing adjusted for latent variables, with an application to the AGEMAP gene expression data. *Ann*. *Appl*. *Stat*. **6** 1664–1688. MR3058679 https://doi.org/10.1214/12-AOAS561

[46] TESCHENDORFF, A. E., ZHUANG, J. and WIDSCHWENDTER, M. (2011). Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. *Bioinformatics* **27** 1496–1505. https://doi.org/10.1093/bioinformatics/btr171

[47] VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge Univ. Press, Cambridge. MR2963170

[48] WANG, J., ZHAO, Q., HASTIE, T. and OWEN, A. B. (2017). Confounder adjustment in multiple hypothesis testing. *Ann*. *Statist*. **45** 1863–1894. MR3718155 https://doi.org/10.1214/16-AOS1511

[49] YU, Y., WANG, T. and SAMWORTH, R. J. (2015). A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika* **102** 315–323. MR3371006 https://doi.org/10.1093/biomet/asv008

[50] YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **68** 49–67. MR2212574 https://doi.org/10.1111/j.1467-9868.2005.00532.x

[51] ZHANG, A., CAI, T. T. and WU, Y. (2018). *Heteroskedastic Pca*: *Algorithm*, *Optimality*, *and Applications*.

The Annals of Statistics
2022, Vol. 50, No. 2, 673–697
https://doi.org/10.1214/21-AOS2122

REFINED CRAMÉR-TYPE MODERATE DEVIATION THEOREMS FOR GENERAL SELF-NORMALIZED SUMS WITH APPLICATIONS TO DEPENDENT RANDOM VARIABLES AND WINSORIZED MEAN

BY LAN GAO[1,3,a], QI-MAN SHAO[2,c] AND JIASHENG SHI[1,4,b]

[1]*Department of Statistics, Chinese University of Hong Kong,* [a]*lgao@link.cuhk.edu.hk,* [b]*jiashengSHI@link.cuhk.edu.hk*

[2]*Department of Statistics and Data Science, SICM, NCAMS, Southern University of Science and Technology,* [c]*shaoqm@sustech.edu.cn*

[3]*Data Sciences and Operations Department, University of Southern California*

[4]*Department of Biostatistics, Epidemiology and informatics, University of Pennsylvania*

Let $\{(X_i, Y_i)\}_{i=1}^n$ be a sequence of independent bivariate random vectors. In this paper, we establish a refined Cramér-type moderate deviation theorem for the general self-normalized sum $\sum_{i=1}^n X_i/(\sum_{i=1}^n Y_i^2)^{1/2}$, which unifies and extends the classical Cramér (*Actual. Sci. Ind.* **736** (1938) 5–23) theorem and the self-normalized Cramér-type moderate deviation theorems by Jing, Shao and Wang (*Ann. Probab.* **31** (2003) 2167–2215) as well as the further refined version by Wang (*J. Theoret. Probab.* **24** (2011) 307–329). The advantage of our result is evidenced through successful applications to weakly dependent random variables and self-normalized winsorized mean. Specifically, by applying our new framework on general self-normalized sum, we significantly improve Cramér-type moderate deviation theorems for one-dependent random variables, geometrically β-mixing random variables and causal processes under geometrical moment contraction. As an additional application, we also derive the Cramér-type moderate deviation theorems for self-normalized winsorized mean.

REFERENCES

BENTKUS, V., BLOZNELIS, M. and GÖTZE, F. (1996). A Berry–Esséen bound for Student's statistic in the non-i.i.d. case. *J. Theoret. Probab.* **9** 765–796. MR1400598 https://doi.org/10.1007/BF02214086

BENTKUS, V. and GÖTZE, F. (1996). The Berry–Esseen bound for Student's statistic. *Ann. Probab.* **24** 491–503. MR1387647 https://doi.org/10.1214/aop/1042644728

BERBEE, H. (1987). Convergence rates in the strong law for bounded mixing sequences. *Probab. Theory Related Fields* **74** 255–270. MR0871254 https://doi.org/10.1007/BF00569992

CHANG, J., TANG, C. Y. and WU, Y. (2016). Local independence feature screening for nonparametric and semiparametric models by marginal empirical likelihood. *Ann. Statist.* **44** 515–539. MR3476608 https://doi.org/10.1214/15-AOS1374

CHEN, X., SHAO, Q.-M., WU, W. B. and XU, L. (2016). Self-normalized Cramér-type moderate deviations under dependence. *Ann. Statist.* **44** 1593–1617. MR3519934 https://doi.org/10.1214/15-AOS1429

CRAMÉR, H. (1938). Sur un nouveau théoreme-limite de la théorie des probabilités. *Actual. Sci. Ind.* **736** 5–23.

DELAIGLE, A. and HALL, P. (2009). Higher criticism in the context of unknown distribution, non-independence and classification. In *Perspectives in Mathematical Sciences. I. Stat. Sci. Interdiscip. Res.* **7** 109–138. World Sci. Publ., Hackensack, NJ. MR2581742 https://doi.org/10.1142/9789814273633_0006

DE LA PEÑA, V. H., LAI, T. L. and SHAO, Q.-M. (2009). *Self-Normalized Processes: Limit Theory and Statistical Applications. Probability and Its Applications (New York)*. Springer, Berlin. MR2488094 https://doi.org/10.1007/978-3-540-85636-8

DIXON, W. J. (1960). Simplified estimation from censored normal samples. *Ann. Math. Stat.* **31** 385–391. MR0119358 https://doi.org/10.1214/aoms/1177705900

FAN, J. and FAN, Y. (2008). High-dimensional classification using features annealed independence rules. *Ann. Statist.* **36** 2605–2637. MR2485009 https://doi.org/10.1214/07-AOS504

MSC2020 subject classifications. Primary 62E17, 62E20; secondary 62J07.

Key words and phrases. Moderate deviation, self-normalized sums, dependent random variables, high dimensional, Winsorized mean.

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

GAO, L., SHAO, Q.-M. and SHI, J. (2022). Supplement to "Refined Cramér-type moderate deviation theorems for general self-normalized sums with applications to dependent random variables and winsorized mean." https://doi.org/10.1214/21-AOS2122SUPP

GINÉ, E., GÖTZE, F. and MASON, D. M. (1997). When is the Student t-statistic asymptotically standard normal? *Ann. Probab.* **25** 1514–1531. MR1457629 https://doi.org/10.1214/aop/1024404523

HALL, P. (1987). Edgeworth expansion for Student's t statistic under minimal moment conditions. *Ann. Probab.* **15** 920–931. MR0893906

HSING, T. and WU, W. B. (2004). On weighted U-statistics for stationary processes. *Ann. Probab.* **32** 1600–1631. MR2060311 https://doi.org/10.1214/009117904000000333

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

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

JING, B.-Y., SHAO, Q.-M. and WANG, Q. (2003). Self-normalized Cramér-type large deviations for independent random variables. *Ann. Probab.* **31** 2167–2215. MR2016616 https://doi.org/10.1214/aop/1068646382

LIN, Z. and BAI, Z. (2010). *Probability Inequalities*. Science Press Beijing, Beijing; Springer, Heidelberg. MR2789096

LIU, W. and SHAO, Q.-M. (2010). Cramér-type moderate deviation for the maximum of the periodogram with application to simultaneous tests in gene expression time series. *Ann. Statist.* **38** 1913–1935. MR2662363 https://doi.org/10.1214/09-AOS774

LIU, W. and SHAO, Q.-M. (2013). A Carmér moderate deviation theorem for Hotelling's T^2-statistic with applications to global tests. *Ann. Statist.* **41** 296–322. MR3059419 https://doi.org/10.1214/12-AOS1082

ROTHENBERG, T. J., FISHER, F. M. and TILANUS, C. B. (1964). A note on estimation from a Cauchy sample. *J. Amer. Statist. Assoc.* **59** 460–463. MR0166872

SHAO, Q.-M. (1999). A Cramér type large deviation result for Student's t-statistic. *J. Theoret. Probab.* **12** 385–398. MR1684750 https://doi.org/10.1023/A:1021626127372

SHAO, Q. (2018). On necessary and sufficient conditions for the self-normalized central limit theorem. *Sci. China Math.* **61** 1741–1748. MR3856964 https://doi.org/10.1007/s11425-018-9368-3

SHAO, Q.-M. and YU, H. (1996). Weak convergence for weighted empirical processes of dependent sequences. *Ann. Probab.* **24** 2098–2127. MR1415243 https://doi.org/10.1214/aop/1041903220

SHAO, Q.-M. and ZHOU, W.-X. (2016). Cramér type moderate deviation theorems for self-normalized processes. *Bernoulli* **22** 2029–2079. MR3498022 https://doi.org/10.3150/15-BEJ719

SHERGIN, V. (1979). The rate of convergence in the central limit theorem for m-dependent random variables. *Theory Probab. Appl.* **24** 782–796.

WANG, Q. (2011). Refined self-normalized large deviations for independent random variables. *J. Theoret. Probab.* **24** 307–329. MR2795041 https://doi.org/10.1007/s10959-011-0347-6

WANG, Q. and JING, B.-Y. (1999). An exponential nonuniform Berry–Esseen bound for self-normalized sums. *Ann. Probab.* **27** 2068–2088. MR1742902 https://doi.org/10.1214/aop/1022677562

WU, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proc. Natl. Acad. Sci. USA* **102** 14150–14154. MR2172215 https://doi.org/10.1073/pnas.0506715102

WU, W. B. (2011). Asymptotic theory for stationary processes. *Stat. Interface* **4** 207–226. MR2812816 https://doi.org/10.4310/SII.2011.v4.n2.a15

WU, W. B. and SHAO, X. (2004). Limit theorems for iterated random functions. *J. Appl. Probab.* **41** 425–436. MR2052582 https://doi.org/10.1239/jap/1082999076

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

The Annals of Statistics
2022, Vol. 50, No. 2, 698–725
https://doi.org/10.1214/21-AOS2124

DISTRIBUTED NONPARAMETRIC FUNCTION ESTIMATION: OPTIMAL RATE OF CONVERGENCE AND COST OF ADAPTATION

BY T. TONY CAI[a] AND HONGJI WEI[b]

Department of Statistics and Data Science, The Wharton School, University Of Pennsylvania, [a]*tcai@wharton.upenn.edu,* [b]*hongjiw@wharton.upenn.edu*

Distributed minimax estimation and distributed adaptive estimation under communication constraints for Gaussian sequence model and white noise model are studied. The minimax rate of convergence for distributed estimation over a given Besov class, which serves as a benchmark for the cost of adaptation, is established. We then quantify the exact communication cost for adaptation and construct an optimally adaptive procedure for distributed estimation over a range of Besov classes.

The results demonstrate significant differences between nonparametric function estimation in the distributed setting and the conventional centralized setting. For global estimation, adaptation in general cannot be achieved for free in the distributed setting. The new technical tools to obtain the exact characterization for the cost of adaptation can be of independent interest.

REFERENCES

ABRAMOVICH, F., BENJAMINI, Y., DONOHO, D. L. and JOHNSTONE, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* **34** 584–653. MR2281879 https://doi.org/10.1214/009053606000000074

BARNES, L. P., HAN, Y. and ÖZGÜR, A. (2019). Learning distributions from their samples under communication constraints. *CoRR*. Available at 1902.02890.

BATTEY, H., FAN, J., LIU, H., LU, J. and ZHU, Z. (2018). Distributed testing and estimation under sparse high dimensional models. *Ann. Statist.* **46** 1352–1382. MR3798006 https://doi.org/10.1214/17-AOS1587

BLAHUT, R. E. and BLAHUT, R. E. (1987). *Principles and Practice of Information Theory* **1**. Addison-Wesley, Reading, MA.

BRAVERMAN, M., GARG, A., MA, T., NGUYEN, H. L. and WOODRUFF, D. P. (2016). Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *STOC'*16—*Proceedings of the* 48*th Annual ACM SIGACT Symposium on Theory of Computing* 1011–1020. ACM, New York. MR3536632 https://doi.org/10.1145/2897518.2897582

BROWN, L. D. and LOW, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24** 2384–2398. MR1425958 https://doi.org/10.1214/aos/1032181159

BROWN, L. D., CAI, T. T., LOW, M. G. and ZHANG, C.-H. (2002). Asymptotic equivalence theory for nonparametric regression with random design. *Ann. Statist.* **30** 688–707. MR1922538 https://doi.org/10.1214/aos/1028674838

BROWN, L. D., CARTER, A. V., LOW, M. G. and ZHANG, C.-H. (2004). Equivalence theory for density estimation, Poisson processes and Gaussian white noise with drift. *Ann. Statist.* **32** 2074–2097. MR2102503 https://doi.org/10.1214/009053604000000012

BROWN, L., CAI, T., ZHANG, R., ZHAO, L. and ZHOU, H. (2010). The root-unroot algorithm for density estimation as implemented via wavelet block thresholding. *Probab. Theory Related Fields* **146** 401–433. MR2574733 https://doi.org/10.1007/s00440-008-0194-2

CAI, T. T. (1999). Adaptive wavelet estimation: A block thresholding and oracle inequality approach. *Ann. Statist.* **27** 898–924. MR1724035 https://doi.org/10.1214/aos/1018031262

CAI, T. T. (2008). On information pooling, adaptability and superefficiency in nonparametric function estimation. *J. Multivariate Anal.* **99** 421–436. MR2396972 https://doi.org/10.1016/j.jmva.2006.11.010

CAI, T. T. and WEI, H. (2020a). Distributed gaussian mean estimation under communication constraints: Optimal rates and communication-efficient algorithms. ArXiv preprint. Available at arXiv:2001.08877.

MSC2020 subject classifications. Primary 62F30; secondary 62B10, 62F12.

Key words and phrases. Communication constraints, distributed learning, nonparametric regression, optimal rate of convergence, adaptation.

CAI, T. T. and WEI, H. (2022). Supplement to "Distributed nonparametric function estimation: Optimal rate of convergence and cost of adaptation." https://doi.org/10.1214/21-AOS2124SUPP

CAI, T. T. and ZHOU, H. H. (2009). A data-driven block thresholding approach to wavelet estimation. *Ann. Statist.* **37** 569–595. MR2502643 https://doi.org/10.1214/07-AOS538

DEISENROTH, M. and NG, J. W. (2015). Distributed Gaussian processes. In *Proceedings of the* 32*nd International Conference on Machine Learning* (F. Bach and D. Blei, eds.). *Proceedings of Machine Learning Research* **37** 1481–1490. PMLR, Lille, France.

DIAKONIKOLAS, I., GRIGORESCU, E., LI, J., NATARAJAN, A., ONAK, K. and SCHMIDT, L. (2017). Communication-efficient distributed learning of discrete distributions. In *Advances in Neural Information Processing Systems* 6391–6401.

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

DONOHO, D. L. and JOHNSTONE, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90** 1200–1224. MR1379464

DONOHO, D. L. and JOHNSTONE, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26** 879–921. MR1635414 https://doi.org/10.1214/aos/1024691081

FAN, J., WANG, D., WANG, K. and ZHU, Z. (2019). Distributed estimation of principal eigenspaces. *Ann. Statist.* **47** 3009–3031. MR4025733 https://doi.org/10.1214/18-AOS1713

GARG, A., MA, T. and NGUYEN, H. (2014). On communication cost of distributed statistical estimation and dimensionality. In *Advances in Neural Information Processing Systems* 2726–2734.

HALL, P., KERKYACHARIAN, G. and PICARD, D. (1999). On the minimax optimality of block thresholded wavelet estimators. *Statist. Sinica* **9** 33–49. MR1678880

HAN, Y., ÖZGÜR, A. and WEISSMAN, T. (2021). Geometric lower bounds for distributed parameter estimation under communication constraints. *IEEE Trans. Inf. Theory* **67** 8248–8263. MR4346086 https://doi.org/10.1109/TIT.2021.3108952

IBRAGIMOV, I. and KHASMINSKII, R. (1997). Some estimation problems in infinite-dimensional Gaussian white noise. In *Festschrift for Lucien Le Cam* 259–274. Springer, New York. MR1462950

JOHNSTONE, I. M. (2017). Gaussian Estimation: Sequence and Wavelet Models. Manuscript.

JORDAN, M. I., LEE, J. D. and YANG, Y. (2019). Communication-efficient distributed statistical inference. *J. Amer. Statist. Assoc.* **114** 668–681. MR3963171 https://doi.org/10.1080/01621459.2018.1429274

KLEINER, A., TALWALKAR, A., SARKAR, P. and JORDAN, M. I. (2014). A scalable bootstrap for massive data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 795–816. MR3248677 https://doi.org/10.1111/rssb.12050

LEE, J. D., LIU, Q., SUN, Y. and TAYLOR, J. E. (2017). Communication-efficient sparse regression. *J. Mach. Learn. Res.* **18** Paper No. 5. MR3625709

MEYER, Y. (1992). *Wavelets and Operators*. *Cambridge Studies in Advanced Mathematics* **37**. Cambridge Univ. Press, Cambridge. MR1228209

MICHALOWICZ, J., NICHOLS, J. and BUCHOLTZ, F. (2008). Calculation of differential entropy for a mixed Gaussian distribution. *Entropy* **10** 200–206.

NUSSBAUM, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.* **24** 2399–2430. MR1425959 https://doi.org/10.1214/aos/1032181160

SHANNON, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* **27** 379–423. MR0026286 https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

SZABÓ, B. and VAN ZANTEN, H. (2020). Adaptive distributed methods under communication constraints. *Ann. Statist.* **48** 2347–2380. MR4134798 https://doi.org/10.1214/19-AOS1890

SZABO, B. and VAN ZANTEN, H. (2020). Distributed function estimation: Adaptation using minimal communication. ArXiv preprint. Available at arXiv:2003.12838.

ZHANG, Y., DUCHI, J., JORDAN, M. I. and WAINWRIGHT, M. J. (2013). Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *Advances in Neural Information Processing Systems* 2328–2336.

ZHU, Y. and LAFFERTY, J. (2018). Distributed nonparametric regression under communication constraints. ArXiv preprint. Available at arXiv:1803.01302.

The Annals of Statistics
2022, Vol. 50, No. 2, 726–753
https://doi.org/10.1214/21-AOS2125

EDGEWORTH EXPANSIONS FOR NETWORK MOMENTS

BY YUAN ZHANG[1,a] AND DONG XIA[2,b]

[1]*Department of Statistics, The Ohio State University,* [a]*yzhanghf@stat.osu.edu*
[2]*Department of Mathematics, The Hong Kong University of Science and Technology,* [b]*madxia@ust.hk*

Network method of moments (*Ann. Statist.* **39** (2011) 2280–2301) is an important tool for nonparametric network inference. However, there has been little investigation on accurate descriptions of the sampling distributions of network moment statistics. In this paper, we present the first higher-order accurate approximation to the sampling CDF of a studentized network moment by Edgeworth expansion. In sharp contrast to classical literature on *noiseless* U-statistics, we show that the Edgeworth expansion of a network moment statistic as a *noisy* U-statistic can achieve higher-order accuracy without non-lattice or smoothness assumptions but just requiring weak regularity conditions. Behind this result is our surprising discovery that the two typically-hated factors in network analysis, namely, sparsity and edgewise observational errors, jointly play a blessing role, contributing a crucial *self-smoothing* effect in the network moment statistic and making it analytically tractable. Our assumptions match the minimum requirements in related literature. For sparse networks, our theory shows that our empirical Edgeworth expansion and a simple normal approximation both achieve the same gradually depreciating Berry–Esseen-type bound as the network becomes sparser. This result also significantly refines the best previous theoretical result.

For practitioners, our empirical Edgeworth expansion is highly accurate and computationally efficient. It is also easy to implement and convenient for parallel computing. We demonstrate the clear advantage of our method by several comprehensive simulation studies. As a byproduct, we also provide a finite-sample analysis of the network jackknife.

We showcase three applications of our results in network inference. We prove, to our knowledge, the first theoretical guarantee of higher-order accuracy for some network bootstrap schemes, and moreover, the first theoretical guidance for selecting the subsample size for network subsampling. We also derive a one-sample test and the Cornish–Fisher confidence interval for a given moment with higher-order accurate controls of confidence level and type I error, respectively.

REFERENCES

[1] ABBE, E. (2017). Community detection and stochastic block models: Recent developments. *J. Mach. Learn. Res.* **18** Paper No. 177, 86 pp. MR3827065

[2] AHMED, N. K., NEVILLE, J., ROSSI, R. A. and DUFFIELD, N. (2015). Efficient graphlet counting for large networks. In 2015 *IEEE International Conference on Data Mining* 1–10. IEEE.

[3] AIROLDI, E. M., BLEI, D. M., FIENBERG, S. E. and XING, E. P. (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9** 1981–2014.

[4] ALDOUS, D. J. (1981). Representations for partially exchangeable arrays of random variables. *J. Multivariate Anal.* **11** 581–598. MR0637937 https://doi.org/10.1016/0047-259X(81)90099-3

[5] ALI, W., WEGNER, A. E., GAUNT, R. E., DEANE, C. M. and REINERT, G. (2016). Comparison of large networks with sub-sampling strategies. *Sci. Rep.* **6** 28955. https://doi.org/10.1038/srep28955

[6] ALON, U. (2007). Network motifs: Theory and experimental approaches. *Nat. Rev. Genet.* **8** 450–461.

[7] AL HASAN, M. and DAVE, V. S. (2018). Triangle counting in large networks: A review. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **8** e1226.

MSC2020 subject classifications. Primary 91D30; secondary 60F05, 62E17.
Key words and phrases. Network inference, method of moments, Edgeworth expansion, noisy U-statistic, network bootstrap, network jackknife.

[8] AMBROISE, C. and MATIAS, C. (2012). New consistent and asymptotically normal parameter estimates for random-graph mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 3–35. MR2885838 https://doi.org/10.1111/j.1467-9868.2011.01009.x

[9] ANGST, J. and POLY, G. (2017). A weak Cramér condition and application to Edgeworth expansions. *Electron. J. Probab.* **22** Paper No. 59, 24 pp. MR3683368 https://doi.org/10.1214/17-EJP77

[10] BABU, G. J. and SINGH, K. (1984). On one term Edgeworth correction by Efron's bootstrap. *Sankhyā Ser. A* **46** 219–232. MR0778872

[11] BABU, G. J. and SINGH, K. (1985). Edgeworth expansions for sampling without replacement from finite populations. *J. Multivariate Anal.* **17** 261–278. MR0813236 https://doi.org/10.1016/0047-259X(85)90084-3

[12] BABU, G. J. and SINGH, K. (1989). On Edgeworth expansions in the mixture cases. *Ann. Statist.* **17** 443–447. MR0981463 https://doi.org/10.1214/aos/1176347029

[13] BENTKUS, V., GÖTZE, F. and VAN ZWET, W. R. (1997). An Edgeworth expansion for symmetric statistics. *Ann. Statist.* **25** 851–896. MR1439326 https://doi.org/10.1214/aos/1031833676

[14] BENTKUS, V., GÖTZE, F. and ZITIKIS, R. (1994). Lower estimates of the convergence rate for U-statistics. *Ann. Probab.* **22** 1707–1714. MR1331199

[15] BENTKUS, V., JING, B.-Y. and ZHOU, W. (2009). On normal approximations to U-statistics. *Ann. Probab.* **37** 2174–2199. MR2573555 https://doi.org/10.1214/09-AOP474

[16] BHATTACHARYA, R. N. and GHOSH, J. K. (1978). On the validity of the formal Edgeworth expansion. *Ann. Statist.* **6** 434–451. MR0471142

[17] BHATTACHARYYA, S. and BICKEL, P. J. (2015). Subsampling bootstrap of count features of networks. *Ann. Statist.* **43** 2384–2411. MR3405598 https://doi.org/10.1214/15-AOS1338

[18] BICKEL, P. J. (1974). Edgeworth expansions in non parametric statistics. *Ann. Statist.* **2** 1–20. MR0350952

[19] BICKEL, P. J. and CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan modularities. *Proc. Natl. Acad. Sci. USA* **106** 21068–21073.

[20] BICKEL, P. J., CHEN, A. and LEVINA, E. (2011). The method of moments and degree distributions for network models. *Ann. Statist.* **39** 2280–2301. MR2906868 https://doi.org/10.1214/11-AOS904

[21] BICKEL, P. J., GÖTZE, F. and VAN ZWET, W. R. (1986). The Edgeworth expansion for U-statistics of degree two. *Ann. Statist.* **14** 1463–1484. MR0868312 https://doi.org/10.1214/aos/1176350170

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

[23] BLOZNELIS, M. (2003). An Edgeworth expansion for Studentized finite population statistics. *Acta Appl. Math.* **78** 51–60.

[24] BLOZNELIS, M. and GÖTZE, F. (2001). Orthogonal decomposition of finite population statistics and its applications to distributional asymptotics. *Ann. Statist.* **29** 899–917. MR1865345 https://doi.org/10.1214/aos/1009210694

[25] BLOZNELIS, M. and GÖTZE, F. (2002). An Edgeworth expansion for symmetric finite population statistics. *Ann. Probab.* **30** 1238–1265. MR1920107 https://doi.org/10.1214/aop/1029867127

[26] BORGS, C., CHAYES, J. and LOVÁSZ, L. (2010). Moments of two-variable functions and the uniqueness of graph limits. *Geom. Funct. Anal.* **19** 1597–1619. MR2594615 https://doi.org/10.1007/s00039-010-0044-0

[27] BROIDO, A. D. and CLAUSET, A. (2019). Scale-free networks are rare. *Nat. Commun.* **10** 1–10.

[28] CAI, T. T. and LI, X. (2015). Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *Ann. Statist.* **43** 1027–1059. MR3346696 https://doi.org/10.1214/14-AOS1290

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

[30] CALLAERT, H. and JANSSEN, P. (1978). The Berry–Esseen theorem for U-statistics. *Ann. Statist.* **6** 417–421. MR0464359

[31] CALLAERT, H., JANSSEN, P. and VERAVERBEKE, N. (1980). An Edgeworth expansion for U-statistics. *Ann. Statist.* **8** 299–312. MR0560731

[32] CALLAERT, H. and VERAVERBEKE, N. (1981). The order of the normal approximation for a studentized U-statistic. *Ann. Statist.* **9** 194–200. MR0600547

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

[34] CHAN, S. and AIROLDI, E. (2014). A consistent histogram estimator for exchangeable graph models. In *International Conference on Machine Learning* 208–216.

[35] CHEN, S. and ONNELA, J.-P. (2019). A bootstrap method for goodness of fit and model selection with a single observed network. *Sci. Rep.* **9** 1–12.

[36] CHEN, S. X. and PENG, L. (2021). Distributed statistical inference for massive data. *Ann. Statist.* To appear.

[37] CHEN, X. and KATO, K. (2019). Randomized incomplete U-statistics in high dimensions. *Ann. Statist.* **47** 3127–3156. MR4025737 https://doi.org/10.1214/18-AOS1773

[38] CHOI, D. (2017). Co-clustering of nonsmooth graphons. *Ann. Statist.* **45** 1488–1515. MR3670186 https://doi.org/10.1214/16-AOS1497

[39] CHOI, D. S., WOLFE, P. J. and AIROLDI, E. M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika* **99** 273–284. MR2931253 https://doi.org/10.1093/biomet/asr053

[40] CHUNG, F. and LU, L. (2002). Connected components in random graphs with given expected degree sequences. *Ann. Comb.* **6** 125–145. MR1955514 https://doi.org/10.1007/PL00012580

[41] CLAUSET, A., NEWMAN, M. E. and MOORE, C. (2004). Finding community structure in very large networks. *Phys. Rev. E* **70** 066111.

[42] CORNISH, E. A. and FISHER, R. A. (1938). Moments and cumulants in the specification of distributions. *Rev. Inst. Int. Stat.* **5** 307–320.

[43] CRAMÉR, H. (1928). On the composition of elementary errors: First paper: Mathematical deductions. *Scand. Actuar. J.* **1928** 13–74.

[44] CRANE, H. (2018). *Probabilistic Foundations of Statistical Network Analysis. Monographs on Statistics and Applied Probability* **157**. CRC Press, Boca Raton, FL. MR3791467

[45] DAVISON, A. C. and HINKLEY, D. V. (1997). *Bootstrap Methods and Their Application. Cambridge Series in Statistical and Probabilistic Mathematics* **1**. Cambridge Univ. Press, Cambridge. MR1478673 https://doi.org/10.1017/CBO9780511802843

[46] DELLA ROSSA, F., DERCOLE, F. and PICCARDI, C. (2013). Profiling core-periphery network structure by random walkers. *Sci. Rep.* **3** 1–8.

[47] DIACONIS, P. and JANSON, S. (2008). Graph limits and exchangeable random graphs. *Rend. Mat. Appl.* (7) **28** 33–61. MR2463439

[48] EDGEWORTH, F. (1905). The law of error. In *Cambridge Philos. Soc.* **20** 36–66 and 113–141.

[49] FISHER, S. R. A. and CORNISH, E. (1960). The percentile points of distributions having known cumulants. *Technometrics* **2** 209–225.

[50] GAO, C. and LAFFERTY, J. (2017). Testing network structure using relations between small subgraph probabilities. Preprint. Available at arXiv:1704.06742.

[51] GAO, C. and LAFFERTY, J. (2017). Testing for global network structure using small subgraph statistics. Preprint. Available at arXiv:1710.00862.

[52] GAO, C., LU, Y., MA, Z. and ZHOU, H. H. (2016). Optimal estimation and completion of matrices with biclustering structures. *J. Mach. Learn. Res.* **17** Paper No. 161, 29 pp. MR3569248

[53] GAO, C., LU, Y. and ZHOU, H. H. (2015). Rate-optimal graphon estimation. *Ann. Statist.* **43** 2624–2652. MR3405606 https://doi.org/10.1214/15-AOS1354

[54] 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

[55] GAO, F., MA, Z. and YUAN, H. (2020). Community detection in sparse latent space models. Preprint. Available at arXiv:2008.01375.

[56] GEL, Y. R., LYUBCHICH, V. and RAMIREZ, L. L. R. (2017). Bootstrap quantification of estimation uncertainties in network degree distributions. *Sci. Rep.* **7** 1–12.

[57] GREEN, A. and SHALIZI, C. R. (2017). Bootstrapping exchangeable random graphs. Preprint. Available at arXiv:1711.00813.

[58] GROVER, A. and LESKOVEC, J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 855–864.

[59] GUÉDON, O. and VERSHYNIN, R. (2016). Community detection in sparse networks via Grothendieck's inequality. *Probab. Theory Related Fields* **165** 1025–1049. MR3520025 https://doi.org/10.1007/s00440-015-0659-z

[60] HAEUSLER, E. (1988). On the rate of convergence in the central limit theorem for martingales with discrete and continuous time. *Ann. Probab.* **16** 275–299. MR0920271

[61] HALL, P. (1983). Inverting an Edgeworth expansion. *Ann. Statist.* **11** 569–576. MR0696068 https://doi.org/10.1214/aos/1176346162

[62] HALL, P. (1987). Edgeworth expansion for Student's t statistic under minimal moment conditions. *Ann. Probab.* **15** 920–931. MR0893906

[63] HALL, P. (1993). On Edgeworth expansion and bootstrap confidence bands in nonparametric curve estimation. *J. Roy. Statist. Soc. Ser. B* **55** 291–304.

[64] HALL, P. (2013). *The Bootstrap and Edgeworth Expansion. Springer Series in Statistics*. Springer, New York. MR1145237 https://doi.org/10.1007/978-1-4612-4384-7

[65] HALL, P. and HEYDE, C. C. (2014). *Martingale Limit Theory and Its Application. Probability and Mathematical Statistics*. Academic Press, New York. MR0624435

[66] HALL, P. and JING, B.-Y. (1995). Uniform coverage bounds for confidence intervals and Berry–Esseen theorems for Edgeworth expansion. *Ann. Statist.* **23** 363–375. MR1332571 https://doi.org/10.1214/aos/1176324525

[67] HELMERS, R. (1991). On the Edgeworth expansion and the bootstrap approximation for a Studentized U-statistic. *Ann. Statist.* **19** 470–484. MR1091863 https://doi.org/10.1214/aos/1176347994

[68] HEYDE, C. C. and BROWN, B. M. (1970). On the departure from normality of a certain class of martingales. *Ann. Math. Stat.* **41** 2161–2165. MR0293702 https://doi.org/10.1214/aoms/1177696722

[69] 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

[70] HOFF, P. D., RAFTERY, A. E. and HANDCOCK, M. S. (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* **97** 1090–1098. MR1951262 https://doi.org/10.1198/016214502388618906

[71] HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. MR0718088 https://doi.org/10.1016/0378-8733(83)90021-7

[72] HOLLAND, P. W. and LEINHARDT, S. (1971). Transitivity in structural models of small groups. *Comp. Group Stud.* **2** 107–124.

[73] HOOVER, D. N. (1979). Relations on probability spaces and arrays of random variables. Preprint **2**, Institute for Advanced Study, Princeton, NJ.

[74] HUNTER, D. R., HANDCOCK, M. S., BUTTS, C. T., GOODREAU, S. M. and MORRIS, M. (2008). ergm: A package to fit, simulate and diagnose exponential-family models for networks. *J. Stat. Softw.* **24** nihpa54860. https://doi.org/10.18637/jss.v024.i03

[75] JANSSEN, P. (1981). Rate of convergence in the central limit theorem and in the strong law of large numbers for von Mises statistics. *Metrika* **28** 35–46. MR0614539 https://doi.org/10.1007/BF01902875

[76] JING, B.-Y., LI, T., LYU, Z. and XIA, D. (2020). Community detection on mixture multi-layer networks via regularized tensor decomposition. Preprint. Available at arXiv:2002.04457.

[77] JING, B.-Y. and WANG, Q. (2003). Edgeworth expansion for U-statistics under minimal conditions. *Ann. Statist.* **31** 1376–1391. MR2001653 https://doi.org/10.1214/aos/1059655916

[78] JING, B.-Y. and WANG, Q. (2010). A unified approach to Edgeworth expansions for a general class of statistics. *Statist. Sinica* **20** 613–636. MR2682633

[79] KARRER, B. and NEWMAN, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* (3) **83** 016107, 10 pp. MR2788206 https://doi.org/10.1103/PhysRevE.83.016107

[80] KLOPP, O., TSYBAKOV, A. B. and VERZELEN, N. (2017). Oracle inequalities for network models and sparse graphon estimation. *Ann. Statist.* **45** 316–354. MR3611494 https://doi.org/10.1214/16-AOS1454

[81] KOLASSA, J. E. and MCCULLAGH, P. (1990). Edgeworth series for lattice distributions. *Ann. Statist.* **18** 981–985. MR1056348 https://doi.org/10.1214/aos/1176347637

[82] KONG, X. and ZHENG, W. (2021). Design based incomplete U-statistics. *Statist. Sinica* **31** 1593–1618. MR4297718 https://doi.org/10.5705/ss.202019.0098

[83] LAHIRI, S. N. (1993). Bootstrapping the Studentized sample mean of lattice variables. *J. Multivariate Anal.* **45** 247–256. MR1221921 https://doi.org/10.1006/jmva.1993.1037

[84] LAI, T. L. and WANG, J. Q. (1993). Edgeworth expansions for symmetric statistics with applications to bootstrap methods. *Statist. Sinica* **3** 517–542. MR1243399

[85] LEI, J. (2016). A goodness-of-fit test for stochastic block models. *Ann. Statist.* **44** 401–424. MR3449773 https://doi.org/10.1214/15-AOS1370

[86] LESKOVEC, J. and KREVL, A. (2014). SNAP datasets: Stanford large network dataset collection. Available at http://snap.stanford.edu/data.

