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

THE ANNALS *of* STATISTICS

An Official Journal of the Institute of Mathematical Statistics

Articles

Vol. 54, No. 2—April 2026

THE ANNALS OF STATISTICS Vol. 54, No. 2, pp. 543–1099 April 2026

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

IMS EDITORS

The Annals of Statistics **[ISSN 0090-5364 (print); ISSN 2168-8966 (online)],** Volume 54, Number 2, April 2026. Published bimonthly by the Institute of Mathematical Statistics, 9760 Smith Road, Waite Hill, OH 44094, USA. Periodicals postage paid at Cleveland, Ohio, and at additional mailing offices.

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

The Annals of Statistics
2026, Vol. 54, No. 2, 543–569
https://doi.org/10.1214/25-AOS2552

RANK TESTS FOR PCA UNDER WEAK IDENTIFIABILITY

BY DAVY PAINDAVEINE[a], LAURA PERALVO MAROTO[b] AND THOMAS VERDEBOUT[c]

ECARES and Department of Mathematics, Université libre de Bruxelles, [a]*Davy.Paindaveine@ulb.be,* [b]*Laura.Peralvo@ulb.be,* [c]*Thomas.Verdebout@ulb.be*

In a triangular array framework where n observations are randomly sampled from a p-dimensional elliptical distribution with shape matrix $\mathbf{V}_n$, we consider the problem of testing the null hypothesis $\mathcal{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against the alternative hypothesis $\mathcal{H}_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}$ is the (fixed) leading unit eigenvector of $\mathbf{V}_n$ and $\boldsymbol{\theta}_0$ is a given unit p-vector. The dependence of the shape matrix on the sample size allows us to consider challenging asymptotic scenarios in which the parameter of interest $\boldsymbol{\theta}$ is unidentified in the limit, because the ratio between both leading eigenvalues of $\mathbf{V}_n$ converges to one. We carefully study the corresponding limiting experiments under such *weak identifiability*, and we show that these may be LAN or non-LAN. While earlier work in the framework was strictly limited to Gaussian distributions, where the study of local log-likelihood ratios could simply rely on explicit expressions, our asymptotic investigation allows for essentially arbitrary elliptical distributions. This requires original results on quadratic mean differentiable families for triangular arrays of observations, which are likely to be of interest in other models, too. Even in non-LAN experiments, our results enable us to investigate, through Le Cam's first and third lemmas, the asymptotic null and nonnull properties of multivariate rank tests. These nonparametric tests are shown to exhibit an excellent behavior under weak identifiability: not only do they maintain the target nominal size irrespective of the amount of weak identifiability, but they also keep their outstanding uniform efficiency properties under such nonstandard scenarios. In particular, Gaussian-score rank tests, under arbitrarily weak identifiability, still *uniformly* dominate their parametric pseudo-Gaussian competitor in terms of asymptotic relative efficiencies. Our theoretical results, which are the first ones to study rank tests in the triangular array framework allowing for weak identifiability, are supported by several Monte Carlo exercises.

REFERENCES

ANDERSON, T. W. (1963). Asymptotic theory for principal component analysis. *Ann. Math. Statist.* **34** 122–148. MR0145620 https://doi.org/10.1214/aoms/1177704248

BANERJEE, D. and MA, Z. (2022). Optimal signal detection in some spiked random matrix models: Likelihood ratio tests and linear spectral statistics. *Ann. Statist.* **50** 1910–1932. MR4474477 https://doi.org/10.1214/21-aos2150

BERNARD, G. and VERDEBOUT, T. (2024a). On testing the equality of latent roots of scatter matrices under ellipticity. *J. Multivariate Anal.* **199** Paper No. 105232, 17 pp. MR4642575 https://doi.org/10.1016/j.jmva.2023.105232

BERNARD, G. and VERDEBOUT, T. (2024b). Power enhancement for dimension detection of Gaussian signals. *Statist. Sinica* **34** 2161–2182. MR4805565

CROUX, C. and HAESBROECK, G. (2000). Principal component analysis based on robust estimators of the covariance or correlation matrix: Influence functions and efficiencies. *Biometrika* **87** 603–618. MR1789812 https://doi.org/10.1093/biomet/87.3.603

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

MSC2020 subject classifications. Primary 62H25, 62F05; secondary 62G10, 62G35.

Key words and phrases. Elliptical densities, limiting experiments, multivariate signs and ranks, principal component analysis, spiked scatter matrices, triangular arrays of observations, weak identifiability.

DÖRNEMANN, N. and DETTE, H. (2026). A CLT for the difference of eigenvalue statistics of sample covariance matrices. *Bernoulli* **32** 615–637. MR5000313 https://doi.org/10.3150/25-bej1872

DÜMBGEN, L. (1995). Likelihood ratio tests for principal components. *J. Multivariate Anal.* **52** 245–258. MR1323332 https://doi.org/10.1006/jmva.1995.1012

FAN, J., FAN, Y., HAN, X. and LV, J. (2022). Asymptotic theory of eigenvectors for random matrices with diverging spikes. *J. Amer. Statist. Assoc.* **117** 996–1009. MR4436328 https://doi.org/10.1080/01621459.2020.1840990

FLURY, B. (1988). *Common Principal Components and Related Multivariate Models*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR0986245 https://doi.org/10.1007/978-3-642-01932-6_28

HALLIN, M. and PAINDAVEINE, D. (2004). Rank-based optimal tests of the adequacy of an elliptic VARMA model. *Ann. Statist.* **32** 2642–2678. MR2153998 https://doi.org/10.1214/009053604000000724

HALLIN, M. and PAINDAVEINE, D. (2006). Semiparametrically efficient rank-based inference for shape. I. Optimal rank-based tests for sphericity. *Ann. Statist.* **34** 2707–2756. MR2329465 https://doi.org/10.1214/009053606000000731

HALLIN, M. and PAINDAVEINE, D. (2009). Optimal tests for homogeneity of covariance, scale, and shape. *J. Multivariate Anal.* **100** 422–444. MR2483429 https://doi.org/10.1016/j.jmva.2008.05.010

HALLIN, M., PAINDAVEINE, D. and VERDEBOUT, T. (2010a). Optimal rank-based testing for principal components. *Ann. Statist.* **38** 3245–3299. MR2766852 https://doi.org/10.1214/10-AOS810

HALLIN, M., PAINDAVEINE, D. and VERDEBOUT, T. (2010b). Testing for common principal components under heterokurticity. *J. Nonparametr. Stat.* **22** 879–895. MR2738873 https://doi.org/10.1080/10485250903548737

HALLIN, M., PAINDAVEINE, D. and VERDEBOUT, T. (2014). Efficient R-estimation of principal and common principal components. *J. Amer. Statist. Assoc.* **109** 1071–1083. MR3265681 https://doi.org/10.1080/01621459.2014.880057

HALLIN, M. and WERKER, B. J. M. (2003). Semi-parametric efficiency, distribution-freeness and invariance. *Bernoulli* **9** 137–165. MR1963675 https://doi.org/10.3150/bj/1068129013

HAN, F. and LIU, H. (2014). Scale-invariant sparse PCA on high-dimensional meta-elliptical data. *J. Amer. Statist. Assoc.* **109** 275–287. MR3180563 https://doi.org/10.1080/01621459.2013.844699

HE, R., HU, B.-G., ZHENG, W.-S. and KONG, X.-W. (2011). Robust principal component analysis based on maximum correntropy criterion. *IEEE Trans. Image Process.* **20** 1485–1494. MR2828599 https://doi.org/10.1109/TIP.2010.2103949

HETTMANSPERGER, T. P. and RANDLES, R. H. (2002). A practical affine equivariant multivariate median. *Biometrika* **89** 851–860. MR1946515 https://doi.org/10.1093/biomet/89.4.851

HUBERT, M., ROUSSEEUW, P. J. and VANDEN BRANDEN, K. (2005). ROBPCA: A new approach to robust principal component analysis. *Technometrics* **47** 64–79. MR2135793 https://doi.org/10.1198/004017004000000563

ILMONEN, P. and PAINDAVEINE, D. (2011). Semiparametrically efficient inference based on signed ranks in symmetric independent component models. *Ann. Statist.* **39** 2448–2476. MR2906874 https://doi.org/10.1214/11-AOS906

JACKSON, J. E. (2005). *A User's Guide to Principal Components*. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ.

JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693. MR2751448 https://doi.org/10.1198/jasa.2009.0121

JOLICOEUR, P. (1984). Principal components, factor analysis, and multivariate allometry: A small-sample direction test. *Biometrics* **40** 685–690.

KREISS, J.-P. (1987). On adaptive estimation in stationary ARMA processes. *Ann. Statist.* **15** 112–133. MR0885727 https://doi.org/10.1214/aos/1176350256

MUIRHEAD, R. J. and WATERNAUX, C. M. (1980). Asymptotic distributions in canonical correlation analysis and other multivariate procedures for nonnormal populations. *Biometrika* **67** 31–43. MR0570502 https://doi.org/10.1093/biomet/67.1.31

OJA, H. (2010). *Multivariate Nonparametric Methods with R*: *An Approach Based on Spatial Signs and Ranks*. *Lecture Notes in Statistics* **199**. Springer, New York. MR2598854 https://doi.org/10.1007/978-1-4419-0468-3

PAINDAVEINE, D. (2006). A Chernoff–Savage result for shape: On the non-admissibility of pseudo-Gaussian methods. *J. Multivariate Anal.* **97** 2206–2220. MR2301635 https://doi.org/10.1016/j.jmva.2005.08.005

PAINDAVEINE, D., PERALVO MAROTO, L. and VERDEBOUT, T. (2026). Supplement to "Rank tests for PCA under weak identifiability." https://doi.org/10.1214/25-AOS2552SUPP

PAINDAVEINE, D., REMY, J. and VERDEBOUT, T. (2020a). Testing for principal component directions under weak identifiability. *Ann. Statist.* **48** 324–345. MR4065164 https://doi.org/10.1214/18-AOS1805

PAINDAVEINE, D., REMY, J. and VERDEBOUT, T. (2020b). Sign tests for weak principal directions. *Bernoulli* **26** 2987–3016. MR4140535 https://doi.org/10.3150/20-BEJ1213

SCHWARTZMAN, A., MASCARENHAS, W. F. and TAYLOR, J. E. (2008). Inference for eigenvalues and eigenvectors of Gaussian symmetric matrices. *Ann. Statist.* **36** 2886–2919. MR2485016 https://doi.org/10.1214/08-AOS628

SHAPIRO, A. and BROWNE, M. W. (1987). Analysis of covariance structures under elliptical distributions. *J. Amer. Statist. Assoc.* **82** 1092–1097. MR0922173

SYLVESTER, A. D., KRAMER, P. A. and JUNGERS, W. L. (2008). Modern humans are not (quite) isometric. *Amer. J. Phys. Anthropol.* **137** 371–383.

TYLER, D. E. (1981). Asymptotic inference for eigenvectors. *Ann. Statist.* **9** 725–736. MR0619278

TYLER, D. E. (1983). A class of asymptotic tests for principal component vectors. *Ann. Statist.* **11** 1243–1250. MR0720269 https://doi.org/10.1214/aos/1176346337

TYLER, D. E. (1987). A distribution-free M-estimator of multivariate scatter. *Ann. Statist.* **15** 234–251. MR0885734 https://doi.org/10.1214/aos/1176350263

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

The Annals of Statistics
2026, Vol. 54, No. 2, 570–596
https://doi.org/10.1214/25-AOS2578

DISTRIBUTIONALLY ROBUST LEARNING FOR MULTISOURCE UNSUPERVISED DOMAIN ADAPTATION

BY ZHENYU WANG[1,a], PETER BÜHLMANN[2,b] AND ZIJIAN GUO[3,c]

[1]*Department of Statistics, Rutgers University,* [a]*zw425@stat.rutgers.edu*
[2]*Seminar for Statistics, ETH Zürich,* [b]*peter.buehlmann@stat.math.ethz.ch*
[3]*Center for Data Science, Zhejiang University,* [c]*zijguo@zju.edu.cn*

Empirical risk minimization often performs poorly when the distribution of the target domain differs from those of the source domains. To address such potential distributional shifts, we develop an unsupervised domain adaptation approach that leverages labeled data from multiple source domains and unlabeled data from the target domain. We introduce a distributionally robust model that optimizes an adversarial reward based on explained variance across a class of target distributions, ensuring generalization to the target domain. We show that the proposed robust model is a weighted average of conditional outcome models from the source domains. This formulation allows us to compute the robust model through the aggregation of source models, which can be estimated using various machine learning algorithms of the user's choice such as random forests, boosting and neural networks. Additionally, we introduce a bias-correction step to obtain a more accurate aggregation weight, which is effective for various machine learning algorithms. Our framework can be interpreted as a distributionally robust federated learning approach that satisfies privacy constraints while providing insights into the importance of each source for prediction on the target domain. The performance of our method is evaluated on both simulated and real data.

REFERENCES

[1] AGARWAL, A. and ZHANG, T. (2022). Minimax regret optimization for robust machine learning under distribution shift. In *Conference on Learning Theory* 2704–2729. PMLR.
[2] BEN-DAVID, S., BLITZER, J., CRAMMER, K. and PEREIRA, F. (2006). Analysis of representations for domain adaptation. *Adv. Neural Inf. Process. Syst.* **19**.
[3] BEN-TAL, A., DEN HERTOG, D., DE WAEGENAERE, A., MELENBERG, B. and RENNEN, G. (2013). Robust solutions of optimization problems affected by uncertain probabilities. *Manag. Sci.* **59** 341–357.
[4] BIAU, G. (2012). Analysis of a random forests model. *J. Mach. Learn. Res.* **13** 1063–1095. MR2930634
[5] BIAU, G., DEVROYE, L. and LUGOSI, G. (2008). Consistency of random forests and other averaging classifiers. *J. Mach. Learn. Res.* **9** 2015–2033. MR2447310
[6] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469 https://doi.org/10.1214/08-AOS620
[7] BÜHLMANN, P. and MEINSHAUSEN, N. (2015). Magging: Maximin aggregation for inhomogeneous large-scale data. *Proc. IEEE* **104** 126–135.
[8] CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2404.
[9] CAO, L., YANG, Q. and YU, P. S. (2021). Data science and AI in FinTech: An overview. *Int. J. Data Sci. Anal.* **12** 81–99.
[10] CHEN, X., HONG, H. and TAROZZI, A. (2008). Semiparametric efficiency in GMM models with auxiliary data. *Ann. Statist.* **36** 808–843. MR2396816 https://doi.org/10.1214/009053607000000947
[11] CHEN, Z., CHEN, D., ZHAO, C., KWAN, M.-P., CAI, J., ZHUANG, Y., ZHAO, B., WANG, X., CHEN, B. et al. (2020). Influence of meteorological conditions on PM2. 5 concentrations across China: A review of methodology and mechanism. *Environ. Int.* **139** 105558.

MSC2020 subject classifications. Primary 62R07; secondary 62G05.
Key words and phrases. Unsupervised domain adaptation, distributionally robust optimization, federated learning, interpretable machine learning.

[12] DENG, Y., KAMANI, M. M. and MAHDAVI, M. (2020). Distributionally robust federated averaging. *Adv. Neural Inf. Process. Syst.* **33** 15111–15122.
[13] DIANA, E., GILL, W., KEARNS, M., KENTHAPADI, K. and ROTH, A. (2021). Minimax group fairness: Algorithms and experiments. In *Proceedings of the* 2021 *AAAI/ACM Conference on AI, Ethics, and Society* 66–76.
[14] DUAN, L., XU, D. and TSANG, I. W.-H. (2012). Domain adaptation from multiple sources: A domain-dependent regularization approach. *IEEE Trans. Neural Netw. Learn. Syst.* **23** 504–518.
[15] ELDAR, Y. C., BECK, A. and TEBOULLE, M. (2008). A minimax Chebyshev estimator for bounded error estimation. *IEEE Trans. Signal Process.* **56** 1388–1397. MR2512472 https://doi.org/10.1109/TSP.2007.908945
[16] FARRELL, M. H., LIANG, T. and MISRA, S. (2021). Deep neural networks for estimation and inference. *Econometrica* **89** 181–213. MR4220387 https://doi.org/10.3982/ecta16901
[17] GANIN, Y., USTINOVA, E., AJAKAN, H., GERMAIN, P., LAROCHELLE, H., LAVIOLETTE, F., MARCHAND, M. and LEMPITSKY, V. (2016). Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **17** Paper No. 59. MR3504619
[18] GRETTON, A., SMOLA, A. J., HUANG, J., SCHMITTFULL, M., BORGWARDT, K. M. and SCHÖLLKOPF, B. (2009). Covariate shift by kernel mean matching. 131–160.
[19] GUO, Z. (2024). Statistical inference for maximin effects: Identifying stable associations across multiple studies. *J. Amer. Statist. Assoc.* **119** 1968–1984. MR4797916 https://doi.org/10.1080/01621459.2023.2233162
[20] HU, J., LU, J. and TAN, Y.-P. (2015). Deep transfer metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 325–333.
[21] HU, W., NIU, G., SATO, I. and SUGIYAMA, M. (2018). Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning* 2029–2037. PMLR.
[22] KANAMORI, T., HIDO, S. and SUGIYAMA, M. (2009). A least-squares approach to direct importance estimation. *J. Mach. Learn. Res.* **10** 1391–1445. MR2534866
[23] KOH, P. W., SAGAWA, S., MARKLUND, H., XIE, S. M., ZHANG, M., BALSUBRAMANI, A., HU, W., YASUNAGA, M., PHILLIPS, R. L. et al. (2021). Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning* 5637–5664. PMLR.
[24] LONG, M., CAO, Y., CAO, Z., WANG, J. and JORDAN, M. I. (2018). Transferable representation learning with deep adaptation networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **41** 3071–3085.
[25] MALININ, A., BAND, N., CHESNOKOV, G., GAL, Y., GALES, M. J., NOSKOV, A., PLOSKONOSOV, A., PROKHORENKOVA, L., PROVILKOV, I. et al. (2021). Shifts: a dataset of real distributional shift across multiple large-scale tasks. arXiv Preprint. Available at arXiv:2107.07455.
[26] MANSOUR, Y., MOHRI, M. and ROSTAMIZADEH, A. (2008). Domain adaptation with multiple sources. *Adv. Neural Inf. Process. Syst.* **21**.
[27] MARTINEZ, N., BERTRAN, M. and SAPIRO, G. (2020). Minimax Pareto fairness: A multi objective perspective. In *International Conference on Machine Learning* 6755–6764. PMLR.
[28] MEINSHAUSEN, N. (2006). Quantile regression forests. *J. Mach. Learn. Res.* **7** 983–999. MR2274394
[29] MEINSHAUSEN, N. and BÜHLMANN, P. (2015). Maximin effects in inhomogeneous large-scale data. *Ann. Statist.* **43** 1801–1830. MR3357879 https://doi.org/10.1214/15-AOS1325
[30] MENON, A. and ONG, C. S. (2016). Linking losses for density ratio and class-probability estimation. In *International Conference on Machine Learning* 304–313. PMLR.
[31] MILANESE, M. and TEMPO, R. (1985). Optimal algorithms theory for robust estimation and prediction. *IEEE Trans. Automat. Control* **30** 730–738. MR0794206 https://doi.org/10.1109/TAC.1985.1104056
[32] MO, W., TANG, W., XUE, S., LIU, Y. and ZHU, J. (2024). Minimax Regret Learning for Data with Heterogeneous Subgroups. arXiv Preprint. Available at arXiv:2405.01709.
[33] MOHRI, M., SIVEK, G. and SURESH, A. T. (2019). Agnostic federated learning. In *International Conference on Machine Learning* 4615–4625. PMLR.
[34] NAMKOONG, H. and DUCHI, J. C. (2017). Variance-based regularization with convex objectives. *Adv. Neural Inf. Process. Syst.* **30**.
[35] NGUYEN, X., WAINWRIGHT, M. J. and JORDAN, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Trans. Inf. Theory* **56** 5847–5861. MR2808937 https://doi.org/10.1109/TIT.2010.2068870
[36] PERONE, C. S., BALLESTER, P., BARROS, R. C. and COHEN-ADAD, J. (2019). Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *NeuroImage* **194** 1–11.
[37] QUINONERO-CANDELA, J., SUGIYAMA, M., SCHWAIGHOFER, A. and LAWRENCE, N. D. (2008). *Dataset Shift in Machine Learning*. MIT Press, Cambridge.
[38] REN, C.-X., XU, X.-L. and YAN, H. (2018). Generalized conditional domain adaptation: A causal perspective with low-rank translators. *IEEE Trans. Cybern.* **50** 821–834.

[39] ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974 https://doi.org/10.1093/biomet/70.1.41
[40] ROTHENHÄUSLER, D., MEINSHAUSEN, N. and BÜHLMANN, P. (2016). Confidence intervals for maximin effects in inhomogeneous large-scale data. In *Statistical Analysis for High-Dimensional Data*. *Abel Symp*. **11** 255–277. Springer, Cham. MR3616272
[41] SAGAWA, S., KOH, P. W., HASHIMOTO, T. B. and LIANG, P. (2019). Distributionally robust neural networks for group shifts: on the importance of regularization for worst-case generalization. arXiv Preprint. Available at arXiv:1911.08731.
[42] SAITO, K., WATANABE, K., USHIKU, Y. and HARADA, T. (2018). Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 3723–3732.
[43] SCHMIDT-HIEBER, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *Ann*. *Statist*. **48** 1875–1897. MR4134774 https://doi.org/10.1214/19-AOS1875
[44] SCHULAM, P. and SARIA, S. (2015). A framework for individualizing predictions of disease trajectories by exploiting multi-resolution structure. *Adv. Neural Inf. Process*. *Syst*. **28**.
[45] SCORNET, E., BIAU, G. and VERT, J.-P. (2015). Consistency of random forests. *Ann*. *Statist*. **43** 1716–1741. MR3357876 https://doi.org/10.1214/15-AOS1321
[46] SINHA, A., NAMKOONG, H., VOLPI, R. and DUCHI, J. (2017). Certifying some distributional robustness with principled adversarial training. arXiv Preprint. Available at arXiv:1710.10571.
[47] SOMA, T., GATMIRY, K. and JEGELKA, S. (2022). Optimal algorithms for group distributionally robust optimization and beyond. arXiv Preprint. Available at arXiv:2212.13669.
[48] SUGIYAMA, M., NAKAJIMA, S., KASHIMA, H., BUENAU, P. and KAWANABE, M. (2007). Direct importance estimation with model selection and its application to covariate shift adaptation. *Adv. Neural Inf. Process*. *Syst*. **20**.
[49] SUGIYAMA, M., SUZUKI, T. and KANAMORI, T. (2012). *Density Ratio Estimation in Machine Learning*. Cambridge Univ. Press, Cambridge. MR2895762 https://doi.org/10.1017/CBO9781139035613
[50] TAI, A. P., MICKLEY, L. J. and JACOB, D. J. (2010). Correlations between fine particulate matter (PM2. 5) and meteorological variables in the United States: Implications for the sensitivity of PM2. 5 to climate change. *Atmos*. *Environ*. **44** 3976–3984.
[51] TOPALOGLU, M. Y., MORRELL, E. M., RAJENDRAN, S. and TOPALOGLU, U. (2021). In the pursuit of privacy: The promises and predicaments of federated learning in healthcare. *Front*. *Artif. Intell*. 147.
[52] TZENG, E., HOFFMAN, J., SAENKO, K. and DARRELL, T. (2017). Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 7167–7176.
[53] WANG, Z., BÜHLMANN, P. and GUO, Z. (2026). Supplement to "Distributionally robust learning for multi-source unsupervised domain adaptation." https://doi.org/10.1214/25-AOS2578SUPP
[54] WEISS, K., KHOSHGOFTAAR, T. M. and WANG, D. (2016). A survey of transfer learning. *J. Big Data* **3** 1–40.
[55] XIA, Y., YANG, M. and WANG, S. (2021). Chebyshev center of the intersection of balls: Complexity, relaxation and approximation. *Math*. *Program*. **187** 287–315. MR4246304 https://doi.org/10.1007/s10107-020-01479-0
[56] XIONG, X., GUO, Z. and CAI, T. (2023). Distributionally robust transfer learning. arXiv preprint. Available at arXiv:2309.06534.
[57] XU, S., FREUND, R. M. and SUN, J. (2003). Solution methodologies for the smallest enclosing circle problem. *Comput*. *Optim*. *Appl*. **25** 283–292.
[58] XU, Y., XUE, W., LEI, Y., ZHAO, Y., CHENG, S., REN, Z. and HUANG, Q. (2018). Impact of meteorological conditions on PM2. 5 pollution in China during winter. *Atmosphere* **9** 429.
[59] ZHANG, J., MENON, A., VEIT, A., BHOJANAPALLI, S., KUMAR, S. and SRA, S. (2020). Coping with label shift via distributionally robust optimisation. arXiv Preprint. Available at arXiv:2010.12230.
[60] ZHANG, S., GUO, B., DONG, A., HE, J., XU, Z. and CHEN, S. X. (2017). Cautionary tales on air-quality improvement in Beijing. *Proc*. *R. Soc*. *A*, *Math*. *Phys*. *Eng*. *Sci*. **473** 20170457.
[61] ZHANG, Y., HUANG, M. and IMAI, K. (2024). Minimax Regret Estimation for Generalizing Heterogeneous Treatment Effects with Multisite Data. arXiv Preprint. Available at arXiv:2412.11136.
[62] ZHUANG, F., QI, Z., DUAN, K., XI, D., ZHU, Y., ZHU, H., XIONG, H. and HE, Q. (2020). A comprehensive survey on transfer learning. *Proc*. *IEEE* **109** 43–76.

The Annals of Statistics
2026, Vol. 54, No. 2, 597–620
https://doi.org/10.1214/25-AOS2579

TRACE TEST FOR HIGH-DIMENSIONAL COINTEGRATION

BY ALEXEI ONATSKI[1,a] AND CHEN WANG[2,b]

[1]*Faculty of Economics, University of Cambridge,* [a]*ao319@cam.ac.uk*
[2]*Department of Statistics and Actuarial Science, University of Hong Kong,* [b]*stacw@hku.hk*

This paper studies Johansen's (*J. Econom. Dynam. Control* **12** (1988) 231–254) trace test for cointegration in high-dimensional data. We show that when both cross-sectional and temporal dimension of the data go to infinity proportionally, the shifted and scaled modified trace statistic converges to a Gaussian random variable. We give explicit formulae for the shift and scale parameters as well as for the mean and variance of the Gaussian limit. Monte Carlo analysis shows excellent size properties of the asymptotic test, which is an improvement over the Bartlett-corrected versions of the original trace test, especially for relatively large ratios of the dimensionality to the sample size. The Monte Carlo also reveals a nonmonotonicity of the power of the test. We comment on the source of such a nonmonotonicity.

REFERENCES

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

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

BAO, Z., HU, J., XU, X. and ZHANG, X. (2024). Spectral statistics of sample block correlation matrices. *Ann. Statist.* **52** 1873–1898. MR4828865 https://doi.org/10.1214/24-AOS2375

BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York. MR0233396

BILLINGSLEY, P. (1995). *Probability and Measure*, 3rd ed. *Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. MR1324786

BYKHOVSKAYA, A. and GORIN, V. (2022). Cointegration in large VARs. *Ann. Statist.* **50** 1593–1617. MR4441133 https://doi.org/10.1214/21-aos2164

BYKHOVSKAYA, A. and GORIN, V. (2024). Asymptotics of cointegration tests for high-dimensional VAR(k). *Rev. Econ. Stat.* forthcoming.

COHEN, A. and SACKROWITZ, H. B. (1998). Directional tests for one-sided alternatives in multivariate models. *Ann. Statist.* **26** 2321–2338. MR1700234 https://doi.org/10.1214/aos/1024691473

DUMITRIU, I. and PAQUETTE, E. (2012). Global fluctuations for linear statistics of β-Jacobi ensembles. *Random Matrices Theory Appl.* **1** 1250013, 60. MR3039374 https://doi.org/10.1142/S201032631250013X

GONZALO, J. and PITARAKIS, J.-Y. (1995). Comovements in Large Systems 10. Universidad Carlos III de Madrid. Statistics and Econometrics Series, Working Paper 95-38.

GONZALO, J. and PITARAKIS, J.-Y. (1999). Dimensionality effect in cointegration analysis. In *Cointegration, Causality, and Forecasting. A Festschrift in Honour of Clive WJ Granger* 212–229. Oxford Univ. Press, Oxford.

HO, M. S. and SORENSEN, B. E. (1996). Finding cointegration rank in high dimensional systems using the Johansen test: An illustration using data based mote Carlo simulations. *Rev. Econ. Stat.* **78** 726–732.

JIN, B., WANG, C., MIAO, B. and LO HUANG, M.-N. (2009). Limiting spectral distribution of large-dimensional sample covariance matrices generated by VARMA. *J. Multivariate Anal.* **100** 2112–2125. MR2543090 https://doi.org/10.1016/j.jmva.2009.06.011

JOHANSEN, S. (1988). Statistical analysis of cointegration vectors. *J. Econom. Dynam. Control* **12** 231–254. MR0986516 https://doi.org/10.1016/0165-1889(88)90041-3

JOHANSEN, S. (1995). *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford Univ. Press, New York. MR1487375 https://doi.org/10.1093/0198774508.001.0001

JOHANSEN, S. (2002). A small sample correction for the test of cointegrating rank in the vector autoregressive model. *Econometrica* **70** 1929–1961. MR1925161 https://doi.org/10.1111/1468-0262.00358

MSC2020 subject classifications. Primary 62M10; secondary 60B20, 60F05.

Key words and phrases. High-dimensional VAR, cointegration test, Johansen's trace statistic, central limit theorem, linear spectral statistics.

JOHANSEN, S., HANSEN, H. and FACHIN, S. (2005). A simulation study of some functionals of random walk. Manuscript available at: http://www.math.ku.dk/~sjo/.
JOHNSTONE, I. M. (2009). Approximate null distribution of the largest root in multivariate analysis. *Ann. Appl. Stat.* **3** 1616–1633. MR2752150 https://doi.org/10.1214/08-AOAS220
LIU, Z., HU, J., BAI, Z. and SONG, H. (2023). A CLT for the LSS of large-dimensional sample covariance matrices with diverging spikes. *Ann. Statist.* **51** 2246–2271. MR4678803 https://doi.org/10.1214/23-aos2333
LOGAN, B. R. (2003). A cone order monotone test for the one-sided multivariate testing problem. *Statist. Probab. Lett.* **63** 315–323. MR1986331 https://doi.org/10.1016/S0167-7152(03)00097-X
ONATSKI, A. and WANG, C. (2018). Alternative asymptotics for cointegration tests in large VARs. *Econometrica* **86** 1465–1478. MR3843495 https://doi.org/10.3982/ECTA14649
ONATSKI, A. and WANG, C. (2019). Extreme canonical correlations and high-dimensional cointegration analysis. *J. Econometrics* **212** 307–322. MR3994019 https://doi.org/10.1016/j.jeconom.2019.04.032
ONATSKI, A. and WANG, C. (2026). Supplement to "Trace test for high-dimensional cointegration." https://doi.org/10.1214/25-AOS2579SUPP
PAN, G. (2014). Comparison between two types of large sample covariance matrices. *Ann. Inst. Henri Poincaré Probab. Stat.* **50** 655–677. MR3189088 https://doi.org/10.1214/12-AIHP506
SAVIN, N. E. and WURTZ, A. H. (1999). Power of tests in binary response models. *Econometrica* **67** 413–421.
SILVERSTEIN, J. W. and BAI, Z. D. (1995). On the empirical distribution of eigenvalues of a class of large-dimensional random matrices. *J. Multivariate Anal.* **54** 175–192. MR1345534 https://doi.org/10.1006/jmva.1995.1051
TRACY, C. A. and WIDOM, H. (1996). On orthogonal and symplectic matrix ensembles. *Comm. Math. Phys.* **177** 727–754. MR1385083
YANG, Y. and PAN, G. (2015). Independence test for high dimensional data based on regularized canonical correlation coefficients. *Ann. Statist.* **43** 467–500. MR3316187 https://doi.org/10.1214/14-AOS1284
ZHENG, S. (2012). Central limit theorems for linear spectral statistics of large dimensional F-matrices. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 444–476. MR2954263 https://doi.org/10.1214/11-AIHP414
ZHENG, S., BAI, Z. and YAO, J. (2015). Substitution principle for CLT of linear spectral statistics of high-dimensional sample covariance matrices with applications to hypothesis testing. *Ann. Statist.* **43** 546–591. MR3316190 https://doi.org/10.1214/14-AOS1292
ZHENG, S., BAI, Z. and YAO, J. (2017). CLT for eigenvalue statistics of large-dimensional general Fisher matrices with applications. *Bernoulli* **23** 1130–1178. MR3606762 https://doi.org/10.3150/15-BEJ772

The Annals of Statistics
2026, Vol. 54, No. 2, 621–646
https://doi.org/10.1214/25-AOS2580

ESTIMATION OF GROUPED TIME-VARYING NETWORK VECTOR AUTOREGRESSIVE MODELS

BY DEGUI LI[1,a], BIN PENG[2,b], SONGQIAO TANG[3,c] AND WEIBIAO WU[4,d]

[1]*Faculty of Business Administration, Asia-Pacific Academy of Economics and Management, and Department of Economics, University of Macau,* [a]*deguili@um.edu.mo*

[2]*Department of Econometrics and Business Statistics, Monash University,* [b]*Bin.Peng@monash.edu*

[3]*School of Mathematical Sciences, Zhejiang University,* [c]*statsq@zju.edu.cn*

[4]*Department of Statistics, University of Chicago,* [d]*wbwu@uchicago.edu*

This paper introduces a flexible time-varying network vector autoregressive model framework for large-scale time series. A latent group structure is imposed on the heterogeneous and node-specific time-varying momentum and network spillover effects so that the number of unknown time-varying coefficients to be estimated can be reduced considerably. A classic agglomerative clustering algorithm with nonparametrically estimated distance matrix is combined with a ratio criterion to consistently estimate the latent group number and membership. A postgrouping local linear smoothing method is proposed to estimate the group-specific time-varying momentum and network effects, substantially improving the convergence rates of the preliminary estimates which ignore the latent structure. We further modify the methodology and theory to allow for structural breaks in either the group membership, group number or group-specific coefficient functions. Numerical studies including Monte-Carlo simulation and an empirical application are presented to examine the finite-sample performance of the developed model and methodology.

