

THE ANNALS *of* STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles

- The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics . . . JOSHUA CAPE, MINH TANG AND CAREY E. PRIEBE 2405
- Isotonic regression in general dimensions QIYANG HAN, TENG YAO WANG, SABYASACHI CHATTERJEE AND RICHARD J. SAMWORTH 2440
- Property testing in high-dimensional Ising models MATEY NEYKOV AND HAN LIU 2472
- A knockoff filter for high-dimensional selective inference
RINA FOYGEL BARBER AND EMMANUEL J. CANDÈS 2504
- Semi-supervised inference: General theory and estimation of means
ANRU ZHANG, LAWRENCE D. BROWN AND T. TONY CAI 2538
- Doubly penalized estimation in additive regression with high-dimensional data
ZHIQIANG TAN AND CUN-HUI ZHANG 2567
- Semiparametrically point-optimal hybrid rank tests for unit roots
BO ZHOU, RAMON VAN DEN AKKER AND BAS J. M. WERKER 2601
- The middle-scale asymptotics of Wishart matrices
DIDIER CHÉTELAT AND MARTIN T. WELLS 2639
- Linear hypothesis testing for high dimensional generalized linear models
CHENGCHUN SHI, RUI SONG, ZHAO CHEN AND RUNZE LI 2671
- An operator theoretic approach to nonparametric mixture models
ROBERT A. VANDERMEULEN AND CLAYTON D. SCOTT 2704
- Phase transition in the spiked random tensor with Rademacher prior WEI-KUO CHEN 2734
- Distance multivariate: New dependence measures for random vectors
BJÖRN BÖTTCHER, MARTIN KELLER-RESSEL AND RENÉ L. SCHILLING 2757
- A unified treatment of multiple testing with prior knowledge using the p-filter
AADITYA K. RAMDAS, RINA F. BARBER, MARTIN J. WAINWRIGHT
AND MICHAEL I. JORDAN 2790
- Exact lower bounds for the agnostic probably-approximately-correct (PAC) machine learning model ARYEH KONTOROVICH AND IOSIF PINELIS 2822
- Eigenvalue distributions of variance components estimators in high-dimensional random effects models ZHOU FAN AND IAIN M. JOHNSTONE 2855
- Test for high-dimensional correlation matrices
SHURONG ZHENG, GUANGHUI CHENG, JIANHUA GUO AND HONGTU ZHU 2887
- Projected spline estimation of the nonparametric function in high-dimensional partially linear models for massive data . . . HENG LIAN, KAIFENG ZHAO AND SHAOGAO LV 2922
- Inference for the mode of a log-concave density
CHARLES R. DOSS AND JON A. WELLNER 2950
- Testing for independence of large dimensional vectors
TARAS BODNAR, HOLGER DETTE AND NESTOR PAROLYA 2977

THE ANNALS OF STATISTICS

Vol. 47, No. 5, pp. 2405–3008 October 2019

INSTITUTE OF MATHEMATICAL STATISTICS

(Organized September 12, 1935)

The purpose of the Institute is to foster the development and dissemination of the theory and applications of statistics and probability.

IMS OFFICERS

President: Susan Murphy, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138-2901, USA

President-Elect: Regina Y. Liu, Department of Statistics, Rutgers University, Piscataway, New Jersey 08854-8019, USA

Past President: Xiao-Li Meng, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138-2901, USA

Executive Secretary: Edsel Peña, Department of Statistics, University of South Carolina, Columbia, South Carolina 29208-001, USA

Treasurer: Zhengjun Zhang, Department of Statistics, University of Wisconsin, Madison, Wisconsin 53706-1510, USA

Program Secretary: Ming Yuan, Department of Statistics, Columbia University, New York, NY 10027-5927, USA

IMS EDITORS

The Annals of Statistics. *Editors:* Richard J. Samworth, Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, Cambridge, CB3 0WB, UK. Ming Yuan, Department of Statistics, Columbia University, New York, NY 10027, USA

The Annals of Applied Statistics. *Editor-in-Chief:* Karen Kafadar, Department of Statistics, University of Virginia, Heidelberg Institute for Theoretical Studies, Charlottesville, VA 22904-4135, USA

The Annals of Probability. *Editor:* Amir Dembo, Department of Statistics and Department of Mathematics, Stanford University, Stanford, California 94305, USA

The Annals of Applied Probability. *Editors:* François Delarue, Laboratoire J. A. Dieudonné, Université de Nice Sophia-Antipolis, France-06108 Nice Cedex 2. Peter Friz, Institut für Mathematik, Technische Universität Berlin, 10623 Berlin, Germany and Weierstrass-Institut für Angewandte Analysis und Stochastik, 10117 Berlin, Germany

Statistical Science. *Editor:* Cun-Hui Zhang, Department of Statistics, Rutgers University, Piscataway, New Jersey 08854, USA

The IMS Bulletin. *Editor:* Vlada Limic, UMR 7501 de l'Université de Strasbourg et du CNRS, 7 rue René Descartes, 67084 Strasbourg Cedex, France

The Annals of Statistics [ISSN 0090-5364 (print); ISSN 2168-8966 (online)], Volume 47, Number 5, October 2019. Published bimonthly by the Institute of Mathematical Statistics, 3163 Somerset Drive, Cleveland, Ohio 44122, 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, 9650 Rockville Pike, Suite L 2310, Bethesda, Maryland 20814-3998, USA.

THE TWO-TO-INFINITY NORM AND SINGULAR SUBSPACE GEOMETRY WITH APPLICATIONS TO HIGH-DIMENSIONAL STATISTICS¹

BY JOSHUA CAPE, MINH TANG AND CAREY E. PRIEBE

Johns Hopkins University

The singular value matrix decomposition plays a ubiquitous role throughout statistics and related fields. Myriad applications including clustering, classification, and dimensionality reduction involve studying and exploiting the geometric structure of singular values and singular vectors.

This paper provides a novel collection of technical and theoretical tools for studying the geometry of singular subspaces using the two-to-infinity norm. Motivated by preliminary deterministic Procrustes analysis, we consider a general matrix perturbation setting in which we derive a new Procrustean matrix decomposition. Together with flexible machinery developed for the two-to-infinity norm, this allows us to conduct a refined analysis of the induced perturbation geometry with respect to the underlying singular vectors even in the presence of singular value multiplicity. Our analysis yields singular vector entrywise perturbation bounds for a range of popular matrix noise models, each of which has a meaningful associated statistical inference task. In addition, we demonstrate how the two-to-infinity norm is the preferred norm in certain statistical settings. Specific applications discussed in this paper include covariance estimation, singular subspace recovery, and multiple graph inference.

Both our Procrustean matrix decomposition and the technical machinery developed for the two-to-infinity norm may be of independent interest.

REFERENCES

- [1] ABBE, E., FAN, J., WANG, K. and ZHONG, Y. (2017). Entrywise eigenvector analysis of random matrices with low expected rank. Preprint. Available at [arXiv:1709.09565](https://arxiv.org/abs/1709.09565).
- [2] BAI, Z. and SILVERSTEIN, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed. *Springer Series in Statistics*. Springer, New York. [MR2567175](#)
- [3] 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](#)
- [4] BENIDIS, K., SUN, Y., BABU, P. and PALOMAR, D. P. (2016). Orthogonal sparse eigenvectors: A Procrustes problem. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 4683–4686.
- [5] BHATIA, R. (1997). *Matrix Analysis*. *Graduate Texts in Mathematics* **169**. Springer, New York. [MR1477662](#)

Primary 62H12, 62H25; secondary 62H30.

Key words and phrases. Singular value decomposition, principal component analysis, eigenvector perturbation, spectral methods, Procrustes analysis, high-dimensional statistics.

- [6] BOJANCZYK, A. W. and LUTOBORSKI, A. (1999). The Procrustes problem for orthogonal Stiefel matrices. *SIAM J. Sci. Comput.* **21** 1291–1304. [MR1740396](#)
- [7] CAI, T. T., MA, Z. and WU, Y. (2013). Sparse PCA: Optimal rates and adaptive estimation. *Ann. Statist.* **41** 3074–3110. [MR3161458](#)
- [8] CAI, T. T. and ZHANG, A. (2018). Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional statistics. *Ann. Statist.* **46** 60–89. [MR3766946](#)
- [9] CANDÈS, E. J. and RECHT, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.* **9** 717–772. [MR2565240](#)
- [10] CAPE, J., TANG, M. and PRIEBE, C. E. (2018). Signal-plus-noise matrix models: Eigenvector deviations and fluctuations. *Biometrika*. To appear. Available at [arXiv:1802.00381](#).
- [11] CHEN, L., VOGELSTEIN, J. T., LYZINSKI, V. and PRIEBE, C. E. (2016). A joint graph inference case study: The *C. elegans* chemical and electrical connectome. *Worm* **5** 1–8.
- [12] CHIKUSE, Y. (2003). *Statistics on Special Manifolds. Lecture Notes in Statistics* **174**. Springer, New York. [MR1960435](#)
- [13] DAVIS, C. and KAHAN, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.* **7** 1–46. [MR0264450](#)
- [14] DRYDEN, I. L., KOLOYDENKO, A. and ZHOU, D. (2009). Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *Ann. Appl. Stat.* **3** 1102–1123. [MR2750388](#)
- [15] DRYDEN, I. L. and MARDIA, K. V. (2016). *Statistical Shape Analysis with Applications in R*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley, Chichester. [MR3559734](#)
- [16] EDELMAN, A., ARIAS, T. A. and SMITH, S. T. (1999). The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.* **20** 303–353. [MR1646856](#)
- [17] ELДАР, Y. C. and KUTYNIOK, G. (2012). *Compressed Sensing: Theory and Applications*. Cambridge Univ. Press, Cambridge. [MR2961961](#)
- [18] ELDRIDGE, J., BELKIN, M. and WANG, Y. (2018). Unperturbed: Spectral analysis beyond Davis-Kahan. In: *Proceedings of Algorithmic Learning Theory. Proceedings of Machine Learning Research* **83**, PMLR 321–358.
- [19] 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](#)
- [20] FAN, J., RIGOLLET, P. and WANG, W. (2015). Estimation of functionals of sparse covariance matrices. *Ann. Statist.* **43** 2706–2737. [MR3405609](#)
- [21] FAN, J., WANG, W. and ZHONG, Y. (2017). An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation. *J. Mach. Learn. Res.* **18** 1–42. [MR3827095](#)
- [22] FISHKIND, D. E., SUSSMAN, D. L., TANG, M., VOGELSTEIN, J. T. and PRIEBE, C. E. (2013). Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. *SIAM J. Matrix Anal. Appl.* **34** 23–39. [MR3032990](#)
- [23] GOWER, J. C. and DIJKSTERHUIS, G. B. (2004). *Procrustes Problems. Oxford Statistical Science Series* **30**. Oxford Univ. Press, Oxford. [MR2051013](#)
- [24] HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. [MR0718088](#)
- [25] HORN, R. A. and JOHNSON, C. R. (2012). *Matrix Analysis*. Cambridge Univ. Press, Cambridge. [MR2978290](#)
- [26] JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. [MR1863961](#)
- [27] JOLLIFFE, I. T. (1986). *Principal Component Analysis. Springer Series in Statistics*. Springer, New York. [MR0841268](#)
- [28] KOLTCHINSKII, V. and LOUNICI, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli* **23** 110–133. [MR3556768](#)

- [29] KOLTCHINSKII, V. and LOUNICI, K. (2017). New asymptotic results in principal component analysis. *Sankhyā A* **79** 254–297. [MR3707422](#)
- [30] LE, C. M., LEVINA, E. and VERSHYNIN, R. (2016). Optimization via low-rank approximation for community detection in networks. *Ann. Statist.* **44** 373–400. [MR3449772](#)
- [31] LEI, J. and RINALDO, A. (2015). Consistency of spectral clustering in stochastic block models. *Ann. Statist.* **43** 215–237. [MR3285605](#)
- [32] LEVIN, K., ATHREYA, A., TANG, M., LYZINSKI, V. and PRIEBE, C. E. (2017). A central limit theorem for an omnibus embedding of random dot product graphs. Preprint. Available at [arXiv:1705.09355](#).
- [33] LU, L. and PENG, X. (2013). Spectra of edge-independent random graphs. *Electron. J. Combin.* **20** 1–18. [MR3158266](#)
- [34] LYZINSKI, V. (2018). Information recovery in shuffled graphs via graph matching. *IEEE Trans. Inform. Theory* **64** 3254–3273. [MR3798375](#)
- [35] LYZINSKI, V., PARK, Y., PRIEBE, C. E. and TROSSET, M. (2017). Fast embedding for JOFC using the raw stress criterion. *J. Comput. Graph. Statist.* **26** 786–802. [MR3765344](#)
- [36] LYZINSKI, V., SUSSMAN, D. L., FISHKIND, D. E., PAO, H., CHEN, L., VOGELSTEIN, J. T., PARK, Y. and PRIEBE, C. E. (2015). Spectral clustering for divide-and-conquer graph matching. *Parallel Comput.* **47** 70–87. [MR3371642](#)
- [37] LYZINSKI, V., SUSSMAN, D. L., TANG, M., ATHREYA, A. and PRIEBE, C. E. (2014). Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. *Electron. J. Stat.* **8** 2905–2922. [MR3299126](#)
- [38] MAO, X., SARKAR, P. and CHAKRABARTI, D. (2017). Estimating mixed memberships with sharp eigenvector deviations. Preprint. Available at [arXiv:1709.00407](#).
- [39] NADLER, B. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach. *Ann. Statist.* **36** 2791–2817. [MR2485013](#)
- [40] O’ROURKE, S., VU, V. and WANG, K. (2016). Eigenvectors of random matrices: A survey. *J. Combin. Theory Ser. A* **144** 361–442. [MR3534074](#)
- [41] O’ROURKE, S., VU, V. and WANG, K. (2018). Random perturbation of low rank matrices: Improving classical bounds. *Linear Algebra Appl.* **540** 26–59. [MR3739989](#)
- [42] PAUL, D. and AUE, A. (2014). Random matrix theory in statistics: A review. *J. Statist. Plann. Inference* **150** 1–29. [MR3206718](#)
- [43] PRIEBE, C. E., MARCHETTE, D. J., MA, Z. and ADALI, S. (2013). Manifold matching: Joint optimization of fidelity and commensurability. *Braz. J. Probab. Stat.* **27** 377–400. [MR3064729](#)
- [44] REBROVA, E. and VERSHYNIN, R. (2018). Norms of random matrices: Local and global problems. *Adv. Math.* **324** 40–83. [MR3733881](#)
- [45] ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915. [MR2893856](#)
- [46] RUDELSON, M. and VERSHYNIN, R. (2015). Delocalization of eigenvectors of random matrices with independent entries. *Duke Math. J.* **164** 2507–2538. [MR3405592](#)
- [47] SARKAR, P. and BICKEL, P. J. (2015). Role of normalization in spectral clustering for stochastic blockmodels. *Ann. Statist.* **43** 962–990. [MR3346694](#)
- [48] STEWART, G. W. and SUN, J. G. (1990). *Matrix Perturbation Theory*. Computer Science and Scientific Computing. Academic Press, Boston, MA. [MR1061154](#)
- [49] SUSSMAN, D. L., TANG, M., FISHKIND, D. E. and PRIEBE, C. E. (2012). A consistent adjacency spectral embedding for stochastic blockmodel graphs. *J. Amer. Statist. Assoc.* **107** 1119–1128. [MR3010899](#)
- [50] SUSSMAN, D. L., TANG, M. and PRIEBE, C. E. (2014). Consistent latent position estimation and vertex classification for random dot product graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* **36** 48–57.

- [51] 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](#)
- [52] TANG, M., ATHREYA, A., SUSSMAN, D. L., LYZINSKI, V. and PRIEBE, C. E. (2017). A non-parametric two-sample hypothesis testing problem for random graphs. *Bernoulli* **23** 1599–1630. [MR3624872](#)
- [53] TANG, M., CAPE, J. and PRIEBE, C. E. (2017). Asymptotically efficient estimators for stochastic blockmodels: The naive MLE, the rank-constrained MLE, and the spectral. Preprint. Available at [arXiv:1710.10936](#).
- [54] TANG, M. and PRIEBE, C. E. (2018). Limit theorems for eigenvectors of the normalized Laplacian for random graphs. *Ann. Statist.* **46** 2360–2415. [MR3845021](#)
- [55] VERSHYNIN, R. (2018). High-dimensional probability: An introduction with applications in data science. Online version 2018-02-09.
- [56] VON LUXBURG, U. (2007). A tutorial on spectral clustering. *Stat. Comput.* **17** 395–416. [MR2409803](#)
- [57] WEDIN, P. (1972). Perturbation bounds in connection with singular value decomposition. *BIT* **12** 99–111. [MR0309968](#)
- [58] WEYL, H. (1912). Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung). *Math. Ann.* **71** 441–479. [MR1511670](#)
- [59] YAO, J., ZHENG, S. and BAI, Z. (2015). *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. *Cambridge Series in Statistical and Probabilistic Mathematics* **39**. Cambridge Univ. Press, New York. [MR3468554](#)
- [60] YOUNG, S. J. and SCHEINERMAN, E. R. (2007). Random dot product graph models for social networks. In *Algorithms and Models for the Web-Graph*. *Lecture Notes in Computer Science* **4863** 138–149. Springer, Berlin. [MR2504912](#)
- [61] YU, Y., WANG, T. and SAMWORTH, R. J. (2015). A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika* **102** 315–323. [MR3371006](#)

ISOTONIC REGRESSION IN GENERAL DIMENSIONS¹

BY QIYANG HAN^{*,2}, TENG YAO WANG^{†,3}, SABYASACHI CHATTERJEE^{‡,§} AND
 RICHARD J. SAMWORTH^{†,3}

University of Washington^{*}, *University of Cambridge*[†], *University of Chicago*[‡]
 and *University of Illinois at Urbana-Champaign*[§]

We study the least squares regression function estimator over the class of real-valued functions on $[0, 1]^d$ that are increasing in each coordinate. For uniformly bounded signals and with a fixed, cubic lattice design, we establish that the estimator achieves the minimax rate of order $n^{-\min\{2/(d+2), 1/d\}}$ in the empirical L_2 loss, up to polylogarithmic factors. Further, we prove a sharp oracle inequality, which reveals in particular that when the true regression function is piecewise constant on k hyperrectangles, the least squares estimator enjoys a faster, adaptive rate of convergence of $(k/n)^{\min(1, 2/d)}$, again up to polylogarithmic factors. Previous results are confined to the case $d \leq 2$. Finally, we establish corresponding bounds (which are new even in the case $d = 2$) in the more challenging random design setting. There are two surprising features of these results: first, they demonstrate that it is possible for a global empirical risk minimisation procedure to be rate optimal up to polylogarithmic factors even when the corresponding entropy integral for the function class diverges rapidly; second, they indicate that the adaptation rate for shape-constrained estimators can be strictly worse than the parametric rate.

REFERENCES

- AMELUNXEN, D., LOTZ, M., MCCOY, M. B. and TROPP, J. A. (2014). Living on the edge: Phase transitions in convex programs with random data. *Inf. Inference* **3** 224–294. [MR3311453](#)
- BACCHETTI, P. (1989). Additive isotonic models. *J. Amer. Statist. Assoc.* **84** 289–294. [MR0999691](#)
- BARLOW, R. E., BARTHOLOMEW, D. J., BREMNER, J. M. and BRUNK, H. D. (1972). *Statistical Inference Under Order Restrictions*. Wiley, New York.
- BELLEÇ, P. C. (2018). Sharp oracle inequalities for least squares estimators in shape restricted regression. *Ann. Statist.* **46** 745–780. [MR3782383](#)
- BIRGÉ, L. and MASSART, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields* **97** 113–150. [MR1240719](#)
- BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford Univ. Press, Oxford. [MR3185193](#)
- BRUNK, H. D. (1955). Maximum likelihood estimates of monotone parameters. *Ann. Math. Stat.* **26** 607–616. [MR0073894](#)
- CHATTERJEE, S. (2014). A new perspective on least squares under convex constraint. *Ann. Statist.* **42** 2340–2381. [MR3269982](#)

MSC2010 subject classifications. 62G05, 62G08, 62C20.

Key words and phrases. Isotonic regression, block increasing functions, adaptation, least squares, sharp oracle inequality, statistical dimension.