[87] LEVIN, K. and LEVINA, E. (2019). Bootstrapping networks with latent space structure. Preprint. Available at arXiv:1907.10821.

[88] LI, Y. and LI, H. (2018). Two-sample test of community memberships of weighted stochastic block models. Preprint. Available at arXiv:1811.12593.

[89] LOVÁSZ, L. (2009). Very large graphs. In *Current Developments in Mathematics*, 2008 67–128. Int. Press, Somerville, MA. MR2555927

[90] MAESONO, Y. (1997). Edgeworth expansions of a Studentized U-statistic and a jackknife estimator of variance. *J. Statist. Plann. Inference* **61** 61–84. MR1455802 https://doi.org/10.1016/S0378-3758(96)00148-6

[91] MATIAS, C., SCHBATH, S., BIRMELÉ, E., DAUDIN, J.-J. and ROBIN, S. (2006). Network motifs: Mean and variance for the count. *REVSTAT* **4** 31–51. MR2259363

[92] MATSUSHITA, Y. and OTSU, T. (2021). Jackknife empirical likelihood: Small bandwidth, sparse network and high-dimensional asymptotics. *Biometrika* **108** 661–674. MR4298770 https://doi.org/10.1093/biomet/asaa081

[93] MATTNER, L. and SHEVTSOVA, I. G. (2017). An optimal Berry–Esseen type inequality for integrals of smooth functions. *Dokl. Akad. Nauk* **474** 535–539. MR3791145 https://doi.org/10.1134/s1064562417030188

[94] MAUGIS, P., PRIEBE, C. E., OLHEDE, S. C. and WOLFE, P. J. (2017). Statistical inference for network samples using subgraph counts. Preprint. Available at arXiv:1701.00505.

[95] MILO, R., SHEN-ORR, S., ITZKOVITZ, S., KASHTAN, N., CHKLOVSKII, D. and ALON, U. (2002). Network motifs: Simple building blocks of complex networks. *Science* **298** 824–827.

[96] NEWMAN, M. E. and LEICHT, E. A. (2007). Mixture models and exploratory analysis in networks. *Proc. Natl. Acad. Sci. USA* **104** 9564–9569.

[97] NEWMAN, M. E. J. (2009). Random graphs with clustering. *Phys. Rev. Lett.* **103** 058701.

[98] NOWICKI, K. and WIERMAN, J. C. (1988). Subgraph counts in random graphs using incomplete U-statistics methods. *Discrete Math.* **72** 299–310.

[99] OUADAH, S., ROBIN, S. and LATOUCHE, P. (2020). Degree-based goodness-of-fit tests for heterogeneous random graph models: Independent and exchangeable cases. *Scand. J. Stat.* **47** 156–181. MR4075233 https://doi.org/10.1111/sjos.12410

[100] OUCHTI, L. (2005). On the rate of convergence in the central limit theorem for martingale difference sequences. *Ann. Inst. Henri Poincaré Probab. Stat.* **41** 35–43. MR2110445 https://doi.org/10.1016/j.anihpb.2004.03.003

[101] PRŽULJ, N. (2007). Biological network comparison using graphlet degree distribution. *Bioinformatics* **23** e177–e183.

[102] PUTTER, H. and VAN ZWET, W. R. (1998). Empirical Edgeworth expansions for symmetric statistics. *Ann. Statist.* **26** 1540–1569. MR1647697 https://doi.org/10.1214/aos/1024691253

[103] RIZZO, M. L. and SZÉKELY, G. J. (2016). Energy distance. *Wiley Interdiscip. Rev.: Comput. Stat.* **8** 27–38. MR3457239 https://doi.org/10.1002/wics.1375

[104] ROHE, K. (2019). A critical threshold for design effects in network sampling. *Ann. Statist.* **47** 556–582. MR3909942 https://doi.org/10.1214/18-AOS1700

[105] ROHE, K. and QIN, T. (2013). The blessing of transitivity in sparse and stochastic networks. Preprint. Available at arXiv:1307.2302.

[106] RUBINOV, M. and SPORNS, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage* **52** 1059–1069.

[107] SAKOV, A. and BICKEL, P. J. (2000). An Edgeworth expansion for the m out of n bootstrapped median. *Statist. Probab. Lett.* **49** 217–223. MR1794738 https://doi.org/10.1016/S0167-7152(00)00050-X

[108] SESHADHRI, C., SHARMA, A., STOLMAN, A. and GOEL, A. (2020). The impossibility of low-rank representations for triangle-rich complex networks. *Proc. Natl. Acad. Sci. USA* **117** 5631–5637. https://doi.org/10.1073/pnas.1911030117

[109] SINGH, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.* **9** 1187–1195. MR0630102

[110] SONG, K. (2016). Ordering-free inference from locally dependent data. Preprint. Available at arXiv:1604.00447.

[111] SONG, K. (2020). A uniform-in-P Edgeworth expansion under weak Cramér conditions. *Statistics* **54** 926–950. MR4178892 https://doi.org/10.1080/02331888.2020.1830094

[112] SONG, Y., CHEN, X. and KATO, K. (2019). Approximating high-dimensional infinite-order U-statistics: Statistical and computational guarantees. *Electron. J. Stat.* **13** 4794–4848. MR4038726 https://doi.org/10.1214/19-EJS1643

[113] STEPHEN, A. T. and TOUBIA, O. (2009). Explaining the power-law degree distribution in a social commerce network. *Soc. Netw.* **31** 262–270.

[114] TANG, M., ATHREYA, A., SUSSMAN, D. L., LYZINSKI, V., PARK, Y. and PRIEBE, C. E. (2017). A semiparametric two-sample hypothesis testing problem for random graphs. *J. Comput. Graph. Statist.* **26** 344–354. MR3640191 https://doi.org/10.1080/10618600.2016.1193505

[115] THOMPSON, M. E., RAMIREZ RAMIREZ, L. L., LYUBCHICH, V. and GEL, Y. R. (2016). Using the bootstrap for statistical inference on random graphs. *Canad. J. Statist.* **44** 3–24. MR3474218 https://doi.org/10.1002/cjs.11271

[116] TSIOTAS, D. (2019). Detecting different topologies immanent in scale-free networks with the same degree distribution. *Proc. Natl. Acad. Sci. USA* **116** 6701–6706. MR3939792 https://doi.org/10.1073/pnas.1816842116

[117] VERSHYNIN, R. (2018). Four lectures on probabilistic methods for data science. In *The Mathematics of Data. IAS/Park City Math. Ser.* **25** 231–271. Amer. Math. Soc., Providence, RI. MR3839170 https://doi.org/10.1090/pcms/025

[118] WALLACE, D. L. (1958). Asymptotic approximations to distributions. *Ann. Math. Stat.* **29** 635–654. MR0097109 https://doi.org/10.1214/aoms/1177706528

[119] WANG, Y. X. R. and BICKEL, P. J. (2017). Likelihood-based model selection for stochastic block models. *Ann. Statist.* **45** 500–528. MR3650391 https://doi.org/10.1214/16-AOS1457

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

[121] WATTS, D. J. and STROGATZ, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* **393** 440.

[122] WEGNER, A. E., OSPINA-FORERO, L., GAUNT, R. E., DEANE, C. M. and REINERT, G. (2018). Identifying networks with common organizational principles. *J. Complex Netw.* **6** 887–913. MR3892299 https://doi.org/10.1093/comnet/cny003

[123] WOODROOFE, M. and JHUN, M. (1988). Singh's theorem in the lattice case. *Statist. Probab. Lett.* **7** 201–205. MR0980922 https://doi.org/10.1016/0167-7152(88)90051-X

[124] XIA, D. and YUAN, M. (2021). Statistical inferences of linear forms for noisy matrix completion. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **83** 58–77. MR4220984 https://doi.org/10.1111/rssb.12400

[125] XU, J. (2018). Rates of convergence of spectral methods for graphon estimation. In *Proceedings of the 35th International Conference on Machine Learning*.

[126] 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

[127] 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

[128] ZHANG, A. Y. and ZHOU, H. H. (2020). Theoretical and computational guarantees of mean field variational inference for community detection. *Ann. Statist.* **48** 2575–2598. MR4152113 https://doi.org/10.1214/19-AOS1898

[129] ZHANG, X., MARTIN, T. and NEWMAN, M. E. (2015). Identification of core-periphery structure in networks. *Phys. Rev. E* **91** 032803.

[130] ZHANG, Y., LEVINA, E. and ZHU, J. (2017). Estimating network edge probabilities by neighbourhood smoothing. *Biometrika* **104** 771–783. MR3737303 https://doi.org/10.1093/biomet/asx042

[131] ZHANG, Y., LEVINA, E. and ZHU, J. (2020). Detecting overlapping communities in networks using spectral methods. *SIAM J. Math. Data Sci.* **2** 265–283. MR4081803 https://doi.org/10.1137/19M1272238

[132] ZHANG, Y. and XIA, D. (2022). Supplement to "Edgeworth expansions for network moments." https://doi.org/10.1214/21-AOS2125SUPP

[133] ZHAO, Y., LEVINA, E. and ZHU, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Statist.* **40** 2266–2292. MR3059083 https://doi.org/10.1214/12-AOS1036

The Annals of Statistics
2022, Vol. 50, No. 2, 754–780
https://doi.org/10.1214/21-AOS2126

PARAMETRIC COPULA ADJUSTED FOR NON- AND SEMIPARAMETRIC REGRESSION

BY YUE ZHAO[1,3,c], IRÈNE GIJBELS[2,b] AND INGRID VAN KEILEGOM[1,a]

[1]*Research Centre for Operations Research and Statistics (ORSTAT), KU Leuven,* [a]*ingrid.vankeilegom@kuleuven.be*
[2]*Department of Mathematics and Leuven Statistics Research Center (LStat), KU Leuven,* [b]*irene.gijbels@kuleuven.be*
[3]*Department of Mathematics, University of York,* [c]*yue.zhao@york.ac.uk*

We consider a multivariate response regression model where each coordinate is described by a location-scale non- or semiparametric regression and where the dependence structure of the "noise term" is described by a parametric copula. Our goal is to estimate the associated Euclidean copula parameter, given a sample of the response and the covariate. In the absence of the copula sample, the usual oracle ranks are no longer computable. Instead, we study the normal scores estimator for the Gaussian copula and generalized pseudo-likelihood estimation for general parametric copulas, both based on residual ranks calculated from preliminary non- or semiparametric estimators of the location and scale functions. We show that the residual-based estimators are asymptotically equivalent to their oracle counterparts and provide explicit rate of convergence. Partially to serve this objective, we also study weighted convergence of the residual empirical process under the non- or semiparametric regression model.

REFERENCES

AKRITAS, M. G. and VAN KEILEGOM, I. (2001). Non-parametric estimation of the residual distribution. *Scand. J. Stat.* **28** 549–567. MR1858417 https://doi.org/10.1111/1467-9469.00254

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York. MR1623559

CHEN, X. and FAN, Y. (2006). Estimation and model selection of semiparametric copula-based multivariate dynamic models under copula misspecification. *J. Econometrics* **135** 125–154. MR2328398 https://doi.org/10.1016/j.jeconom.2005.07.027

CHEN, X., HUANG, Z. and YI, Y. (2021). Efficient estimation of multivariate semi-nonparametric GARCH filtered copula models. *J. Econometrics* **222** 484–501. MR4234829 https://doi.org/10.1016/j.jeconom.2020.07.012

CHEN, G. and LOCKHART, R. A. (2001). Weak convergence of the empirical process of residuals in linear models with many parameters. *Ann. Statist.* **29** 748–762. MR1865339 https://doi.org/10.1214/aos/1009210688

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. *Monographs on Statistics and Applied Probability* **66**. CRC Press, London. MR1383587

GENEST, C., GHOUDI, K. and RIVEST, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika* **82** 543–552. MR1366280 https://doi.org/10.1093/biomet/82.3.543

GIJBELS, I., OMELKA, M. and VERAVERBEKE, N. (2015). Estimation of a copula when a covariate affects only marginal distributions. *Scand. J. Stat.* **42** 1109–1126. MR3426313 https://doi.org/10.1111/sjos.12154

HÁJEK, J., ŠIDÁK, Z. and SEN, P. K. (1999). *Theory of Rank Tests*, 2nd ed. *Probability and Mathematical Statistics*. Academic Press, San Diego, CA. MR1680991

KLAASSEN, C. A. J. and WELLNER, J. A. (1997). Efficient estimation in the bivariate normal copula model: Normal margins are least favourable. *Bernoulli* **3** 55–77. MR1466545 https://doi.org/10.2307/3318652

MAMMEN, E. (1996). Empirical process of residuals for high-dimensional linear models. *Ann. Statist.* **24** 307–335. MR1389892 https://doi.org/10.1214/aos/1033066211

MSC2020 subject classifications. Primary 62H05; secondary 62E20, 62G20, 62G30.
Key words and phrases. Non- and semiparametric regression, normal scores rank correlation estimator, parametric copula, pseudo-likelihood, residual empirical processes.

MAMMEN, E., LINTON, O. and NIELSEN, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.* **27** 1443–1490. MR1742496 https://doi.org/10.1214/aos/1017939137

MÜLLER, U. U., SCHICK, A. and WEFELMEYER, W. (2009). Estimating the error distribution function in nonparametric regression with multivariate covariates. *Statist. Probab. Lett.* **79** 957–964. MR2509488 https://doi.org/10.1016/j.spl.2008.11.024

MÜLLER, U. U., SCHICK, A. and WEFELMEYER, W. (2012). Estimating the error distribution function in semiparametric additive regression models. *J. Statist. Plann. Inference* **142** 552–566. MR2843057 https://doi.org/10.1016/j.jspi.2011.08.013

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

NEUMEYER, N., OMELKA, M. and HUDECOVÁ, Š. (2019). A copula approach for dependence modeling in multivariate nonparametric time series. *J. Multivariate Anal.* **171** 139–162. MR3892031 https://doi.org/10.1016/j.jmva.2018.11.016

NEUMEYER, N. and VAN KEILEGOM, I. (2010). Estimating the error distribution in nonparametric multiple regression with applications to model testing. *J. Multivariate Anal.* **101** 1067–1078. MR2595293 https://doi.org/10.1016/j.jmva.2010.01.007

OAKES, D. (1994). Multivariate survival distributions. *J. Nonparametr. Stat.* **3** 343–354. MR1291555 https://doi.org/10.1080/10485259408832593

OECD (2015). Programme for International Student Assessment. Available at https://databank.worldbank.org/source/education-statistics:-learning-outcomes.

OMELKA, M., HUDECOVÁ, Š. and NEUMEYER, N. (2021). Maximum pseudo-likelihood estimation based on estimated residuals in copula semiparametric models. *Scand. J. Stat.* **48** 1433–1473. https://doi.org/10.1111/sjos.12498

PAKES, A. and POLLARD, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica* **57** 1027–1057. MR1014540 https://doi.org/10.2307/1913622

RUYMGAART, F. H. (1974). Asymptotic normality of nonparametric tests for independence. *Ann. Statist.* **2** 892–910. MR0386140

SEGERS, J., VAN DEN AKKER, R. and WERKER, B. J. M. (2014). Semiparametric Gaussian copula models: Geometry and efficient rank-based estimation. *Ann. Statist.* **42** 1911–1940. MR3262472 https://doi.org/10.1214/14-AOS1244

SKLAR, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Stat. Univ. Paris* **8** 229–231. MR0125600

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

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

VERAVERBEKE, N., GIJBELS, I. and OMELKA, M. (2014). Preadjusted non-parametric estimation of a conditional distribution function. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 399–438. MR3164872 https://doi.org/10.1111/rssb.12041

VERAVERBEKE, N., OMELKA, M. and GIJBELS, I. (2011). Estimation of a conditional copula and association measures. *Scand. J. Stat.* **38** 766–780. MR2859749 https://doi.org/10.1111/j.1467-9469.2011.00744.x

XIE, H. and HUANG, J. (2009). SCAD-penalized regression in high-dimensional partially linear models. *Ann. Statist.* **37** 673–696. MR2502647 https://doi.org/10.1214/07-AOS580

ZHAO, Y. and GENEST, C. (2019). Inference for elliptical copula multivariate response regression models. *Electron. J. Stat.* **13** 911–984. MR3934620 https://doi.org/10.1214/19-EJS1534

ZHAO, Y., GIJBELS, I. and VAN KEILEGOM, I. (2020). Inference for semiparametric Gaussian copula model adjusted for linear regression using residual ranks. *Bernoulli* **26** 2815–2846. MR4140530 https://doi.org/10.3150/20-BEJ1208

ZHAO, Y., GIJBELS, I. and VAN KEILEGOM, I. (2022). Supplement to "Parametric copula adjusted for non- and semiparametric regression." https://doi.org/10.1214/21-AOS2126SUPP

The Annals of Statistics
2022, Vol. 50, No. 2, 781–806
https://doi.org/10.1214/21-AOS2127

INFERENCE FOR CHANGE POINTS IN HIGH-DIMENSIONAL DATA VIA SELFNORMALIZATION

BY RUNMIN WANG[1,a], CHANGBO ZHU[2,b], STANISLAV VOLGUSHEV[3,c] AND XIAOFENG SHAO[4,d]

[1]*Department of Statistical Science, Southern Methodist University,* [a]*runminw@smu.edu*
[2]*Department of Statistics, University of California at Davis,* [b]*cbzhu@ucdavis.edu*
[3]*Department of Statistical Sciences, University of Toronto,* [c]*stanislav.volgushev@utoronto.ca*
[4]*Department of Statistics, University of Illinois at Urbana–Champaign,* [d]*xshao@illinois.edu*

This article considers change-point testing and estimation for a sequence of high-dimensional data. In the case of testing for a mean shift for high-dimensional independent data, we propose a new test which is based on U-statistic in Chen and Qin (*Ann. Statist.* **38** (2010) 808–835) and utilizes the self-normalization principle (Shao *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** (2010) 343–366; Shao and Zhang *J. Amer. Statist. Assoc.* **105** (2010) 1228–1240). Our test targets dense alternatives in the high-dimensional setting and involves no tuning parameters. To extend to change-point testing for high-dimensional time series, we introduce a trimming parameter and formulate a self-normalized test statistic with trimming to accommodate the weak temporal dependence. On the theory front we derive the limiting distributions of self-normalized test statistics under both the null and alternatives for both independent and dependent high-dimensional data. At the core of our asymptotic theory, we obtain weak convergence of a sequential U-statistic based process for high-dimensional independent data, and weak convergence of sequential trimmed U-statistic based processes for high-dimensional linear processes, both of which are of independent interests. Additionally, we illustrate how our tests can be used in combination with wild binary segmentation to estimate the number and location of multiple change points. Numerical simulations demonstrate the competitiveness of our proposed testing and estimation procedures in comparison with several existing methods in the literature.

REFERENCES

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

AUE, A., HÖRMANN, S., HORVÁTH, L. and REIMHERR, M. (2009). Break detection in the covariance structure of multivariate time series models. *Ann. Statist.* **37** 4046–4087. MR2572452 https://doi.org/10.1214/09-AOS707

AVANESOV, V. and BUZUN, N. (2018). Change-point detection in high-dimensional covariance structure. *Electron. J. Stat.* **12** 3254–3294. MR3861282 https://doi.org/10.1214/18-EJS1484

CHAN, J., HORVÁTH, L. and HUŠKOVÁ, M. (2013). Darling–Erdős limit results for change-point detection in panel data. *J. Statist. Plann. Inference* **143** 955–970. MR3011306 https://doi.org/10.1016/j.jspi.2012.11.004

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

CHO, H. (2016). Change-point detection in panel data via double CUSUM statistic. *Electron. J. Stat.* **10** 2000–2038. MR3522667 https://doi.org/10.1214/16-EJS1155

DETTE, H. and GÖSMANN, J. (2018). Relevant change points in high dimensional time series. *Electron. J. Stat.* **12** 2578–2636. MR3849896 https://doi.org/10.1214/18-EJS1464

ENIKEEVA, F. and HARCHAOUI, Z. (2019). High-dimensional change-point detection under sparse alternatives. *Ann. Statist.* **47** 2051–2079. MR3953444 https://doi.org/10.1214/18-AOS1740

MSC2020 subject classifications. Primary 62H15, 60K35; secondary 62G10, 62G20.
Key words and phrases. CUSUM, segmentation, selfnormalization, structural break, time series, U-statistic.

FAN, Z. and MACKEY, L. (2017). Empirical Bayesian analysis of simultaneous changepoints in multiple data sequences. *Ann. Appl. Stat.* **11** 2200–2221. MR3743294 https://doi.org/10.1214/17-AOAS1075

FRYZLEWICZ, P. (2014). Wild binary segmentation for multiple change-point detection. *Ann. Statist.* **42** 2243–2281. MR3269979 https://doi.org/10.1214/14-AOS1245

HORVÁTH, L. and HUŠKOVÁ, M. (2012). Change-point detection in panel data. *J. Time Series Anal.* **33** 631–648. MR2944843 https://doi.org/10.1111/j.1467-9892.2012.00796.x

HSING, T. and WU, W. B. (2004). On weighted U-statistics for stationary processes. *Ann. Probab.* **32** 1600–1631. MR2060311 https://doi.org/10.1214/009117904000000333

JIRAK, M. (2012). Change-point analysis in increasing dimension. *J. Multivariate Anal.* **111** 136–159. MR2944411 https://doi.org/10.1016/j.jmva.2012.05.007

JIRAK, M. (2015). Uniform change point tests in high dimension. *Ann. Statist.* **43** 2451–2483. MR3405600 https://doi.org/10.1214/15-AOS1347

KIEFER, N. M. and VOGELSANG, T. J. (2005). A new asymptotic theory for heteroskedasticity-autocorrelation robust tests. *Econometric Theory* **21** 1130–1164. MR2200988 https://doi.org/10.1017/S0266466605050565

KIRCH, C., MUHSAL, B. and OMBAO, H. (2015). Detection of changes in multivariate time series with application to EEG data. *J. Amer. Statist. Assoc.* **110** 1197–1216. MR3420695 https://doi.org/10.1080/01621459.2014.957545

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

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

LI, J., XU, M., ZHONG, P.-S. and LI, L. (2019). Change point detection in the mean of high-dimensional time series data under dependence. arXiv preprint. Available at arXiv:1903.07006.

MATTESON, D. S. and JAMES, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *J. Amer. Statist. Assoc.* **109** 334–345. MR3180567 https://doi.org/10.1080/01621459.2013.849605

PERRON, P. (2006). Dealing with structural breaks. In *Palgrave Handbook of Econometrics* **1** 278–352.

PHILLIPS, P. C. B. and SOLO, V. (1992). Asymptotics for linear processes. *Ann. Statist.* **20** 971–1001. MR1165602 https://doi.org/10.1214/aos/1176348666

SHAO, X. (2010). A self-normalized approach to confidence interval construction in time series. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 343–366. MR2758116 https://doi.org/10.1111/j.1467-9868.2009.00737.x

SHAO, X. (2015). Self-normalization for time series: A review of recent developments. *J. Amer. Statist. Assoc.* **110** 1797–1817. MR3449074 https://doi.org/10.1080/01621459.2015.1050493

SHAO, X. and WU, W. B. (2007). Local Whittle estimation of fractional integration for nonlinear processes. *Econometric Theory* **23** 899–929. MR2396737 https://doi.org/10.1017/S0266466607070387

SHAO, X. and ZHANG, X. (2010). Testing for change points in time series. *J. Amer. Statist. Assoc.* **105** 1228–1240. MR2752617 https://doi.org/10.1198/jasa.2010.tm10103

WANG, T. and SAMWORTH, R. J. (2018). High dimensional change point estimation via sparse projection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 57–83. MR3744712 https://doi.org/10.1111/rssb.12243

WANG, R. and SHAO, X. (2020). Hypothesis testing for high-dimensional time series via self-normalization. *Ann. Statist.* **48** 2728–2758. MR4152119 https://doi.org/10.1214/19-AOS1904

WANG, R., ZHU, C., VOLGUSHEV, S. and SHAO, X. (2022). Supplement to "Inference for change points in high-dimensional data via selfnormalization." https://doi.org/10.1214/21-AOS2127SUPP

WU, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proc. Natl. Acad. Sci. USA* **102** 14150–14154. MR2172215 https://doi.org/10.1073/pnas.0506715102

WU, W. B. and SHAO, X. (2004). Limit theorems for iterated random functions. *J. Appl. Probab.* **41** 425–436. MR2052582 https://doi.org/10.1239/jap/1082999076

YU, M. and CHEN, X. (2021). Finite sample change point inference and identification for high-dimensional mean vectors. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **83** 247–270. MR4250275 https://doi.org/10.1111/rssb.12406

ZHANG, X. and CHENG, G. (2018). Gaussian approximation for high dimensional vector under physical dependence. *Bernoulli* **24** 2640–2675. MR3779697 https://doi.org/10.3150/17-BEJ939

ZHANG, T. and LAVITAS, L. (2018). Unsupervised self-normalized change-point testing for time series. *J. Amer. Statist. Assoc.* **113** 637–648. MR3832215 https://doi.org/10.1080/01621459.2016.1270214

ZHURBENKO, I. G. and ZUEV, N. M. (1975). On higher spectral densities of stationary processes with mixing. *Ukrainian Math. J.* **27** 364–373.

The Annals of Statistics
2022, Vol. 50, No. 2, 807–857
https://doi.org/10.1214/21-AOS2128

OPTIMAL FALSE DISCOVERY RATE CONTROL FOR LARGE SCALE MULTIPLE TESTING WITH AUXILIARY INFORMATION

BY HONGYUAN CAO[1,a], JUN CHEN[2,b] AND XIANYANG ZHANG[3,c]

[1]*Department of Statistics, Florida State University,* [a]*hcao@fsu.edu*
[2]*Department of Quantitative Health Sciences, Mayo Clinic,* [b]*Chen.Jun2@mayo.edu*
[3]*Department of Statistics, Texas A&M University,* [c]*zhangxiany@stat.tamu.edu*

Large-scale multiple testing is a fundamental problem in high dimensional statistical inference. It is increasingly common that various types of auxiliary information, reflecting the structural relationship among the hypotheses, are available. Exploiting such auxiliary information can boost statistical power. To this end, we propose a framework based on a two-group mixture model with varying probabilities of being null for different hypotheses *a priori*, where a shape-constrained relationship is imposed between the auxiliary information and the prior probabilities of being null. An optimal rejection rule is designed to maximize the expected number of true positives when average false discovery rate is controlled. Focusing on the ordered structure, we develop a robust EM algorithm to estimate the prior probabilities of being null and the distribution of p-values under the alternative hypothesis simultaneously. We show that the proposed method has better power than state-of-the-art competitors while controlling the false discovery rate, both empirically and theoretically. Extensive simulations demonstrate the advantage of the proposed method. Datasets from genome-wide association studies are used to illustrate the new methodology.

REFERENCES

[1] AYER, M., BRUNK, H. D., EWING, G. M., REID, W. T. and SILVERMAN, E. (1955). An empirical distribution function for sampling with incomplete information. *Ann. Math. Stat.* **26** 641–647. MR0073895 https://doi.org/10.1214/aoms/1177728423

[2] BARBER, R. F. and CANDÈS, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.* **43** 2055–2085. MR3375876 https://doi.org/10.1214/15-AOS1337

[3] BARLOW, R. E. and BRUNK, H. D. (1972). The isotonic regression problem and its dual. *J. Amer. Statist. Assoc.* **67** 140–147. MR0314205

[4] BASU, P., CAI, T. T., DAS, K. and SUN, W. (2018). Weighted false discovery rate control in large-scale multiple testing. *J. Amer. Statist. Assoc.* **113** 1172–1183. MR3862348 https://doi.org/10.1080/01621459.2017.1336443

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

[6] BOCA, S. M. and LEEK, J. T. (2018). A direct approach to estimating false discovery rates conditional on covariates. *PeerJ* **6** e6035. https://doi.org/10.7717/peerj.6035

[7] CAI, T. T. and SUN, W. (2009). Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *J. Amer. Statist. Assoc.* **104** 1467–1481. MR2597000 https://doi.org/10.1198/jasa.2009.tm08415

[8] CAO, H., SUN, W. and KOSOROK, M. R. (2013). The optimal power puzzle: Scrutiny of the monotone likelihood ratio assumption in multiple testing. *Biometrika* **100** 495–502. MR3068449 https://doi.org/10.1093/biomet/ast001

[9] CORONARY ARTERY DISEASE (C4D) GENETICS CONSORTIUM (2011). A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. *Nat. Genet.* **43** 339–344.

MSC2020 subject classifications. Primary 62G07, 62G10; secondary 62C12.
Key words and phrases. EM algorithm, false discovery rate, isotonic regression, local false discovery rate, multiple testing, Pool-Adjacent-Violators algorithm.

[10] DEB, N., SAHA, S., GUNTUBOYINA, A. and SEN, B. (2019). Two-component mixture model in the presence of covariates. Preprint. Available at arXiv:1810.07897.

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

[12] DOBRIBAN, E. (2018). Weighted mining of massive collections of P-values by convex optimization. *Inf. Inference* **7** 251–275. MR3812121 https://doi.org/10.1093/imaiai/iax013

[13] EFRON, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.* **23** 1–22. MR2431866 https://doi.org/10.1214/07-STS236

[14] EFRON, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. Institute of Mathematical Statistics (IMS) Monographs* **1**. Cambridge Univ. Press, Cambridge. MR2724758 https://doi.org/10.1017/CBO9780511761362

[15] EFRON, B., TIBSHIRANI, R., STOREY, J. D. and TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151–1160. MR1946571 https://doi.org/10.1198/016214501753382129

[16] FERKINGSTAD, E., FRIGESSI, A., RUE, H., THORLEIFSSON, G. and KONG, A. (2008). Unsupervised empirical Bayesian multiple testing with external covariates. *Ann. Appl. Stat.* **2** 714–735. MR2524353 https://doi.org/10.1214/08-AOAS158

[17] GENOVESE, C. R., ROEDER, K. and WASSERMAN, L. (2006). False discovery control with p-value weighting. *Biometrika* **93** 509–524. MR2261439 https://doi.org/10.1093/biomet/93.3.509

[18] GRAZIER G'SELL, M., WAGER, S., CHOULDECHOVA, A. and TIBSHIRANI, R. (2016). Sequential selection procedures and false discovery rate control. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 423–444. MR3454203 https://doi.org/10.1111/rssb.12122

[19] GRENANDER, U. (1956). On the theory of mortality measurement. II. *Skand. Aktuarietidskr.* **39** 125–153. MR0093415 https://doi.org/10.1080/03461238.1956.10414944

[20] HU, J. X., ZHAO, H. and ZHOU, H. H. (2010). False discovery rate control with groups. *J. Amer. Statist. Assoc.* **105** 1215–1227. MR2752616 https://doi.org/10.1198/jasa.2010.tm09329

[21] HUANG, J. Y., BAI, L., CUI, B. W., WU, L., WANG, L. W., AN, Z. Y., RUAN, YU Y, S. L., ZHANG, X. Y. et al. (2020). Leveraging biological and statistical covariates improves the detection power in epigenome-wide association testing. *Genome Biol.* **21** 1–19.

[22] IGNATIADIS, N. and HUBER, W. (2017). Covariate-powered weighted multiple testing with false discovery rate control. Preprint. Available at arXiv:1701.05179.

[23] IGNATIADIS, N., KLAUS, B., ZAUGG, J. B. and HUBER, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat. Methods* **13** 577–580.

[24] LANGAAS, M., LINDQVIST, B. H. and FERKINGSTAD, E. (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 555–572. MR2168204 https://doi.org/10.1111/j.1467-9868.2005.00515.x

[25] LEI, L. and FITHIAN, W. (2016). Power of Ordered Hypothesis Testing. Preprint. Available at arXiv:1606.01969.

[26] LEI, L. and FITHIAN, W. (2018). AdaPT: An interactive procedure for multiple testing with side information. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 649–679. MR3849338 https://doi.org/10.1111/rssb.12253

[27] LEI, L., RAMDAS, A. and FITHIAN, W. (2021). A general interactive framework for FDR control under structural constraints. *Biometrika* **108** 253–267. MR4259130 https://doi.org/10.1093/biomet/asaa064

[28] LI, A. and BARBER, R. F. (2017). Accumulation tests for FDR control in ordered hypothesis testing. *J. Amer. Statist. Assoc.* **112** 837–849. MR3671774 https://doi.org/10.1080/01621459.2016.1180989

[29] LI, A. and BARBER, R. F. (2019). Multiple testing with the structure-adaptive Benjamini-Hochberg algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 45–74. MR3904779

[30] LOVE, M. I., HUBER, W. and ANDERS, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15** 550. https://doi.org/10.1186/s13059-014-0550-8

[31] LYNCH, G., GUO, W., SARKAR, S. K. and FINNER, H. (2017). The control of the false discovery rate in fixed sequence multiple testing. *Electron. J. Stat.* **11** 4649–4673. MR3724971 https://doi.org/10.1214/17-EJS1359

[32] ROBERTSON, T. and WALTMAN, P. (1968). On estimating monotone parameters. *Ann. Math. Stat.* **39** 1030–1039. MR0230408 https://doi.org/10.1214/aoms/1177698335

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

[34] SCHUNKERT, H., KONIG, I. R., KATHIRESAN, S., REILLY, M. P., ASSIMES, T. L., HOLM, H. et al. (2011). Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat. Genet.* **43** 333–338.

[35] SCOTT, J. G., KELLY, R. C., SMITH, M. A., ZHOU, P. and KASS, R. E. (2015). False discovery rate regression: An application to neural synchrony detection in primary visual cortex. *J. Amer. Statist. Assoc.* **110** 459–471. MR3367240 https://doi.org/10.1080/01621459.2014.990973

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

[37] STOREY, J. D., TAYLOR, J. E. and SIEGMUND, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 187–205. MR2035766 https://doi.org/10.1111/j.1467-9868.2004.00439.x

[38] STOREY, J. D. and TIBSHIRANI, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100** 9440–9445. MR1994856 https://doi.org/10.1073/pnas.1530509100

[39] SUN, W. and CAI, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.* **102** 901–912. MR2411657 https://doi.org/10.1198/016214507000000545

[40] SUN, W., REICH, B. J., CAI, T. T., GUINDANI, M. and SCHWARTZMAN, A. (2015). False discovery control in large-scale spatial multiple testing. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 59–83. MR3299399 https://doi.org/10.1111/rssb.12064

[41] TANG, W. and ZHANG, C. (2005). Bayes and empirical bayes approaches to controlling the false discovery rate Technical report, Dept. Statistics and Biostatistics, Rutgers Univ.

[42] TANSEY, W., KOYEJO, O., POLDRACK, R. A. and SCOTT, J. G. (2018). False discovery rate smoothing. *J. Amer. Statist. Assoc.* **113** 1156–1171. MR3862347 https://doi.org/10.1080/01621459.2017.1319838

[43] VAN DE GEER, S. (2000). *Empirical Processes in M-Estimation*. Cambridge Univ. Press, Cambridge.

[44] VAN DER VAART, A. W. and WELLNER, J. A. (2000). *Weak Convergence and Empirical Processes: With Applications to Statistics. Springer Series in Statistics*. Springer, New York.