REFERENCES

ANDO, T. and BAI, J. (2017). Clustering huge number of financial time series: A panel data approach with high-dimensional predictors and factor structures. *J. Amer. Statist. Assoc.* **112** 1182–1198. MR3735369 https://doi.org/10.1080/01621459.2016.1195743

BAI, J., LI, K. and LU, L. (2016). Estimation and inference of FAVAR models. *J. Bus. Econom. Statist.* **34** 620–641. MR3548000 https://doi.org/10.1080/07350015.2015.1111222

BAI, J. and NG, S. (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica* **74** 1133–1150. MR2238213 https://doi.org/10.1111/j.1468-0262.2006.00696.x

BARIGOZZI, M. and BROWNLEES, C. (2019). NETS: Network estimation for time series. *J. Appl. Econometrics* **34** 347–364. MR3948470 https://doi.org/10.1002/jae.2676

BASU, S. and MICHAILIDIS, G. (2015). Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.* **43** 1535–1567. MR3357870 https://doi.org/10.1214/15-AOS1315

BERNANKE, B., BOIVIN, J. and ELIASZ, P. S. (2005). Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach. *Q. J. Econ.* **120** 387–422.

BONHOMME, S. and MANRESA, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica* **83** 1147–1184. MR3357486 https://doi.org/10.3982/ECTA11319

BUCHWALTER, B., DIEBOLD, F. and YILMAZ, K. (2025). Clustered network connectedness: A new measurement framework with application to global equity markets. Working paper. Available at arXiv:2502.15458.

CHEN, E. Y., FAN, J. and ZHU, X. (2023). Community network auto-regression for high-dimensional time series. *J. Econometrics* **235** 1239–1256. MR4602909 https://doi.org/10.1016/j.jeconom.2022.10.005

CHEN, J. (2019). Estimating latent group structure in time-varying coefficient panel data models. *Econom. J.* **22** 223–240. MR4021122 https://doi.org/10.1093/ectj/utz008

MSC2020 subject classifications. Primary 62M10; secondary 62G05, 62G10.

Key words and phrases. Cluster analysis, network VAR, latent groups, local linear estimator, time-varying coefficients.

CHEN, J., LI, D., LI, Y.-N. and LINTON, O. (2025). Estimating time-varying networks for high-dimensional time series. *J. Econometrics* **249** Paper No. 105941, 21 pp. MR4906021 https://doi.org/10.1016/j.jeconom.2024.105941

CHEN, L., WANG, W. and WU, W. B. (2022). Inference of breakpoints in high-dimensional time series. *J. Amer. Statist. Assoc.* **117** 1951–1963. MR4528482 https://doi.org/10.1080/01621459.2021.1893178

CHO, H. and FRYZLEWICZ, P. (2012). Multiscale and multilevel technique for consistent segmentation of nonstationary time series. *Statist. Sinica* **22** 207–229. MR2933173 https://doi.org/10.5705/ss.2009.280

CHO, H. and FRYZLEWICZ, P. (2015). Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 475–507. MR3310536 https://doi.org/10.1111/rssb.12079

DAVIS, R. A., ZANG, P. and ZHENG, T. (2016). Sparse vector autoregressive modeling. *J. Comput. Graph. Statist.* **25** 1077–1096. MR3572029 https://doi.org/10.1080/10618600.2015.1092978

DEMIRER, M., DIEBOLD, F. X., LIU, L. and YILMAZ, K. (2018). Estimating global bank network connectedness. *J. Appl. Econometrics* **33** 1–15. MR3771571 https://doi.org/10.1002/jae.2585

DIEBOLD, F. X. and YILMAZ, K. (2014). On the network topology of variance decompositions: Measuring the connectedness of financial firms. *J. Econometrics* **182** 119–134. MR3212765 https://doi.org/10.1016/j.jeconom.2014.04.012

DING, X., QIU, Z. and CHEN, X. (2017). Sparse transition matrix estimation for high-dimensional and locally stationary vector autoregressive models. *Electron. J. Stat.* **11** 3871–3902. MR3714301 https://doi.org/10.1214/17-EJS1325

ENIKEEVA, F., KLOPP, O. and ROUSSELOT, M. (2025). Change-point detection in low-rank VAR processes. *Bernoulli* **31** 1058–1083. MR4863067 https://doi.org/10.3150/24-bej1760

EVERITT, B. S., LANDAU, S., LEESE, M. and STAHL, D. (2011). *Cluster Analysis*, 5th ed. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR3155074 https://doi.org/10.1002/9780470977811

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

GUÐMUNDSSON, G. M. S. and BROWNLEES, C. (2021). Detecting groups in large vector autoregressions. *J. Econometrics* **225** 2–26. MR4314058 https://doi.org/10.1016/j.jeconom.2021.03.012

HANLON, H. M., BERNIE, D., CARIGI, G. and LOWE, J. A. (2021). Future changes to high impact weather in the UK. *Clim. Change* **166** 50.

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

KE, Y., LI, J. and ZHANG, W. (2016). Structure identification in panel data analysis. *Ann. Statist.* **44** 1193–1233. MR3485958 https://doi.org/10.1214/15-AOS1403

KENDON, M., MCCARTHY, M., JEVREJEVA, S., MATTHEWS, A., SPARKS, T. and GARFORTH, J. (2020). State of the UK climate 2019. *Int. J. Climatol.* **40** 1–69.

KILIAN, K. and LÜTKEPOHL, H. (2017). *Structural Vector Autoregressive Analysis*. Cambridge Univ. Press, Cambridge.

KOCK, A. B. and CALLOT, L. (2015). Oracle inequalities for high dimensional vector autoregressions. *J. Econometrics* **186** 325–344. MR3343790 https://doi.org/10.1016/j.jeconom.2015.02.013

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

LI, D., PENG, B., TANG, S. and WU, W. (2026). Supplement to "Estimation of grouped time-varying network vector autoregressive models." https://doi.org/10.1214/25-AOS2580SUPP

LI, D., ROBINSON, P. M. and SHANG, H. L. (2020). Long-range dependent curve time series. *J. Amer. Statist. Assoc.* **115** 957–971. MR4107692 https://doi.org/10.1080/01621459.2019.1604362

LUMSDAINE, R. L., OKUI, R. and WANG, W. (2023). Estimation of panel group structure models with structural breaks in group memberships and coefficients. *J. Econometrics* **233** 45–65. MR4554726 https://doi.org/10.1016/j.jeconom.2022.01.001

LÜTKEPOHL, H. (2006). *New Introduction to Multiple Time Series Analysis*. Springer, Berlin. MR2172368 https://doi.org/10.1007/978-3-540-27752-1

MIAO, K., PHILLIPS, P. C. B. and SU, L. (2023). High-dimensional VARs with common factors. *J. Econometrics* **233** 155–183. MR4554731 https://doi.org/10.1016/j.jeconom.2022.02.002

PORTMANN, R. W., SOLOMON, S. and HEGERL, G. C. (2009). Spatial and seasonal patterns in climate change, temperatures, and precipitation across the United States. *Proc. Natl. Acad. Sci. USA* **106** 7324–7329.

SAFIKHANI, A. and SHOJAIE, A. (2022). Joint structural break detection and parameter estimation in high-dimensional nonstationary VAR models. *J. Amer. Statist. Assoc.* **117** 251–264. MR4399083 https://doi.org/10.1080/01621459.2020.1770097

SCOTT, J. (2017). *Social Network Analysis*, 4th ed. Sage, London.
SIMS, C. A. (1980). Macroeconomics and reality. *Econometrica* **48** 1–48.
SU, L., SHI, Z. and PHILLIPS, P. C. B. (2016). Identifying latent structures in panel data. *Econometrica* **84** 2215–2264. MR3580267 https://doi.org/10.3982/ECTA12560
SU, L. and WANG, X. (2017). On time-varying factor models: Estimation and testing. *J. Econometrics* **198** 84–101. MR3628100 https://doi.org/10.1016/j.jeconom.2016.12.004
SUN, Y. (2016). Functional-coefficient spatial autoregressive models with nonparametric spatial weights. *J. Econometrics* **195** 134–153. MR3545298 https://doi.org/10.1016/j.jeconom.2016.07.005
SUN, Y. and MALIKOV, E. (2018). Estimation and inference in functional-coefficient spatial autoregressive panel data models with fixed effects. *J. Econometrics* **203** 359–378. MR3770832 https://doi.org/10.1016/j.jeconom.2017.12.006
TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. Springer, New York. MR2724359 https://doi.org/10.1007/b13794
VOGT, M. and LINTON, O. (2017). Classification of non-parametric regression functions in longitudinal data models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 5–27. MR3597962 https://doi.org/10.1111/rssb.12155
VOGT, M. and LINTON, O. (2020). Multiscale clustering of nonparametric regression curves. *J. Econometrics* **216** 305–325. MR4077396 https://doi.org/10.1016/j.jeconom.2020.01.020
WANG, Y., PHILLIPS, P. C. B. and SU, L. (2024). Panel data models with time-varying latent group structures. *J. Econometrics* **240** Paper No. 105685, 24 pp. MR4695637 https://doi.org/10.1016/j.jeconom.2024.105685
WU, B. (2019). Time-varying network vector autoregression model. Working paper. Available at http://dx.doi.org/10.2139/ssrn.3491608.
WU, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proc. Natl. Acad. Sci. USA* **102** 14150–14154. MR2172215 https://doi.org/10.1073/pnas.0506715102
XU, M., CHEN, X. and WU, W. B. (2020). Estimation of dynamic networks for high-dimensional nonstationary time series. *Entropy* **22** Paper No. 55, 27 pp. MR4057818 https://doi.org/10.3390/e22010055
YANG, X., CHEN, J., LI, D. and LI, R. (2024). Functional-coefficient quantile regression for panel data with latent group structure. *J. Bus. Econom. Statist.* **42** 1026–1040. MR4757123 https://doi.org/10.1080/07350015.2023.2277172
YIN, H., SAFIKHANI, A. and MICHAILIDIS, G. (2023). A general modeling framework for network autoregressive processes. *Technometrics* **65** 579–589. MR4662690 https://doi.org/10.1080/00401706.2023.2203184
YIN, H., SAFIKHANI, A. and MICHAILIDIS, G. (2026). A functional coefficients network autoregressive model. *Statist. Sinica* **36** 305–329. MR5005317 https://doi.org/10.5705/ss.202022.0402
ZHANG, D. and WU, W. B. (2021). Convergence of covariance and spectral density estimates for high-dimensional locally stationary processes. *Ann. Statist.* **49** 233–254. MR4206676 https://doi.org/10.1214/20-AOS1954
ZHANG, T. (2013). Clustering high-dimensional time series based on parallelism. *J. Amer. Statist. Assoc.* **108** 577–588. MR3174643 https://doi.org/10.1080/01621459.2012.760458
ZHU, X. and PAN, R. (2020). Grouped network vector autoregression. *Statist. Sinica* **30** 1437–1462. MR4257540 https://doi.org/10.5705/ss.202017.0533
ZHU, X., PAN, R., LI, G., LIU, Y. and WANG, H. (2017). Network vector autoregression. *Ann. Statist.* **45** 1096–1123. MR3662449 https://doi.org/10.1214/16-AOS1476
ZHU, X., XU, G. and FAN, J. (2025). Simultaneous estimation and group identification for network vector autoregressive model with heterogeneous nodes. *J. Econometrics* **249** Paper No. 105564, 19 pp. MR4905996 https://doi.org/10.1016/j.jeconom.2023.105564

The Annals of Statistics
2026, Vol. 54, No. 2, 647–666
https://doi.org/10.1214/25-AOS2581

LARGE-SCALE MULTIPLE TESTING: FUNDAMENTAL LIMITS OF FALSE DISCOVERY RATE CONTROL AND COMPOUND ORACLE

BY YUTONG NIE[a] AND YIHONG WU[b]

Department of Statistics and Data Science, Yale University, [a]*ytnie1998@gmail.com,* [b]*yihong.wu@yale.edu*

The false discovery rate (FDR) and the false nondiscovery rate (FNR), defined as the expected false discovery proportion (FDP) and the false nondiscovery proportion (FNP), are the most popular benchmarks for multiple testing. Despite the theoretical and algorithmic advances in recent years, the optimal trade-off between the FDR and the FNR has been largely unknown, except for certain restricted classes of decision rules, for example, separable rules, or for other performance metrics, for example, the marginal FDR and the marginal FNR (mFDR and mFNR). In this paper we determine the asymptotically optimal FDR-FNR trade-off under the two-group random mixture model when the number of hypotheses tends to infinity. Distinct from the optimal mFDR-mFNR trade-off, which is achieved by separable decision rules, the optimal FDR-FNR trade-off requires compound rules, even in the large-sample limit and for models as simple as the Gaussian location model. This suboptimality of separable rules also holds for other objectives, such as maximizing the expected number of true discoveries. Finally, to address the limitation of the FDR, which only controls the expectation but not the fluctuation of the FDP, we also determine the optimal tradeoff when the FDP is controlled with high probability and show it coincides with that of the mFDR and the mFNR. Extensions to models with a fixed nonnull proportion are also obtained.

REFERENCES

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

[2] ARIAS-CASTRO, E. and CHEN, S. (2017). Distribution-free multiple testing. *Electron. J. Stat.* **11** 1983–2001. MR3651021 https://doi.org/10.1214/17-EJS1277

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

[4] BASU, P., FU, L., SARETTO, A. and SUN, W. (2024). An empirical Bayes approach to controlling the false discovery exceedance. *J. Bus. Econom. Statist.* **42** 1041–1052. MR4757124 https://doi.org/10.1080/07350015.2023.2277857

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

[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 https://doi.org/10.1093/biomet/93.3.491

[7] 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.

[8] BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245 https://doi.org/10.1214/aos/1013699998

[9] BIRGÉ, L. (1989). The Grenander estimator: A nonasymptotic approach. *Ann. Statist.* **17** 1532–1549. MR1026298 https://doi.org/10.1214/aos/1176347380

[10] BLAHUT, R. E. (1974). Hypothesis testing and information theory. *IEEE Trans. Inf. Theory* **IT-20** 405–417. MR0396072 https://doi.org/10.1109/tit.1974.1055254

MSC2020 subject classifications. Primary 62G10, 62C10; secondary 62C12.
Key words and phrases. False discovery rate, multiple testing, compound decision rule.

[11] CAI, T. T. and SUN, W. (2017). Optimal screening and discovery of sparse signals with applications to multistage high throughput studies. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 197–223. MR3597970 https://doi.org/10.1111/rssb.12171

[12] CAO, H., CHEN, J. and ZHANG, X. (2022). Optimal false discovery rate control for large scale multiple testing with auxiliary information. *Ann. Statist.* **50** 807–857. MR4404920 https://doi.org/10.1214/21-aos2128

[13] CHI, Z. and TAN, Z. (2008). Positive false discovery proportions: Intrinsic bounds and adaptive control. *Statist. Sinica* **18** 837–860. MR2440397

[14] DÖHLER, S. and ROQUAIN, E. (2020). Controlling the false discovery exceedance for heterogeneous tests. *Electron. J. Stat.* **14** 4244–4272. MR4185850 https://doi.org/10.1214/20-EJS1771

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

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

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

[18] 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 https://doi.org/10.1214/07-AOS586

[19] GENOVESE, C. and WASSERMAN, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 499–517. MR1924303 https://doi.org/10.1111/1467-9868.00347

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

[21] GENOVESE, C. R. and WASSERMAN, L. (2006). Exceedance control of the false discovery proportion. *J. Amer. Statist. Assoc.* **101** 1408–1417. MR2279468 https://doi.org/10.1198/016214506000000339

[22] HELLER, R. and ROSSET, S. (2021). Optimal control of false discovery criteria in the two-group model. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **83** 133–155. MR4220987 https://doi.org/10.1111/rssb.12403

[23] HOEFFDING, W. (1965). Asymptotically optimal tests for multinomial distributions. *Ann. Math. Statist.* **36** 369–408. MR0173322 https://doi.org/10.1214/aoms/1177700150

[24] LEHMANN, E. L. and ROMANO, J. P. (2012). Generalizations of the familywise error rate. In *Selected Works of EL Lehmann* 719–735. Springer, Berlin.

[25] NEYMAN, J. and PEARSON, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. Lond., Ser. A, Contain. Pap. Math. Phys. Character* **231** 289–337.

[26] NIE, Y. and WU, Y. (2026). Supplement to "Large-scale Multiple Testing: Fundamental Limits of False Discovery Rate Control and Compound Oracle." https://doi.org/10.1214/25-AOS2581SUPP

[27] RABINOVICH, M., RAMDAS, A., JORDAN, M. I. and WAINWRIGHT, M. J. (2020). Optimal rates and tradeoffs in multiple testing. *Statist. Sinica* **30** 741–762. MR4214160

[28] ROBBINS, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 1950 131–148. Univ. California Press, Berkeley, CA. MR0044803

[29] ROSSET, S., HELLER, R., PAINSKY, A. and AHARONI, E. (2022). Optimal and maximin procedures for multiple testing problems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 1105–1128. MR4494154 https://doi.org/10.1111/rssb.12507

[30] SARKAR, S. K. (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Ann. Statist.* **30** 239–257. MR1892663 https://doi.org/10.1214/aos/1015362192

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

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

[33] SUN, W. and CAI, T. T. (2009). Large-scale multiple testing under dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 393–424. MR2649603 https://doi.org/10.1111/j.1467-9868.2008.00694.x

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

[35] XIE, J., CAI, T. T., MARIS, J. and LI, H. (2011). Optimal false discovery rate control for dependent data. *Stat. Interface* **4** 417–430. MR2868825 https://doi.org/10.4310/SII.2011.v4.n4.a1

The Annals of Statistics
2026, Vol. 54, No. 2, 667–691
https://doi.org/10.1214/25-AOS2582

ANALYSIS OF SINGULAR SUBSPACES UNDER RANDOM PERTURBATIONS

BY KE WANG[a]

Department of Mathematics, The Hong Kong University of Science and Technology, [a]*kewang@ust.hk*

We present a comprehensive analysis of singular vector and singular subspace perturbations in the signal-plus-noise matrix model with random Gaussian noise. Assuming a low-rank signal matrix, we extend the Davis–Kahan–Wedin theorem in a fully generalized manner, applicable to any unitarily invariant matrix norm, building on previous results by O'Rourke, Vu, and the author. Our analysis provides fine-grained insights, including ℓ_∞ bounds for singular vectors, $\ell_{2,\infty}$ bounds for singular subspaces, and results for linear and bilinear functions of singular vectors. Additionally, we derive $\ell_{2,\infty}$ bounds on perturbed singular vectors, taking into account the weighting by their corresponding singular values. Finally, we explore practical implications of these results in the Gaussian mixture model and the submatrix localization problem.

REFERENCES

[1] ABBE, E., FAN, J. and WANG, K. (2022). An ℓ_p theory of PCA and spectral clustering. *Ann. Statist.* **50** 2359–2385. https://doi.org/10.1214/22-aos2196

[2] ABBE, E., FAN, J., WANG, K. and ZHONG, Y. (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *Ann. Statist.* **48** 1452–1474.

[3] AGTERBERG, J. (2023). Distributional theory and statistical inference for linear functions of eigenvectors with small eigengaps. arXiv preprint. Available at arXiv:2308.02480.

[4] AGTERBERG, J., LUBBERTS, Z. and PRIEBE, C. E. (2022). Entrywise estimation of singular vectors of low-rank matrices with heteroskedasticity and dependence. *IEEE Trans. Inf. Theory* **68** 4618–4650.

[5] ALLEZ, R. and BOUCHAUD, J.-P. (2014). Eigenvector dynamics under free addition. *Random Matrices Theory Appl.* **3** 1450010. MR3256861 https://doi.org/10.1142/S2010326314500105

[6] ALON, N., KRIVELEVICH, M. and SUDAKOV, B. (1998). Finding a large hidden clique in a random graph. *Random Structures Algorithms* **13** 457–466.

[7] ARIAS-CASTRO, E. and VERZELEN, N. (2014). Community detection in dense random networks. *Ann. Statist.* **42** 940–969.

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

[9] BAIK, J., AROUS, G. B., PÉCHÉ, S. et al. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.* **33** 1643–1697.

[10] BAIK, J. and SILVERSTEIN, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *J. Multivariate Anal.* **97** 1382–1408.

[11] BALAKRISHNAN, S., KOLAR, M., RINALDO, A., SINGH, A. and WASSERMAN, L. (2011). Statistical and computational tradeoffs in biclustering. In *NIPS* 2011 *Workshop on Computational Trade-Offs in Statistical Learning* 4.

[12] BAO, Z., DING, X., WANG, J. and WANG, K. (2022). Statistical inference for principal components of spiked covariance matrices. *Ann. Statist.* **50** 1144–1169. MR4404931 https://doi.org/10.1214/21-aos2143

[13] BAO, Z., DING, X. and WANG, K. (2021). Singular vector and singular subspace distribution for the matrix denoising model. *Ann. Statist.* **49** 370–392.

[14] BAO, Z. and WANG, D. (2021). Eigenvector distribution in the critical regime of BBP transition. *Probab. Theory Related Fields* 1–81.

[15] BENAYCH-GEORGES, F., ENRIQUEZ, N. and MICHAÏL, A. (2021). Eigenvectors of a matrix under random perturbation. *Random Matrices Theory Appl.* **10** 2150023. MR4260218 https://doi.org/10.1142/S2010326321500234

MSC2020 subject classifications. Primary 60B20; secondary 62H25, 62H30.

Key words and phrases. Singular vector perturbation, singular subspace perturbation, low-rank structures, random matrices, spectral clustering, mixture models, submatrix localization.

[16] BENAYCH-GEORGES, F., GUIONNET, A. and MAIDA, M. (2011). Fluctuations of the extreme eigenvalues of finite rank deformations of random matrices. *Electron. J. Probab.* **16** 1621–1662.
[17] BENAYCH-GEORGES, F. and NADAKUDITI, R. R. (2011). The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Adv. Math.* **227** 494–521. https://doi.org/10.1016/j.aim.2011.02.007
[18] BENAYCH-GEORGES, F. and NADAKUDITI, R. R. (2012). The singular values and vectors of low rank perturbations of large rectangular random matrices. *J. Multivariate Anal.* **111** 120–135. https://doi.org/10.1016/j.jmva.2012.04.019
[19] BENIGNI, L. (2020). Eigenvectors distribution and quantum unique ergodicity for deformed Wigner matrices. *Ann. Inst. Henri Poincaré Probab. Stat.* **56** 2822–2867.
[20] BHARDWAJ, A. and VU, V. (2023). Matrix perturbation: Davis-Kahan in the infinity norm. arXiv preprint. Available at arXiv:2304.00328.
[21] BHATIA, R. (1997). *Matrix Analysis. Graduate Texts in Mathematics* **169**. Springer, New York. MR1477662 https://doi.org/10.1007/978-1-4612-0653-8
[22] BLOEMENDAL, A., ERDŐS, L., KNOWLES, A., YAU, H.-T. and YIN, J. (2014). Isotropic local laws for sample covariance and generalized Wigner matrices. *Electron. J. Probab.* **19** 1–53. MR3183577 https://doi.org/10.1214/ejp.v19-3054
[23] BLOEMENDAL, A., KNOWLES, A., YAU, H.-T. and YIN, J. (2016). On the principal components of sample covariance matrices. *Probab. Theory Related Fields* **164** 459–552. MR3449395 https://doi.org/10.1007/s00440-015-0616-x
[24] BRENNAN, M., BRESLER, G. and HULEIHEL, W. (2018). Reducibility and computational lower bounds for problems with planted sparse structure. In *Conference on Learning Theory* 48–166. PMLR.
[25] BUTUCEA, C. and INGSTER, Y. I. (2013). Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli* **19** 2652–2688.
[26] BUTUCEA, C., INGSTER, Y. I. and SUSLINA, I. A. (2015). Sharp variable selection of a sparse submatrix in a high-dimensional noisy matrix. *ESAIM Probab. Stat.* **19** 115–134.
[27] CAI, C., LI, G., CHI, Y., POOR, H. V. and CHEN, Y. (2021). Subspace estimation from unbalanced and incomplete data matrices: $\ell_{2,\infty}$ statistical guarantees. *Ann. Statist.* **49** 944–967.
[28] CAI, T. T., LIANG, T. and RAKHLIN, A. (2017). Computational and statistical boundaries for submatrix localization in a large noisy matrix. *Ann. Statist.* **45** 1403–1430.
[29] 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.
[30] CANDÈS, E. J. and PLAN, Y. (2010). Matrix completion with noise. *Proc. IEEE* **98** 925–936.
[31] CANDÈS, E. J. and RECHT, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.* **9** 717–772. MR2565240 https://doi.org/10.1007/s10208-009-9045-5
[32] CAPE, J., TANG, M. and PRIEBE, C. E. (2019). The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *Ann. Statist.* **47** 2405–2439.
[33] CAPITAINE, M. (2018). Limiting eigenvectors of outliers for spiked information-plus-noise type matrices. *Sémin. Probab.* **XLIX** 119–164.
[34] CAPITAINE, M. and DONATI-MARTIN, C. (2021). Non universality of fluctuations of outlier eigenvectors for block diagonal deformations of Wigner matrices. *ALEA Lat. Amer. J. Probab. Math. Stat.* **18** 129–165.
[35] CHEN, Y., CHI, Y., FAN, J. and MA, C. (2021). Spectral methods for data science: A statistical perspective. *Found. Trends Mach. Learn.* **14** 566–806.
[36] CHEN, Y., FAN, J., MA, C. and WANG, K. (2019). Spectral method and regularized MLE are both optimal for top-K ranking. *Ann. Statist.* **47** 2204–2235.
[37] CHEN, Y. and XU, J. (2016). Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *J. Mach. Learn. Res.* **17** 882–938.
[38] CHENG, C., WEI, Y. and CHEN, Y. (2021). Tackling small eigen-gaps: Fine-grained eigenvector estimation and inference under heteroscedastic noise. *IEEE Trans. Inf. Theory* **67** 7380–7419.
[39] DADON, M., HULEIHEL, W. and BENDORY, T. (2024). Detection and recovery of hidden submatrices. *IEEE Trans. Signal Inf. Process. Netw.*
[40] DAVIS, C. and KAHAN, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.* **7** 1–46. MR0264450 https://doi.org/10.1137/0707001
[41] DEKEL, Y., GUREL-GUREVICH, O. and PERES, Y. (2014). Finding hidden cliques in linear time with high probability. *Combin. Probab. Comput.* **23** 29–49.
[42] EL KAROUI, N. (2007). Tracy-Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices. *Ann. Probab.* **35** 663–714.
[43] ELDRIDGE, J., BELKIN, M. and WANG, Y. (2018). Unperturbed: Spectral analysis beyond Davis-Kahan. In *Algorithmic Learning Theory* 321–358. PMLR.

[44] FAN, J., FAN, Y., HAN, X. and LV, J. (2022). Asymptotic theory of eigenvectors for large random matrices. *J. Amer. Statist. Assoc.* **117** 996–1009.

[45] FAN, J., WANG, W. and ZHONG, Y. (2018). An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation. *J. Mach. Learn. Res.* **18** 207–207. MR3827095

[46] FEIGE, U. and RON, D. (2010). Finding hidden cliques in linear time. In *Discrete Mathematics and Theoretical Computer Science. Discrete Mathematics and Theoretical Computer Science*. 189–204.

[47] HAJEK, B., WU, Y. and XU, J. (2016). Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Trans. Inf. Theory* **62** 2788–2797.

[48] JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693.

[49] KESHAVAN, R., MONTANARI, A. and OH, S. (2009). Matrix completion from noisy entries. *Adv. Neural Inf. Process. Syst.* **22**.

[50] KNOWLES, A. and YIN, J. (2013). The isotropic semicircle law and deformation of Wigner matrices. *Comm. Pure Appl. Math.* **66** 1663–1750. MR3103909 https://doi.org/10.1002/cpa.21450

[51] KOLAR, M., BALAKRISHNAN, S., RINALDO, A. and SINGH, A. (2011). Minimax localization of structural information in large noisy matrices. *Adv. Neural Inf. Process. Syst.* **24**.

[52] KOLTCHINSKII, V. and XIA, D. (2016). Perturbation of linear forms of singular vectors under Gaussian noise. In *High Dimensional Probability VII. Progress in Probability* **71** 397–423. Springer, Cham. MR3565274 https://doi.org/10.1007/978-3-319-40519-3_18

[53] LEI, L. (2019). Unified $\ell_{2\to\infty}$ eigenspace perturbation theory for symmetric random matrices. arXiv preprint. Available at arXiv:1909.04798.

[54] LI, G., CAI, C., POOR, H. V. and CHEN, Y. (2021). Minimax estimation of linear functions of eigenvectors in the face of small eigen-gaps. arXiv preprint. Available at arXiv:2104.03298.

[55] LÖFFLER, M., ZHANG, A. Y. and ZHOU, H. H. (2021). Optimality of spectral clustering in the Gaussian mixture model. *Ann. Statist.* **49** 2506–2530.

[56] LUO, Y., HAN, R. and ZHANG, A. R. (2021). A Schatten-q low-rank matrix perturbation analysis via perturbation projection error bound. *Linear Algebra Appl.* **630** 225–240.

[57] MA, Z. and WU, Y. (2015). Computational barriers in minimax submatrix detection. *Ann. Statist.* 1089–1116.

[58] MCSHERRY, F. (2001). Spectral partitioning of random graphs. In *Proceedings* 42*nd IEEE Symposium on Foundations of Computer Science* 529–537. IEEE.

[59] MONTANARI, A., REICHMAN, D. and ZEITOUNI, O. (2015). On the limitation of spectral methods: From the Gaussian hidden clique problem to rank-one perturbations of Gaussian tensors. *Adv. Neural Inf. Process. Syst.* **28**.

[60] 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 https://doi.org/10.1016/j.laa.2017.11.014

[61] O'ROURKE, S., VU, V. and WANG, K. (2024). Matrices with Gaussian noise: Optimal estimates for singular subspace perturbation. *IEEE Trans. Inf. Theory* **70** 1978–2002.

[62] PAUL, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17** 1617–1642.

[63] RUDELSON, M. and VERSHYNIN, R. (2010). Non-asymptotic theory of random matrices: Extreme singular values. In *Proceedings of the International Congress of Mathematicians. Vol. III* 1576–1602. Hindustan Book Agency, New Delhi.

[64] STEWART, G. W. and SUN, J. G. (1990). *Matrix Perturbation Theory. Computer Science and Scientific Computing*. Academic Press, Boston, MA. MR1061154

[65] TROPP, J. A. (2015). An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.* **8** 1–230. https://doi.org/10.1561/2200000048

[66] VON LUXBURG, U., BELKIN, M. and BOUSQUET, O. (2008). Consistency of spectral clustering. *Ann. Statist.* **36** 555–586.