- CHATTERJEE, S., GUNTUBOYINA, A. and SEN, B. (2015). On risk bounds in isotonic and other shape restricted regression problems. *Ann. Statist.* **43** 1774–1800. [MR3357878](#)
- CHATTERJEE, S., GUNTUBOYINA, A. and SEN, B. (2018). On matrix estimation under monotonicity constraints. *Bernoulli* **24** 1072–1100. [MR3706788](#)
- CHATTERJEE, S. and LAFFERTY, J. (2017). Adaptive risk bounds in unimodal regression. Preprint. Available at [arxiv:1512.02956v5](#).
- CHEN, Y. and SAMWORTH, R. J. (2016). Generalized additive and index models with shape constraints. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 729–754. [MR3534348](#)
- DONOHU, D. (1991). Gelfand n -widths and the method of least squares. Technical report, Univ. California, Berkeley, Berkeley, CA.
- DUROT, C. (2007). On the $\mathbb{L}_p$ -error of monotonicity constrained estimators. *Ann. Statist.* **35** 1080–1104. [MR2341699](#)
- DUROT, C. (2008). Monotone nonparametric regression with random design. *Math. Methods Statist.* **17** 327–341. [MR2483461](#)
- DYKSTRA, R. L. (1983). An algorithm for restricted least squares regression. *J. Amer. Statist. Assoc.* **78** 837–842. [MR0727568](#)
- DYKSTRA, R. L. and ROBERTSON, T. (1982). An algorithm for isotonic regression for two or more independent variables. *Ann. Statist.* **10** 708–716. [MR0663427](#)
- EICHLER, E. E., FLINT, J., GIBSON, G., KONG, A., LEAL, S. M., MOORE, J. H. and NADEAU, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* **11** 446–450.
- ELENA, S. F. and LENSKI, R. E. (1997). Test of synergistic interactions among deleterious mutations in bacteria. *Nature* **390** 395–398.
- GAO, F. and WELLNER, J. A. (2007). Entropy estimate for high-dimensional monotonic functions. *J. Multivariate Anal.* **98** 1751–1764. [MR2392431](#)
- GOLDSTEIN, D. B. (2009). Common genetic variation and human traits. *N. Engl. J. Med.* **360** 1696–1698.
- GROENEBOOM, P. and JONGBLOED, G. (2014). *Nonparametric Estimation Under Shape Constraints: Estimators, Algorithms and Asymptotics*. Cambridge Series in Statistical and Probabilistic Mathematics **38**. Cambridge Univ. Press, New York. [MR3445293](#)
- GUNTUBOYINA, A. and SEN, B. (2015). Global risk bounds and adaptation in univariate convex regression. *Probab. Theory Related Fields* **163** 379–411. [MR3405621](#)
- HAN, Q., WANG, T., CHATTERJEE, S. and SAMWORTH, R. J. (2019). Supplement to “Isotonic regression in general dimensions.” DOI:[10.1214/18-AOS1753SUPP](#).
- KIM, A. K. H., GUNTUBOYINA, A. and SAMWORTH, R. J. (2018). Adaptation in log-concave density estimation. *Ann. Statist.* **46** 2279–2306. [MR3845018](#)
- KIM, A. K. H. and SAMWORTH, R. J. (2016). Global rates of convergence in log-concave density estimation. *Ann. Statist.* **44** 2756–2779. [MR3576560](#)
- KYNG, R., RAO, A. and SACHDEVA, S. (2015). Fast, provable algorithms for isotonic regression in all ℓ_p -norms. In *Advances in Neural Information Processing Systems* 2719–2727.
- LUSS, R., ROSSET, S. and SHAHAR, M. (2012). Efficient regularized isotonic regression with application to gene–gene interaction search. *Ann. Appl. Stat.* **6** 253–283. [MR2951537](#)
- MAMMEN, E. and YU, K. (2007). Additive isotone regression. In *Asymptotics: Particles, Processes and Inverse Problems* (E. A. Cator et al., eds.). Institute of Mathematical Statistics Lecture Notes—Monograph Series **55** 179–195. IMS, Beachwood, OH. [MR2459939](#)
- MANI, R., ONGE, R. P. S., HARTMAN, J. L., GIAEVER, G. and ROTH, F. P. (2008). Defining genetic interaction. *Proc. Natl. Acad. Sci. USA* **105** 3461–3466.
- MASSART, P. (2000). About the constants in Talagrand’s concentration inequalities for empirical processes. *Ann. Probab.* **28** 863–884. [MR1782276](#)
- MEYER, M. and WOODROOFE, M. (2000). On the degrees of freedom in shape-restricted regression. *Ann. Statist.* **28** 1083–1104. [MR1810920](#)

- MORTON-JONES, T., DIGGLE, P., PARKER, L., DICKINSON, H. O. and BINKS, K. (2000). Additive isotonic regression models in epidemiology. *Stat. Med.* **19** 849–859.
- PISIER, G. (1989). *The Volume of Convex Bodies and Banach Space Geometry*. *Cambridge Tracts in Mathematics* **94**. Cambridge Univ. Press, Cambridge. [MR1036275](#)
- POLLARD, D. (2002). *A User's Guide to Measure Theoretic Probability*. *Cambridge Series in Statistical and Probabilistic Mathematics* **8**. Cambridge Univ. Press, Cambridge. [MR1873379](#)
- RAKHLIN, A., SRIDHARAN, K. and TSYBAKOV, A. B. (2017). Empirical entropy, minimax regret and minimax risk. *Bernoulli* **23** 789–824. [MR3606751](#)
- ROMIK, D. (2015). *The Surprising Mathematics of Longest Increasing Subsequences*. *Institute of Mathematical Statistics Textbooks* **4**. Cambridge Univ. Press, New York. [MR3468738](#)
- ROTH, F. P., LIPSHITZ, H. D. and ANDREWS, B. J. (2009). Q&A: Epistasis. *J. Biol.* **8** 35.
- SANJUÁN, R. and ELENA, S. F. (2006). Epistasis correlates to genomic complexity. *Proc. Natl. Acad. Sci. USA* **103** 14402–14405.
- SCHELL, M. J. and SINGH, B. (1997). The reduced monotonic regression method. *J. Amer. Statist. Assoc.* **92** 128–135.
- SHAO, H., BURRAGE, L. C., SINASAC, D. S., HILL, A. E., ERNEST, S. R., O'BRIEN, W., COURTLAND, H.-W., JEPSEN, K. J., KIRBY, A., KULBOKAS, E. J., DALY, M. J., BROMAN, K. W., LANDER, E. S. and NADEAU, J. H. (2008). Genetic architecture of complex traits: Large phenotypic effects and pervasive epistasis. *Proc. Natl. Acad. Sci. USA* **105** 19910–19914.
- STOUT, Q. F. (2015). Isotonic regression for multiple independent variables. *Algorithmica* **71** 450–470. [MR3331888](#)
- TALAGRAND, M. (1996). New concentration inequalities in product spaces. *Invent. Math.* **126** 505–563. [MR1419006](#)
- TONG, A. H., EVANGELISTA, M., PARSONS, A. B., XU, H., BADER, G. D., PAGÉ, N., ROBINSON, M., RAGHIBIZADEH, S., HOGUE, C. W. V., BUSSEY, H., ANDREWS, B., TYERS, M. and BOONE, C. (2001). Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294** 2364–2368.
- VAN DE GEER, S. (1990). Estimating a regression function. *Ann. Statist.* **18** 907–924. [MR1056343](#)
- VAN DE GEER, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.* **21** 14–44. [MR1212164](#)
- VAN DE GEER, S. A. (2000). *Applications of Empirical Process Theory*. *Cambridge Series in Statistical and Probabilistic Mathematics* **6**. Cambridge Univ. Press, Cambridge. [MR1739079](#)
- VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York. [MR1385671](#)
- VAN EEDEN, C. (1958). *Testing and Estimating Ordered Parameters of Probability Distributions*. Mathematical Centre, Amsterdam. [MR0102874](#)
- YANG, F. and BARBER, R. F. (2017). Uniform convergence of isotonic regression. Preprint. Available at [arxiv:1706.01852](#).
- YANG, Y. and BARRON, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.* **27** 1564–1599. [MR1742500](#)
- YU, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam* (D. Pollard, E. Torgersen and G. L. Yang, eds.) 423–435. Springer, New York. [MR1462963](#)
- ZHANG, C.-H. (2002). Risk bounds in isotonic regression. *Ann. Statist.* **30** 528–555. [MR1902898](#)

PROPERTY TESTING IN HIGH-DIMENSIONAL ISING MODELS

BY MATEY NEYKOV AND HAN LIU¹

Carnegie Mellon University and Northwestern University

This paper explores the information-theoretic limitations of graph property testing in zero-field Ising models. Instead of learning the entire graph structure, sometimes testing a basic graph property such as connectivity, cycle presence or maximum clique size is a more relevant and attainable objective. Since property testing is more fundamental than graph recovery, any necessary conditions for property testing imply corresponding conditions for graph recovery, while custom property tests can be statistically and/or computationally more efficient than graph recovery based algorithms. Understanding the statistical complexity of property testing requires the distinction of *ferromagnetic* (i.e., positive interactions only) and general Ising models. Using combinatorial constructs such as graph packing and strong monotonicity, we characterize how target properties affect the corresponding minimax upper and lower bounds within the realm of ferromagnets. On the other hand, by studying the detection of an *antiferromagnetic* (i.e., negative interactions only) Curie–Weiss model buried in Rademacher noise, we show that property testing is strictly more challenging over general Ising models. In terms of methodological development, we propose two types of correlation based tests: computationally efficient screening for ferromagnets, and score type tests for general models, including a fast cycle presence test. Our correlation screening tests match the information-theoretic bounds for property testing in ferromagnets in certain regimes.

REFERENCES

- ADDARIO-BERRY, L., BROUTIN, N., DEVROYE, L. and LUGOSI, G. (2010). On combinatorial testing problems. *Ann. Statist.* **38** 3063–3092. [MR2722464](#)
- AHMED, A. and XING, E. P. (2009). Estimating time-varying networks. In *Proc. Natl. Acad. Sci. USA* **106** 11878–11883.
- ANANDKUMAR, A., TAN, V. Y. F., HUANG, F. and WILLSKY, A. S. (2012). High-dimensional structure estimation in Ising models: Local separation criterion. *Ann. Statist.* **40** 1346–1375. [MR3015028](#)
- ARIAS-CASTRO, E., BUBECK, S. and LUGOSI, G. (2012). Detection of correlations. *Ann. Statist.* **40** 412–435. [MR3014312](#)
- ARIAS-CASTRO, E., BUBECK, S. and LUGOSI, G. (2015). Detecting positive correlations in a multivariate sample. *Bernoulli* **21** 209–241. [MR3322317](#)
- ARIAS-CASTRO, E., BUBECK, S., LUGOSI, G. and VERZELEN, N. (2018). Detecting Markov random fields hidden in white noise. *Bernoulli* **24** 3628–3656. [MR3788184](#)

MSC2010 subject classifications. 62F03, 62H15.

Key words and phrases. Ising models, minimax testing, Dobrushin’s comparison theorem, anti-ferromagnetic Curie–Weiss detection, two-point function bounds, correlation tests.

- BERTHET, Q., RIGOLLET, P. and SRIVASTAVA, P. (2016). Exact recovery in the Ising blockmodel. Preprint. Available at [arXiv:1612.03880](https://arxiv.org/abs/1612.03880).
- BHATTACHARYA, B. B. and MUKHERJEE, S. (2018). Inference in Ising models. *Bernoulli* **24** 493–525. [MR3706767](#)
- BOLLOBÁS, B. (2004). *Extremal Graph Theory*. Dover Publications, Inc., Mineola, NY. [MR2078877](#)
- BRESLER, G. (2015). Efficiently learning Ising models on arbitrary graphs. In *STOC'15—Proceedings of the 2015 ACM Symposium on Theory of Computing* 771–782. ACM, New York. [MR3388257](#)
- BRESLER, G., GAMARNIK, D. and SHAH, D. (2014). Structure learning of antiferromagnetic Ising models. In *Advances in Neural Information Processing Systems*.
- BRESLER, G., MOSSEL, E. and SLY, A. (2008). Reconstruction of Markov random fields from samples: Some observations and algorithms. In *Approximation, Randomization and Combinatorial Optimization. Lecture Notes in Computer Science* **5171** 343–356. Springer, Berlin. [MR2538799](#)
- CAI, T., LIU, W. and LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. [MR2847973](#)
- CHOW, C. and LIU, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Trans. Inform. Theory* **14** 462–467.
- DASKALAKIS, C., DIKKALA, N. and KAMATH, G. (2018). Testing Ising models. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms* 1989–2007. SIAM, Philadelphia, PA. [MR3775918](#)
- FÖLLMER, H. (1988). Random fields and diffusion processes. In *École D'Été de Probabilités de Saint-Flour XV–XVII, 1985–87. Lecture Notes in Math.* **1362** 101–203. Springer, Berlin. [MR0983373](#)
- GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6** 721–741.
- GHEISSARI, R., LUBETZKY, E. and PERES, Y. (2017). Concentration inequalities for polynomials of contracting Ising models. Preprint. Available at [arXiv:1706.00121](https://arxiv.org/abs/1706.00121).
- GRIMMETT, G. (2018). *Probability on Graphs: Random Processes on Graphs and Lattices. Institute of Mathematical Statistics Textbooks* **8**. Cambridge Univ. Press, Cambridge. [MR3751350](#)
- ISING, E. (1925). Beitrag zur theorie des ferromagnetismus. *Z. Phys. A, Hadrons Nucl.* **31** 253–258.
- KOCHMAŃSKI, M., PASZKIEWICZ, T. and WOLSKI, S. (2013). Curie–Weiss magnet—A simple model of phase transition. *Eur. J. Phys.* **34** 1555.
- LIU, H., LAFFERTY, J. and WASSERMAN, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.* **10** 2295–2328. [MR2563983](#)
- MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. [MR2278363](#)
- MONTANARI, A. and PEREIRA, J. A. (2009). Which graphical models are difficult to learn? In *Advances in Neural Information Processing Systems*.
- MUKHERJEE, R., MUKHERJEE, S. and YUAN, M. (2018). Global testing against sparse alternatives under Ising models. *Ann. Statist.* **46** 2062–2093. [MR3845011](#)
- NEYKOV, M. and LIU, H. (2019). Supplement to “Property testing in high dimensional Ising models.” DOI:[10.1214/18-AOS1754SUPP](https://doi.org/10.1214/18-AOS1754SUPP).
- NEYKOV, M., LU, J. and LIU, H. (2016). Combinatorial inference for graphical models. Preprint. Available at [arXiv:1608.03045](https://arxiv.org/abs/1608.03045).
- RAVIKUMAR, P., WAINWRIGHT, M. J. and LAFFERTY, J. D. (2010). High-dimensional Ising model selection using ℓ_1 -regularized logistic regression. *Ann. Statist.* **38** 1287–1319. [MR2662343](#)
- RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1 -penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980. [MR2836766](#)

- ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2** 494–515. [MR2417391](#)
- SANTHANAM, N. P. and WAINWRIGHT, M. J. (2012). Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Trans. Inform. Theory* **58** 4117–4134. [MR2943079](#)
- SHANMUGAM, K., TANDON, R., RAVIKUMAR, P. K. and DIMAKIS, A. G. (2014). On the information theoretic limits of learning Ising models. In *Advances in Neural Information Processing Systems*.
- YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. [MR2367824](#)

A KNOCKOFF FILTER FOR HIGH-DIMENSIONAL SELECTIVE INFERENCE

BY RINA FOYGEL BARBER¹ AND EMMANUEL J. CANDÈS²

University of Chicago and Stanford University

This paper develops a framework for testing for associations in a possibly high-dimensional linear model where the number of features/variables may far exceed the number of observational units. In this framework, the observations are split into two groups, where the first group is used to screen for a set of potentially relevant variables, whereas the second is used for inference over this reduced set of variables; we also develop strategies for leveraging information from the first part of the data at the inference step for greater power. In our work, the inferential step is carried out by applying the recently introduced knockoff filter, which creates a knockoff copy—a fake variable serving as a control—for each screened variable. We prove that this procedure controls the *directional* false discovery rate (FDR) in the reduced model controlling for all screened variables; this says that our high-dimensional knockoff procedure “discovers” important variables as well as the directions (signs) of their effects, in such a way that the expected proportion of wrongly chosen signs is below the user-specified level (thereby controlling a notion of Type S error averaged over the selected set). This result is nonasymptotic, and holds for any distribution of the original features and any values of the unknown regression coefficients, so that inference is not calibrated under hypothesized values of the effect sizes. We demonstrate the performance of our general and flexible approach through numerical studies, showing more power than existing alternatives. Finally, we apply our method to a genome-wide association study to find locations on the genome that are possibly associated with a continuous phenotype.

REFERENCES

- [1] BARBER, R. F. and CANDÈS, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.* **43** 2055–2085. [MR3375876](#)
- [2] BARBER, R. F. and CANDÈS, E. J. (2019). Supplement to “A knockoff filter for high-dimensional selective inference.” DOI:[10.1214/18-AOS1755SUPP](#).
- [3] 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](#)
- [4] BELLONI, A., CHERNOZHUKOV, V. and WANG, L. (2011). Square-root Lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika* **98** 791–806. [MR2860324](#)
- [5] BENJAMINI, Y. and BRAUN, H. (2002). John Tukey’s contributions to multiple comparisons. ETS Research Report Series.

MSC2010 subject classifications. 62F03, 62J05.

Key words and phrases. Knockoffs, variable selection, false discovery rate (FDR), high-dimensional regression.