[45] XIAO, J., CAO, H. and CHEN, J. (2017). False discovery rate control incorporating phylogenetic tree increases detection power in microbiome wide multiple testing. *Bioinformatics* **33** 2873–2881.

[46] ZHANG, X. and CHEN, J. (2020). Covariate adaptive false discovery rate control with applications to Omics-Wide multiple testing. *J. Amer. Statist. Assoc.* To appear. https://doi.org/10.1080/01621459.2020.1783273

The Annals of Statistics
2022, Vol. 50, No. 2, 858–879
https://doi.org/10.1214/21-AOS2129

ADAPTIVE TEST OF INDEPENDENCE BASED ON HSIC MEASURES

BY MÉLISANDE ALBERT[1,a], BÉATRICE LAURENT[1,b], AMANDINE MARREL[2,d] AND ANOUAR MEYNAOUI[1,c]

[1]*Institut de Mathématiques de Toulouse; UMR 5219 Université de Toulouse; CNRS, INSA,* [a]*melisande.albert@insa-toulouse.fr,* [b]*beatrice.laurent@insa-toulouse.fr,* [c]*anouar.meynaoui@gmail.com*

[2]*CEA, DES, IRESNE, DER, Cadarache Center and Institut de Mathématiques de Toulouse; UMR 5219 Université de Toulouse; CNRS,* [d]*amandine.marrel@cea.fr*

The Hilbert–Schmidt Independence Criterion (HSIC) is a dependence measure based on reproducing kernel Hilbert spaces that is widely used to test independence between two random vectors. Remains the delicate choice of the kernel. In this work, we develop a new HSIC-based aggregated procedure which avoids such a kernel choice, and provide theoretical guarantees for this procedure. To achieve this, on the one hand, we introduce non-asymptotic single tests based on Gaussian kernels with a given bandwidth, which are of prescribed level. Then, we aggregate several single tests with different bandwidths, and prove sharp upper bounds for the uniform separation rate of the aggregated procedure over Sobolev balls. On the other hand, we provide a lower bound for the non-asymptotic minimax separation rate of testing over Sobolev balls, and deduce that the aggregated procedure is adaptive in the minimax sense over such regularity spaces. Finally, from a practical point of view, we perform numerical studies in order to assess the efficiency of our aggregated procedure and compare it to existing tests in the literature.

REFERENCES

AHMAD, I. A. and LI, Q. (1997). Testing independence by nonparametric kernel method. *Statist. Probab. Lett.* **34** 201–210. MR1457514 https://doi.org/10.1016/S0167-7152(96)00183-6

ALBERT, M. (2015). Tests of independence by bootstrap and permutation: An asymptotic and non-asymptotic study. Application to neurosciences Ph.D. thesis, Univ. Nice Sophia Antipolis.

ALBERT, M., LAURENT, B., MARREL, A. and MEYNAOUI, A. (2022). Supplement to “Adaptive test of independence based on HSIC measures.” https://doi.org/10.1214/21-AOS2129SUPP

ARONSZAJN, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68** 337–404. MR0051437 https://doi.org/10.2307/1990404

BACH, F. R. and JORDAN, M. I. (2002). Kernel independent component analysis. *J. Mach. Learn. Res.* **3** 1–48.

BAKER, C. R. (1973). Joint measures and cross-covariance operators. *Trans. Amer. Math. Soc.* **186** 273–289. MR0336795 https://doi.org/10.2307/1996566

BALAKRISHNAN, S. and WASSERMAN, L. (2019). Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates. *Ann. Statist.* **47** 1893–1927. MR3953439 https://doi.org/10.1214/18-AOS1729

BARAUD, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli* **8** 577–606. MR1935648

BARAUD, Y., HUET, S. and LAURENT, B. (2003). Adaptive tests of linear hypotheses by model selection. *Ann. Statist.* **31** 225–251. MR1962505 https://doi.org/10.1214/aos/1046294463

BERRETT, T. B., KONTOYIANNIS, I. and SAMWORTH, R. J. (2021). Optimal rates for independence testing via U-statistic permutation tests. *Ann. Statist.* **49** 2457–2490. MR4338371 https://doi.org/10.1214/20-aos2041

BERRETT, T. B. and SAMWORTH, R. J. (2019). Nonparametric independence testing via mutual information. *Biometrika* **106** 547–566. MR3992389 https://doi.org/10.1093/biomet/asz024

BUTUCEA, C. (2007). Goodness-of-fit testing and quadratic functional estimation from indirect observations. *Ann. Statist.* **35** 1907–1930. MR2363957 https://doi.org/10.1214/009053607000000118

MSC2020 subject classifications. Primary 62G10; secondary 62G09.

Key words and phrases. Nonparametric test of independence, Hilbert–Schmidt Independence Criterion, permutation methods, uniform separation rates, aggregated tests, non-asymptotic minimax and adaptive tests.

DE LOZZO, M. and MARREL, A. (2017). Sensitivity analysis with dependence and variance-based measures for spatio-temporal numerical simulators. *Stoch. Environ. Res. Risk Assess.* **31** 1437–1453.

FROMONT, M. and LAURENT, B. (2006). Adaptive goodness-of-fit tests in a density model. *Ann. Statist.* **34** 680–720. MR2281881 https://doi.org/10.1214/009053606000000119

FROMONT, M., LAURENT, B. and REYNAUD-BOURET, P. (2013). The two-sample problem for Poisson processes: Adaptive tests with a nonasymptotic wild bootstrap approach. *Ann. Statist.* **41** 1431–1461. MR3113817 https://doi.org/10.1214/13-AOS1114

FUKUMIZU, K., BACH, F. R. and JORDAN, M. I. (2004). Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *J. Mach. Learn. Res.* **5** 73–99. MR2247974 https://doi.org/10.1162/153244303768966111

FUKUMIZU, K., GRETTON, A., SUN, X. and SCHÖLKOPF, B. (2008). Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems* 489–496.

GRETTON, A., HERBRICH, R. and SMOLA, A. J. (2003). The kernel mutual information. In *Acoustics, Speech, and Signal Processing*, 2003. *Proceedings*. (*ICASSP'*03). 2003 *IEEE International Conference on* **4** IV–880. IEEE Press, New York.

GRETTON, A., BOUSQUET, O., SMOLA, A. and SCHÖLKOPF, B. (2005a). Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory*. *Lecture Notes in Computer Science* **3734** 63–77. Springer, Berlin. MR2255909 https://doi.org/10.1007/11564089_7

GRETTON, A., HERBRICH, R., SMOLA, A., BOUSQUET, O. and SCHÖLKOPF, B. (2005b). Kernel methods for measuring independence. *J. Mach. Learn. Res.* **6** 2075–2129. MR2249882

GRETTON, A., SMOLA, A. J., BOUSQUET, O., HERBRICH, R., BELITSKI, A., AUGATH, M., MURAYAMA, Y., PAULS, J., SCHÖLKOPF, B. et al. (2005c). Kernel constrained covariance for dependence measurement. In *AISTATS* **10** 112–119.

GRETTON, A., FUKUMIZU, K., TEO, C. H., SONG, L., SCHÖLKOPF, B. and SMOLA, A. J. (2008). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems* 585–592.

HELLER, R., HELLER, Y., KAUFMAN, S., BRILL, B. and GORFINE, M. (2016). Consistent distribution-free K-sample and independence tests for univariate random variables. *J. Mach. Learn. Res.* **17** Paper No. 29, 54. MR3491123

HOEFFDING, W. (1948). A non-parametric test of independence. *Ann. Math. Stat.* **19** 546–557. MR0029139 https://doi.org/10.1214/aoms/1177730150

INGSTER, YU. I. (1989). An asymptotically minimax test of the hypothesis of independence. *J. Sov. Math.* **44** 466–476.

INGSTER, YU. I. (1993a). Asymptotically minimax hypothesis testing for nonparametric alternatives. I. *Math. Methods Statist.* **2** 85–114. MR1257978

INGSTER, YU. I. (1993b). Minimax testing of the hypothesis of independence for ellipsoids in l_p. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov.* (*POMI*) **207** 77–97, 159. MR1227976 https://doi.org/10.1007/BF02362345

INGSTER, YU. I. (2000). Adaptive chi-square tests. *J. Math. Sciences* **99** 1110–1119.

INGSTER, YU. I. and SUSLINA, I. A. (1998). Minimax signal detection for Besov bodies and balls. *Problemy Peredachi Informatsii* **34** 56–68. MR1654822

ISHIGAMI, T. and HOMMA, T. (1990). An importance quantification technique in uncertainty analysis for computer models. In [1990] *Proceedings*. *First International Symposium on Uncertainty Modeling and Analysis* 398–403. IEEE Comput. Soc., Los Alamitos, CA.

JACOD, J. and PROTTER, P. (2012). *Probability Essentials*. Springer Science & Business Media, Berlin.

KIM, I., BALAKRISHNAN, S. and WASSERMAN, L. (2020). Minimax optimality of permutation tests. Preprint. Available at arXiv:2003.13208.

LAURENT, B., LOUBES, J.-M. and MARTEAU, C. (2012). Non asymptotic minimax rates of testing in signal detection with heterogeneous variances. *Electron. J. Stat.* **6** 91–122. MR2879673 https://doi.org/10.1214/12-EJS667

LEE, D., ZHANG, K. and KOSOROK, M. R. (2019). Testing independence with the binary expansion randomized ensemble test. Preprint. Available at arXiv:1912.03662.

LI, T. and YUAN, M. (2019). On the optimality of gaussian kernel based nonparametric tests against smooth alternatives. Preprint. Available at arXiv:1909.03302.

MARREL, A., RAGUET, H. and CHABRIDON, V. (2020). Statistical developments for target and conditional sensitivity analysis: Application on safety studies for nuclear reactor. HAL preprint hal-02541142v2.

MICCHELLI, C. A., XU, Y. and ZHANG, H. (2006). Universal kernels. *J. Mach. Learn. Res.* **7** 2651–2667. MR2274454

PARZEN, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Stat.* **33** 1065–1076. MR0143282 https://doi.org/10.1214/aoms/1177704472

PFISTER, N., BÜHLMANN, P., SCHÖLKOPF, B. and PETERS, J. (2018). Kernel-based tests for joint independence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 5–31. MR3744710 https://doi.org/10.1111/rssb.12235

PÓCZOS, B., GHAHRAMANI, Z. and SCHNEIDER, J. (2012). Copula-based kernel dependency measures. Preprint. Available at arXiv:1206.4682.

RAMDAS, A., ISENBERG, D., SINGH, A. and WASSERMAN, L. (2016). Minimax lower bounds for linear independence testing. In 2016 *IEEE International Symposium on Information Theory* (*ISIT*) 965–969. IEEE Press, New York.

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

ROSENBLATT, M. (1975). A quadratic measure of deviation of two-dimensional density estimates and a test of independence. *Ann. Statist.* **3** 1–14. MR0428579

SPOKOINY, V. G. (1996). Adaptive hypothesis testing using wavelets. *Ann. Statist.* **24** 2477–2498. MR1425962 https://doi.org/10.1214/aos/1032181163

SRIPERUMBUDUR, B. K., GRETTON, A., FUKUMIZU, K., SCHÖLKOPF, B. and LANCKRIET, G. R. G. (2010). Hilbert space embeddings and metrics on probability measures. *J. Mach. Learn. Res.* **11** 1517–1561. MR2645460

STEINWART, I. (2001). On the influence of the kernel on the consistency of support vector machines. *J. Mach. Learn. Res.* **2** 67–93.

SZÉKELY, G. J. and RIZZO, M. L. (2013). The distance correlation t-test of independence in high dimension. *J. Multivariate Anal.* **117** 193–213. MR3053543 https://doi.org/10.1016/j.jmva.2013.02.012

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

WEIHS, L., DRTON, M. and MEINSHAUSEN, N. (2018). Symmetric rank covariances: A generalized framework for nonparametric measures of dependence. *Biometrika* **105** 547–562. MR3842884 https://doi.org/10.1093/biomet/asy021

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

YODÉ, A. F. (2004). Asymptotically minimax test of independence. *Math. Methods Statist.* **13** 201–234. MR2090472

YODÉ, A. F. (2011). Adaptive minimax test of independence. *Math. Methods Statist.* **20** 246–268. MR2908761 https://doi.org/10.3103/S1066530711030069

ZHANG, K. (2019). BET on independence. *J. Amer. Statist. Assoc.* **114** 1620–1637. MR4047288 https://doi.org/10.1080/01621459.2018.1537921

ZHANG, K., PETERS, J., JANZING, D. and SCHÖLKOPF, B. (2011). Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence* (*UAI*) 804–813. AUAI Press, USA.

The Annals of Statistics
2022, Vol. 50, No. 2, 880–903
https://doi.org/10.1214/21-AOS2130

SPARSE HIGH-DIMENSIONAL LINEAR REGRESSION. ESTIMATING SQUARED ERROR AND A PHASE TRANSITION

BY DAVID GAMARNIK[a] AND ILIAS ZADIK[b]

Massachusetts Institute of Technology, [a]*gamarnik@mit.com,* [b]*izadik@mit.com*

We consider a sparse high-dimensional regression model where the goal is to recover a k-sparse unknown binary vector β^* from n noisy linear observations of the form $Y = X\beta^* + W \in \mathbb{R}^n$ where $X \in \mathbb{R}^{n\times p}$ has i.i.d. $N(0,1)$ entries and $W \in \mathbb{R}^n$ has i.i.d. $N(0,\sigma^2)$ entries. In the high signal-to-noise ratio regime and sublinear sparsity regime, while the order of the sample size needed to recover the unknown vector information-theoretically is known to be $n^* := 2k \log p/\log(k/\sigma^2 + 1)$, no polynomial-time algorithm is known to succeed unless $n > n_{\text{alg}} := (2k + \sigma^2)\log p$.

In this work, we offer a series of results investigating multiple computational and statistical aspects of the recovery task in the regime $n \in [n^*, n_{\text{alg}}]$. First, we establish a novel information-theoretic property of the MLE of the problem happening around $n = n^*$ samples, which we coin as an "all-or-nothing behavior": when $n > n^*$ it recovers almost perfectly the support of β^*, while if $n < n^*$ it fails to recover any fraction of it correctly. Second, at an attempt to understand the computational hardness in the regime $n \in [n^*, n_{\text{alg}}]$, we prove that at order n_{alg} samples there is an Overlap Gap Property (OGP) phase transition occurring at the landscape of the MLE: for constants $c, C > 0$ when $n < cn_{\text{alg}}$ OGP appears in the landscape of MLE while if $n > Cn_{\text{alg}}$ OGP disappears. OGP is a geometric "disconnectivity" property, which initially appeared in the theory of spin glasses and is known to suggest algorithmic hardness when it occurs. Finally, using certain technical results obtained to establish the OGP phase transition, we additionally establish various novel positive and negative algorithmic results for the recovery task of interest, including the failure of LASSO with access to $n < cn_{\text{alg}}$ samples and the success of a simple local search method with access to $n > Cn_{\text{alg}}$ samples.

REFERENCES

[1] ACHLIOPTAS, D. and COJA-OGHLAN, A. (2008). Algorithmic barriers from phase transitions. In *Foundations of Computer Science*, 2008. *FOCS'*08 *IEEE* 49*th Annual IEEE Symposium on* 793–802. IEEE, New York.

[2] ACHLIOPTAS, D., COJA-OGHLAN, A. and RICCI-TERSENGHI, F. (2011). On the solution-space geometry of random constraint satisfaction problems. *Random Structures Algorithms* **38** 251–268. MR2663730 https://doi.org/10.1002/rsa.20323

[3] ARIAS-CASTRO, E., CANDÈS, E. J. and PLAN, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *Ann. Statist.* **39** 2533–2556. MR2906877 https://doi.org/10.1214/11-AOS910

[4] BALISTER, P., BOLLOBÁS, B., SAHASRABUDHE, J. and VEREMYEV, A. (2019). Dense subgraphs in random graphs. *Discrete Appl. Math.* **260** 66–74. MR3944609 https://doi.org/10.1016/j.dam.2019.01.032

[5] BANKS, J., MOORE, C., VERSHYNIN, R., VERZELEN, N. and XU, J. (2018). Information-theoretic bounds and phase transitions in clustering, sparse PCA, and submatrix localization. *IEEE Trans. Inf. Theory* **64** 4872–4994. MR3819345 https://doi.org/10.1109/tit.2018.2810020

[6] BARANIUK, R., DAVENPORT, M., DEVORE, R. and WAKIN, M. (2008). A simple proof of the restricted isometry property for random matrices. *Constr. Approx.* **28** 253–263. MR2453366 https://doi.org/10.1007/s00365-007-9003-x

MSC2020 subject classifications. Primary 62J05; secondary 62B10, 68Q87.

Key words and phrases. High-dimensional linear regression, sparsity, overlap gap property, all-or-nothing phenomenon, second moment method.

[7] BARBIER, J. and KRZAKALA, F. (2014). Replica analysis and approximate message passing decoder for superposition codes. In 2014 *IEEE International Symposium on Information Theory* 1494–1498.

[8] BARBIER, J., MACRIS, N. and RUSH, C. (2020). All-or-nothing statistical and computational phase transitions in sparse spiked matrix estimation.

[9] BEN ATITALLAH, I., THRAMPOULIDIS, C., KAMMOUN, A., AL-NAFFOURI, T. Y., ALOUINI, M. and HASSIBI, B. (2017). The box-lasso with application to gssk modulation in massive mimo systems. In 2017 *IEEE International Symposium on Information Theory* (*ISIT*) 1082–1086.

[10] BERTSIMAS, D. and VAN PARYS, B. (2020). Sparse high-dimensional regression: Exact scalable algorithms and phase transitions. *Ann. Statist.* **48** 300–323. MR4065163 https://doi.org/10.1214/18-AOS1804

[11] BICKEL, P. J., BROWN, J. B., HUANG, H. and LI, Q. (2009). An overview of recent developments in genomics and associated statistical methods. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **367** 4313–4337. MR2546390 https://doi.org/10.1098/rsta.2009.0164

[12] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469 https://doi.org/10.1214/08-AOS620

[13] BLUMENSATH, T. and DAVIES, M. E. (2009). Iterative hard thresholding for compressed sensing. *Appl. Comput. Harmon. Anal.* **27** 265–274. MR2559726 https://doi.org/10.1016/j.acha.2009.04.002

[14] BRUNEL, L. and BOUTROS, J. J. (1999). Euclidean space lattice decoding for joint detection in CDMA systems. In *IEEE Information Theory and Communications Workshop*.

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

[16] CAI, T. T. and GUO, Z. (2018). Accuracy assessment for high-dimensional linear regression. *Ann. Statist.* **46** 1807–1836. MR3819118 https://doi.org/10.1214/17-AOS1604

[17] CAI, T. T. and WANG, L. (2011). Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Trans. Inf. Theory* **57** 4680–4688. MR2840484 https://doi.org/10.1109/TIT.2011.2146090

[18] CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644 https://doi.org/10.1214/009053606000001523

[19] CANDÈS, E. J., ROMBERG, J. K. and TAO, T. (2006). Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.* **59** 1207–1223. MR2230846 https://doi.org/10.1002/cpa.20124

[20] CANDES, E. J. and TAO, T. (2005). Decoding by linear programming. *IEEE Trans. Inf. Theory* **51** 4203–4215. MR2243152 https://doi.org/10.1109/TIT.2005.858979

[21] CARVALHO, C. M., CHANG, J., LUCAS, J. E., NEVINS, J. R., WANG, Q. and WEST, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *J. Amer. Statist. Assoc.* **103** 1438–1456. MR2655722 https://doi.org/10.1198/016214508000000869

[22] CHEN, V. C. and LING, H. (1999). Joint time-frequency analysis for radar signal and image processing. *IEEE Signal Process. Mag.* **16** 81–93.

[23] COJA-OGHLAN, A. and EFTHYMIOU, C. (2011). On independent sets in random graphs. In *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms* 136–144. SIAM, Philadelphia, PA. MR2857116

[24] COVER, T. M. and THOMAS, J. A. (2006). *Elements of Information Theory*, 2nd ed. Wiley Interscience, Hoboken, NJ. MR2239987

[25] DONOHO, D. L. (2006). Compressed sensing. *IEEE Trans. Inf. Theory* **52** 1289–1306. MR2241189 https://doi.org/10.1109/TIT.2006.871582

[26] DONOHO, D. L., JAVANMARD, A. and MONTANARI, A. (2013). Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing. *IEEE Trans. Inf. Theory* **59** 7434–7464. MR3124654 https://doi.org/10.1109/TIT.2013.2274513

[27] DONOHO, D. L. and TANNER, J. (2010). Counting the faces of randomly-projected hypercubes and orthants, with applications. *Discrete Comput. Geom.* **43** 522–541. MR2587835 https://doi.org/10.1007/s00454-009-9221-z

[28] DUDCZYK, J. (2017). A method of feature selection in the aspect of specific identification of radar signals. *Bull. Pol. Acad. Sci., Tech. Sci.* **65** 113–119.

[29] 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

[30] GAMARNIK, D. and LI, Q. (2018). Finding a large submatrix of a Gaussian random matrix. *Ann. Statist.* **46** 2511–2561. MR3851747 https://doi.org/10.1214/17-AOS1628

[31] GAMARNIK, D. and SUDAN, M. (2017). Limits of local algorithms over sparse random graphs. *Ann. Probab.* **45** 2353–2376. MR3693964 https://doi.org/10.1214/16-AOP1114

[32] GAMARNIK, D. and SUDAN, M. (2017). Performance of sequential local algorithms for the random NAE-K-SAT problem. *SIAM J. Comput.* **46** 590–619. MR3620150 https://doi.org/10.1137/140989728

[33] GAMARNIK, D. and ZADIK, I. (2018). High dimensional linear regression using lattice basis reduction. In *Neural Information Processing Systems* (*NeurIPS*).

[34] GAMARNIK, D. and ZADIK, I. (2022). Supplement to "Sparse high-dimensional linear regression. Estimating squared error and a phase transition." https://doi.org/10.1214/21-AOS2130SUPP

[35] GEORGE, E. I. (2000). The variable selection problem. *J. Amer. Statist. Assoc.* **95** 1304–1308. MR1825282 https://doi.org/10.2307/2669776

[36] HANSON, D. L. and WRIGHT, F. T. (1971). A bound on tail probabilities for quadratic forms in independent random variables. *Ann. Math. Stat.* **42** 1079–1083. MR0279864 https://doi.org/10.1214/aoms/1177693335

[37] HASSIBI, A. and BOYD, S. (1998). Integer parameter estimation in linear models with applications to GPS. *IEEE Trans. Signal Process.* **46** 2938–2952. MR1719935 https://doi.org/10.1109/78.726808

[38] HASSIBI, B. and VIKALO, H. (2002). On the expected complexity of integer least-squares problems. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*.

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

[40] HO, J. W. K., STEFANI, M., DOS REMEDIOS, C. G. and CHARLESTON, M. A. (2008). Differential variability analysis of gene expression and its application to human diseases. *Bioinformatics* **24** i390–i398.

[41] HU, R., QIU, X. and GLAZKO, G. (2010). A new gene selection procedure based on the covariance distance. *Bioinformatics* **26** 348–354.

[42] HU, R., QIU, X., GLAZKO, G. KLEVANOV, L. and YAKOVLEV, A. (2009). Detecting intergene correlation changes in microarray analysis: A new approach to gene selection. *BMC Bioinform.* **10** 20.

[43] JANSON, L., FOYGEL BARBER, R. and CANDÈS, E. (2017). EigenPrism: Inference for high dimensional signal-to-noise ratios. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1037–1065. MR3689308 https://doi.org/10.1111/rssb.12203

[44] JOSEPH, A. and BARRON, A. R. (2012). Least squares superposition codes of moderate dictionary size are reliable at rates up to capacity. *IEEE Trans. Inf. Theory* **58** 2541–2557. MR2952500 https://doi.org/10.1109/TIT.2012.2184847

[45] JOSEPH, A. and BARRON, A. R. (2014). Fast sparse superposition codes have near exponential error probability for $R < \mathcal{C}$. *IEEE Trans. Inf. Theory* **60** 919–942. MR3164953 https://doi.org/10.1109/TIT.2013.2289865

[46] LAN, TRUONG, V., ALDRIDGE, M. and SCARLETT, J. (2020). On the all-or-nothing behavior of Bernoulli group testing.

[47] LUNEAU, C., BARBIER, J. and MACRIS, N. (2020). Information theoretic limits of learning a sparse rule.

[48] LUSTIG, M., DONOHO, D. L., SANTOS, J. M. and PAULY, J. M. (2008). Compressed sensing MRI. *IEEE Signal Process. Mag.* **25** 72–82.

[49] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363 https://doi.org/10.1214/009053606000000281

[50] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363 https://doi.org/10.1214/009053606000000281

[51] MONTANARI, A., RESTREPO, R. and TETALI, P. (2011). Reconstruction and clustering in random constraint satisfaction problems. *SIAM J. Discrete Math.* **25** 771–808. MR2823097 https://doi.org/10.1137/090755862

[52] NICKL, R. and VAN DE GEER, S. (2013). Confidence sets in sparse regression. *Ann. Statist.* **41** 2852–2876. MR3161450 https://doi.org/10.1214/13-AOS1170

[53] NILES-WEED, J. and ZADIK, I. (2020). The all-or-nothing phenomenon in sparse tensor PCA.

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

[55] OYMAK, S., THRAMPOULIDIS, C. and HASSIBI, B. (2013). The squared-error of generalized lasso: A precise analysis. In 51*st Annual Allerton Conference on Communication, Control, and Computing* (*Allerton*) 1002–1009.

[56] PENG, B., ZHAO, Z., HAN, G. and SHEN, J. (2016). Consensus-based sparse signal reconstruction algorithm for wireless sensor networks. *Int. J. Distrib. Sens. Netw.* **12**.

[57] QUER, G., MASIERO, R., PILLONETTO, G., ROSSI, M. and ZORZI, M. (2012). Sensing, compression, and recovery for WSNs: Sparse signal modeling and monitoring framework. *IEEE Trans. Wirel. Commun.* **11** 3447–3461.

[58] RAD, K. R. (2011). Nearly sharp sufficient conditions on exact sparsity pattern recovery. *IEEE Trans. Inf. Theory* **57** 4672–4679. MR2840483 https://doi.org/10.1109/TIT.2011.2145670

[59] RAHMAN, M. and VIRÁG, B. (2017). Local algorithms for independent sets are half-optimal. *Ann. Probab.* **45** 1543–1577. MR3650409 https://doi.org/10.1214/16-AOP1094

[60] REEVES, G. and GASTPAR, M. (2012). The sampling rate-distortion tradeoff for sparsity pattern recovery in compressed sensing. *IEEE Trans. Inf. Theory* **58** 3065–3092. MR2952533 https://doi.org/10.1109/TIT.2012.2184848

[61] REEVES, G. and GASTPAR, M. C. (2013). Approximate sparsity pattern recovery: Information-theoretic lower bounds. *IEEE Trans. Inf. Theory* **59** 3451–3465. MR3061258 https://doi.org/10.1109/TIT.2013.2253852

[62] REEVES, G., XU, J. and ZADIK, I. (2019). All-or-nothing phenomena: From single-letter to high dimensions. In 8*th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, *CAMSAP* 2019, *Le Gosier*, *Guadeloupe*, *December* 15–18, 2019 654–658. IEEE, New York.

[63] REEVES, G., XU, J. and ZADIK, I. (2019). The all-or-nothing phenomenon in sparse linear regression. In *Proceedings of Machine Learning Research*, 25–28 *Jun* 2019 **99** 2652–2663. PMLR.

[64] RUSH, C., GREIG, A. and VENKATARAMANAN, R. (2017). Capacity-achieving sparse superposition codes via approximate message passing decoding. *IEEE Trans. Inf. Theory* **63** 1476–1500. MR3625975 https://doi.org/10.1109/TIT.2017.2649460

[65] RUSH, C. and VENKATARAMANAN, R. (2019). The error probability of sparse superposition codes with approximate message passing decoding. *IEEE Trans. Inf. Theory* **65** 3278–3303. MR3951396 https://doi.org/10.1109/TIT.2018.2882177

[66] SCARLETT, J. and CEVHER, V. (2017). Limits on support recovery with probabilistic models: An information-theoretic framework. *IEEE Trans. Inf. Theory* **63** 593–620. MR3599962 https://doi.org/10.1109/TIT.2016.2606605

[67] SHI, Y., ZHANG, X. and BAO, Z. (2001). A new feature vector using selected bispectra for signal classification with application in radar target recognition. *IEEE Trans. Signal Process.* **49** 1875–1885.

[68] THRAMPOULIDIS, C., ZADIK, I. and POLYANSKYI, Y. (2019). A simple bound on the ber of the map decoder for massive mimo systems. In *IEEE International Conference on Acoustics*, *Speech*, *and Signal Processing*.

[69] 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

[70] VENKATARAMANAN, R., TATIKONDA, S. and BARRON, A. (2019). Sparse regression codes. *Found. Trends Commun. Inf. Theory* **15** 1–195.

[71] WAINWRIGHT, M. J. (2009). Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inf. Theory* **55** 5728–5741. MR2597190 https://doi.org/10.1109/TIT.2009.2032816

[72] WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inf. Theory* **55** 2183–2202. MR2729873 https://doi.org/10.1109/TIT.2009.2016018

[73] WANG, W., WAINWRIGHT, M. J. and RAMCHANDRAN, K. (2010). Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices. *IEEE Trans. Inf. Theory* **56** 2967–2979. MR2683451 https://doi.org/10.1109/TIT.2010.2046199

[74] ZADIK, I., THRAMPOULIDIS, C. and POLYANSKYI, Y. (2019). Improved bounds on Gaussian mac and sparse regression via Gaussian inequalities. In *IEEE International Symposium on Information Theory* (*ISIT*).

[75] ZHANG, P. (1993). Model selection via multifold cross validation. *Ann. Statist.* **21** 299–313. MR1212178 https://doi.org/10.1214/aos/1176349027

[76] ZHAO, B., LU, W., HITCHENS, T. K., LAM, F., HO, C. and LIANG, Z.-P. (2015). Accelerated MR parameter mapping with low-rank and sparsity constraints. *Magn. Reson. Med.* **74** 489–498. https://doi.org/10.1002/mrm.25421

[77] ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

The Annals of Statistics
2022, Vol. 50, No. 2, 904–929
https://doi.org/10.1214/21-AOS2131

FUNCTIONAL SUFFICIENT DIMENSION REDUCTION THROUGH AVERAGE FRÉCHET DERIVATIVES

BY KUANG-YAO LEE[1,a] AND LEXIN LI[2,b]

[1]*Department of Statistical Science, Temple University,* [a]*kuang-yao.lee@temple.edu*
[2]*Department of Biostatistics, University of California, Berkeley,* [b]*lexinli@berkeley.edu*

Sufficient dimension reduction (SDR) embodies a family of methods that aim for reduction of dimensionality without loss of information in a regression setting. In this article, we propose a new method for nonparametric function-on-function SDR, where both the response and the predictor are a function. We first develop the notions of functional central mean subspace and functional central subspace, which form the population targets of our functional SDR. We then introduce an average Fréchet derivative estimator, which extends the gradient of the regression function to the operator level and enables us to develop estimators for our functional dimension reduction spaces. We show the resulting functional SDR estimators are unbiased and exhaustive, and more importantly, without imposing any distributional assumptions such as the linearity or the constant variance conditions that are commonly imposed by all existing functional SDR methods. We establish the uniform convergence of the estimators for the functional dimension reduction spaces, while allowing both the number of Karhunen–Loève expansions and the intrinsic dimension to diverge with the sample size. We demonstrate the efficacy of the proposed methods through both simulations and two real data examples.

REFERENCES

AMINI, A. A. and WAINWRIGHT, M. J. (2012). Sampled forms of functional PCA in reproducing kernel Hilbert spaces. *Ann*. *Statist*. **40** 2483–2510. MR3097610 https://doi.org/10.1214/12-AOS1033

CONWAY, J. B. (2010). *A Course in Functional Analysis*, 2nd ed. Springer, New York.

COOK, R. D. and LI, B. (2002). Dimension reduction for conditional mean in regression. *Ann*. *Statist*. **30** 455–474. MR1902895 https://doi.org/10.1214/aos/1021379861

COOK, R. D. and WEISBERG, S. (1991). Discussion of "Sliced inverse regression for dimension reduction". *J*. *Amer*. *Statist*. *Assoc*. **86** 328–332.

FANAEE-T, H. and GAMA, J. (2014). Event labeling combining ensemble detectors and background knowledge. *Prog*. *Artif*. *Intell*. **2** 113–127. https://doi.org/10.1007/s13748-013-0040-3

FERRÉ, L. and YAO, A. F. (2003). Functional sliced inverse regression analysis. *Statistics* **37** 475–488. MR2022235 https://doi.org/10.1080/0233188031000112845

FERRÉ, L. and YAO, A.-F. (2005). Smoothed functional inverse regression. *Statist*. *Sinica* **15** 665–683. MR2233905

FUKUMIZU, K., BACH, F. R. and JORDAN, M. I. (2009). Kernel dimension reduction in regression. *Ann*. *Statist*. **37** 1871–1905. MR2533474 https://doi.org/10.1214/08-AOS637

FUKUMIZU, K. and LENG, C. (2014). Gradient-based kernel dimension reduction for regression. *J*. *Amer*. *Statist*. *Assoc*. **109** 359–370. MR3180569 https://doi.org/10.1080/01621459.2013.838167

HARDLE, W. (1989). Investigating smooth multiple regression by the method of average derivatives. *J*. *Amer*. *Statist*. *Assoc*. **84** 986–995.

HÄRDLE, W. and STOKER, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *J*. *Amer*. *Statist*. *Assoc*. **84** 986–995. MR1134488

HSING, T. (1999). Nearest neighbor inverse regression. *Ann*. *Statist*. **27** 697–731. MR1714711 https://doi.org/10.1214/aos/1018031213

MSC2020 subject classifications. 62B05, 62G08, 62G20, 62R10.
Key words and phrases. Functional central mean subspace, functional central subspace, function-on-function regression, unbiasedness, exhaustiveness, consistency, reproducing kernel Hilbert space.