[67] VU, V. (2018). A simple SVD algorithm for finding hidden partitions. *Combin. Probab. Comput.* **27** 124–140.

[68] VU, V. Q. and LEI, J. (2013). Minimax sparse principal subspace estimation in high dimensions. *Ann. Statist.* **41** 2905–2947.

[69] WANG, K. (2026). Supplement to "Analysis of singular subspaces under random perturbations." https://doi.org/10.1214/25-AOS2582SUPP

[70] WANG, R. (2015). Singular vector perturbation under Gaussian noise. *SIAM J. Matrix Anal. Appl.* **36** 158–177. MR3310977 https://doi.org/10.1137/130938177

[71] WEDIN, P.-A. (1972). Perturbation bounds in connection with singular value decomposition. *BIT* **12** 99–111. MR309968 https://doi.org/10.1007/bf01932678

[72] XIA, D. and ZHOU, F. (2019). The sup-norm perturbation of HOSVD and low rank tensor denoising. *J. Mach. Learn. Res.* **20** 1–42. MR3960915

[73] YAN, Y., CHEN, Y. and FAN, J. (2021). Inference for heteroskedastic PCA with missing data. arXiv preprint. Available at arXiv:2107.12365.
[74] YAN, Y. and WAINWRIGHT, M. J. (2024). Entrywise inference for missing panel data: A simple and instance-optimal approach. arXiv preprint. Available at arXiv:2401.13665.
[75] YU, Y., WANG, T. and SAMWORTH, R. J. (2015). A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika* **102** 315–323.
[76] ZHANG, A. Y. and ZHOU, H. H. (2022). Leave-one-out singular subspace perturbation analysis for spectral clustering. arXiv preprint. Available at arXiv:2205.14855.
[77] ZHONG, Y. (2017). Eigenvector under random perturbation: A nonasymptotic Rayleigh-Schrödinger theory. arXiv preprint. Available at arXiv:1702.00139.
[78] ZHONG, Y. and BOUMAL, N. (2018). Near-optimal bounds for phase synchronization. *SIAM J. Optim.* **28** 989–1016. MR3782406 https://doi.org/10.1137/17M1122025

The Annals of Statistics
2026, Vol. 54, No. 2, 692–717
https://doi.org/10.1214/25-AOS2583

VERSATILE DIFFERENTIALLY PRIVATE LEARNING FOR GENERAL LOSS FUNCTIONS

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

[1]*Guanghua School of Management, Peking University,* [a]*lu_qilong@stu.pku.edu.cn*
[2]*Department of Statistics and Data Science, Tsinghua University,* [b]*sxchen@tsinghua.edu.cn*
[3]*School of Mathematical Sciences and Center for Statistical Science, Peking University,* [c]*qiuyumou@math.pku.edu.cn*

This paper aims to provide a versatile privacy-preserving release mechanism along with a unified approach for subsequent parameter estimation and statistical inference. We propose a privacy mechanism based on zero-inflated symmetric multivariate Laplace (ZIL) noise, which requires no prior specification of subsequent analysis tasks, allows for general loss functions under minimal conditions, imposes no limit on the number of analyses, and is adaptable to increasing data volume in online scenarios. We derive the trade-off function for the proposed ZIL mechanism, which characterizes its privacy protection level. Furthermore, to formalize the local differential privacy (LDP) property of the ZIL mechanism, we extend the classical ε-LDP to a more general f-LDP framework. To address scenarios where only individual attribute values require protection, we propose attribute-level differential privacy (ADP) and its local version. Within the M-estimation framework, we introduce a novel doubly random (DR) corrected loss for the ZIL mechanism, which yields consistent and asymptotically normal M-estimates under differential privacy constraints. The proposed approach is computationally efficient and does not require numerical integration or differentiation for noisy data. It applies to a broad class of loss functions, including nonsmooth ones. Two alternative estimators for smooth loss are also proposed with asymptotic properties. The cost of privacy in terms of estimation efficiency for these three estimators is evaluated both theoretically and numerically.

REFERENCES

AMORINO, C. and GLOTER, A. (2025). Minimax rate for multivariate data under componentwise local differential privacy constraints. *Ann. Statist.* **53** 1176–1202. MR4925120 https://doi.org/10.1214/25-aos2497

APPLE DIFFERENTIAL PRIVACY TEAM (2017). Learning with privacy at scale.

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

AVELLA-MEDINA, M., BRADSHAW, C. and LOH, P.-L. (2023). Differentially private inference via noisy optimization. *Ann. Statist.* **51** 2067–2092. MR4678796 https://doi.org/10.1214/23-aos2321

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

BEIMEL, A., NISSIM, K. and STEMMER, U. (2016). Private learning and sanitization: Pure vs. approximate differential privacy. *Theory Comput.* **12** Paper No. 1, 61. MR3518176 https://doi.org/10.4086/toc.2016.v012a001

BIE, A., KAMATH, G. and SINGHAL, V. (2022). Private estimation with public data. In *Advances in Neural Information Processing Systems* **35** 18653–18666. Curran Associates, Red Hook.

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

CARROLL, R. J. and HALL, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.* **83** 1184–1186. MR0997599

MSC2020 subject classifications. Primary 62-11; secondary 68P27.

Key words and phrases. Differential privacy, M-estimation, symmetric multivariate Laplace distribution, zero-inflated symmetric multivariate Laplace distribution.

CHANG, J., HU, Q., KOLACZYK, E. D., YAO, Q. and YI, F. (2024). Edge differentially private estimation in the β-model via jittering and method of moments. *Ann. Statist.* **52** 708–728. MR4744193 https://doi.org/10.1214/24-aos2365

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

DING, B., KULKARNI, J. and YEKHANIN, S. (2017). Collecting telemetry data privately. In *Advances in Neural Information Processing Systems*, *Vol.* 30. Curran Associates, Red Hook.

DONG, J., ROTH, A. and SU, W. J. (2022). Gaussian differential privacy. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 3–54. With discussions and a reply by the authors. MR4400389

DUCHI, J. C., JORDAN, M. I. and WAINWRIGHT, M. J. (2013). Local privacy and minimax bounds: Sharp rates for probability estimation. In *Proceedings of the* 27*th International Conference on Neural Information Processing Systems*, *Vol.* 1 1529–1537. Curran Associates, Red Hook.

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

DUCHI, J. C. and RUAN, F. (2024). The right complexity measure in locally private estimation: It is not the Fisher information. *Ann. Statist.* **52** 1–51. MR4718406 https://doi.org/10.1214/22-aos2227

DWORK, C., KENTHAPADI, K., MCSHERRY, F., MIRONOV, I. and NAOR, M. (2006a). Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology—EUROCRYPT* 2006. *Lecture Notes in Computer Science* **4004** 486–503. Springer, Berlin. MR2423560 https://doi.org/10.1007/11761679_29

DWORK, C., MCSHERRY, F., NISSIM, K. and SMITH, A. (2006b). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*. *Lecture Notes in Computer Science* **3876** 265–284. Springer, Berlin. MR2241676 https://doi.org/10.1007/11681878_14

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

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

FAN, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.* **19** 1257–1272. MR1126324 https://doi.org/10.1214/aos/1176348248

FIRPO, S., GALVAO, A. F. and SONG, S. (2017). Measurement errors in quantile regression models. *J. Econometrics* **198** 146–164. MR3628103 https://doi.org/10.1016/j.jeconom.2017.02.002

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

KENT, D. and RUPPERT, D. (2024). Smoothness-penalized deconvolution (SPeD) of a density estimate. *J. Amer. Statist. Assoc.* **119** 2407–2417. MR4797950 https://doi.org/10.1080/01621459.2023.2259028

KIFER, D. and MACHANAVAJJHALA, A. (2014). Pufferfish: A framework for mathematical privacy definitions. *ACM Trans. Database Syst.* **39** Art. 3, 36. MR3238192 https://doi.org/10.1145/2514689

KOTZ, S., KOZUBOWSKI, T. J. and PODGÓRSKI, K. (2001). *The Laplace Distribution and Generalizations*: *A Revisit with Applications to Communications*, *Economics*, *Engineering*, *and Finance*. Birkhäuser, Boston, MA. MR1935481 https://doi.org/10.1007/978-1-4612-0173-1

LEI, J. (2011). Differentially private M-estimators. In *Advances in Neural Information Processing Systems*, *Vol.* 24. Curran Associates, Red Hook.

LU, Q., CHEN, S. X. and QIU, Y. (2026). Supplement to "Versatile differentially private learning for general loss functions." https://doi.org/10.1214/25-AOS2583SUPP

NISSIM, K., RASKHODNIKOVA, S. and SMITH, A. (2007). Smooth sensitivity and sampling in private data analysis. In *STOC'*07*—Proceedings of the* 39*th Annual ACM Symposium on Theory of Computing* 75–84. ACM, New York. MR2402430 https://doi.org/10.1145/1250790.1250803

PAN, R., REN, T., GUO, B., LI, F., LI, G. and WANG, H. (2022). A note on distributed quantile regression by pilot sampling and one-step updating. *J. Bus. Econom. Statist.* **40** 1691–1700. MR4492062 https://doi.org/10.1080/07350015.2021.1961789

RAJKUMAR, A. and AGARWAL, S. (2012). A differentially private stochastic gradient descent algorithm for multiparty classification. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*. *Proceedings of Machine Learning Research* **22** 933–941. PMLR.

STEFANSKI, L. and CARROLL, R. J. (1990). Deconvoluting kernel density estimators. *Statistics* **21** 169–184. MR1054861 https://doi.org/10.1080/02331889008802238

STEFANSKI, L. A. (1989). Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models. *Comm. Statist. Theory Methods* **18** 4335–4358. MR1046712 https://doi.org/10.1080/03610928908830159

STEINBERGER, L. (2024). Efficiency in local differential privacy. *Ann. Statist.* **52** 2139–2166. MR4829483 https://doi.org/10.1214/24-aos2425

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

WANG, D. and XU, J. (2019). On sparse linear regression in the local differential privacy model. In *International Conference on Machine Learning*, *Vol*. 97 6628–6637. PMLR.

WANG, H. J., STEFANSKI, L. A. and ZHU, Z. (2012). Corrected-loss estimation for quantile regression with covariate measurement errors. *Biometrika* **99** 405–421. MR2931262 https://doi.org/10.1093/biomet/ass005

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

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

YANG, R., APLEY, D. W., STAUM, J. and RUPPERT, D. (2020). Density deconvolution with additive measurement errors using quadratic programming. *J. Comput*. *Graph*. *Statist*. **29** 580–591. MR4153183 https://doi.org/10.1080/10618600.2019.1704294

ZHANG, Y., XU, Q., TANG, N. and QU, A. (2024). Differentially private data release for mixed-type data via latent factor models. *J. Mach*. *Learn*. *Res*. **25** Paper No. [116], 37. MR4749768

The Annals of Statistics
2026, Vol. 54, No. 2, 718–744
https://doi.org/10.1214/25-AOS2584

OPTIMAL EIGENVALUE SHRINKAGE IN THE SEMICIRCLE LIMIT

BY DAVID L. DONOHO[1,a] AND MICHAEL J. FELDMAN[2,b]

[1]*Department of Statistics, Stanford University,* [a]*donoho@stanford.edu*
[2]*Stern School of Business, New York University,* [b]*mjf529@stern.nyu.edu*

In response to the increasing dimensionality of modern datasets, recent theoretical studies of covariance estimation frequently adopt the *proportional-growth* asymptotic framework in which the sample size n and dimension p are comparable, with $n, p \to \infty$ and $\gamma_n := p/n \to \gamma > 0$. However, in many datasets—perhaps most—the sample size and dimension are highly imbalanced. To address this, we consider the *disproportional-growth* asymptotic framework, where $n, p \to \infty$ and $\gamma_n \to 0$ or $\gamma_n \to \infty$. These regimes give rise to novel behavior distinct from those in the proportional-growth and classical fixed-p settings.

We work under the spiked covariance model in which the theoretical covariance matrix is a low-rank perturbation of the identity. For each of 15 loss functions, we derive closed-form optimal shrinkage and thresholding rules; for several losses, optimality takes the particularly strong form of *unique asymptotic admissibility*. These optimal procedures involve substantial eigenvalue shrinkage and yield significant improvements over the standard empirical covariance estimator.

Practitioners may ask whether their data is better modeled under the proportional or disproportional frameworks and which of the corresponding procedures to apply. Fortunately, it is possible to remain framework-agnostic: a single, unified set of shrinkage rules—depending only on the aspect ratio γ_n of the given data—achieves asymptotic optimality in both regimes.

At the heart of these phenomena is the spiked Wigner model in which a low-rank matrix is perturbed by symmetric noise. Under both the (scaled) spiked covariance model as $\gamma_n \to 0$ and the spiked Wigner model, the empirical spectral distributions converge to the semicircle law. Exploiting this connection, we derive optimal spiked-Wigner shrinkage rules, which are of independent and fundamental interest.

REFERENCES

[1] ARNOLD, L. (1967). On the asymptotic distribution of the eigenvalues of random matrices. *J. Math. Anal. Appl.* **20** 262–268. MR0217833 https://doi.org/10.1016/0022-247X(67)90089-3

[2] BAI, Z. D. and YIN, Y. Q. (1988). Convergence to the semicircle law. *Ann. Probab.* **16** 863–875. MR0929083

[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 https://doi.org/10.1214/009117905000000233

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

[5] BENAYCH-GEORGES, F. and NADAKUDITI, R. R. (2011). The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Adv. Math.* **227** 494–521. MR2782201 https://doi.org/10.1016/j.aim.2011.02.007

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

MSC2020 subject classifications. 62H25, 60B20.

Key words and phrases. Covariance estimation, optimal shrinkage, spiked covariance, high-dimensional asymptotics, Wigner semicircle law.

[7] CAPITAINE, M., DONATI-MARTIN, C. and FÉRAL, D. (2009). The largest eigenvalues of finite rank deformation of large Wigner matrices: Convergence and nonuniversality of the fluctuations. *Ann. Probab.* **37** 1–47. MR2489158 https://doi.org/10.1214/08-AOP394

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

[9] DONOHO, D., GAVISH, M. and ROMANOV, E. (2023). *ScreeNOT*: Exact MSE-optimal singular value thresholding in correlated noise. *Ann. Statist.* **51** 122–148. MR4564851 https://doi.org/10.1214/22-aos2232

[10] DONOHO, D. L and FELDMAN, M. J (2026). Supplement to "Optimal eigenvalue shrinkage in the semicircle limit." https://doi.org/10.1214/25-AOS2584SUPP

[11] EL KAROUI, N. (2009). Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond. *Ann. Appl. Probab.* **19** 2362–2405. MR2588248 https://doi.org/10.1214/08-AAP548

[12] GAVISH, M. and DONOHO, D. L. (2014). The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Trans. Inf. Theory* **60** 5040–5053. MR3245370 https://doi.org/10.1109/TIT.2014.2323359

[13] GAVISH, M. and DONOHO, D. L. (2017). Optimal shrinkage of singular values. *IEEE Trans. Inf. Theory* **63** 2137–2152. MR3626861 https://doi.org/10.1109/TIT.2017.2653801

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

[15] LEDOIT, O. and WOLF, M. (2020). Analytical nonlinear shrinkage of large-dimensional covariance matrices. *Ann. Statist.* **48** 3043–3065. MR4152634 https://doi.org/10.1214/19-AOS1921

[16] LEDOIT, O. and WOLF, M. (2020). The power of (non-)linear shrinking: A review and guide to covariance matrix estimation. *J. Financ. Econom.* **1**.

[17] MAÏDA, M. (2007). Large deviations for the largest eigenvalue of rank one deformations of Gaussian ensembles. *Electron. J. Probab.* **12** 1131–1150. MR2336602 https://doi.org/10.1214/EJP.v12-438

[18] MARCHENKO, V. A. and PASTUR, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Math. USSR, Sb.* **1** 457–483.

[19] PAUL, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17** 1617–1642. MR2399865

[20] PERLMAN, M. D. (2018). *STAT* 542: *Multivariate Statistical Analysis*. Univ. Washington.

[21] SHEN, D., SHEN, H., ZHU, H. and MARRON, J. S. (2016). The statistics and mathematics of high dimension low sample size asymptotics. *Statist. Sinica* **26** 1747–1770. MR3586237

[22] STEIN, C. (1956). Some problems in multivariate analysis Technical report, Department of Statistics, Stanford Univ.

[23] STEIN, C. (1986). Lectures on the theory of estimation of many parameters. *J. Math. Sci.* **34** 1373–1403.

[24] WIGNER, E. P. (1955). Characteristic vectors of bordered matrices with infinite dimensions. *Ann. of Math.* (2) **62** 548–564. MR0077805 https://doi.org/10.2307/1970079

[25] WIGNER, E. P. (1958). On the distribution of the roots of certain symmetric matrices. *Ann. of Math.* (2) **67** 325–327. MR0095527 https://doi.org/10.2307/1970008

The Annals of Statistics
2026, Vol. 54, No. 2, 745–770
https://doi.org/10.1214/25-AOS2586

SPECTRUM-AWARE DEBIASING: A MODERN INFERENCE FRAMEWORK WITH APPLICATIONS TO PRINCIPAL COMPONENTS REGRESSION

BY YUFAN LI[a] AND PRAGYA SUR[b]

Department of Statistics, Harvard University, [a]*yufan_li@g.harvard.edu,* [b]*pragya@fas.harvard.edu*

Debiasing is a fundamental concept in high-dimensional statistics. While degrees-of-freedom adjustment is the state-of-the-art debiasing technique in high-dimensional linear regression, it largely remains limited to independent, identically distributed samples and sub-Gaussian covariates. These limitations hinder its wider practical use. In this paper we break this barrier and introduce Spectrum-Aware Debiasing—a novel inference method that applies to challenging high-dimensional regression problems with structured row-column dependencies, heavy tails, asymmetric properties, and latent low-rank structures. Our method achieves debiasing through a rescaled gradient descent step, where the rescaling factor is derived from the spectral properties of the sample covariance matrix. This spectrum-based approach enables accurate debiasing in much broader contexts. We study the common modern regime where the number of features and samples scale proportionally. We establish asymptotic normality of our proposed estimator (suitably centered and scaled) under various convergence notions when the covariates are right-rotationally invariant. We further prove a spectral universality result, extending our guarantees to a much broader class of covariate distributions. Furthermore, we devise a consistent estimator for the asymptotic variance.

Our work has two notable by-products: First, Spectrum-Aware Debiasing rectifies the bias in principal components regression (PCR), providing the first debiased PCR estimator in high dimensions. Second, we introduce a principled test for checking the presence of alignment between the signal and the eigenvectors of the sample covariance matrix. This test is independently valuable for statistical methods developed using approximate message passing, leave-one-out, random matrix theory, or convex Gaussian min-max theorems. We demonstrate the utility of our method through diverse simulated and real data experiments.

REFERENCES

[1] ADLAM, B. and PENNINGTON, J. (2020). Understanding double descent requires a fine-grained bias-variance decomposition. *Adv. Neural Inf. Process. Syst.* **33** 11022–11032.

[2] AGARWAL, A., SHAH, D., SHEN, D. and SONG, D. (2021). On robustness of principal component regression. *J. Amer. Statist. Assoc.* **116** 1731–1745. MR4353710 https://doi.org/10.1080/01621459.2021.1928513

[3] ANASTASIOU, A., BARP, A., BRIOL, F.-X., EBNER, B., GAUNT, R. E., GHADERINEZHAD, F., GORHAM, J., GRETTON, A., LEY, C. et al. (2023). Stein's method meets computational statistics: A review of some recent developments. *Statist. Sci.* **38** 120–139. MR4534646 https://doi.org/10.1214/22-sts863

[4] BAIR, E., HASTIE, T., PAUL, D. and TIBSHIRANI, R. (2006). Prediction by supervised principal components. *J. Amer. Statist. Assoc.* **101** 119–137. MR2252436 https://doi.org/10.1198/016214505000000628

[5] BARBIER, J., KRZAKALA, F., MACRIS, N., MIOLANE, L. and ZDEBOROVÁ, L. (2019). Optimal errors and phase transitions in high-dimensional generalized linear models. *Proc. Natl. Acad. Sci. USA* **116** 5451–5460. MR3939767 https://doi.org/10.1073/pnas.1802705116

[6] BAYATI, M. and MONTANARI, A. (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. Inf. Theory* **57** 764–785. MR2810285 https://doi.org/10.1109/TIT.2010.2094817

MSC2020 subject classifications. Primary62E20; secondary 62F12.

Key words and phrases. High-dimensional inference, debiasing, spectral properties, principal components regression, debiased PCR, right-rotationally invariant designs, vector approximate message passing.

[7] BAYATI, M. and MONTANARI, A. (2012). The LASSO risk for Gaussian matrices. *IEEE Trans. Inf. Theory* **58** 1997–2017. MR2951312 https://doi.org/10.1109/TIT.2011.2174612
[8] BEAN, D., BICKEL, P. J., EL KAROUI, N. and YU, B. (2013). Optimal M-estimation in high-dimensional regression. *Proc. Natl. Acad. Sci. USA* **110** 14563–14568.
[9] BELLEC, P. C. and ZHANG, C.-H. (2022). De-biasing the lasso with degrees-of-freedom adjustment. *Bernoulli* **28** 713–743. MR4389062 https://doi.org/10.3150/21-BEJ1348
[10] BELLEC, P. C. and ZHANG, C.-H. (2023). Debiasing convex regularized estimators and interval estimation in linear models. *Ann. Statist.* **51** 391–436. MR4600987 https://doi.org/10.1214/22-aos2243
[11] 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, Methodol.* **57** 289–300. MR1325392
[12] BICKEL, P. J., LI, B., TSYBAKOV, A. B., VAN DE GEER, S. A., YU, B., VALDÉS, T., RIVERO, C., FAN, J. and VAN DER VAART, A. (2006). Regularization in statistics. *TEST* **15** 271–344.
[13] BING, X., BUNEA, F., STRIMAS-MACKEY, S. and WEGKAMP, M. (2021). Prediction under latent factor regression: Adaptive PCR, interpolating predictors and beyond. *J. Mach. Learn. Res.* **22** Paper No. 177, 50 pp. MR4318533
[14] CADEMARTORI, C. and RUSH, C. (2024). A non-asymptotic analysis of generalized vector approximate message passing algorithms with rotationally invariant designs. *IEEE Trans. Inf. Theory* **70** 5811–5856. MR4773211 https://doi.org/10.1109/tit.2024.3396472
[15] CAI, T. T. and GUO, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Ann. Statist.* **45** 615–646. MR3650395 https://doi.org/10.1214/16-AOS1461
[16] CANDÈS, E. J. and SUR, P. (2020). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *Ann. Statist.* **48** 27–42. MR4065151 https://doi.org/10.1214/18-AOS1789
[17] CELENTANO, M., MONTANARI, A. and WEI, Y. (2023). The Lasso with general Gaussian designs with applications to hypothesis testing. *Ann. Statist.* **51** 2194–2220. MR4678801 https://doi.org/10.1214/23-aos2327
[18] CELENTANO, M. and WAINWRIGHT, M. J. (2023). Challenges of the inconsistency regime: Novel debiasing methods for missing data models.
[19] CHATTERJEE, S. (2010). Spin glasses and Stein's method. *Probab. Theory Related Fields* **148** 567–600. MR2678899 https://doi.org/10.1007/s00440-009-0240-8
[20] CHEN, Y., CHI, Y., FAN, J. and MA, C. (2021). Spectral methods for data science: A statistical perspective. *Found. Trends Mach. Learn.* **14** 566–806.
[21] CHENG, C. and MONTANARI, A. (2024). Dimension free ridge regression. *Ann. Statist.* **52** 2879–2912. MR4842830 https://doi.org/10.1214/24-aos2449
[22] DICKER, L. H. (2016). Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli* **22** 1–37. MR3449775 https://doi.org/10.3150/14-BEJ609
[23] DOBRIBAN, E. and WAGER, S. (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *Ann. Statist.* **46** 247–279. MR3766952 https://doi.org/10.1214/17-AOS1549
[24] DONOHO, D. and MONTANARI, A. (2016). High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *Probab. Theory Related Fields* **166** 935–969. MR3568043 https://doi.org/10.1007/s00440-015-0675-z
[25] DONOHO, D. and TANNER, J. (2009). Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **367** 4273–4293. MR2546388 https://doi.org/10.1098/rsta.2009.0152
[26] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2009). Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci. USA* **106** 18914–18919.
[27] DRUILHET, P. and MOM, A. (2008). Shrinkage structure in biased regression. *J. Multivariate Anal.* **99** 232–244. MR2432327 https://doi.org/10.1016/j.jmva.2006.06.011
[28] DUDEJA, R., BAKHSHIZADEH, M., MA, J. and MALEKI, A. (2020). Analysis of spectral methods for phase retrieval with random orthogonal matrices. *IEEE Trans. Inf. Theory* **66** 5182–5203. MR4130669 https://doi.org/10.1109/TIT.2020.2981910
[29] DUDEJA, R., SEN, S. and LU, Y. M. (2024). Spectral universality in regularized linear regression with nearly deterministic sensing matrices. *IEEE Trans. Inf. Theory* **70** 7923–7951. MR4818369 https://doi.org/10.1109/tit.2024.3458953
[30] EL KAROUI, N. (2018). On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probab. Theory Related Fields* **170** 95–175. MR3748322 https://doi.org/10.1007/s00440-016-0754-9
[31] EL KAROUI, N., BEAN, D., BICKEL, P. J., LIM, C. and YU, B. (2013). On robust regression with high-dimensional predictors. *Proc. Natl. Acad. Sci. USA* **110** 14557–14562.

[32] Fan, J., Liao, Y. and Wang, W. (2016). Projected principal component analysis in factor models. *Ann. Statist.* **44** 219–254. MR3449767 https://doi.org/10.1214/15-AOS1364
[33] Fan, Z. (2022). Approximate message passing algorithms for rotationally invariant matrices. *Ann. Statist.* **50** 197–224. MR4382014 https://doi.org/10.1214/21-aos2101
[34] Farebrother, R. W. (1978). A class of shrinkage estimators. *J. Roy. Statist. Soc. Ser. B, Methodol.* **40** 47–49. MR0512142
[35] Feng, O. Y., Venkataramanan, R., Rush, C. and Samworth, R. J. (2022). A unifying tutorial on approximate message passing. *Found. Trends Mach. Learn.* **15** 335–536.
[36] Fletcher, A. K., Pandit, P., Rangan, S., Sarkar, S. and Schniter, P. (2018). Plug-in estimation in high-dimensional linear inverse problems: A rigorous analysis. *Adv. Neural Inf. Process. Syst.* **31**.
[37] Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35** 109–135.
[38] George, E. I. and Oman, S. D. (1996). Multiple-shrinkage principal component regression. *J. R. Stat. Soc., Ser. D, Stat.* **45** 111–124.
[39] Gerbelot, C., Abbara, A. and Krzakala, F. (2020). Asymptotic errors for high-dimensional convex penalized linear regression beyond Gaussian matrices. In *Proceedings of Thirty Third Conference on Learning Theory* (J. Abernethy and S. Agarwal, eds.). *Proceedings of Machine Learning Research* **125** 1682–1713. PMLR.
[40] Gerbelot, C., Abbara, A. and Krzakala, F. (2023). Asymptotic errors for teacher-student convex generalized linear models (or: How to prove Kabashima's replica formula). *IEEE Trans. Inf. Theory* **69** 1824–1852. MR4564683 https://doi.org/10.1109/tit.2022.3222913
[41] Goldstein, M. and Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLoS ONE* **11** e0152173.
[42] Han, Q. and Shen, Y. (2023). Universality of regularized regression estimators in high dimensions. *Ann. Statist.* **51** 1799–1823. MR4658577 https://doi.org/10.1214/23-aos2309
[43] Hanin, B. and Nica, M. (2020). Products of many large random matrices and gradients in deep neural networks. *Comm. Math. Phys.* **376** 287–322. MR4093863 https://doi.org/10.1007/s00220-019-03624-z
[44] Hastie, T., Montanari, A., Rosset, S. and Tibshirani, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *Ann. Statist.* **50** 949–986. MR4404925 https://doi.org/10.1214/21-aos2133
[45] Howley, T., Madden, M. G., O'Connell, M.-L. and Ryder, A. G. (2006). The effect of principal component analysis on machine learning accuracy with high dimensional spectral data. In *Applications and Innovations in Intelligent Systems XIII*: *Proceedings of AI*-2005, *the Twenty-Fifth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*, *Cambridge*, *UK*, *December* 2005 209–222. Springer, Berlin.
[46] Huang, G. B., Ramesh, M., Berg, T. and Learned-Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report No. 07-49, Univ. Massachusetts, Amherst.
[47] Hubert, M. and Verboven, S. (2003). A robust PCR method for high-dimensional regressors. *J. Chemom.* **17** 438–452.
[48] 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 https://doi.org/10.1093/imaiai/iat004
[49] Javanmard, A. and Montanari, A. (2014). Hypothesis testing in high-dimensional regression under the Gaussian random design model: Asymptotic theory. *IEEE Trans. Inf. Theory* **60** 6522–6554. MR3265038 https://doi.org/10.1109/TIT.2014.2343629
[50] Javanmard, A. and Montanari, A. (2018). Debiasing the Lasso: Optimal sample size for Gaussian designs. *Ann. Statist.* **46** 2593–2622. MR3851749 https://doi.org/10.1214/17-AOS1630
[51] Jiang, K., Mukherjee, R., Sen, S. and Sur, P. (2025). A new central limit theorem for the augmented IPW estimator: Variance inflation, cross-fit covariance and beyond. *Ann. Statist.* **53** 647–675. MR4900161 https://doi.org/10.1214/24-aos2476
[52] Jolliffe, I. T. (1982). A note on the use of principal components in regression. *J. R. Stat. Soc., Ser. C, Appl. Stat.* **31** 300–303.
[53] Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philos. Trans. R. Soc. Lond. A* **374** 20150202, 16 pp. MR3479904 https://doi.org/10.1098/rsta.2015.0202
[54] Li, G. and Wei, Y. (2022). A non-asymptotic framework for approximate message passing in spiked models. Preprint. Available at arXiv:2208.03313.
[55] Li, G. and Wei, Y. (2024). A non-asymptotic distributional theory of approximate message passing for sparse and robust regression. Preprint. Available at arXiv:2401.03923.

[56] LI, Y., FAN, Z., SEN, S. and WU, Y. (2024). Random linear estimation with rotationally-invariant designs: Asymptotics at high temperature. *IEEE Trans. Inf. Theory* **70** 2118–2153. MR4709776 https://doi.org/10.1109/tit.2023.3321575

[57] LI, Y., SEN, S. and ADLAM, B. (2024). Understanding optimal feature transfer via a fine-grained bias-variance analysis. Preprint. Available at arXiv:2404.12481.

[58] LI, Y. and SUR, P. (2025). Optimal and provable calibration in high-dimensional binary classification: Angular calibration and platt scaling. Preprint. Available at arXiv:2502.15131.

[59] LI, Y. and SUR, P. (2026). Supplement to "Spectrum-aware debiasing: A modern inference framework with applications to principal components regression." https://doi.org/10.1214/25-AOS2586SUPPA, https://doi.org/10.1214/25-AOS2586SUPPB

[60] LIANG, T., SEN, S. and SUR, P. (2023). High-dimensional asymptotics of Langevin dynamics in spiked matrix models. *Inf. Inference* **12** Paper No. iaad042, 33 pp. MR4655761 https://doi.org/10.1093/imaiai/iaad042

[61] LIANG, T. and SUR, P. (2022). A precise high-dimensional asymptotic theory for boosting and minimum-ℓ1-norm interpolated classifiers. *Ann. Statist.* **50** 1669–1695. MR4441136 https://doi.org/10.1214/22-aos2170

[62] LIU, L., HUANG, S. and KURKOSKI, B. M. (2022). Memory AMP. *IEEE Trans. Inf. Theory* **68** 8015–8039. MR4544929 https://doi.org/10.1109/tit.2022.3186166

[63] S&P DOW JONES INDICES LLC (2024). S&P 500 [SP500]. Retrieved from FRED, Federal Reserve Bank of St. Louis. May 3, 2024.

[64] LUO, K., LI, Y. and SUR, P. (2025). ROTI-GCV: Generalized cross-validation for right-ROTationally invariant data. In *Proceedings of the* 28*th International Conference on Artificial Intelligence and Statistics* (Y. Li, S. Mandt, S. Agrawal and E. Khan, eds.). *Proceedings of Machine Learning Research* **258** 1603–1611. PMLR.

[65] MA, J. and PING, L. (2017). Orthogonal amp. *IEEE Access* **5** 2020–2033.