- [6] BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. [MR1325392](#)
- [7] BENJAMINI, Y. and YEKUTIELI, D. (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters. *J. Amer. Statist. Assoc.* **100** 71–93. [MR2156820](#)
- [8] BERK, R., BROWN, L., BUJA, A., ZHANG, K. and ZHAO, L. (2013). Valid post-selection inference. *Ann. Statist.* **41** 802–837. [MR3099122](#)
- [9] CANDÈS, E., FAN, Y., JANSON, L. and LV, J. (2018). Panning for gold: ‘Model-X’ knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 551–577. [MR3798878](#)
- [10] DAI, R. and BARBER, R. (2016). The knockoff filter for fdr control in group-sparse and multi-task regression. In *Proceedings of the 33rd International Conference on Machine Learning* 1851–1859.
- [11] FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911. [MR2530322](#)
- [12] FITHIAN, W., SUN, D. and TAYLOR, J. (2014). Optimal inference after model selection. Preprint. Available at [arXiv:1410.2597](#).
- [13] G’SELL, M. G., HASTIE, T. and TIBSHIRANI, R. (2013). False variable selection rates in regression. Preprint. Available at [arXiv:1302.2303](#).
- [14] G’SELL, M. G., 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](#)
- [15] GELMAN, A. and TUERLINCKX, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Comput. Statist.* **15** 373–390.
- [16] HUANG, J., MA, S., ZHANG, C.-H. and ZHOU, Y. (2013). Semi-penalized inference with direct false discovery rate control in high-dimensions. Preprint. Available at [arXiv:1311.7455](#).
- [17] JANSON, L., BARBER, R. F. 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](#)
- [18] JANSON, L. and SU, W. (2016). Familywise error rate control via knockoffs. *Electron. J. Stat.* **10** 960–975. Available at [arXiv:1505.06549](#). [MR3486422](#)
- [19] JÄRVELIN, M.-R., SOVIO, U., KING, V., LAUREN, L., XU, B., MCCARTHY, M. I., HARTIKAINEN, A.-L., LAITINEN, J., ZITTING, P. et al. (2004). Early life factors and blood pressure at age 31 years in the 1966 northern Finland birth cohort. *Hypertension* **44** 838–846.
- [20] JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909. [MR3277152](#)
- [21] JONES, L. V. and TUKEY, J. W. (2000). A sensible formulation of the significance test. *Psychol. Methods* **5** 411.
- [22] LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927. Available at [arXiv:1311.6238](#). [MR3485948](#)
- [23] LEEB, H. and PÖTSCHER, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *Ann. Statist.* **34** 2554–2591. [MR2291510](#)
- [24] LOCKHART, R., TAYLOR, J., TIBSHIRANI, R. J. and TIBSHIRANI, R. (2014). A significance test for the Lasso. *Ann. Statist.* **42** 413–468. [MR3210970](#)
- [25] MILLER, A. (2002). *Subset Selection in Regression*, 2nd ed. *Monographs on Statistics and Applied Probability* **95**. CRC Press/CRC, Boca Raton, FL. [MR2001193](#)

- [26] PATI, Y. C., REZAIIFAR, R. and KRISHNAPRASAD, P. S. (1993). Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers* 40–44. IEEE, New York.
- [27] PRICE, A. L., PATTERSON, N. J., PLENGE, R. M., WEINBLATT, M. E., SHADICK, N. A. and REICH, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38** 904–909.
- [28] SABATTI, C., SERVICE, S. K., HARTIKAINEN, A.-L., POUTA, A., RIPATTI, S., BRODSKY, J., JONES, C. G., ZAITLEN, N. A., VARILLO, T. et al. (2009). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat. Genet.* **41** 35–46.
- [29] SHAFFER, J. P. (2002). Multiplicity, directional (type III) errors, and the null hypothesis. *Psychol. Methods* **7** 356–369.
- [30] SU, W., BOGDAN, M. and CANDÈS, E. (2017). False discoveries occur early on the Lasso path. *Ann. Statist.* **45** 2133–2150. [MR3718164](#)
- [31] TAYLOR, J. T. (2017). Selective-inference. Available at <https://github.com/jonathan-taylor/selective-inference>.
- [32] TIAN, X., LOFTUS, J. R. and TAYLOR, J. E. (2018). Selective inference with unknown variance via the square-root Lasso. *Biometrika* **105** 755–768. Available at [arXiv:1504.08031](#). [MR3877864](#)
- [33] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. [MR1379242](#)
- [34] TIBSHIRANI, R. J., TAYLOR, J., LOCKHART, R. and TIBSHIRANI, R. (2016). Exact post-selection inference for sequential regression procedures. *J. Amer. Statist. Assoc.* **111** 600–620. [MR3538689](#)
- [35] TUKEY, J. W. (1991). The philosophy of multiple comparisons. *Statist. Sci.* **6** 100–116.
- [36] VOORMAN, A., SHOJAIE, A. and WITTEN, D. (2014). Inference in high dimensions with the penalized score test. Preprint. Available at [arXiv:1401.2678](#).
- [37] WASSERMAN, L. and ROEDER, K. (2009). High-dimensional variable selection. *Ann. Statist.* **37** 2178–2201. [MR2543689](#)
- [38] WILLER, C. J., SCHMIDT, E. M., SENGUPTA, S., PELOSO, G. M., GUSTAFSSON, S., KANONI, S., GANNA, A., CHEN, J., BUCHKOVICH, M. L. et al. (2013). Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45** 1274–1283.
- [39] WU, J., DEVLIN, B., RINGQUIST, S., TRUCCO, M. and ROEDER, K. (2010). Screen and clean: A tool for identifying interactions in genome-wide association studies. *Genet. Epidemiol.* **34** 275–285.
- [40] WU, Y., BOOS, D. D. and STEFANSKI, L. A. (2007). Controlling variable selection by the addition of pseudovariables. *J. Amer. Statist. Assoc.* **102** 235–243. [MR2345541](#)
- [41] 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](#)
- [42] 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](#)
- [43] ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. [MR2274449](#)

SEMI-SUPERVISED INFERENCE: GENERAL THEORY AND ESTIMATION OF MEANS

BY ANRU ZHANG^{*,1}, LAWRENCE D. BROWN^{†,2} AND T. TONY CAI^{†,3}

University of Wisconsin—Madison^{} and University of Pennsylvania[†]*

In memory of Lawrence D. Brown

We propose a general semi-supervised inference framework focused on the estimation of the population mean. As usual in semi-supervised settings, there exists an unlabeled sample of covariate vectors and a labeled sample consisting of covariate vectors along with real-valued responses (“labels”). Otherwise, the formulation is “assumption-lean” in that no major conditions are imposed on the statistical or functional form of the data. We consider both the ideal semi-supervised setting where infinitely many unlabeled samples are available, as well as the ordinary semi-supervised setting in which only a finite number of unlabeled samples is available.

Estimators are proposed along with corresponding confidence intervals for the population mean. Theoretical analysis on both the asymptotic distribution and ℓ_2 -risk for the proposed procedures are given. Surprisingly, the proposed estimators, based on a simple form of the least squares method, outperform the ordinary sample mean. The simple, transparent form of the estimator lends confidence to the perception that its asymptotic improvement over the ordinary sample mean also nearly holds even for moderate size samples. The method is further extended to a nonparametric setting, in which the oracle rate can be achieved asymptotically. The proposed estimators are further illustrated by simulation studies and a real data example involving estimation of the homeless population.

REFERENCES

- ANDO, R. K. and ZHANG, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Mach. Learn. Res.* **6** 1817–1853. [MR2249873](#)
- ANDO, R. K. and ZHANG, T. (2007). Two-view feature generation model for semi-supervised learning. In *Proceedings of the 24th International Conference on Machine Learning* 25–32. ACM, New York.
- AZRIEL, D., BROWN, L. D., SKLAR, M., BERK, R., BUJA, A. and ZHAO, L. (2016). Semi-supervised linear regression. Preprint. Available at [arXiv:1612.02391](#).
- BICKEL, P. J., RITOV, Y. and WELLNER, J. A. (1991). Efficient estimation of linear functionals of a probability measure P with known marginal distributions. *Ann. Statist.* **19** 1316–1346. [MR1126327](#)

MSC2010 subject classifications. Primary 62F10, 62J05; secondary 62F12, 62G08.

Key words and phrases. Confidence interval, efficiency, estimation of mean, limiting distribution, semi-supervised inference.

- BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York. Reprint of the 1993 original. [MR1623559](#)
- BLUM, A. and MITCHELL, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory (Madison, WI, 1998)* 92–100. ACM, New York. [MR1811574](#)
- BRATLEY, P., FOX, B. L. and SCHRAGE, L. E. (1987). *A Guide to Simulation*. Springer, New York.
- BUJA, A., BERK, R., BROWN, L., GEORGE, E., PITKIN, E., TRASKIN, M., ZHAN, K. and ZHAO, L. (2014). Models as approximations, Part I: A conspiracy of nonlinearity and random regressors in linear regression. Preprint. Available at [arXiv:1404.1578](#).
- BUJA, A., BERK, R., BROWN, L., GEORGE, E., KUCHIBHOTLA, A. K. and ZHAO, L. (2016). Models as approximations—Part II: A general theory of model-robust regression. Preprint. Available at [ArXiv:1612.03257](#).
- CHAKRABORTTY, A. and CAI, T. (2018). Efficient and adaptive linear regression in semi-supervised settings. *Ann. Statist.* **46** 1541–1572. [MR3819109](#)
- COCHRAN, W. G. (1953). *Sampling Techniques*. Wiley, New York. [MR0054199](#)
- DENG, L.-Y. and WU, C.-F. J. (1987). Estimation of variance of the regression estimator. *J. Amer. Statist. Assoc.* **82** 568–576. [MR0898360](#)
- FISHMAN, G. S. (1996). *Monte Carlo: Concepts, Algorithms, and Applications*. Springer, New York. [MR1392474](#)
- GRAHAM, B. S. (2011). Efficiency bounds for missing data models with semiparametric restrictions. *Econometrica* **79** 437–452. [MR2809376](#)
- HANSEN, B. E. (2017). Econometrics. Book draft. Available at <http://www.ssc.wisc.edu/~bhansen/econometrics/>.
- HASMINSKII, R. Z. and IBRAGIMOV, I. A. (1983). On asymptotic efficiency in the presence of an infinite-dimensional nuisance parameter. In *Probability Theory and Mathematical Statistics (Tbilisi, 1982)*. *Lecture Notes in Math.* **1021** 195–229. Springer, Berlin. [MR0735986](#)
- HICKERNELL, F. J., LEMIEUX, C. and OWEN, A. B. (2005). Control variates for quasi-Monte Carlo. *Statist. Sci.* **20** 1–31. [MR2182985](#)
- JOHNSON, R. and ZHANG, T. (2008). Graph-based semi-supervised learning and spectral kernel design. *IEEE Trans. Inform. Theory* **54** 275–288. [MR2446753](#)
- KRIEGLER, B. and BERK, R. (2010). Small area estimation of the homeless in Los Angeles: An application of cost-sensitive stochastic gradient boosting. *Ann. Appl. Stat.* **4** 1234–1255. [MR2751340](#)
- KUCHIBHOTLA, A. (2017). Research notes on efficiency in semi-supervised problems. Available from the author at arunku@wharton.upenn.edu.
- LAFFERTY, J. D. and WASSERMAN, L. (2008). Statistical analysis of semi-supervised regression. In *Advances in Neural Information Processing Systems* 801–808.
- LOHR, S. (2009). *Sampling: Design and Analysis*. Nelson Education.
- PENG, H. and SCHICK, A. (2002). On efficient estimation of linear functionals of a bivariate distribution with known marginals. *Statist. Probab. Lett.* **59** 83–91. [MR1925291](#)
- PITKIN, E., BERK, R., BROWN, L., BUJA, A., GEORGE, E., ZHANG, K. and ZHAO, L. (2013). Improved precision in estimating average treatment effects. Preprint. Available at [arXiv:1311.0291](#).
- ROSSI, P. H. (1991). Strategies for homeless research in the 1990s. *Hous. Policy Debate* **2** 1027–1055.
- RUBIN, D. B. (1990). Comment on J. Neyman and causal inference in experiments and observational studies: “On the application of probability theory to agricultural experiments. Essay on principles. Section 9” [*Ann. Agric. Sci.* **10** (1923), 1–51]. *Statist. Sci.* **5** 472–480. [MR1092987](#)
- SPELAWA-NEYMAN, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–472. Translated from the Polish and edited by D. M. Dabrowska and T. P. Speed. [MR1092986](#)

- VAN DER VAART, A. (2002). Semiparametric statistics. In *Lectures on Probability Theory and Statistics (Saint-Flour, 1999)*. *Lecture Notes in Math.* **1781** 331–457. Springer, Berlin. [MR1915446](#)
- VAPNIK, V. N. (2013). *The Nature of Statistical Learning Theory*. Springer, Berlin.
- WANG, J. and SHEN, X. (2007). Large margin semi-supervised learning. *J. Mach. Learn. Res.* **8** 1867–1891. [MR2353822](#)
- WANG, J., SHEN, X. and LIU, Y. (2008). Probability estimation for large-margin classifiers. *Biometrika* **95** 149–167. [MR2409720](#)
- WANG, J., SHEN, X. and PAN, W. (2009). On efficient large margin semisupervised learning: Method and theory. *J. Mach. Learn. Res.* **10** 719–742. [MR2491755](#)
- YASKOV, P. (2014). Lower bounds on the smallest eigenvalue of a sample covariance matrix. *Electron. Commun. Probab.* **19** no. 83, 10. [MR3291620](#)
- ZHANG, A., BROWN, L. D. and CAI, T. T. (2019). Supplement to “Semi-supervised inference: General theory and estimation of means.” DOI:[10.1214/18-AOS1756SUPP](#).
- ZHU, X. (2008). Semi-supervised learning literature survey. Technical report.
- ZHU, X. and GOLDBERG, A. B. (2009). Introduction to semi-supervised learning. *Synth. Lect. Artif. Intell. Mach. Learn.* **3** 1–130. DOI:[10.2200/S00196ED1V01Y200906AIM006](#).

DOUBLY PENALIZED ESTIMATION IN ADDITIVE REGRESSION WITH HIGH-DIMENSIONAL DATA

BY ZHIQIANG TAN¹ AND CUN-HUI ZHANG²

Rutgers University

Additive regression provides an extension of linear regression by modeling the signal of a response as a sum of functions of covariates of relatively low complexity. We study penalized estimation in high-dimensional nonparametric additive regression where functional semi-norms are used to induce smoothness of component functions and the empirical L_2 norm is used to induce sparsity. The functional semi-norms can be of Sobolev or bounded variation types and are allowed to be different amongst individual component functions. We establish oracle inequalities for the predictive performance of such methods under three simple technical conditions: a sub-Gaussian condition on the noise, a compatibility condition on the design and the functional classes under consideration and an entropy condition on the functional classes. For random designs, the sample compatibility condition can be replaced by its population version under an additional condition to ensure suitable convergence of empirical norms. In homogeneous settings where the complexities of the component functions are of the same order, our results provide a spectrum of minimax convergence rates, from the so-called slow rate without requiring the compatibility condition to the fast rate under the hard sparsity or certain L_q sparsity to allow many small components in the true regression function. These results significantly broaden and sharpen existing ones in the literature.

REFERENCES

- BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. [MR2533469](#)
- BUNEA, F., TSYBAKOV, A. and WEGKAMP, M. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.* **1** 169–194. [MR2312149](#)
- DALALYAN, A., INGSTER, Y. and TSYBAKOV, A. B. (2014). Statistical inference in compound functional models. *Probab. Theory Related Fields* **158** 513–532. [MR3176357](#)
- DEVORE, R. A. and LORENTZ, G. G. (1993). *Constructive Approximation. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]* **303**. Springer, Berlin. [MR1261635](#)
- GREENSHTEIN, E. and RITOV, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10** 971–988. [MR2108039](#)
- GU, C. (2002). *Smoothing Spline ANOVA Models. Springer Series in Statistics*. Springer, New York. [MR1876599](#)

MSC2010 subject classifications. Primary 62E20, 62F25, 62F35; secondary 62J05, 62J12.

Key words and phrases. Additive model, bounded variation space, ANOVA model, high-dimensional data, metric entropy, penalized estimation, reproducing kernel Hilbert space, Sobolev space, total variation, trend filtering.

- HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models. Monographs on Statistics and Applied Probability* **43**. CRC Press, London. [MR1082147](#)
- HUANG, J., HOROWITZ, J. L. and WEI, F. (2010). Variable selection in nonparametric additive models. *Ann. Statist.* **38** 2282–2313. [MR2676890](#)
- KIM, S.-J., KOH, K., BOYD, S. and GORINEVSKY, D. (2009). l_1 trend filtering. *SIAM Rev.* **51** 339–360. [MR2505584](#)
- 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](#)
- KOLTCHINSKII, V. and YUAN, M. (2010). Sparsity in multiple kernel learning. *Ann. Statist.* **38** 3660–3695. [MR2766864](#)
- LEDoux, M. and TALAGRAND, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes. Ergebnisse der Mathematik und Ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]* **23**. Springer, Berlin. [MR1102015](#)
- LIN, Y. and ZHANG, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.* **34** 2272–2297. [MR2291500](#)
- LORENTZ, G. G., GOLITSCHKE, M. V. and MAKOVZ, Y. (1996). *Constructive Approximation: Advanced Problems. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]* **304**. Springer, Berlin. [MR1393437](#)
- MAMMEN, E. (1991). Nonparametric regression under qualitative smoothness assumptions. *Ann. Statist.* **19** 741–759. [MR1105842](#)
- MAMMEN, E. and VAN DE GEER, S. (1997). Locally adaptive regression splines. *Ann. Statist.* **25** 387–413. [MR1429931](#)
- MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2009). High-dimensional additive modeling. *Ann. Statist.* **37** 3779–3821. [MR2572443](#)
- MÜLLER, P. and VAN DE GEER, S. (2015). The partial linear model in high dimensions. *Scand. J. Stat.* **42** 580–608. [MR3345123](#)
- NIRENBERG, L. (1966). An extended interpolation inequality. *Ann. Sc. Norm. Super. Pisa Cl. Sci. (3)* **20** 733–737. [MR0208360](#)
- PETERSEN, A., WITTEN, D. and SIMON, N. (2016). Fused lasso additive model. *J. Comput. Graph. Statist.* **25** 1005–1025. [MR3572026](#)
- RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Mach. Learn. Res.* **13** 389–427. [MR2913704](#)
- RAVIKUMAR, P., LAFFERTY, J., LIU, H. and WASSERMAN, L. (2009). Sparse additive models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 1009–1030. [MR2750255](#)
- SADHANALA, V. and TIBSHIRANI, R. J. (2017). Additive models with trend filtering. Preprint. Available at [arXiv:1702.05037](#).
- STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053. [MR0673642](#)
- STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705. [MR0790566](#)
- SUZUKI, T. and SUGIYAMA, M. (2013). Fast learning rate of multiple kernel learning: Trade-off between sparsity and smoothness. *Ann. Statist.* **41** 1381–1405. [MR3113815](#)
- TAN, Z. and ZHANG, C.-H. (2019). Supplement to “Doubly penalized estimation in additive regression with high-dimensional data.” DOI:[10.1214/18-AOS1757SUPP](#).
- TIBSHIRANI, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *Ann. Statist.* **42** 285–323. [MR3189487](#)
- VAN DE GEER, S. (2000). *Empirical Processes in M-Estimation*. Cambridge Univ. Press, Cambridge.
- 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](#)

- 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](#)
- YANG, T. and TAN, Z. (2018). Backfitting algorithms for total-variation and empirical-norm penalized additive modelling with high-dimensional data. *Stat* **7** e198. [MR3905854](#)
- YANG, Y. and TOKDAR, S. T. (2015). Minimax-optimal nonparametric regression in high dimensions. *Ann. Statist.* **43** 652–674. [MR3319139](#)
- YUAN, M. and ZHOU, D.-X. (2016). Minimax optimal rates of estimation in high dimensional additive models. *Ann. Statist.* **44** 2564–2593. [MR3576554](#)

SEMPARAMETRICALLY POINT-OPTIMAL HYBRID RANK TESTS FOR UNIT ROOTS

BY BO ZHOU*, RAMON VAN DEN AKKER[†] AND BAS J. M. WERKER[†]

*Hong Kong University of Science and Technology** and *Tilburg University[†]*

We propose a new class of unit root tests that exploits invariance properties in the Locally Asymptotically Brownian Functional limit experiment associated to the unit root model. The invariance structures naturally suggest tests that are based on the ranks of the increments of the observations, their average and an assumed reference density for the innovations. The tests are semiparametric in the sense that they are valid, that is, have the correct (asymptotic) size, irrespective of the true innovation density. For a correctly specified reference density, our test is point-optimal and nearly efficient. For arbitrary reference densities, we establish a Chernoff–Savage-type result, that is, our test performs as well as commonly used tests under Gaussian innovations but has improved power under other, for example, fat-tailed or skewed, innovation distributions. To avoid nonparametric estimation, we propose a simplified version of our test that exhibits the same asymptotic properties, except for the Chernoff–Savage result that we are only able to demonstrate by means of simulations.

REFERENCES

- AHN, S. K., FOTOPOULOS, S. B. and HE, L. (2001). Unit root tests with infinite variance errors. *Econometric Rev.* **20** 461–483. [MR1868343](#)
- BICKEL, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10** 647–671. [MR0663424](#)
- BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York. Reprint of the 1993 original. [MR1623559](#)
- BROCKWELL, P. J. and DAVIS, R. A. (2016). *Introduction to Time Series and Forecasting*, 3rd ed. *Springer Texts in Statistics*. Springer, Cham. [MR3526245](#)
- CALLEGARI, F., CAPPUCCIO, N. and LUBIAN, D. (2003). Asymptotic inference in time series regressions with a unit root and infinite variance errors. *J. Statist. Plann. Inference* **116** 277–303. [MR1997142](#)
- CASSART, D., HALLIN, M. and PAINDAVEINE, D. (2010). On the estimation of cross-information quantities in rank-based inference. In *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in Honor of Professor Jana Jurečková*. *Inst. Math. Stat. (IMS) Collect.* **7** 35–45. IMS, Beachwood, OH. [MR2808364](#)
- CHAN, N. H. and WEI, C. Z. (1988). Limiting distributions of least squares estimates of unstable autoregressive processes. *Ann. Statist.* **16** 367–401. [MR0924877](#)
- CHERNOFF, H. and SAVAGE, I. R. (1958). Asymptotic normality and efficiency of certain nonparametric test statistics. *Ann. Math. Stat.* **29** 972–994. [MR0100322](#)

MSC2010 subject classifications. Primary 62G10, 62G20; secondary 62P20, 62M10.

Key words and phrases. Unit root test, semiparametric power envelope, limit experiment, LABF, maximal invariant, rank statistic.