HSING, T. and REN, H. (2009). An RKHS formulation of the inverse regression dimension-reduction problem. *Ann. Statist.* **37** 726–755. MR2502649 https://doi.org/10.1214/07-AOS589

JIANG, C.-R., YU, W. and WANG, J.-L. (2014). Inverse regression for longitudinal data. *Ann. Statist.* **42** 563–591. MR3210979 https://doi.org/10.1214/13-AOS1193

KIM, J. S., STAICU, A.-M., MAITY, A., CARROLL, R. J. and RUPPERT, D. (2018). Additive function-on-function regression. *J. Comput. Graph. Statist.* **27** 234–244. MR3788315 https://doi.org/10.1080/10618600.2017.1356730

LEE, Y.-J. and HUANG, S.-Y. (2007). Reduced support vector machines: A statistical theory. *IEEE Trans. Neural Netw.* **18** 1–13. https://doi.org/10.1109/TNN.2006.883722

LEE, K.-Y. and LI, L. (2022). Supplement to "Functional sufficient dimension reduction through average Fréchet derivatives." https://doi.org/10.1214/21-AOS2131SUPP

LI, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86** 316–342. MR1137117

LI, B. (2018a). *Sufficient Dimension Reduction: Methods and Applications with R. Monographs on Statistics and Applied Probability* **161**. CRC Press, Boca Raton, FL. MR3838449 https://doi.org/10.1201/9781315119427

LI, B. (2018b). Linear operator-based statistical analysis: A useful paradigm for big data. *Canad. J. Statist.* **46** 79–103. MR3767167 https://doi.org/10.1002/cjs.11329

LI, B. and SONG, J. (2017). Nonlinear sufficient dimension reduction for functional data. *Ann. Statist.* **45** 1059–1095. MR3662448 https://doi.org/10.1214/16-AOS1475

LI, B. and WANG, S. (2007). On directional regression for dimension reduction. *J. Amer. Statist. Assoc.* **102** 997–1008. MR2354409 https://doi.org/10.1198/016214507000000536

LI, B., WEN, S. and ZHU, L. (2008). On a projective resampling method for dimension reduction with multivariate responses. *J. Amer. Statist. Assoc.* **103** 1177–1186. MR2462891 https://doi.org/10.1198/016214508000000445

LI, B., ZHA, H. and CHIAROMONTE, F. (2005). Contour regression: A general approach to dimension reduction. *Ann. Statist.* **33** 1580–1616. MR2166556 https://doi.org/10.1214/009053605000000192

LIN, Q., ZHAO, Z. and LIU, J. S. (2019). Sparse sliced inverse regression via Lasso. *J. Amer. Statist. Assoc.* **114** 1726–1739. MR4047295 https://doi.org/10.1080/01621459.2018.1520115

LIN, Q., LI, X., HUANG, D. and LIU, J. S. (2017). On the optimality of sliced inverse regression in high dimensions.

LUO, R. and QI, X. (2017). Function-on-function linear regression by signal compression. *J. Amer. Statist. Assoc.* **112** 690–705. MR3671763 https://doi.org/10.1080/01621459.2016.1164053

LUO, R. and QI, X. (2019). Interaction model and model selection for function-on-function regression. *J. Comput. Graph. Statist.* **28** 309–322. MR3974882 https://doi.org/10.1080/10618600.2018.1514310

MA, Y. and ZHU, L. (2012). A semiparametric approach to dimension reduction. *J. Amer. Statist. Assoc.* **107** 168–179. MR2949349 https://doi.org/10.1080/01621459.2011.646925

MA, Y. and ZHU, L. (2013). Efficient estimation in sufficient dimension reduction. *Ann. Statist.* **41** 250–268. MR3059417 https://doi.org/10.1214/12-AOS1072

MÜLLER, H.-G. and YAO, F. (2008). Functional additive models. *J. Amer. Statist. Assoc.* **103** 1534–1544. MR2504202 https://doi.org/10.1198/016214508000000751

RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2168993

REIMHERR, M., SRIPERUMBUDUR, B. and TAOUFIK, B. (2018). Optimal prediction for additive function-on-function regression. *Electron. J. Stat.* **12** 4571–4601. MR3893421 https://doi.org/10.1214/18-EJS1505

STEINWART, I. and CHRISTMANN, A. (2008). *Support Vector Machines*. Springer, New York.

SUN, X., DU, P., WANG, X. and MA, P. (2018). Optimal penalized function-on-function regression under a reproducing kernel Hilbert space framework. *J. Amer. Statist. Assoc.* **113** 1601–1611. MR3902232 https://doi.org/10.1080/01621459.2017.1356320

WANG, J. L., CHIOU, J. M. and MULLER, H. G. (2016). Functional data analysis. *Annu. Rev. Stat. Appl.* **3** 257–295.

WANG, G., LIN, N. and ZHANG, B. (2013). Functional contour regression. *J. Multivariate Anal.* **116** 1–13. MR3049886 https://doi.org/10.1016/j.jmva.2012.11.005

WANG, G., ZHOU, Y., FENG, X.-N. and ZHANG, B. (2015). The hybrid method of FSIR and FSAVE for functional effective dimension reduction. *Comput. Statist. Data Anal.* **91** 64–77. MR3368006 https://doi.org/10.1016/j.csda.2015.05.011

WEIDMANN, J. (1980). *Linear Operators in Hilbert Spaces. Graduate Texts in Mathematics* **68**. Springer, New York–Berlin. MR0566954

XIA, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *Ann. Statist.* **35** 2654–2690. MR2382662 https://doi.org/10.1214/009053607000000352

XIA, Y., TONG, H., LI, W. K. and ZHU, L.-X. (2002). An adaptive estimation of dimension reduction space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 363–410. MR1924297 https://doi.org/10.1111/1467-9868.03411
YAO, F., LEI, E. and WU, Y. (2015). Effective dimension reduction for sparse functional data. *Biometrika* **102** 421–437. MR3371014 https://doi.org/10.1093/biomet/asv006
YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005). Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33** 2873–2903. MR2253106 https://doi.org/10.1214/009053605000000660
YIN, X. and LI, B. (2011). Sufficient dimension reduction based on an ensemble of minimum average variance estimators. *Ann. Statist.* **39** 3392–3416. MR3012413 https://doi.org/10.1214/11-AOS950
YIN, X., LI, B. and COOK, R. D. (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. *J. Multivariate Anal.* **99** 1733–1757. MR2444817 https://doi.org/10.1016/j.jmva.2008.01.006

The Annals of Statistics
2022, Vol. 50, No. 2, 930–948
https://doi.org/10.1214/21-AOS2132

GENERAL AND FEASIBLE TESTS WITH MULTIPLY-IMPUTED DATASETS

BY KIN WAI CHAN[a]

Department of Statistics, The Chinese University of Hong Kong, [a]*kinwaichan@cuhk.edu.hk*

Multiple imputation (MI) is a technique especially designed for handling missing data in public-use datasets. It allows analysts to perform incomplete-data inference straightforwardly by using several already imputed datasets released by the dataset owners. However, the existing MI tests require either a restrictive assumption on the missing-data mechanism, known as equal odds of missing information (EOMI), or an infinite number of imputations. Some of them also require analysts to have access to restrictive or nonstandard computer subroutines. Besides, the existing MI testing procedures cover only Wald's tests and likelihood ratio tests but not Rao's score tests, therefore, these MI testing procedures are not general enough. In addition, the MI Wald's tests and MI likelihood ratio tests are not procedurally identical, so analysts need to resort to distinct algorithms for implementation. In this paper, we propose a general MI procedure, called stacked multiple imputation (SMI), for performing Wald's tests, likelihood ratio tests and Rao's score tests by a unified algorithm. SMI requires neither EOMI nor an infinite number of imputations. It is particularly feasible for analysts as they just need to use a complete-data testing device for performing the corresponding incomplete-data test.

REFERENCES

AGRESTI, A. (2015). *Foundations of Linear and Generalized Linear Models*. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR3308143

BARNARD, J. and RUBIN, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika* **86** 948–955. MR1741991 https://doi.org/10.1093/biomet/86.4.948

BERA, A. K. and BILIAS, Y. (2001). Rao's score, Neyman's $C(\alpha)$ and Silvey's LM tests: An essay on historical developments and some new results. *J. Statist. Plann. Inference* **97** 9–44.

BIRNBAUM, A. (1961). Confidence curves: An omnibus technique for estimation and testing statistical hypotheses. *J. Amer. Statist. Assoc.* **56** 246–249. MR0121904

CARPENTER, J. R. and KENWARD, M. G. (2013). *Multiple Imputation and Its Application*. Wiley, New York.

CHAN, K. W. (2022). Supplement to "General and feasible tests with multiply-imputed datasets." https://doi.org/10.1214/21-AOS2132SUPP

CHAN, K. W. and MENG, X.-L. (2021+). Multiple improvements of multiple imputation likelihood ratio tests. *Statist. Sinica* To appear.

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26. MR0515681

FISHER, R. A. (1934). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh. 5 edn.

FRASER, D. A. S. (1991). Statistical inference: Likelihood to significance. *J. Amer. Statist. Assoc.* **86** 258–265. MR1137116

FRASER, D. A. S. (2019). The p-value function and statistical inference. *Amer. Statist.* **73** 135–147. MR3925719 https://doi.org/10.1080/00031305.2018.1556735

HAREL, O. and ZHOU, X.-H. (2007). Multiple imputation: Review of theory, implementation and software. *Stat. Med.* **26** 3057–3077. MR2380504 https://doi.org/10.1002/sim.2787

HEARD, N. A. and RUBIN-DELANCHY, P. (2018). Choosing between methods of combining p-values. *Biometrika* **105** 239–246. MR3768879 https://doi.org/10.1093/biomet/asx076

HOLAN, S. H., TOTH, D., FERREIRA, M. A. R. and KARR, A. F. (2010). Bayesian multiscale multiple imputation with implications for data confidentiality. *J. Amer. Statist. Assoc.* **105** 564–577. MR2759932 https://doi.org/10.1198/jasa.2009.ap08629

MSC2020 subject classifications. Primary 62D05; secondary 62F03, 62E20.

Key words and phrases. Fraction of missing information, hypothesis testing, jackknife, missing data, stacking.

HORTON, N. J. and KLEINMAN, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *Amer. Statist.* **61** 79–90. MR2339152 https://doi.org/10.1198/000313007X172556

INFANGER, D. and SCHMIDT-TRUCKSÄSS, A. (2019). P value functions: An underused method to present research results and to promote quantitative reasoning. *Stat. Med.* **38** 4189–4197. MR3999273 https://doi.org/10.1002/sim.8293

KENWARD, M. G. and CARPENTER, J. (2007). Multiple imputation: Current perspectives. *Stat. Methods Med. Res.* **16** 199–218. MR2371006 https://doi.org/10.1177/0962280206075304

KIM, J. K. and SHAO, J. (2013). *Statistical Methods for Handling Incomplete Data*. CRC Press/CRC, Boca Raton, FL.

KIM, J. K. and YANG, S. (2017). A note on multiple imputation under complex sampling. *Biometrika* **104** 221–228. MR3626470 https://doi.org/10.1093/biomet/asw058

KING, G., HONAKER, J., JOSEPH, A. and SCHEVE, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *Am. Polit. Sci. Rev.* **95** 49–69.

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

LI, K. H., RAGHUNATHAN, T. E. and RUBIN, D. B. (1991). Large-sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution. *J. Amer. Statist. Assoc.* **86** 1065–1073. MR1146352

MARTIN, R. (2015). Plausibility functions and exact frequentist inference. *J. Amer. Statist. Assoc.* **110** 1552–1561. MR3449054 https://doi.org/10.1080/01621459.2014.983232

MARTIN, R. (2017). A statistical inference course based on p-values. *Amer. Statist.* **71** 128–136. MR3668699 https://doi.org/10.1080/00031305.2016.1208629

MENG, X.-L. (1990). *Towards Complete Results for Some Incomplete-Data Problems*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)—Harvard University. MR2685717

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

MENG, X.-L. and RUBIN, D. B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika* **79** 103–111. MR1158520 https://doi.org/10.1093/biomet/79.1.103

PARKER, J. D. and SCHENKER, N. (2007). Multiple imputation for national public-use datasets and its possible application for gestational age in United States Natality files. *Paediatr. Perinat. Epidemiol.* **21** 97–105. https://doi.org/10.1111/j.1365-3016.2007.00866.x

PEUGH, J. L. and ENDERS, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Rev. Educ. Res.* **74** 525–556.

POLITIS, D. N., ROMANO, J. P. and WOLF, M. (1999). *Subsampling*. *Springer Series in Statistics*. Springer, New York. MR1707286 https://doi.org/10.1007/978-1-4612-1554-7

QUENOUILLE, M. H. (1956). Notes on bias in estimation. *Biometrika* **43** 353–360. MR0081040 https://doi.org/10.1093/biomet/43.3-4.353

RAO, C. R. (2005). Score test: Historical review and recent developments. In *Advances in Ranking and Selection, Multiple Comparisons, and Reliability*. *Stat. Ind. Technol.* 3–20. Birkhäuser, Boston, MA. MR2111500 https://doi.org/10.1007/0-8176-4422-9_1

REITER, J. P. (2007). Small-sample degrees of freedom for multi-component significance tests for multiple imputation for missing data. *Biometrika* **94** 502–508. MR2380575 https://doi.org/10.1093/biomet/asm028

ROSE, R. A. and FRASER, M. W. (2008). A simplified framework for using multiple imputation in social work research. *Soc. Work Res.* **32** 171–178.

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

RUBIN, D. B. (1978). Multiple imputations in sample surveys—a phenomenological Bayesian approach to nonresponse. *Proc. Surv. Res. Methods Sect. Am. Stat. Assoc.* 20–34.

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

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

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

SCHAFER, J. L. (1999). Multiple imputation: A primer. *Stat. Methods Med. Res.* **8** 3–15. https://doi.org/10.1177/096228029900800102

SCHENKER, N., RAGHUNATHAN, T. E., CHIU, P.-L., MAKUC, D. M., ZHANG, G. and COHEN, A. J. (2006). Multiple imputation of missing income data in the National Health interview survey. *J. Amer. Statist. Assoc.* **101** 924–933. MR2324093 https://doi.org/10.1198/016214505000001375

SERFLING, R. J. (2001). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

SHAO, J. (1999). *Mathematical Statistics*. *Springer Texts in Statistics*. Springer, New York. MR1670883

TU, X. M., MENG, X.-L. and PAGANO, M. (1993). The AIDS epidemic: Estimating survival after AIDS diagnosis from surveillance data. *J. Amer. Statist. Assoc.* **88** 26–36.

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

VAN DER VAART, A. W. (2000). *Asymptotic Statistics*. Cambridge Univ. Press, Cambridge.

WANG, N. and ROBINS, J. M. (1998). Large-sample theory for parametric multiple imputation procedures. *Biometrika* **85** 935–948. MR1666715 https://doi.org/10.1093/biomet/85.4.935

XIE, X. and MENG, X.-L. (2017). Dissecting multiple imputation from a multi-phase inference perspective: What happens when God's, imputer's and analyst's models are uncongenial? (with discussion). *Statist. Sinica* **27** 1485–1594.

XIE, M. and SINGH, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review. *Int. Stat. Rev.* **81** 3–39. MR3047496 https://doi.org/10.1111/insr.12000

XIE, M., SINGH, K. and STRAWDERMAN, W. E. (2011). Confidence distributions and a unifying framework for meta-analysis. *J. Amer. Statist. Assoc.* **106** 320–333. MR2816724 https://doi.org/10.1198/jasa.2011.tm09803

YU, P. S. Y., CHAN, K. W., LAU, R. W. H., WAN, I. Y. P., CHEN, G. G. and NG, C. S. H. (2021). Uniportal video-assisted thoracic surgery for major lung resection is associated with less immunochemokine disturbances than multiportal approach. *Sci. Rep.* **11** 10369.

The Annals of Statistics
2022, Vol. 50, No. 2, 949–986
https://doi.org/10.1214/21-AOS2133

SURPRISES IN HIGH-DIMENSIONAL RIDGELESS LEAST SQUARES INTERPOLATION

BY TREVOR HASTIE[1,a], ANDREA MONTANARI[2,b], SAHARON ROSSET[3,c] AND RYAN J. TIBSHIRANI[4,d]

[1]*Department of Statistics and Department of Biomedical Data Science, Stanford University,* [a]*hastie@stanford.edu*

[2]*Department of Statistics and Department of Electrical Engineering, Stanford University,* [b]*montanar@stanford.edu*

[3]*School of Mathematical Sciences, Tel Aviv University,* [c]*saharon@post.tau.ac.il*

[4]*Department of Statistics and Department of Machine Learning, Carnegie Mellon University,* [d]*ryantibs@cmu.edu*

Interpolators—estimators that achieve zero training error—have attracted growing attention in machine learning, mainly because state-of-the art neural networks appear to be models of this type. In this paper, we study minimum ℓ_2 norm ("ridgeless") interpolation least squares regression, focusing on the high-dimensional regime in which the number of unknown parameters p is of the same order as the number of samples n. We consider two different models for the feature distribution: a linear model, where the feature vectors $x_i \in \mathbb{R}^p$ are obtained by applying a linear transform to a vector of i.i.d. entries, $x_i = \Sigma^{1/2} z_i$ (with $z_i \in \mathbb{R}^p$); and a nonlinear model, where the feature vectors are obtained by passing the input through a random one-layer neural network, $x_i = \varphi(W z_i)$ (with $z_i \in \mathbb{R}^d$, $W \in \mathbb{R}^{p \times d}$ a matrix of i.i.d. entries, and φ an activation function acting componentwise on $W z_i$). We recover—in a precise quantitative way—several phenomena that have been observed in large-scale neural networks and kernel machines, including the "double descent" behavior of the prediction risk, and the potential benefits of overparametrization.

REFERENCES

[1] ADLAM, B. and PENNINGTON, J. (2020). The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. Available at http://proceedings.mlr.press/v119/adlam20a/adlam20a.pdf.

[2] ADVANI, M. S. and SAXE, A. M. (2017). High-dimensional dynamics of generalization error in neural networks. Available at arXiv:1710.03667.

[3] ALI, A., KOLTER, J. Z. and TIBSHIRANI, R. J. (2019). A continuous-time view of early stopping for least squares. *Int. Conf. Artif. Intell. Stat.* **22**.

[4] ALLEN-ZHU, Z., LI, Y. and SONG, Z. (2019). A convergence theory for deep learning via over-parameterization. International Conference on Machine Learning. PMLR. Availablae at http://proceedings.mlr.press/v97/allen-zhu19a/allen-zhu19a.pdf.

[5] 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

[6] 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

[7] 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

[8] BELKIN, M., HSU, D., MA, S. and MANDAL, S. (2019). Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proc. Natl. Acad. Sci. USA* **116** 15849–15854. MR3997901 https://doi.org/10.1073/pnas.1903070116

[9] BELKIN, M., HSU, D. and XU, J. (2020). Two models of double descent for weak features. *SIAM J. Math. Data Sci.* **2** 1167–1180. MR4186534 https://doi.org/10.1137/20M1336072

MSC2020 subject classifications. Primary 62J05, 62J07; secondary 62J02, 62F12.
Key words and phrases. Regression, interpolation, overparametrization, ridge regression, random matrix theory.

[10] BELKIN, M., MA, S. and MANDAL, S. (2018). To understand deep learning we need to understand kernel learning. 22nd International Conference on Artificial Intelligence and Statistics. PMLR. Available at http://proceedings.mlr.press/v80/belkin18a/belkin18a.pdf.

[11] BELKIN, M., RAKHLIN, A. and TSYBAKOV, A. B. (2018). Does data interpolation contradict statistical optimality? Available at arXiv:1806.09471.

[12] BING, X., BUNEA, F., STRIMAS-MACKEY, S. and WEGKAMP, M. (2021). Prediction under latent factor regression: Adaptive PCR, interpolating predictors and beyond. *J. Mach. Learn. Res.* **22** 177. MR4318533 https://doi.org/10.22405/2226-8383-2021-22-1-177-187

[13] BISHOP, C. M. (1995). Training with noise is equivalent to Tikhonov regularization. *Neural Comput.* **7** 108–116.

[14] BUNEA, F., STRIMAS-MACKEY, S. and WEGKAMP, M. (2020). Interpolating predictors in high-dimensional factor regression. Available at arXiv:2002.02525.

[15] CHATTERJI, N. S. and LONG, P. M. (2021). Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *J. Mach. Learn. Res.* **22** 129. MR4279780

[16] CHEN, S. S., DONOHO, D. L. and SAUNDERS, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* **20** 33–61. MR1639094 https://doi.org/10.1137/S1064827596304010

[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] 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**.

[19] CHIZAT, L. and BACH, F. (2019). A note on lazy training in supervised differentiable programming. *Advances in Neural Information Processing Systems* **32** 2933–2943.

[20] CRAVEN, P. and WAHBA, G. (1978/79). Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31** 377–403. MR0516581 https://doi.org/10.1007/BF01404567

[21] DAUBECHIES, I. (1988). Time-frequency localization operators: A geometric phase space approach. *IEEE Trans. Inf. Theory* **34** 605–612. MR0966733 https://doi.org/10.1109/18.9761

[22] DICKER, L. H. (2016). Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli* **22** 1–37. MR3449775 https://doi.org/10.3150/14-BEJ609

[23] DOBRIBAN, E. and WAGER, S. (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *Ann. Statist.* **46** 247–279. MR3766952 https://doi.org/10.1214/17-AOS1549

[24] DU, S. S., LEE, J. D., LI, H., WANG, L. and ZHAI, X. (2018). Gradient descent finds global minima of deep neural networks. International conference on machine learning. PMLR.

[25] DU, S. S., ZHAI, X., POCZOS, B. and SINGH, A. (2018). Gradient descent provably optimizes over-parameterized neural networks. Available at arXiv:1810.02054.

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

[27] 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

[28] GEIGER, M., JACOT, A., SPIGLER, S., GABRIEL, F., SAGUN, L., D'ASCOLI, S., BIROLI, G., HONGLER, C. and WYART, M. (2020). Scaling description of generalization with number of parameters in deep learning. *J. Stat. Mech. Theory Exp.* **2** 023401. MR4137317 https://doi.org/10.1088/1742-5468/ab633c

[29] GERACE, F., LOUREIRO, B., KRZAKALA, F., MÉZARD, M. and ZDEBOROVÁ, L. (2020). Generalisation error in learning with random features and the hidden manifold model. International Conference on Machine Learning. PMLR.

[30] 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

[31] GOLDT, S., REEVES, G., MEZARD, M., KRZAKALA, F. and ZDEBOROVÁ, L. (2020). The gaussian equivalence of generative models for learning with two-layer neural networks. Available at arXiv:2006.14709.

[32] GOLUB, G. H., HEATH, M. and WAHBA, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21** 215–223. MR0533250 https://doi.org/10.2307/1268518

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

[34] GUNASEKAR, S., LEE, J., SOUDRY, D. and SREBRO, N. (2018). Characterizing implicit bias in terms of optimization geometry. In 35*th International Conference on Machine Learning*, *ICML* 2018 2932–2955. International Machine Learning Society (IMLS).

[35] GUNASEKAR, S., LEE, J. D., SOUDRY, D. and SREBRO, N. (2018). Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems* 9461–9471.

[36] HASTIE, T., MONTANARI, A., ROSSET, S. and TIBSHIRANI, R. J (2022). Supplement to "Surprises in high-dimensional ridgeless least squares interpolation." https://doi.org/10.1214/21-AOS2133SUPP

[37] HASTIE, T., ROSSET, S., TIBSHIRANI, R. and ZHU, J. (2003/04). The entire regularization path for the support vector machine. *J. Mach. Learn. Res.* **5** 1391–1415. MR2248021

[38] HU, H. and LU, Y. M. (2020). Universality laws for high-dimensional learning with random features. Available at arXiv:2009.07669.

[39] JACOT, A., GABRIEL, F. and HONGLER, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *Adv. Neural Inf. Process. Syst.* **31**.

[40] KE, W. and THRAMPOULIDIS, C. (2020). Benign overfitting in binary classification of gaussian mixtures. arXiv preprint. Available at arXiv:2011.09148.

[41] KNOWLES, A. and YIN, J. (2017). Anisotropic local laws for random matrices. *Probab. Theory Related Fields* **169** 257–352. MR3704770 https://doi.org/10.1007/s00440-016-0730-4

[42] KOBAK, D., LOMOND, J. and SANCHEZ, B. (2020). The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *J. Mach. Learn. Res.* **21** 169. MR4209455

[43] LEDOIT, O. and PÉCHÉ, S. (2011). Eigenvectors of some large sample covariance matrix ensembles. *Probab. Theory Related Fields* **151** 233–264. MR2834718 https://doi.org/10.1007/s00440-010-0298-3

[44] LEE, J., XIAO, L., SCHOENHOLZ, S. S., BAHRI, Y., NOVAK, R., SOHL-DICKSTEIN, J. and PENNINGTON, J. (2020). Wide neural networks of any depth evolve as linear models under gradient descent. *J. Stat. Mech. Theory Exp.* **12** 124002. MR4241355 https://doi.org/10.1088/1742-5468/abc62b

[45] LI, K.-C. (1986). Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. *Ann. Statist.* **14** 1101–1112. MR0856808 https://doi.org/10.1214/aos/1176350052

[46] LI, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *Ann. Statist.* **15** 958–975. MR0902239 https://doi.org/10.1214/aos/1176350486

[47] LIANG, T. and RAKHLIN, A. (2020). Just interpolate: Kernel "ridgeless" regression can generalize. *Ann. Statist.* **48** 1329–1347. MR4124325 https://doi.org/10.1214/19-AOS1849

[48] LIANG, T., RAKHLIN, A. and ZHAI, X. (2020). On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory* 2683–2711.

[49] MEI, S. and MONTANARI, A. (2019). The generalization error of random features regression: Precise asymptotics and double descent curve. *Comm. Pure Appl. Math.* To appear.

[50] 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

[51] MIOLANE, L. and MONTANARI, A. (2021). The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning. *Ann. Statist.* **49** 2313–2335. MR4319252 https://doi.org/10.1214/20-aos2038

[52] MONTANARI, A., RUAN, F., SOHN, Y. and YAN, J. (2019). The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. Available at arXiv:1911.01544.

[53] MUTHUKUMAR, V., NARANG, A., SUBRAMANIAN, V., BELKIN, M., HSU, D. and SAHAI, A. (2021). Classification vs regression in overparameterized regimes: Does the loss function matter? *J. Mach. Learn. Res.* **22** 222. MR4329801

[54] PATIL, P., WEI, Y., RINALDO, A. and TIBSHIRANI, R. (2021). Uniform consistency of cross-validation estimators for high-dimensional ridge regression. In *International Conference on Artificial Intelligence and Statistics* 3178–3186. PMLR.

[55] PENNINGTON, J. and WORAH, P. (2017). Nonlinear random matrix theory for deep learning. *Adv. Neural Inf. Process. Syst.* **30**.

[56] RAKHLIN, A. and ZHAI, X. (2019). Consistency of interpolation with Laplace kernels is a high-dimensional phenomenon. In *Conference on Learning Theory* 2595–2623. PMLR.

[57] RICHARDS, D., MOURTADA, J. and ROSASCO, L. (2020). Asymptotics of ridge (less) regression under general source condition. Available at arXiv:2006.06386.

[58] ROSSET, S., ZHU, J. and HASTIE, T. (2003/04). Boosting as a regularized path to a maximum margin classifier. *J. Mach. Learn. Res.* **5** 941–973. MR2248005

[59] ROTSKOFF, G. M. and VANDEN-EIJNDEN, E. (2018). Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. Available at arXiv:1805.00915.

[60] RUBIO, F. and MESTRE, X. (2011). Spectral convergence for a general class of random matrices. *Statist. Probab. Lett.* **81** 592–602. MR2772917 https://doi.org/10.1016/j.spl.2011.01.004

[61] SERDOBOLSKII, V. I. (2008). *Multiparametric Statistics*. Elsevier, Amsterdam. MR2531357

[62] 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

[63] SPIGLER, S., GEIGER, M., D'ASCOLI, S., SAGUN, L., BIROLI, G. and WYART, M. (2019). A jamming transition from under- to over-parametrization affects generalization in deep learning. *J. Phys. A* **52** 474001. MR4028949 https://doi.org/10.1088/1751-8121/ab4c8b

[64] TAO, T. (2012). *Topics in Random Matrix Theory*. *Graduate Studies in Mathematics* **132**. Amer. Math. Soc., Providence, RI. MR2906465 https://doi.org/10.1090/gsm/132

[65] TIBSHIRANI, R. J. (2015). A general framework for fast stagewise algorithms. *J. Mach. Learn. Res.* **16** 2543–2588. MR3450516

[66] TSIGLER, A. and BARTLETT, P. L. (2020). Benign overfitting in ridge regression. Available at arXiv:2009.14286.

[67] TULINO, A. M. and VERDU, S. (2004). Random matrix theory and wireless communications. *Foundations and Trends in Communications and Information Theory* **1** 1–182.

[68] WEBB, A. R. (1994). Functional approximation by feed-forward networks: A least-squares approach to generalization. *IEEE Trans. Neural Netw.* **5** 363–371.

[69] WU, D. and XU, J. (2020). On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *Adv. Neural Inf. Process. Syst.* **33** 10112–10123.

[70] XU, J., MALEKI, A., RAD, K. R. and HSU, D. (2021). Consistent risk estimation in moderately high-dimensional linear regression. *IEEE Trans. Inf. Theory* **67** 5997–6030. MR4345048 https://doi.org/10.1109/TIT.2021.3095375

[71] ZHANG, C., BENGIO, S., HARDT, M., RECHT, B. and VINYALS, O. (2016). Understanding deep learning requires rethinking generalization. Available at arXiv:1611.03530.

[72] ZHANG, C., BENGIO, S. and SINGER, Y. (2019). Are all layers created equal? Available at arXiv:1902.01996.

[73] ZOU, D., CAO, Y., ZHOU, D. and GU, Q. (2018). Stochastic gradient descent optimizes over-parameterized deep ReLU networks. Available at arXiv:1811.08888.

The Annals of Statistics
2022, Vol. 50, No. 2, 987–1011
https://doi.org/10.1214/21-AOS2134

MOTIF ESTIMATION VIA SUBGRAPH SAMPLING: THE FOURTH-MOMENT PHENOMENON

BY BHASWAR B. BHATTACHARYA[1,a], SAYAN DAS[2,b] AND SUMIT MUKHERJEE[3,c]

[1]*Department of Statistics, University of Pennsylvania,* [a]*bhaswar@wharton.upenn.edu*
[2]*Department of Mathematics, Columbia University,* [b]*sayan.das@columbia.edu*
[3]*Department of Statistics, Columbia University,* [c]*sm3949@columbia.edu*

Network sampling is an indispensable tool for understanding features of large complex networks where it is practically impossible to search over the entire graph. In this paper, we develop a framework for statistical inference for counting network motifs, such as edges, triangles and wedges, in the widely used subgraph sampling model, where each vertex is sampled independently, and the subgraph induced by the sampled vertices is observed. We derive necessary and sufficient conditions for the consistency and the asymptotic normality of the natural Horvitz–Thompson (HT) estimator, which can be used for constructing confidence intervals and hypothesis testing for the motif counts based on the sampled graph. In particular, we show that the asymptotic normality of the HT estimator exhibits an interesting fourth-moment phenomenon, which asserts that the HT estimator (appropriately centered and rescaled) converges in distribution to the standard normal whenever its fourth-moment converges to 3 (the fourth-moment of the standard normal distribution). As a consequence, we derive the exact thresholds for consistency and asymptotic normality of the HT estimator in various natural graph ensembles, such as sparse graphs with bounded degree, Erdős–Rényi random graphs, random regular graphs and dense graphons.

REFERENCES

[1] AIROLDI, E. M., COSTA, T. B. and CHAN, S. H. (2013). Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems* 692–700.

[2] ALIAKBARPOUR, M., SHANKHA BISWAS, A., GOULEAKIS, T., PEEBLES, J., RUBINFELD, R. and YODPINYANEE, A. (2018). Sublinear-time algorithms for counting star subgraphs via edge sampling. *Algorithmica* **80** 668–697. MR3757567 https://doi.org/10.1007/s00453-017-0287-3

[3] APICELLA, C. L., MARLOWE, F. W., FOWLER, J. H. and CHRISTAKIS, N. A. (2012). Social networks and cooperation in hunter-gatherers. *Nature* **481** 497–501. https://doi.org/10.1038/nature10736

[4] BANDIERA, O. and RASUL, I. (2006). Social networks and technology adoption in northern Mozambique. *Econ. J.* **116** 869–902.

[5] BARBOUR, A. D. and CHEN, L. H. Y., eds. (2005). *An Introduction to Stein's Method. Lecture Notes Series. Institute for Mathematical Sciences. National University of Singapore* **4**. World Scientific, Hackensack, NJ. MR2235447 https://doi.org/10.1142/9789812567680

[6] BERTAIL, P., CHAUTRU, E. and CLÉMENÇON, S. (2017). Empirical processes in survey sampling with (conditional) Poisson designs. *Scand. J. Stat.* **44** 97–111. MR3619696 https://doi.org/10.1111/sjos.12243

[7] BHATTACHARYA, A., BISHNU, A., GHOSH, A. and MISHRA, G. (2019). Triangle estimation using tripartite independent set queries. In 30*th International Symposium on Algorithms and Computation* (*ISAAC* 2019). Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern. MR4042092

[8] BHATTACHARYA, B. B, DAS, S. and MUKHERJEE, S. (2022). Supplement to "Motif estimation via subgraph sampling: The fourth-moment phenomenon." https://doi.org/10.1214/21-AOS2134SUPP

MSC2020 subject classifications. 62G05, 62G20, 05C30.
Key words and phrases. Fourth moment phenomenon, Motif counting, network analysis, asymptotic inference, Stein's method, random graphs.

[9] BHATTACHARYA, B. B., DIACONIS, P. and MUKHERJEE, S. (2017). Universal limit theorems in graph coloring problems with connections to extremal combinatorics. *Ann. Appl. Probab.* **27** 337–394. MR3619790 https://doi.org/10.1214/16-AAP1205

[10] BICKEL, P. J., CHEN, A. and LEVINA, E. (2011). The method of moments and degree distributions for network models. *Ann. Statist.* **39** 2280–2301. MR2906868 https://doi.org/10.1214/11-AOS904

[11] BORGS, C., CHAYES, J. T., LOVÁSZ, L., SÓS, V. T. and VESZTERGOMBI, K. (2008). Convergent sequences of dense graphs. I. Subgraph frequencies, metric properties and testing. *Adv. Math.* **219** 1801–1851. MR2455626 https://doi.org/10.1016/j.aim.2008.07.008

[12] BORGS, C., CHAYES, J. T., LOVÁSZ, L., SÓS, V. T. and VESZTERGOMBI, K. (2012). Convergent sequences of dense graphs II. Multiway cuts and statistical physics. *Ann. of Math.* (2) **176** 151–219. MR2925382 https://doi.org/10.4007/annals.2012.176.1.2

[13] CAPOBIANCO, M. (1972). Estimating the connectivity of a graph. In *Graph Theory and Applications* (*Proc. Conf., Western Michigan Univ., Kalamazoo, Mich.*, 1972; *Dedicated to the Memory of J. W. T. Youngs*). *Lecture Notes in Math.* **303** 65–74. MR0332542

[14] CHANDRASEKHAR, A. and LEWIS, R. (2011). Econometrics of sampled networks.

[15] CHATTERJEE, S. and DIACONIS, P. (2013). Estimating and understanding exponential random graph models. *Ann. Statist.* **41** 2428–2461. MR3127871 https://doi.org/10.1214/13-AOS1155

[16] 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

[17] CHEN, L. H. Y. and SHAO, Q.-M. (2004). Normal approximation under local dependence. *Ann. Probab.* **32** 1985–2028. MR2073183 https://doi.org/10.1214/009117904000000450

[18] CRANE, H. (2018). *Probabilistic Foundations of Statistical Network Analysis*. *Monographs on Statistics and Applied Probability* **157**. CRC Press, Boca Raton, FL. MR3791467

[19] DIACONIS, P. and FREEDMAN, D. (1980). Finite exchangeable sequences. *Ann. Probab.* **8** 745–764. MR0577313

[20] EDEN, T., LEVI, A., RON, D. and SESHADHRI, C. (2017). Approximately counting triangles in sublinear time. *SIAM J. Comput.* **46** 1603–1646. MR3709896 https://doi.org/10.1137/15M1054389

[21] FEIGE, U. (2006). On sums of independent random variables with unbounded variance and estimating the average degree in a graph. *SIAM J. Comput.* **35** 964–984. MR2203734 https://doi.org/10.1137/S0097539704447304

[22] FRANK, O. (1977). Estimation of graph totals. *Scand. J. Stat.* **4** 81–89. MR0458659

[23] FRANK, O. (1978). Estimation of the number of connected components in a graph by using a sampled subgraph. *Scand. J. Stat.* **5** 177–188. MR0515656

[24] GAO, C. and MA, Z. (2021). Minimax rates in network analysis: Graphon estimation, community detection and hypothesis testing. *Statist. Sci.* **36** 16–33. MR4194201 https://doi.org/10.1214/19-STS736

[25] GOLDREICH, O. and RON, D. (2008). Approximating average parameters of graphs. *Random Structures Algorithms* **32** 473–493. MR2422391 https://doi.org/10.1002/rsa.20203

[26] GOLDREICH, O. and RON, D. (2011). On testing expansion in bounded-degree graphs. In *Studies in Complexity and Cryptography. Miscellanea on the Interplay Between Randomness and Computation* 68–75. Springer.