[66] MÉZARD, M., PARISI, G. and VIRASORO, M. A. (1987). *Spin Glass Theory and Beyond*: *An Introduction to the Replica Method and Its Applications*. *World Scientific Lecture Notes in Physics* **9**. World Scientific Co., Inc., Teaneck, NJ. MR1026102

[67] NOORDEWIER, M., TOWELL, G. and SHAVLIK, J. (1990). Training knowledge-based neural networks to recognize genes in DNA sequences. *Adv. Neural Inf. Process. Syst.* **3**.

[68] RANGAN, S., SCHNITER, P. and FLETCHER, A. K. (2019). Vector approximate message passing. *IEEE Trans. Inf. Theory* **65** 6664–6684. MR4009222 https://doi.org/10.1109/TIT.2019.2916359

[69] REDMOND, M. (2009). Communities and crime. UCI Machine Learning Repository. https://doi.org/10.24432/C53W3X

[70] SCHNITER, P., RANGAN, S. and FLETCHER, A. K. (2016). Vector approximate message passing for the generalized linear model. In 2016 50*th Asilomar Conference on Signals, Systems and Computers* 1525–1529. IEEE.

[71] SILIN, I. and FAN, J. (2022). Canonical thresholding for nonsparse high-dimensional linear regression. *Ann. Statist.* **50** 460–486. MR4382024 https://doi.org/10.1214/21-aos2116

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

[73] STOJNIC, M. (2013). A framework to characterize performance of LASSO algorithms.

[74] SUR, P. and CANDÈS, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proc. Natl. Acad. Sci. USA* **116** 14516–14525. MR3984492 https://doi.org/10.1073/pnas.1810420116

[75] SUR, P., CHEN, Y. and CANDÈS, E. J. (2019). The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probab. Theory Related Fields* **175** 487–558. MR4009715 https://doi.org/10.1007/s00440-018-00896-9

[76] TAKEDA, K., UDA, S. and KABASHIMA, Y. (2006). Analysis of CDMA systems that are characterized by eigenvalue spectrum. *Europhys. Lett.* **76** 1193.

[77] TAKEUCHI, K. (2020). Rigorous dynamics of expectation-propagation-based signal recovery from unitarily invariant measurements. *IEEE Trans. Inf. Theory* **66** 368–386. MR4053400 https://doi.org/10.1109/TIT.2019.2947058

[78] TAKEUCHI, K. (2020). Convolutional approximate message-passing. *IEEE Signal Process. Lett.* **27** 416–420.

[79] TAKEUCHI, K. (2021). Bayes-optimal convolutional AMP. *IEEE Trans. Inf. Theory* **67** 4405–4428. MR4306276 https://doi.org/10.1109/TIT.2021.3077471

[80] TALAGRAND, M. (2003). *Spin Glasses*: *A Challenge for Mathematicians*: *Cavity and Mean Field Models*. *Ergebnisse der Mathematik und Ihrer Grenzgebiete*. 3. *Folge*. *A Series of Modern Surveys in Mathemat-*

ics [*Results in Mathematics and Related Areas*. 3*rd Series*. *A Series of Modern Surveys in Mathematics*] **46**. Springer, Berlin. MR1993891

[81] THRAMPOULIDIS, C., OYMAK, S. and HASSIBI, B. (2015). The Gaussian min-max theorem in the presence of convexity.

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

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

[84] VENKATARAMANAN, R., KÖGLER, K. and MONDELLI, M. (2022). Estimation in rotationally invariant generalized linear models via approximate message passing. In *International Conference on Machine Learning* 22120–22144. PMLR.

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

[86] XU, Y., LIU, Y., LIANG, S., WU, T., BAI, B., BARBIER, J. and HOU, T. (2023). Capacity-achieving sparse regression codes via vector approximate message passing. In 2023 *IEEE International Symposium on Information Theory* (*ISIT*) 785–790. IEEE.

[87] ZADOROZHNYI, O., BENECKE, G., MANDT, S., SCHEFFER, T. and KLOFT, M. (2016). Huber-norm regularization for linear prediction models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 714–730. Springer, Berlin.

[88] ZDEBOROVÁ, L. and KRZAKALA, F. (2016). Statistical physics of inference: Thresholds and algorithms. *Adv. Phys*. **65** 453–552.

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

[90] ZHAO, Q., SUR, P. and CANDÈS, E. J. (2022). The asymptotic distribution of the MLE in high-dimensional logistic models: Arbitrary covariance. *Bernoulli* **28** 1835–1861. MR4411513 https://doi.org/10.3150/21-bej1401

[91] ZHOU, L., KOEHLER, F., SUR, P., SUTHERLAND, D. J. and SREBRO, N. (2022). A non-asymptotic Moreau envelope theory for high-dimensional generalized linear models. *Adv. Neural Inf. Process. Syst*. **35** 21286–21299.

[92] ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2007). On the "degrees of freedom" of the lasso. *Ann*. *Statist*. **35** 2173–2192. MR2363967 https://doi.org/10.1214/009053607000000127

The Annals of Statistics
2026, Vol. 54, No. 2, 771–795
https://doi.org/10.1214/25-AOS2587

SCALABLE INFERENCE FOR NONPARAMETRIC STOCHASTIC APPROXIMATION IN REPRODUCING KERNEL HILBERT SPACES

BY MEIMEI LIU[1,a], ZUOFENG SHANG[2,b] AND YUN YANG[3,c]

[1]*Department of Statistics, Virginia Tech,* [a]*meimeiliu@vt.edu*
[2]*Department of Mathematical Science, New Jersey Institute of Technology,* [b]*zuofeng.shang@njit.edu*
[3]*Department of Mathematics, University of Maryland, College Park,* [c]*yy84@umd.edu*

Stochastic approximation (SA) is a powerful and scalable computational method for iteratively estimating the solution of optimization problems in the presence of randomness, particularly well suited for large-scale and streaming data settings. In this work we propose a theoretical framework for stochastic approximation (SA) applied to nonparametric least squares in reproducing kernel Hilbert spaces (RKHS), enabling online statistical inference in nonparametric regression models. We achieve this by constructing asymptotically valid pointwise (and simultaneous) confidence intervals (bands) for local (and global) inference of the nonlinear regression function, via employing an online multiplier bootstrap approach to a functional stochastic gradient descent (SGD) algorithm in the RKHS. Our main theoretical contributions consist of a unified framework for characterizing the nonasymptotic behavior of the functional SGD estimator and demonstrating the consistency of the multiplier bootstrap method. The proof techniques involve the development of a higher-order expansion of the functional SGD estimator under the supremum norm metric and the Gaussian approximation of suprema of weighted and non-identically distributed empirical processes. Our theory specifically reveals an interesting relationship between the tuning of step sizes in SGD for estimation and the accuracy of uncertainty quantification.

REFERENCES

[1] ALAOUI, A. and MAHONEY, M. W. (2015). Fast randomized kernel ridge regression with statistical guarantees. *Adv. Neural Inf. Process. Syst.* **28**.
[2] ARMSTRONG, T. B. and KOLESÁR, M. (2020). Simple and honest confidence intervals in nonparametric regression. *Quant. Econ.* **11** 1–39. MR4084655 https://doi.org/10.3982/qe1199
[3] ARONSZAJN, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68** 337–404. MR0051437 https://doi.org/10.2307/1990404
[4] AVRON, H., CLARKSON, K. L. and WOODRUFF, D. P. (2017). Faster kernel ridge regression using sketching and preconditioning. *SIAM J. Matrix Anal. Appl.* **38** 1116–1138. MR3713904 https://doi.org/10.1137/16M1105396
[5] AVRON, H., KAPRALOV, M., MUSCO, C., MUSCO, C., VELINGKER, A. and ZANDIEH, A. (2017). Random Fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In *International Conference on Machine Learning* 253–262. PMLR.
[6] BACH, F. (2013). Sharp analysis of low-rank kernel matrix approximations. In *Conference on Learning Theory* 185–209. PMLR.
[7] BACH, F. (2017). On the equivalence between kernel quadrature rules and random feature expansions. *J. Mach. Learn. Res.* **18** Paper No. 21. MR3634888
[8] BERLINET, A. and THOMAS-AGNAN, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic, Boston, MA. MR2239907 https://doi.org/10.1007/978-1-4419-9096-9
[9] BOTTOU, L. (1998). Online learning and stochastic approximations. *Online Learn. Neural Netw.* **17** 142.
[10] BOTTOU, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'*2010 177–186. Physica-Verlag/Springer, Heidelberg. MR3362066

MSC2020 subject classifications. Primary 62G20, 68W27; secondary 62L20, 62F40.
Key words and phrases. Bootstrap, functional stochastic gradient descent, nonparametric regression, online inference, RKHS.

[11] BOTTOU, L. and BOUSQUET, O. (2007). The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems* **20**.
[12] BOTTOU, L. and BOUSQUET, O. (2008). The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems* 20 (J. C. Platt, D. Koller, Y. Singer and S. Roweis, eds.) 161–168. Curran Associates, Red Hook.
[13] CAO, Y. and GU, Q. (2019). Generalization bounds of stochastic gradient descent for wide and deep neural networks. *Adv. Neural Inf. Process. Syst.* **32**.
[14] CHEN, X., LEE, J. D., TONG, X. T. and ZHANG, Y. (2020). Statistical inference for model parameters in stochastic gradient descent. *Ann. Statist.* **48** 251–273. MR4065161 https://doi.org/10.1214/18-AOS1801
[15] CHEN, X., LIU, Q. and TONG, X. T. (2022). Dimension independent excess risk by stochastic gradient descent. *Electron. J. Stat.* **16** 4547–4603. MR4489235 https://doi.org/10.1214/22-ejs2055
[16] CHERIDITO, P., JENTZEN, A. and ROSSMANNEK, F. (2021). Non-convergence of stochastic gradient descent in the training of deep neural networks. *J. Complexity* **64** Paper No. 101540. MR4232648 https://doi.org/10.1016/j.jco.2020.101540
[17] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2014). Gaussian approximation of suprema of empirical processes. *Ann. Statist.* **42** 1564–1597. MR3262461 https://doi.org/10.1214/14-AOS1230
[18] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2014). Anti-concentration and honest, adaptive confidence bands. *Ann. Statist.* **42** 1787–1818. MR3262468 https://doi.org/10.1214/14-AOS1235
[19] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2016). Empirical and multiplier bootstraps for suprema of empirical processes of increasing complexity, and related Gaussian couplings. *Stoch. Process. Appl.* **126** 3632–3651. MR3565470 https://doi.org/10.1016/j.spa.2016.04.009
[20] DAI, B., XIE, B., HE, N., LIANG, Y., RAJ, A., BALCAN, M.-F. F. and SONG, L. (2014). Scalable kernel methods via doubly stochastic gradients. *Adv. Neural Inf. Process. Syst.* **27**.
[21] DICICCIO, T. J. and ROMANO, J. P. (1988). A review of bootstrap confidence intervals. *J. Roy. Statist. Soc. Ser. B, Methodol.* **50** 338–354. MR0970972
[22] DIEULEVEUT, A. and BACH, F. (2016). Nonparametric stochastic approximation with large step-sizes. *Ann. Statist.* **44** 1363–1399. MR3519927 https://doi.org/10.1214/15-AOS1391
[23] EFRON, B. and TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap*. *Monographs on Statistics and Applied Probability* **57**. CRC Press, New York. MR1270903 https://doi.org/10.1007/978-1-4899-4541-9
[24] FAN, J. and ZHANG, W. (2000). Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scand. J. Stat.* **27** 715–731. MR1804172 https://doi.org/10.1111/1467-9469.00218
[25] FANG, Y., XU, J. and YANG, L. (2018). Online bootstrap confidence intervals for the stochastic gradient descent estimator. *J. Mach. Learn. Res.* **19** Paper No. 78. MR3899780
[26] GU, C. (2013). *Smoothing Spline ANOVA Models*, 2nd ed. *Springer Series in Statistics* **297**. Springer, New York. MR3025869 https://doi.org/10.1007/978-1-4614-5369-7
[27] HÄRDLE, W. and MARRON, J. S. (1991). Bootstrap simultaneous error bars for nonparametric regression. *Ann. Statist.* **19** 778–796. MR1105844 https://doi.org/10.1214/aos/1176348120
[28] KIM, S., PASUPATHY, R. and HENDERSON, S. G. (2015). A guide to sample average approximation. In *Handbook of Simulation Optimization* 207–243. Springer, Berlin.
[29] KLEYWEGT, A. J., SHAPIRO, A. and HOMEM-DE-MELLO, T. (2001/02). The sample average approximation method for stochastic discrete optimization. *SIAM J. Optim.* **12** 479–502. MR1885572 https://doi.org/10.1137/S1052623499363220
[30] KOLTCHINSKII, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.* **34** 2593–2656. MR2329442 https://doi.org/10.1214/009053606000001019
[31] LE, Q. V., NGIAM, J., COATES, A., LAHIRI, A., PROCHNOW, B. and NG, A. Y. (2011). On optimization methods for deep learning. In *Proceedings of the* 28*th International Conference on Machine Learning* 265–272.
[32] LIU, M., SHANG, Z. and CHENG, G. (2020). Nonparametric distributed learning under general designs. *Electron. J. Stat.* **14** 3070–3102. MR4137596 https://doi.org/10.1214/20-EJS1733
[33] LIU, M., SHANG, Z. and YANG, Y. (2026). Supplement to "Scalable inference for nonparametric stochastic approximation in reproducing kernel Hilbert spaces." https://doi.org/10.1214/25-AOS2587SUPP
[34] MENDELSON, S. (2002). Geometric parameters of kernel machines. In *Computational Learning Theory* (*Sydney*, 2002). *Lecture Notes in Computer Science* **2375** 29–43. Springer, Berlin. MR2040403 https://doi.org/10.1007/3-540-45435-7_3
[35] MENDELSON, S. and NEEMAN, J. (2010). Regularization in kernel learning. *Ann. Statist.* **38** 526–565. MR2590050 https://doi.org/10.1214/09-AOS728

[36] NEMIROVSKI, A., JUDITSKY, A., LAN, G. and SHAPIRO, A. (2008). Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.* **19** 1574–1609. MR2486041 https://doi.org/10.1137/070704277

[37] NESTEROV, Y. and VIAL, J.-P. (2008). Confidence level solutions for stochastic programming. *Automatica J. IFAC* **44** 1559–1568. MR2531843 https://doi.org/10.1016/j.automatica.2008.01.017

[38] NEUMANN, M. H. and POLZEHL, J. (1998). Simultaneous bootstrap confidence bands in nonparametric regression. *J. Nonparametr. Stat.* **9** 307–333. MR1646905 https://doi.org/10.1080/10485259808832748

[39] POLYAK, B. T. and JUDITSKY, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.* **30** 838–855. MR1167814 https://doi.org/10.1137/0330046

[40] RAHIMI, A. and RECHT, B. (2007). Random features for large-scale kernel machines. *Adv. Neural Inf. Process. Syst.* **20**.

[41] RAKHLIN, A., SHAMIR, O. and SRIDHARAN, K. (2012). Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the* 29*th International Conference on Machine Learning* (*ICML*) 449–456.

[42] ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Stat.* **22** 400–407. MR0042668 https://doi.org/10.1214/aoms/1177729586

[43] RUPPERT, D. (1988). Efficient estimations from a slowly convergent Robbins-Monro process. Technical Report, Cornell Univ. Operations Research and Industrial Engineering Ithaca, NY.

[44] SAUNDERS, C., GAMMERMAN, A. and VOVK, V. (1998). Ridge regression learning algorithm in dual variables. In *International Conference on Machine Learning* 515–521. Springer, Berlin.

[45] STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053. MR0673642

[46] SU, W. J. and ZHU, Y. (2023). HiGrad: Uncertainty quantification for online learning and stochastic approximation. *J. Mach. Learn. Res.* **24** Paper No. [124]. MR4596071

[47] SUN, J. and LOADER, C. R. (1994). Simultaneous confidence bands for linear regression and smoothing. *Ann. Statist.* **22** 1328–1345. MR1311978 https://doi.org/10.1214/aos/1176325631

[48] WAHBA, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B, Methodol.* **45** 133–150. MR0701084

[49] WAHBA, G. (1990). *Spline Models for Observational Data. CBMS-NSF Regional Conference Series in Applied Mathematics* **59**. SIAM, Philadelphia, PA. MR1045442 https://doi.org/10.1137/1.9781611970128

[50] WANG, Y. and WAHBA, G. (1995). Bootstrap confidence intervals for smoothing splines and their comparison to Bayesian confidence intervals. *J. Stat. Comput. Simul.* **51** 263–279.

[51] YANG, Y. and BARRON, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.* **27** 1564–1599. MR1742500 https://doi.org/10.1214/aos/1017939142

[52] YANG, Y., PILANCI, M. and WAINWRIGHT, M. J. (2017). Randomized sketches for kernels: Fast and optimal nonparametric regression. *Ann. Statist.* **45** 991–1023. MR3662446 https://doi.org/10.1214/16-AOS1472

[53] ZHANG, T. and SIMON, N. (2022). A sieve stochastic gradient descent estimator for online nonparametric regression in Sobolev ellipsoids. *Ann. Statist.* **50** 2848–2871. MR4500627 https://doi.org/10.1214/22-aos2212

[54] 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

[55] ZHU, W., CHEN, X. and WU, W. B. (2023). Online covariance matrix estimation in stochastic gradient descent. *J. Amer. Statist. Assoc.* **118** 393–404. MR4571129 https://doi.org/10.1080/01621459.2021.1933498

The Annals of Statistics
2026, Vol. 54, No. 2, 796–833
https://doi.org/10.1214/25-AOS2590

PRECISE ASYMPTOTICS OF BAGGING REGULARIZED M-ESTIMATORS

BY TAKUYA KORIYAMA[1,a], PRATIK PATIL[2,b], JIN-HONG DU[3,c], KAI TAN[4,d] AND PIERRE C. BELLEC[5,e]

[1]*Econometrics and Statistics, Booth School of Business, The University of Chicago,* [a]*tkoriyam@uchicago.edu*
[2]*Department of Statistics and Data Sciences, University of Texas at Austin,* [b]*pratikpatil@utexas.edu*
[3]*Institute of Data Science, University of Hong Kong,* [c]*jinhongd@hku.hk*
[4]*Department of Statistics, Stanford University,* [d]*kaitan9@stanford.edu*
[5]*Department of Statistics, Rutgers University,* [e]*pierre.bellec@rutgers.edu*

We characterize the squared prediction risk of ensemble estimators obtained through subagging (subsample bootstrap aggregating) regularized M-estimators and construct a consistent estimator for the risk. Specifically, we consider a heterogeneous collection of $M \geq 1$ regularized M-estimators, each trained with (possibly different) subsample sizes, convex differentiable losses, and convex regularizers. We operate under the proportional asymptotics regime, where the sample size n, feature size p, and subsample sizes k_m for $m \in [M]$ all diverge with fixed limiting ratios n/p and k_m/n. Key to our analysis is a new result on the joint asymptotic behavior of correlations between the estimator and residual errors on overlapping subsamples, governed through a (provably) contractive nonlinear system of equations. Of independent interest we also establish convergence of trace functionals related to degrees of freedom in the nonensemble setting (with $M = 1$) along the way, extending previously known cases for squared loss with ridge and lasso regularizers. When specialized to homogeneous ensembles trained with a common loss, regularizer, and subsample size, the risk characterization sheds some light on the (implicitly) induced regularization effect due to the ensemble and subsample sizes (M, k). For any ensemble size M, optimally tuning subsample size yields samplewise monotonic risk. For the full-ensemble estimator (when $M \to \infty$), the optimal subsample size $k^\star$ tends to be in the overparameterized regime ($k^\star \leq \min\{n, p\}$), when explicit regularization is vanishing. Finally, joint optimization of subsample size, ensemble size, and regularization can significantly outperform regularizer optimization alone on the full data (without any subagging).

REFERENCES

[1] ADLAM, B. and PENNINGTON, J. (2020). Understanding double descent requires a fine-grained bias-variance decomposition. In *Advances in Neural Information Processing Systems*.

[2] ALLEN-ZHU, Z., LI, Y. and SONG, Z. (2019). A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning* 242–252.

[3] ANDO, R. and KOMAKI, F. (2023). On high-dimensional asymptotic properties of model averaging estimators. arXiv preprint. Available at arXiv:2308.09476.

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

[5] BAYATI, M. and MONTANARI, A. (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. Inf. Theory* **57** 764–785. MR2810285 https://doi.org/10.1109/TIT.2010.2094817

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

[7] BEAN, D., BICKEL, P. J., EL KAROUI, N. and YU, B. (2013). Optimal M-estimation in high-dimensional regression. *Proc. Natl. Acad. Sci. USA* **110** 14563–14568.

MSC2020 subject classifications. 62F40, 62F12.
Key words and phrases. Bagging, proportional asymptotics, overparameterization.

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

[9] BELKIN, M., MA, S. and MANDAL, S. (2018). To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*.

[10] BELLEC, P. C. (2018). Optimal bounds for aggregation of affine estimators. *Ann. Statist.* **46** 30–59. MR3766945 https://doi.org/10.1214/17-AOS1540

[11] BELLEC, P. C. (2023). Out-of-sample error estimation for M-estimators with convex penalty. *Inf. Inference* **12** 2782–2817.

[12] BELLEC, P. C. (2025). Observable adjustments in single-index models for regularized M-estimators with bounded p/n. *Ann. Statist.* **53** 531–560. MR4900157 https://doi.org/10.1214/24-aos2464

[13] BELLEC, P. C., DU, J.-H., KORIYAMA, T., PATIL, P. and TAN, K. (2025). Corrected generalized cross-validation for finite ensembles of penalized estimators. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **87** 289–318. MR4896648 https://doi.org/10.1093/jrsssb/qkae092

[14] BELLEC, P. C. and KORIYAMA, T. (2023). Existence of solutions to the nonlinear equations characterizing the precise error of M-estimators. arXiv preprint. Available at arXiv:2312.13254.

[15] BELLEC, P. C. and KORIYAMA, T. (2024). Asymptotics of resampling without replacement in robust and logistic regression. arXiv preprint. Available at arXiv:2404.02070.

[16] BELLEC, P. C. and KORIYAMA, T. (2025). Error estimation and adaptive tuning for unregularized robust M-estimator. *J. Mach. Learn. Res.* **26** 1–40.

[17] BELLEC, P. C. and SHEN, Y. (2022). Derivatives and residual distribution of regularized M-estimators with application to adaptive tuning. In *Conference on Learning Theory*.

[18] BERTSEKAS, D. P. (2016). *Nonlinear Programming*, 3rd ed. *Athena Scientific Optimization and Computation Series*. Athena Scientific, Belmont, MA. MR3587371

[19] BREIMAN, L. (1996). Bagging predictors. *Mach. Learn.* **24** 123–140.

[20] BREIMAN, L. (2001). Using iterated bagging to debias regressions. *Mach. Learn.* **45** 261–277.

[21] BÜHLMANN, P. and YU, B. (2002). Analyzing bagging. *Ann. Statist.* **30** 927–961. MR1926165 https://doi.org/10.1214/aos/1031689014

[22] BUJA, A. and STUETZLE, W. (2006). Observations on bagging. *Statist. Sinica* **16** 323–351. MR2267238

[23] CELENTANO, M. and MONTANARI, A. (2024). Correlation adjusted debiased Lasso: Debiasing the Lasso with inaccurate covariate model. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **86** 1455–1482. MR4896643 https://doi.org/10.1093/jrsssb/qkae039

[24] CELENTANO, M., MONTANARI, A. and WEI, Y. (2023). The Lasso with general Gaussian designs with applications to hypothesis testing. *Ann. Statist.* **51** 2194–2220. MR4678801 https://doi.org/10.1214/23-aos2327

[25] CHEN, X., ZENG, Y., YANG, S. and SUN, Q. (2023). Sketched ridgeless linear regression: The role of downsampling. In *International Conference on Machine Learning*.

[26] CHIZAT, L. and BACH, F. (2018). On the global convergence of gradient descent for overparameterized models using optimal transport. *Adv. Neural Inf. Process. Syst.*

[27] CLARTÉ, L., VANDENBROUCQUE, A., DALLE, G., LOUREIRO, B., KRZAKALA, F. and ZDEBOROVÁ, L. (2024). Analysis of bootstrap and subsampling in high-dimensional regularized regression. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence* **244** 787–819. PMLR.

[28] D'ASCOLI, S., REFINETTI, M., BIROLI, G. and KRZAKALA, F. (2020). Double trouble in double descent: Bias and variance (s) in the lazy regime. In *International Conference on Machine Learning*.

[29] DAI, D., RIGOLLET, P., XIA, L. and ZHANG, T. (2014). Aggregation of affine estimators. *Electron. J. Stat.* **8** 302–327. MR3192554 https://doi.org/10.1214/14-EJS886

[30] DAI, D., RIGOLLET, P. and ZHANG, T. (2012). Deviation optimal learning using greedy Q-aggregation. *Ann. Statist.* **40** 1878–1905. MR3015047 https://doi.org/10.1214/12-AOS1025

[31] DALALYAN, A. and TSYBAKOV, A. B. (2008). Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Mach. Learn.* **72** 39–61.

[32] DALALYAN, A. S., GRAPPIN, E. and PARIS, Q. (2018). On the exponentially weighted aggregate with the Laplace prior. *Ann. Statist.* **46** 2452–2478. MR3845023 https://doi.org/10.1214/17-AOS1626

[33] DALALYAN, A. S. and SALMON, J. (2012). Sharp oracle inequalities for aggregation of affine estimators. *Ann. Statist.* **40** 2327–2355. MR3059085 https://doi.org/10.1214/12-AOS1038

[34] DENG, Z., KAMMOUN, A. and THRAMPOULIDIS, C. (2022). A model of double descent for high-dimensional binary linear classification. *Inf. Inference* **11** 435–495. MR4474343 https://doi.org/10.1093/imaiai/iaab002

[35] DOBRIBAN, E. and SHENG, Y. (2020). WONDER: Weighted one-shot distributed ridge regression in high dimensions. *J. Mach. Learn. Res.* **21** 1–52.

[36] DOBRIBAN, E. and SHENG, Y. (2021). Distributed linear regression by averaging. *Ann. Statist.* **49** 918–943. MR4255113 https://doi.org/10.1214/20-aos1984
[37] DOBRIBAN, E. and WAGER, S. (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *Ann. Statist.* **46** 247–279. MR3766952 https://doi.org/10.1214/17-AOS1549
[38] DONOHO, D. and MONTANARI, A. (2016). High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *Probab. Theory Related Fields* **166** 935–969. MR3568043 https://doi.org/10.1007/s00440-015-0675-z
[39] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2009). Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci. USA* **106** 18914–18919.
[40] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2011). The noise-sensitivity phase transition in compressed sensing. *IEEE Trans. Inf. Theory* **57** 6920–6941. MR2882271 https://doi.org/10.1109/TIT.2011.2165823
[41] DU, J.-H. and PATIL, P. (2024). Implicit regularization paths of weighted neural representations. *Adv. Neural Inf. Process. Syst.* **37** 30261–30299.
[42] DU, J.-H. and PATIL, P. (2024). Python package sklearn_ensemble_cv v0.2.3. PyPI.
[43] DU, J.-H., PATIL, P. and KUCHIBHOTLA, A. K. (2023). Subsample ridge ensembles: Equivalences and generalized cross-validation. In *International Conference on Machine Learning*.
[44] DU, J.-H., PATIL, P., ROEDER, K. and KUCHIBHOTLA, A. K. (2024). Extrapolated cross-validation for randomized ensembles. *J. Comput. Graph. Statist.* **33** 1061–1072. MR4785806 https://doi.org/10.1080/10618600.2023.2288194
[45] DU, S., ZHAI, X., POCZOS, B. and SINGH, A. (2018). Gradient descent provably optimizes overparameterized neural networks. arXiv preprint. Available at arXiv:1810.02054.
[46] EL KAROUI, N. (2013). Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: Rigorous results. arXiv preprint. Available at arXiv:1311.2445.
[47] EL KAROUI, N. (2018). On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probab. Theory Related Fields* **170** 95–175. MR3748322 https://doi.org/10.1007/s00440-016-0754-9
[48] EL KAROUI, N., BEAN, D., BICKEL, P. J., LIM, C. and YU, B. (2013). On robust regression with high-dimensional predictors. *Proc. Natl. Acad. Sci. USA* **110** 14557–14562.
[49] FRANK, I. E. and FRIEDMAN, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35** 109–135.
[50] FRIEDMAN, J. H. and HALL, P. (2007). On bagging and nonlinear estimation. *J. Statist. Plann. Inference* **137** 669–683. MR2301708 https://doi.org/10.1016/j.jspi.2006.06.002
[51] FU, W. J. (1998). Penalized regressions: The bridge versus the lasso. *J. Comput. Graph. Statist.* **7** 397–416. MR1646710 https://doi.org/10.2307/1390712
[52] GERBELOT, C. and BERTHIER, R. (2023). Graph-based approximate message passing iterations. *Inf. Inference* **12** 2562–2628.
[53] GUNASEKAR, S., LEE, J., SOUDRY, D. and SREBRO, N. (2018). Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*.
[54] GUNASEKAR, S., LEE, J., SOUDRY, D. and SREBRO, N. (2018). Implicit bias of gradient descent on linear convolutional networks. *Adv. Neural Inf. Process. Syst.*
[55] HALL, P. and SAMWORTH, R. J. (2005). Properties of bagged nearest neighbour classifiers. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 363–379. MR2155343 https://doi.org/10.1111/j.1467-9868.2005.00506.x
[56] HAN, Q. and SHEN, Y. (2023). Universality of regularized regression estimators in high dimensions. *Ann. Statist.* **51** 1799–1823. MR4658577 https://doi.org/10.1214/23-aos2309
[57] HAN, Q. and XU, X. (2023). The distribution of Ridgeless least squares interpolators. arXiv preprint. Available at arXiv:2307.02044.
[58] HANSEN, L. and SALAMON, P. (1990). Neural network ensembles. *IEEE Trans. Pattern Anal. Mach. Intell.* **12** 993–1001.
[59] HASTIE, T., MONTANARI, A., ROSSET, S. and TIBSHIRANI, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *Ann. Statist.* **50** 949–986. MR4404925 https://doi.org/10.1214/21-aos2133
[60] HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2722294 https://doi.org/10.1007/978-0-387-84858-7
[61] 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 https://doi.org/10.1093/imaiai/iat004
[62] KORIYAMA, T., PATIL, P., DU, J.-H, TAN, K. and BELLEC, P. C (2026). Supplement to "Precise asymptotics of bagging regularized M-estimators." https://doi.org/10.1214/25-AOS2590SUPP

[63] KROGH, A. and SOLLICH, P. (1997). Statistical mechanics of ensemble learning. *Phys. Rev. E* **55** 811.
[64] LECUÉ, G. and RIGOLLET, P. (2014). Optimal learning with Q-aggregation. *Ann. Statist.* **42** 211–224. MR3178462 https://doi.org/10.1214/13-AOS1190
[65] LEE, J., XIAO, L., SCHOENHOLZ, S., BAHRI, Y., NOVAK, R., SOHL-DICKSTEIN, J. and PENNINGTON, J. (2019). Wide neural networks of any depth evolve as linear models under gradient descent. *Adv. Neural Inf. Process. Syst.*
[66] LEJEUNE, D., JAVADI, H. and BARANIUK, R. (2020). The implicit regularization of ordinary least squares ensembles. In *International Conference on Artificial Intelligence and Statistics*.
[67] LEUNG, G. and BARRON, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Trans. Inf. Theory* **52** 3396–3410. MR2242356 https://doi.org/10.1109/TIT.2006.878172
[68] LI, Y. and WEI, Y. (2021). Minimum ℓ_1-norm interpolators: Precise asymptotics and multiple descent. arXiv preprint. Available at arXiv:2110.09502.
[69] LIANG, T. and SUR, P. (2022). A precise high-dimensional asymptotic theory for boosting and minimum-ℓ1-norm interpolated classifiers. *Ann. Statist.* **50** 1669–1695. MR4441136 https://doi.org/10.1214/22-aos2170
[70] LOUREIRO, B., GERBELOT, C., CUI, H., GOLDT, S., KRZAKALA, F., MEZARD, M. and ZDEBOROVÁ, L. (2021). Learning curves of generic features maps for realistic datasets with a teacher-student model. *Adv. Neural Inf. Process. Syst.* **34** 18137–18151.
[71] LOUREIRO, B., GERBELOT, C., REFINETTI, M., SICURO, G. and KRZAKALA, F. (2022). Fluctuations, bias, variance and ensemble of learners: Exact asymptotics for convex losses in high-dimension. In *International Conference on Machine Learning*.
[72] MAI, X., LIAO, Z. and COUILLET, R. (2019). A large scale analysis of logistic regression: Asymptotic performance and new insights. In *International Conference on Acoustics, Speech and Signal Processing*.
[73] MIOLANE, L. and MONTANARI, A. (2021). The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning. *Ann. Statist.* **49** 2313–2335. MR4319252 https://doi.org/10.1214/20-aos2038
[74] MONTANARI, A., RUAN, F., SOHN, Y. and YAN, J. (2025). The generalization error of max-margin linear classifiers: Benign overfitting and high dimensional asymptotics in the overparametrized regime. *Ann. Statist.* **53** 822–853. MR4900168 https://doi.org/10.1214/25-aos2489
[75] MÜCKE, N., REISS, E., RUNGENHAGEN, J. and KLEIN, M. (2022). Data-splitting improves statistical performance in overparameterized regimes. In *International Conference on Artificial Intelligence and Statistics*.
[76] OYMAK, S. and HASSIBI, B. (2016). Sharp MSE bounds for proximal denoising. *Found. Comput. Math.* **16** 965–1029. MR3529131 https://doi.org/10.1007/s10208-015-9278-4
[77] OYMAK, S., THRAMPOULIDIS, C. and HASSIBI, B. (2013). The squared-error of generalized lasso: A precise analysis. In *Allerton Conference on Communication, Control, and Computing*.
[78] PARIKH, N., BOYD, S. et al. (2014). Proximal algorithms. *Found. Trends Optim.* **1** 127–239.
[79] PATIL, P. and DU, J.-H. (2023). Generalized equivalences between subsampling and ridge regularization. In *Advances in Neural Information Processing Systems*.
[80] PATIL, P., DU, J.-H. and KUCHIBHOTLA, A. K. (2023). Bagging in overparameterized learning: Risk characterization and risk monotonization. *J. Mach. Learn. Res.* **24** 319. MR4664756
[81] PATIL, P., DU, J.-H. and TIBSHIRANI, R. J. (2024). Revisiting optimism and model complexity in the wake of overparameterized machine learning. arXiv preprint. Available at arXiv:2410.01259.
[82] PATIL, P., KUCHIBHOTLA, A. K., WEI, Y. and RINALDO, A. (2022). Mitigating multiple descents: A model-agnostic framework for risk monotonization. arXiv preprint. Available at arXiv:2205.12937.
[83] PATIL, P. and LEJEUNE, D. (2024). Asymptotically free sketched ridge ensembles: Risks, cross-validation, and tuning. In *International Conference on Learning Representations*.
[84] PATIL, P., WEI, Y., RINALDO, A. and TIBSHIRANI, R. (2021). Uniform consistency of cross-validation estimators for high-dimensional ridge regression. In *International Conference on Artificial Intelligence and Statistics*.
[85] PERRONE, M. (1993). Putting it all together: Methods for combining neural networks. In *Advances in Neural Information Processing Systems*.
[86] PESCE, L., KRZAKALA, F., LOUREIRO, B. and STEPHAN, L. (2023). Are Gaussian data all you need? The extents and limits of universality in high-dimensional generalized linear estimation. In *International Conference on Machine Learning*.
[87] RIGOLLET, P. (2012). Kullback-Leibler aggregation and misspecified generalized linear models. *Ann. Statist.* **40** 639–665. MR2933661 https://doi.org/10.1214/11-AOS961
[88] SALEHI, F., ABBASI, E. and HASSIBI, B. (2019). The impact of regularization on high-dimensional logistic regression. In *Advances in Neural Information Processing Systems*.