- CHOI, I. (2015). *Almost All About Unit Roots. Themes in Modern Econometrics*. Cambridge Univ. Press, New York. [MR3558861](#)
- DICKEY, D. A. and FULLER, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *J. Amer. Statist. Assoc.* **74** 427–431. [MR0548036](#)
- DICKEY, D. A. and FULLER, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica* **49** 1057–1072. [MR0625773](#)
- DUFOUR, J.-M. and KING, M. L. (1991). Optimal invariant tests for the autocorrelation coefficient in linear regressions with stationary or nonstationary AR(1) errors. *J. Econometrics* **47** 115–143. [MR1087209](#)
- ELLIOTT, G. and MÜLLER, U. K. (2006). Minimizing the impact of the initial condition on testing for unit roots. *J. Econometrics* **135** 285–310. [MR2328403](#)
- ELLIOTT, G., ROTHENBERG, T. J. and STOCK, J. H. (1996). Efficient tests for an autoregressive unit root. *Econometrica* **64** 813–836. [MR1399219](#)
- HÁJEK, J. and ŠIDÁK, Z. (1967). *Theory of Rank Tests*. Academic Press, New York. [MR0229351](#)
- HALLIN, M. and PURI, M. L. (1988). Optimal rank-based procedures for time series analysis: Testing an ARMA model against other ARMA models. *Ann. Statist.* **16** 402–432. [MR0924878](#)
- HALLIN, M. and PURI, M. L. (1994). Aligned rank tests for linear models with autocorrelated error terms. *J. Multivariate Anal.* **50** 175–237. [MR1293044](#)
- HALLIN, M., VAN DEN AKKER, R. and WERKER, B. J. M. (2011). A class of simple distribution-free rank-based unit root tests. *J. Econometrics* **163** 200–214. [MR2812866](#)
- HALLIN, M., VAN DEN AKKER, R. and WERKER, B. J. M. (2015). On quadratic expansions of log-likelihoods and a general asymptotic linearity result. In *Mathematical Statistics and Limit Theorems* 147–165. Springer, Cham. [MR3380735](#)
- HASAN, M. N. (2001). Rank tests of unit root hypothesis with infinite variance errors. *J. Econometrics* **104** 49–65. [MR1862029](#)
- JANSSON, M. (2008). Semiparametric power envelopes for tests of the unit root hypothesis. *Econometrica* **76** 1103–1142. [MR2455121](#)
- JEGANATHAN, P. (1995). Some aspects of asymptotic theory with applications to time series models. *Econometric Theory* **11** 818–887. [MR1458943](#)
- JEGANATHAN, P. (1997). On asymptotic inference in linear cointegrated time series systems. *Econometric Theory* **13** 692–745. [MR1491257](#)
- KLAASSEN, C. A. J. (1987). Consistent estimation of the influence function of locally asymptotically linear estimators. *Ann. Statist.* **15** 1548–1562. [MR0913573](#)
- KREISS, J.-P. (1987). On adaptive estimation in stationary ARMA processes. *Ann. Statist.* **15** 112–133. [MR0885727](#)
- LE CAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory. Springer Series in Statistics*. Springer, New York. [MR0856411](#)
- MÜLLER, U. K. (2011). Efficient tests under a weak convergence assumption. *Econometrica* **79** 395–435. [MR2809375](#)
- MÜLLER, U. K. and ELLIOTT, G. (2003). Tests for unit roots and the initial condition. *Econometrica* **71** 1269–1286. [MR1995831](#)
- MÜLLER, U. K. and WATSON, M. W. (2008). Testing models of low-frequency variability. *Econometrica* **76** 979–1016. [MR2455119](#)
- PATTERSON, K. (2011). *Unit Root Tests in Time Series. Vol. 1. Palgrave Texts in Econometrics*. Palgrave Macmillan, New York. [MR3184412](#)
- PATTERSON, K. (2012). *Unit Root Tests in Time Series. Vol. 2. Palgrave Texts in Econometrics*. Palgrave Macmillan, New York. [MR3237671](#)
- PHILLIPS, P. C. B. (1987). Time series regression with a unit root. *Econometrica* **55** 277–301. [MR0882096](#)
- PHILLIPS, P. C. B. and PERRON, P. (1988). Testing for a unit root in time series regression. *Biometrika* **75** 335–346. [MR0946054](#)

- ROTHENBERG, T. J. and STOCK, J. H. (1997). Inference in a nearly integrated autoregressive model with nonnormal innovations. *J. Econometrics* **80** 269–286. [MR1483180](#)
- RUDIN, W. (1987). *Real and Complex Analysis*, 3rd ed. McGraw-Hill, New York. [MR0924157](#)
- SAIKKONEN, P. and LUUKKONEN, R. (1993). Point optimal tests for testing the order of differencing in ARIMA models. *Econometric Theory* **9** 343–362. [MR1241978](#)
- SCHICK, A. (1986). On asymptotically efficient estimation in semiparametric models. *Ann. Statist.* **14** 1139–1151. [MR0856811](#)
- VAN DER VAART, A. W. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. [MR1652247](#)
- WHITE, J. S. (1958). The limiting distribution of the serial correlation coefficient in the explosive case. *Ann. Math. Stat.* **29** 1188–1197. [MR0100952](#)
- ZHOU, B., VAN DEN AKKER, R. and WERKER, B. J. M. (2019). Supplement to “Semiparametrically point-optimal hybrid rank tests for unit roots.” DOI:[10.1214/18-AOS1758SUPP](#).

THE MIDDLE-SCALE ASYMPTOTICS OF WISHART MATRICES

BY DIDIER CHÉTELAT AND MARTIN T. WELLS¹

École Polytechnique de Montréal and Cornell University

We study the behavior of a real p -dimensional Wishart random matrix with n degrees of freedom when $n, p \rightarrow \infty$ but $p/n \rightarrow 0$. We establish the existence of phase transitions when p grows at the order $n^{(K+1)/(K+3)}$ for every $K \in \mathbb{N}$, and derive expressions for approximating densities between every two phase transitions. To do this, we make use of a novel tool we call the $\mathcal{F}$ -conjugate of an absolutely continuous distribution, which is obtained from the Fourier transform of the square root of its density. In the case of the normalized Wishart distribution, this represents an extension of the t -distribution to the space of real symmetric matrices.

REFERENCES

- ANDERSON, G. W., GUIONNET, A. and ZEITOUNI, O. (2010). *An Introduction to Random Matrices*. Cambridge Studies in Advanced Mathematics **118**. Cambridge Univ. Press, Cambridge. [MR2760897](#)
- BARTLETT, M. S. (1933). On the theory of statistical regression. *Proc. Roy. Soc. Edinburgh Sect. A* **53** 260–283.
- BUBECK, S. and GANGULY, S. (2018). Entropic CLT and phase transition in high-dimensional Wishart matrices. *Int. Math. Res. Not. IMRN* **2** 588–606. [MR3801440](#)
- BUBECK, S., DING, J., ELDAN, R. and RÁČZ, M. Z. (2016). Testing for high-dimensional geometry in random graphs. *Random Structures Algorithms* **49** 503–532. [MR3545825](#)
- HARDY, G. H., LITTLEWOOD, J. E. and PÓLYA, G. (1967). *Inequalities*, 2nd ed. Cambridge University Press.
- HUA, L. K. (1963). *Harmonic Analysis of Functions of Several Complex Variables in the Classical Domains*. Amer. Math. Soc., Providence, RI. [MR0171936](#)
- JIANG, T. and LI, D. (2015). Approximation of rectangular beta-Laguerre ensembles and large deviations. *J. Theoret. Probab.* **28** 804–847. [MR3413957](#)
- KNUTH, D. (1976). Big omicron and big omega and big theta. *SIGACT News* 18–24.
- LEDoux, M. (2009). A recursion formula for the moments of the Gaussian orthogonal ensemble. *Ann. Inst. Henri Poincaré Probab. Stat.* **45** 754–769. [MR2548502](#)
- LETAC, G. and MASSAM, H. (2004). All invariant moments of the Wishart distribution. *Scand. J. Stat.* **31** 295–318. [MR2066255](#)
- MARČENKO, V. A. and PASTUR, L. A. (1967). Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb.* **72** 507–536. [MR0208649](#)
- MATSUMOTO, S. (2012). General moments of the inverse real Wishart distribution and orthogonal Weingarten functions. *J. Theoret. Probab.* **25** 798–822. [MR2956213](#)
- MUIRHEAD, R. J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley, New York. [MR0652932](#)

MSC2010 subject classifications. Primary 60B20, 60B10; secondary 60E10.

Key words and phrases. Covariance estimation, Gaussian orthogonal ensemble, high-dimensional asymptotics, phase transitions, random matrix theory.

- RÁCZ, M. Z. and RICHEY, J. (2018). A smooth transition from Wishart to GOE. *J. Theoret. Probab.* 1–9.
- STEIN, E. M. and WEISS, G. (1971). *Introduction to Fourier Analysis on Euclidean Spaces*. Princeton Mathematical Series **32**. Princeton Univ. Press, Princeton, NJ. [MR0304972](#)
- WISHART, J. (1928). The generalised product moment distribution in samples from a normal multivariate population. *Biometrika* **20A** 32–52.

LINEAR HYPOTHESIS TESTING FOR HIGH DIMENSIONAL GENERALIZED LINEAR MODELS

BY CHENGCHUN SHI^{*,1}, RUI SONG^{*,1}, ZHAO CHEN^{†,2} AND RUNZE LI^{‡,3}

North Carolina State University^{}, Fudan University[†] and
 Pennsylvania State University[‡]*

This paper is concerned with testing linear hypotheses in high dimensional generalized linear models. To deal with linear hypotheses, we first propose the constrained partial regularization method and study its statistical properties. We further introduce an algorithm for solving regularization problems with folded-concave penalty functions and linear constraints. To test linear hypotheses, we propose a partial penalized likelihood ratio test, a partial penalized score test and a partial penalized Wald test. We show that the limiting null distributions of these three test statistics are χ^2 distribution with the same degrees of freedom, and under local alternatives, they asymptotically follow noncentral χ^2 distributions with the same degrees of freedom and noncentral parameter, provided the number of parameters involved in the test hypothesis grows to ∞ at a certain rate. Simulation studies are conducted to examine the finite sample performance of the proposed tests. Empirical analysis of a real data example is used to illustrate the proposed testing procedures.

REFERENCES

- BENTKUS, V. (2004). A Lyapunov type bound in $\mathbf{R}^d$. *Teor. Veroyatn. Primen.* **49** 400–410. [MR2144310](#)
- BOYD, S., PARIKH, N., CHU, E., PELEATO, B. and ECKSTEIN, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **3** 1–122.
- BREHENY, P. and HUANG, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Stat.* **5** 232–253. [MR2810396](#)
- CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n . *Ann. Statist.* **35** 2313–2351. [MR2382644](#)
- DEZEURE, R., BÜHLMANN, P., MEIER, L. and MEINSHAUSEN, N. (2015). High-dimensional inference: Confidence intervals, p -values and R-software `hdi`. *Statist. Sci.* **30** 533–558. [MR3432840](#)
- FAN, J., GUO, S. and HAO, N. (2012). Variance estimation using refitted cross-validation in ultra-high dimensional regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 37–65. [MR2885839](#)
- FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. [MR1946581](#)
- FAN, J. and LV, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica* **20** 101–148. [MR2640659](#)

MSC2010 subject classifications. Primary 62F03; secondary 62J12.

Key words and phrases. High dimensional testing, linear hypothesis, likelihood ratio statistics, score test, Wald test.

- FAN, J. and LV, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Inform. Theory* **57** 5467–5484. [MR2849368](#)
- FAN, J. and PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32** 928–961. [MR2065194](#)
- FAN, Y. and TANG, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 531–552. [MR3065478](#)
- FANG, E. X., NING, Y. and LIU, H. (2017). Testing and confidence intervals for high dimensional proportional hazards models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1415–1437. [MR3731669](#)
- GHOSH, B. K. (1973). Some monotonicity theorems for χ^2 , F and t distributions with applications. *J. Roy. Statist. Soc. Ser. B* **35** 480–492. [MR0336872](#)
- LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927. [MR3485948](#)
- LOCKHART, R., TAYLOR, J., TIBSHIRANI, R. J. and TIBSHIRANI, R. (2014). A significance test for the lasso. *Ann. Statist.* **42** 413–468. [MR3210970](#)
- MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. [of MR0727836]. Chapman & Hall, London. [MR3223057](#)
- NING, Y. and LIU, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Statist.* **45** 158–195. [MR3611489](#)
- SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. [MR0468014](#)
- SHI, C., SONG, R., CHEN, Z. and LI, R. (2019). Supplement to “Linear hypothesis testing for high dimensional generalized linear models.” DOI:[10.1214/18-AOS1761SUPP](#).
- SUN, T. and ZHANG, C.-H. (2013). Sparse matrix inversion with scaled lasso. *J. Mach. Learn. Res.* **14** 3385–3418. [MR3144466](#)
- TAYLOR, J., LOCKHART, R., TIBSHIRANI, R. J. and TIBSHIRANI, R. (2015). Exact Post-Selection Inference for Sequential Regression Procedures. Preprint. Available at [arXiv:1401.3889](#).
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. [MR1379242](#)
- 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](#)
- WANG, S. and CUI, H. (2013). Partial penalized likelihood ratio test under sparse case. Preprint. Available at [arXiv:1312.3723](#).
- ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. [MR2604701](#)
- ZHANG, X. and CHENG, G. (2017). Simultaneous inference for high-dimensional linear models. *J. Amer. Statist. Assoc.* **112** 757–768. [MR3671768](#)

AN OPERATOR THEORETIC APPROACH TO NONPARAMETRIC MIXTURE MODELS¹

BY ROBERT A. VANDERMEULEN AND CLAYTON D. SCOTT

Technische Universität Kaiserslautern and University of Michigan

When estimating finite mixture models, it is common to make assumptions on the mixture components, such as parametric assumptions. In this work, we make no distributional assumptions on the mixture components and instead assume that observations from the mixture model are grouped, such that observations in the same group are known to be drawn from the same mixture component. We precisely characterize the number of observations n per group needed for the mixture model to be identifiable, as a function of the number m of mixture components. In addition to our assumption-free analysis, we also study the settings where the mixture components are either linearly independent or jointly irreducible. Furthermore, our analysis considers two kinds of identifiability, where the mixture model is the simplest one explaining the data, and where it is the only one. As an application of these results, we precisely characterize identifiability of multinomial mixture models. Our analysis relies on an operator-theoretic framework that associates mixture models in the grouped-sample setting with certain infinite-dimensional tensors. Based on this framework, we introduce a general spectral algorithm for recovering the mixture components.

REFERENCES

- [1] ALLMAN, E. S., MATIAS, C. and RHODES, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.* **37** 3099–3132. [MR2549554](#)
- [2] ANANDKUMAR, A., GE, R., HSU, D., KAKADE, S. M. and TELGARSKY, M. (2014). Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.* **15** 2773–2832. [MR3270750](#)
- [3] ANDERSON, J., BELKIN, M., GOYAL, N., RADEMACHER, L. and VOSS, J. (2014). The more, the merrier: The blessing of dimensionality for learning large Gaussian mixtures. In *Proceedings of the 27th Conference on Learning Theory* 1135–1164.
- [4] ARORA, S., GE, R., KANNAN, R. and MOITRA, A. (2012). Computing a nonnegative matrix factorization—provably. In *STOC’12—Proceedings of the 2012 ACM Symposium on Theory of Computing* 145–161. ACM, New York. [MR2961503](#)
- [5] ARORA, S., GE, R. and MOITRA, A. (2012). Learning topic models—going beyond SVD. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science—FOCS 2012* 1–10. IEEE Computer Soc., Los Alamitos, CA. [MR3185945](#)
- [6] BALAN, R., CASAZZA, P. and EDIDIN, D. (2006). On signal reconstruction without phase. *Appl. Comput. Harmon. Anal.* **20** 345–356. [MR2224902](#)

MSC2010 subject classifications. Primary 62E10; secondary 62G05.

Key words and phrases. Mixture model, nonparametric mixture, identifiability, tensor factorization, multinomial mixture, topic model, joint irreducibility.

- [7] BHASKARA, A., CHARIKAR, M., MOITRA, A. and VIJAYARAGHAVAN, A. (2014). Smoothed analysis of tensor decompositions. In *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing, STOC'14* 594–603. ACM, New York.
- [8] BLANCHARD, G. and SCOTT, C. (2014). Decontamination of mutually contaminated models. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics* 1–9.
- [9] BRUNI, C. and KOCH, G. (1985). Identifiability of continuous mixtures of unknown Gaussian distributions. *Ann. Probab.* **13** 1341–1357. [MR0806230](#)
- [10] COMON, P., GOLUB, G., LIM, L.-H. and MOURRAIN, B. (2008). Symmetric tensors and symmetric tensor rank. *SIAM J. Matrix Anal. Appl.* **30** 1254–1279. [MR2447451](#)
- [11] DONOHO, D. and STODDEN, V. (2004). When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems* (S. Thrun, L. K. Saul and B. Schölkopf, eds.) **16** 1141–1148. MIT Press, Cambridge, MA.
- [12] ELMORE, R. and WANG, S. (2003). Identifiability and estimation in finite mixture models with multinomial components. Technical Report 03-04, Dept. Statistics, Pennsylvania State Univ., State College, PA.
- [13] FOLLAND, G. B. (1999). *Real Analysis: Modern Techniques and Their Applications*, 2nd ed. *Pure and Applied Mathematics (New York)*. Wiley, New York. [MR1681462](#)
- [14] GOLUB, G. H. and VAN LOAN, C. F. (1996). *Matrix Computations*, 3rd ed. *Johns Hopkins Studies in the Mathematical Sciences*. Johns Hopkins Univ. Press, Baltimore, MD. [MR1417720](#)
- [15] KADISON, R. V. and RINGROSE, J. R. (1983). *Fundamentals of the Theory of Operator Algebras. Vol. I: Elementary Theory. Pure and Applied Mathematics* **100**. Academic Press, New York. [MR0719020](#)
- [16] KALLENBERG, O. (2002). *Foundations of Modern Probability*, 2nd ed. *Probability and Its Applications (New York)*. Springer, New York. [MR1876169](#)
- [17] KIM, B. S. (1984). Studies of multinomial mixture models. Ph.D. thesis, Univ. North Carolina, Chapel Hill.
- [18] KRUSKAL, J. B. (1977). Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra Appl.* **18** 95–138. [MR0444690](#)
- [19] MICCHELLI, C. A., XU, Y. and ZHANG, H. (2006). Universal kernels. *J. Mach. Learn. Res.* **7** 2651–2667. [MR2274454](#)
- [20] PACHTER, L. and SPEYER, D. (2004). Reconstructing trees from subtree weights. *Appl. Math. Lett.* **17** 615–621. [MR2064171](#)
- [21] PAZ, A. (1971). *Introduction to Probabilistic Automata*. Academic Press, New York. [MR0289222](#)
- [22] RABANI, Y., SCHULMAN, L. J. and SWAMY, C. (2014). Learning mixtures of arbitrary distributions over large discrete domains. In *ITCS'14—Proceedings of the 2014 Conference on Innovations in Theoretical Computer Science* 207–223. ACM, New York. [MR3359477](#)
- [23] SONG, L., ANANDKUMAR, A., DAI, B. and XIE, B. (2014). Nonparametric estimation of multi-view latent variable models. In *Proceedings of the 31st International Conference on Machine Learning, ICML 2014* 640–648.
- [24] TEICHER, H. (1963). Identifiability of finite mixtures. *Ann. Math. Stat.* **34** 1265–1269. [MR0155376](#)
- [25] VANDERMEULEN, R. A. and SCOTT, C. D. (2019). Supplement to “An operator theoretic approach to nonparametric mixture models.” DOI:[10.1214/18-AOS1762SUPP](#).
- [26] YAKOWITZ, S. J. and SPRAGINS, J. D. (1968). On the identifiability of finite mixtures. *Ann. Math. Stat.* **39** 209–214. [MR0224204](#)

PHASE TRANSITION IN THE SPIKED RANDOM TENSOR WITH RADEMACHER PRIOR¹

BY WEI-KUO CHEN

University of Minnesota

We consider the problem of detecting a deformation from a symmetric Gaussian random p -tensor ($p \geq 3$) with a rank-one spike sampled from the Rademacher prior. Recently, in Lesieur et al. (Barbier, Krzakala, Macris, Miolane and Zdeborová (2017)), it was proved that there exists a critical threshold β_p so that when the signal-to-noise ratio exceeds β_p , one can distinguish the spiked and unspiked tensors and weakly recover the prior via the minimal mean-square-error method. On the other side, Perry, Wein and Bandeira (Perry, Wein and Bandeira (2017)) proved that there exists a $\beta'_p < \beta_p$ such that any statistical hypothesis test cannot distinguish these two tensors, in the sense that their total variation distance asymptotically vanishes, when the signal-to-noise ratio is less than β'_p . In this work, we show that β_p is indeed the critical threshold that strictly separates the distinguishability and indistinguishability between the two tensors under the total variation distance. Our approach is based on a subtle analysis of the high temperature behavior of the pure p -spin model with Ising spin, arising initially from the field of spin glasses. In particular, we identify the signal-to-noise criticality β_p as the critical temperature, distinguishing the high and low temperature behavior, of the Ising pure p -spin mean-field spin glass model.