[27] GONEN, M., RON, D. and SHAVITT, Y. (2011). Counting stars and other small subgraphs in sublinear-time. *SIAM J. Discrete Math.* **25** 1365–1411. MR2837605 https://doi.org/10.1137/100783066

[28] GOODMAN, L. A. (1961). Snowball sampling. *Ann. Math. Stat.* **32** 148–170. MR0124140 https://doi.org/10.1214/aoms/1177705148

[29] GOVINDAN, R. and TANGMUNARUNKIT, H. (2000). Heuristics for Internet map discovery. In *Proceedings IEEE INFOCOM* 2000. *Conference on Computer Communications*. *Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies* **3** 1371–1380. IEEE, New York.

[30] GRIMMETT, G. (1999). *Percolation*, 2nd ed. *Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **321**. Springer, Berlin. MR1707339 https://doi.org/10.1007/978-3-662-03981-6

[31] GUPTA, P. and BHATTACHARJEE, G. P. (1984). An efficient algorithm for random sampling without replacement. *Int. J. Comput. Math.* **16** 201–209. MR0775085 https://doi.org/10.1080/00207168408803438

[32] HANDCOCK, M. S. and GILE, K. J. (2010). Modeling social networks from sampled data. *Ann. Appl. Stat.* **4** 5–25. MR2758082 https://doi.org/10.1214/08-AOAS221

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

[34] JANSON, S., ŁUCZAK, T. and RUCINSKI, A. (2000). *Random Graphs*. *Wiley-Interscience Series in Discrete Mathematics and Optimization* **45**. Wiley Interscience, New York. MR1782847 https://doi.org/10.1002/9781118032718

[35] KIM, J. H., SUDAKOV, B. and VU, V. (2007). Small subgraphs of random regular graphs. *Discrete Math.* **307** 1961–1967. MR2320201 https://doi.org/10.1016/j.disc.2006.09.032

[36] KLUSOWSKI, J. M. and WU, Y. (2018). Counting motifs with graph sampling. In *Conference on Learning Theory* 1966–2011.

[37] KLUSOWSKI, J. M. and WU, Y. (2020). Estimating the number of connected components in a graph via subgraph sampling. *Bernoulli* **26** 1635–1664. MR4091087 https://doi.org/10.3150/19-BEJ1147

[38] KOLACZYK, E. D. (2009). *Statistical Analysis of Network Data*: *Methods and Models*. *Springer Series in Statistics*. Springer, New York. MR2724362 https://doi.org/10.1007/978-0-387-88146-1

[39] KOLACZYK, E. D. (2017). *Topics at the Frontier of Statistics and Network Analysis*: (*Re*)*Visiting the Foundations*. *SemStat Elements*. Cambridge Univ. Press, Cambridge. MR3702038 https://doi.org/10.1017/9781108290159

[40] KRIVELEVICH, M., SUDAKOV, B., VU, V. H. and WORMALD, N. C. (2001). Random regular graphs of high degree. *Random Structures Algorithms* **18** 346–363. MR1839497 https://doi.org/10.1002/rsa.1013

[41] LESKOVEC, J. and FALOUTSOS, C. (2006). Sampling from large graphs. In *Proceedings of the* 12*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 631–636.

[42] LOVÁSZ, L. (2012). *Large Networks and Graph Limits*. *American Mathematical Society Colloquium Publications* **60**. Amer. Math. Soc., Providence, RI. MR3012035 https://doi.org/10.1090/coll/060

[43] MARINUCCI, D. and PECCATI, G. (2011). *Random Fields on the Sphere*: *Representation*, *Limit Theorems and Cosmological Applications*. *London Mathematical Society Lecture Note Series* **389**. Cambridge Univ. Press, Cambridge. MR2840154 https://doi.org/10.1017/CBO9780511751677

[44] MILO, R., SHEN-ORR, S., ITZKOVITZ, S., KASHTAN, N., CHKLOVSKII, D. and ALON, U. (2002). Network motifs: Simple building blocks of complex networks. *Science* **298** 824–827.

[45] NOURDIN, I. and PECCATI, G. (2009). Stein's method on Wiener chaos. *Probab. Theory Related Fields* **145** 75–118. MR2520122 https://doi.org/10.1007/s00440-008-0162-x

[46] NOURDIN, I. and PECCATI, G. (2012). *Normal Approximations with Malliavin Calculus*: *From Stein's Method to Universality*. *Cambridge Tracts in Mathematics* **192**. Cambridge Univ. Press, Cambridge. MR2962301 https://doi.org/10.1017/CBO9781139084659

[47] NOURDIN, I., PECCATI, G. and REINERT, G. (2010). Invariance principles for homogeneous sums: Universality of Gaussian Wiener chaos. *Ann. Probab.* **38** 1947–1985. MR2722791 https://doi.org/10.1214/10-AOP531

[48] NUALART, D. and PECCATI, G. (2005). Central limit theorems for sequences of multiple stochastic integrals. *Ann. Probab.* **33** 177–193. MR2118863 https://doi.org/10.1214/009117904000000621

[49] OVE, F. (2005). Network sampling and model fitting. In *Structural Analysis in the Social Sciences* 31–56. Cambridge Univ. Press, Cambridge.

[50] PRŽULJ, N., CORNEIL, D. G. and JURISICA, I. (2004). Modeling interactome: Scale-free or geometric? *Bioinformatics* **20** 3508–3515.

[51] RIBEIRO, B. and TOWSLEY, D. (2010). Estimating and sampling graphs with multidimensional random walks. In *Proceedings of the* 10*th ACM SIGCOMM Conference on Internet Measurement* 390–403.

[52] ROUZANKIN, P. S. and VOYTISHEK, A. V. (1999). On the cost of algorithms for random selection. *Monte Carlo Methods Appl.* **5** 39–54. MR1684992 https://doi.org/10.1515/mcma.1999.5.1.39

[53] RUAL, J.-F., VENKATESAN, K., HAO, T., HIROZANE-KISHIKAWA, T., DRICOT, A., LI, N., BERRIZ, G. F., GIBBONS, F. D., DREZE, M. et al. (2005). Towards a proteome-scale map of the human protein–protein interaction network. *Nature* **437** 1173–1178.

[54] STEIN, C. (1972). A bound for the error in the normal approximation to the distribution of a sum 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* 583–602. MR0402873

[55] TANG, M., SUSSMAN, D. L. and PRIEBE, C. E. (2013). Universally consistent vertex classification for latent positions graphs. *Ann. Statist.* **41** 1406–1430. MR3113816 https://doi.org/10.1214/13-AOS1112

[56] TILLÉ, Y. (2006). *Sampling Algorithms*. *Springer Series in Statistics*. Springer, New York. MR2225036

[57] UETZ, P., GIOT, L., CAGNEY, G., MANSFIELD, T. A., JUDSON, R. S., KNIGHT, J. R., LOCKSHON, D., NARAYAN, V., SRINIVASAN, M. et al. (2000). A comprehensive analysis of protein–protein interactions in saccharomyces cerevisiae. *Nature* **403** 623–627.

[58] YANG, J. and LESKOVEC, J. (2015). Defining and evaluating network communities based on ground-truth. *Knowl. Inf. Syst.* **42** 181–213.

[59] ZHANG, Y., KOLACZYK, E. D. and SPENCER, B. D. (2015). Estimating network degree distributions under sampling: An inverse problem, with applications to monitoring social media networks. *Ann. Appl. Stat.* **9** 166–199. MR3341112 https://doi.org/10.1214/14-AOAS800

The Annals of Statistics
2022, Vol. 50, No. 2, 1012–1037
https://doi.org/10.1214/21-AOS2136

MULTIVARIATE RANKS AND QUANTILES USING OPTIMAL TRANSPORT: CONSISTENCY, RATES AND NONPARAMETRIC TESTING

BY PROMIT GHOSAL[1,a] AND BODHISATTVA SEN[2,b]

[1]*Department of Mathematics, Massachusetts Institute of Technology,* [a]*promit@mit.edu*
[2]*Department of Statistics, Columbia University,* [b]*bodhi@stat.columbia.edu*

In this paper, we study multivariate ranks and quantiles, defined using the theory of optimal transport, and build on the work of Chernozhukov et al. (*Ann. Statist.* **45** (2017) 223–256) and Hallin et al. (*Ann. Statist.* **49** (2021) 1139–1165). We study the characterization, computation and properties of the multivariate rank and quantile functions and their empirical counterparts. We derive the uniform consistency of these empirical estimates to their population versions, under certain assumptions. In fact, we prove a Glivenko–Cantelli type theorem that shows the asymptotic stability of the empirical rank map in any direction. Under mild structural assumptions, we provide global and local rates of convergence of the empirical quantile and rank maps. We also provide a sub-Gaussian tail bound for the global L_2-loss of the empirical quantile function. Further, we propose tuning parameter-free multivariate nonparametric tests—a two-sample test and a test for mutual independence—based on our notion of multivariate quantiles/ranks. Asymptotic consistency of these tests are shown and the rates of convergence of the associated test statistics are derived, both under the null and alternative hypotheses.

REFERENCES

[1] AURENHAMMER, F. (1987). Power diagrams: Properties, algorithms and applications. *SIAM J. Comput*. **16** 78–96. MR0873251 https://doi.org/10.1137/0216006

[2] AURENHAMMER, F., HOFFMANN, F. and ARONOV, B. (1998). Minkowski-type theorems and least-squares clustering. *Algorithmica* **20** 61–76. MR1483422 https://doi.org/10.1007/PL00009187

[3] AURENHAMMER, F., KLEIN, R. and LEE, D.-T. (2013). *Voronoi Diagrams and Delaunay Triangulations*. World Scientific Co. Pte. Ltd., Hackensack, NJ. MR3186045 https://doi.org/10.1142/8685

[4] BARINGHAUS, L. and FRANZ, C. (2004). On a new multivariate two-sample test. *J. Multivariate Anal*. **88** 190–206. MR2021870 https://doi.org/10.1016/S0047-259X(03)00079-4

[5] BERRETT, T. B. and SAMWORTH, R. J. (2019). Nonparametric independence testing via mutual information. *Biometrika* **106** 547–566. MR3992389 https://doi.org/10.1093/biomet/asz024

[6] BHATTACHARYA, B. B. (2020). Asymptotic distribution and detection thresholds for two-sample tests based on geometric graphs. *Ann. Statist*. **48** 2879–2903. MR4152627 https://doi.org/10.1214/19-AOS1913

[7] BICKEL, P. J. (1968). A distribution free version of the Smirnov two sample test in the p-variate case. *Ann. Math. Stat*. **40** 1–23. MR0256519 https://doi.org/10.1214/aoms/1177697800

[8] BLOMQVIST, N. (1950). On a measure of dependence between two random variables. *Ann. Math. Stat*. **21** 593–600. MR0039190 https://doi.org/10.1214/aoms/1177729754

[9] BLUM, J. R., KIEFER, J. and ROSENBLATT, M. (1961). Distribution free tests of independence based on the sample distribution function. *Ann. Math. Stat*. **32** 485–498. MR0125690 https://doi.org/10.1214/aoms/1177705055

[10] BOBKOV, S. and LEDOUX, M. (2019). One-dimensional empirical measures, order statistics, and Kantorovich transport distances. *Mem. Amer. Math. Soc*. **261** v+126. MR4028181 https://doi.org/10.1090/memo/1259

[11] BOECKEL, M., SPOKOINY, V. and SUVORIKOVA, A. (2018). Multivariate Brenier cumulative distribution functions and their application to non-parametric testing. Preprint. Available at arXiv:1809.04090.

MSC2020 subject classifications. Primary 62G30, 62G20; secondary 60F15, 35J96.
Key words and phrases. Brenier–McCann's theorem, convergence of subdifferentials of convex functions, Glivenko–Cantelli type theorem, Legendre–Fenchel dual, local uniform rate of convergence, semidiscrete optimal transport, testing mutual independence, two-sample goodness-of-fit testing.

[12] 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

[13] CAFFARELLI, L. A. (1992). The regularity of mappings with a convex potential. *J. Amer. Math. Soc.* **5** 99–104. MR1124980 https://doi.org/10.2307/2152752

[14] CAFFARELLI, L. A. (1992). Boundary regularity of maps with convex potentials. *Comm. Pure Appl. Math.* **45** 1141–1151. MR1177479 https://doi.org/10.1002/cpa.3160450905

[15] CAFFARELLI, L. A. (1996). Boundary regularity of maps with convex potentials. II. *Ann. of Math.* (2) **144** 453–496. MR1426885 https://doi.org/10.2307/2118564

[16] CAFFARELLI, L. A., KOCHENGIN, S. A. and OLIKER, V. I. (1999). On the numerical solution of the problem of reflector design with given far-field scattering data. In *Monge Ampère Equation: Applications to Geometry and Optimization* (*Deerfield Beach, FL*, 1997). *Contemp. Math.* **226** 13–32. Amer. Math. Soc., Providence, RI. MR1660740 https://doi.org/10.1090/conm/226/03233

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

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

[19] CHIZAT, L., ROUSSILLON, P., LÉGER, F., VIALARD, F.-X. and PEYRÉ, G. (2020). Faster Wasserstein distance estimation with the Sinkhorn divergence. *Adv. Neural Inf. Process. Syst.* **33**.

[20] CORDERO-ERAUSQUIN, D. and FIGALLI, A. (2019). Regularity of monotone transport maps between unbounded domains. *Discrete Contin. Dyn. Syst.* **39** 7101–7112. MR4026183 https://doi.org/10.3934/dcds.2019297

[21] DE PHILIPPIS, G. and FIGALLI, A. (2014). The Monge-Ampère equation and its link to optimal transportation. *Bull. Amer. Math. Soc.* (*N.S.*) **51** 527–580. MR3237759 https://doi.org/10.1090/S0273-0979-2014-01459-4

[22] DE VALK, C. and SEGERS, J. (2018). Tails of optimal transport plans for regularly varying probability measures. Preprint. Available at arXiv:1811.12061.

[23] DEB, N., BHATTACHARYA, B. B. and SEN, B. (2021). Efficiency lower bounds for distribution-free Hotelling-type two-sample tests based on optimal transport. Preprint. Available at arXiv:2104.01986.

[24] DEB, N. and SEN, B. (2019). Multivariate rank-based distribution-free nonparametric testing using measure transportation. Preprint. Available at arXiv:1909.08733.

[25] DEL BARRIO, E., GINÉ, E. and UTZET, F. (2005). Asymptotics for L_2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances. *Bernoulli* **11** 131–189. MR2121458 https://doi.org/10.3150/bj/1110228245

[26] DEL BARRIO, E., GONZÁLEZ-SANZ, A. and HALLIN, M. (2020). A note on the regularity of optimal-transport-based center-outward distribution and quantile functions. *J. Multivariate Anal.* **180** 104671, 13. MR4147635 https://doi.org/10.1016/j.jmva.2020.104671

[27] DEL BARRIO, E., GORDALIZA, P., LESCORNEL, H. and LOUBES, J.-M. (2019). Central limit theorem and bootstrap procedure for Wasserstein's variations with an application to structural relationships between distributions. *J. Multivariate Anal.* **169** 341–362. MR3875604 https://doi.org/10.1016/j.jmva.2018.09.014

[28] DEL BARRIO, E. and LOUBES, J.-M. (2019). Central limit theorems for empirical transportation cost in general dimension. *Ann. Probab.* **47** 926–951. MR3916938 https://doi.org/10.1214/18-AOP1275

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

[30] FIGALLI, A. (2018). On the continuity of center-outward distribution and quantile functions. *Nonlinear Anal.* **177** 413–421. MR3886582 https://doi.org/10.1016/j.na.2018.05.008

[31] 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

[32] FRIEDMAN, J. H. and RAFSKY, L. C. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *Ann. Statist.* **7** 697–717. MR0532236

[33] GHOSAL, P. and SEN, B. (2022). Supplement to "Multivariate ranks and quantiles using optimal transport: Consistency, rates and nonparametric testing." https://doi.org/10.1214/21-AOS2136SUPP

[34] GIGLI, N. (2011). On Hölder continuity-in-time of the optimal transport map towards measures along a curve. *Proc. Edinb. Math. Soc.* (2) **54** 401–409. MR2794662 https://doi.org/10.1017/S001309150800117X

[35] GRETTON, A., HERBRICH, R., SMOLA, A., BOUSQUET, O. and SCHÖLKOPF, B. (2005). Kernel methods for measuring independence. *J. Mach. Learn. Res.* **6** 2075–2129. MR2249882

[36] GU, X., LUO, F., SUN, J. and YAU, S.-T. (2016). Variational principles for Minkowski type problems, discrete optimal transport, and discrete Monge-Ampere equations. *Asian J. Math.* **20** 383–398. MR3480024 https://doi.org/10.4310/AJM.2016.v20.n2.a7

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

[38] HALLIN, M., MORDANT, G. and SEGERS, J. (2021). Multivariate goodness-of-fit tests based on Wasserstein distance. *Electron. J. Stat.* **15** 1328–1371. MR4255302 https://doi.org/10.1214/21-ejs1816

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

[40] HALLIN, M. and WERKER, B. J. M. (2003). Semi-parametric efficiency, distribution-freeness and invariance. *Bernoulli* **9** 137–165. MR1963675 https://doi.org/10.3150/bj/1068129013

[41] HOEFFDING, W. (1952). The large-sample power of tests based on permutations of observations. *Ann. Math. Stat.* **23** 169–192. MR0057521 https://doi.org/10.1214/aoms/1177729436

[42] HOLLANDER, M. and WOLFE, D. A. (1999). *Nonparametric Statistical Methods*, 2nd ed. *Wiley Series in Probability and Statistics*: *Texts and References Section*. Wiley, New York. MR1666064

[43] 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

[44] KIM, I., BALAKRISHNAN, S. and WASSERMAN, L. (2020). Minimax optimality of permutation tests. Preprint. Available at arXiv:2003.13208.

[45] KITAGAWA, J. (2014). An iterative scheme for solving the optimal transportation problem. *Calc. Var. Partial Differential Equations* **51** 243–263. MR3247388 https://doi.org/10.1007/s00526-013-0673-x

[46] KITAGAWA, J. and MCCANN, R. (2019). Free discontinuities in optimal transport. *Arch. Ration. Mech. Anal.* **232** 1505–1541. MR3928755 https://doi.org/10.1007/s00205-018-01348-3

[47] KITAGAWA, J., MÉRIGOT, Q. and THIBERT, B. (2019). Convergence of a Newton algorithm for semi-discrete optimal transport. *J. Eur. Math. Soc.* (*JEMS*) **21** 2603–2651. MR3985609 https://doi.org/10.4171/JEMS/889

[48] KLATT, M., TAMELING, C. and MUNK, A. (2020). Empirical regularized optimal transport: Statistical theory and applications. *SIAM J. Math. Data Sci.* **2** 419–443. MR4105566 https://doi.org/10.1137/19M1278788

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

[50] LEHMANN, E. L. (1975). *Nonparametrics*: *Statistical Methods Based on Ranks*. *Holden-Day Series in Probability and Statistics*. Holden-Day, San Francisco, CA; McGraw-Hill, New York–Düsseldorf. With the special assistance of H. J. M. d'Abrera. MR0395032

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

[52] LÉVY, B. (2015). A numerical algorithm for L_2 semi-discrete optimal transport in 3D. *ESAIM Math. Model. Numer. Anal.* **49** 1693–1715. MR3423272 https://doi.org/10.1051/m2an/2015055

[53] LI, W. and NOCHETTO, R. H. (2021). Quantitative stability and error estimates for optimal transport plans. *IMA J. Numer. Anal.* **41** 1941–1965. MR4286252 https://doi.org/10.1093/imanum/draa045

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

[55] LYONS, R. (2013). Distance covariance in metric spaces. *Ann. Probab.* **41** 3284–3305. MR3127883 https://doi.org/10.1214/12-AOP803

[56] MCCANN, R. J. (1995). Existence and uniqueness of monotone measure-preserving maps. *Duke Math. J.* **80** 309–323. MR1369395 https://doi.org/10.1215/S0012-7094-95-08013-2

[57] MÉRIGOT, Q. (2011). A multiscale approach to optimal transport. In *Computer Graphics Forum* **30** 1583–1592. Wiley, New York.

[58] MÉRIGOT, Q. and THIBERT, B. (2021). Optimal transport: Discretization and algorithms. In *Geometric Partial Differential Equations*. *Part II*. *Handb. Numer. Anal.* **22** 133–212. Elsevier/North-Holland, Amsterdam. MR4254135 https://doi.org/10.1016/bs.hna.2020.10.001

[59] MÓRI, T. F. and SZÉKELY, G. J. (2019). Four simple axioms of dependence measures. *Metrika* **82** 1–16. MR3897521 https://doi.org/10.1007/s00184-018-0670-3

[60] OJA, H. (1983). Descriptive statistics for multivariate distributions. *Statist. Probab. Lett.* **1** 327–332. MR0721446 https://doi.org/10.1016/0167-7152(83)90054-8

[61] OLIKER, V. I. and PRUSSNER, L. D. (1988). On the numerical solution of the equation $(\partial^2 z/\partial x^2)(\partial^2 z/\partial y^2) - ((\partial^2 z/\partial x \partial y))^2 = f$ and its discretizations. I. *Numer. Math.* **54** 271–293. MR0971703 https://doi.org/10.1007/BF01396762

[62] Panaretos, V. M. and Zemel, Y. (2019). Statistical aspects of Wasserstein distances. *Annu. Rev. Stat. Appl.* **6** 405–431. MR3939527 https://doi.org/10.1146/annurev-statistics-030718-104938

[63] Peyré, G., Cuturi, M. et al. (2019). Computational optimal transport: With applications to data science. *Found. Trends Mach. Learn.* **11** 355–607.

[64] Pfister, N., Bühlmann, P., Schölkopf, B. and Peters, J. (2018). Kernel-based tests for joint independence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 5–31. MR3744710 https://doi.org/10.1111/rssb.12235

[65] Ramdas, A., García Trillos, N. and Cuturi, M. (2017). On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy* **19** Paper No. 47, 15. MR3608466 https://doi.org/10.3390/e19020047

[66] Rigollet, P. and Weed, J. (2018). Entropic optimal transport is maximum-likelihood deconvolution. *C. R. Math. Acad. Sci. Paris* **356** 1228–1235. MR3907589 https://doi.org/10.1016/j.crma.2018.10.010

[67] Rigollet, P. and Weed, J. (2019). Uncoupled isotonic regression via minimum Wasserstein deconvolution. *Inf. Inference* **8** 691–717. MR4045481 https://doi.org/10.1093/imaiai/iaz006

[68] Rosenbaum, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 515–530. MR2168202 https://doi.org/10.1111/j.1467-9868.2005.00513.x

[69] Rousseeuw, P. J. and Ruts, I. (1999). The depth function of a population distribution. *Metrika* **49** 213–244. MR1731769

[70] Schilling, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *J. Amer. Statist. Assoc.* **81** 799–806. MR0860514

[71] Sejdinovic, D., Sriperumbudur, B., Gretton, A. and Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Statist.* **41** 2263–2291. MR3127866 https://doi.org/10.1214/13-AOS1140

[72] Serfling, R. (2010). Equivariance and invariance properties of multivariate quantile and related functions, and the role of standardisation. *J. Nonparametr. Stat.* **22** 915–936. MR2738875 https://doi.org/10.1080/10485250903431710

[73] Shi, H., Drton, M. and Han, F. (2019). Distribution-free consistent independence tests via Hallin's multivariate rank. Preprint. Available at arXiv:1909.10024.

[74] Shi, H., Hallin, M., Drton, M. and Han, F. (2020). Rate-optimality of consistent distribution-free tests of independence based on center-outward ranks and signs. Preprint. Available at arXiv:2007.02186.

[75] Székely, G. J. and Rizzo, M. L. (2009). Brownian distance covariance. *Ann. Appl. Stat.* **3** 1236–1265. MR2752127 https://doi.org/10.1214/09-AOAS312

[76] Székely, G. J. and Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *J. Statist. Plann. Inference* **143** 1249–1272. MR3055745 https://doi.org/10.1016/j.jspi.2013.03.018

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

[78] van der Vaart, A. W. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

[79] 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

[80] 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

[81] Wang, X.-J. (2017). Monge-Ampère equation and optimal transportation. In *Proceedings of the Sixth International Congress of Chinese Mathematicians. Vol. i. Adv. Lect. Math.* (*ALM*) **36** 153–172. Int. Press, Somerville, MA. MR3702074

[82] 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

[83] Weihs, L., Drton, M. and Meinshausen, N. (2018). Symmetric rank covariances: A generalized framework for nonparametric measures of dependence. *Biometrika* **105** 547–562. MR3842884 https://doi.org/10.1093/biomet/asy021

[84] Weiss, L. (1960). Two-sample tests for multivariate distributions. *Ann. Math. Stat.* **31** 159–164. MR0119305 https://doi.org/10.1214/aoms/1177705995

[85] Xu, P. (2019). testOTM: Multivariate Ranks and Quantiles using Optimal Transportation. R package version 1.00.0.

[86] ZEMEL, Y. and PANARETOS, V. M. (2019). Fréchet means and Procrustes analysis in Wasserstein space. *Bernoulli* **25** 932–976. MR3920362 https://doi.org/10.3150/17-bej1009
[87] ZUO, Y. (2003). Projection-based depth functions and associated medians. *Ann*. *Statist*. **31** 1460–1490. MR2012822 https://doi.org/10.1214/aos/1065705115

The Annals of Statistics
2022, Vol. 50, No. 2, 1038–1065
https://doi.org/10.1214/21-AOS2138

NECESSARY AND SUFFICIENT CONDITIONS FOR ASYMPTOTICALLY OPTIMAL LINEAR PREDICTION OF RANDOM FIELDS ON COMPACT METRIC SPACES

BY KRISTIN KIRCHNER[1,a] AND DAVID BOLIN[2,b]

[1] *Delft Institute of Applied Mathematics, Delft University of Technology,* [a] *k.kirchner@tudelft.nl*
[2] *CEMSE Division, King Abdullah University of Science and Technology,* [b] *david.bolin@kaust.edu.sa*

Optimal linear prediction (aka. kriging) of a random field $\{Z(x)\}_{x\in\mathcal{X}}$ indexed by a compact metric space $(\mathcal{X}, d_{\mathcal{X}})$ can be obtained if the mean value function $m\colon \mathcal{X} \to \mathbb{R}$ and the covariance function $\varrho\colon \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ of Z are known. We consider the problem of predicting the value of $Z(x^*)$ at some location $x^* \in \mathcal{X}$ based on observations at locations $\{x_j\}_{j=1}^n$, which accumulate at x^* as $n \to \infty$ (or, more generally, predicting $\varphi(Z)$ based on $\{\varphi_j(Z)\}_{j=1}^n$ for linear functionals $\varphi, \varphi_1, \dots, \varphi_n$). Our main result characterizes the asymptotic performance of linear predictors (as n increases) based on an incorrect second-order structure $(\widetilde{m}, \widetilde{\varrho})$, without any restrictive assumptions on $\varrho, \widetilde{\varrho}$ such as stationarity. We, for the first time, provide necessary and sufficient conditions on $(\widetilde{m}, \widetilde{\varrho})$ for asymptotic optimality of the corresponding linear predictor holding uniformly with respect to φ. These general results are illustrated by weakly stationary random fields on $\mathcal{X} \subset \mathbb{R}^d$ with Matérn or periodic covariance functions, and on the sphere $\mathcal{X} = \mathbb{S}^2$ for the case of two isotropic covariance functions.

REFERENCES

[1] ANDERES, E., MØLLER, J. and RASMUSSEN, J. G. (2020). Isotropic covariance functions on graphs and their edges. *Ann. Statist.* **48** 2478–2503. MR4134803 https://doi.org/10.1214/19-AOS1896

[2] ARAFAT, A., PORCU, E., BEVILACQUA, M. and MATEU, J. (2018). Equivalence and orthogonality of Gaussian measures on spheres. *J. Multivariate Anal.* **167** 306–318. MR3830648 https://doi.org/10.1016/j.jmva.2018.05.005

[3] BERLINET, A. and THOMAS-AGNAN, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic, Boston, MA. MR2239907 https://doi.org/10.1007/978-1-4419-9096-9

[4] BEVILACQUA, M., FAOUZI, T., FURRER, R. and PORCU, E. (2019). Estimation and prediction using generalized Wendland covariance functions under fixed domain asymptotics. *Ann. Statist.* **47** 828–856. MR3909952 https://doi.org/10.1214/17-AOS1652

[5] CLEVELAND, W. S. (1971). Projection with the wrong inner product and its application to regression with correlated errors and linear filtering of time series. *Ann. Math. Stat.* **42** 616–624. MR0286236 https://doi.org/10.1214/aoms/1177693411

[6] GNEITING, T. (2013). Strictly and non-strictly positive definite functions on spheres. *Bernoulli* **19** 1327–1349. MR3102554 https://doi.org/10.3150/12-BEJSP06

[7] GUINNESS, J. and FUENTES, M. (2016). Isotropic covariance functions on spheres: Some properties and modeling considerations. *J. Multivariate Anal.* **143** 143–152. MR3431424 https://doi.org/10.1016/j.jmva.2015.08.018

[8] JANSON, S. (1997). *Gaussian Hilbert Spaces. Cambridge Tracts in Mathematics* **129**. Cambridge Univ. Press, Cambridge. MR1474726 https://doi.org/10.1017/CBO9780511526169

[9] KIRCHNER, K. and BOLIN, D. (2022). Supplement to "Necessary and sufficient conditions for asymptotically optimal linear prediction of random fields on compact metric spaces." https://doi.org/10.1214/21-AOS2138SUPP

[10] KRASNITSKII, S. M. (1990). Conditions for the equivalence of distributions of homogeneous Gaussian fields. *Theory Probab. Appl.* **34** 720–725.

MSC2020 subject classifications. Primary 62M20; secondary 60G12, 60G25, 60G60.
Key words and phrases. Kriging, approximation in Hilbert spaces, spatial statistics.

[11] KUKUSH, A. (2019). *Gaussian Measures in Hilbert Space*: *Construction and Properties*. *Mathematics and Statistics*. ISTE Ltd, London.

[12] LINDGREN, F., RUE, H. and LINDSTRÖM, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 423–498. MR2853727 https://doi.org/10.1111/j.1467-9868.2011.00777.x

[13] MARINUCCI, D. and PECCATI, G. (2011). *Random Fields on the Sphere*. *London Mathematical Society Lecture Note Series* **389**. Cambridge Univ. Press, Cambridge. MR2840154 https://doi.org/10.1017/CBO9780511751677

[14] MATÉRN, B. (1960). *Spatial Variation*: *Stochastic Models and Their Application to Some Problems in Forest Surveys and Other Sampling Investigations*. Meddelanden Från Statens Skogsforskningsinstitut, Band 49, Nr. 5, Stockholm. MR0169346

[15] MERCER, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philos. Trans. R. Soc. A* **209** 415–446.

[16] PARZEN, E. (1959). Statistical inference on time series by Hilbert space methods, I. Technical Report No. 23, Department of Statistics, Stanford Univ.

[17] SCHOENBERG, I. J. (1942). Positive definite functions on spheres. *Duke Math. J.* **9** 96–108. MR0005922

[18] SHUBIN, M. A. (2001). *Pseudodifferential Operators and Spectral Theory*, 2nd ed. Springer, Berlin. MR1852334 https://doi.org/10.1007/978-3-642-56579-3

[19] SKOROKHOD, A. V. and YADRENKO, M. I. (1973). On absolute continuity of measures corresponding to homogeneous Gaussian fields. *Theory Probab. Appl.* **18** 27–40.

[20] STEIN, M. L. (1988). Asymptotically efficient prediction of a random field with a misspecified covariance function. *Ann. Statist.* **16** 55–63. MR0924856 https://doi.org/10.1214/aos/1176350690

[21] STEIN, M. L. (1990). Uniform asymptotic optimality of linear predictions of a random field using an incorrect second-order structure. *Ann. Statist.* **18** 850–872. MR1056340 https://doi.org/10.1214/aos/1176347629

[22] STEIN, M. L. (1993). A simple condition for asymptotic optimality of linear predictions of random fields. *Statist. Probab. Lett.* **17** 399–404. MR1237787 https://doi.org/10.1016/0167-7152(93)90261-G

[23] STEIN, M. L. (1997). Efficiency of linear predictors for periodic processes using an incorrect covariance function. *J. Statist. Plann. Inference* **58** 321–331. MR1450019 https://doi.org/10.1016/S0378-3758(96)00088-2

[24] STEIN, M. L. (1999). Predicting random fields with increasing dense observations. *Ann. Appl. Probab.* **9** 242–273. MR1682572 https://doi.org/10.1214/aoap/1029962604

[25] STEIN, M. L. (1999). *Interpolation of Spatial Data*: *Some Theory for Kriging*. *Springer Series in Statistics*. Springer, New York. MR1697409 https://doi.org/10.1007/978-1-4612-1494-6

[26] 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

[27] WHITTLE, P. (1963). Stochastic processes in several dimensions. *Bull. Int. Stat. Inst.* **40** 974–994. MR0173287

[28] ZHANG, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *J. Amer. Statist. Assoc.* **99** 250–261. MR2054303 https://doi.org/10.1198/016214504000000241

The Annals of Statistics
2022, Vol. 50, No. 2, 1066–1094
https://doi.org/10.1214/21-AOS2140

ITERATIVE ALGORITHM FOR DISCRETE STRUCTURE RECOVERY

BY CHAO GAO[1,a] AND ANDERSON Y. ZHANG[2,b]

[1]*Department of Statistics, University of Chicago,* [a]*chaogao@uchicago.edu*
[2]*Department of Statistics, University of Pennsylvania,* [b]*ayz@wharton.upenn.edu*

We propose a general modeling and algorithmic framework for discrete structure recovery that can be applied to a wide range of problems. Under this framework, we are able to study the recovery of clustering labels, ranks of players, signs of regression coefficients, cyclic shifts and even group elements from a unified perspective. A simple iterative algorithm is proposed for discrete structure recovery, which generalizes methods including Lloyd's algorithm and the power method. A linear convergence result for the proposed algorithm is established in this paper under appropriate abstract conditions on stochastic errors and initialization. We illustrate our general theory by applying it on several representative problems: (1) clustering in Gaussian mixture model, (2) approximate ranking, (3) sign recovery in compressed sensing, (4) multireference alignment and (5) group synchronization, and show that minimax rate is achieved in each case.