[89] SAMWORTH, R. J. (2012). Optimal weighted nearest neighbour classifiers. *Ann. Statist.* **40** 2733–2763. MR3097618 https://doi.org/10.1214/12-AOS1049

[90] SOLLICH, P. and KROGH, A. (1995). Learning with ensembles: How overfitting can be useful. In *Advances in Neural Information Processing Systems*.

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

[92] SUR, P. and CANDÈS, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proc. Natl. Acad. Sci. USA* **116** 14516–14525. MR3984492 https://doi.org/10.1073/pnas.1810420116

[93] THRAMPOULIDIS, C., ABBASI, E. and HASSIBI, B. (2018). Precise error analysis of regularized M-estimators in high dimensions. *IEEE Trans. Inf. Theory* **64** 5592–5628. MR3832326 https://doi.org/10.1109/TIT.2018.2840720

[94] THRAMPOULIDIS, C., OYMAK, S. and HASSIBI, B. (2015). Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*.

[95] TIBSHIRANI, R. J. (2013). The lasso problem and uniqueness. *Electron. J. Stat.* **7** 1456–1490. MR3066375 https://doi.org/10.1214/13-EJS815

[96] WANG, S., WENG, H. and MALEKI, A. (2020). Which bridge estimator is the best for variable selection? *Ann. Statist.* **48** 2791–2823. MR4152121 https://doi.org/10.1214/19-AOS1906

[97] WENG, H., MALEKI, A. and ZHENG, L. (2018). Overcoming the limitations of phase transition by higher order analysis of regularization techniques. *Ann. Statist.* **46** 3099–3129. MR3851766 https://doi.org/10.1214/17-AOS1651

The Annals of Statistics
2026, Vol. 54, No. 2, 834–860
https://doi.org/10.1214/25-AOS2591

FINITE- AND LARGE SAMPLE INFERENCE FOR MODEL AND COEFFICIENTS IN HIGH-DIMENSIONAL LINEAR REGRESSION WITH REPRO SAMPLES

BY PENG WANG[1,a], MIN-GE XIE[2,b] AND LINJUN ZHANG[2,c]

[1]*Department of Operations, Business Analytics and Information Systems, University of Cincinnati,* [a]*wangp9@ucmail.uc.edu*
[2]*Department of Statistics, Rutgers University,* [b]*mxie@stat.rutgers.edu,* [c]*linjun.zhang@rutgers.edu*

In this paper, we present a novel and effective inference approach to conduct both finite and large sample inference for high-dimensional linear regression models. This approach is developed under the so-called *repro samples* framework, in which we conduct statistical inference by creating and studying the behavior of artificial samples that are obtained by mimicking the sampling mechanism of the data. We construct confidence sets for (a) the true model corresponding to the nonzero coefficients, (b) a single or any collection of regression coefficients and (c) both the model and regression coefficients jointly. To facilitate the constructions of these confidence sets and overcome computational difficulties of searching all possible models, we use an innovative Fisher inversion technique to construct a model candidate set that includes the true sparse model with probability close to 1 for models with both Gaussian and non-Gaussian errors. The proposed approach fills in two major gaps in the high-dimensional regression literature: (1) lack of effective approaches to addressing model selection uncertainty and providing valid inference for the underlying true model; (2) lack of effective inference approaches to guarantee finite sample performance. We provide both finite sample and asymptotic results to theoretically guarantee the performance of the proposed methods. In addition, our numerical results demonstrate that the proposed methods are valid and achieve better coverage with smaller confidence sets than the current state-of-the-art approaches, such as debiasing and bootstrap approaches.

REFERENCES

[1] ATHEY, S., IMBENS, G. W. and WAGER, S. (2018). Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 597–623. MR3849336 https://doi.org/10.1111/rssb.12268

[2] BEAUMONT, M. A., ZHANG, W. and BALDING, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics* **162** 2025–2035.

[3] BELLONI, A., CHERNOZHUKOV, V., CHETVERIKOV, D., HANSEN, C. and KATO, K. (2018). High-dimensional econometrics and regularized GMM. arXiv preprint. Available at arXiv:1806.01888.

[4] BERTSIMAS, D., KING, A. and MAZUMDER, R. (2016). Best subset selection via a modern optimization lens. *Ann. Statist.* **44** 813–852. MR3476618 https://doi.org/10.1214/15-AOS1388

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

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

[7] CANDÈS, E., FAN, Y., JANSON, L. and LV, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 551–577. MR3798878 https://doi.org/10.1111/rssb.12265

MSC2020 subject classifications. Primary 62J86, 62F40; secondary 62F07, 62J07.
Key words and phrases. High-dimensional inference, model selection uncertainty, irregular inference problem, finite sample inference.

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

[9] CHATTERJEE, A. and LAHIRI, S. N. (2011). Bootstrapping lasso estimators. *J. Amer. Statist. Assoc.* **106** 608–625. MR2847974 https://doi.org/10.1198/jasa.2011.tm10159

[10] CHEN, J. and CHEN, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95** 759–771. MR2443189 https://doi.org/10.1093/biomet/asn034

[11] CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C. and NEWEY, W. (2017). Double/debiased/Neyman machine learning of treatment effects. *Amer. Econ. Rev.* **107** 261–65.

[12] CHERNOZHUKOV, V., HANSEN, C. and SPINDLER, M. (2015). Post-selection and post-regularization inference in linear models with many controls and instruments. *Amer. Econ. Rev.* **105** 486–90.

[13] CRAIU, R. V. and LEVI, E. (2023). Approximate methods for Bayesian computation. *Annu. Rev. Stat. Appl.* **10** 379–399. MR4567798 https://doi.org/10.1146/annurev-statistics-033121-110254

[14] DAI, H. and CHARNIGO, R. (2007). Inferences in contaminated regression and density models. *Sankhyā* **69** 842–869. MR2521235

[15] DAS, D., GREGORY, K. and LAHIRI, S. N. (2019). Perturbation bootstrap in adaptive Lasso. *Ann. Statist.* **47** 2080–2116. MR3953445 https://doi.org/10.1214/18-AOS1741

[16] DAS, D. and LAHIRI, S. N. (2019). Distributional consistency of the lasso by perturbation bootstrap. *Biometrika* **106** 957–964. MR4031208 https://doi.org/10.1093/biomet/asz029

[17] DEZEURE, R., BÜHLMANN, P. and ZHANG, C.-H. (2017). High-dimensional simultaneous inference with the bootstrap. *TEST* **26** 685–719. MR3713586 https://doi.org/10.1007/s11749-017-0554-2

[18] EFRON, B. (1992). Bootstrap methods: Another look at the jackknife. In *Breakthroughs in Statistics* 569–593. Springer, Berlin.

[19] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581 https://doi.org/10.1198/016214501753382273

[20] 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 https://doi.org/10.1111/j.1467-9868.2008.00674.x

[21] FERRARI, D. and YANG, Y. (2015). Confidence sets for model selection by F-testing. *Statist. Sinica* **25** 1637–1658. MR3409085

[22] GUO, Z., RENAUX, C., BÜHLMANN, P. and CAI, T. (2021). Group inference in high dimensions with applications to hierarchical testing. *Electron. J. Stat.* **15** 6633–6676. MR4357274 https://doi.org/10.1214/21-ejs1955

[23] GUO, Z., WANG, W., CAI, T. T. and LI, H. (2019). Optimal estimation of genetic relatedness in high-dimensional linear models. *J. Amer. Statist. Assoc.* **114** 358–369. MR3941260 https://doi.org/10.1080/01621459.2017.1407774

[24] HANNIG, J., IYER, H., LAI, R. C. S. and LEE, T. C. M. (2016). Generalized fiducial inference: A review and new results. *J. Amer. Statist. Assoc.* **111** 1346–1361. MR3561954 https://doi.org/10.1080/01621459.2016.1165102

[25] HANSEN, P. R., LUNDE, A. and NASON, J. M. (2011). The model confidence set. *Econometrica* **79** 453–497. MR2809377 https://doi.org/10.3982/ECTA5771

[26] HOU, X., WANG, P., XIE, M. and ZHANG, L. (2025). Repro samples method for model-free inference in high-dimensional binary classification. arXiv preprint. Available at arXiv:2510.01468.

[27] HOU, X., ZHANG, L., WANG, P. and XIE, M. (2024). Repro samples method for high-dimensional logistic model. arXiv preprint. Available at arXiv:2403.09984.

[28] JAVANMARD, A. and LEE, J. D. (2020). A flexible framework for hypothesis testing in high dimensions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 685–718. MR4112781

[29] JAVANMARD, A. and MONTANARI, A. Confidence Intervals and Hypothesis Testing for High-Dimensional Regression. Available at https://web.stanford.edu/~montanar/sslasso/code.html.

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

[31] KOROBILIS, D. (2021). High-dimensional macroeconomic forecasting using message passing algorithms. *J. Bus. Econom. Statist.* **39** 493–504. MR4235191 https://doi.org/10.1080/07350015.2019.1677472

[32] LANGE, K. L., LITTLE, R. J. A. and TAYLOR, J. M. G. (1989). Robust statistical modeling using the t distribution. *J. Amer. Statist. Assoc.* **84** 881–896. MR1134486

[33] LEI, J. (2020). Cross-validation with confidence. *J. Amer. Statist. Assoc.* **115** 1978–1997. MR4189771 https://doi.org/10.1080/01621459.2019.1672556

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

[35] LEMDANI, M. and PONS, O. (1999). Likelihood ratio tests in contamination models. *Bernoulli* **5** 705–719. MR1704563 https://doi.org/10.2307/3318698

[36] LI, Y., LUO, Y., FERRARI, D., HU, X. and QIN, Y. (2019). Model confidence bounds for variable selection. *Biometrics* **75** 392–403. MR3999165 https://doi.org/10.1111/biom.13024
[37] LI, Y., WANG, W., HOU, X., HUANG, W., ZHANG, P., HE, Y., WANG, B., DUAN, Q., MAO, F. et al. (2023). Glioma-derived LRIG3 interacts with NETO2 in tumor-associated macrophages to modulate microenvironment and suppress tumor growth. *Cell Death Dis.* **14** 28.
[38] LIU, W. (2013). Gaussian graphical model estimation with false discovery rate control. *Ann. Statist.* **41** 2948–2978. MR3161453 https://doi.org/10.1214/13-AOS1169
[39] LIU, Y. and WU, Y. (2007). Variable selection via a combination of the L_0 and L_1 penalties. *J. Comput. Graph. Statist.* **16** 782–798. MR2412482 https://doi.org/10.1198/106186007X255676
[40] MARTIN, R. and LIU, C. (2015). *Inferential Models*: *Reasoning with Uncertainty*. *Monographs on Statistics and Applied Probability* **147**. CRC Press, Boca Raton, FL. MR3618727
[41] MEINSHAUSEN, N., MEIER, L. and BÜHLMANN, P. (2009). p-values for high-dimensional regression. *J. Amer. Statist. Assoc.* **104** 1671–1681. MR2750584 https://doi.org/10.1198/jasa.2009.tm08647
[42] NICKL, R. and VAN DE GEER, S. (2013). Confidence sets in sparse regression. *Ann. Statist.* **41** 2852–2876. MR3161450 https://doi.org/10.1214/13-AOS1170
[43] PEK, J., WONG, O. and WONG, A. C. M. (2018). How to address non-normality: A taxonomy of approaches, reviewed, and illustrated. *Front. Psychol.* **9**. Publisher: Frontiers. https://doi.org/10.3389/fpsyg.2018.02104
[44] SARCAR, B., KAHALI, S. and CHINNAIYAN, P. (2010). Vorinostat enhances the cytotoxic effects of the topoisomerase I inhibitor SN38 in glioblastoma cell lines. *J. Neuro-Oncol.* **99** 201–207.
[45] SHAFER, G. (1976). *A Mathematical Theory of Evidence*. Princeton Univ. Press, Princeton, NJ. MR0464340
[46] SHEN, X., PAN, W. and ZHU, Y. (2012). Likelihood-based selection and sharp parameter estimation. *J. Amer. Statist. Assoc.* **107** 223–232. MR2949354 https://doi.org/10.1080/01621459.2011.645783
[47] SHEN, X., PAN, W., ZHU, Y. and ZHOU, H. (2013). On constrained and regularized high-dimensional regression. *Ann. Inst. Statist. Math.* **65** 807–832. MR3105798 https://doi.org/10.1007/s10463-012-0396-3
[48] TAYLOR, J. and TIBSHIRANI, R. J. (2015). Statistical learning and selective inference. *Proc. Natl. Acad. Sci. USA* **112** 7629–7634. MR3371123 https://doi.org/10.1073/pnas.1507583112
[49] THORNTON, S. and XIE, M. (2024). Bridging Bayesian, frequentist and fiducial inferences using confidence distributions. In *Handbook of Bayesian*, *Fiducial*, *and Frequentist Inference* 106–131. CRC Press, Boca Raton.
[50] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B, Methodol.* **58** 267–288. MR1379242
[51] 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 https://doi.org/10.1080/01621459.2015.1108848
[52] VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285 https://doi.org/10.1214/14-AOS1221
[53] WANG, H., LENGERICH, B. J., ARAGAM, B. and XING, E. P. (2019). Precision Lasso: Accounting for correlations and linear dependencies in high-dimensional genomic data. *Bioinformatics* **35** 1181–1187.
[54] WANG, P., XIE, M. and ZHANG, L. (2026). Supplement to "Finite- and large sample inference for model and coefficients in high-dimensional linear regression with repro samples." https://doi.org/10.1214/25-AOS2591SUPP
[55] WANG, S., NAN, B., ROSSET, S. and ZHU, J. (2011). Random Lasso. *Ann. Appl. Stat.* **5** 468–485. MR2810406 https://doi.org/10.1214/10-AOAS377
[56] WANG, W., SILVA, M. R., CHAMBERS, J. and TSE-DINH, Y.-C. (2017). Exth-09. Tdp1/Top1 Ratio as a Predictive Indicator for the Response of Glioblastoma Cancer Cells to Irinotecan Treatment. Neuro-oncology 19 vi74.
[57] WANG, X., BENESTY, J., CHEN, J. and COHEN, I. (2020). Beamforming with small-spacing microphone arrays using constrained/generalized LASSO. *IEEE Signal Process. Lett.* **27** 356–360.
[58] WILLIAMS, M. N., GRAJALES, C. A. G. and KURKIEWICZ, D. (2019). Assumptions of multiple regression: Correcting two misconceptions. *Pract. Assess. Res. Eval.* **18** 11.
[59] XIE, M. and WANG, P. (2022). Repro Samples Method for Finite- and Large-Sample Inferences. arXiv e-prints. Available at arXiv:2402.15004 (Invited revision for the Journal of the American Statistical Association). https://doi.org/10.48550/arXiv.2206.06421
[60] ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701 https://doi.org/10.1214/09-AOS729
[61] ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940 https://doi.org/10.1111/rssb.12026

[62] ZHANG, X. and CHENG, G. (2017). Simultaneous inference for high-dimensional linear models. *J. Amer. Statist. Assoc.* **112** 757–768. MR3671768 https://doi.org/10.1080/01621459.2016.1166114

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

[64] ZHOU, K., LI, K.-C. and ZHOU, Q. (2023). Honest confidence sets for high-dimensional regression by projection and shrinkage. *J. Amer. Statist. Assoc.* **118** 469–488. MR4571135 https://doi.org/10.1080/01621459.2021.1938581

[65] ZHU, Y. and BRADIC, J. (2017). A Projection Pursuit Framework for Testing General High-Dimensional Hypothesis. Available at arXiv:1705.01024 [math, stat].

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

[67] ZHU, Y., SHEN, X. and PAN, W. (2020). On high-dimensional constrained maximum likelihood inference. *J. Amer. Statist. Assoc.* **115** 217–230. MR4078458 https://doi.org/10.1080/01621459.2018.1540986

[68] ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469 https://doi.org/10.1198/016214506000000735

The Annals of Statistics
2026, Vol. 54, No. 2, 861–881
https://doi.org/10.1214/25-AOS2592

INFERRING THE DEPENDENCE GRAPH DENSITY OF BINARY GRAPHICAL MODELS IN HIGH DIMENSION

BY JULIEN CHEVALLIER[1,a], EVA LÖCHERBACH[2,b] AND GUILHERME OST[3,c]

[1]*Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP (Institute of Engineering Univ. Grenoble Alpes), LJK,* [a]*julien.chevallier1@univ-grenoble-alpes.fr*

[2]*CMAP, Ecole Polytechnique, Institut Polytechnique de Paris,* [b]*eva.loecherbach@polytechnique.edu*

[3]*Institute of Mathematics, Federal University of Rio de Janeiro,* [c]*guilhermeost@im.ufrj.br*

We consider a system of binary interacting chains describing the dynamics of a group of N components that, at each time unit, either send some signal to the others or remain silent otherwise. The interactions among the chains are encoded by a directed Erdös–Rényi random graph with unknown parameter $p \in (0, 1)$. Moreover, the system is structured within two populations (excitatory chains versus inhibitory ones), which are coupled via a mean field interaction on the underlying Erdös–Rényi graph. In this paper, we address the question of inferring the connectivity parameter p based only on the observation of the interacting chains over T time units. In our main result, we show that the connectivity parameter p can be estimated with rate $N^{-1/2} + N^{1/2}/T + (\log(T)/T)^{1/2}$ through an easy-to-compute estimator. Our analysis relies on a precise study of the spatiotemporal decay of correlations of the interacting chains. This is done through the study of coalescing random walks defining a backward regeneration representation of the system. Interestingly, we also show that this backward regeneration representation allows us to perfectly sample the system of interacting chains (conditionally on each realization of the underlying Erdös–Rényi graph) from its stationary distribution. These probabilistic results have an interest in its own.

REFERENCES

BASU, S. and MICHAILIDIS, G. (2015). Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.* **43** 1535–1567. MR3357870 https://doi.org/10.1214/15-AOS1315

BRESLER, G. (2015). Efficiently learning Ising models on arbitrary graphs [extended abstract]. In *STOC'*15*—Proceedings of the* 2015 *ACM Symposium on Theory of Computing* 771–782. ACM, New York. MR3388257

CHEVALLIER, J., LÖCHERBACH, E. and OST, G. (2026). Supplement to "Inferring the dependence graph density of binary graphical models in high dimension." https://doi.org/10.1214/25-AOS2592SUPP

CHEVALLIER, J. and OST, G. (2024). Community detection for binary graphical models in high dimension. arXiv preprint. Available at arXiv:2411.15627.

COMETS, F., FERNÁNDEZ, R. and FERRARI, P. A. (2002). Processes with long memory: Regenerative construction and perfect simulation. *Ann. Appl. Probab.* **12** 921–943. MR1925446 https://doi.org/10.1214/aoap/1031863175

DE SANTIS, E. and PICCIONI, M. (2015). One-dimensional infinite memory imitation models with noise. *J. Stat. Phys.* **161** 346–364. MR3401021 https://doi.org/10.1007/s10955-015-1335-5

DELATTRE, S. and FOURNIER, N. (2016). Statistical inference versus mean field limit for Hawkes processes. *Electron. J. Stat.* **10** 1223–1295. MR3499526 https://doi.org/10.1214/16-EJS1142

DUARTE, A., GALVES, A., LÖCHERBACH, E. and OST, G. (2019). Estimating the interaction graph of stochastic neural dynamics. *Bernoulli* **25** 771–792. MR3892336 https://doi.org/10.3150/17-bej1006

EICHLER, M. (2012). Graphical modelling of multivariate time series. *Probab. Theory Related Fields* **153** 233–268. MR2925574 https://doi.org/10.1007/s00440-011-0345-8

FERNÁNDEZ, R., FERRARI, P. A. and GALVES, A. (2001). Coupling, renewal and perfect simulation of chains of infinite order. Notes for a minicourse given at the vth Brazilian school of probability.

MSC2020 subject classifications. Primary 62M05; secondary 60J10, 60K35, 62F12.
Key words and phrases. Dependence graph inference, Markov chain, mean field limit, perfect simulation.

FERRARI, P. A. (1990). Ergodicity for spin systems with stirrings. *Ann. Probab.* **18** 1523–1538. MR1071806

FERRARI, P. A., MAASS, A., MARTÍNEZ, S. and NEY, P. (2000). Cesàro mean distribution of group automata starting from measures with summable decay. *Ergod. Theory Dyn. Syst.* **20** 1657–1670. MR1804951 https://doi.org/10.1017/S0143385700000924

GAITONDE, J., MOITRA, A. and MOSSEL, E. (2025). Bypassing the noisy parity barrier: Learning higher-order Markov random fields from dynamics. In *STOC'25—Proceedings of the 57th Annual ACM Symposium on Theory of Computing* 348–359. ACM, New York. MR4928431 https://doi.org/10.1145/3717823.3718231

KIM, B., LIU, S. and KOLAR, M. (2021). Two-sample inference for high-dimensional Markov networks. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **83** 939–962. MR4349123 https://doi.org/10.1111/rssb.12446

LAURITZEN, S. L. (1996). *Graphical Models*. *Oxford Statistical Science Series* **17**. The Clarendon Press, New York. MR1419991

LERASLE, M. and TAKAHASHI, D. Y. (2016). Sharp oracle inequalities and slope heuristic for specification probabilities estimation in discrete random fields. *Bernoulli* **22** 325–344. MR3449785 https://doi.org/10.3150/14-BEJ660

LIU, C. (2020). Statistical inference for a partially observed interacting system of Hawkes processes. *Stoch. Process. Appl.* **130** 5636–5694. MR4127342 https://doi.org/10.1016/j.spa.2020.04.003

MONTANARI, A. and PEREIRA, J. (2009). Which graphical models are difficult to learn? In *Advances in Neural Information Processing Systems* (Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams and A. Culotta, eds.). **22**. Curran Associates, Red Hook.

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 https://doi.org/10.1214/09-AOS691

REYNAUD-BOURET, P., RIVOIRARD, V. and TULEAU-MALOT, C. (2013). Inference of functional connectivity in neurosciences via Hawkes processes. In 2013 *IEEE Global Conference on Signal and Information Processing* 317–320.

STROGATZ, S. H. (2001). Exploring complex networks. *Nature* **410** 268–276.

The Annals of Statistics
2026, Vol. 54, No. 2, 882–909
https://doi.org/10.1214/25-AOS2593

EIGENVECTOR OVERLAPS IN LARGE SAMPLE COVARIANCE MATRICES AND NONLINEAR SHRINKAGE ESTIMATORS

BY ZEQIN LIN[a] AND GUANGMING PAN[b]

School of Physical and Mathematical Sciences, Nanyang Technological University, [a]*ZEQIN001@e.ntu.edu.sg,* [b]*GMPAN@ntu.edu.sg*

Consider a data matrix $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$ of size $M \times N$, where the columns are independent observations from a random vector $\mathbf{y}$ with zero mean and population covariance Σ. Let $\mathbf{u}_i$ and $\mathbf{v}_j$ denote the left and right singular vectors of Y, respectively. This study investigates the eigenvector/singular vector overlaps $\langle \mathbf{u}_i, D_1 \mathbf{u}_j \rangle$, $\langle \mathbf{v}_i, D_2 \mathbf{v}_j \rangle$ and $\langle \mathbf{u}_i, D_3 \mathbf{v}_j \rangle$, where D_k are general deterministic matrices with bounded operator norms. In the high-dimensional regime, where the dimension M scales proportionally with the sample size N, we establish the convergence in probability of these eigenvector overlaps towards their deterministic counterparts with explicit convergence rates. Building upon these findings, we offer a more precise characterization of the loss associated with Ledoit and Wolf's nonlinear shrinkage estimators of the population covariance Σ.

REFERENCES

[1] ADHIKARI, A., DUBOVA, S., XU, C. and YIN, J. (2024). Eigenstate thermalization hypothesis for generalized Wigner matrices. *Electron. J. Probab.* **29** Paper No. 141, 33. MR4798613 https://doi.org/10.1214/24-ejp1186

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

[3] BAI, Z. D. and SILVERSTEIN, J. W. (1998). No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *Ann. Probab.* **26** 316–345. MR1617051 https://doi.org/10.1214/aop/1022855421

[4] BAO, Z., DING, X., WANG, J. and WANG, K. (2022). Statistical inference for principal components of spiked covariance matrices. *Ann. Statist.* **50** 1144–1169. MR4404931 https://doi.org/10.1214/21-aos2143

[5] BAO, Z., PAN, G. and ZHOU, W. (2015). Universality for the largest eigenvalue of sample covariance matrices with general population. *Ann. Statist.* **43** 382–421. MR3311864 https://doi.org/10.1214/14-AOS1281

[6] BENAYCH-GEORGES, F. (2023). A Short Proof of Ledoit–Péché's RIE Formula for Covariance Matrices. https://doi.org/10.48550/arXiv.2201.05690

[7] BENAYCH-GEORGES, F., BOUCHAUD, J.-P. and POTTERS, M. (2023). Optimal cleaning for singular values of cross-covariance matrices. *Ann. Appl. Probab.* **33** 1295–1326. MR4564427 https://doi.org/10.1214/22-aap1842

[8] BENAYCH-GEORGES, F. and KNOWLES, A. (2018). Lectures on the Local Semicircle Law for Wigner Matrices. https://doi.org/10.48550/arXiv.1601.04055

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

[10] CAI, T. T., REN, Z. and ZHOU, H. H. (2016). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electron. J. Stat.* **10** 1–59. MR3466172 https://doi.org/10.1214/15-EJS1081

[11] CIPOLLONI, G., ERDŐS, L., HENHEIK, J. and KOLUPAIEV, O. (2023). Gaussian fluctuations in the equipartition principle for Wigner matrices. *Forum Math. Sigma* **11** Paper No. e74, 40. MR4634106 https://doi.org/10.1017/fms.2023.70

MSC2020 subject classifications. Primary 62E20; secondary 60B20.
Key words and phrases. Sample covariance matrices, nonlinear shrinkage estimators, eigenvector overlaps.