REFERENCES

- [1] AIZENMAN, M., LEBOWITZ, J. L. and RUELLE, D. (1987). Some rigorous results on the Sherrington–Kirkpatrick spin glass model. *Comm. Math. Phys.* **112** 3–20. [MR0904135](#)
- [2] AUFFINGER, A. and CHEN, W.-K. (2015). The Parisi formula has a unique minimizer. *Comm. Math. Phys.* **335** 1429–1444. [MR3320318](#)
- [3] 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](#)
- [4] BARBIER, J., DIA, M., MACRIS, N. and KRZAKALA, F. (2016). The mutual information in random linear estimation. In *54th Annual Allerton Conference on Communication, Control, and Computing* 625–632.
- [5] BARBIER, J., DIA, M., MACRIS, N., KRZAKALA, F., LESIEUR, T. and ZDEBOROVÁ, L. (2016). Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula. *Adv. Neural Inf. Process. Syst.* **29** 424–432.
- [6] BARBIER, J., KRZAKALA, F., MACRIS, N., MIOLANE, L. and ZDEBOROVÁ, L. (2017). Phase transitions, optimal errors and optimality of message-passing in generalized linear models. ArXiv e-prints.

MSC2010 subject classifications. Primary 93E10; secondary 60K35, 82B44.

Key words and phrases. BBP transition, signal detection, Parisi formula, replica symmetry breaking, spin glass, spiked tensor.

- [7] BARBIER, J. and MACRIS, N. (2017). The stochastic interpolation method: A simple scheme to prove replica formulas in bayesian inference. ArXiv e-prints.
- [8] BARBIER, J., MACRIS, N., DIA, M. and KRZAKALA, F. (2017). Mutual information and optimality of approximate message-passing in random linear estimation. In ArXiv E-prints.
- [9] BARBIER, J., MACRIS, N. and MIOLANE, L. (2017). The layered structure of tensor estimation and its mutual information. In *55th Annual Allerton Conference on Communication, Control, and Computing*.
- [10] BARDINA, X., MÁRQUEZ-CARRERAS, D., ROVIRA, C. and TINDEL, S. (2004). The p -spin interaction model with external field. *Potential Anal.* **21** 311–362. [MR2081143](#)
- [11] BAYATI, M. and MONTANARI, A. (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. Inform. Theory* **57** 764–785. [MR2810285](#)
- [12] BEN AROUS, G., MEI, S., MONTANARI, A. and NICA, M. (2017). The landscape of the spiked tensor model. ArXiv e-prints.
- [13] 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](#)
- [14] BOVIER, A., KURKOVA, I. and LÖWE, M. (2002). Fluctuations of the free energy in the REM and the p -spin SK models. *Ann. Probab.* **30** 605–651. [MR1905853](#)
- [15] 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](#)
- [16] CHEN, W.-K. (2014). Chaos in the mixed even-spin models. *Comm. Math. Phys.* **328** 867–901. [MR3201215](#)
- [17] CHEN, W.-K. (2017). Variational representations for the Parisi functional and the two-dimensional Guerra–Talagrand bound. *Ann. Probab.* **45** 3929–3966. [MR3729619](#)
- [18] CHEN, W.-K. (2019). Supplement to “Phase transition in the spiked random tensor with Rademacher prior.” DOI:[10.1214/18-AOS1763SUPP](#).
- [19] DESHPANDE, Y., ABBE, E. and MONTANARI, A. (2017). Asymptotic mutual information for the balanced binary stochastic block model. *Inf. Inference* **6** 125–170. [MR3671474](#)
- [20] DESHPANDE, Y. and MONTANARI, A. (2014). Information-theoretically optimal sparse pca. In *IEEE Internation Symposium on Information Theory*. 2197–2201.
- [21] DONNOHO, D. L., MALEKI, A. and MONTANARI, A. (2009). Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci. USA* **106** 18914–18919.
- [22] DONNOHO, D. L., MALEKI, A. and MONTANARI, A. (2010). Message-passing algorithms for compressed sensing: I. motivation and construction. In *IEEE Information Theory Workshop (ITW)*. 115–144.
- [23] EL ALAOUI, A., KRZAKALA, F. and JORDAN, M. I. (2017). Finite-size corrections and likelihood ratio fluctuations in the spiked wigner model. ArXiv e-prints.
- [24] JAGANNATH, A. and TOBASCO, I. (2016). A dynamic programming approach to the Parisi functional. *Proc. Amer. Math. Soc.* **144** 3135–3150. [MR3487243](#)
- [25] JAVANMARD, A. and MONTANARI, A. (2013). State evolution for general approximate message passing algorithms, with applications to spatial coupling. *Inf. Inference* **2** 115–144. [MR3311445](#)
- [26] KORADA, S. B. and MACRIS, N. (2009). Exact solution of the gauge symmetric p -spin glass model on a complete graph. *J. Stat. Phys.* **136** 205–230. [MR2525244](#)
- [27] KRZAKALA, F., XU, J. and ZDEBOROVÁ, L. (2016). Mutual information in rank-one matrix estimation. *IEEE Information Theory Workshop (ITW)* 71–75.
- [28] LELARGE, M. and MIOLANE, L. Fundamental limits of symmetric low-rank matrix estimation. In *Proceedings of the 2017 Conference on Learning Theory*. PMLR 65.

- [29] LESIEUR, T., KRZAKALA, F. and ZDEBOROVÁ, L. (2017). Constrained low-rank matrix estimation: Phase transitions, approximate message passing and applications. *J. Stat. Mech. Theory Exp.* **7** 073403. [MR3683819](#)
- [30] MÉZARD, M., PARISI, G. and VIRASORO, M. A. (1987). *Spin Glass Theory and Beyond. World Scientific Lecture Notes in Physics* **9**. World Scientific, Teaneck, NJ. [MR1026102](#)
- [31] MIOLANE, L. (2017). Fundamental limits of symmetric low-rank matrix estimation: The non-symmetric case. ArXiv e-prints.
- [32] MONTANARI, A., REICHMAN, D. and ZEITOUNI, O. (2017). On the limitation of spectral methods: From the Gaussian hidden clique problem to rank one perturbations of Gaussian tensors. *IEEE Trans. Inform. Theory* **63** 1572–1579. [MR3625981](#)
- [33] MONTANARI, A. and RICHARD, E. (2014). A statistical model for tensor pca. In *Neural Information Processing Systems*. 2897–2905.
- [34] MONTANARI, A. and RICHARD, E. (2016). Non-negative principal component analysis: Message passing algorithms and sharp asymptotics. *IEEE Trans. Inform. Theory* **62** 1458–1484. [MR3472260](#)
- [35] ONATSKI, A., MOREIRA, M. J. and HALLIN, M. (2013). Asymptotic power of sphericity tests for high-dimensional data. *Ann. Statist.* **41** 1204–1231. [MR3113808](#)
- [36] PANCHENKO, D. (2013). The Parisi ultrametricity conjecture. *Ann. of Math. (2)* **177** 383–393. [MR2999044](#)
- [37] PANCHENKO, D. (2013). *The Sherrington–Kirkpatrick Model. Springer Monographs in Mathematics*. Springer, New York. [MR3052333](#)
- [38] PANCHENKO, D. (2014). The Parisi formula for mixed p -spin models. *Ann. Probab.* **42** 946–958. [MR3189062](#)
- [39] PANCHENKO, D. (2015). The free energy in a multi-species Sherrington–Kirkpatrick model. *Ann. Probab.* **43** 3494–3513. [MR3433586](#)
- [40] PANCHENKO, D. (2018). Free energy in the mixed p -spin models with vector spins. *Ann. Probab.* **46** 865–896. [MR3773376](#)
- [41] PÉCHÉ, S. (2006). The largest eigenvalue of small rank perturbations of Hermitian random matrices. *Probab. Theory Related Fields* **134** 127–173. [MR2221787](#)
- [42] PERRY, A., WEIN, A. S. and BANDEIRA, A. (2017). Statistical limits of spiked tensor models. ArXiv e-prints.
- [43] 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](#)
- [44] REEVES, G. and PFISTER, H. D. (2016). The replica-symmetric prediction for compressed sensing with gaussian matrices is exact. ArXiv e-prints.
- [45] TALAGRAND, M. (2006). The Parisi formula. *Ann. of Math. (2)* **163** 221–263. [MR2195134](#)
- [46] TALAGRAND, M. (2011). *Mean Field Models for Spin Glasses. Volume I. Basic Examples*. Springer, Berlin. [MR2731561](#)
- [47] TALAGRAND, M. (2011). *Mean Field Models for Spin Glasses. Volume II. Advanced Replica-Symmetry and Low Temperature*. Springer, Heidelberg. [MR3024566](#)

DISTANCE MULTIVARIANCE: NEW DEPENDENCE MEASURES FOR RANDOM VECTORS

BY BJÖRN BÖTTCHER, MARTIN KELLER-RESSEL¹ AND RENÉ L. SCHILLING

TU Dresden

We introduce two new measures for the dependence of $n \geq 2$ random variables: *distance multivariate* and *total distance multivariate*. Both measures are based on the weighted L^2 -distance of quantities related to the characteristic functions of the underlying random variables. These extend distance covariance (introduced by Székely, Rizzo and Bakirov) from pairs of random variables to n -tuples of random variables. We show that total distance multivariate can be used to detect the independence of n random variables and has a simple finite-sample representation in terms of distance matrices of the sample points, where distance is measured by a continuous negative definite function. Under some mild moment conditions, this leads to a test for independence of multiple random vectors which is consistent against all alternatives.

REFERENCES

- [1] BAKIROV, N. K. and SZÉKELY, G. J. (2010). Brownian covariance central limit theorem for stationary sequences. *Teor. Veroyatn. Primen.* **55** 462–488. All references to this paper are based on the english translation (Nail K. Bakirov and Gábor J. Székely. Brownian covariance and central limit theorem for stationary sequences. *Theory of Probability & Its Applications*, 55(3):371–394, 2011). [MR2768533](#)
- [2] BARTHOLOMÄUS, J. (2018). Private communication, TU Dresden.
- [3] BERG, C. and FORST, G. (1975). *Potential Theory on Locally Compact Abelian Groups. Ergebnisse der Mathematik und ihrer Grenzgebiete* **87**. Springer, New York. [MR0481057](#)
- [4] BERRETT, T. B. and SAMWORTH, R. J. (2017). Nonparametric independence testing via mutual information. Available at [arXiv:1711.06642v1](#).
- [5] BERSCHNEIDER, G. and BÖTTCHER, B. (2018). On complex Gaussian random fields, Gaussian quadratic forms and sample distance multivariate. Available at [arXiv:1808.07280v1](#).
- [6] BÖTTCHER, B. (2017). Dependence structures—estimation and visualization using distance multivariate. Available at [arXiv:1712.06532v1](#).
- [7] BÖTTCHER, B. (2017). multivariate: Measuring multivariate dependence using distance multivariate. R package version 1.0.5.
- [8] BÖTTCHER, B., KELLER-RESSEL, M. and SCHILLING, R. L. (2018). Detecting independence of random vectors: Generalized distance covariance and Gaussian covariance. *Mod. Stoch. Theory Appl.* **5** 353–383.
- [9] BÖTTCHER, B., KELLER-RESSEL, M. and SCHILLING, R. L. (2019). Supplement to “Distance multivariate: New dependence measures for random vectors.” DOI:[10.1214/18-AOS1764SUPP](#).

MSC2010 subject classifications. Primary 62H20; secondary 60E10, 62G10, 62G15, 62G20.

Key words and phrases. Dependence measure, stochastic independence, negative definite function, characteristic function, Gaussian random field, statistical test of independence.

- [10] CHOW, Y. S. and TEICHER, H. (1997). *Probability Theory: Independence, Interchangeability, Martingales*, 3rd ed. *Springer Texts in Statistics*. Springer, New York. [MR1476912](#)
- [11] COMON, P. (1994). Independent component analysis, a new concept? *Signal Process.* **36** 287–314.
- [12] COPE, L. (2009). Discussion of: Brownian distance covariance [MR2752127]. *Ann. Appl. Stat.* **3** 1279–1281. [MR2752130](#)
- [13] CSÖRGÖ, S. (1985). Testing for independence by the empirical characteristic function. *J. Multivariate Anal.* **16** 290–299. [MR0793494](#)
- [14] FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications. Vol. II.* 2nd ed. Wiley, New York. [MR0270403](#)
- [15] GRETTON, A., BOUSQUET, O., SMOLA, A. and SCHÖLKOPF, B. (2005). Measuring statistical dependence with Hilbert–Schmidt norms. In *Algorithmic Learning Theory. Lecture Notes in Computer Science* **3734** 63–77. Springer, Berlin. [MR2255909](#)
- [16] JACOB, N. (2001). *Pseudo Differential Operators and Markov Processes. Vol. I: Fourier Analysis and Semigroups*. Imperial College Press, London. [MR1873235](#)
- [17] JIN, Z. and MATTESON, D. S. (2018). Generalizing distance covariance to measure and test multivariate mutual dependence via complete and incomplete V-statistics. *J. Multivariate Anal.* **168** 304–322. Available at [arXiv:1709.02532](#). All references to this paper are based on the arxiv version 1709.02532v5 which differs from the published version. [MR3858367](#)
- [18] 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](#)
- [19] SATO, K. (1999). *Lévy Processes and Infinitely Divisible Distributions. Cambridge Studies in Advanced Mathematics* **68**. Cambridge Univ. Press, Cambridge. Translated from the 1990 Japanese original, revised by the author. [MR1739520](#)
- [20] SCHILLING, R. L. (2017). *Measures, Integrals and Martingales*, 2nd ed. Cambridge Univ. Press, Cambridge. [MR3644418](#)
- [21] 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](#)
- [22] SZÉKELY, G. J. and BAKIROV, N. K. (2003). Extremal probabilities for Gaussian quadratic forms. *Probab. Theory Related Fields* **126** 184–202. [MR1990053](#)
- [23] SZÉKELY, G. J. and RIZZO, M. L. (2009). Brownian distance covariance. *Ann. Appl. Stat.* **3** 1236–1265. [MR2752127](#)
- [24] SZÉKELY, G. J. and RIZZO, M. L. (2009). Rejoinder: Brownian distance covariance. *Ann. Appl. Stat.* **3** 1303–1308. [MR2752135](#)
- [25] 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](#)
- [26] 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](#)

A UNIFIED TREATMENT OF MULTIPLE TESTING WITH PRIOR KNOWLEDGE USING THE P-FILTER¹

BY AADITYA K. RAMDAS^{*}, RINA F. BARBER[†],
MARTIN J. WAINWRIGHT[‡] AND MICHAEL I. JORDAN[‡]

Carnegie Mellon University^{}, University of Chicago[†] and
University of California, Berkeley[‡]*

There is a significant literature on methods for incorporating knowledge into multiple testing procedures so as to improve their power and precision. Some common forms of prior knowledge include (a) beliefs about which hypotheses are null, modeled by nonuniform prior weights; (b) differing importances of hypotheses, modeled by differing penalties for false discoveries; (c) multiple arbitrary partitions of the hypotheses into (possibly overlapping) groups and (d) knowledge of independence, positive or arbitrary dependence between hypotheses or groups, suggesting the use of more aggressive or conservative procedures. We present a unified algorithmic framework called *p-filter* for global null testing and false discovery rate (FDR) control that allows the scientist to incorporate all four types of prior knowledge (a)–(d) simultaneously, recovering a variety of known algorithms as special cases.

REFERENCES

- [1] BARBER, R. F. and RAMDAS, A. (2017). The *p*-filter: Multilayer false discovery rate control for grouped hypotheses. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1247–1268. [MR3689317](#)
- [2] BENJAMINI, Y. and BOGOMOLOV, M. (2014). Selective inference on multiple families of hypotheses. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 297–318. [MR3153943](#)
- [3] 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](#)
- [4] BENJAMINI, Y. and HOCHBERG, Y. (1997). Multiple hypotheses testing with weights. *Scand. J. Stat.* **24** 407–418. [MR1481424](#)
- [5] BENJAMINI, Y. and HOCHBERG, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Stat.* **25** 60–83.
- [6] BENJAMINI, Y., KRIEGER, A. M. and YEKUTIELI, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* **93** 491–507. [MR2261438](#)
- [7] BENJAMINI, Y. and LIU, W. (1999). A distribution-free multiple test procedure that controls the false discovery rate. Technical report, Tel Aviv Univ.
- [8] BENJAMINI, Y. and LIU, W. (1999). A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *J. Statist. Plann. Inference* **82** 163–170. [MR1736441](#)

MSC2010 subject classifications. Primary 62J15, 60G10; secondary 62F03.

Key words and phrases. Multiple testing, false discovery rate, prior knowledge, Simes, Benjamini–Hochberg–Yekutieli, adaptivity, group FDR.

- [9] BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. [MR1869245](#)
- [10] BLANCHARD, G. and ROQUAIN, E. (2008). Two simple sufficient conditions for FDR control. *Electron. J. Stat.* **2** 963–992. [MR2448601](#)
- [11] BLANCHARD, G. and ROQUAIN, É. (2009). Adaptive false discovery rate control under independence and dependence. *J. Mach. Learn. Res.* **10** 2837–2871. [MR2579914](#)
- [12] BOGOMOLOV, M., PETERSON, C. B., BENJAMINI, Y. and SABATTI, C. (2017). Testing hypotheses on a tree: New error rates and controlling strategies. Preprint. Available at [arXiv:1705.07529](#).
- [13] BRZYSKI, D., PETERSON, C. B., SOBCZYK, P., CANDÈS, E. J., BOGDAN, M. and SABATTI, C. (2017). Controlling the rate of GWAS false discoveries. *Genetics* **205** 61–75.
- [14] GABRIEL, K. R. (1969). Simultaneous test procedures—Some theory of multiple comparisons. *Ann. Math. Stat.* **40** 224–250. [MR0240931](#)
- [15] GAVRILOV, Y., BENJAMINI, Y. and SARKAR, S. K. (2009). An adaptive step-down procedure with proven FDR control under independence. *Ann. Statist.* **37** 619–629. [MR2502645](#)
- [16] GENOVESE, C. R., ROEDER, K. and WASSERMAN, L. (2006). False discovery control with p -value weighting. *Biometrika* **93** 509–524. [MR2261439](#)
- [17] HEARD, N. A. and RUBIN-DELANCHY, P. (2018). Choosing between methods of combining p -values. *Biometrika* **105** 239–246. [MR3768879](#)
- [18] HOCHBERG, Y. and BENJAMINI, Y. (1990). More powerful procedures for multiple significance testing. *Stat. Med.* **9** 811–818.
- [19] HOCHBERG, Y. and LIBERMAN, U. (1994). An extended Simes’ test. *Statist. Probab. Lett.* **21** 101–105. [MR1309882](#)
- [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](#)
- [21] JAVANMARD, A. and MONTANARI, A. (2018). Online rules for control of false discovery rate and false discovery exceedance. *Ann. Statist.* **46** 526–554. [MR3782376](#)
- [22] KATSEVICH, E. and RAMDAS, A. (2018). Towards “simultaneous selective inference”: Post-hoc bounds on the false discovery proportion. Preprint. Available at [arXiv:1803.06790](#).
- [23] KATSEVICH, E. and SABATTI, C. (2019). Multilayer knockoff filter: Controlled variable selection at multiple resolutions. *Ann. Appl. Stat.* **13** 1–33. [MR3937419](#)
- [24] KATSEVICH, E., SABATTI, C. and BOGOMOLOV, M. (2018). Controlling FDR while highlighting distinct discoveries. Preprint. Available at [arXiv:1809.01792](#).
- [25] LEHMANN, E. L. (1966). Some concepts of dependence. *Ann. Math. Stat.* **37** 1137–1153. [MR0202228](#)
- [26] PETERSON, C. B., BOGOMOLOV, M., BENJAMINI, Y. and SABATTI, C. (2016). Many phenotypes without many false discoveries: Error controlling strategies for multitrait association studies. *Genet. Epidemiol.* **40** 45–56.
- [27] RAMDAS, A., CHEN, J., WAINWRIGHT, M. J. and JORDAN, M. I. (2017). QuTE: Decentralized multiple testing on sensor networks with false discovery rate control. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)* 6415–6421. IEEE, New York.
- [28] RAMDAS, A., CHEN, J., WAINWRIGHT, M. J. and JORDAN, M. I. (2019). A sequential algorithm for false discovery rate control on directed acyclic graphs. *Biometrika* **106** 69–86. [MR3912384](#)
- [29] RAMDAS, A., YANG, F., WAINWRIGHT, M. J. and JORDAN, M. I. (2017). Online control of the false discovery rate with decaying memory. In *Advances in Neural Information Processing Systems* 5655–5664.
- [30] RAMDAS, A., ZRNIC, T., WAINWRIGHT, M. J. and JORDAN, M. I. (2018). SAFFRON: An adaptive algorithm for online control of the false discovery rate. In *Proceedings of the 35th International Conference on Machine Learning* 4286–4294.