REFERENCES

[1] ABBE, E., BANDEIRA, A. S. and HALL, G. (2016). Exact recovery in the stochastic block model. *IEEE Trans. Inf. Theory* **62** 471–487. MR3447993 https://doi.org/10.1109/TIT.2015.2490670

[2] ABBE, E., BENDORY, T., LEEB, W., PEREIRA, J. M., SHARON, N. and SINGER, A. (2019). Multireference alignment is easier with an aperiodic translation distribution. *IEEE Trans. Inf. Theory* **65** 3565–3584. MR3959006 https://doi.org/10.1109/TIT.2018.2889674

[3] ABBE, E., FAN, J., WANG, K. and ZHONG, Y. (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *Ann. Statist.* **48** 1452–1474. MR4124330 https://doi.org/10.1214/19-AOS1854

[4] ABBE, E., MASSOULIÉ, L., MONTANARI, A., SLY, A. and SRIVASTAVA, N. (2018). Group synchronization on grids. *Math. Stat. Learn.* **1** 227–256. MR4059722 https://doi.org/10.4171/msl/6

[5] AERON, S., SALIGRAMA, V. and ZHAO, M. (2010). Information theoretic bounds for compressed sensing. *IEEE Trans. Inf. Theory* **56** 5111–5130. MR2808668 https://doi.org/10.1109/TIT.2010.2059891

[6] AGUERREBERE, C., DELBRACIO, M., BARTESAGHI, A. and SAPIRO, G. (2016). Fundamental limits in multi-image alignment. *IEEE Trans. Signal Process.* **64** 5707–5722. MR3548763 https://doi.org/10.1109/TSP.2016.2600517

[7] ALOISE, D., DESHPANDE, A., HANSEN, P. and POPAT, P. (2009). NP-hardness of Euclidean sum-of-squares clustering. *Mach. Learn.* **75** 245–248.

[8] ARTHUR, D. and VASSILVITSKII, S. (2007). k-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms* 1027–1035. ACM, New York. MR2485254

[9] AWASTHI, P. and SHEFFET, O. (2012). Improved spectral-norm bounds for clustering. In *Approximation, Randomization, and Combinatorial Optimization. Lecture Notes in Computer Science* **7408** 37–49. Springer, Heidelberg. MR3003539 https://doi.org/10.1007/978-3-642-32512-0_4

[10] BAGARIA, V., DING, J., TSE, D., WU, Y. and XU, J. (2020). Hidden Hamiltonian cycle recovery via linear programming. *Oper. Res.* **68** 53–70. MR4059492 https://doi.org/10.1287/opre.2019.1886

[11] BALAKRISHNAN, S., WAINWRIGHT, M. J. and YU, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Ann. Statist.* **45** 77–120. MR3611487 https://doi.org/10.1214/16-AOS1435

[12] BANDEIRA, A. S., BOUMAL, N. and SINGER, A. (2017). Tightness of the maximum likelihood semidefinite relaxation for angular synchronization. *Math. Program.* **163** 145–167. MR3632977 https://doi.org/10.1007/s10107-016-1059-6

MSC2020 subject classifications. 62F07.
Key words and phrases. k-means clustering, approximate ranking, high-dimensional statistics, multireference alignment, group synchronization.

[13] BANDEIRA, A. S., CHARIKAR, M., SINGER, A. and ZHU, A. (2014). Multireference alignment using semidefinite programming. In *ITCS'14—Proceedings of the* 2014 *Conference on Innovations in Theoretical Computer Science* 459–470. ACM, New York. MR3359498

[14] BANDEIRA, A. S., NILES-WEED, J. and RIGOLLET, P. (2019). Optimal rates of estimation for multi-reference alignment. *Math. Stat. Learn.* **2** 25–75. MR4073147 https://doi.org/10.4171/msl/11

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

[16] BENDORY, T., BOUMAL, N., MA, C., ZHAO, Z. and SINGER, A. (2018). Bispectrum inversion with application to multireference alignment. *IEEE Trans. Signal Process.* **66** 1037–1050. MR3771661 https://doi.org/10.1109/TSP.2017.2775591

[17] BLUMENSATH, T. and DAVIES, M. E. (2009). Iterative hard thresholding for compressed sensing. *Appl. Comput. Harmon. Anal.* **27** 265–274. MR2559726 https://doi.org/10.1016/j.acha.2009.04.002

[18] BOGDAN, M., VAN DEN BERG, E., SABATTI, C., SU, W. and CANDÈS, E. J. (2015). SLOPE—Adaptive variable selection via convex optimization. *Ann. Appl. Stat.* **9** 1103–1140. MR3418717 https://doi.org/10.1214/15-AOAS842

[19] 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

[20] BRAVERMAN, M. and MOSSEL, E. (2008). Noisy sorting without resampling. In *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms* 268–276. ACM, New York. MR2485312

[21] BRODER, A. Z., FRIEZE, A. M. and SHAMIR, E. (1994). Finding hidden Hamiltonian cycles. *Random Structures Algorithms* **5** 395–410. MR1277610 https://doi.org/10.1002/rsa.3240050303

[22] BUTUCEA, C., NDAOUD, M., STEPANOVA, N. A. and TSYBAKOV, A. B. (2018). Variable selection with Hamming loss. *Ann. Statist.* **46** 1837–1875. MR3845003 https://doi.org/10.1214/17-AOS1572

[23] CHEN, Y. and CANDÈS, E. J. (2018). The projected power method: An efficient algorithm for joint alignment from pairwise differences. *Comm. Pure Appl. Math.* **71** 1648–1714. MR3847751 https://doi.org/10.1002/cpa.21760

[24] CHEN, Y., GUIBAS, L. J. and HUANG, Q.-X. (2014). Near-optimal joint object matching via convex relaxation. Preprint. Available at arXiv:1402.1473.

[25] COLLIER, O. and DALALYAN, A. S. (2016). Minimax rates in permutation estimation for feature matching. *J. Mach. Learn. Res.* **17** Paper No. 6, 31 pp. MR3482926

[26] CONTE, D., FOGGIA, P., SANSONE, C. and VENTO, M. (2004). Thirty years of graph matching in pattern recognition. *Int. J. Pattern Recognit. Artif. Intell.* **18** 265–298.

[27] DASGUPTA, S. (2008). *The Hardness of k-Means Clustering*. Department of Computer Science and Engineering, Univ. California.

[28] DASKALAKIS, C., TZAMOS, C. and ZAMPETAKIS, M. (2016). Ten steps of EM suffice for mixtures of two Gaussians. Preprint. Available at arXiv:1609.00368.

[29] DAWID, A. P. and SKENE, A. M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **28** 20–28.

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

[31] DERUMIGNY, A. (2018). Improved bounds for square-root Lasso and square-root slope. *Electron. J. Stat.* **12** 741–766. MR3769194 https://doi.org/10.1214/18-EJS1410

[32] DING, J., MA, Z., WU, Y. and XU, J. (2021). Efficient random graph matching via degree profiles. *Probab. Theory Related Fields* **179** 29–115. MR4221654 https://doi.org/10.1007/s00440-020-00997-4

[33] DWIVEDI, R., HO, N., KHAMARU, K., WAINWRIGHT, M. J., JORDAN, M. I. and YU, B. (2020). Singularity, misspecification and the convergence rate of EM. *Ann. Statist.* **48** 3161–3182. MR4185804 https://doi.org/10.1214/19-AOS1924

[34] FEI, Y. and CHEN, Y. (2019). Exponential error rates of SDP for block models: Beyond Grothendieck's inequality. *IEEE Trans. Inf. Theory* **65** 551–571. MR3901009 https://doi.org/10.1109/TIT.2018.2839677

[35] FEI, Y. and CHEN, Y. (2020). Achieving the Bayes error rate in synchronization and block models by SDP, robustly. *IEEE Trans. Inf. Theory* **66** 3929–3953. MR4115142 https://doi.org/10.1109/TIT.2020.2966438

[36] FLETCHER, A. K., RANGAN, S. and GOYAL, V. K. (2009). Necessary and sufficient conditions for sparsity pattern recovery. *IEEE Trans. Inf. Theory* **55** 5758–5772. MR2597192 https://doi.org/10.1109/TIT.2009.2032726

[37] FOUCART, S. (2011). Hard thresholding pursuit: An algorithm for compressive sensing. *SIAM J. Numer. Anal.* **49** 2543–2563. MR2873246 https://doi.org/10.1137/100806278

[38] FRANK, J. (2006). *Three-Dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in Their Native State*. Oxford Univ. Press, London.

[39] FRIEDRICH, F., KEMPE, A., LIEBSCHER, V. and WINKLER, G. (2008). Complexity penalized M-estimation: Fast computation. *J. Comput. Graph. Statist.* **17** 201–224. MR2424802 https://doi.org/10.1198/106186008X285591
[40] GAO, C. (2017). Phase transitions in approximate ranking. Preprint. Available at arXiv:1711.11189.
[41] GAO, C., LU, Y. and ZHOU, D. (2016). Exact exponent in optimal rates for crowdsourcing. In *International Conference on Machine Learning* 603–611.
[42] GAO, C. and MA, Z. (2021). Minimax rates in network analysis: Graphon estimation, community detection and hypothesis testing. *Statist. Sci.* **36** 16–33. MR4194201 https://doi.org/10.1214/19-STS736
[43] GAO, C., MA, Z., ZHANG, A. Y. and ZHOU, H. H. (2017). Achieving optimal misclassification proportion in stochastic block models. *J. Mach. Learn. Res.* **18** Paper No. 60, 45 pp. MR3687603
[44] 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
[45] GAO, C., VAN DER VAART, A. W. and ZHOU, H. H. (2020). A general framework for Bayes structured linear models. *Ann. Statist.* **48** 2848–2878. MR4152123 https://doi.org/10.1214/19-AOS1909
[46] GAO, C. and ZHANG, A. Y. (2022). Supplement to "Iterative algorithm for discrete structure recovery." https://doi.org/10.1214/21-AOS2140SUPP
[47] GIRAUD, C. and VERZELEN, N. (2018). Partial recovery bounds for clustering with the relaxed K-means. *Math. Stat. Learn.* **1** 317–374. MR4059724
[48] GIRVAN, M. and NEWMAN, M. E. J. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **99** 7821–7826. MR1908073 https://doi.org/10.1073/pnas.122653799
[49] HAJEK, B., WU, Y. and XU, J. (2016). Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Trans. Inf. Theory* **62** 2788–2797. MR3493879 https://doi.org/10.1109/TIT.2016.2546280
[50] HAJEK, B., WU, Y. and XU, J. (2016). Achieving exact cluster recovery threshold via semidefinite programming: Extensions. *IEEE Trans. Inf. Theory* **62** 5918–5937. MR3552431 https://doi.org/10.1109/TIT.2016.2594812
[51] HARTIGAN, J. A. (1975). *Clustering Algorithms. Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. MR0405726
[52] JI, P. and JIN, J. (2012). UPS delivers optimal phase diagram in high-dimensional variable selection. *Ann. Statist.* **40** 73–103. MR3013180 https://doi.org/10.1214/11-AOS947
[53] KANUNGO, T., MOUNT, D. M., NETANYAHU, N. S., PIATKO, C. D., SILVERMAN, R. and WU, A. Y. (2004). A local search approximation algorithm for k-means clustering. *Comput. Geom.* **28** 89–112. MR2062789 https://doi.org/10.1016/j.comgeo.2004.03.003
[54] KUMAR, A. and KANNAN, R. (2010). Clustering with spectral norm and the k-means algorithm. In 2010 *IEEE* 51*st Annual Symposium on Foundations of Computer Science—FOCS* 2010 299–308. IEEE Computer Soc., Los Alamitos, CA. MR3025203
[55] KUMAR, A., SABHARWAL, Y. and SEN, S. (2004). A simple linear time $(1+\epsilon)$-approximation algorithm for k-means clustering in any dimensions. In *Annual Symposium on Foundations of Computer Science* **45** 454–462. IEEE Computer Society Press.
[56] LERMAN, G. and SHI, Y. (2019). Robust group synchronization via cycle-edge message passing. Preprint. Available at arXiv:1912.11347.
[57] LESKOVEC, J., LANG, K. J. and MAHONEY, M. (2010). Empirical comparison of algorithms for network community detection. In *Proceedings of the* 19*th International Conference on World Wide Web* 631–640. ACM, New York.
[58] LING, S. (2020). Solving orthogonal group synchronization via convex and low-rank optimization: Tightness and landscape analysis. Preprint. Available at arXiv:2006.00902.
[59] LING, S. (2020). Near-optimal performance bounds for orthogonal and permutation group synchronization via spectral methods. Preprint. Available at arXiv:2008.05341.
[60] 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
[61] LOUNICI, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electron. J. Stat.* **2** 90–102. MR2386087 https://doi.org/10.1214/08-EJS177
[62] LU, Y. and ZHOU, H. H. (2016). Statistical and computational guarantees of Lloyd's algorithm and its variants. Preprint. Available at arXiv:1612.02099.
[63] LUCE, R. D. (2012). *Individual Choice Behavior: A Theoretical Analysis*. Wiley, New York. MR0108411
[64] MAHAJAN, M., NIMBHORKAR, P. and VARADARAJAN, K. (2012). The planar k-means problem is NP-hard. *Theoret. Comput. Sci.* **442** 13–21. MR2927097 https://doi.org/10.1016/j.tcs.2010.05.034
[65] MAO, C., WEED, J. and RIGOLLET, P. (2018). Minimax rates and efficient algorithms for noisy sorting. In *Algorithmic Learning Theory* 2018. *Proc. Mach. Learn. Res.* (*PMLR*) **83** 27. Proceedings of Machine Learning Research PMLR. MR3857331

[66] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34** 1436–1462. MR2278363 https://doi.org/10.1214/009053606000000281

[67] MONTANARI, A. and SEN, S. (2016). Semidefinite programs on sparse random graphs and their application to community detection. In *STOC'16—Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing* 814–827. ACM, New York. MR3536616 https://doi.org/10.1145/2897518.2897548

[68] MONTEILLER, P., CLAICI, S., CHIEN, E., MIRZAZADEH, F., SOLOMON, J. M. and YUROCHKIN, M. (2019). Alleviating label switching with optimal transport. In *Advances in Neural Information Processing Systems* 13634–13644.

[69] MOSSEL, E., NEEMAN, J. and SLY, A. (2014). Consistency thresholds for binary symmetric block models. Preprint. Available at arXiv:1407.1591.

[70] 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

[71] NDAOUD, M. (2018). Sharp optimal recovery in the two Gaussian mixture model. Preprint. Available at arXiv:1812.08078.

[72] NDAOUD, M. and TSYBAKOV, A. B. (2020). Optimal variable selection and adaptive noisy compressed sensing. *IEEE Trans. Inf. Theory* **66** 2517–2532. MR4087700 https://doi.org/10.1109/TIT.2020.2965738

[73] PACHAURI, D., KONDOR, R. and SINGH, V. (2013). Solving the multi-way matching problem by permutation synchronization. In *Advances in Neural Information Processing Systems* 1860–1868.

[74] PANANJADY, A., WAINWRIGHT, M. J. and COURTADE, T. A. (2018). Linear regression with shuffled data: Statistical and computational limits of permutation recovery. *IEEE Trans. Inf. Theory* **64** 3286–3300. MR3798377 https://doi.org/10.1109/TIT.2017.2776217

[75] PERRY, A., WEED, J., BANDEIRA, A. S., RIGOLLET, P. and SINGER, A. (2019). The sample complexity of multireference alignment. *SIAM J. Math. Data Sci.* **1** 497–517. MR4002723 https://doi.org/10.1137/18M1214317

[76] PERRY, A., WEIN, A. S., BANDEIRA, A. S. and MOITRA, A. (2016). Optimality and sub-optimality of pca for spiked random matrices and synchronization. Preprint. Available at arXiv:1609.05573.

[77] RAD, K. R. (2011). Nearly sharp sufficient conditions on exact sparsity pattern recovery. *IEEE Trans. Inf. Theory* **57** 4672–4679. MR2840483 https://doi.org/10.1109/TIT.2011.2145670

[78] SALIGRAMA, V. and ZHAO, M. (2011). Thresholded basis pursuit: LP algorithm for order-wise optimal support recovery for sparse and approximately sparse signals from noisy random measurements. *IEEE Trans. Inf. Theory* **57** 1567–1586. MR2815835 https://doi.org/10.1109/TIT.2011.2104512

[79] SIGWORTH, F. J. (1998). A maximum-likelihood approach to single-particle image refinement. *J. Struct. Biol.* **122** 328–339.

[80] SINGER, A. (2011). Angular synchronization by eigenvectors and semidefinite programming. *Appl. Comput. Harmon. Anal.* **30** 20–36. MR2737931 https://doi.org/10.1016/j.acha.2010.02.001

[81] SU, W. and CANDÈS, E. (2016). SLOPE is adaptive to unknown sparsity and asymptotically minimax. *Ann. Statist.* **44** 1038–1068. MR3485953 https://doi.org/10.1214/15-AOS1397

[82] WAINWRIGHT, M. J. (2009). Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inf. Theory* **55** 5728–5741. MR2597190 https://doi.org/10.1109/TIT.2009.2032816

[83] WANG, W., WAINWRIGHT, M. J. and RAMCHANDRAN, K. (2010). Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices. *IEEE Trans. Inf. Theory* **56** 2967–2979. MR2683451 https://doi.org/10.1109/TIT.2010.2046199

[84] WASSERMAN, L. and ROEDER, K. (2009). High-dimensional variable selection. *Ann. Statist.* **37** 2178–2201. MR2543689 https://doi.org/10.1214/08-AOS646

[85] WU, C.-F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11** 95–103. MR0684867 https://doi.org/10.1214/aos/1176346060

[86] WU, Y. and ZHOU, H. H. (2019). Randomly initialized EM algorithm for two-component Gaussian mixture achieves near optimality in $\sqrt{n}$ iterations. Preprint. Available at arXiv:1908.10935.

[87] XU, J., HSU, D. J. and MALEKI, A. (2016). Global analysis of expectation maximization for mixtures of two Gaussians. In *Advances in Neural Information Processing Systems* 2676–2684.

[88] XU, J., HSU, D. J. and MALEKI, A. (2018). Benefits of over-parameterization with EM. In *Advances in Neural Information Processing Systems* 10662–10672.

[89] YAN, J., CHO, M., ZHA, H., YANG, X. and CHU, S. M. (2015). Multi-graph matching via affinity optimization with graduated consistency regularization. *IEEE Trans. Pattern Anal. Mach. Intell.* **38** 1228–1242.

[90] YUN, S.-Y. and PROUTIERE, A. (2014). Accurate community detection in the stochastic block model via spectral algorithms. Preprint. Available at arXiv:1412.7335.

[91] 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
[92] ZHANG, A. Y. and ZHOU, H. H. (2020). Theoretical and computational guarantees of mean field variational inference for community detection. *Ann. Statist.* **48** 2575–2598. MR4152113 https://doi.org/10.1214/19-AOS1898
[93] ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449
[94] ZHONG, Y. and BOUMAL, N. (2018). Near-optimal bounds for phase synchronization. *SIAM J. Optim.* **28** 989–1016. MR3782406 https://doi.org/10.1137/17M1122025
[95] ZHOU, X., ZHU, M. and DANIILIDIS, K. (2015). Multi-image matching via fast alternating minimization. In *Proceedings of the IEEE International Conference on Computer Vision* 4032–4040.

The Annals of Statistics
2022, Vol. 50, No. 2, 1095–1123
https://doi.org/10.1214/21-AOS2141

FALSE DISCOVERY RATE CONTROL WITH UNKNOWN NULL DISTRIBUTION: IS IT POSSIBLE TO MIMIC THE ORACLE?

BY ETIENNE ROQUAIN[1,a] AND NICOLAS VERZELEN[2,b]

[1]*Université de Paris and Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation,* [a]*etienne.roquain@sorbonne-universite.fr*

[2]*INRAE, L'institut Agro, Université de Montpellier, MISTEA,* [b]*nicolas.verzelen@inrae.fr*

Classical multiple testing theory prescribes the null distribution, which is often too stringent an assumption for nowadays large scale experiments. This paper presents theoretical foundations to understand the limitations caused by ignoring the null distribution, and how it can be properly learned from the same data set, when possible. We explore this issue in the setting where the null distributions are Gaussian with unknown rescaling parameters (mean and variance) whereas the alternative distributions are let arbitrary. In that case, an oracle procedure is the Benjamini–Hochberg procedure applied with the true (unknown) null distribution and we aim at building a procedure that asymptotically mimics the performances of the oracle (AMO in short). Our main result establishes a phase transition at the sparsity boundary $n/\log(n)$: an AMO procedure exists if and only if the number of false nulls is of order less than $n/\log(n)$, where n is the total number of tests. Further sparsity boundaries are derived for general location models where the shape of the null distribution is not necessarily Gaussian. In light of our impossibility results, we also pursue the less stringent aim of building a nonparametric confidence region for the null distribution. From a practical perspective, this provides goodness-of-fit tests for the null distribution and allows to assess the reliability of empirical null procedures via novel diagnostic graphs. Our results are illustrated on numerical experiments and real data sets, as detailed in a companion vignette (Roquain and Verzelen (2021)).

REFERENCES

AMAR, D., SHAMIR, R. and YEKUTIELI, D. (2017). Extracting replicable associations across multiple studies: Empirical Bayes algorithms for controlling the false discovery rate. *PLoS Comput. Biol.* **13** e1005700. https://doi.org/10.1371/journal.pcbi.1005700

ARIAS-CASTRO, E. and CHEN, S. (2017). Distribution-free multiple testing. *Electron. J. Stat.* **11** 1983–2001. MR3651021 https://doi.org/10.1214/17-EJS1277

AZRIEL, D. and SCHWARTZMAN, A. (2015). The empirical distribution of a large number of correlated normal variables. *J. Amer. Statist. Assoc.* **110** 1217–1228. MR3420696 https://doi.org/10.1080/01621459.2014.958156

BARBER, R. F. and CANDÈS, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.* **43** 2055–2085. MR3375876 https://doi.org/10.1214/15-AOS1337

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

BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245 https://doi.org/10.1214/aos/1013699998

BLANCHARD, G., LEE, G. and SCOTT, C. (2010). Semi-supervised novelty detection. *J. Mach. Learn. Res.* **11** 2973–3009. MR2746544

BOURGON, R., GENTLEMAN, R. and HUBER, W. (2010). Independent filtering increases detection power for high-throughput experiments. *Proc. Natl. Acad. Sci. USA* **107** 9546–9551. https://doi.org/10.1073/pnas.0914005107

MSC2020 subject classifications. Primary 62G10; secondary 62C20.

Key words and phrases. Benjamini–Hochberg procedure, false discovery rate, minimax, multiple testing, phase transition, sparsity, null distribution.

CAI, T. T. and JIN, J. (2010). Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. *Ann. Statist.* **38** 100–145. MR2589318 https://doi.org/10.1214/09-AOS696

CAI, T. T. and SUN, W. (2009). Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *J. Amer. Statist. Assoc.* **104** 1467–1481. MR2597000 https://doi.org/10.1198/jasa.2009.tm08415

CAI, T. T., SUN, W. and WANG, W. (2019). Covariate-assisted ranking and screening for large-scale two-sample inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 187–234. MR3928141

CARPENTIER, A., DELATTRE, S., ROQUAIN, E. and VERZELEN, N. (2021). Estimating minimum effect with outlier selection. *Ann. Statist.* **49** 272–294. MR4206678 https://doi.org/10.1214/20-AOS1956

CASTILLO, I. and ROQUAIN, É. (2020). On spike and slab empirical Bayes multiple testing. *Ann. Statist.* **48** 2548–2574. MR4152112 https://doi.org/10.1214/19-AOS1897

CHEN, M., GAO, C. and REN, Z. (2018). Robust covariance and scatter matrix estimation under Huber's contamination model. *Ann. Statist.* **46** 1932–1960. MR3845006 https://doi.org/10.1214/17-AOS1607

CONSORTIUM, E. P. et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature* **447** 799.

DICKHAUS, T. (2014). *Simultaneous Statistical Inference*. Springer, Heidelberg. With applications in the life sciences. MR3184277 https://doi.org/10.1007/978-3-642-45182-9

DONOHO, D. and JIN, J. (2006). Asymptotic minimaxity of false discovery rate thresholding for sparse exponential data. *Ann. Statist.* **34** 2980–3018. MR2329475 https://doi.org/10.1214/009053606000000920

EFRON, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *J. Amer. Statist. Assoc.* **99** 96–104. MR2054289 https://doi.org/10.1198/016214504000000089

EFRON, B. (2007). Doing thousands of hypothesis tests at the same time. *Metron* **LXV** 3–21.

EFRON, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.* **23** 1–22. MR2431866 https://doi.org/10.1214/07-STS236

EFRON, B. (2009). Empirical Bayes estimates for large-scale prediction problems. *J. Amer. Statist. Assoc.* **104** 1015–1028. MR2562003 https://doi.org/10.1198/jasa.2009.tm08523

EFRON, B., TIBSHIRANI, R., STOREY, J. D. and TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96** 1151–1160. MR1946571 https://doi.org/10.1198/016214501753382129

FAN, J. and HAN, X. (2017). Estimation of the false discovery proportion with unknown dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1143–1164. MR3689312 https://doi.org/10.1111/rssb.12204

FAN, J., HAN, X. and GU, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *J. Amer. Statist. Assoc.* **107** 1019–1035. MR3010887 https://doi.org/10.1080/01621459.2012.720478

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

FRIGUET, C., KLOAREG, M. and CAUSEUR, D. (2009). A factor model approach to multiple testing under dependence. *J. Amer. Statist. Assoc.* **104** 1406–1415. MR2750571 https://doi.org/10.1198/jasa.2009.tm08332

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

GHOSH, D. (2012). Incorporating the empirical null hypothesis into the Benjamini–Hochberg procedure. *Stat. Appl. Genet. Mol. Biol.* **11** 11. MR2958610 https://doi.org/10.1515/1544-6115.1735

GOLUB, T. R., SLONIM, D. K., TAMAYO, P., HUARD, C., GAASENBEEK, M., MESIROV, J. P., COLLER, H., LOH, M. L., DOWNING, J. R. et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286** 531–537.

HEDENFALK, I., DUGGAN, D., CHEN, Y., RADMACHER, M., BITTNER, M., SIMON, R., MELTZER, P., GUSTERSON, B., ESTELLER, M. et al. (2001). Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.* **344** 539–548.

HELLER, R. and YEKUTIELI, D. (2014). Replicability analysis for genome-wide association studies. *Ann. Appl. Stat.* **8** 481–498. MR3191999 https://doi.org/10.1214/13-AOAS697

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

HUBER, P. J. (2011). Robust statistics. In *International Encyclopedia of Statistical Science* 1248–1251. Springer, Berlin.

JIANG, W. and YU, W. (2016). Controlling the joint local false discovery rate is more powerful than meta-analysis methods in joint analysis of summary statistics from multiple genome-wide association studies. *Bioinformatics* **33** 500–507.

JIN, J. and CAI, T. T. (2007). Estimating the null and the proportional of nonnull effects in large-scale multiple comparisons. *J. Amer. Statist. Assoc.* **102** 495–506. MR2325113 https://doi.org/10.1198/016214507000000167

JING, B.-Y., KONG, X.-B. and ZHOU, W. (2014). FDR control in multiple testing under non-normality. *Statist. Sinica* **24** 1879–1899. MR3308667

LEE, N., KIM, A.-Y., PARK, C.-H. and KIM, S.-H. (2016). An improvement on local fdr analysis applied to functional mri data. *J. Neurosci. Methods* **267** 115–125.

LEEK, J. T. and STOREY, J. D. (2008). A general framework for multiple testing dependence. *Proc. Natl. Acad. Sci. USA* **105** 18718–18723.

MARY, D. and ROQUAIN, E. (2021). Semi-supervised multiple testing. arXiv preprint. Available at arXiv:2106.13501.

MERLEVÈDE, F., PELIGRAD, M. and RIO, E. (2009). Bernstein inequality and moderate deviations under strong mixing conditions. In *High Dimensional Probability V: The Luminy Volume. Inst. Math. Stat. (IMS) Collect.* **5** 273–292. IMS, Beachwood, OH. MR2797953 https://doi.org/10.1214/09-IMSCOLL518

MILLER, C. J., GENOVESE, C., NICHOL, R. C., WASSERMAN, L., CONNOLLY, A., REICHART, D., HOPKINS, A., SCHNEIDER, J. and MOORE, A. (2001). Controlling the false-discovery rate in astrophysical data analysis. *Astron. J.* **122** 3492–3505.

MURALIDHARAN, O. (2010). An empirical Bayes mixture method for effect size and false discovery rate estimation. *Ann. Appl. Stat.* **4** 422–438. MR2758178 https://doi.org/10.1214/09-AOAS276

NAAMAN, M. (2021). On the tight constant in the multivariate Dvoretzky–Kiefer–Wolfowitz inequality. *Statist. Probab. Lett.* **173** 109088. MR4233456 https://doi.org/10.1016/j.spl.2021.109088

NGUYEN, V. H. and MATIAS, C. (2014). On efficient estimators of the proportion of true null hypotheses in a multiple testing setup. *Scand. J. Stat.* **41** 1167–1194. MR3277044 https://doi.org/10.1111/sjos.12091

PADILLA, M. and BICKEL, D. R. (2012). Estimators of the local false discovery rate designed for small numbers of tests. *Stat. Appl. Genet. Mol. Biol.* **11** 4. MR2990984 https://doi.org/10.1515/1544-6115.1807

POLLARD, K. S. and VAN DER LAAN, M. J. (2004). Choice of a null distribution in resampling-based multiple testing. *J. Statist. Plann. Inference* **125** 85–100. MR2086890 https://doi.org/10.1016/j.jspi.2003.07.019

RABINOVICH, M., RAMDAS, A., JORDAN, M. I. and WAINWRIGHT, M. J. (2020). Optimal rates and trade-offs in multiple testing. *Statist. Sinica* **30** 741–762. MR4214160

REBAFKA, T., ROQUAIN, E. and VILLERS, F. (2019). Graph inference with clustering and false discovery rate control. arXiv preprint. Available at arXiv:1907.10176.

ROQUAIN, E. and VAN DE WIEL, M. A. (2009). Optimal weighting for false discovery rate control. *Electron. J. Stat.* **3** 678–711. MR2521216 https://doi.org/10.1214/09-EJS430

ROQUAIN, E. and VERZELEN, N. (2021). False discovery rate control with unknown null distribution: Illustrations on real data sets. Available at https://github.com/eroquain/empiricalnull/blob/main/vignette.pdf.

ROQUAIN, E. and VERZELEN, N. (2022). Supplement to "False discovery rate control with unknown null distribution: is it possible to mimic the oracle?." https://doi.org/10.1214/21-AOS2141SUPP

ROSSET, S., HELLER, R., PAINSKY, A. and AHARONI, E. (2020). Optimal and maximin procedures for multiple testing problems.

SCHWARTZMAN, A. (2008). Empirical null and false discovery rate inference for exponential families. *Ann. Appl. Stat.* **2** 1332–1359. MR2655662 https://doi.org/10.1214/08-AOAS184

SCHWARTZMAN, A. (2010). Comment: "Correlated z-values and the accuracy of large-scale statistical estimates". *J. Amer. Statist. Assoc.* **105** 1059–1063. MR2752600 https://doi.org/10.1198/jasa.2010.tm10237

STEPHENS, M. (2017). False discovery rates: A new deal. *Biostatistics* **18** 275–294. MR3824755 https://doi.org/10.1093/biostatistics/kxw041

SULIS, S., MARY, D. and BIGOT, L. (2017). A study of periodograms standardized using training datasets and application to exoplanet detection. *IEEE Trans. Signal Process.* **65** 2136–2150. MR3608115 https://doi.org/10.1109/TSP.2017.2652391

SUN, W. and CAI, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.* **102** 901–912. MR2411657 https://doi.org/10.1198/016214507000000545

SUN, W. and CAI, T. T. (2009). Large-scale multiple testing under dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 393–424. MR2649603 https://doi.org/10.1111/j.1467-9868.2008.00694.x

SUN, L. and STEPHENS, M. (2018). Solving the empirical bayes normal means problem with correlated noise. arXiv preprint. Available at arXiv:1812.07488.

SZALAY, A. S., CONNOLLY, A. J. and SZOKOLY, G. P. (1999). Simultaneous multicolor detection of faint galaxies in the hubble deep field. *Astron. J.* **117** 68–74.

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

VAN'T WOUT, A. B., LEHRMAN, G. K., MIKHEEVA, S. A., O'KEEFFE, G. C., KATZE, M. G., BUMGARNER, R. E., GEISS, G. K. and MULLINS, J. I. (2003). Cellular gene expression upon human immunodeficiency virus type 1 infection of cd4+-t-cell lines. *J. Virol.* **77** 1392–1402.

ZABLOCKI, R. W., SCHORK, A. J., LEVINE, R. A., ANDREASSEN, O. A., DALE, A. M. and THOMPSON, W. K. (2014). Covariate-modulated local false discovery rate for genome-wide association studies. *Bioinformatics* **30** 2098–2104.

The Annals of Statistics
2022, Vol. 50, No. 2, 1124–1143
https://doi.org/10.1214/21-AOS2142

MAX-SUM TESTS FOR CROSS-SECTIONAL INDEPENDENCE OF HIGH-DIMENSIONAL PANEL DATA

BY LONG FENG[1,a], TIEFENG JIANG[2,b], BINGHUI LIU[3,c] AND WEI XIONG[4,d]

[1]*School of Statistics and Data Science, LPMC & KLMDASR, Nankai University,* [a]*flnankai@gmail.com*

[2]*School of Statistics, University of Minnesota,* [b]*jiang040@umn.edu*

[3]*Key Laboratory of Applied Statistics of MOE & School of Mathematics and Statistics, Northeast Normal University,* [c]*liubh100@nenu.edu.cn*

[4]*School of Statistics, University of International Business and Economics,* [d]*xiongwei@uibe.edu.cn*

We consider a testing problem for cross-sectional independence for high-dimensional panel data, where the number of cross-sectional units is potentially much larger than the number of observations. The cross-sectional independence is described through linear regression models. We study three tests named the sum, the max and the max-sum tests, where the latter two are new. The sum test is initially proposed by Breusch and Pagan (1980). We design the max and sum tests for sparse and nonsparse correlation coefficients of random errors between the linear regression models, respectively. And the max-sum test is devised to compromise both situations on the correlation coefficients. Indeed, our simulation shows that the max-sum test outperforms the previous two tests. This makes the max-sum test very useful in practice where sparsity or not for a set of numbers is usually vague. Toward the theoretical analysis of the three tests, we have settled two conjectures regarding the sum of squares of sample correlation coefficients asked by Pesaran (2004 and 2008). In addition, we establish the asymptotic theory for maxima of sample correlation coefficients appeared in the linear regression model for panel data, which is also the first successful attempt to our knowledge. To study the max-sum test, we create a novel method to show asymptotic independence between maxima and sums of dependent random variables. We expect the method itself is useful for other problems of this nature. Finally, an extensive simulation study as well as a case study are carried out. They demonstrate advantages of our proposed methods in terms of both empirical powers and robustness for correlation coefficients of residuals regardless of sparsity or not.