[12] CIPOLLONI, G., ERDŐS, L., HENHEIK, J. and SCHRÖDER, D. (2024). Optimal lower bound on eigenvector overlaps for non-Hermitian random matrices. *J. Funct. Anal.* **287** Paper No. 110495, 90. MR4748758 https://doi.org/10.1016/j.jfa.2024.110495

[13] CIPOLLONI, G., ERDŐS, L. and SCHRÖDER, D. (2021). Eigenstate thermalization hypothesis for Wigner matrices. *Comm. Math. Phys.* **388** 1005–1048. MR4334253 https://doi.org/10.1007/s00220-021-04239-z

[14] CIPOLLONI, G., ERDŐS, L. and SCHRÖDER, D. (2022). Optimal multi-resolvent local laws for Wigner matrices. *Electron. J. Probab.* **27** Paper No. 117, 38. MR4479913 https://doi.org/10.1214/22-ejp838

[15] CIPOLLONI, G., ERDŐS, L. and SCHRÖDER, D. (2022). Rank-uniform local law for Wigner matrices. *Forum Math. Sigma* **10** Paper No. e96, 43. MR4502022 https://doi.org/10.1017/fms.2022.86

[16] CIPOLLONI, G., ERDŐS, L. and SCHRÖDER, D. (2022). Thermalisation for Wigner matrices. *J. Funct. Anal.* **282** Paper No. 109394, 37. MR4372147 https://doi.org/10.1016/j.jfa.2022.109394

[17] CIPOLLONI, G., ERDŐS, L. and SCHRÖDER, D. (2022). Normal fluctuation in quantum ergodicity for Wigner matrices. *Ann. Probab.* **50** 984–1012. MR4413210 https://doi.org/10.1214/21-aop1552

[18] CIPOLLONI, G., ERDŐS, L. and SCHRÖDER, D. (2023). Eigenstate Thermalisation at the Edge for Wigner Matrices. https://doi.org/10.48550/arXiv.2309.05488

[19] CIPOLLONI, G., ERDŐS, L. and XU, Y. (2024). Universality of Extremal Eigenvalues of Large Random Matrices. https://doi.org/10.48550/arXiv.2312.08325

[20] DING, X., LI, Y. and YANG, F. (2024). Eigenvector Distributions and Optimal Shrinkage Estimators for Large Covariance and Precision Matrices. https://doi.org/10.48550/arXiv.2404.14751

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

[22] EL KAROUI, N. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Ann. Statist.* **36** 2757–2790. MR2485012 https://doi.org/10.1214/07-AOS581

[23] ERDŐS, L. (2019). The Matrix Dyson Equation and Its Applications for Random Matrices. https://doi.org/10.48550/arXiv.1903.10060

[24] ERDŐS, L., KNOWLES, A., YAU, H.-T. and YIN, J. (2013). The local semicircle law for a general class of random matrices. *Electron. J. Probab.* **18** no. 59, 58. MR3068390 https://doi.org/10.1214/EJP.v18-2473

[25] ERDŐS, L. and RIABOV, V. (2024). Eigenstate thermalization hypothesis for Wigner-type matrices. *Comm. Math. Phys.* **405** Paper No. 282, 70. MR4819934 https://doi.org/10.1007/s00220-024-05143-y

[26] ERDŐS, L., YAU, H.-T. and YIN, J. (2012). Rigidity of eigenvalues of generalized Wigner matrices. *Adv. Math.* **229** 1435–1515. MR2871147 https://doi.org/10.1016/j.aim.2011.12.010

[27] FAN, J., LIAO, Y. and LIU, H. (2016). An overview of the estimation of large covariance and precision matrices. *Econom. J.* **19** C1–C32. MR3501529 https://doi.org/10.1111/ectj.12061

[28] HACHEM, W., HARDY, A. and NAJIM, J. (2016). Large complex correlated Wishart matrices: Fluctuations and asymptotic independence at the edges. *Ann. Probab.* **44** 2264–2348. MR3502605 https://doi.org/10.1214/15-AOP1022

[29] HACHEM, W., HARDY, A. and NAJIM, J. (2016). Large complex correlated Wishart matrices: The Pearcey kernel and expansion at the hard edge. *Electron. J. Probab.* **21** Paper No. 1, 36. MR3485343 https://doi.org/10.1214/15-EJP4441

[30] HE, Y. and KNOWLES, A. (2017). Mesoscopic eigenvalue statistics of Wigner matrices. *Ann. Appl. Probab.* **27** 1510–1550. MR3678478 https://doi.org/10.1214/16-AAP1237

[31] JING, B.-Y., PAN, G., SHAO, Q.-M. and ZHOU, W. (2010). Nonparametric estimate of spectral density functions of sample covariance matrices: A first step. *Ann. Statist.* **38** 3724–3750. MR2766866 https://doi.org/10.1214/10-AOS833

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

[33] KONG, W. and VALIANT, G. (2017). Spectrum estimation from samples. *Ann. Statist.* **45** 2218–2247. MR3718167 https://doi.org/10.1214/16-AOS1525

[34] LAM, C. (2020). High-dimensional covariance matrix estimation. *Wiley Interdiscip. Rev.: Comput. Stat.* **12** Paper No. e1485, 21. MR4072468 https://doi.org/10.1002/wics.1485

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

[36] LEDOIT, O. and WOLF, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.* **88** 365–411. MR2026339 https://doi.org/10.1016/S0047-259X(03)00096-4

[37] LEDOIT, O. and WOLF, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Ann. Statist.* **40** 1024–1060. MR2985942 https://doi.org/10.1214/12-AOS989

[38] LEDOIT, O. and WOLF, M. (2015). Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions. *J. Multivariate Anal.* **139** 360–384. MR3349498 https://doi.org/10.1016/j.jmva.2015.04.006

[39] LEDOIT, O. and WOLF, M. (2018). Optimal estimation of a large-dimensional covariance matrix under Stein's loss. *Bernoulli* **24** 3791–3832. MR3788189 https://doi.org/10.3150/17-BEJ979
[40] LEDOIT, O. and WOLF, M. (2020). Analytical nonlinear shrinkage of large-dimensional covariance matrices. *Ann. Statist.* **48** 3043–3065. MR4152634 https://doi.org/10.1214/19-AOS1921
[41] LEDOIT, O. and WOLF, M. (2021). Shrinkage estimation of large covariance matrices: Keep it simple, statistician? *J. Multivariate Anal.* **186** Paper No. 104796, 24. MR4308803 https://doi.org/10.1016/j.jmva.2021.104796
[42] LEDOIT, O. and WOLF, M. (2022). The power of (non-)linear shrinking: a review and guide to covariance matrix estimation. *J. Financ. Econom.* **20** 187–218. https://doi.org/10.1093/jjfinec/nbaa007
[43] LEDOIT, O. and WOLF, M. (2022). Quadratic shrinkage for large covariance matrices. *Bernoulli* **28** 1519–1547. MR4411501 https://doi.org/10.3150/20-bej1315
[44] LI, W., CHEN, J., QIN, Y., BAI, Z. and YAO, J. (2013). Estimation of the population spectral distribution from a large dimensional sample covariance matrix. *J. Statist. Plann. Inference* **143** 1887–1897. MR3095079 https://doi.org/10.1016/j.jspi.2013.06.017
[45] LI, W. and YAO, J. (2014). A local moment estimator of the spectrum of a large dimensional covariance matrix. *Statist. Sinica* **24** 919–936. MR3235405
[46] LIN, Z. and PAN, G. (2026). Supplement to "Eigenvector overlaps in large sample covariance matrices and nonlinear shrinkage estimators." https://doi.org/10.1214/25-AOS2593SUPP
[47] LYTOVA, A. and PASTUR, L. (2009). Central limit theorem for linear eigenvalue statistics of random matrices with independent entries. *Ann. Probab.* **37** 1778–1840. MR2561434 https://doi.org/10.1214/09-AOP452
[48] MARČENKO, V. A. and PASTUR, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Mat. Sb.* **114** 507–536. https://doi.org/10.1070/SM1967v001n04ABEH001994
[49] PAUL, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17** 1617–1642. MR2399865
[50] POURAHMADI, M. (2013). *High-Dimensional Covariance Estimation. Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR3235948 https://doi.org/10.1002/9781118573617
[51] 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 https://doi.org/10.1214/07-AOS583
[52] SILVERSTEIN, J. W. (1984). Some limit theorems on the eigenvectors of large-dimensional sample covariance matrices. *J. Multivariate Anal.* **15** 295–324. MR0768500 https://doi.org/10.1016/0047-259X(84)90054-X
[53] SILVERSTEIN, J. W. (1989). On the eigenvectors of large-dimensional sample covariance matrices. *J. Multivariate Anal.* **30** 1–16. MR1003705 https://doi.org/10.1016/0047-259X(89)90084-5
[54] SILVERSTEIN, J. W. (1990). Weak convergence of random functions defined by the eigenvectors of sample covariance matrices. *Ann. Probab.* **18** 1174–1194. MR1062064
[55] SILVERSTEIN, J. W. and CHOI, S.-I. (1995). Analysis of the limiting spectral distribution of large-dimensional random matrices. *J. Multivariate Anal.* **54** 295–309. MR1345541 https://doi.org/10.1006/jmva.1995.1058
[56] STEIN, C. (1975). Estimation of a covariance matrix. In 39*th Annual Meeting IMS, Atlanta, GA*.
[57] STEIN, C. (1986). Lectures on the theory of estimation of many parameters. *J. Sov. Math.* **34** 1373–1403. https://doi.org/10.1007/BF01085007
[58] XI, H., YANG, F. and YIN, J. (2020). Convergence of eigenvector empirical spectral distribution of sample covariance matrices. *Ann. Statist.* **48** 953–982. MR4102683 https://doi.org/10.1214/19-AOS1832
[59] YANG, F. (2020). Linear Spectral Statistics of Eigenvectors of Anisotropic Sample Covariance Matrices. https://doi.org/10.48550/arXiv.2005.00999

The Annals of Statistics
2026, Vol. 54, No. 2, 910–928
https://doi.org/10.1214/25-AOS2595

OBJECT DETECTION UNDER THE LINEAR SUBSPACE MODEL WITH APPLICATION TO CRYO-EM IMAGES

BY KEREN MOR WAKNIN[1,b], AMITAY ELDAR[1,a], SAMUEL DAVENPORT[2,d], TAMIR BENDORY[3,f], ARMIN SCHWARTZMAN[2,e] AND YOEL SHKOLNISKY[1,c]

[1]*Department of Applied Mathematics, Tel Aviv University,* [a]*amitayeldar@tauex.tau.ac.il,* [b]*kerenmor@tauex.tau.ac.il,* [c]*yoelsh@tauex.tau.ac.il*

[2]*Halıcıoğlu Data Science Institute and Division of Biostatistics, University of California,* [d]*sdavenport@health.ucsd.edu,* [e]*armins@ucsd.edu*

[3]*School of Electrical Engineering, Tel Aviv University,* [f]*bendory@tauex.tau.ac.il*

Detecting multiple unknown objects in noisy data is a key problem in many scientific fields, such as electron microscopy imaging. A common model for the unknown objects is the linear subspace model, which assumes that the objects can be expanded in some known basis (such as the Fourier basis). In this paper, we develop an object detection algorithm that under the linear subspace model is asymptotically guaranteed to detect all objects, while controlling the familywise error rate or the false discovery rate. Numerical simulations show that the algorithm also controls the error rate with high power in the nonasymptotic regime, even in highly challenging regimes. We apply the proposed algorithm to an experimental electron microscopy data set, and show that it outperforms existing standard software.

REFERENCES

[1] AIGER, D. and TALBOT, H. (2010). The phase only transform for unsupervised surface defect detection. In 2010 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 295–302. IEEE.

[2] AN, J. and CHO, S. (2015). Variational autoencoder based anomaly detection using reconstruction probability. In *Special Lecture on IE* **2** 1–18.

[3] BALANOV, A., HULEIHEL, W. and BENDORY, T. (2024). Einstein from noise: Statistical analysis. bioRxiv 2024–07.

[4] BENDORY, T., BARTESAGHI, A. and SINGER, A. (2020). Single-particle cryo-electron microscopy: Mathematical theory, computational challenges, and opportunities. *IEEE Signal Process. Mag.* **37** 58–76.

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

[6] BEPLER, T., MORIN, A., RAPP, M., BRASCH, J., SHAPIRO, L., NOBLE, A. J. and BERGER, B. (2019). Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. *Nat. Methods* **16** 1153–1160.

[7] BORACCHI, G., CARRERA, D. and WOHLBERG, B. (2014). Novelty detection in images by sparse representations. In 2014 *IEEE Symposium on Intelligent Embedded Systems* (*IES*) 47–54. IEEE.

[8] BRUTTI, P., GENOVESE, C. R., MILLER, C. J., NICHOL, R. C. and WASSERMAN, L. H. (2005). Spike Hunting in Galaxy Spectra Technical report, Carnegie Mellon Univ.

[9] CHENG, D. and SCHWARTZMAN, A. (2017). Multiple testing of local maxima for detection of peaks in random fields. *Ann. Statist.* **45** 529–556. MR3650392 https://doi.org/10.1214/16-AOS1458

[10] DAVENPORT, S., NICHOLS, T. E. and SCHWARZMAN, A. (2022). Confidence regions for the location of peaks of a smooth random field. arXiv preprint. Available at arXiv:2208.00251.

[11] DAVY, A., EHRET, T., MOREL, J.-M. and DELBRACIO, M. (2018). Reducing anomaly detection in images to detection in noise. In 2018 25*th IEEE International Conference on Image Processing* (*ICIP*) 1058–1062. IEEE.

MSC2020 subject classifications. Primary 6008, 60G15, 60G10, 60G35, 62M20, 62M40; secondary G.3, I.4.

Key words and phrases. Object detection, multiple hypothesis testing, false discovery rate, familywise error rate, matched filter.

[12] DE LA ROSA-TREVÍN, J. M., OTÓN, J., MARABINI, R., ZALDÍVAR, A., VARGAS, J., CARAZO, J. M. and SORZANO, C. O. S. (2013). Xmipp 3.0: An improved software suite for image processing in electron microscopy. *J. Struct. Biol.* **184** 321–328.

[13] DHAKAL, A., GYAWALI, R., WANG, L. and CHENG, J. (2023). A large expert-curated cryo-EM image dataset for machine learning protein particle picking. *Sci. Data* **10** 392. https://doi.org/10.1038/s41597-023-02280-2

[14] EDITORIAL (2016). Method of the year 2015. *Nat. Methods* **13** 1.

[15] EHRET, T., DAVY, A., MOREL, J.-M. and DELBRACIO, M. (2019). Image anomalies: A review and synthesis of detection methods. *J. Math. Imaging Vision* **61** 710–743. MR3958816 https://doi.org/10.1007/s10851-019-00885-0

[16] ELDAR, A., AMOS, I. and SHKOLNISKY, Y. (2022). ASOCEM: Automatic Segmentation Of Contaminations in cryo-EM. *J. Struct. Biol.* **214** 107871.

[17] ELDAR, A., LANDA, B. and SHKOLNISKY, Y. (2020). KLT picker: Particle picking using data-driven optimal templates. *J. Struct. Biol.* **210** 107473.

[18] ELDAR, A., MOR WAKNIN, K., DAVENPORT, S., BENDORY, T., SCHWARTZMAN, A. and SHKOLNISKY, Y. (2026). Supplement to "Object detection under the linear subspace model with application to cryo-EM images." https://doi.org/10.1214/25-AOS2595SUPPA, https://doi.org/10.1214/25-AOS2595SUPPB

[19] GEISLER, C., SCHÖNLE, A., VON MIDDENDORFF, C., BOCK, H., EGGELING, C., EGNER, A. and HELL, S. W. (2007). Resolution of $\lambda/10$ in fluorescence microscopy using fast single molecule photo-switching. *Appl. Phys. A* **88** 223–226.

[20] GENOVESE, C. R., LAZAR, N. A. and NICHOLS, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage* **15** 870–878.

[21] GROSJEAN, B. and MOISAN, L. (2009). A-contrario detectability of spots in textured backgrounds. *J. Math. Imaging Vision* **33** 313–337. MR2480965 https://doi.org/10.1007/s10851-008-0111-4

[22] HENDERSON, R. (2013). Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein from noise. *Proc. Natl. Acad. Sci. USA* **110** 18037–18041.

[23] IUDIN, A., KORIR, P. K., SOMASUNDHARAM, S., WEYAND, S., CATTAVITELLO, C., FONSECA, N., SALIH, O., KLEYWEGT, G. J. and PATWARDHAN, A. (2023). EMPIAR: The electron microscopy public image archive. *Nucleic Acids Res.* **51** D1503–D1511.

[24] LYUMKIS, D. (2019). Challenges and opportunities in cryo-EM single-particle analysis. *J. Biol. Chem.* **294** 5181–5197.

[25] MARANDON, A., LEI, L., MARY, D. and ROQUAIN, E. (2022). Machine learning meets false discovery rate. arXiv preprint. Available at arXiv:2208.06685.

[26] PERNG, D.-B., CHEN, S.-H. and CHANG, Y.-S. (2010). A novel internal thread defect auto-inspection system. *Int. J. Adv. Manuf. Technol.* **47** 731–743.

[27] RUPERT, G. JR et al. (2012). *Simultaneous Statistical Inference*. Springer, Berlin.

[28] SCHWARTZMAN, A., GAVRILOV, Y. and ADLER, R. J. (2011). Multiple testing of local maxima for detection of peaks in 1D. *Ann. Statist.* **39** 3290–3319. MR3012409 https://doi.org/10.1214/11-AOS943

[29] TANG, G., PENG, L., BALDWIN, P. R., MANN, D. S., JIANG, W., REES, I. and LUDTKE, S. J. (2007). EMAN2: An extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157** 38–46.

[30] TANG, G., PENG, L., BALDWIN, P. R., MANN, D. S., JIANG, W., REES, I. and LUDTKE, S. J. (2007). EMAN2: An extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157** 38–46. Software tools for macromolecular microscopy. https://doi.org/10.1016/j.jsb.2006.05.009

[31] TARTAKOVSKY, A., NIKIFOROV, I. and BASSEVILLE, M. (2014). *Sequential Analysis*: *Hypothesis Testing and Changepoint Detection*. *Monographs on Statistics and Applied Probability* **136**. CRC Press, Boca Raton, FL. MR3241619

[32] TAYLOR, J. E. and WORSLEY, K. J. (2007). Detecting sparse signals in random fields, with an application to brain mapping. *J. Amer. Statist. Assoc.* **102** 913–928. MR2354405 https://doi.org/10.1198/016214507000000815

[33] TSAI, D.-M. and HSIEH, C.-Y. (1999). Automated surface inspection for directional textures. *Image Vis. Comput.* **18** 49–62.

[34] WONG, W., BAI, X.-C., BROWN, A., FERNANDEZ, I. S., HANSSEN, E., CONDRON, M., TAN, Y. H., BAUM, J. and SCHERES, S. H. (2014). Cryo-EM structure of the Plasmodium Falciparum 80S ribosome bound to the anti-protozoan drug emetine. *eLife* **3** e03080. https://doi.org/10.7554/eLife.03080

[35] WORSLEY, K. J., MARRETT, S., NEELIN, P., VANDAL, A. C., FRISTON, K. J. and EVANS, A. C. (1996). A unified statistical approach for determining significant signals in images of cerebral activation. *Hum. Brain Mapp.* **4** 58–73.

[36] XIE, X. and MIRMEHDI, M. (2007). TEXEMS: Texture exemplars for defect detection on random textured surfaces. *IEEE Trans. Pattern Anal. Mach. Intell.* **29** 1454–1464.

[37] ZHAO, Z. and SINGER, A. (2013). Fourier–Bessel rotational invariant eigenimages. *J. Opt. Soc. Amer. A* **30** 871–877.

[38] ZIVANOV, J., NAKANE, T., FORSBERG, B. O., KIMANIUS, D., HAGEN, W. J., LINDAHL, E. and SCHERES, S. H. (2018). New tools for automated high-resolution cryo-EM structure determination in RELION. *eLife* **7** e42166.

[39] ZIVANOV, J., NAKANE, T., FORSBERG, B. O., KIMANIUS, D., HAGEN, W. J., LINDAHL, E. and SCHERES, S. H. (2018). New tools for automated high-resolution cryo-EM structure determination in RELION. *eLife* **7** e42166. https://doi.org/10.7554/eLife.42166

[40] ZONTAK, M. and COHEN, I. (2009). Defect detection in patterned wafers using multichannel scanning electron microscope. *Signal Process.* **89** 1511–1520.

[41] ZONTAK, M. and COHEN, I. (2010). Defect detection in patterned wafers using anisotropic kernels. *Mach. Vis. Appl.* **21** 129–141.

The Annals of Statistics
2026, Vol. 54, No. 2, 929–956
https://doi.org/10.1214/25-AOS2596

PCA FOR POINT PROCESSES

BY FRANCK PICARD[1,a], VINCENT RIVOIRARD[2,b], ANGELINA ROCHE[2,c] AND VICTOR M. PANARETOS[3,d]

[1]*Laboratoire de Biologie et Modélisation de la Cellule, CNRS ENS de Lyon,* [a]*franck.picard@ens-lyon.fr*
[2]*CEREMADE, Université Paris Dauphine,* [b]*Vincent.Rivoirard@dauphine.fr,* [c]*roche@ceremade.dauphine.fr*
[3]*Institut de Mathématiques, Ecole Polytechnique Fédérale de Lausanne,* [d]*victor.panaretos@epfl.ch*

We introduce a novel statistical framework for the analysis of replicated point processes that allows for the study of point pattern variability at a population level. By treating point process realizations as random measures, we adopt a functional analysis perspective and propose a form of functional Principal Component Analysis (fPCA) for point processes. The originality of our method is to base our analysis on the cumulative mass functions of the random measures, which gives us a direct and interpretable analysis. Key theoretical contributions include establishing a Karhunen–Loève expansion for the random measures and a Mercer theorem for covariance measures. We establish convergence in a strong sense, and introduce the concept of principal measures, which can be seen as latent processes governing the dynamics of the observed point patterns. We propose an easy-to-implement estimation strategy of eigenelements for which parametric rates are achieved. We fully characterize the solutions of our approach to Poisson and Hawkes processes and validate our methodology via simulations and diverse applications in seismology, single-cell biology and neurosiences, demonstrating its versatility and effectiveness. Our method is implemented in the `pppca` R-package.

REFERENCES

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

[2] ASH, R. B. and GARDNER, M. F. (1975). *Topics in Stochastic Processes*. *Probability and Mathematical Statistics*, *Vol*. 27. Academic Press, New York–London. MR0448463

[3] BACRY, E., BOMPAIRE, M., GAÏFFAS, S. and MUZY, J.-F. (2020). Sparse and low-rank multivariate Hawkes processes. *J. Mach. Learn. Res*. **21** Paper No. 50, 32. MR4095329

[4] BELITSER, E., SERRA, P. and VAN ZANTEN, H. (2015). Rate-optimal Bayesian intensity smoothing for inhomogeneous Poisson processes. *J. Statist. Plann. Inference* **166** 24–35. MR3390131 https://doi.org/10.1016/j.jspi.2014.03.009

[5] BONNET, A., DION-BLANC, C., GINDRAUD, F. and LEMLER, S. (2022). Neuronal network inference and membrane potential model using multivariate Hawkes processes. *J. Neurosci. Methods* **372** 109550. https://doi.org/10.1016/j.jneumeth.2022.109550

[6] BONNET, A., MARTINEZ HERRERA, M. and SANGNIER, M. (2023). Inference of multivariate exponential Hawkes processes with inhibition and application to neuronal activity. *Stat. Comput*. **33** Paper No. 91, 26. MR4606276 https://doi.org/10.1007/s11222-023-10264-w

[7] BOUZAS, P. R., VALDERRAMA, M. J., AGUILERA, A. M. and RUIZ-FUENTES, N. (2006). Modelling the mean of a doubly stochastic Poisson process by functional data analysis. *Comput. Statist. Data Anal*. **50** 2655–2667. MR2227341 https://doi.org/10.1016/j.csda.2005.04.015

[8] BRÉMAUD, P. and MASSOULIÉ, L. (1996). Stability of nonlinear Hawkes processes. *Ann. Probab.* **24** 1563–1588. MR1411506 https://doi.org/10.1214/aop/1065725193

[9] BRÉMAUD, P. and MASSOULIÉ, L. (2001). Hawkes branching point processes without ancestors. *J. Appl. Probab.* **38** 122–135. MR1816118 https://doi.org/10.1017/s0021900200018556

MSC2020 subject classifications. Primary 62M, 62P; secondary 60G55, 62H25.
Key words and phrases. Point processes, Karhunen–Loève expansion, PCA.

[10] BREZIS, H. (2011). *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. *Universitext*. Springer, New York. MR2759829
[11] CARRIZO VERGARA, R. (2022). Karhunen-Loève expansion of Random Measures. Available at arXiv:2203.14202.
[12] CARSTENSEN, L., SANDELIN, A., WINTHER, O. and HANSEN, N. R. (2010). Multivariate Hawkes process models of the occurrence of regulatory elements. *BMC Bioinform*. **11** 456.
[13] CHEN, S., WITTEN, D. and SHOJAIE, A. (2017). Nearly assumptionless screening for the mutually-exciting multivariate Hawkes process. *Electron. J. Stat*. **11** 1207–1234. MR3634334 https://doi.org/10.1214/17-EJS1251
[14] CHEYSSON, F. (2023). hawkesbow: Estimation of Hawkes Processes from Binned Observations R package version 1.0.2.
[15] CHIANG, W.-H., LIU, X. and MOHLER, G. (2022). Hawkes process modeling of COVID-19 with mobility leading indicators and spatial covariates. *Int. J. Forecast*. **38** 505–520. https://doi.org/10.1016/j.ijforecast.2021.07.001
[16] CHIU, S. N., STOYAN, D., KENDALL, W. S. and MECKE, J. (2013). *Stochastic Geometry and Its Applications*, 3rd ed. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR3236788 https://doi.org/10.1002/9781118658222
[17] CHORNOBOY, E. S., SCHRAMM, L. P. and KARR, A. F. (1988). Maximum likelihood identification of neural point process systems. *Biol. Cybernet*. **59** 265–275. MR0961117 https://doi.org/10.1007/BF00332915
[18] COHN, D. L. (1993). *Measure Theory*. Birkhäuser, Boston, MA. MR1454121
[19] CORLAY, S. and PAGÈS, G. (2015). Functional quantization-based stratified sampling methods. *Monte Carlo Methods Appl*. **21** 1–32. MR3318550 https://doi.org/10.1515/mcma-2014-0010
[20] CRANE, R. and SORNETTE, D. (2008). Robust dynamic classes revealed by measuring the response function of a social system. *Proc. Natl. Acad. Sci. USA* **105** 15649–15653.
[21] CUNNINGHAM, J. P. and YU, B. M. (2014). Dimensionality reduction for large-scale neural recordings. *Nat. Neurosci*. **17** 1500–1509.
[22] DALEY, D. J. and VERE-JONES, D. (2003). *An Introduction to the Theory of Point Processes. Vol. I. Elementary Theory and Methods*, 2nd ed. *Probability and Its Applications* (*New York*). Springer, New York. MR1950431
[23] DALEY, D. J. and VERE-JONES, D. (2008). *An Introduction to the Theory of Point Processes. Vol. II. General Theory and Structure*, 2nd ed. *Probability and Its Applications* (*New York*). Springer, New York. MR2371524 https://doi.org/10.1007/978-0-387-49835-5
[24] DELAIGLE, A., HALL, P. and BATHIA, N. (2012). Componentwise classification and clustering of functional data. *Biometrika* **99** 299–313. MR2931255 https://doi.org/10.1093/biomet/ass003
[25] DONNET, S., RIVOIRARD, V. and ROUSSEAU, J. (2020). Nonparametric Bayesian estimation for multivariate Hawkes processes. *Ann. Statist*. **48** 2698–2727. MR4152118 https://doi.org/10.1214/19-AOS1903
[26] EMBRECHTS, P., LINIGER, T. and LIN, L. (2011). Multivariate Hawkes processes: An application to financial data. *J. Appl. Probab*. **48A** 367–378. MR2865638 https://doi.org/10.1239/jap/1318940477
[27] ESCABIAS, M., AGUILERA, A. M. and VALDERRAMA, M. J. (2004). Principal component estimation of functional logistic regression: Discussion of two different approaches. *J. Nonparametr. Stat*. **16** 365–384. MR2073031 https://doi.org/10.1080/10485250310001624738
[28] FARAJTABAR, M., WANG, Y., GOMEZ-RODRIGUEZ, M., LI, S., ZHA, H. and SONG, L. (2017). COEVOLVE: A joint point process model for information diffusion and network evolution. *J. Mach. Learn. Res*. **18** Paper No. 41, 49. MR3655306
[29] GAO, X. and ZHU, L. (2018). Functional central limit theorems for stationary Hawkes processes and application to infinite-server queues. *Queueing Syst*. **90** 161–206. MR3850052 https://doi.org/10.1007/s11134-018-9570-5
[30] GUSTO, G. and SCHBATH, S. (2005). FADO: A statistical method to detect favored or avoided distances between occurrences of motifs using the Hawkes' model. *Stat. Appl. Genet. Mol. Biol*. **4** Art. 24, 28. MR2170440 https://doi.org/10.2202/1544-6115.1119
[31] HANSEN, N. R., REYNAUD-BOURET, P. and RIVOIRARD, V. (2015). Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli* **21** 83–143. MR3322314 https://doi.org/10.3150/13-BEJ562
[32] HAPP, C. and GREVEN, S. (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *J. Amer. Statist. Assoc*. **113** 649–659. MR3832216 https://doi.org/10.1080/01621459.2016.1273115
[33] HAWKES, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika* **58** 83–90. MR0278410 https://doi.org/10.1093/biomet/58.1.83
[34] HAWKES, A. G. (1971). Point spectra of some mutually exciting point processes. *J. Roy. Statist. Soc. Ser. B, Methodol*. **33** 438–443. MR0358976

[35] HILGERT, N., MAS, A. and VERZELEN, N. (2013). Minimax adaptive tests for the functional linear model. *Ann. Statist.* **41** 838–869. MR3099123 https://doi.org/10.1214/13-AOS1093

[36] HSING, T. and EUBANK, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators. Wiley Series in Probability and Statistics*. Wiley, Chichester. MR3379106 https://doi.org/10.1002/9781118762547

[37] ILLIAN, J., BENSON, E., CRAWFORD, J. and STAINES, H. (2006). Principal component analysis for spatial point processes—assessing the appropriateness of the approach in an ecological context. In *Case Studies in Spatial Point Process Modeling. Lect. Notes Stat.* **185** 135–150. Springer, New York. MR2232127 https://doi.org/10.1007/0-387-31144-0_7

[38] JACQUES, J. and PREDA, C. (2014). Model-based clustering for multivariate functional data. *Comput. Statist. Data Anal.* **71** 92–106. MR3131956 https://doi.org/10.1016/j.csda.2012.12.004

[39] KARASÖZEN, E., NISSEN, E., BÜYÜKAKPINAR, P., CAMBAZ, M. D., KAHRAMAN, M., KALKAN ERTAN, E., ABGARMI, B., BERGMAN, E., GHODS, A. et al. (2018). The 2017 July 20 Mw 6.6 Bodrum–Kos earthquake illuminates active faulting in the Gulf of Gökova, SW Turkey. *Geophys. J. Int.* **214** 185–199.

[40] KARR, A. F. (1991). *Point Processes and Their Statistical Inference*, 2nd ed. *Probability: Pure and Applied* **7**. Dekker, New York. MR1113698

[41] KINGMAN, J. F. C. (1993). *Poisson Processes. Oxford Studies in Probability* **3**. The Clarendon Press, New York. MR1207584

[42] KOLACZYK, E. D. (1999). Wavelet shrinkage estimation of certain Poisson intensity signals using corrected thresholds. *Statist. Sinica* **9** 119–135. MR1678884

[43] LAMBERT, R., TULEAU-MALOT, C., BESSAIH, T., RIVOIRARD, V., BOURET, Y., LERESCHE, N. and REYNAUD-BOURET, P. (2017). Reconstructing the functional connectivity of multiple spike trains using Hawkes models. *J. Neurosci. Methods* **297**. https://doi.org/10.1016/j.jneumeth.2017.12.026

[44] LI, Y. and GUAN, Y. (2014). Functional principal component analysis of spatiotemporal point processes with applications in disease surveillance. *J. Amer. Statist. Assoc.* **109** 1205–1215. MR3265691 https://doi.org/10.1080/01621459.2014.885434

[45] LI, Y., WANG, N. and CARROLL, R. J. (2013). Selecting the number of principal components in functional data. *J. Amer. Statist. Assoc.* **108** 1284–1294. MR3174708 https://doi.org/10.1080/01621459.2013.788980

[46] MANTÉ, C., YAO, A.-F. and DEGIOVANNI, C. (2007). Principal component analysis of measures, with special emphasis on grain-size curves. *Comput. Statist. Data Anal.* **51** 4969–4983. MR2364553 https://doi.org/10.1016/j.csda.2006.08.003

[47] MARSOLIER, J., PROMPSY, P. and VALLOT, C. (2022). H3K27me3 conditions chemotolerance in triple-negative breast cancer. *Nat. Genet.* **54** 459–468.

[48] MEERS, M. P., LLAGAS, G., JANSSENS, D. H., CODOMO, C. A. and HENIKOFF, S. (2023). Multifactorial profiling of epigenetic landscapes at single-cell resolution using MulTI-Tag. *Nat. Biotechnol.* **41** 708–716.

[49] MOHLER, G. O., SHORT, M. B., BRANTINGHAM, P. J., SCHOENBERG, F. P. and TITA, G. E. (2011). Self-exciting point process modeling of crime. *J. Amer. Statist. Assoc.* **106** 100–108. MR2816705 https://doi.org/10.1198/jasa.2011.ap09546

[50] OAKES, D. (1975). The Markovian self-exciting process. *J. Appl. Probab.* **12** 69–77. MR0362522 https://doi.org/10.1017/s0021900200033106

[51] OGATA, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *J. Amer. Statist. Assoc.* **83** 9–27.

[52] PANARETOS, V. M. and ZEMEL, Y. (2016). Amplitude and phase variation of point processes. *Ann. Statist.* **44** 771–812. MR3476617 https://doi.org/10.1214/15-AOS1387

[53] PICARD, F., RIVOIRARD, V., ROCHE, A. and PANARETOS, V. (2026). Supplement to "PCA for point processes." https://doi.org/10.1214/25-AOS2596SUPP

[54] RAMSAY, J. O. and SILVERMAN, B. W. (2002). *Applied Functional Data Analysis: Methods and Case Studies. Springer Series in Statistics*. Springer, New York. MR1910407 https://doi.org/10.1007/b98886

[55] RASMUSSEN, J. G. (2013). Bayesian inference for Hawkes processes. *Methodol. Comput. Appl. Probab.* **15** 623–642. MR3085883 https://doi.org/10.1007/s11009-011-9272-5

[56] REISS, P. T. and OGDEN, R. T. (2007). Functional principal component regression and functional partial least squares. *J. Amer. Statist. Assoc.* **102** 984–996. MR2411660 https://doi.org/10.1198/016214507000000527

[57] REYNAUD-BOURET, P. (2003). Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. *Probab. Theory Related Fields* **126** 103–153. MR1981635 https://doi.org/10.1007/s00440-003-0259-1

[58] REYNAUD-BOURET, P. and SCHBATH, S. (2010). Adaptive estimation for Hawkes processes; application to genome analysis. *Ann. Statist.* **38** 2781–2822. MR2722456 https://doi.org/10.1214/10-AOS806

[59] SULEM, D., RIVOIRARD, V. and ROUSSEAU, J. (2024). Bayesian estimation of nonlinear Hawkes processes. *Bernoulli* **30** 1257–1286. MR4699552 https://doi.org/10.3150/23-bej1631

[60] WARD, O. G., WU, J., ZHENG, T., SMITH, A. L. and CURLEY, J. P. (2022). Network Hawkes process models for exploring latent hierarchy in social animal interactions. *J. R. Stat. Soc., Ser. C, Appl. Stat.* **71** 1402–1426. MR4511116 https://doi.org/10.1111/rssc.12581

[61] WILLETT, R. M. and NOWAK, R. D. (2007). Multiscale Poisson intensity and density estimation. *IEEE Trans. Inf. Theory* **53** 3171–3187. MR2417680 https://doi.org/10.1109/TIT.2007.903139

[62] WU, S., MÜLLER, H.-G. and ZHANG, Z. (2013). Functional data analysis for point processes with rare events. *Statist. Sinica* **23** 1–23. MR3076156

[63] ZETTL, A. (2005). *Sturm-Liouville Theory*. *Mathematical Surveys and Monographs* **121**. Amer. Math. Soc., Providence, RI. MR2170950 https://doi.org/10.1090/surv/121

[64] ZHENG, G. X. Y., TERRY, J. M. and BIELAS, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8** 14049.