- [31] RAMDAS, A. K, BARBER, R. F, WAINWRIGHT, M. J and JORDAN, M. I (2019). Supplement to “A unified treatment of multiple testing with prior knowledge using the p-filter.” DOI:[10.1214/18-AOS1765SUPP](https://doi.org/10.1214/18-AOS1765SUPP).
- [32] ROMANO, J. P., SHAIKH, A. and WOLF, M. (2011). Consonance and the closure method in multiple testing. *Int. J. Biostat.* **7** Art. 12, 27. [MR2775079](#)
- [33] ROMANO, J. P. and SHAIKH, A. M. (2006). On stepdown control of the false discovery proportion. In *Optimality. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **49** 33–50. IMS, Beachwood, OH. [MR2337829](#)
- [34] SARKAR, S. K. (2008). On methods controlling the false discovery rate. *Sankhyā* **70** 135–168. [MR2551809](#)
- [35] SARKAR, S. K. (2008). Two-stage stepup procedures controlling FDR. *J. Statist. Plann. Inference* **138** 1072–1084. [MR2384506](#)
- [36] SARKAR, T. K. (1969). Some lower bounds of reliability. Technical report, Stanford Univ. [MR2618992](#)
- [37] SEEGER, P. (1968). A note on a method for the analysis of significances en masse. *Technometrics* **10** 586–593.
- [38] SIMES, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73** 751–754. [MR0897872](#)
- [39] SONNEMANN, E. (1982). Allgemeine Lösungen multipler Testprobleme. Institut für Mathematische Statistik und Versicherungslehre, Univ. Bern.
- [40] SONNEMANN, E. (2008). General solutions to multiple testing problems. *Biom. J.* **50** 641–656. [MR2542333](#)
- [41] SONNEMANN, E. and FINNER, H. (1988). Vollständigkeitssätze für multiple testprobleme. In *Multiple Hypothesenprüfung/Multiple Hypotheses Testing* 121–135. Springer, Berlin.
- [42] STOREY, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 479–498. [MR1924302](#)
- [43] 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](#)
- [44] STOUTER, S. A., SUCHMAN, E. A., DEVINNEY, L. C., STAR, S. A. and WILLIAMS, R. M. JR. (1949). *The American Soldier: Adjustment During Army Life*, Vol. 1. *Studies in Social Psychology in World War II*.
- [45] TUKEY, J. (1953). *The Problem of Multiple Comparisons: Introduction and Parts A, B, and C*. Princeton Univ., Princeton, NJ.
- [46] TUKEY, J. W. (1994). *The Collected Works of John W. Tukey. Vol. VIII: Multiple Comparisons: 1948–1983*. CRC Press, New York. [MR1263027](#)
- [47] VOVK, V. and WANG, R. (2018). Combining p-values via averaging. Preprint. Available at [arXiv:1212.4966v4](https://arxiv.org/abs/1212.4966v4).

EXACT LOWER BOUNDS FOR THE AGNOSTIC PROBABLY-APPROXIMATELY-CORRECT (PAC) MACHINE LEARNING MODEL

BY ARYEH KONTOROVICH AND IOSIF PINELIS

Ben-Gurion University and Michigan Technological University

We provide an exact nonasymptotic lower bound on the minimax expected excess risk (EER) in the agnostic probably-approximately-correct (PAC) machine learning classification model and identify minimax learning algorithms as certain maximally symmetric and minimally randomized “voting” procedures. Based on this result, an exact asymptotic lower bound on the minimax EER is provided. This bound is of the simple form $c_\infty/\sqrt{v}$ as $v \rightarrow \infty$, where $c_\infty = 0.16997\dots$ is a universal constant, $v = m/d$, m is the size of the training sample and d is the Vapnik–Chervonenkis dimension of the hypothesis class. It is shown that the differences between these asymptotic and nonasymptotic bounds, as well as the differences between these two bounds and the maximum EER of any learning algorithms that minimize the empirical risk, are asymptotically negligible, and all these differences are due to ties in the mentioned “voting” procedures. A few easy to compute nonasymptotic lower bounds on the minimax EER are also obtained, which are shown to be close to the exact asymptotic lower bound $c_\infty/\sqrt{v}$ even for rather small values of the ratio $v = m/d$. As an application of these results, we substantially improve existing lower bounds on the tail probability of the excess risk. Among the tools used are Bayes estimation and apparently new identities and inequalities for binomial distributions.

REFERENCES

- [1] ANTHONY, M. and BARTLETT, P. L. (1999). *Neural Network Learning: Theoretical Foundations*. Cambridge Univ. Press, Cambridge. [MR1741038](#)
- [2] AUDIBERT, J.-Y. (2009). Fast learning rates in statistical inference through aggregation. *Ann. Statist.* **37** 1591–1646. [MR2533466](#)
- [3] BEREND, D. and KONTOROVICH, A. (2015). A finite sample analysis of the naive Bayes classifier. *J. Mach. Learn. Res.* **16** 1519–1545. [MR3417789](#)
- [4] BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities. A Nonasymptotic Theory of Independence*. Oxford Univ. Press, Oxford. [MR3185193](#)
- [5] DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition. Applications of Mathematics (New York)* **31**. Springer, New York. [MR1383093](#)
- [6] DEVROYE, L. and LUGOSI, G. (1995). Lower bounds in pattern recognition and learning. *Pattern Recognit.* **28** 1011–1018.

MSC2010 subject classifications. Primary 68T05, 62C20, 62C10, 62C12, 62G20, 62H30; secondary 62G10, 62C20, 91A35, 60C05.

Key words and phrases. PAC learning theory, classification, generalization error, minimax decision rules, Bayes decision rules, empirical estimators, binomial distribution.

- [7] FERGUSON, T. S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. Probability and Mathematical Statistics **1**. Academic Press, New York. [MR0215390](#)
- [8] HARDY, G. H., LITTLEWOOD, J. E. and PÓLYA, G. (1988). *Inequalities*. Cambridge Univ. Press, Cambridge. [MR0944909](#)
- [9] HAUSSLER, D. (1992). Decision-theoretic generalizations of the PAC model for neural net and other learning applications. *Inform. and Comput.* **100** 78–150. [MR1175977](#)
- [10] HAUSSLER, D. (1995). Sphere packing numbers for subsets of the Boolean n -cube with bounded Vapnik–Chervonenkis dimension. *J. Combin. Theory Ser. A* **69** 217–232. [MR1313896](#)
- [11] KEARNS, M. J. and SCHAPIRE, R. E. (1994). Efficient distribution-free learning of probabilistic concepts. *J. Comput. System Sci.* **48** 464–497. [MR1279411](#)
- [12] KEARNS, M. J., SCHAPIRE, R. E. and SELLIE, L. (1994). Toward efficient agnostic learning. *Mach. Learn.* **17** 115–141.
- [13] KINGMAN, J. F. C. (1961). A convexity property of positive matrices. *Quart. J. Math. Oxford Ser. (2)* **12** 283–284. [MR0138632](#)
- [14] KONTOROVICH, A., SABATO, S. and URNER, R. (2016). Active nearest-neighbor learning in metric spaces. In *NIPS 2016*. Available at [arXiv:1605.06792](#).
- [15] LONG, P. M. (1999). The complexity of learning according to two models of a drifting environment. *Mach. Learn.* **37** 337–354.
- [16] PINELIS, I. (2016). Optimal binomial, Poisson, and normal left-tail domination for sums of nonnegative random variables. *Electron. J. Probab.* **21** Paper No. 20. [MR3485362](#)
- [17] PINELIS, I. F. (1991). Criteria for complete determinateness for concave-convex games. *Mat. Zametki* **49** 73–76, 159. [MR1110310](#)
- [18] PINELIS, I. S. and UTEV, S. A. (1989). Sharp exponential estimates for sums of independent random variables. *Teor. Veroyatn. Primen.* **34** 384–390. [MR1005745](#)
- [19] SHALEV-SHWARTZ, S. and BEN-DAVID, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge Univ. Press, Cambridge.
- [20] SIMON, H. U. (1996). General bounds on the number of examples needed for learning probabilistic concepts. *J. Comput. System Sci.* **52** 239–254. [MR1393992](#)
- [21] SION, M. (1958). On general minimax theorems. *Pacific J. Math.* **8** 171–176. [MR0097026](#)
- [22] TALAGRAND, M. (1994). Sharper bounds for Gaussian and empirical processes. *Ann. Probab.* **22** 28–76. [MR1258865](#)
- [23] V. NEUMANN, J. (1928). Zur Theorie der Gesellschaftsspiele. *Math. Ann.* **100** 295–320. [MR1512486](#)
- [24] VALIANT, L. G. (1984). A theory of the learnable. *Commun. ACM* **27** 1134–1142.
- [25] VAPNIK, V. N. and ČERVONENKIS, A. JA. (1971). The uniform convergence of frequencies of the appearance of events to their probabilities. *Teor. Veroyatn. Primen.* **16** 264–279. [MR0288823](#)

EIGENVALUE DISTRIBUTIONS OF VARIANCE COMPONENTS ESTIMATORS IN HIGH-DIMENSIONAL RANDOM EFFECTS MODELS

BY ZHOU FAN¹ AND IAIN M. JOHNSTONE²

Yale University and Stanford University

We study the spectra of MANOVA estimators for variance component covariance matrices in multivariate random effects models. When the dimensionality of the observations is large and comparable to the number of realizations of each random effect, we show that the empirical spectra of such estimators are well approximated by deterministic laws. The Stieltjes transforms of these laws are characterized by systems of fixed-point equations, which are numerically solvable by a simple iterative procedure. Our proof uses operator-valued free probability theory, and we establish a general asymptotic freeness result for families of rectangular orthogonally invariant random matrices, which is of independent interest. Our work is motivated in part by the estimation of components of covariance between multiple phenotypic traits in quantitative genetics, and we specialize our results to common experimental designs that arise in this application.

REFERENCES

- BAI, Z., CHEN, J. and YAO, J. (2010). On estimation of the population spectral distribution from a high-dimensional sample covariance matrix. *Aust. N. Z. J. Stat.* **52** 423–437. [MR2791528](#)
- BAI, Z. D. and SILVERSTEIN, J. W. (2004). CLT for linear spectral statistics of large-dimensional sample covariance matrices. *Ann. Probab.* **32** 553–605. [MR2040792](#)
- BAI, Z. and YAO, J. (2012). On sample eigenvalues in a generalized spiked population model. *J. Multivariate Anal.* **106** 167–177. [MR2887686](#)
- 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](#)
- BAIK, J. and SILVERSTEIN, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *J. Multivariate Anal.* **97** 1382–1408. [MR2279680](#)
- BARTON, N. H. (1990). Pleiotropic models of quantitative variation. *Genetics* **124** 773–782.
- BENAYCH-GEORGES, F. (2009). Rectangular random matrices, related convolution. *Probab. Theory Related Fields* **144** 471–515. [MR2496440](#)
- 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](#)
- BLOWS, M. W. (2007). A tale of two matrices: Multivariate approaches in evolutionary biology. *J. Evol. Biol.* **20** 1–8.
- BLOWS, M. W. and MCGUIGAN, K. (2015). The distribution of genetic variance across phenotypic space and the response to selection. *Mol. Ecol.* **24** 2056–2072.

MSC2010 subject classifications. 62E20.

Key words and phrases. Random matrix theory, free probability, deterministic equivalents, covariance estimation.

- BLOWS, M. W., ALLEN, S. L., COLLET, J. M., CHENOWETH, S. F. and MCGUIGAN, K. (2015). The phenome-wide distribution of genetic variance. *Amer. Nat.* **186** 15–30.
- COLLINS, B. (2003). Moments and cumulants of polynomial random variables on unitary groups, the Itzykson–Zuber integral, and free probability. *Int. Math. Res. Not.* **2003** 953–982. [MR1959915](#)
- COLLINS, B. and ŚNIADY, P. (2006). Integration with respect to the Haar measure on unitary, orthogonal and symplectic group. *Comm. Math. Phys.* **264** 773–795. [MR2217291](#)
- COUILLET, R., DEBBAH, M. and SILVERSTEIN, J. W. (2011). A deterministic equivalent for the analysis of correlated MIMO multiple access channels. *IEEE Trans. Inform. Theory* **57** 3493–3514. [MR2817033](#)
- DOBRIBAN, E. (2015). Efficient computation of limit spectra of sample covariance matrices. *Random Matrices Theory Appl.* **4** 1550019, 36. [MR3418848](#)
- DOBRIBAN, E. (2017). Sharp detection in PCA under correlations: All eigenvalues matter. *Ann. Statist.* **45** 1810–1833. [MR3670197](#)
- DUPUY, F. and LOUBATON, P. (2011). On the capacity achieving covariance matrix for frequency selective MIMO channels using the asymptotic approach. *IEEE Trans. Inform. Theory* **57** 5737–5753. [MR2857932](#)
- DYKEMA, K. (1993). On certain free product factors via an extended matrix model. *J. Funct. Anal.* **112** 31–60. [MR1207936](#)
- EL KAROUI, N. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Ann. Statist.* **36** 2757–2790. [MR2485012](#)
- FAN, Z. and JOHNSTONE, I. M. (2017). Tracy–Widom at each edge of real covariance estimators. Unpublished manuscript.
- FAN, Z. and JOHNSTONE, I. M. (2019). Supplement to “Eigenvalue distributions of variance components estimators in high-dimensional random effects models.” DOI:[10.1214/18-AOS1767SUPP](#).
- FAN, Z., JOHNSTONE, I. M. and SUN, Y. (2018). Spiked covariances and principal components analysis in high-dimensional random effects models. Available at [arXiv:1806.09529](#).
- FISHER, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.* **52** 399–433.
- HACHEM, W., LOUBATON, P. and NAJIM, J. (2007). Deterministic equivalents for certain functionals of large random matrices. *Ann. Appl. Probab.* **17** 875–930. [MR2326235](#)
- HIAI, F. and PETZ, D. (2000). Asymptotic freeness almost everywhere for random matrices. *Acta Sci. Math. (Szeged)* **66** 809–834. [MR1804226](#)
- HINE, E., MCGUIGAN, K. and BLOWS, M. W. (2014). Evolutionary constraints in high-dimensional trait sets. *Amer. Nat.* **184** 119–131.
- JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. [MR1863961](#)
- KIRKPATRICK, M. (2009). Patterns of quantitative genetic variation in multiple dimensions. *Genetica* **136** 271–284.
- LAMOTTE, L. R. (1973). Quadratic estimation of variance components. *Biometrics* **29** 311–330. [MR0329142](#)
- LANDE, R. (1979). Quantitative genetic analysis of multivariate evolution, applied to brain: Body size allometry. *Evolution* **33** 402–416.
- LANDE, R. and ARNOLD, S. J. (1983). The measurement of selection on correlated characters. *Evolution* **37** 1210–1226.
- LEDOIT, O. and PÉCHÉ, S. (2011). Eigenvectors of some large sample covariance matrix ensembles. *Probab. Theory Related Fields* **151** 233–264. [MR2834718](#)
- LEDOIT, O. and WOLF, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Ann. Statist.* **40** 1024–1060. [MR2985942](#)
- LOH, P.-R., TUCKER, G., BULIK-SULLIVAN, B. K., VILHJÁLMSOHN, B. J., FINUCANE, H. K., SALEM, R. M., CHASMAN, D. I., RIDKER, P. M., NEALE, B. M. et al. (2015). Efficient

- Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47** 284–290.
- LUSH, J. L. (1937). *Animal Breeding Plans*. Iowa State College Press, Ames, IA.
- LYNCH, M. and WALSH, B. (1998). *Genetics and Analysis of Quantitative Traits* **1**. Sinauer, Sunderland, MA.
- MARČENKO, V. A. and PASTUR, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Sb. Math.* **1** 457–483.
- MCGUIGAN, K., COLLET, J. M., MCGRAW, E. A., YE YIXIN, H., ALLEN, S. L., CHENOWETH, S. F. and BLOWS, M. W. (2014). The nature and extent of mutational pleiotropy in gene expression of male drosophila serrata. *Genetics* **196** 911–921.
- MESTRE, X. (2008). Improved estimation of eigenvalues and eigenvectors of covariance matrices using their sample estimates. *IEEE Trans. Inform. Theory* **54** 5113–5129. [MR2589886](#)
- MOUSTAKAS, A. L. and SIMON, S. H. (2007). On the outage capacity of correlated multiple-path MIMO channels. *IEEE Trans. Inform. Theory* **53** 3887–3903. [MR2446543](#)
- NICA, A., SHLYAKHTENKO, D. and SPEICHER, R. (2002). Operator-valued distributions. I. Characterizations of freeness. *Int. Math. Res. Not.* **2002** 1509–1538. [MR1907203](#)
- NICA, A. and SPEICHER, R. (2006). *Lectures on the Combinatorics of Free Probability*. London Mathematical Society Lecture Note Series **335**. Cambridge Univ. Press, Cambridge. [MR2266879](#)
- ONATSKI, A., MOREIRA, M. J. and HALLIN, M. (2014). Signal detection in high dimension: The multispiked case. *Ann. Statist.* **42** 225–254. [MR3189485](#)
- PAUL, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17** 1617–1642. [MR2399865](#)
- PAUL, D. and AUE, A. (2014). Random matrix theory in statistics: A review. *J. Statist. Plann. Inference* **150** 1–29. [MR3206718](#)
- PHILLIPS, P. C. and ARNOLD, S. J. (1989). Visualizing multivariate selection. *Evolution* 1209–1222.
- RAO, C. R. (1971). Minimum variance quadratic unbiased estimation of variance components. *J. Multivariate Anal.* **1** 445–456. [MR0301870](#)
- RAO, C. R. (1972). Estimation of variance and covariance components in linear models. *J. Amer. Statist. Assoc.* **67** 112–115. [MR0314185](#)
- RAO, N. R., MINGO, J. A., SPEICHER, R. and EDELMAN, A. (2008). Statistical eigen-inference from large Wishart matrices. *Ann. Statist.* **36** 2850–2885. [MR2485015](#)
- ROBERTSON, A. (1959a). The sampling variance of the genetic correlation coefficient. *Biometrics* **15** 469–485. [MR0107568](#)
- ROBERTSON, A. (1959b). The sampling variance of the genetic correlation coefficient. *Biometrics* **15** 469–485. [MR0107568](#)
- SEARLE, S. R., CASELLA, G. and MCCULLOCH, C. E. (2006). *Variance Components*. Wiley Series in Probability and Statistics. Wiley-Interscience, Hoboken, NJ. [MR2298115](#)
- SILVERSTEIN, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. *J. Multivariate Anal.* **55** 331–339. [MR1370408](#)
- SOSHNIKOV, A. (2002). A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices. *J. Stat. Phys.* **108** 1033–1056. [MR1933444](#)
- SPEICHER, R. (1998). Combinatorial theory of the free product with amalgamation and operator-valued free probability theory. *Mem. Amer. Math. Soc.* **132** x+88. [MR1407898](#)
- SPEICHER, R. and VARGAS, C. (2012). Free deterministic equivalents, rectangular random matrix models, and operator-valued free probability theory. *Random Matrices Theory Appl.* **1** 1150008, 26. [MR2934714](#)
- VOICULESCU, D. (1991). Limit laws for random matrices and free products. *Invent. Math.* **104** 201–220. [MR1094052](#)
- VOICULESCU, D. (1995). Operations on certain non-commutative operator-valued random variables. Recent advances in operator algebras (Orléans, 1992). *Astérisque* **232** 243–275. [MR1372537](#)

- VOICULESCU, D. (1998). A strengthened asymptotic freeness result for random matrices with applications to free entropy. *Int. Math. Res. Not.* **1998** 41–63. [MR1601878](#)
- VOICULESCU, D. V., DYKEMA, K. J. and NICA, A. (1992). *Free Random Variables: A Noncommutative Probability Approach to Free Products with Applications to Random Matrices, Operator Algebras and Harmonic Analysis on Free Groups*. CRM Monograph Series **1**. Amer. Math. Soc., Providence, RI. [MR1217253](#)
- WALSH, B. and BLOWS, M. W. (2009). Abundant genetic variation + strong selection = multivariate genetic constraints: A geometric view of adaptation. *Annu. Rev. Ecol. Evol. Syst.* **40** 41–59.
- WIGNER, E. P. (1955). Characteristic vectors of bordered matrices with infinite dimensions. *Ann. of Math.* (2) **62** 548–564. [MR0077805](#)
- WRIGHT, S. (1935). The analysis of variance and the correlations between relatives with respect to deviations from an optimum. *J. Genet.* **30** 243–256.
- YANG, M., GOLDSTEIN, H., BROWNE, W. and WOODHOUSE, G. (2002). Multivariate multilevel analyses of examination results. *J. Roy. Statist. Soc. Ser. A* **165** 137–153. [MR1909740](#)
- YANG, J., LEE, S. H., GODDARD, M. E. and VISSCHER, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88** 76–82.
- YAO, J., ZHENG, S. and BAI, Z. (2015). *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. Cambridge Series in Statistical and Probabilistic Mathematics **39**. Cambridge Univ. Press, New York. [MR3468554](#)
- ZHANG, L. (2006). Spectral analysis of large dimensional random matrices. Ph.D. thesis, National Univ. Singapore.