REFERENCES

[1] BALTAGI, B. H. (2013). *Econometric Analysis of Panel Data*, 5 ed. Wiley, New York.

[2] BALTAGI, B. H., KAO, C. and PENG, B. (2016). Testing cross-sectional correlation in large panel data models with serial correlation. *Econometrics* **4** 1–24.

[3] BEAULIEU, M.-C., DUFOUR, J.-M. and KHALAF, L. (2007). Multivariate tests of mean-variance efficiency with possibly non-Gaussian errors: An exact simulation-based approach. *J. Bus. Econom. Statist.* **25** 398–410. MR2410025 https://doi.org/10.1198/073500106000000468

[4] BREUSCH, T. S. and PAGAN, A. R. (1980). The Lagrange multiplier test and its applications to model specification in econometrics. *Rev. Econ. Stud.* **47** 239–253. MR0611118 https://doi.org/10.2307/2297111

[5] CAI, T., FAN, J. and JIANG, T. (2013). Distributions of angles in random packing on spheres. *J. Mach. Learn. Res.* **14** 1837–1864. MR3104497

[6] CAI, T. and LIU, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.* **106** 672–684. MR2847949 https://doi.org/10.1198/jasa.2011.tm10560

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

MSC2020 subject classifications. 62F03, 62H15.

Key words and phrases. High-dimensional data, panel data models, hypothesis tests, cross-sectional independence, asymptotic normality, extreme-value distribution, asymptotic independence, max-sum test.

[8] CAI, T. T., LIU, W. and XIA, Y. (2014). Two-sample test of high dimensional means under dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 349–372. MR3164870 https://doi.org/10.1111/rssb.12034

[9] CAI, T. T. and ZHANG, A. (2016). Inference for high-dimensional differential correlation matrices. *J. Multivariate Anal.* **143** 107–126. MR3431422 https://doi.org/10.1016/j.jmva.2015.08.019

[10] CHANG, J., YAO, Q. and ZHOU, W. (2017). Testing for high-dimensional white noise using maximum cross-correlations. *Biometrika* **104** 111–127. MR3626482 https://doi.org/10.1093/biomet/asw066

[11] CHUDIK, A. and PESARAN, M. H. (2015). Large panel data models with cross-sectional dependence: A survey. In *The Oxford Handbook of Panel Data* 3–45.

[12] FAMA, E. and FRENCH, K. (1993). Common risk factors in the returns on stocks and bonds. *J. Financ. Econ.* **33** 3–56.

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

[14] FENG, L., JIANG, T., LIU, B. and XIONG, W. (2022). Supplement to "Max-sum tests for cross-sectional independence of high-dimensional panel data." https://doi.org/10.1214/21-AOS2142SUPP

[15] GAGLIARDINI, P., OSSOLA, E. and SCAILLET, O. (2016). Time-varying risk premium in large cross-sectional equity data sets. *Econometrica* **84** 985–1046. MR3502510 https://doi.org/10.3982/ECTA11069

[16] GIBBONS, M. R., ROSS, S. A. and SHANKEN, J. (1989). A test of the efficiency of a given portfolio. *Econometrica* **57** 1121–1152. MR1014543 https://doi.org/10.2307/1913625

[17] HSIAO, C. (2014). *Analysis of Panel Data*, 3rd ed. *Econometric Society Monographs* **54**. Cambridge Univ. Press, New York. MR3560738 https://doi.org/10.1017/CBO9781139839327

[18] HSIAO, C., PESARAN, M. H. and PICK, A. (2012). Diagnostic tests of cross-sectional independence for limited dependent variable panel data models. *Oxf. Bull. Econ. Stat.* **74** 253–277.

[19] HSING, T. (1995). A note on the asymptotic independence of the sum and maximum of strongly mixing stationary random variables. *Ann. Probab.* **23** 938–947. MR1334178

[20] JAMES, B., JAMES, K. and QI, Y. (2007). Limit distribution of the sum and maximum from multivariate Gaussian sequences. *J. Multivariate Anal.* **98** 517–532. MR2293012 https://doi.org/10.1016/j.jmva.2006.06.009

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

[22] JIANG, T. (2019). Determinant of sample correlation matrix with application. *Ann. Appl. Probab.* **29** 1356–1397. MR3914547 https://doi.org/10.1214/17-AAP1362

[23] JIANG, T. and QI, Y. (2015). Likelihood ratio tests for high-dimensional normal distributions. *Scand. J. Stat.* **42** 988–1009. MR3426306 https://doi.org/10.1111/sjos.12147

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

[25] LI, D. and XUE, L. (2015). Joint limiting laws for high-dimensional independence tests. Available at arXiv:1512.08819.

[26] LI, Z., LAM, C., YAO, J. and YAO, Q. (2019). On testing for high-dimensional white noise. *Ann. Statist.* **47** 3382–3412. MR4025746 https://doi.org/10.1214/18-AOS1782

[27] LIU, W.-D., LIN, Z. and SHAO, Q.-M. (2008). The asymptotic distribution and Berry–Esseen bound of a new test for independence in high dimension with an application to stochastic optimization. *Ann. Appl. Probab.* **18** 2337–2366. MR2474539 https://doi.org/10.1214/08-AAP527

[28] MACKINLAY, A. R. and RICHARDSON, M. P. (1991). Using generalized method of moments to test mean ariance efficiency. *J. Finance* **46** 511–527.

[29] MOSCONE, F. and TOSETTI, E. (2009). A review and comparison of tests of cross-section independence in panels. *J. Econ. Surv.* **23** 528–561.

[30] PESARAN, M. H. (2004). General diagnostic test for cross section dependence in panels. IZA Discussion Paper No. 1240.

[31] PESARAN, M. H. (2015). Testing weak cross-sectional dependence in large panels. *Econometric Rev.* **34** 1088–1116. MR3292151 https://doi.org/10.1080/07474938.2014.956623

[32] PESARAN, M. H. (2015). *Time Series and Panel Data Econometrics*. Oxford University Press, London.

[33] PESARAN, M. H., ULLAH, A. and YAMAGATA, T. (2008). A bias-adjusted LM test of error cross-section independence. *Econom. J.* **11** 105–127. MR2400271 https://doi.org/10.1111/j.1368-423X.2007.00227.x

[34] SARAFIDIS, V. and WANSBEEK, T. (2012). Cross-sectional dependence in panel data analysis. *Econometric Rev.* **31** 483–531. MR2903105 https://doi.org/10.1080/07474938.2011.611458

[35] SARAFIDIS, V., YAMAGATA, T. and ROBERTSON, D. (2009). A test of cross section dependence for a linear dynamic panel model with regressors. *J. Econometrics* **148** 149–161. MR2500653 https://doi.org/10.1016/j.jeconom.2008.10.006

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

[37] SHAPIRO, S. S. and WILK, M. B. (1965). An analysis of variance test for normality: Complete samples. *Biometrika* **52** 591–611. MR0205384 https://doi.org/10.1093/biomet/52.3-4.591

[38] STEPHAN, F. F. (1934). Sampling errors and interpretations of social data ordered in time and space. *J. Amer. Statist. Assoc.* **29** 165–166.

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

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

[41] ZHENG, S., BAI, Z. and YAO, J. (2015). Substitution principle for CLT of linear spectral statistics of high-dimensional sample covariance matrices with applications to hypothesis testing. *Ann. Statist.* **43** 546–591. MR3316190 https://doi.org/10.1214/14-AOS1292

The Annals of Statistics
2022, Vol. 50, No. 2, 1144–1169
https://doi.org/10.1214/21-AOS2143

STATISTICAL INFERENCE FOR PRINCIPAL COMPONENTS OF SPIKED COVARIANCE MATRICES

BY ZHIGANG BAO[1,a], XIUCAI DING[2,d], JINGMING WANG[1,b] AND KE WANG[1,c]

[1]*Department of Mathematics, Hong Kong University of Science and Technology,* [a]*mazgbao@ust.hk,* [b]*jwangdm@connect.ust.hk,* [c]*kewang@ust.hk*

[2]*Department of Statistics, University of California, Davis,* [d]*xcading@ucdavis.edu*

In this paper, we study the asymptotic behavior of the extreme eigenvalues and eigenvectors of the high-dimensional spiked sample covariance matrices, in the supercritical case when a reliable detection of spikes is possible. In particular, we derive the joint distribution of the extreme eigenvalues and the generalized components of the associated eigenvectors, that is, the projections of the eigenvectors onto arbitrary given direction, assuming that the dimension and sample size are comparably large. In general, the joint distribution is given in terms of linear combinations of finitely many Gaussian and Chi-square variables, with parameters depending on the projection direction and the spikes. Our assumption on the spikes is fully general. First, the strengths of spikes are only required to be slightly above the critical threshold and no upper bound on the strengths is needed. Second, multiple spikes, that is, spikes with the same strength, are allowed. Third, no structural assumption is imposed on the spikes. Thanks to the general setting, we can then apply the results to various high dimensional statistical hypothesis testing problems involving both the eigenvalues and eigenvectors. Specifically, we propose accurate and powerful statistics to conduct hypothesis testing on the principal components. These statistics are data-dependent and adaptive to the underlying true spikes. Numerical simulations also confirm the accuracy and powerfulness of our proposed statistics and illustrate significantly better performance compared to the existing methods in the literature. In particular, our methods are accurate and powerful even when either the spikes are small or the dimension is large.

REFERENCES

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

[2] AVELLANEDA, M., HEALY, B., PAPANICOLAOU, A., PAPANICOLAOU, G. and XU, T. (2020). Principal Eigenportfolios for U.S. Equities. Available at SSRN.

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

[4] BAI, J. and NG, S. (2006). Evaluating latent and observed factors in macroeconomics and finance. *J. Econometrics* **131** 507–537. MR2276009 https://doi.org/10.1016/j.jeconom.2005.01.015

[5] BAI, J. and NG, S. (2013). Principal components estimation and identification of static factors. *J. Econometrics* **176** 18–29. MR3067022 https://doi.org/10.1016/j.jeconom.2013.03.007

[6] BAI, J. and NG, S. (2019). Rank regularized estimation of approximate factor models. *J. Econometrics* **212** 78–96. MR3994008 https://doi.org/10.1016/j.jeconom.2019.04.021

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

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

MSC2020 subject classifications. Primary 60B20, 62G10; secondary 62H10, 15B52, 62H25.

Key words and phrases. Random matrix, sample covariance matrix, eigenvector, spiked model, principal component, adaptive estimator.

[9] BAIK, J., BEN AROUS, G. and PÉCHÉ, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.* **33** 1643–1697. MR2165575 https://doi.org/10.1214/009117905000000233

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

[11] BAO, Z., DING, X. and WANG, K. (2021). Singular vector and singular subspace distribution for the matrix denoising model. *Ann. Statist.* **49** 370–392. MR4206682 https://doi.org/10.1214/20-AOS1960

[12] BAO, Z. and WANG, D. (2022). Eigenvector distribution in the critical regime of BBP transition. *Probab. Theory Related Fields* **182** 399–479. MR4367951 https://doi.org/10.1007/s00440-021-01062-4

[13] BAO, Z., DING, X., WANG, J. and WANG, K. (2022). Supplement to "Statistical inference for principal components of spiked covariance matrices." https://doi.org/10.1214/21-AOS2143SUPP

[14] BELINSCHI, S. T., BERCOVICI, H. and CAPITAINE, M. (2021). On the outlying eigenvalues of a polynomial in large independent random matrices. *Int. Math. Res. Not. IMRN* **4** 2588–2641. MR4218332 https://doi.org/10.1093/imrn/rnz080

[15] BELINSCHI, S. T., BERCOVICI, H., CAPITAINE, M. and FÉVRIER, M. (2017). Outliers in the spectrum of large deformed unitarily invariant models. *Ann. Probab.* **45** 3571–3625. MR3729610 https://doi.org/10.1214/16-AOP1144

[16] BENAYCH-GEORGES, F., GUIONNET, A. and MAIDA, M. (2011). Fluctuations of the extreme eigenvalues of finite rank deformations of random matrices. *Electron. J. Probab.* **16** 1621–1662. MR2835249 https://doi.org/10.1214/EJP.v16-929

[17] BENAYCH-GEORGES, F. and NADAKUDITI, R. R. (2011). The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Adv. Math.* **227** 494–521. MR2782201 https://doi.org/10.1016/j.aim.2011.02.007

[18] BENAYCH-GEORGES, F. and NADAKUDITI, R. R. (2012). The singular values and vectors of low rank perturbations of large rectangular random matrices. *J. Multivariate Anal.* **111** 120–135. MR2944410 https://doi.org/10.1016/j.jmva.2012.04.019

[19] BENIDIS, K., SUN, Y., BABU, P. and PALOMAR, D. P. (2016). Orthogonal sparse PCA and covariance estimation via Procrustes reformulation. *IEEE Trans. Signal Process.* **64** 6211–6226. MR3562717 https://doi.org/10.1109/TSP.2016.2605073

[20] BIRNBAUM, A., JOHNSTONE, I. M., NADLER, B. and PAUL, D. (2013). Minimax bounds for sparse PCA with noisy high-dimensional data. *Ann. Statist.* **41** 1055–1084. MR3113803 https://doi.org/10.1214/12-AOS1014

[21] BLOEMENDAL, A., ERDŐS, L., KNOWLES, A., YAU, H.-T. and YIN, J. (2014). Isotropic local laws for sample covariance and generalized Wigner matrices. *Electron. J. Probab.* **19** no. 33. MR3183577 https://doi.org/10.1214/ejp.v19-3054

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

[23] BLOEMENDAL, A. and VIRÁG, B. (2013). Limits of spiked random matrices I. *Probab. Theory Related Fields* **156** 795–825. MR3078286 https://doi.org/10.1007/s00440-012-0443-2

[24] BLOEMENDAL, A. and VIRÁG, B. (2016). Limits of spiked random matrices II. *Ann. Probab.* **44** 2726–2769. MR3531679 https://doi.org/10.1214/15-AOP1033

[25] BUN, J., BOUCHAUD, J.-P. and POTTERS, M. (2017). Cleaning large correlation matrices: Tools from random matrix theory. *Phys. Rep.* **666** 1–109. MR3590056 https://doi.org/10.1016/j.physrep.2016.10.005

[26] CAPITAINE, M. (2018). Limiting eigenvectors of outliers for spiked information-plus-noise type matrices. In *Séminaire de Probabilités XLIX. Lecture Notes in Math.* **2215** 119–164. Springer, Cham. MR3837103

[27] CAPITAINE, M. and DONATI-MARTIN, C. (2021). Non universality of fluctuations of outlier eigenvectors for block diagonal deformations of Wigner matrices. *ALEA Lat. Am. J. Probab. Math. Stat.* **18** 129–165. MR4198872 https://doi.org/10.30757/alea.v18-07

[28] CAPITAINE, M., DONATI-MARTIN, C. and FÉRAL, D. (2009). The largest eigenvalues of finite rank deformation of large Wigner matrices: Convergence and nonuniversality of the fluctuations. *Ann. Probab.* **37** 1–47. MR2489158 https://doi.org/10.1214/08-AOP394

[29] CAPITAINE, M., DONATI-MARTIN, C. and FÉRAL, D. (2012). Central limit theorems for eigenvalues of deformations of Wigner matrices. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 107–133. MR2919200 https://doi.org/10.1214/10-AIHP410

[30] DING, X. (2020). High dimensional deformed rectangular matrices with applications in matrix denoising. *Bernoulli* **26** 387–417. MR4036038 https://doi.org/10.3150/19-BEJ1129

[31] DING, X. and YANG, F. (2021). Spiked separable covariance matrices and principal components. *Ann. Statist.* **49** 1113–1138. MR4255121 https://doi.org/10.1214/20-aos1995

[32] DING, X. C. and JI, H. C. (2020). Local laws for multiplication of random matrices and spiked invariant model. Available at arXiv:2010.16083.

[33] DOBRIBAN, E. (2017). Sharp detection in PCA under correlations: All eigenvalues matter. *Ann. Statist.* **45** 1810–1833. MR3670197 https://doi.org/10.1214/16-AOS1514

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

[35] DOBRIBAN, E. and OWEN, A. B. (2019). Deterministic parallel analysis: An improved method for selecting factors and principal components. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 163–183. MR3904784

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

[37] ERDŐS, L., KNOWLES, A. and YAU, H.-T. (2013). Averaging fluctuations in resolvents of random band matrices. *Ann. Henri Poincaré* **14** 1837–1926. MR3119922 https://doi.org/10.1007/s00023-013-0235-y

[38] FAN, J., FAN, Y., HAN, X. and LV, J. (2019). Asymptotic theory of eigenvectors for large random matrices. Available at arXiv:1902.06846.

[39] FAN, J., LIAO, Y. and MINCHEVA, M. (2011). High-dimensional covariance matrix estimation in approximate factor models. *Ann. Statist.* **39** 3320–3356. MR3012410 https://doi.org/10.1214/11-AOS944

[40] FAWCETT, T. (2006). An introduction to ROC analysis. *Pattern Recogn. Lett.* **27** 861–874.

[41] FÉRAL, D. and PÉCHÉ, S. (2007). The largest eigenvalue of rank one deformation of large Wigner matrices. *Comm. Math. Phys.* **272** 185–228. MR2291807 https://doi.org/10.1007/s00220-007-0209-3

[42] HALLIN, M., PAINDAVEINE, D. and VERDEBOUT, T. (2010). Optimal rank-based testing for principal components. *Ann. Statist.* **38** 3245–3299. MR2766852 https://doi.org/10.1214/10-AOS810

[43] JAMES, G., WITTEN, D., HASTIE, T. and TIBSHIRANI, R. (2013). *An Introduction to Statistical Learning: With Applications in R. Springer Texts in Statistics* **103**. Springer, New York. MR3100153 https://doi.org/10.1007/978-1-4614-7138-7

[44] JOHNSTONE, I. and YANG, J. (2018). Notes on asymptotics of sample eigenstructure for spiked covariance models with non-Gaussian data. Available at arXiv:1810.10427.

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

[46] JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693. MR2751448 https://doi.org/10.1198/jasa.2009.0121

[47] JOLLIFFE, I. T. (2002). *Principal Component Analysis*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2036084

[48] KE, Z., MA, Y. and LIN, X. (2020). Estimation of the number of spiked eigenvalues in a covariance matrix by bulk eigenvalue matching analysis. ArXiv preprint. Available at arXiv:2006.00436.

[49] KNOWLES, A. and YIN, J. (2013). The isotropic semicircle law and deformation of Wigner matrices. *Comm. Pure Appl. Math.* **66** 1663–1750. MR3103909 https://doi.org/10.1002/cpa.21450

[50] KNOWLES, A. and YIN, J. (2014). The outliers of a deformed Wigner matrix. *Ann. Probab.* **42** 1980–2031. MR3262497 https://doi.org/10.1214/13-AOP855

[51] KNOWLES, A. and YIN, J. (2017). Anisotropic local laws for random matrices. *Probab. Theory Related Fields* **169** 257–352. MR3704770 https://doi.org/10.1007/s00440-016-0730-4

[52] KOLTCHINSKII, V. and LOUNICI, K. (2017). New asymptotic results in principal component analysis. *Sankhya A* **79** 254–297. MR3707422 https://doi.org/10.1007/s13171-017-0106-6

[53] LAMBERT, Z. V., WILDT, A. R. and DURAND, R. M. (1991). Approximating confidence intervals for factor loadings. *Multivar. Behav. Res.* **26** 421–434.

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

[55] LOUBATON, P. and VALLET, P. (2011). Almost sure localization of the eigenvalues in a Gaussian information plus noise model—application to the spiked models. *Electron. J. Probab.* **16** 1934–1959. MR2851051 https://doi.org/10.1214/EJP.v16-943

[56] MA, Z. (2013). Sparse principal component analysis and iterative thresholding. *Ann. Statist.* **41** 772–801. MR3099121 https://doi.org/10.1214/13-AOS1097

[57] MONASSON, R. and VILLAMAINA, D. (2015). Estimating the principal components of correlation matrices from all their empirical eigenvectors. *Europhys. Lett.* **112** 50001.

[58] MORALES-JIMENEZ, D., JOHNSTONE, I. M., MCKAY, M. R. and YANG, J. (2021). Asymptotics of eigenstructure of sample correlation matrices for high-dimensional spiked models. *Statist. Sinica* **31** 571–601. MR4286186 https://doi.org/10.5705/ss.20

[59] NADAKUDITI, R. R. and EDELMAN, A. (2008). Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples. *IEEE Trans. Signal Process.* **56** 2625–2638. MR1500236 https://doi.org/10.1109/TSP.2008.917356

[60] NADAKUDITI, R. R. and SILVERSTEIN, J. W. (2010). Fundamental limit of sample generalized eigenvalue based detection of signals in noise using relatively few signal-bearing and noise-only samples. *IEEE J. Sel. Top. Signal Process.* **4** 468–480.

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

[62] NADLER, B. and JOHNSTONE, I. M. (2011). Detection performance of Roy's largest root test when the noise covariance matrix is arbitrary. In 2011 IEEE Statistical Signal Processing Workshop 681–684.

[63] NAUMOV, A., SPOKOINY, V. and ULYANOV, V. (2019). Bootstrap confidence sets for spectral projectors of sample covariance. *Probab. Theory Related Fields* **174** 1091–1132. MR3980312 https://doi.org/10.1007/s00440-018-0877-2

[64] ONATSKI, A. (2009). Testing hypotheses about the numbers of factors in large factor models. *Econometrica* **77** 1447–1479. MR2561070 https://doi.org/10.3982/ECTA6964

[65] ONATSKI, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *J. Econometrics* **168** 244–258. MR2923766 https://doi.org/10.1016/j.jeconom.2012.01.034

[66] ONATSKI, A., MOREIRA, M. J. and HALLIN, M. (2013). Asymptotic power of sphericity tests for high-dimensional data. *Ann. Statist.* **41** 1204–1231. MR3113808 https://doi.org/10.1214/13-AOS1100

[67] ONATSKI, A., MOREIRA, M. J. and HALLIN, M. (2014). Signal detection in high dimension: The multispiked case. *Ann. Statist.* **42** 225–254. MR3189485 https://doi.org/10.1214/13-AOS1181

[68] OORT, F. J. (2011). Likelihood-based confidence intervals in exploratory factor analysis. *Struct. Equ. Model.* **18** 383–396. MR2822246 https://doi.org/10.1080/10705511.2011.582009

[69] PASSEMIER, D. and YAO, J. (2014). Estimation of the number of spikes, possibly equal, in the high-dimensional case. *J. Multivariate Anal.* **127** 173–183. MR3188885 https://doi.org/10.1016/j.jmva.2014.02.017

[70] PAUL, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17** 1617–1642. MR2399865

[71] PÉCHÉ, S. (2006). The largest eigenvalue of small rank perturbations of Hermitian random matrices. *Probab. Theory Related Fields* **134** 127–173. MR2221787 https://doi.org/10.1007/s00440-005-0466-z

[72] 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

[73] PIZZO, A., RENFREW, D. and SOSHNIKOV, A. (2013). On finite rank deformations of Wigner matrices. *Ann. Inst. Henri Poincaré Probab. Stat.* **49** 64–94. MR3060148 https://doi.org/10.1214/11-AIHP459

[74] SHEN, H. and HUANG, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *J. Multivariate Anal.* **99** 1015–1034. MR2419336 https://doi.org/10.1016/j.jmva.2007.06.007

[75] SILIN, I. and FAN, J. (2020). Hypothesis testing for eigenspaces of covariance matrix. Available at arXiv:2020.09810.

[76] SILIN, I. and SPOKOINY, V. (2018). Bayesian inference for spectral projectors of the covariance matrix. *Electron. J. Stat.* **12** 1948–1987. MR3815302 https://doi.org/10.1214/18-EJS1451

[77] TYLER, D. E. (1981). Asymptotic inference for eigenvectors. *Ann. Statist.* **9** 725–736. MR0619278

[78] TYLER, D. E. (1983). A class of asymptotic tests for principal component vectors. *Ann. Statist.* **11** 1243–1250. MR0720269 https://doi.org/10.1214/aos/1176346337

[79] XI, H., YANG, F. and YIN, J. (2020). Convergence of eigenvector empirical spectral distribution of sample covariance matrices. *Ann. Statist.* **48** 953–982. MR4102683 https://doi.org/10.1214/19-AOS1832

[80] YAMAMOTO, H., FUJIMORI, T., SATO, H., ISHIKAWA, G., KAMI, K. and OHASHI, Y. (2014). Statistical hypothesis testing of factor loading in principal component analysis and its application to metabolite set enrichment analysis. *BMC Bioinform.* **15** 51.

[81] YANG, F. (2020). Linear spectral statistics of eigenvectors of anisotropic sample covariance matrices. Available at arXiv:2005.00999.

[82] ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.* **15** 265–286. MR2252527 https://doi.org/10.1198/106186006X113430

The Annals of Statistics
2022, Vol. 50, No. 2, 1170–1192
https://doi.org/10.1214/21-AOS2144

RECONCILING DESIGN-BASED AND MODEL-BASED CAUSAL INFERENCES FOR SPLIT-PLOT EXPERIMENTS

BY ANQI ZHAO[1,a] AND PENG DING[2,b]

[1]*Department of Statistics and Data Science, National University of Singapore,* [a]*staza@nus.edu.sg*
[2]*Department of Statistics, University of California, Berkeley,* [b]*pengdingpku@berkeley.edu*

The split-plot design arose from agricultural science with experimental units, also known as the subplots, nested within groups known as the whole plots. It assigns different interventions at the whole-plot and subplot levels, respectively, providing a convenient way to accommodate hard-to-change factors. By design, subplots within the same whole plot receive the same level of the whole-plot intervention, and thereby induce a group structure on the final treatment assignments. A common strategy is to run an ordinary least squares (OLS) regression of the outcome on the treatment indicators coupled with the robust standard errors clustered at the whole-plot level. It does not give consistent estimators for the treatment effects of interest when the whole-plot sizes vary. Another common strategy is to fit a linear mixed-effects model of the outcome with normal random effects and errors. It is a purely model-based approach and can be sensitive to violations of the parametric assumptions. In contrast, design-based inference assumes no outcome models and relies solely on the controllable randomization mechanism determined by the physical experiment. We first extend the existing design-based inference based on the Horvitz–Thompson estimator to the Hajek estimator, and establish the finite-population central limit theorem for both under split-plot randomization. We then reconcile the results with those under the model-based approach, and propose two regression strategies, namely (i) the weighted least squares (WLS) fit of the unit-level data based on the inverse probability weighting and (ii) the OLS fit of the aggregate data based on whole-plot total outcomes, to reproduce the Hajek and Horvitz–Thompson estimators, respectively. This, together with the asymptotic conservativeness of the corresponding cluster-robust covariances for estimating the true design-based covariances as we establish in the process, justifies the validity of the regression estimators for design-based inference. In light of the flexibility of regression formulation for covariate adjustment, we further extend the theory to the case with covariates, and demonstrate the efficiency gain by regression-based covariate adjustment via both asymptotic theory and simulation. Importantly, all our theories are either numeric or design-based, and hold regardless of how well the regression equations represent the true data generating process.

REFERENCES

ABADIE, A., ATHEY, S., IMBENS, G. W. and WOOLDRIDGE, J. M. (2017). When should you adjust standard errors for clustering? Technical Report, National Bureau of Economic Research.

ABADIE, A., ATHEY, S., IMBENS, G. W. and WOOLDRIDGE, J. M. (2020). Sampling-based versus design-based uncertainty in regression analysis. *Econometrica* **88** 265–296. MR4084709 https://doi.org/10.3982/ecta12675

BASSE, G. and FELLER, A. (2018). Analyzing two-stage experiments in the presence of interference. *J. Amer. Statist. Assoc.* **113** 41–55. MR3803438 https://doi.org/10.1080/01621459.2017.1323641

BELL, R. M. and MCCAFFREY, D. F. (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Surv. Methodol.* **28** 169–181.

MSC2020 subject classifications. 62K15, 62J05, 62G05.

Key words and phrases. Cluster randomization, cluster-robust standard error, covariate adjustment, inverse probability weighting, potential outcome, randomization inference.

BOX, G. E. P. and ANDERSEN, S. L. (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumption. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **17** 1–34.

CAMERON, A. and MILLER, D. (2015). A practitioner's guide to cluster-robust inference. *J. Hum. Resour.* **50** 317–372.

CHONG, A., GONZALEZ-NAVARRO, M., KARLAN, D. and VALDIVIA, M. (2020). Do information technologies improve teenagers' sexual education? Evidence from a randomized evaluation in Colombia. *World Bank Econ. Rev.* **34** 371–392.

COX, D. R. and REID, N. (2000). *The Theory of the Design of Experiments*. Chapman and Hall/CRC, London.

DASGUPTA, T., PILLAI, N. S. and RUBIN, D. B. (2015). Causal inference from 2^K factorial designs by using potential outcomes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 727–753. MR3382595 https://doi.org/10.1111/rssb.12085

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

FOGARTY, C. B. (2018a). On mitigating the analytical limitations of finely stratified experiments. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 1035–1056. MR3874309 https://doi.org/10.1111/rssb.12290

FOGARTY, C. B. (2018b). Regression-assisted inference for the average treatment effect in paired experiments. *Biometrika* **105** 994–1000. MR3877880 https://doi.org/10.1093/biomet/asy034

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

FRICANO, C. J., DESPENZA JR., T., FRAZEL, P. W., LI, M., O'MALLEY, J. A., WESTBROOK, G. L. and LUIKART, B. W. (2014). Fatty acids increase neuronal hypertrophy of Pten knockdown neurons. *Front. Mol. Neuroscience* **7** 30.

FULLER, W. A. (2009). *Sampling Statistics*. Wiley, New York.

HINKELMANN, K. and KEMPTHORNE, O. (2008). *Design and Analysis of Experiments. Vol.* 1. Wiley, New York.

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

IMAI, K., JIANG, Z. and MALANI, A. (2021). Causal inference with interference and noncompliance in two-stage randomized experiments. *J. Amer. Statist. Assoc.* **116** 632–644. MR4270009 https://doi.org/10.1080/01621459.2020.1775612

IMAI, K., KING, G. and NALL, C. (2009). The essential role of pair matching in cluster-randomized experiments, with application to the Mexican universal health insurance evaluation. *Statist. Sci.* **24** 29–53. MR2561126 https://doi.org/10.1214/08-STS274

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

JI, X., FINK, G., ROBYN, P. J. and SMALL, D. S. (2017). Randomization inference for stepped-wedge cluster-randomized trials: An application to community-based health insurance. *Ann. Appl. Stat.* **11** 1–20. MR3634312 https://doi.org/10.1214/16-AOAS969

JONES, B. and NACHTSHEIM, C. J. (2009). Split-plot designs: What, why, and how. *J. Qual. Technol.* **41** 340–361.

KEMPTHORNE, O. (1952). *The Design and Analysis of Experiments*. Wiley, New York; CRC Press, London. MR0045368

LI, X. and DING, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *J. Amer. Statist. Assoc.* **112** 1759–1769. MR3750897 https://doi.org/10.1080/01621459.2017.1295865

LIANG, K. Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22. MR0836430 https://doi.org/10.1093/biomet/73.1.13

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

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

LIU, H. and YANG, Y. (2020). Regression-adjusted average treatment effect estimates in stratified randomized experiments. *Biometrika* **107** 935–948. MR4186497 https://doi.org/10.1093/biomet/asaa038

MACKINNON, J. G., NIELSEN, M. and WEBB, M. D. (2021). Cluster-robust inference: A guide to empirical practice. Working Paper No. 1456, Economics Department, Queen's University.

MIDDLETON, J. A. and ARONOW, P. M. (2015). Unbiased estimation of the average treatment effect in cluster-randomized experiments. *Statistics, Politics and Policy* **6** 39–75.

MIRATRIX, L. W. (2013). Adjusting treatment effect estimates by post-stratification in randomized experiments. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 369–396. MR3021392 https://doi.org/10.1111/j.1467-9868.2012.01048.x

MIRATRIX, L. W., WEISS, M. J. and HENDERSON, B. (2021). An applied researcher's guide to estimating effects from multisite individually randomized trials: Estimands, estimators, and estimates. *J. Res. Educ. Eff.* **14** 270–308.

MOEN, E. L., FRICANO-KUGLER, C. J., LUIKART, B. W. and O'MALLEY, A. J. (2016). Analyzing clustered data: Why and how to account for multiple observations nested within a study participant? *PLoS ONE* **11** e0146721. https://doi.org/10.1371/journal.pone.0146721

MUKERJEE, R. and DASGUPTA, T. (2021). Causal inference from possibly unbalanced split-plot designs: A randomization-based perspective. *Statist. Sinica*.

MUKERJEE, R., DASGUPTA, T. and RUBIN, D. B. (2018). Using standard tools from finite population sampling to improve causal inference for complex experiments. *J. Amer. Statist. Assoc.* **113** 868–881. MR3832233 https://doi.org/10.1080/01621459.2017.1294076

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

OHLSSON, E. (1989). Asymptotic normality for two-stage sampling from a finite population. *Probab. Theory Related Fields* **81** 341–352. MR0983089 https://doi.org/10.1007/BF00340058

OLKEN, B. A. (2007). Monitoring corruption: Evidence from a field experiment in Indonesia. *J. Polit. Econ.* **115** 200–249.

PASHLEY, N. E. and MIRATRIX, L. W. (2021). Insights on variance estimation for blocked and matched pairs designs. *J. Educ. Behav. Stat.* **46** 271–296.

PUSTEJOVSKY, J. E. and TIPTON, E. (2018). Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *J. Bus. Econom. Statist.* **36** 672–683. MR3871709 https://doi.org/10.1080/07350015.2016.1247004

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

SABBAGHI, A. and RUBIN, D. B. (2014). Comments on the Neyman-Fisher controversy and its consequences. *Statist. Sci.* **29** 267–284. MR3264542 https://doi.org/10.1214/13-STS454

SCHOCHET, P. Z. (2010). Is regression adjustment supported by the Neyman model for causal inference? *J. Statist. Plann. Inference* **140** 246–259. MR2568136 https://doi.org/10.1016/j.jspi.2009.07.008

SCHOCHET, P. Z., PASHLEY, N. E., MIRATRIX, L. W. and KAUTZ, T. (2021). Design-based ratio estimators and central limit theorems for clustered, blocked RCTs. *J. Amer. Statist. Assoc.* https://doi.org/10.1080/01621459.2021.1906685

NEYMAN, J. (1923). On the application of probability theory to agricultural experiments. *Statist. Sci.* **5** 465–472.

SU, F. and DING, P. (2021). Model-assisted analyses of cluster-randomized experiments. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **83** 994–1015. MR4349125 https://doi.org/10.1111/rssb.12468

WU, C.-F. (1981). On the robustness and efficiency of some randomized designs. *Ann. Statist.* **9** 1168–1177. MR0630100

WU, J. and DING, P. (2021). Randomization Tests for Weak Null Hypotheses in Randomized Experiments. *J. Amer. Statist. Assoc.* **116** 1898–1913. MR4353721 https://doi.org/10.1080/01621459.2020.1750415

WU, C. F. J. and HAMADA, M. S. (2009). *Experiments: Planning, Analysis, and Optimization*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR2583259

YATES, F. (1935). Complex experiments. *Suppl. J. R. Stat. Soc.* **2** 181–247.

YATES, F. (1937). *The Design and Analysis of Factorial Experiments*. Imperial Bureau of Soil Science London, Harpenden.