The Annals of Statistics
2026, Vol. 54, No. 2, 957–981
https://doi.org/10.1214/25-AOS2597

PARAMETER IDENTIFICATION IN LINEAR NON-GAUSSIAN CAUSAL MODELS UNDER GENERAL CONFOUNDING

BY DANIELE TRAMONTANO[a], MATHIAS DRTON[b] AND JALAL ETESAMI[c]

School of Computation, Information and Technology, Technical University of Munich, [a]*daniele.tramontano@tum.de,* [b]*mathias.drton@tum.de,* [c]*j.etesami@tum.de*

Linear non-Gaussian causal models postulate that each random variable is a linear function of parent variables and non-Gaussian exogenous error terms. We study identification of the linear coefficients when such models contain latent variables. Our focus is on the commonly studied acyclic setting, where each model corresponds to a directed acyclic graph (DAG). For this case, prior literature has demonstrated that connections to overcomplete independent component analysis yield effective criteria to decide parameter identifiability in latent variable models. However, this connection is based on the assumption that the observed variables linearly depend on the latent variables. Departing from this assumption, we treat models that allow for arbitrary nonlinear latent confounding. Our main result is a graphical criterion that is necessary and sufficient for deciding the generic identifiability of direct causal effects. Moreover, we provide an algorithmic implementation of the criterion with a run time that is polynomial in the number of observed variables. Finally, we report on estimation heuristics based on the identification result and explore a generalization to models with feedback loops.

REFERENCES

BARBER, R. F., DRTON, M., STURMA, N. and WEIHS, L. (2022). Half-trek criterion for identifiability of latent variable models. *Ann. Statist.* **50** 3174–3196. MR4524493 https://doi.org/10.1214/22-aos2221

BRITO, C. (2004). Graphical Models for Identification in Structural Equation Models. Ph.D. thesis, UCLA Computer Science Dept.

CHEN, L., KYNG, R., LIU, Y. P., PENG, R., GUTENBERG, M. P. and SACHDEVA, S. (2022). Maximum flow and minimum-cost flow in almost-linear time. In 2022 *IEEE* 63*rd Annual Symposium on Foundations of Computer Science—FOCS* 2022, *Denver, CO, USA, October* 31–*November* 3, 2022 612–623. IEEE Comput. Soc., Los Alamitos, CA. MR4537240

COMON, P. and JUTTEN, C. (2010). *Handbook of Blind Source Separation*: *Independent Component Analysis and Applications*, 1st ed. Academic Press, USA.

CORMEN, T. H., LEISERSON, C. E., RIVEST, R. L. and STEIN, C. (2009). *Introduction to Algorithms*, 3rd ed. MIT Press, Cambridge, MA. MR2572804

COVER, T. M. and THOMAS, J. A. (1991). *Elements of Information Theory*. *Wiley Series in Telecommunications*. Wiley, New York. A Wiley-Interscience Publication. MR1122806 https://doi.org/10.1002/0471200611

DINITS, E. A. (1970). Algorithm for solution of a problem of maximum flow in a network with power estimation. *Sov. Math., Dokl.* **11** 1277–1280.

DRTON, M. (2018). Algebraic problems in structural equation modeling. In *The* 50*th Anniversary of Gröbner Bases*. *Adv. Stud. Pure Math.* **77** 35–86. Math. Soc. Japan, Tokyo. MR3839705 https://doi.org/10.2969/aspm/07710035

DRTON, M., FOYGEL, R. and SULLIVANT, S. (2011). Global identifiability of linear structural equation models. *Ann. Statist.* **39** 865–886. MR2816341 https://doi.org/10.1214/10-AOS859

DRTON, M. and MAATHUIS, M. H. (2017). Structure learning in graphical modeling. *Annu. Rev. Stat. Appl.* **4** 365–393.

DRTON, M. and RICHARDSON, T. S. (2008). Binary models for marginal independence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 287–309. MR2424754 https://doi.org/10.1111/j.1467-9868.2007.00636.x

MSC2020 subject classifications. Primary 62H22, 62A09, 62J05; secondary 62R01.

Key words and phrases. Causal effect, graphical model, independent component analysis, latent variable model, structural causal model.

ERIKSSON, J. and KOIVUNEN, V. (2004). Identifiability, separability, and uniqueness of linear ICA models. *IEEE Signal Process. Lett.* **11** 601–604.

EVANS, W. N. and RINGEL, J. S. (1999). Can higher cigarette taxes improve birth outcomes? *J. Public Econ.* **72** 135–154.

FOYGEL, R., DRAISMA, J. and DRTON, M. (2012). Half-trek criterion for generic identifiability of linear structural equation models. *Ann. Statist.* **40** 1682–1713. MR3015040 https://doi.org/10.1214/12-AOS1012

FUKUMIZU, K., GRETTON, A., SUN, X. and SCHÖLKOPF, B. (2007). Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems* 20. *Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems*, *Vancouver*, *British Columbia*, *Canada*, *December* 3–6, 2007 489–496. Curran Associates, Red Hook.

GARROTE-LÓPEZ, M. and STEPHENSON, M. (2024). Cumulant Tensors in Partitioned Independent Component Analysis. Available at arXiv:2402.10089.

GEENENS, G. and LAFAYE DE MICHEAUX, P. (2022). The Hellinger correlation. *J. Amer. Statist. Assoc.* **117** 639–653. MR4436302 https://doi.org/10.1080/01621459.2020.1791132

GRETTON, A., FUKUMIZU, K., TEO, C. H., SONG, L., SCHÖLKOPF, B. and SMOLA, A. J. (2007). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems* 20. *Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems*, *Vancouver*, *British Columbia*, *Canada*, *December* 3–6, 2007 585–592. Curran Associates, Red Hook.

IMBENS, G. W. and NEWEY, W. K. (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica* **77** 1481–1512. MR2561069 https://doi.org/10.3982/ECTA7108

KIVVA, Y., ETESAMI, J. and KIYAVASH, N. (2023). On identifiability of conditional causal effects. In *Uncertainty in Artificial Intelligence*, *UAI* 2023. *Proceedings of Machine Learning Research* **216** 1078–1086. PMLR, Pittsburgh, PA, USA.

KIVVA, Y., MOKHTARIAN, E., ETESAMI, J. and KIYAVASH, N. (2022). Revisiting the general identifiability problem. In *Uncertainty in Artificial Intelligence*, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, *UAI* 2022. *Proceedings of Machine Learning Research* **180** 1022–1030. PMLR, Eindhoven, The Netherlands.

KIVVA, Y., SALEHKALEYBAR, S. and KIYAVASH, N. (2023). A cross-moment approach for causal effect estimation. In *Advances in Neural Information Processing Systems* 36: *Annual Conference on Neural Information Processing Systems* 2023, *NeurIPS* 2023, *New Orleans*, *LA*, *USA*, 2023.

KRASKOV, A., STÖGBAUER, H. and GRASSBERGER, P. (2004). Estimating mutual information. *Phys. Rev. E* **69** 066138, 16. MR2096503 https://doi.org/10.1103/PhysRevE.69.066138

KUMOR, D., CINELLI, C. and BAREINBOIM, E. (2020). Efficient identification in linear structural causal models with auxiliary cutsets. In *Proceedings of the* 37*th International Conference on Machine Learning*, *ICML* 2020, 13–18 *July* 2020, *Virtual Event*. *Proceedings of Machine Learning Research* **119** 5501–5510. PMLR.

LAURITZEN, S. L. (1996). *Graph. Models. Oxford Statistical Science Series* **17**. Oxford Univ. Press, New York. Oxford Science Publications. MR1419991

LEE, S., CORREA, J. D. and BAREINBOIM, E. (2020). General identifiability with arbitrary surrogate experiments. In *Proceedings of the* 35*th Uncertainty in Artificial Intelligence Conference* (R. P. Adams and V. Gogate, eds.). *Proceedings of Machine Learning Research* **115** 389–398. PMLR.

LEWICKI, M. S. and SEJNOWSKI, T. J. (2000). Learning overcomplete representations. *Neural Comput.* **12** 337–365.

LIU, D. C. and NOCEDAL, J. (1989). On the limited memory BFGS method for large scale optimization. *Math. Program.* **45** 503–528. MR1038245 https://doi.org/10.1007/BF01589116

LIU, Y., ROBEVA, E. and WANG, H. (2021). Learning linear non-Gaussian graphical models with multidirected edges. *J. Causal Inference* **9** 250–263. MR4315952 https://doi.org/10.1515/jci-2020-0027

LONDSCHIEN, M. and BÜHLMANN, P. (2024). Weak-instrument-robust subvector inference in instrumental variables regression: a subvector Lagrange multiplier test and properties of subvector Anderson–Rubin confidence sets. Available at arXiv:2407.15256.

MAATHUIS, M., DRTON, M., LAURITZEN, S. and WAINWRIGHT, M., eds. (2019). *Handbook of Graphical Models*. *Chapman & Hall/CRC Handbooks of Modern Statistical Methods*. CRC Press, Boca Raton, FL. MR3889064

MESTERS, G. and ZWIERNIK, P. (2024). Nonndependent components analysis. *Ann. Statist.* **52** 2506–2528. MR4842816 https://doi.org/10.1214/24-aos2373

MOOIJ, J. M., JANZING, D., PETERS, J. and SCHÖLKOPF, B. (2009). Regression by dependence minimization and its application to causal inference in additive noise models. In *Proceedings of the* 26*th Annual International Conference on Machine Learning*, *ICML* 2009, *Montreal*, *Quebec*, *Canada*, 2009. *ACM International Conference Proceeding Series* **382** 745–752. ACM, New York.

OKAMOTO, M. (1973). Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *Ann. Statist.* **1** 763–765. MR0331643

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

PEARL, J. (2017). A linear 'microscope' for interventions and counterfactuals. *J. Causal Inference* **5** Art. No. 20170003. MR4323814 https://doi.org/10.1515/jci-2017-0003

PETERS, J., JANZING, D. and SCHÖLKOPF, B. (2017). *Elements of Causal Inference*: *Foundations and Learning Algorithms*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR3822088

RIBOT, A., SEIGAL, A. and ZWIERNIK, P. (2025). Beyond independent component analysis: Identifiability and algorithms. Available at arXiv:2510.07525.

RICHARDSON, T. (2003). Markov properties for acyclic directed mixed graphs. *Scand. J. Stat.* **30** 145–157. MR1963898 https://doi.org/10.1111/1467-9469.00323

RICHARDSON, T. and SPIRTES, P. (2002). Ancestral graph Markov models. *Ann. Statist.* **30** 962–1030. MR1926166 https://doi.org/10.1214/aos/1031689015

RICHARDSON, T. S., EVANS, R. J., ROBINS, J. M. and SHPITSER, I. (2023). Nested Markov properties for acyclic directed mixed graphs. *Ann. Statist.* **51** 334–361. MR4564859 https://doi.org/10.1214/22-aos2253

SAENGKYONGAM, S., HENCKEL, L., PFISTER, N. and PETERS, J. (2022). Exploiting independent instruments: Identification and distribution generalization. In *International Conference on Machine Learning*, *ICML* 2022, 17–23 *July* 2022, *Baltimore*, *Maryland*, *USA*. *Proceedings of Machine Learning Research* **162** 18935–18958. PMLR.

SALEHKALEYBAR, S., GHASSAMI, A., KIYAVASH, N. and ZHANG, K. (2020). Learning linear non-Gaussian causal models in the presence of latent variables. *J. Mach. Learn. Res.* **21** Paper No. 39, 24. MR4073772

SCHKODA, D. and DRTON, M. (2025). Goodness-of-fit tests for linear non-Gaussian structural equation models. *Biometrika* **112** Paper No. asaf046, 16. MR5002277 https://doi.org/10.1093/biomet/asaf046

SCHÖLKOPF, B. and SMOLA, A. J. (2018). *Learning with Kernels*: *Support Vector Machines*, *Regularization*, *Optimization*, *and Beyond*. MIT Press, Cambridge.

SHI, H., DRTON, M. and HAN, F. (2022). On the power of Chatterjee's rank correlation. *Biometrika* **109** 317–333. MR4430960 https://doi.org/10.1093/biomet/asab028

SHIMIZU, S. (2022). *Statistical Causal Discovery*: *LiNGAM Approach*. *SpringerBriefs in Statistics*. Springer, Tokyo. MR4501962

SHPITSER, I. (2023). When does the ID algorithm fail? Available at arXiv:2307.03750.

SHPITSER, I. and PEARL, J. (2006). Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the* 21*st National Conference on Artificial Intelligence*, *Vol.* 2 *AAAI'*06 1219–1226. AAAI Press, Menlo Park.

SHUAI, K., LUO, S., ZHANG, Y., XIE, F. and HE, Y. (2023). Identification and Estimation of Causal Effects Using non-Gaussianity and Auxiliary Covariates. Available at arXiv:2304.14895.

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation*, *Prediction*, *and Search*, 2nd ed. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. With additional material by David Heckerman, Christopher Meek, Gregory F. Cooper and Thomas Richardson. A Bradford Book. MR1815675

STURMA, N. and DRTON, M. (2025). Trek-Based Parameter Identification for Linear Causal Models with Arbitrarily Structured Latent Variables. Available at arXiv:2507.18170.

SULLIVANT, S., TALASKA, K. and DRAISMA, J. (2010). Trek separation for Gaussian graphical models. *Ann. Statist.* **38** 1665–1685. MR2662356 https://doi.org/10.1214/09-AOS760

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

TRAMONTANO, D., DRTON, M. and ETESAMI, J. (2026). Supplement to "Parameter identification in linear non-Gaussian causal models under general confounding." https://doi.org/10.1214/25-AOS2597SUPPA, https://doi.org/10.1214/25-AOS2597SUPPB

TRAMONTANO, D., KIVVA, Y., SALEHKALEYBAR, S., DRTON, M. and KIYAVASH, N. (2024). Causal effect identification in LiNGAM models with latent confounders. In *Forty-First International Conference on Machine Learning*, *ICML* 2024. PMLR.

TRAMONTANO, D., KIVVA, Y., SALEHKALEYBAR, S., KIYAVASH, N. and DRTON, M. (2025). Causal effect identification in lvLiNGAM from higher-order cumulants. In *Forty-Second International Conference on Machine Learning*, *ICML* 2025. PMLR.

VERMA, T. and PEARL, J. (1990). Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*. *UAI'*90 255–270. Elsevier, USA.

WANG, K. and SEIGAL, A. (2024). Identifiability of overcomplete independent component analysis. Available at arXiv:2401.14709.

WANG, Y. S. and DRTON, M. (2017). Empirical likelihood for linear structural equation models with dependent errors. *Stat* **6** 434–447. MR3722940 https://doi.org/10.1002/sta4.169

WANG, Y. S. and DRTON, M. (2023). Causal discovery with unobserved confounding and non-Gaussian data. *J. Mach. Learn. Res.* **24** Paper No. [271], 61. MR4664708

ZHAO, Q. (2025). On statistical and causal models associated with acyclic directed mixed graphs. Available at arXiv:2501.03048.

The Annals of Statistics
2026, Vol. 54, No. 2, 982–1005
https://doi.org/10.1214/25-AOS2598

GENERALIZED MULTILINEAR MODELS FOR SUFFICIENT DIMENSION REDUCTION ON TENSOR-VALUED PREDICTORS

BY DANIEL KAPLA[a] AND EFSTATHIA BURA[b]

[1] *Institute of Statistics and Mathematical Methods in Economics, Faculty of Mathematics and Geoinformation, TU Wien,*
[a] *daniel.kapla@tuwien.ac.at,* [b] *efstathia.bura@tuwien.ac.at*

We consider supervised learning problems with tensor-valued input. We derive multilinear sufficient reductions for the regression or classification problem by modeling the conditional distribution of the predictors given the response as a member of the quadratic exponential family. We develop estimation procedures of sufficient reductions for both continuous and binary tensor-valued predictors. We prove the consistency and asymptotic normality of the estimated sufficient reduction using manifold theory. For multilinear normal predictors, the estimation algorithm is highly computationally efficient and is also applicable to situations where the dimension of the reduction exceeds the sample size. Our method outperforms competing techniques in both simulated settings and real-world datasets involving continuous and binary tensor-valued predictors.

REFERENCES

ABADIR, K. M. and MAGNUS, J. R. (2005). *Matrix Algebra*. *Econometric Exercises* **1**. Cambridge Univ. Press, Cambridge. MR2408356 https://doi.org/10.1017/CBO9780511810800

ABSIL, P.-A., MAHONY, R. and SEPULCHRE, R. (2008). *Optimization Algorithms on Matrix Manifolds*. Princeton Univ. Press, Princeton, NJ. Available at https://press.princeton.edu/absil. MR2364186 https://doi.org/10.1515/9781400830244

ARNOLD, S. F. (1981). *The Theory of Linear Models and Multivariate Analysis*. *Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. MR0606011

BURA, E., FORZANI, L., GARCÍA ARANCIBIA, R., LLOP, P. and TOMASSI, D. (2022). Sufficient reductions in regression with mixed predictors. *J. Mach. Learn. Res.* **23** 102. MR4576687

BURDICK, D. S. (1995). An introduction to tensor products with applications to multiway data analysis. *Chemom. Intell. Lab. Syst.* **28** 229–237. https://doi.org/10.1016/0169-7439(95)80060-M

CHEN, Y.-L., KOLAR, M. and TSAY, R. S. (2021). Tensor canonical correlation analysis with convergence and statistical guarantees. *J. Comput. Graph. Statist.* **30** 728–744. MR4313472 https://doi.org/10.1080/10618600.2020.1856118

CHENG, J., LEVINA, E., WANG, P. and ZHU, J. (2014). A sparse Ising model with covariates. *Biometrics* **70** 943–953. MR3295755 https://doi.org/10.1111/biom.12202

COMON, P. (2009). Tensors versus matrices usefulness and unexpected properties. In 2009 *IEEE/SP* 15*th Workshop on Statistical Signal Processing* 781–788. https://doi.org/10.1109/SSP.2009.5278471

COOK, R. D. (1998). *Regression Graphics*. *Wiley Series in Probability and Statistics*: *Probability and Statistics*. Wiley, New York. Ideas for studying regressions through graphics, a Wiley-Interscience Publication. MR1645673 https://doi.org/10.1002/9780470316931

COOK, R. D. (2007). Fisher lecture: Dimension reduction in regression. *Statist. Sci.* **22** 1–26. MR2408655 https://doi.org/10.1214/088342306000000682

COOK, R. D., LI, B. and CHIAROMONTE, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statist. Sinica* **20** 927–960. MR2729839

COX, D. R. and WERMUTH, N. (1994). A note on the quadratic exponential binary distribution. *Biometrika* **81** 403–408. MR1294901 https://doi.org/10.1093/biomet/81.2.403

DAWID, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika* **68** 265–274. MR0614963 https://doi.org/10.1093/biomet/68.1.265

MSC2020 subject classifications. Primary 62E20, 62J05, 62F12; secondary 62F30, 62B05, 15A69.

Key words and phrases. Regression, asymptotics, manifold theory, constrained optimization, maximum likelihood estimation.

DE ALMEIDA, A. L. F., FAVIER, G. and MOTA, J. C. M. (2007). PARAFAC-based unified tensor modeling for wireless communication systems with application to blind multiuser equalization. *Signal Process*. **87** 337–351. Tensor Signal Processing. https://doi.org/10.1016/j.sigpro.2005.12.014

DE LATHAUWER, L. and CASTAING, J. (2007). Tensor-based techniques for the blind separation of DS-CDMA signals. *Signal Process*. **87** 322–336. Tensor Signal Processing. 10.1016/j.sigpro.2005.12.015.

DING, S. and COOK, R. D. (2014). Dimension folding PCA and PFC for matrix-valued predictors. *Statist*. *Sinica* **24** 463–492. MR3183694

DING, S. and COOK, R. D. (2015). Tensor sliced inverse regression. *J*. *Multivariate Anal*. **133** 216–231. MR3282027 https://doi.org/10.1016/j.jmva.2014.08.015

DRTON, M., KURIKI, S. and HOFF, P. (2021). Existence and uniqueness of the Kronecker covariance MLE. *Ann*. *Statist*. **49** 2721–2754. MR4338381 https://doi.org/10.1214/21-aos2052

DUTILLEUL, P. (1999). The mle algorithm for the matrix normal distribution. *J*. *Stat*. *Comput*. *Simul*. **64** 105–123. https://doi.org/10.1080/00949659908811970

GIRKA, F., GLOAGUEN, A., LE BRUSQUET, L., ZUJOVIC, V. and TENENHAUS, A. (2024). Tensor generalized canonical correlation analysis. *Inf*. *Fusion* **102**. https://doi.org/10.1016/j.inffus.2023.102045

HAO, B., WANG, B., WANG, P., ZHANG, J., YANG, J. and SUN, W. W. (2021). Sparse tensor additive regression. *J*. *Mach*. *Learn*. *Res*. **22** 64. MR4253757

HARVILLE, D. A. (1997). *Matrix Algebra from a Statistician's Perspective*. Springer, New York. MR1467237 https://doi.org/10.1007/b98818

HILLAR, C. J. and LIM, L.-H. (2013). Most tensor problems are NP-hard. *J*. *ACM* **60** 45. MR3144915 https://doi.org/10.1145/2512329

HOFF, P. D. (2011). Separable covariance arrays via the Tucker product, with applications to multivariate relational data. *Bayesian Anal*. **6** 179–196. MR2806238 https://doi.org/10.1214/11-BA606

HOFF, P. D. (2015). Multilinear tensor regression for longitudinal relational data. *Ann*. *Appl*. *Stat*. **9** 1169–1193. MR3418719 https://doi.org/10.1214/15-AOAS839

ISING, E. (1925). Beitrag zur Theorie des Ferromagnetismus. *Z*. *Phys*. **31** 253–258. https://doi.org/10.1007/BF02980577

JOHNSON, N. L., KOTZ, S. and BALAKRISHNAN, N. (1997). *Discrete Multivariate Distributions*. *Wiley Series in Probability and Statistics*: *Applied Probability and Statistics*. Wiley, New York. A Wiley-Interscience Publication. MR1429617

JUNG, S., AHN, J. and JEON, Y. (2019). Penalized orthogonal iteration for sparse estimation of generalized eigenvalue problem. *J*. *Comput*. *Graph*. *Statist*. **28** 710–721. MR4007752 https://doi.org/10.1080/10618600.2019.1568014

KALTENBÄCK, M. (2021). *Aufbau Analysis*, 27th ed. *Berliner Studienreihe zur Mathematik*. Heldermann Verlag.

KAPLA, D. and BURA, E. (2026). Supplement to "Generalized multilinear models for sufficient dimension reduction on tensor-valued predictors." https://doi.org/10.1214/25-AOS2598SUPP

KOFIDIS, E. and REGALIA, P. A. (2001). Tensor approximation and signal processing applications. In *Structured Matrices in Mathematics*, *Computer Science*, *and Engineering*, *I* (*Boulder*, *CO*, 1999). *Contemp*. *Math*. **280** 103–133. Amer. Math. Soc., Providence, RI. MR1850404 https://doi.org/10.1090/conm/280/04625

KOLDA, T. G. (2006). Multilinear operators for higher-order decompositions. https://doi.org/10.2172/923081

KOLDA, T. G. and BADER, B. W. (2009). Tensor decompositions and applications. *SIAM Rev*. **51** 455–500. MR2535056 https://doi.org/10.1137/07070111X

KOLLO, T. and VON ROSEN, D. (2005). *Advanced Multivariate Statistics with Matrices*. *Mathematics and Its Applications* (*New York*) **579**. Springer, Dordrecht. MR2162145 https://doi.org/10.1007/1-4020-3419-9

LANDGRAF, A. J. and LEE, Y. (2020). Dimensionality reduction for binary data through the projection of natural parameters. *J*. *Multivariate Anal*. **180** 104668. MR4147633 https://doi.org/10.1016/j.jmva.2020.104668

LEE, J. M. (2012). *Introduction to Smooth Manifolds*. Springer, New York. https://doi.org/10.1007/978-1-4419-9982-5

LEE, J. M. (2018). *Introduction to Riemannian Manifolds*. *Graduate Texts in Mathematics* **176**. Springer, Cham. MR3887684

LENZ, W. (1920). Beitrag zum Verständnis der magnetischen Erscheinungen in festen Körpern. *Eur*. *Phys*. *J*. *A* **21** 613–615.

LEURGANS, S. and ROSS, R. T. (1992). Multilinear models: Applications in spectroscopy. *Statist*. *Sci*. **7** 289–319. MR1181414

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

LI, B., KIM, M. K. and ALTMAN, N. (2010). On dimension folding of matrix- or array-valued statistical objects. *Ann*. *Statist*. **38** 1094–1121. MR2604706 https://doi.org/10.1214/09-AOS737

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

LI, L. and ZHANG, X. (2017). Parsimonious tensor response regression. *J. Amer. Statist. Assoc.* **112** 1131–1146. MR3735365 https://doi.org/10.1080/01621459.2016.1193022

LLOSA-VITE, C. and MAITRA, R. (2023). Reduced-rank tensor-on-tensor regression and tensor-variate analysis of variance. *IEEE Trans. Pattern Anal. Mach. Intell.* **45** 2282–2296. https://doi.org/10.1109/tpami.2022.3164836

LOCK, E. F. (2018). Tensor-on-tensor regression. *J. Comput. Graph. Statist.* **27** 638–647. MR3863764 https://doi.org/10.1080/10618600.2017.1401544

LU, N. and ZIMMERMAN, D. L. (2005). The likelihood ratio test for a separable covariance matrix. *Statist. Probab. Lett.* **73** 449–457. MR2187860 https://doi.org/10.1016/j.spl.2005.04.020

MANCEUR, A. M. and DUTILLEUL, P. (2013). Maximum likelihood estimation for the tensor normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion. *J. Comput. Appl. Math.* **239** 37–49. MR2991957 https://doi.org/10.1016/j.cam.2012.09.017

MARDIA, K. V. and GOODALL, C. R. (1993). Spatial-temporal analysis of multivariate environmental monitoring data. In *Multivariate Environmental Statistics*. *North-Holland Ser. Statist. Probab.* **6** 347–386. North-Holland, Amsterdam. MR1268443

NISS, M. (2005). History of the Lenz-Ising model 1920–1950: From ferromagnetic to cooperative phenomena. *Arch. Hist. Exact Sci.* **59** 267–318. MR2124728 https://doi.org/10.1007/s00407-004-0088-3

OHLSON, M., AHMAD, R. M. and VON ROSEN, D. (2013). The multilinear normal distribution: Introduction and some basic properties. *J. Multivariate Anal.* **113** 37–47. MR2984354 https://doi.org/10.1016/j.jmva.2011.05.015

PEPE, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. *Oxford Statistical Science Series* **28**. Oxford Univ. Press, Oxford. MR2260483

PFEIFFER, R., KAPLA, D. and BURA, E. (2021). Least squares and maximum likelihood estimation of sufficient reductions in regressions with matrix-valued predictors. *Int. J. Data Sci. Anal.* **11**. https://doi.org/10.1007/s41060-020-00228-y

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

RABUSSEAU, G. and KADRI, H. (2016). Low-rank regression with tensor responses. In *Advances in Neural Information Processing Systems* (D. Lee, M. Sugiyama, U. Luxburg, I. Guyon and R. Garnett, eds.) **29**. Curran Associates.

SUN, W. W. and LI, L. (2017). STORE: Sparse tensor response regression and neuroimaging analysis. *J. Mach. Learn. Res.* **18** 135. MR3763769

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

WANG, Y., SUN, Z., SONG, D. and HERO, A. (2022). Kronecker-structured covariance models for multiway data. *Stat. Surv.* **16** 238–270. MR4522372 https://doi.org/10.1214/22-ss139

XU, Y. and MUKHERJEE, S. (2023). Inference in Ising models on dense regular graphs. *Ann. Statist.* **51** 1183–1206. MR4630945 https://doi.org/10.1214/23-aos2286

ZHANG, D. and ZHOU, Z.-H. (2005). (2D)2PCA: Two-directional two-dimensional PCA for efficient face representation and recognition. *Neurocomputing* **69** 224–231. Neural Networks in Signal Processing. https://doi.org/10.1016/j.neucom.2005.06.004

ZHOU, H., LI, L. and ZHU, H. (2013). Tensor regression with applications in neuroimaging data analysis. *J. Amer. Statist. Assoc.* **108** 540–552. MR3174640 https://doi.org/10.1080/01621459.2013.776499

ZHOU, J., SUN, W. W., ZHANG, J. and LI, L. (2023). Partially observed dynamic tensor response regression. *J. Amer. Statist. Assoc.* **118** 424–439. MR4571132 https://doi.org/10.1080/01621459.2021.1938082

The Annals of Statistics
2026, Vol. 54, No. 2, 1006–1027
https://doi.org/10.1214/25-AOS2599

THE DISTRIBUTIONALLY ROBUST PREDICTION ERROR OF THE √LASSO AND RELATED ESTIMATORS

BY JOSÉ LUIS MONTIEL OLEA[1,a], CYNTHIA RUSH[2,c], AMILCAR VELEZ[1,b] AND JOHANNES WIESEL[3,d]

[1]*Department of Economics, Cornell University,* [a]*montiel.olea@gmail.com,* [b]*amilcare@cornell.edu*
[2]*Department of Statistics, Columbia University,* [c]*cynthia.rush@columbia.edu*
[3]*Department of Mathematics, University of Copenhagen,* [d]*wiesel@math.ku.dk*

We study the classical problem of predicting an outcome variable, Y, using a linear combination of a d-dimensional covariate vector, $\mathbf{X}$. We are interested in linear predictors whose coefficients solve

$$\inf_{\boldsymbol{\beta}\in\mathbb{R}^d} \left(\mathbb{E}_{\mathbb{P}_n}\left[\left|Y-\mathbf{X}^\top\boldsymbol{\beta}\right|^r\right]\right)^{1/r} + \delta\rho(\boldsymbol{\beta}),$$

where $\delta > 0$ is a regularization parameter, $\rho : \mathbb{R}^d \to \mathbb{R}_+$ is a convex penalty function, $\mathbb{P}_n$ is the empirical distribution of the data and $r \geq 1$. Our main contribution is a new bound on the out-of-distribution prediction error of such estimators.

The new bound is obtained by combining three new sets of results. First, we provide conditions under which linear predictors based on these estimators solve a *distributionally robust optimization* problem: they minimize the worst-case prediction error over distributions that are close to each other in a type of *max-sliced Wasserstein metric*. Second, we provide a detailed finite-sample and asymptotic analysis of the statistical properties of the balls of distributions over which the worst-case prediction error is analyzed. Third, we present an oracle recommendation for the choice of regularization parameter, δ, that guarantees good out-of-distribution prediction error.

REFERENCES

[1] ADJAHO, C. and CHRISTENSEN, T. (2022). Externally valid treatment choice. arXiv preprint. Available at arXiv:2205.05561.

[2] AGARWAL, D., LI, L. and SMOLA, A. (2011). Linear-time estimators for propensity scores. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. *JMLR Workshop and Conference Proceedings*. 93–100.

[3] ANDREWS, I., FUDENBERG, D., LIANG, A. and WU, C. (2022). The transfer performance of economic models. arXiv preprint. Available at arXiv:2202.04796.