TEST FOR HIGH-DIMENSIONAL CORRELATION MATRICES

BY SHURONG ZHENG^{*,1}, GUANGHUI CHENG[†], JIANHUA GUO^{*,2} AND
 HONGTU ZHU^{‡,3}

Northeast Normal University^{}, Guangzhou University[†] and University of North
 Carolina, Chapel Hill[‡]*

Testing correlation structures has attracted extensive attention in the literature due to both its importance in real applications and several major theoretical challenges. The aim of this paper is to develop a general framework of testing correlation structures for the one, two and multiple sample testing problems under a high-dimensional setting when both the sample size and data dimension go to infinity. Our test statistics are designed to deal with both the dense and sparse alternatives. We systematically investigate the asymptotic null distribution, power function and unbiasedness of each test statistic. Theoretically, we make great efforts to deal with the nonindependency of all random matrices of the sample correlation matrices. We use simulation studies and real data analysis to illustrate the versatility and practicability of our test statistics.

REFERENCES

- AITKIN, M. A. (1969). Some tests for correlation matrices. *Biometrika* **56** 443–446.
- ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. [MR1990662](#)
- BAI, Z. D. and SILVERSTEIN, J. W. (2004). CLT for linear spectral statistics of large-dimensional sample covariance matrices. *Ann. Probab.* **32** 553–605. [MR2040792](#)
- BARTLETT, M. S. and RAJALAKSHMAN, D. V. (1953). Goodness of fit tests for simultaneous autoregressive series. *J. Roy. Statist. Soc. Ser. B* **15** 107–124. [MR0058181](#)
- BODWIN, K., ZHANG, K. and NOBEL, A. (2018). A testing based approach to the discovery of differentially correlated variable sets. *Ann. Appl. Stat.* **12** 1180–1203. [MR3834299](#)
- BROWNE, M. W. (1978). The likelihood ratio test for the equality of correlation matrices. *Br. J. Math. Stat. Psychol.* **31** 209–217. [MR0553468](#)
- CAI, T. T. (2017). Global testing and large-scale multiple testing for high-dimensional covariance structures. *Ann. Rev. Stat. Appl.* **4** 423–446.
- CAI, T. T. and JIANG, T. (2011). Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *Ann. Statist.* **39** 1496–1525. [MR2850210](#)
- 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](#)
- CAI, T. T. and MA, Z. (2013). Optimal hypothesis testing for high dimensional covariance matrices. *Bernoulli* **19** 2359–2388. [MR3160557](#)

MSC2010 subject classifications. Primary 62H15; secondary 62H10.

Key words and phrases. Dense alternatives, global testing, sample correlation matrices, sparse alternatives.

- CAI, T. T. and ZHANG, A. (2016). Inference for high-dimensional differential correlation matrices. *J. Multivariate Anal.* **143** 107–126. [MR3431422](#)
- CHEN, S. X., ZHANG, L.-X. and ZHONG, P.-S. (2010). Tests for high-dimensional covariance matrices. *J. Amer. Statist. Assoc.* **105** 810–819. [MR2724863](#)
- COLE, N. (1968). The likelihood ratio test of the equality of correlation matrices. 1968-65: The L. L. Thurstone Psychometric Laboratory. Univ. North Carolina, Chapel Hill, NC.
- DEBASHIS, P. and ALEXANDER, A. (2014). Random matrix theory in statistics: A review. *J. Statist. Plann. Inference* **150** 1–29. [MR3206718](#)
- GAO, J., HAN, X., PAN, G. and YANG, Y. (2017). High dimensional correlation matrices: The central limit theorem and its applications. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 677–693. [MR3641402](#)
- GUPTA, A. K., JOHNSON, B. E. and NAGAR, D. K. (2013). Testing equality of several correlation matrices. *Rev. Colombiana Estadist.* **36** 237–258. [MR3242200](#)
- JENNRICH, R. I. (1970). An asymptotic χ^2 test for the equality of two correlation matrices. *J. Amer. Statist. Assoc.* **65** 904–912. [MR0269023](#)
- KULLBACK, S. (1967). On testing correlation matrices. *J. Roy. Statist. Soc. Ser. C* **16** 80–85.
- LARNTZ, K. and PERLMAN, M. D. (1985). A simple test for the equality of correlation matrices Technical Report 63.
- LI, D. N. and XUE, L. Z. (2015). Joint limiting laws for high-dimensional independence tests. Available at [arXiv:1512.08819v1](#).
- QIU, Y. and CHEN, S. X. (2012). Test for bandedness of high-dimensional covariance matrices and bandwidth estimation. *Ann. Statist.* **40** 1285–1314. [MR3015026](#)
- SCHOTT, J. R. (1996). Testing for the equality of several correlation matrices. *Statist. Probab. Lett.* **27** 85–89. [MR1394180](#)
- SCHOTT, J. R. (2005). Testing for complete independence in high dimensions. *Biometrika* **92** 951–956. [MR2234197](#)
- SHAO, Q.-M. and ZHOU, W.-X. (2014). Necessary and sufficient conditions for the asymptotic distributions of coherence of ultra-high dimensional random matrices. *Ann. Probab.* **42** 623–648. [MR3178469](#)
- ZHENG, S., CHENG, G., GUO, J. and ZHU, H. (2019). Supplement to “Test for high-dimensional correlation matrices.” DOI:[10.1214/18-AOS1768SUPP](#).
- ZHENG, S., CHEN, Z., CUI, H. and LI, R. (2018). Hypothesis testing on linear structures of high dimensional covariance matrix. *Ann. Statist.* To appear.
- ZHOU, C., HAN, F., ZHANG, X. S. and LIU, H. (2015). An Extreme-value approach for testing the equality of large U-statistic based correlation matrices. Available at [arXiv:1502.03211](#).

PROJECTED SPLINE ESTIMATION OF THE NONPARAMETRIC FUNCTION IN HIGH-DIMENSIONAL PARTIALLY LINEAR MODELS FOR MASSIVE DATA

BY HENG LIAN¹, KAIFENG ZHAO AND SHAOGAO LV²

City University of Hong Kong, Philips Research China and Nanjing Audit University

In this paper, we consider the local asymptotics of the nonparametric function in a partially linear model, within the framework of the divide-and-conquer estimation. Unlike the fixed-dimensional setting in which the parametric part does not affect the nonparametric part, the high-dimensional setting makes the issue more complicated. In particular, when a sparsity-inducing penalty such as lasso is used to make the estimation of the linear part feasible, the bias introduced will propagate to the nonparametric part. We propose a novel approach for estimation of the nonparametric function and establish the local asymptotics of the estimator. The result is useful for massive data with possibly different linear coefficients in each subpopulation but common nonparametric function. Some numerical illustrations are also presented.

REFERENCES

- BANERJEE, M., DUROT, C. and SEN, B. (2017). Divide and conquer in non-standard problems and the super-efficiency phenomenon. *Ann. Statist.* To appear.
- BELLONI, A. and CHERNOZHUKOV, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19** 521–547. [MR3037163](#)
- BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. [MR2533469](#)
- CHEN, X. and XIE, M. (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statist. Sinica* **24** 1655–1684. [MR3308656](#)
- DE BOOR, C. (2001). *A Practical Guide to Splines*, Revised ed. *Applied Mathematical Sciences* **27**. Springer, New York. [MR1900298](#)
- DEMKO, S. (1977). Inverses of band matrices and local convergence of spline projections. *SIAM J. Numer. Anal.* **14** 616–619. [MR0455281](#)
- FAN, J. and YAO, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. *Springer Series in Statistics*. Springer, New York. [MR1964455](#)
- HAERDLE, W., LIANG, H. and GAO, J. (2007). *Partially linear models*. Springer, New York.
- HUANG, J. Z. (2003). Local asymptotics for polynomial spline regression. *Ann. Statist.* **31** 1600–1635. [MR2012827](#)
- HUANG, J. Z., ZHANG, L. and ZHOU, L. (2007). Efficient estimation in marginal partially linear models for longitudinal/clustering data using splines. *Scand. J. Stat.* **34** 451–477. [MR2368793](#)
- 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](#)

MSC2010 subject classifications. Primary 62G08; secondary 62G20.

Key words and phrases. Asymptotic normality, B-splines, local asymptotics, profiled estimation.

- LI, Q. (2000). Efficient estimation of additive partially linear models. *Internat. Econom. Rev.* **41** 1073–1092. [MR1790072](#)
- LIANG, H. and LI, R. (2009). Variable selection for partially linear models with measurement errors. *J. Amer. Statist. Assoc.* **104** 234–248. [MR2504375](#)
- LV, S. and LIAN, H. (2017). A debiased distributed estimation for sparse partially linear models in diverging dimensions. Available at [arXiv:1708.05487](#).
- SCHUMAKER, L. L. (2007). *Spline Functions: Basic Theory*, 3rd ed. *Cambridge Mathematical Library*. Cambridge Univ. Press, Cambridge. [MR2348176](#)
- SHI, C., LU, W. and SONG, R. (2017). A massive data framework for M-estimators with cubic-rate. *J. Amer. Statist. Assoc.* To appear.
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. [MR1379242](#)
- WAINWRIGHT, M. (2018). *High-Dimensional Statistics: A Non-asymptotic Viewpoint*. Draft book.
- WANG, L., LIU, X., LIANG, H. and CARROLL, R. J. (2011). Estimation and variable selection for generalized additive partial linear models. *Ann. Statist.* **39** 1827–1851. [MR2893854](#)
- XIE, H. and HUANG, J. (2009). SCAD-penalized regression in high-dimensional partially linear models. *Ann. Statist.* **37** 673–696. [MR2502647](#)
- ZHANG, Y., DUCHI, J. and WAINWRIGHT, M. (2015). Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *J. Mach. Learn. Res.* **16** 3299–3340. [MR3450540](#)
- ZHAO, T., CHENG, G. and LIU, H. (2016). A partially linear framework for massive heterogeneous data. *Ann. Statist.* **44** 1400–1437. [MR3519928](#)
- ZHOU, S., SHEN, X. and WOLFE, D. A. (1998). Local asymptotics for regression splines and confidence regions. *Ann. Statist.* **26** 1760–1782. [MR1673277](#)

INFERENCE FOR THE MODE OF A LOG-CONCAVE DENSITY

BY CHARLES R. DOSS¹ AND JON A. WELLNER²

University of Minnesota and University of Washington

We study a likelihood ratio test for the location of the mode of a log-concave density. Our test is based on comparison of the log-likelihoods corresponding to the unconstrained maximum likelihood estimator of a log-concave density and the constrained maximum likelihood estimator where the constraint is that the mode of the density is fixed, say at m . The constrained estimation problem is studied in detail in Doss and Wellner (2018). Here, the results of that paper are used to show that, under the null hypothesis (and strict curvature of $-\log f$ at the mode), the likelihood ratio statistic is asymptotically pivotal: that is, it converges in distribution to a limiting distribution which is free of nuisance parameters, thus playing the role of the χ^2_1 distribution in classical parametric statistical problems. By inverting this family of tests, we obtain new (likelihood ratio based) confidence intervals for the mode of a log-concave density f . These new intervals do not depend on any smoothing parameters. We study the new confidence intervals via Monte Carlo methods and illustrate them with two real data sets. The new intervals seem to have several advantages over existing procedures. Software implementing the test and confidence intervals is available in the R package `logcondens.mode`.

REFERENCES

- ABRAHAM, C., BIAU, G. and CADRE, B. (2003). Simple estimation of the mode of a multivariate density. *Canad. J. Statist.* **31** 23–34. [MR1985502](#)
- BAHADUR, R. R. and SAVAGE, L. J. (1956). The nonexistence of certain statistical procedures in nonparametric problems. *Ann. Math. Stat.* **27** 1115–1122. [MR0084241](#)
- BALABDAOUI, F., RUFIBACH, K. and WELLNER, J. A. (2009). Limit distribution theory for maximum likelihood estimation of a log-concave density. *Ann. Statist.* **37** 1299–1331. [MR2509075](#)
- BANERJEE, M. (2007). Likelihood based inference for monotone response models. *Ann. Statist.* **35** 931–956. [MR2341693](#)
- BANERJEE, M. and WELLNER, J. A. (2001). Likelihood ratio tests for monotone functions. *Ann. Statist.* **29** 1699–1731. [MR1891743](#)
- BIRGÉ, L. (1997). Estimation of unimodal densities without smoothness assumptions. *Ann. Statist.* **25** 970–981. [MR1447736](#)
- BLUM, J. R. and ROSENBLATT, J. (1966). On some statistical problems requiring purely sequential sampling schemes. *Ann. Inst. Statist. Math.* **18** 351–355. [MR0205414](#)
- CHEN, Y. and SAMWORTH, R. J. (2016). Generalized additive and index models with shape constraints. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 729–754. [MR3534348](#)

MSC2010 subject classifications. Primary 62G07; secondary 62G15, 62G10, 62G20.

Key words and phrases. Mode, empirical processes, likelihood ratio, pivot, convex optimization, log-concave, shape constraints.

- CULE, M. and SAMWORTH, R. (2010). Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electron. J. Stat.* **4** 254–270. [MR2645484](#)
- CULE, M., SAMWORTH, R. and STEWART, M. (2010). Maximum likelihood estimation of a multidimensional log-concave density. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 545–607. [MR2758237](#)
- DHARMADHIKARI, S. and JOAG-DEV, K. (1988). *Unimodality, Convexity, and Applications. Probability and Mathematical Statistics*. Academic Press, Boston, MA. [MR0954608](#)
- DONOHU, D. L. (1988). One-sided inference about functionals of a density. *Ann. Statist.* **16** 1390–1420. [MR0964930](#)
- DONOHU, D. L. and LIU, R. C. (1991). Geometrizing rates of convergence. II, III. *Ann. Statist.* **19** 633–667, 668–701. [MR1105839](#)
- DOSS, C. R. (2013a). logcondens.mode: Compute MLE of log-concave density on R with fixed mode, and perform inference for the mode. R package version 1.0.1. Available at <http://CRAN.R-project.org/package=logcondens.mode>.
- DOSS, C. R. (2013b). Shape-constrained inference for concave-transformed densities and their modes. Ph.D. thesis, Univ. Washington, Seattle, WA.
- DOSS, C. R. (2018). Concave regression: Value-constrained estimation and likelihood ratio-based inference. *Math. Program.* To appear. DOI:[10.1007/s10107-018-1338-5](#).
- DOSS, C. R. and WELLNER, J. A. (2016a). Global rates of convergence of the MLEs of log-concave and s -concave densities. *Ann. Statist.* **44** 954–981. [MR3485950](#)
- DOSS, C. R. and WELLNER, J. A. (2016b). Inference for the mode of a log-concave density. Available at [arXiv:1611.10348v2](#).
- DOSS, C. R. and WELLNER, J. A. (2018). Log-concave density estimation with symmetry or modal constraints. Available at [arXiv:1611.10335v3](#).
- DOSS, C. R. and WELLNER, J. A. (2019). Supplement to “Inference for the mode of a log-concave density.” DOI:[10.1214/18-AOS1770SUPP](#).
- DÜMBGEN, L. and RUFIBACH, K. (2009). Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli* **15** 40–68. [MR2546798](#)
- DÜMBGEN, L., SAMWORTH, R. and SCHUHMACHER, D. (2011). Approximation by log-concave distributions, with applications to regression. *Ann. Statist.* **39** 702–730. [MR2816336](#)
- EDDY, W. F. (1980). Optimum kernel estimators of the mode. *Ann. Statist.* **8** 870–882. [MR0572631](#)
- EHM, W. (1996). Adaptive kernel estimation of a cusp-shaped mode. In *Applied Mathematics and Parallel Computing* 109–120. Physica, Heidelberg. [MR1469264](#)
- GROENEBOOM, P. and JONGBLOED, G. (2015). Nonparametric confidence intervals for monotone functions. *Ann. Statist.* **43** 2019–2054. [MR3375875](#)
- GROENEBOOM, P., JONGBLOED, G. and WELLNER, J. A. (2001a). A canonical process for estimation of convex functions: The “envelope” of integrated Brownian motion $+t^4$. *Ann. Statist.* **29** 1620–1652. [MR1891741](#)
- GROENEBOOM, P., JONGBLOED, G. and WELLNER, J. A. (2001b). Estimation of a convex function: Characterizations and asymptotic theory. *Ann. Statist.* **29** 1653–1698. [MR1891742](#)
- HALL, P. (1984). On unimodality and rates of convergence for stable laws. *J. Lond. Math. Soc.* (2) **30** 371–384. [MR0771431](#)
- HALL, P. and JOHNSTONE, I. (1992). Empirical functional and efficient smoothing parameter selection. *J. Roy. Statist. Soc. Ser. B* **54** 475–530. With discussion and a reply by the authors. [MR1160479](#)
- HAN, Q. and WELLNER, J. A. (2016). Approximation and estimation of s -concave densities via Rényi divergences. *Ann. Statist.* **44** 1332–1359. [MR3485962](#)
- HANSON, D. L. and PLEDGER, G. (1976). Consistency in concave regression. *Ann. Statist.* **4** 1038–1050. [MR0426273](#)
- HÄRDLE, W., MARRON, J. S. and WAND, M. P. (1990). Bandwidth choice for density derivatives. *J. Roy. Statist. Soc. Ser. B* **52** 223–232. [MR1049312](#)