ZHAO, A. and DING, P. (2021a). Regression-based causal inference with factorial experiments: Estimands, model specifications, and design-based properties. *Biometrika*. arXiv:2101.02400.

ZHAO, A. and DING, P. (2021b). Covariate-adjusted Fisher randomization tests for the average treatment effect. *J. Econometrics* **225** 278–294. MR4328643 https://doi.org/10.1016/j.jeconom.2021.04.007

ZHAO, A., DING, P. (2022). Supplement to "Reconciling design-based and model-based causal inferences for split-plot experiments." https://doi.org/10.1214/21-AOS2144SUPP

ZHAO, A., DING, P., MUKERJEE, R. and DASGUPTA, T. (2018). Randomization-based causal inference from split-plot designs. *Ann. Statist.* **46** 1876–1903. MR3845004 https://doi.org/10.1214/17-AOS1605

The Annals of Statistics
2022, Vol. 50, No. 2, 1193–1219
https://doi.org/10.1214/21-AOS2145

ALL-IN-ONE ROBUST ESTIMATOR OF THE GAUSSIAN MEAN

BY ARNAK S. DALALYAN[1,a] AND ARSHAK MINASYAN[2,b]

[1]*ENSAE-CREST,* [a]*arnak.dalalyan@ensae.fr*
[2]*Department of Mathematics, Yerevan State University, YerevaNN,* [b]*minasyan@yerevann.com*

The goal of this paper is to show that a single robust estimator of the mean of a multivariate Gaussian distribution can enjoy five desirable properties. First, it is computationally tractable in the sense that it can be computed in a time, which is at most polynomial in dimension, sample size and the logarithm of the inverse of the contamination rate. Second, it is equivariant by translations, uniform scaling and orthogonal transformations. Third, it has a high breakdown point equal to 0.5, and a nearly minimax rate breakdown point approximately equal to 0.28. Fourth, it is minimax rate optimal, up to a logarithmic factor, when data consists of independent observations corrupted by adversarially chosen outliers. Fifth, it is asymptotically efficient when the rate of contamination tends to zero. The estimator is obtained by an iterative reweighting approach. Each sample point is assigned a weight that is iteratively updated by solving a convex optimization problem. We also establish a dimension-free nonasymptotic risk bound for the expected error of the proposed estimator. It is the first result of this kind in the literature and involves only the effective rank of the covariance matrix. Finally, we show that the obtained results can be extended to sub-Gaussian distributions, as well as to the cases of unknown rate of contamination or unknown covariance matrix.

REFERENCES

BALAKRISHNAN, S., DU, S. S., LI, J. and SINGH, A. (2017). Computationally efficient robust sparse estimation in high dimensions. In *COLT* 2017 169–212.

BATENI, A.-H. and DALALYAN, A. S. (2020). Confidence regions and minimax rates in outlier-robust estimation on the probability simplex. *Electron. J. Stat.* **14** 2653–2677. MR4124558 https://doi.org/10.1214/20-EJS1731

CAI, T. T. and LI, X. (2015). Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *Ann. Statist.* **43** 1027–1059. MR3346696 https://doi.org/10.1214/14-AOS1290

CANNINGS, T. I., FAN, Y. and SAMWORTH, R. J. (2020). Classification with imperfect training labels. *Biometrika* **107** 311–330. MR4108933 https://doi.org/10.1093/biomet/asaa011

CHEN, M., GAO, C. and REN, Z. (2016). A general decision theory for Huber's ϵ-contamination model. *Electron. J. Stat.* **10** 3752–3774. MR3579675 https://doi.org/10.1214/16-EJS1216

CHEN, M., GAO, C. and REN, Z. (2018). Robust covariance and scatter matrix estimation under Huber's contamination model. *Ann. Statist.* **46** 1932–1960. MR3845006 https://doi.org/10.1214/17-AOS1607

CHENG, Y., DIAKONIKOLAS, I. and GE, R. (2019). High-dimensional robust mean estimation in nearly-linear time. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms* 2755–2771. SIAM, Philadelphia, PA. MR3909640 https://doi.org/10.1137/1.9781611975482.171

CHERAPANAMJERI, Y., FLAMMARION, N. and BARTLETT, P. L. (2019). Fast mean estimation with sub-gaussian rates. In *Proceedings of COLT Proceedings of Machine Learning Research* **99** 786–806. PMLR.

CHINOT, G. (2020). ERM and RERM are optimal estimators for regression problems when malicious outliers corrupt the labels. *Electron. J. Stat.* **14** 3563–3605. MR4155965 https://doi.org/10.1214/20-EJS1754

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

COMMINGES, L. and DALALYAN, A. S. (2012). Tight conditions for consistency of variable selection in the context of high dimensionality. *Ann. Statist.* **40** 2667–2696. MR3097616 https://doi.org/10.1214/12-AOS1046

MSC2020 subject classifications. Primary 62H12; secondary 62F35.
Key words and phrases. Gaussian mean, robust estimation, breakdown point, minimax rate, computational tractability.

COMMINGES, L., COLLIER, O., NDAOUD, M. and TSYBAKOV, A. B. (2021). Adaptive robust estimation in sparse vector model. *Ann. Statist.* **49** 1347–1377. MR4298867 https://doi.org/10.1214/20-aos2002

COX, B., JUDITSKY, A. and NEMIROVSKI, A. (2014). Dual subgradient algorithms for large-scale nonsmooth learning problems. *Math. Program.* **148** 143–180. MR3274848 https://doi.org/10.1007/s10107-013-0725-1

DALALYAN, A. S. and MINASYAN, A. (2022). Supplement to "All-In-One Robust Estimator of the Gaussian Mean." https://doi.org/10.1214/21-AOS2145SUPP

DALALYAN, A. and THOMPSON, P. (2019). Outlier-robust estimation of a sparse linear model using ℓ_1-penalized Huber's M-estimator. In *NeurIPS* 32 13188–13198.

DEPERSIN, J. and LECUÉ, G. (2019). Robust subgaussian estimation of a mean vector in nearly linear time. Available at 1906.03058.

DEVROYE, L., LERASLE, M., LUGOSI, G. and OLIVEIRA, R. I. (2016). Sub-Gaussian mean estimators. *Ann. Statist.* **44** 2695–2725. MR3576558 https://doi.org/10.1214/16-AOS1440

DIAKONIKOLAS, I. and KANE, D. M. (2019). Recent advances in algorithmic high-dimensional robust statistics. *CoRR*. Available at 1911.05911.

DIAKONIKOLAS, I., KANE, D. M. and STEWART, A. (2017). Statistical query lower bounds for robust estimation of high-dimensional Gaussians and Gaussian mixtures (extended abstract). In 58*th Annual IEEE Symposium on Foundations of Computer Science—FOCS* 2017 73–84. IEEE Computer Soc., Los Alamitos, CA. MR3734219 https://doi.org/10.1109/FOCS.2017.16

DIAKONIKOLAS, I., KAMATH, G., KANE, D. M., LI, J., MOITRA, A. and STEWART, A. (2016). Robust estimators in high dimensions without the computational intractability. In 57*th Annual IEEE Symposium on Foundations of Computer Science—FOCS* 2016 655–664. IEEE Computer Soc., Los Alamitos, CA. MR3631028 https://doi.org/10.1109/FOCS.2016.85

DIAKONIKOLAS, I., KAMATH, G., KANE, D. M., LI, J., MOITRA, A. and STEWART, A. (2018). Robustly learning a Gaussian: Getting optimal error, efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms* 2683–2702. SIAM, Philadelphia, PA. MR3775959 https://doi.org/10.1137/1.9781611975031.171

DIAKONIKOLAS, I., KAMATH, G., KANE, D., LI, J., MOITRA, A. and STEWART, A. (2019a). Robust estimators in high-dimensions without the computational intractability. *SIAM J. Comput.* **48** 742–864. MR3945261 https://doi.org/10.1137/17M1126680

DIAKONIKOLAS, I., KANE, D., KARMALKAR, S., PRICE, E. and STEWART, A. (2019b). Outlier-robust high-dimensional sparse estimation via iterative filtering. In *NeurIPS* 2019 10688–10699.

DONG, Y., HOPKINS, S. B. and LI, J. (2019). Quantum entropy scoring for fast robust mean estimation and improved outlier detection. In *NeurIPS* 2019 6065–6075.

DONOHO, D. (1982). Breakdown properties of multivariate location estimators. Phd thesis, Harvard Univ.

DONOHO, D. L. and GASKO, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann. Statist.* **20** 1803–1827. MR1193313 https://doi.org/10.1214/aos/1176348890

DONOHO, D. and HUBER, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann. Wadsworth Statist./Probab. Ser.* 157–184. Wadsworth, Belmont, CA. MR0689745

ELSENER, A. and VAN DE GEER, S. (2018). Robust low-rank matrix estimation. *Ann. Statist.* **46** 3481–3509. MR3852659 https://doi.org/10.1214/17-AOS1666

GAO, C. (2020). Robust regression via mutivariate regression depth. *Bernoulli* **26** 1139–1170. MR4058363 https://doi.org/10.3150/19-BEJ1144

HAMPEL, F. R. (1968). Contributions to the theory of robust estimation. PhD thesis, Univ. California, Berkeley.

HOPKINS, S. B. (2018). Sub-gaussian mean estimation in polynomial time. *CoRR*. Available at 1809.07425.

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

KLOPP, O., LOUNICI, K. and TSYBAKOV, A. B. (2017). Robust matrix completion. *Probab. Theory Related Fields* **169** 523–564. MR3704775 https://doi.org/10.1007/s00440-016-0736-y

KOLTCHINSKII, V. and LOUNICI, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli* **23** 110–133. MR3556768 https://doi.org/10.3150/15-BEJ730

LAI, K. A., RAO, A. B. and VEMPALA, S. (2016). Agnostic estimation of mean and covariance. In 57*th Annual IEEE Symposium on Foundations of Computer Science—FOCS* 2016 665–674. IEEE Computer Soc., Los Alamitos, CA. MR3631029

LECUÉ, G. and LERASLE, M. (2019). Learning from MOM's principles: Le Cam's approach. *Stochastic Process. Appl.* **129** 4385–4410. MR4013866 https://doi.org/10.1016/j.spa.2018.11.024

LECUÉ, G. and LERASLE, M. (2020). Robust machine learning by median-of-means: Theory and practice. *Ann. Statist.* **48** 906–931. MR4102681 https://doi.org/10.1214/19-AOS1828

LEPSKI, O. V. and SPOKOINY, V. G. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *Ann. Statist.* **25** 2512–2546. MR1604408 https://doi.org/10.1214/aos/1030741083

LEPSKII, O. V. (1992). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.* **36** 682–697.

LI, A. H. and BRADIC, J. (2018). Boosting in the presence of outliers: Adaptive classification with nonconvex loss functions. *J. Amer. Statist. Assoc.* **113** 660–674. MR3832217 https://doi.org/10.1080/01621459.2016.1273116

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

LOPUHAÄ, H. P. and ROUSSEEUW, P. J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *Ann. Statist.* **19** 229–248. MR1091847 https://doi.org/10.1214/aos/1176347978

LUGOSI, G. and MENDELSON, S. (2019). Near-optimal mean estimators with respect to general norms. *Probab. Theory Related Fields* **175** 957–973. MR4026610 https://doi.org/10.1007/s00440-019-00906-4

LUGOSI, G. and MENDELSON, S. (2020). Risk minimization by median-of-means tournaments. *J. Eur. Math. Soc.* (*JEMS*) **22** 925–965. MR4055993 https://doi.org/10.4171/jems/937

LUGOSI, G. and MENDELSON, S. (2021). Robust multivariate mean estimation: The optimality of trimmed mean. *Ann. Statist.* **49** 393–410. MR4206683 https://doi.org/10.1214/20-AOS1961

MARONNA, R. A., MARTIN, R. D. and YOHAI, V. J. (2006). *Robust Statistics*: *Theory and Methods*. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR2238141 https://doi.org/10.1002/0470010940

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

POLZEHL, J. and SPOKOINY, V. G. (2000). Adaptive weights smoothing with applications to image restoration. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 335–354. MR1749543 https://doi.org/10.1111/1467-9868.00235

ROUSSEEUW, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.* **79** 871–880. MR0770281

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

ROUSSEEUW, P. and HUBERT, M. (2013). High-breakdown estimators of multivariate location and scatter. In *Robustness and Complex Data Structures* 49–66. Springer, Heidelberg. MR3135873 https://doi.org/10.1007/978-3-642-35494-6_4

SOLTANOLKOTABI, M. and CANDÉS, E. J. (2012). A geometric analysis of subspace clustering with outliers. *Ann. Statist.* **40** 2195–2238. MR3059081 https://doi.org/10.1214/12-AOS1034

STAHEL, W. (1981). Robuste Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen. Phd thesis, ETH, Zurich.

VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge Univ. Press, Cambridge. MR2963170

ZHU, B., JIAO, J. and STEINHARDT, J. (2020). When does the Tukey median work?. In *IEEE International Symposium on Information Theory* (*ISIT*).

The Annals of Statistics
2022, Vol. 50, No. 2, 1220–1245
https://doi.org/10.1214/21-AOS2146

INFERENCE FOR LOW-RANK TENSORS—NO NEED TO DEBIAS

BY DONG XIA[1,a], ANRU R. ZHANG[2,b] AND YUCHEN ZHOU[3,c]

[1]*Department of Mathematics, Hong Kong University of Science and Technology,* [a]*madxia@ust.hk*

[2]*Departments of Biostatistics & Bioinformatics, Computer Science, Mathematics, and Statistical Science, Duke University,* [b]*anru.zhang@duke.edu*

[3]*Department of Statistics and Data Science, The Wharton School, University of Pennsylvania,* [c]*yczhou@wharton.upenn.edu*

In this paper, we consider the statistical inference for several low-rank tensor models. Specifically, in the Tucker low-rank tensor PCA or regression model, provided with any estimates achieving some attainable error rate, we develop the data-driven confidence regions for the singular subspace of the parameter tensor based on the asymptotic distribution of an updated estimate by two-iteration alternating minimization. The asymptotic distributions are established under some essential conditions on the signal-to-noise ratio (in PCA model) or sample size (in regression model). If the parameter tensor is further orthogonally decomposable, we develop the methods and nonasymptotic theory for inference on each individual singular vector. For the rank-one tensor PCA model, we establish the asymptotic distribution for general linear forms of principal components and confidence interval for each entry of the parameter tensor. Finally, numerical simulations are presented to corroborate our theoretical discoveries.

In all of these models, we observe that different from many matrix/vector settings in existing work, debiasing is not required to establish the asymptotic distribution of estimates or to make statistical inference on low-rank tensors. In fact, due to the widely observed statistical-computational-gap for low-rank tensor estimation, one usually requires stronger conditions than the statistical (or information-theoretic) limit to ensure the computationally feasible estimation is achievable. Surprisingly, such conditions "incidentally" render a feasible low-rank tensor inference without debiasing.

REFERENCES

[1] 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

[2] ANANDKUMAR, A., GE, R. and JANZAMIN, M. (2014). Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates. Preprint. Available at arXiv:1402.5180.

[3] ANANDKUMAR, A., HSU, D. and KAKADE, S. M. (2012). A method of moments for mixture models and hidden Markov models. In *Conference on Learning Theory* 33–1.

[4] AUDDY, A. and YUAN, M. (2020). Perturbation bounds for orthogonally decomposable tensors and their applications in high dimensional data analysis. Preprint. Available at arXiv:2007.09024.

[5] BAO, Z., DING, X. and WANG, K. (2021). Singular vector and singular subspace distribution for the matrix denoising model. *Ann. Statist.* **49** 370–392. MR4206682 https://doi.org/10.1214/20-AOS1960

[6] BARAK, B. and MOITRA, A. (2016). Noisy tensor completion via the sum-of-squares hierarchy. In *Conference on Learning Theory* 417–445.

[7] BELKIN, M., RADEMACHER, L. and VOSS, J. (2018). Eigenvectors of orthogonally decomposable functions. *SIAM J. Comput.* **47** 547–615. MR3794351 https://doi.org/10.1137/17M1122013

[8] BEN AROUS, G., MEI, S., MONTANARI, A. and NICA, M. (2019). The landscape of the spiked tensor model. *Comm. Pure Appl. Math.* **72** 2282–2330. MR4011861 https://doi.org/10.1002/cpa.21861

[9] BI, X., QU, A. and SHEN, X. (2018). Multilayer tensor factorization with applications to recommender systems. *Ann. Statist.* **46** 3308–3333. MR3852653 https://doi.org/10.1214/17-AOS1659

MSC2020 subject classifications. 62H10, 62H25.
Key words and phrases. Asymptotic distribution, confidence region, tensor principal component analysis, tensor regression, statistical inference.

[10] BRENNAN, M. and BRESLER, G. (2020). Reducibility and statistical-computational gaps from secret leakage. Preprint. Available at arXiv:2005.08099.

[11] CAI, C., LI, G., POOR, H. V. and CHEN, Y. (2019). Nonconvex low-rank tensor completion from noisy data. In *Advances in Neural Information Processing Systems* 1863–1874.

[12] CAI, C., POOR, H. V. and CHEN, Y. (2020). Uncertainty quantification for nonconvex tensor completion: Confidence intervals, heteroscedasticity and optimality. Preprint. Available at arXiv:2006.08580.

[13] CAI, T. T., LIANG, T. and RAKHLIN, A. (2016). Geometric inference for general high-dimensional linear inverse problems. *Ann. Statist.* **44** 1536–1563. MR3519932 https://doi.org/10.1214/15-AOS1426

[14] CAI, T. T. and ZHANG, A. (2014). Sparse representation of a polytope and recovery in sparse signals and low-rank matrices. *IEEE Trans. Inf. Theory* **60** 122–132. MR3150915 https://doi.org/10.1109/TIT.2013.2288639

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

[16] CANDÈS, E. J., STROHMER, T. and VORONINSKI, V. (2013). PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming. *Comm. Pure Appl. Math.* **66** 1241–1274. MR3069958 https://doi.org/10.1002/cpa.21432

[17] CARPENTIER, A., EISERT, J., GROSS, D. and NICKL, R. (2019). Uncertainty quantification for matrix compressed sensing and quantum tomography problems. In *High Dimensional Probability VIII—the Oaxaca Volume. Progress in Probability* **74** 385–430. Birkhäuser/Springer, Cham. MR4181374 https://doi.org/10.1007/978-3-030-26391-1_18

[18] CHEN, H., RASKUTTI, G. and YUAN, M. (2019). Non-convex projected gradient descent for generalized low-rank tensor regression. *J. Mach. Learn. Res.* **20** Paper No. 5, 37. MR3911412

[19] CHEN, J. and SAAD, Y. (2008/09). On the tensor SVD and the optimal low rank orthogonal approximation of tensors. *SIAM J. Matrix Anal. Appl.* **30** 1709–1734. MR2486861 https://doi.org/10.1137/070711621

[20] CHEN, W.-K. (2019). Phase transition in the spiked random tensor with Rademacher prior. *Ann. Statist.* **47** 2734–2756. MR3988771 https://doi.org/10.1214/18-AOS1763

[21] 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

[22] CHI, E. C., GAINES, B. R., SUN, W. W., ZHOU, H. and YANG, J. (2020). Provable convex co-clustering of tensors. *J. Mach. Learn. Res.* **21** Paper No. 214, 58. MR4209500

[23] DE LATHAUWER, L., DE MOOR, B. and VANDEWALLE, J. (2000). On the best rank-1 and rank-$(R_1, R_2, \dots, R_N)$ approximation of higher-order tensors. *SIAM J. Matrix Anal. Appl.* **21** 1324–1342. MR1780276 https://doi.org/10.1137/S0895479898346995

[24] DE SILVA, V. and LIM, L.-H. (2008). Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM J. Matrix Anal. Appl.* **30** 1084–1127. MR2447444 https://doi.org/10.1137/06066518X

[25] DE DOMENICO, M., NICOSIA, V., ARENAS, A. and LATORA, V. (2015). Structural reducibility of multilayer networks. *Nat. Commun.* **6** 1–9.

[26] DUDEJA, R. and HSU, D. (2021). Statistical query lower bounds for tensor PCA. *J. Mach. Learn. Res.* **22** Paper No. 83, 51. MR4253776

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

[28] FAN, J., GONG, W. and ZHU, Z. (2019). Generalized high-dimensional trace regression via nuclear norm regularization. *J. Econometrics* **212** 177–202. MR3994013 https://doi.org/10.1016/j.jeconom.2019.04.026

[29] FEIZI, S., JAVADI, H. and TSE, D. (2017). Tensor biclustering. In *Advances in Neural Information Processing Systems* 1311–1320.

[30] FOUCART, S., NEEDELL, D., PLAN, Y. and WOOTTERS, M. (2017). De-biasing low-rank projection for matrix completion. In *Wavelets and Sparsity XVII* **10394** 1039417. International Society for Optics and Photonics. https://doi.org/10.1117/12.2275004.

[31] HAN, R., LUO, Y., WANG, M. and ZHANG, A. R. (2020). Exact clustering in tensor block model: Statistical optimality and computational limit. Preprint. Available at arXiv:2012.09996.

[32] HAN, R., WILLETT, R. and ZHANG, A. (2020). An optimal statistical and computational framework for generalized tensor estimation. Preprint. Available at arXiv:2002.11255.

[33] HAO, B., ZHANG, A. and CHENG, G. (2020). Sparse and low-rank tensor estimation via cubic sketchings. *IEEE Trans. Inf. Theory* **66** 5927–5964. MR4158653 https://doi.org/10.1109/TIT.2020.2982499

[34] HILLAR, C. J. and LIM, L.-H. (2013). Most tensor problems are NP-hard. *J. ACM* **60** Art. 45, 39. MR3144915 https://doi.org/10.1145/2512329

[35] HOPKINS, S. B., SHI, J. and STEURER, D. (2015). Tensor principal component analysis via sum-of-square proofs. In *Conference on Learning Theory* 956–1006.

[36] HUANG, J., HUANG, D. Z., YANG, Q. and CHENG, G. (2020). Power iteration for tensor PCA. Preprint. Available at arXiv:2012.13669.

[37] JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909. MR3277152

[38] JING, B.-Y., LI, T., LYU, Z. and XIA, D. (2021). Community detection on mixture multilayer networks via regularized tensor decomposition. *Ann. Statist.* **49** 3181–3205. MR4352527 https://doi.org/10.1214/21-aos2079

[39] KARATZOGLOU, A., AMATRIAIN, X., BALTRUNAS, L. and OLIVER, N. (2010). Multiverse recommendation: N-dimensional tensor factorization for context-aware collaborative filtering. In *Proceedings of the Fourth ACM Conference on Recommender Systems* 79–86.

[40] KE, Z. T., SHI, F. and XIA, D. (2019). Community detection for hypergraph networks via regularized tensor power iteration. Preprint. Available at arXiv:1909.06503.

[41] KOLDA, T. G. (2001). Orthogonal tensor decompositions. *SIAM J. Matrix Anal. Appl.* **23** 243–255. MR1856608 https://doi.org/10.1137/S0895479800368354

[42] KOLDA, T. G. and BADER, B. W. (2009). Tensor decompositions and applications. *SIAM Rev.* **51** 455–500. MR2535056 https://doi.org/10.1137/07070111X

[43] KOLTCHINSKII, V., LOUNICI, K. and TSYBAKOV, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.* **39** 2302–2329. MR2906869 https://doi.org/10.1214/11-AOS894

[44] KOLTCHINSKII, V. and XIA, D. (2015). Optimal estimation of low rank density matrices. *J. Mach. Learn. Res.* **16** 1757–1792. MR3417798

[45] LI, X., LING, S., STROHMER, T. and WEI, K. (2019). Rapid, robust, and reliable blind deconvolution via nonconvex optimization. *Appl. Comput. Harmon. Anal.* **47** 893–934. MR3994997 https://doi.org/10.1016/j.acha.2018.01.001

[46] LI, X., XU, D., ZHOU, H. and LI, L. (2018). Tucker tensor regression and neuroimaging analysis. *Stat. Biosci.* **10** 520–545.

[47] LIU, T., YUAN, M. and ZHAO, H. (2017). Characterizing spatiotemporal transcriptome of human brain via low rank tensor decomposition. Preprint. Available at arXiv:1702.07449.

[48] LUO, Y. and ZHANG, A. R. (2020). Tensor clustering with planted structures: Statistical optimality and computational limits. Preprint. Available at arXiv:2005.10743.

[49] LUO, Y. and ZHANG, A. R. (2020). Open problem: Average-case hardness of hypergraphic planted clique detection. In *Conference of Learning Theory* (*COLT*) **125** 3852–3856.

[50] LYNCH, R. E., RICE, J. R. and THOMAS, D. H. (1964). Tensor product analysis of partial difference equations. *Bull. Amer. Math. Soc.* **70** 378–384. MR0169390 https://doi.org/10.1090/S0002-9904-1964-11105-8

[51] MONTANARI, A. and SUN, N. (2018). Spectral algorithms for tensor completion. *Comm. Pure Appl. Math.* **71** 2381–2425. MR3862094 https://doi.org/10.1002/cpa.21748

[52] PERRY, A., WEIN, A. S. and BANDEIRA, A. S. (2020). Statistical limits of spiked tensor models. *Ann. Inst. Henri Poincaré Probab. Stat.* **56** 230–264. MR4058987 https://doi.org/10.1214/19-AIHP960

[53] RASKUTTI, G., YUAN, M. and CHEN, H. (2019). Convex regularization for high-dimensional multiresponse tensor regression. *Ann. Statist.* **47** 1554–1584. MR3911122 https://doi.org/10.1214/18-AOS1725

[54] RAUHUT, H., SCHNEIDER, R. and STOJANAC, Ž. (2017). Low rank tensor recovery via iterative hard thresholding. *Linear Algebra Appl.* **523** 220–262. MR3624675 https://doi.org/10.1016/j.laa.2017.02.028

[55] RICHARD, E. and MONTANARI, A. (2014). A statistical model for tensor PCA. In *Advances in Neural Information Processing Systems* 2897–2905.

[56] ROBEVA, E. (2016). Orthogonal decomposition of symmetric tensors. *SIAM J. Matrix Anal. Appl.* **37** 86–102. MR3530392 https://doi.org/10.1137/140989340

[57] SHAH, D. and YU, C. L. (2019). Iterative collaborative filtering for sparse noisy tensor estimation. In 2019 *IEEE International Symposium on Information Theory* (*ISIT*) 41–45. IEEE.

[58] SUN, W. W. and LI, L. (2017). STORE: Sparse tensor response regression and neuroimaging analysis. *J. Mach. Learn. Res.* **18** Paper No. 135, 37. MR3763769

[59] SUN, W. W. and LI, L. (2019). Dynamic tensor clustering. *J. Amer. Statist. Assoc.* **114** 1894–1907. MR4047308 https://doi.org/10.1080/01621459.2018.1527701

[60] SUN, W. W., LU, J., LIU, H. and CHENG, G. (2017). Provable sparse tensor decomposition. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 899–916. MR3641413 https://doi.org/10.1111/rssb.12190

[61] TOMIOKA, R. and SUZUKI, T. (2013). Convex tensor decomposition via structured Schatten norm regularization. In *Advances in Neural Information Processing Systems* 1331–1339.

[62] TUCKER, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika* **31** 279–311. MR0205395 https://doi.org/10.1007/BF02289464

[63] VANNIEUWENHOVEN, N., VANDEBRIL, R. and MEERBERGEN, K. (2012). A new truncation strategy for the higher-order singular value decomposition. *SIAM J. Sci. Comput.* **34** A1027–A1052. MR2914314 https://doi.org/10.1137/110836067

[64] VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285 https://doi.org/10.1214/14-AOS1221

[65] VON NEUMANN, J. (1936). The uniqueness of Haar's measure. *Mat. Sb.* **1** 721–734.

[66] WEIL, A. (1940). *L'intégration dans les Groupes Topologiques et Ses Applications*. *Actualités Scientifiques et Industrielles* [*Current Scientific and Industrial Topics*], *No*. 869. Hermann & Cie, Paris. MR0005741

[67] WILMOTH, J. R. and SHKOLNIKOV, V. (2006). Human mortality database.

[68] WU, T., BENSON, A. R. and GLEICH, D. F. (2016). General tensor spectral co-clustering for higher-order data. In *Advances in Neural Information Processing Systems* 2559–2567.

[69] XIA, D. (2019). Confidence region of singular subspaces for low-rank matrix regression. *IEEE Trans. Inf. Theory* **65** 7437–7459. MR4030894 https://doi.org/10.1109/TIT.2019.2924900

[70] XIA, D. (2021). Normal approximation and confidence region of singular subspaces. *Electron. J. Stat.* **15** 3798–3851. MR4298986 https://doi.org/10.1214/21-ejs1876

[71] XIA, D. and YUAN, M. (2019). On polynomial time methods for exact low-rank tensor completion. *Found. Comput. Math.* **19** 1265–1313. MR4029842 https://doi.org/10.1007/s10208-018-09408-6

[72] XIA, D. and YUAN, M. (2021). Statistical inferences of linear forms for noisy matrix completion. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **83** 58–77. MR4220984 https://doi.org/10.1111/rssb.12400

[73] XIA, D., ZHANG, A. R. and ZHOU, Y. (2022). Supplement to "Inference for Low-rank Tensors—No Need to Debias." https://doi.org/10.1214/21-AOS2146SUPP

[74] YUAN, M. and ZHANG, C.-H. (2016). On tensor completion via nuclear norm minimization. *Found. Comput. Math.* **16** 1031–1068. MR3529132 https://doi.org/10.1007/s10208-015-9269-5

[75] ZHANG, A. (2019). Cross: Efficient low-rank tensor completion. *Ann. Statist.* **47** 936–964. MR3909956 https://doi.org/10.1214/18-AOS1694

[76] ZHANG, A. and HAN, R. (2019). Optimal sparse singular value decomposition for high-dimensional high-order data. *J. Amer. Statist. Assoc.* **114** 1708–1725. MR4047294 https://doi.org/10.1080/01621459.2018.1527227

[77] ZHANG, A. and XIA, D. (2018). Tensor SVD: Statistical and computational limits. *IEEE Trans. Inf. Theory* **64** 7311–7338. MR3876445 https://doi.org/10.1109/TIT.2018.2841377

[78] ZHANG, A. R., LUO, Y., RASKUTTI, G. and YUAN, M. (2020). ISLET: Fast and optimal low-rank tensor regression via importance sketching. *SIAM J. Math. Data Sci.* **2** 444–479. MR4106613 https://doi.org/10.1137/19M126476X

[79] ZHANG, C., HAN, R., ZHANG, A. R. and VOYLES, P. M. (2020). Denoising atomic resolution 4D scanning transmission electron microscopy data with tensor singular value decomposition. *Ultramicroscopy* 113123.

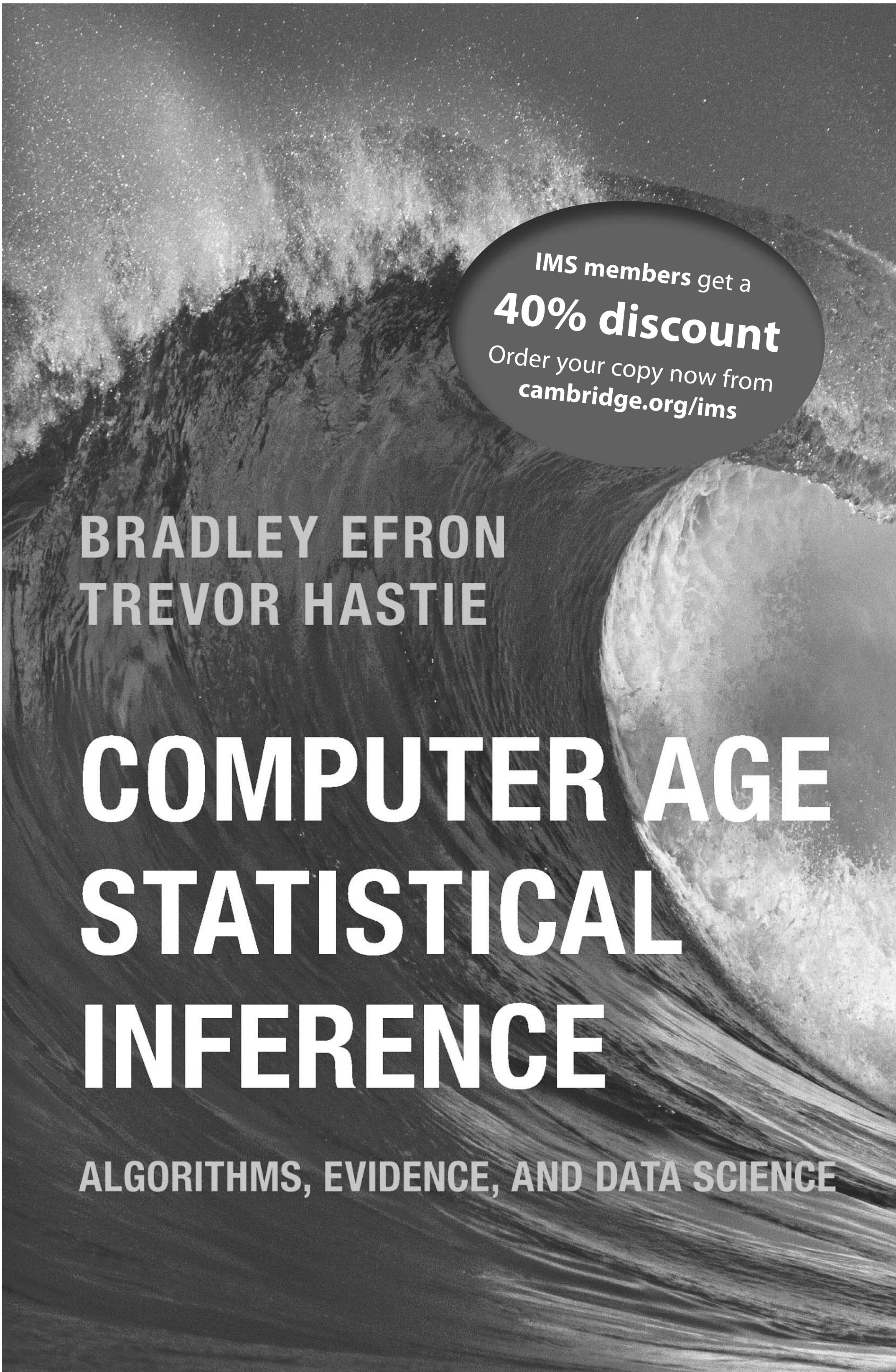
[80] ZHANG, C.-H. and HUANG, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. MR2435448 https://doi.org/10.1214/07-AOS520

[81] ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940 https://doi.org/10.1111/rssb.12026

[82] ZHANG, T. and GOLUB, G. H. (2001). Rank-one approximation to high order tensors. *SIAM J. Matrix Anal. Appl.* **23** 534–550. MR1871328 https://doi.org/10.1137/S0895479899352045

[83] ZHENG, Q. and TOMIOKA, R. (2015). Interpolating convex and non-convex tensor decompositions via the subspace norm. In *Advances in Neural Information Processing Systems* 3106–3113.

[84] ZHOU, H., LI, L. and ZHU, H. (2013). Tensor regression with applications in neuroimaging data analysis. *J. Amer. Statist. Assoc.* **108** 540–552. MR3174640 https://doi.org/10.1080/01621459.2013.776499


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