- HAS'MINSKII, R. Z. (1979). Lower bound for the risks of nonparametric estimates of the mode. In *Contribution to Statistics: Jaroslav Hájek Memorial Volume* 91–97. Reidel, Dordrecht. [MR0561262](#)
- HAZELTON, M. (1996a). Bandwidth selection for local density estimators. *Scand. J. Stat.* **23** 221–232. [MR1394655](#)
- HAZELTON, M. L. (1996b). Optimal rates for local bandwidth selection. *J. Nonparametr. Stat.* **7** 57–66. [MR1618191](#)
- HERRMANN, E. and ZIEGLER, K. (2004). Rates on consistency for nonparametric estimation of the mode in absence of smoothness assumptions. *Statist. Probab. Lett.* **68** 359–368. [MR2081114](#)
- HILDRETH, C. (1954). Point estimates of ordinates of concave functions. *J. Amer. Statist. Assoc.* **49** 598–619. [MR0065093](#)
- HOFFLEIT, D. and WARREN, W. H. JR. (1991). *Bright Star Catalogue*. Yale Univ. Observatory, New Haven, CT. Available at <http://adsabs.harvard.edu/abs/1995yCat.5050....0H>.
- HOGG, R. V. and CRAIG, A. T. (1970). *Introduction to Mathematical Statistics*. 3rd ed. Macmillan, New York. [MR0251823](#)
- IBRAGIMOV, I. A. (1956). On the composition of unimodal distributions. *Theory Probab. Appl.* **1** 225–260. [MR0087249](#)
- JANKOWSKI, H. K. and WELLNER, J. A. (2009). Nonparametric estimation of a convex bathtub-shaped hazard function. *Bernoulli* **15** 1010–1035. [MR2597581](#)
- KIM, J. (1994). Asymptotic theory for multi-dimensional mode estimator. *J. Korean Statist. Soc.* **23** 251–269. [MR1327618](#)
- KIM, A. K. H. and SAMWORTH, R. J. (2016). Global rates of convergence in log-concave density estimation. *Ann. Statist.* **44** 2756–2779. [MR3576560](#)
- KLEMELÄ, J. (2005). Adaptive estimation of the mode of a multivariate density. *J. Nonparametr. Stat.* **17** 83–105. [MR2112688](#)
- KOENKER, R. and MIZERA, I. (2010). Quasi-concave density estimation. *Ann. Statist.* **38** 2998–3027. [MR2722462](#)
- KONAKOV, V. D. (1973). The asymptotic normality of the mode of multivariate distributions. *Teor. Veroyatn. Primen.* **18** 836–842. [MR0336874](#)
- KUCHIBHOTLA, A. K., PATRA, R. K. and SEN, B. (2017). Efficient estimation in convex single index models. Available at [arXiv:1708.00145v2](https://arxiv.org/abs/1708.00145v2).
- LÉGER, C. and ROMANO, J. P. (1990). Bootstrap choice of tuning parameters. *Ann. Inst. Statist. Math.* **42** 709–735. [MR1089472](#)
- MALLOWS, C. L. (1972). A note on asymptotic joint normality. *Ann. Math. Stat.* **43** 508–515. [MR0298812](#)
- MAMMEN, E. (1991). Nonparametric regression under qualitative smoothness assumptions. *Ann. Statist.* **19** 741–759. [MR1105842](#)
- MAMMEN, E. and PARK, B. U. (1997). Optimal smoothing in adaptive location estimation. *J. Statist. Plann. Inference* **58** 333–348. [MR1450020](#)
- MÜLLER, H.-G. (1989). Adaptive nonparametric peak estimation. *Ann. Statist.* **17** 1053–1069. [MR1015137](#)
- PAL, J. K., WOODROOFE, M. and MEYER, M. (2007). Estimating a Pólya frequency function₂. In *Complex Datasets and Inverse Problems. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **54** 239–249. IMS, Beachwood, OH. [MR2459192](#)
- PARZEN, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Stat.* **33** 1065–1076. [MR0143282](#)
- PFANZAGL, J. (1998). The nonexistence of confidence sets for discontinuous functionals. *J. Statist. Plann. Inference* **75** 9–20. [MR1671670](#)
- PFANZAGL, J. (2000). On local uniformity for estimators and confidence limits. *J. Statist. Plann. Inference* **84** 27–53. [MR1747496](#)

- R CORE TEAM (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- ROCKAFELLAR, R. T. (1970). *Convex Analysis*. *Princeton Mathematical Series* **28**. Princeton Univ. Press, Princeton, NJ. [MR0274683](#)
- ROMANO, J. P. (1988a). Bootstrapping the mode. *Ann. Inst. Statist. Math.* **40** 565–586. [MR0964293](#)
- ROMANO, J. P. (1988b). On weak convergence and optimality of kernel density estimates of the mode. *Ann. Statist.* **16** 629–647. [MR0947566](#)
- RUFIBACH, K. (2006). Log-concave density estimation and bump hunting for IID observations. Ph.D. thesis, Univ. Bern and Gottingen.
- SAGER, T. W. (1978). Estimation of a multivariate mode. *Ann. Statist.* **6** 802–812. [MR0491553](#)
- SAMANTA, M. (1973). Nonparametric estimation of the mode of a multivariate density. *South African Statist. J.* **7** 109–117. [MR0331618](#)
- SAMWORTH, R. J. and WAND, M. P. (2010). Asymptotics and optimal bandwidth selection for highest density region estimation. *Ann. Statist.* **38** 1767–1792. [MR2662359](#)
- SAUMARD, A. and WELLNER, J. A. (2014). Log-concavity and strong log-concavity: A review. *Stat. Surv.* **8** 45–114. [MR3290441](#)
- SHEATHER, S. J. and JONES, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B* **53** 683–690. [MR1125725](#)
- SILVERMAN, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.* **10** 795–810. [MR0663433](#)
- SINGH, R. (1963). Existence of bounded length confidence intervals. *Ann. Math. Stat.* **34** 1474–1485. [MR0155394](#)
- SINGH, R. (1968). Non-existence of estimates of prescribed accuracy in fixed sample size. *Canad. Math. Bull.* **11** 135–139. [MR0232487](#)
- TSYBAKOV, A. B. (1990). Recurrent estimation of the mode of a multidimensional distribution. *Problemy Peredachi Informatsii* **26** 38–45. [MR1051586](#)
- WALTHER, G. (2002). Detecting the presence of mixing with multiscale maximum likelihood. *J. Amer. Statist. Assoc.* **97** 508–513. [MR1941467](#)
- WILKS, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.* **9** 60–62.
- ZIEGLER, K. (2001). On bootstrapping the mode in the nonparametric regression model with random design. *Metrika* **53** 141–170. [MR1839003](#)

TESTING FOR INDEPENDENCE OF LARGE DIMENSIONAL VECTORS

BY TARAS BODNAR, HOLGER DETTE¹ AND NESTOR PAROLYA

Stockholm University, Ruhr University Bochum and Leibniz University Hannover

In this paper, new tests for the independence of two high-dimensional vectors are investigated. We consider the case where the dimension of the vectors increases with the sample size and propose multivariate analysis of variance-type statistics for the hypothesis of a block diagonal covariance matrix. The asymptotic properties of the new test statistics are investigated under the null hypothesis and the alternative hypothesis using random matrix theory. For this purpose, we study the weak convergence of linear spectral statistics of central and (conditionally) noncentral Fisher matrices. In particular, a central limit theorem for linear spectral statistics of large dimensional (conditionally) noncentral Fisher matrices is derived which is then used to analyse the power of the tests under the alternative.

The theoretical results are illustrated by means of a simulation study where we also compare the new tests with several alternative, in particular with the commonly used corrected likelihood ratio test. It is demonstrated that the latter test does not keep its nominal level, if the dimension of one sub-vector is relatively small compared to the dimension of the other sub-vector. On the other hand, the tests proposed in this paper provide a reasonable approximation of the nominal level in such situations. Moreover, we observe that one of the proposed tests is most powerful under a variety of correlation scenarios.

REFERENCES

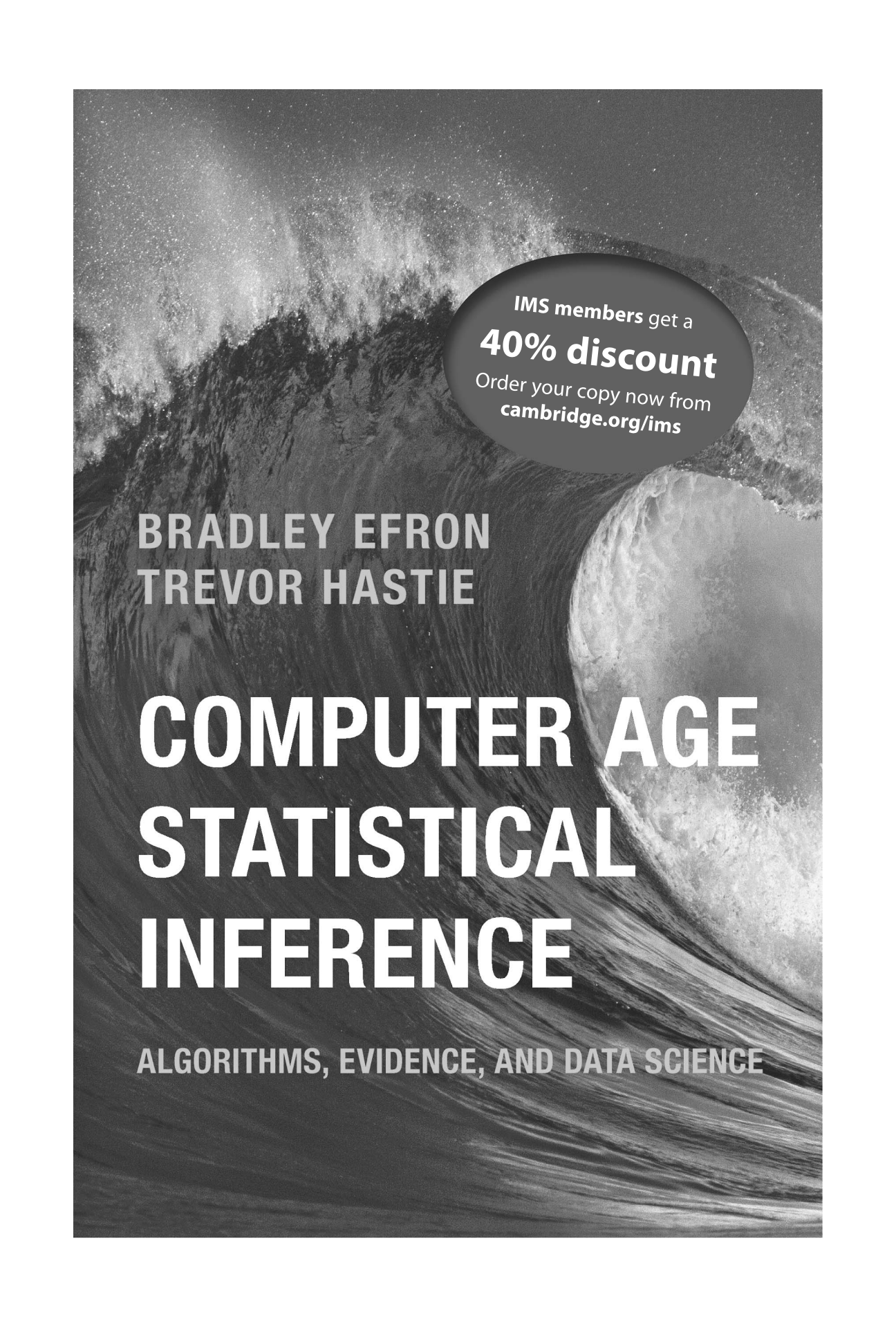
- ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. [MR1990662](#)
- BAI, Z. D. and SILVERSTEIN, J. W. (2004). CLT for linear spectral statistics of large-dimensional sample covariance matrices. *Ann. Probab.* **32** 553–605. [MR2040792](#)
- BAI, Z. and SILVERSTEIN, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed. *Springer Series in Statistics*. Springer, New York. [MR2567175](#)
- BAI, Z., JIANG, D., YAO, J.-F. and ZHENG, S. (2009). Corrections to LRT on large-dimensional covariance matrix by RMT. *Ann. Statist.* **37** 3822–3840. [MR2572444](#)
- BICKEL, P. J. and LEVINA, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. [MR2387969](#)
- BIRKE, M. and DETTE, H. (2005). A note on testing the covariance matrix for large dimension. *Statist. Probab. Lett.* **74** 281–289. [MR2189467](#)
- BODNAR, T., DETTE, H. and PAROLYA, N. (2019). Supplement to “Testing for independence of large dimensional vectors.” DOI:[10.1214/18-AOS1771SUPP](#).

MSC2010 subject classifications. Primary 60B20, 60F05, 62H15; secondary 62H20, 62F05.

Key words and phrases. Testing for independence, large dimensional covariance matrix, noncentral Fisher random matrix, linear spectral statistics, asymptotic normality.

- BODNAR, T., GUPTA, A. K. and PAROLYA, N. (2014). On the strong convergence of the optimal linear shrinkage estimator for large dimensional covariance matrix. *J. Multivariate Anal.* **132** 215–228. [MR3266272](#)
- BODNAR, T., GUPTA, A. K. and PAROLYA, N. (2016). Direct shrinkage estimation of large dimensional precision matrix. *J. Multivariate Anal.* **146** 223–236. [MR3477661](#)
- CAI, T., LIU, W. and LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. [MR2847973](#)
- CAI, T. T., REN, Z. and ZHOU, H. H. (2013). Optimal rates of convergence for estimating Toeplitz covariance matrices. *Probab. Theory Related Fields* **156** 101–143. [MR3055254](#)
- CAI, T. T. and SHEN, X., eds. (2011). *High-Dimensional Data Analysis*. *Frontiers of Statistics* **2**. World Scientific Co. Pte. Ltd., Singapore; Higher Education Press, Beijing. [MR2866981](#)
- CAI, T. T. and ZHOU, H. H. (2012). Minimax estimation of large covariance matrices under ℓ_1 -norm. *Statist. Sinica* **22** 1319–1349. [MR3027084](#)
- CHEN, S. X., ZHANG, L.-X. and ZHONG, P.-S. (2010). Tests for high-dimensional covariance matrices. *J. Amer. Statist. Assoc.* **105** 810–819. [MR2724863](#)
- DEVIJVER, E. and GALLOPIN, M. (2018). Block-diagonal covariance selection for high-dimensional Gaussian graphical models. *J. Amer. Statist. Assoc.* **113** 306–314. [MR3803466](#)
- DOZIER, R. B. and SILVERSTEIN, J. W. (2007). On the empirical distribution of eigenvalues of large dimensional information-plus-noise-type matrices. *J. Multivariate Anal.* **98** 678–694. [MR2322123](#)
- FAN, J. and LI, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. In *International Congress of Mathematicians. Vol. III* 595–622. Eur. Math. Soc., Zürich. [MR2275698](#)
- FISHER, R. A. (1939). The sampling distribution of some statistics obtained from non-linear equations. *Ann. Eugenics* **9** 238–249. [MR0001499](#)
- FISHER, T. J., SUN, X. and GALLAGHER, C. M. (2010). A new test for sphericity of the covariance matrix for high dimensional data. *J. Multivariate Anal.* **101** 2554–2570. [MR2719881](#)
- FUJIKOSHI, Y., HIMENO, T. and WAKAKI, H. (2004). Asymptotic results of a high dimensional Manova test and power comparison when the dimension is large compared to the sample size. *J. Japan Statist. Soc.* **34** 19–26. [MR2084057](#)
- GUPTA, A. K. and XU, J. (2006). On some tests of the covariance matrix under general conditions. *Ann. Inst. Statist. Math.* **58** 101–114. [MR2281208](#)
- HYODO, M., SHUTOH, N., NISHIYAMA, T. and PAVLENKO, T. (2015). Testing block-diagonal covariance structure for high-dimensional data. *Stat. Neerl.* **69** 460–482. [MR3414707](#)
- JIANG, D., BAI, Z. and ZHENG, S. (2013). Testing the independence of sets of large-dimensional variables. *Sci. China Math.* **56** 135–147. [MR3016588](#)
- 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](#)
- JOHN, S. (1971). Some optimal multivariate tests. *Biometrika* **58** 123–127. [MR0275568](#)
- JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. [MR1863961](#)
- JOHNSTONE, I. (2006). High dimensional statistical inference and random matrices. Instituto de Ciencias Matemáticas (ICMAT). Available at: http://www.icm2006.org/proceedings/Vol_I/17.pdf.
- JOHNSTONE, I. M. (2008). Multivariate analysis and Jacobi ensembles: Largest eigenvalue, Tracy–Widom limits and rates of convergence. *Ann. Statist.* **36** 2638–2716. [MR2485010](#)
- KAKIZAWA, Y. and IWASHITA, T. (2008). A comparison of higher-order local powers of a class of one-way MANOVA tests under general distributions. *J. Multivariate Anal.* **99** 1128–1153. [MR2432484](#)
- LEDOIT, O. and WOLF, M. (2002). Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Ann. Statist.* **30** 1081–1102. [MR1926169](#)

- LEDOIT, O. and WOLF, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Ann. Statist.* **40** 1024–1060. [MR2985942](#)
- MARKOWITZ, H. (1952). Portfolio selection. *J. Finance* **7** 77–91.
- MAUCHLY, J. W. (1940). Significance test for sphericity of a normal n -variate distribution. *Ann. Math. Stat.* **11** 204–209. [MR0002084](#)
- MUIRHEAD, R. J. (1982). *Aspects of Multivariate Statistical Theory*. *Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. [MR0652932](#)
- NAGAO, H. (1973). On some test criteria for covariance matrix. *Ann. Statist.* **1** 700–709. [MR0339405](#)
- PILLAI, K. C. S. and JAYACHANDRAN, K. (1967). Power comparisons of tests of two multivariate hypotheses based on four criteria. *Biometrika* **54** 195–210. [MR0216641](#)
- SCHOTT, J. R. (2007). Some high-dimensional tests for a one-way MANOVA. *J. Multivariate Anal.* **98** 1825–1839. [MR2392435](#)
- WANG, C., PAN, G., TONG, T. and ZHU, L. (2015). Shrinkage estimation of large dimensional precision matrix using random matrix theory. *Statist. Sinica* **25** 993–1008. [MR3409734](#)
- YAMADA, Y., HYODO, M. and NISHIYAMA, T. (2017). Testing block-diagonal covariance structure for high-dimensional data under non-normality. *J. Multivariate Anal.* **155** 305–316. [MR3607897](#)
- YANG, Y. and PAN, G. (2015). Independence test for high dimensional data based on regularized canonical correlation coefficients. *Ann. Statist.* **43** 467–500. [MR3316187](#)
- YAO, J. (2013). Estimation and fluctuations of functionals of large random matrices. Telecom Paris-Tech, tel-00909521v1.
- YAO, J., ZHENG, S. and BAI, Z. (2015). *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. *Cambridge Series in Statistical and Probabilistic Mathematics* **39**. Cambridge Univ. Press, New York. [MR3468554](#)
- ZHENG, S. (2012). Central limit theorems for linear spectral statistics of large dimensional F -matrices. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 444–476. [MR2954263](#)
- ZHENG, S., BAI, Z. and YAO, J. (2015a). CLT for linear spectral statistics of a rescaled sample precision matrix. *Random Matrices Theory Appl.* **4** 1550014, 43. [MR3418843](#)
- ZHENG, S., BAI, Z. and YAO, J. (2015b). 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](#)
- ZHENG, S., BAI, Z. and YAO, J. (2017). CLT for eigenvalue statistics of large-dimensional general Fisher matrices with applications. *Bernoulli* **23** 1130–1178. [MR3606762](#)



IMS members get a
40% discount

Order your copy now from
cambridge.org/ims

BRADLEY EFRON
TREVOR HASTIE

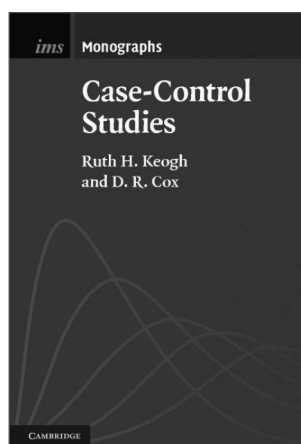
COMPUTER AGE STATISTICAL INFERENCE

ALGORITHMS, EVIDENCE, AND DATA SCIENCE



The Institute of Mathematical Statistics presents

IMS MONOGRAPHS



Case-Control Studies

Ruth H. Keogh
and D. R. Cox

The case-control approach is a powerful method for investigating factors that may explain a particular event. It is extensively used in epidemiology to study disease incidence, one of the best-known examples being Bradford Hill and Doll's investigation of the possible connection between cigarette smoking and lung cancer. More recently, case-control studies have been increasingly used in other fields, including sociology and econometrics.

With a particular focus on statistical analysis, this book is ideal for applied and theoretical statisticians wanting an up-to-date introduction to the field. It covers the fundamentals of case-control study design and analysis as well as more recent developments, including two-stage studies, case-only studies and methods for case-control sampling in time. The latter have important applications in large prospective cohorts which require case-control sampling designs to make efficient use of resources. More theoretical background is provided in an appendix for those new to the field.

**IMS member? Claim
your 40% discount:
www.cambridge.org/ims**

**Hardback price
US\$48.00
(non-member price
\$80.00)**

Cambridge University Press, in conjunction with the Institute of Mathematical Statistics, established the IMS Monographs and IMS Textbooks series of high-quality books. The Series Editors are Xiao-Li Meng, Susan Holmes, Ben Hambly, D. R. Cox and Alan Agresti.