[4] BARTL, D. and MENDELSON, S. (2025). Structure preservation via the Wasserstein distance. *J. Funct. Anal.* **288** 110810. MR4851901 https://doi.org/10.1016/j.jfa.2024.110810

[5] BELLONI, A., CHERNOZHUKOV, V. and WANG, L. (2011). Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika* **98** 791–806. MR2860324 https://doi.org/10.1093/biomet/asr043

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

[7] BEN-DAVID, S., BLITZER, J., CRAMMER, K., KULESZA, A., PEREIRA, F. and VAUGHAN, J. W. (2010). A theory of learning from different domains. *Mach. Learn.* **79** 151–175. MR3108150 https://doi.org/10.1007/s10994-009-5152-4

[8] BERTSIMAS, D. and COPENHAVER, M. S. (2018). Characterization of the equivalence of robustification and regularization in linear and matrix regression. *European J. Oper. Res.* **270** 931–942. MR3814540 https://doi.org/10.1016/j.ejor.2017.03.051

MSC2020 subject classifications. 62J07, 62J20.

Key words and phrases. Distributional robust optimization, max-sliced Wasserstein, square-root LASSO, out-of-distribution prediction error.

[9] BIAU, G., CADRE, B. and PELLETIER, B. (2008). Exact rates in density support estimation. *J. Multivariate Anal.* **99** 2185–2207. MR2463383 https://doi.org/10.1016/j.jmva.2008.02.021

[10] BLANCHET, J., KANG, Y. and MURTHY, K. (2019). Robust Wasserstein profile inference and applications to machine learning. *J. Appl. Probab.* **56** 830–857. MR4015639 https://doi.org/10.1017/jpr.2019.49

[11] BLANCHET, J., KANG, Y., MURTHY, K. and ZHANG, F. (2019). Data-driven optimal transport cost selection for distributionally robust optimization. In 2019 *Winter Simulation Conference* (*WSC*) 3740–3751. IEEE.

[12] BLANCHET, J., KANG, Y., OLEA, J. L. M., NGUYEN, V. A. and ZHANG, X. (2020). Machine learning's dropout training is distributionally robust optimal. arXiv preprint. Available at arXiv:2009.06111.

[13] BLANCHET, J. and MURTHY, K. (2019). Quantifying distributional model risk via optimal transport. *Math. Oper. Res.* **44** 565–600. MR3959085 https://doi.org/10.1287/moor.2018.0936

[14] BOBKOV, S. and LEDOUX, M. (2019). *One-Dimensional Empirical Measures*, *Order Statistics*, *and Kantorovich Transport Distances* **261**. Amer. Math. Soc., Providence.

[15] BOISSARD, E. and LE GOUIC, T. (2014). On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. *Ann. Inst. Henri Poincaré Probab. Stat.* **50** 539–563. MR3189084 https://doi.org/10.1214/12-AIHP517

[16] BONNEEL, N., RABIN, J., PEYRÉ, G. and PFISTER, H. (2015). Sliced and Radon Wasserstein barycenters of measures. *J. Math. Imaging Vision* **51** 22–45. MR3300482 https://doi.org/10.1007/s10851-014-0506-3

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

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

[19] CANER, M. and ELIAZ, K. (2024). Should humans Lie to machines? The incentive compatibility of Lasso and GLM structured sparsity estimators. *J. Bus. Econom. Statist.* **42** 1379–1388. MR4799141 https://doi.org/10.1080/07350015.2024.2316102

[20] CHEN, X., MONFORT, M., LIU, A. and ZIEBART, B. D. (2016). Robust covariate shift regression. In *Artificial Intelligence and Statistics* 1270–1279. PMLR.

[21] CHERNOZHUKOV, V., CHETVERIKOV, D., KATO, K. and KOIKE, Y. (2023). High-dimensional data bootstrap. *Annu. Rev. Stat. Appl.* **10** 427–449. MR4567800 https://doi.org/10.1146/annurev-statistics-040120-022239

[22] CHETVERIKOV, D., LIAO, Z. and CHERNOZHUKOV, V. (2021). On cross-validated Lasso in high dimensions. *Ann. Statist.* **49** 1300–1317. MR4298865 https://doi.org/10.1214/20-aos2000

[23] CHIZAT, L., ROUSSILLON, P., LÉGER, F., VIALARD, F.-X. and PEYRÉ, G. (2020). Faster Wasserstein distance estimation with the Sinkhorn divergence. *Adv. Neural Inf. Process. Syst.* **33** 2257–2269.

[24] CHRISTENSEN, T. and CONNAULT, B. (2023). Counterfactual sensitivity and robustness. *Econometrica* **91** 263–298. MR4562952 https://doi.org/10.3982/ecta17232

[25] CHU, H. T. M., TOH, K.-C. and ZHANG, Y. (2022). On regularized square-root regression problems: Distributionally robust interpretation and fast computations. *J. Mach. Learn. Res.* **23** 308. MR4577747

[26] CUEVAS, A. and FRAIMAN, R. (1997). A plug-in approach to support estimation. *Ann. Statist.* **25** 2300–2312. MR1604449 https://doi.org/10.1214/aos/1030741073

[27] DEREICH, S., SCHEUTZOW, M. and SCHOTTSTEDT, R. (2013). Constructive quantization: Approximation by empirical measures. *Ann. Inst. Henri Poincaré Probab. Stat.* **49** 1183–1203. MR3127919 https://doi.org/10.1214/12-AIHP489

[28] DESHPANDE, I., HU, Y.-T., SUN, R., PYRROS, A., SIDDIQUI, N., KOYEJO, S., ZHAO, Z., FORSYTH, D. and SCHWING, A. G. (2019). Max-sliced Wasserstein distance and its use for GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 10648–10656.

[29] DUCHI, J. C., GLYNN, P. W. and NAMKOONG, H. (2021). Statistics of robust optimization: A generalized empirical likelihood approach. *Math. Oper. Res.* **46** 946–969. MR4312583 https://doi.org/10.1287/moor.2020.1085

[30] DUCHI, J. C. and NAMKOONG, H. (2021). Learning models with uniform performance via distributionally robust optimization. *Ann. Statist.* **49** 1378–1406. MR4298868 https://doi.org/10.1214/20-aos2004

[31] FISHER, R. A., CORBET, A. S. and WILLIAMS, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* 42–58.

[32] FOURNIER, N. (2023). Convergence of the empirical measure in expected Wasserstein distance: Non-asymptotic explicit bounds in $\mathbb{R}^d$. *ESAIM Probab. Stat.* **27** 749–775. MR4624322 https://doi.org/10.1051/ps/2023011

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

[34] FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2017). *The Elements of Statistical Learning*: *Data Mining*, *Inference and Prediction*, 2nd ed. *Series in Statistics* **1**. Springer, New York.
[35] GAO, R., CHEN, X. and KLEYWEGT, A. J. (2024). Wasserstein distributionally robust optimization and variation regularization. *Oper. Res.* **72** 1177–1191. MR4780740
[36] GAO, R. and KLEYWEGT, A. (2023). Distributionally robust stochastic optimization with Wasserstein distance. *Math. Oper. Res.* **48** 603–655. MR4588934 https://doi.org/10.1287/moor.2022.1275
[37] GIANNONE, D., LENZA, M. and PRIMICERI, G. E. (2021). Economic predictions with big data: The illusion of sparsity. *Econometrica* **89** 2409–2437.
[38] GOLDFELD, Z., KATO, K., RIOUX, G. and SADHU, R. (2024). Statistical inference with regularized optimal transport. *Inf. Inference* **13** 13. MR4701828 https://doi.org/10.1093/imaiai/iaad056
[39] GOODFELLOW, I., BENGIO, Y. and COURVILLE, A. (2016). *Deep Learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR3617773
[40] GOODFELLOW, I. J., SHLENS, J. and SZEGEDY, C. (2014). Explaining and harnessing adversarial examples. CoRR. Available at arXiv:1412.6572.
[41] KOLOURI, S., NADJAHI, K., SIMSEKLI, U., BADEAU, R. and ROHDE, G. (2019). Generalized sliced Wasserstein distances. *Adv. Neural Inf. Process. Syst.* **32**.
[42] KPOTUFE, S. and MARTINET, G. (2021). Marginal singularity and the benefits of labels in covariate-shift. *Ann. Statist.* **49** 3299–3323. MR4352531 https://doi.org/10.1214/21-aos2084
[43] KUHN, D., ESFAHANI, P. M., NGUYEN, V. A. and SHAFIEEZADEH-ABADEH, S. (2019). Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics* 130–166. Informs.
[44] KURAKIN, A., GOODFELLOW, I. and BENGIO, S. (2016). Adversarial machine learning at scale. arXiv preprint. Available at arXiv:1611.01236.
[45] LAM, H. (2016). Robust sensitivity analysis for stochastic systems. *Math. Oper. Res.* **41** 1248–1275. MR3544795 https://doi.org/10.1287/moor.2015.0776
[46] LEE, J. and RAGINSKY, M. (2018). Minimax statistical learning with Wasserstein distances. *Adv. Neural Inf. Process. Syst.* **31**.
[47] LEI, J. (2020). Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces. *Bernoulli* **26** 767–798. MR4036051 https://doi.org/10.3150/19-BEJ1151
[48] LI, C. and MÜLLER, U. K. (2021). Linear regression with many controls of limited explanatory power. *Quant. Econ.* **12** 405–442. MR4325590 https://doi.org/10.3982/qe1577
[49] LIN, T., ZHENG, Z., CHEN, E., CUTURI, M. and JORDAN, M. I. (2021). On projection robust optimal transport: Sample complexity and model misspecification. In *International Conference on Artificial Intelligence and Statistics* 262–270. PMLR.
[50] MANSOUR, Y., MOHRI, M. and ROSTAMIZADEH, A. (2009). Domain adaptation: Learning bounds and algorithms. In *Proceedings of the* 22*nd Annual Conference on Learning Theory* (*COLT* 2009).
[51] MOHAJERIN ESFAHANI, P. and KUHN, D. (2018). Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Math. Program.* **171** 115–166. MR3844536 https://doi.org/10.1007/s10107-017-1172-1
[52] MONTIEL OLEA, J. L., RUSH, C., VELEZ, A. and WIESEL, J. (2026). Supplement to "The distributionally robust prediction error of the $\sqrt{\text{LASSO}}$ and related estimators." https://doi.org/10.1214/25-AOS2599SUPP
[53] NGUYEN, V. A., ABADEH, S. S., FILIPOVIĆ, D. and KUHN, D. (2021). Mean-covariance robust risk measurement. Swiss Finance Institute Research Paper 21-93.
[54] NIETERT, S., GOLDFELD, Z., SADHU, R. and KATO, K. (2022). Statistical, robustness, and computational guarantees for sliced Wasserstein distances. *Adv. Neural Inf. Process. Syst.* **35** 28179–28193.
[55] NILES-WEED, J. and RIGOLLET, P. (2022). Estimation of Wasserstein distances in the spiked transport model. *Bernoulli* **28** 2663–2688. MR4474558 https://doi.org/10.3150/21-bej1433
[56] PATY, F.-P. and CUTURI, M. (2019). Subspace robust Wasserstein distances. In *International Conference on Machine Learning* 5072–5081. PMLR.
[57] PATY, F.-P. and CUTURI, M. (2019). Subspace robust Wasserstein distances. In *International Conference on Machine Learning* 5072–5081. PMLR.
[58] QUINONERO-CANDELA, J., SUGIYAMA, M., SCHWAIGHOFER, A. and LAWRENCE, N. D. (2008). *Dataset Shift in Machine Learning*. MIT Press, Cambridge.
[59] RABIN, J., PEYRÉ, G., DELON, J. and BERNOT, M. (2011). Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision* 435–446. Springer, Berlin.
[60] REDDI, S., POCZOS, B. and SMOLA, A. (2015). Doubly robust covariate shift correction. In *Proceedings of the AAAI Conference on Artificial Intelligence* **29**.

[61] RÉNYI, A. (1961). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* **1**. Berkeley, CA, USA.
[62] ROCKAFELLAR, R. T. (1997). *Convex Analysis*. *Princeton Landmarks in Mathematics*. Princeton Univ. Press, Princeton, NJ. MR1451876
[63] SAHOO, R., LEI, L. and WAGER, S. (2022). Learning from a biased sample. arXiv preprint. Available at arXiv:2209.01754.
[64] SHAFIEEZADEH ABADEH, S., MOHAJERIN ESFAHANI, P. M. and KUHN, D. (2015). Distributionally robust logistic regression. *Adv. Neural Inf. Process. Syst.* **28**.
[65] SHIMODAIRA, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Statist. Plann. Inference* **90** 227–244. MR1795598 https://doi.org/10.1016/S0378-3758(00)00115-4
[66] SINGH, S. and PÓCZOS, B. (2018). Minimax distribution estimation in Wasserstein distance. arXiv preprint. Available at arXiv:1802.08855.
[67] SINHA, A., NAMKOONG, H. and DUCHI, J. C. (2018). Certifying some distributional robustness with principled adversarial training. In 6*th International Conference on Learning Representations*, *ICLR* 2018, *Vancouver*, *BC*, *Canada*, *April* 30–*May* 3, 2018, *Conference Track Proceedings*. OpenReview.net.
[68] STUCKY, B. and VAN DE GEER, S. (2017). Sharp oracle inequalities for square root regularization. *J. Mach. Learn. Res.* **18** 67. MR3714230
[69] SUGIYAMA, M., KRAULEDAT, M. and MÜLLER, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *J. Mach. Learn. Res.* **8**.
[70] SUGIYAMA, M. and MÜLLER, K.-R. (2005). Input-dependent estimation of generalization error under covariate shift. *Statist. Decisions* **23** 249–279. MR2255627 https://doi.org/10.1524/stnd.2005.23.4.249
[71] VILLANI, C. (2008). *Optimal Transport*: *Old and New* 338. Springer, Berlin.
[72] WANG, H., LI, G. and JIANG, G. (2007). Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *J. Bus. Econom. Statist.* **25** 347–355. MR2380753 https://doi.org/10.1198/073500106000000251
[73] WEED, J. and BACH, F. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli* **25** 2620–2648. MR4003560 https://doi.org/10.3150/18-BEJ1065
[74] WEN, J., YU, C.-N. and GREINER, R. (2014). Robust learning under uncertain test distributions: Relating covariate shift to model misspecification. In *International Conference on Machine Learning* 631–639. PMLR.
[75] WU, Q., LI, J. Y.-M. and MAO, T. (2025). On generalization and regularization via Wasserstein distributionally robust optimization. *Manag. Sci.*
[76] WU, Y. and WANG, L. (2020). A survey of tuning parameter selection for high-dimensional regression. *Annu. Rev. Stat. Appl.* **7** 209–226. MR4104191 https://doi.org/10.1146/annurev-statistics-030718-105038
[77] WU, Y. and YANG, P. (2019). Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *Ann. Statist.* **47** 857–883. MR3909953 https://doi.org/10.1214/17-AOS1665

The Annals of Statistics
2026, Vol. 54, No. 2, 1028–1052
https://doi.org/10.1214/25-AOS2601

GENERALIZED LINEAR SPECTRAL STATISTICS OF HIGH-DIMENSIONAL SAMPLE COVARIANCE MATRICES AND ITS APPLICATIONS

BY YANLIN HU[a], QING YANG[b] AND XIAO HAN[c]

Department of Statistics & Finance, School of Management, University of Science and Technology of China,
[a]*hyl11@mail.ustc.edu.cn*, [b]*yangq@ustc.edu.cn*, [c]*xhan011@ustc.edu.cn*

In this paper, we introduce the **G**eneralized **L**inear **S**pectral **S**tatistics (GLSS) of a high-dimensional sample covariance matrix $\boldsymbol{S}_n$, denoted as $\operatorname{tr} f(\boldsymbol{S}_n)\boldsymbol{B}_n$, which effectively captures distinct spectral properties of $\boldsymbol{S}_n$ by incorporating an ancillary matrix $\boldsymbol{B}_n$ and a test function f. The joint asymptotic normality of GLSS associated with different test functions is established under mild assumptions on $\boldsymbol{B}_n$ and the underlying distribution, when the dimension n and sample size N are comparable. The convergence rate of GLSS is determined by $\sqrt{N/\operatorname{rank}(\boldsymbol{B}_n)}$. Subsequently, we propose a novel functional projection approach based on GLSS for hypothesis testing on eigenspaces of "population-spiked" covariance matrices, showcasing a universality phenomenon in the magnitude of the spikes. The theoretical accuracy of our results established for GLSS and the advantages of the newly suggested testing procedure are demonstrated through various numerical studies.

REFERENCES

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

[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 https://doi.org/10.1007/978-1-4419-0661-8

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

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

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

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

[7] BAO, Z., DING, X., WANG, J. and WANG, K. (2022). Statistical inference for principal components of spiked covariance matrices. *Ann. Statist.* **50** 1144–1169. MR4404931 https://doi.org/10.1214/21-aos2143

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

[9] BODNAR, T. and PAROLYA, N. (2024). Reviving pseudo-inverses: Asymptotic properties of large dimensional Moore-Penrose and Ridge-type inverses with applications. Preprint. Available at arXiv:2403.15792.

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

[11] CIPOLLONI, G., ERDŐS, L. and SCHRÖDER, D. (2023). Functional central limit theorems for Wigner matrices. *Ann. Appl. Probab.* **33** 447–489. MR4551555 https://doi.org/10.1214/22-aap1820

MSC2020 subject classifications. Primary 62H10, 60B20; secondary 62H15, 60F05.

Key words and phrases. Sample covariance matrix, random matrix theory, eigenspaces, generalized linear spectral statistics.

[12] EL KAROUI, N. (2007). Tracy-Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices. *Ann. Probab.* **35** 663–714. MR2308592 https://doi.org/10.1214/009117906000000917

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

[14] HALLIN, M., PAINDAVEINE, D. and VERDEBOUT, T. (2010). Optimal rank-based testing for principal components. *Ann. Statist.* **38** 3245–3299. MR2766852 https://doi.org/10.1214/10-AOS810

[15] HAN, X., TONG, X. and FAN, Y. (2023). Eigen selection in spectral clustering: A theory-guided practice. *J. Amer. Statist. Assoc.* **118** 109–121. MR4571110 https://doi.org/10.1080/01621459.2021.1917418

[16] HU, Y., YANG, Q. and HAN, X. (2026). Supplement to "Generalized linear spectral statistics of high-dimensional sample covariance matrices and its applications." https://doi.org/10.1214/25-AOS2601SUPP

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

[18] JOHNSTONE, I. M. and LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc.* **104** 682–693. MR2751448 https://doi.org/10.1198/jasa.2009.0121

[19] JOHNSTONE, I. M. and YANG, J. (2018). Notes on asymptotics of sample eigenstructure for spiked covariance models with non-Gaussian data. Preprint. Available at arXiv:1810.10427.

[20] KOLTCHINSKII, V. and LOUNICI, K. (2017). New asymptotic results in principal component analysis. *Sankhya A* **79** 254–297. MR3707422 https://doi.org/10.1007/s13171-017-0106-6

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

[22] LEE, J. O. and SCHNELLI, K. (2016). Tracy-Widom distribution for the largest eigenvalue of real sample covariance matrices with general population. *Ann. Appl. Probab.* **26** 3786–3839. MR3582818 https://doi.org/10.1214/16-AAP1193

[23] LI, Q., CHENG, G., FAN, J. and WANG, Y. (2018). Embracing the blessing of dimensionality in factor models. *J. Amer. Statist. Assoc.* **113** 380–389. MR3803472 https://doi.org/10.1080/01621459.2016.1256815

[24] LIU, X., LIU, Y., PAN, G., ZHANG, L. and ZHANG, Z. (2025). Asymptotic limits of spiked eigenvalues and eigenvectors of signal-plus-noise matrices with weak signals and heteroskedastic noise. *Bernoulli* **31** 2351–2376. MR4890836 https://doi.org/10.3150/24-bej1808

[25] MESTRE, X. (2008). On the asymptotic behavior of the sample estimates of eigenvalues and eigenvectors of covariance matrices. *IEEE Trans. Signal Process.* **56** 5353–5368. MR2472837 https://doi.org/10.1109/TSP.2008.929662

[26] NAJIM, J. and YAO, J. (2016). Gaussian fluctuations for linear spectral statistics of large random covariance matrices. *Ann. Appl. Probab.* **26** 1837–1887. MR3513608 https://doi.org/10.1214/15-AAP1135

[27] NAUMOV, A., SPOKOINY, V. and ULYANOV, V. (2019). Bootstrap confidence sets for spectral projectors of sample covariance. *Probab. Theory Related Fields* **174** 1091–1132. MR3980312 https://doi.org/10.1007/s00440-018-0877-2

[28] PAN, G. M. and ZHOU, W. (2008). Central limit theorem for signal-to-interference ratio of reduced rank linear receiver. *Ann. Appl. Probab.* **18** 1232–1270. MR2418244 https://doi.org/10.1214/07-AAP477

[29] PAUL, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17** 1617–1642. MR2399865

[30] PEARSON, K. (1901). On lines and planes of closest fit to systems of points in space. *Philos. Mag. Ser.* **1** 559–572.

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

[32] SILIN, I. and FAN, J. (2020). Hypothesis testing for eigenspaces of covariance matrix.

[33] SILIN, I. and SPOKOINY, V. (2018). Bayesian inference for spectral projectors of the covariance matrix. *Electron. J. Stat.* **12** 1948–1987. MR3815302 https://doi.org/10.1214/18-EJS1451

[34] WACHTER, K. W. (1978). The strong limits of random matrix spectra for sample matrices of independent elements. *Ann. Probab.* **6** 1–18. MR0467894 https://doi.org/10.1214/aop/1176995607

[35] YAO, J., ZHENG, S. and BAI, Z. (2015). *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. *Cambridge Series in Statistical and Probabilistic Mathematics* **39**. Cambridge Univ. Press, New York. MR3468554 https://doi.org/10.1017/CBO9781107588080

[36] YIN, Y. and ZHOU, W. (2023). Limiting behavior of bilinear forms for the resolvent of sample covariance matrices under elliptical distribution with applications. Preprint. Available at arXiv:2312.16373.

[37] YIN, Y. Q., BAI, Z. D. and KRISHNAIAH, P. R. (1988). On the limit of the largest eigenvalue of the large-dimensional sample covariance matrix. *Probab. Theory Related Fields* **78** 509–521. MR0950344 https://doi.org/10.1007/BF00353874

[38] ZHENG, S., BAI, Z. and YAO, J. (2015). Substitution principle for CLT of linear spectral statistics of high-dimensional sample covariance matrices with applications to hypothesis testing. *Ann. Statist.* **43** 546–591. MR3316190 https://doi.org/10.1214/14-AOS1292

[39] ZHENG, S., CHEN, Z., CUI, H. and LI, R. (2019). Hypothesis testing on linear structures of high-dimensional covariance matrix. *Ann. Statist.* **47** 3300–3334. MR4025743 https://doi.org/10.1214/18-AOS1779

The Annals of Statistics
2026, Vol. 54, No. 2, 1053–1079
https://doi.org/10.1214/25-AOS2602

REVIVING PSEUDO-INVERSES: ASYMPTOTIC PROPERTIES OF LARGE DIMENSIONAL MOORE–PENROSE AND RIDGE-TYPE INVERSES WITH APPLICATIONS

BY TARAS BODNAR[1,a] AND NESTOR PAROLYA[2,b]

[1]*Department of Management and Engineering, Linköping University,* [a]*taras.bodnar@liu.se*
[2]*Department of Applied Mathematics, Delft University of Technology,* [b]*n.parolya@tudelft.nl*

In this paper, we derive high-dimensional asymptotic properties of the Moore–Penrose inverse and, as a byproduct, of various ridge-type inverses of the sample covariance matrix. In particular, the analytical expressions of the asymptotic behavior of the weighted sample trace moments of generalized inverse matrices are deduced in terms of the partial exponential Bell polynomials, which can be easily computed in practice. The existent results for pseudo-inverses are extended in several directions: (i) First, the population covariance matrix is not assumed to be a multiple of the identity matrix; (ii) Second, the assumption of normality is not used in the derivation; (iii) Third, the asymptotic results are derived under the high-dimensional asymptotic regime. Our findings provide universal methodology for construction of fully data-driven improved shrinkage estimators of the precision matrix, optimal portfolio weights and beyond. It is found that the Moore–Penrose inverse acts asymptotically as a certain regularizer of the true covariance matrix and it seems that its proper transformation (shrinkage) performs similar to or even outperforms the existing benchmarks in many applications, while keeping the computational time as minimal as possible.

REFERENCES

AHLFORS, L. V. (1953). *Complex Analysis*. *An Introduction to the Theory of Analytic Functions of One Complex Variable*. McGraw-Hill, New York. MR0054016

AO, M., YINGYING, L. and ZHENG, X. (2019). Approaching mean-variance efficiency for large portfolios. *Rev. Financ. Stud.* **32** 2890–2919.

BAI, Z., FAHEY, M. and GOLUB, G. (1996). Some large-scale matrix computation problems. *J. Comput. Appl. Math.* **74** 71–89. MR1430368 https://doi.org/10.1016/0377-0427(96)00018-0

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

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

BELL, E. T. (1927/28). Partition polynomials. *Ann. of Math.* (2) **29** 38–46. MR1502817 https://doi.org/10.2307/1967979

BELL, E. T. (1934). Exponential polynomials. *Ann. of Math.* (2) **35** 258–277. MR1503161 https://doi.org/10.2307/1968431

BEN-ISRAEL, A. and GREVILLE, T. N. E. (2003). *Generalized Inverses*: *Theory and Applications*, 2nd ed. *CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC* **15**. Springer, New York. MR1987382

BODNAR, T., DETTE, H. and PAROLYA, N. (2016). Spectral analysis of the Moore–Penrose inverse of a large dimensional sample covariance matrix. *J. Multivariate Anal.* **148** 160–172. MR3493027 https://doi.org/10.1016/j.jmva.2016.03.001

BODNAR, T., DETTE, H. and PAROLYA, N. (2019). Testing for independence of large dimensional vectors. *Ann. Statist.* **47** 2977–3008. MR3988779 https://doi.org/10.1214/18-AOS1771

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 https://doi.org/10.1016/j.jmva.2014.08.006

MSC2020 subject classifications. Primary 60B20, 15A09, 62R07; secondary 62H12, 62F12.

Key words and phrases. Moore–Penrose inverse, Bell polynomials, sample covariance matrix, random matrix theory, high-dimensional asymptotics.

BODNAR, T., GUPTA, A. K. and PAROLYA, N. (2016). Direct shrinkage estimation of large dimensional precision matrix. *J. Multivariate Anal.* **146** 223–236. MR3477661 https://doi.org/10.1016/j.jmva.2015.09.010

BODNAR, T. and OKHRIN, Y. (2008). Properties of the singular, inverse and generalized inverse partitioned Wishart distributions. *J. Multivariate Anal.* **99** 2389–2405. MR2463397 https://doi.org/10.1016/j.jmva.2008.02.024

BODNAR, T., OKHRIN, Y. and PAROLYA, N. (2023). Optimal shrinkage-based portfolio selection in high dimensions. *J. Bus. Econom. Statist.* **41** 140–156. MR4522158 https://doi.org/10.1080/07350015.2021.2004897

BODNAR, T. and PAROLYA, N. (2026). Supplement to "Reviving pseudo-inverses: Asymptotic properties of large dimensional Moore–Penrose and Ridge-type inverses with applications." https://doi.org/10.1214/25-AOS2602SUPP

BODNAR, T., PAROLYA, N. and SCHMID, W. (2018). Estimation of the global minimum variance portfolio in high dimensions. *European J. Oper. Res.* **266** 371–390. MR3737003 https://doi.org/10.1016/j.ejor.2017.09.028

BODNAR, T., PAROLYA, N. and THORSÉN, E. (2023). Is the empirical out-of-sample variance an informative risk measure for the high-dimensional portfolios? *Finance Res. Lett.* **54** 103807.

BODNAR, T., PAROLYA, N. and THORSÉN, E. (2024). Two is better than one: Regularized shrinkage of large minimum variance portfolios. *J. Mach. Learn. Res.* **25** 1–32. MR4777415

CAI, T. T., HU, J., LI, Y. and ZHENG, X. (2020). High-dimensional minimum variance portfolio estimation based on high-frequency data. *J. Econometrics* **214** 482–494. MR4057056 https://doi.org/10.1016/j.jeconom.2019.04.039

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

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

COOK, R. D. and FORZANI, L. (2011). On the mean and variance of the generalized inverse of a singular Wishart matrix. *Electron. J. Stat.* **5** 146–158. MR2786485 https://doi.org/10.1214/11-EJS602

DERUMIGNY, A., PAROLYA, N. and BODNAR, T. (2026). UniversalShrink: Estimation of Covariance Matrices and Functions Thereof by Shrinkage. R package version 0.0.1. Available at https://github.com/AlexisDerumigny/UniversalShrink.

DI NARDO, E., GUARINO, G. and SENATO, D. (2008). A unifying framework for k-statistics, polykays and their multivariate generalizations. *Bernoulli* **14** 440–468. MR2544096 https://doi.org/10.3150/07-BEJ6163

FENG, Y. and PALOMAR, D. P. (2016). A signal processing perspective on financial engineering. *Found. Trends Signal Process.* **9** 1–231. MR3540800 https://doi.org/10.1561/2000000072

FRAHM, G. and MEMMEL, C. (2010). Dominating estimators for minimum-variance portfolios. *J. Econometrics* **159** 289–302. MR2733122 https://doi.org/10.1016/j.jeconom.2010.07.007

GOLOSNOY, V. and OKHRIN, Y. (2007). Multivariate shrinkage for optimal portfolio weights. *Eur. J. Finance* **13** 441–458.

GOLUB, G. H. and STRAKOŠ, Z. (1994). Estimates in quadratic formulas. *Numer. Algorithms* **8** 241–268. MR1309223 https://doi.org/10.1007/BF02142693

HAFF, L. R. (1979). An identity for the Wishart distribution with applications. *J. Multivariate Anal.* **9** 531–544. MR0556910 https://doi.org/10.1016/0047-259X(79)90056-3

HARVILLE, D. A. (1997). *Matrix Algebra from a Statistician's Perspective*. Springer, New York. MR1467237 https://doi.org/10.1007/b98818

HEINY, J. (2019). Random matrix theory for heavy-tailed time series. *J. Math. Sci. (N. Y.)* **237** 652–666. MR3924297 https://doi.org/10.1007/s10958-019-04191-3

HEINY, J. and MIKOSCH, T. (2021). Large sample autocovariance matrices of linear processes with heavy tails. *Stoch. Process. Appl.* **141** 344–375. MR4301551 https://doi.org/10.1016/j.spa.2021.07.010

HEINY, J. and YAO, J. (2022). Limiting distributions for eigenvalues of sample correlation matrices from heavy-tailed populations. *Ann. Statist.* **50** 3249–3280. MR4524496 https://doi.org/10.1214/22-aos2226

HILLE, E. (2002). *Analytic Function Theory, Vol.* 2. Amer. Math. Soc., Providence.

IMORI, S. and VON ROSEN, D. (2020). On the mean and dispersion of the Moore–Penrose generalized inverse of a Wishart matrix. *Electron. J. Linear Algebra* **36** 124–133. MR4089045 https://doi.org/10.13001/ela.2020.5091

KAN, R., WANG, X. and ZHOU, G. (2022). Optimal portfolio choice with estimation risk: No risk-free asset case. *Manag. Sci.* **68** 2047–2068.

KRISHNAMOORTHY, K. and GUPTA, A. K. (1989). Improved minimax estimation of a normal precision matrix. *Canad. J. Statist.* **17** 91–102. MR1014094 https://doi.org/10.2307/3314766

KUBOKAWA, T. and SRIVASTAVA, M. S. (2008). Estimation of the precision matrix of a singular Wishart distribution and its application in high-dimensional data. *J. Multivariate Anal.* **99** 1906–1928. MR2466543 https://doi.org/10.1016/j.jmva.2008.01.016

LAM, C. (2020). High-dimensional covariance matrix estimation. *Wiley Interdiscip. Rev.: Comput. Stat.* **12** Paper No. e1485, 21. MR4072468 https://doi.org/10.1002/wics.1485
LASSANCE, N., VANDERVEKEN, R. and VRINS, F. (2024). On the combination of naive and mean-variance portfolio strategies. *J. Bus. Econom. Statist.* **42** 875–889. MR4757112 https://doi.org/10.1080/07350015.2023.2256801
LEDOIT, O. and PÉCHÉ, S. (2011). Eigenvectors of some large sample covariance matrix ensembles. *Probab. Theory Related Fields* **151** 233–264. MR2834718 https://doi.org/10.1007/s00440-010-0298-3
LEDOIT, O. and WOLF, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.* **88** 365–411. MR2026339 https://doi.org/10.1016/S0047-259X(03)00096-4
LEDOIT, O. and WOLF, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Ann. Statist.* **40** 1024–1060. MR2985942 https://doi.org/10.1214/12-AOS989
LEDOIT, O. and WOLF, M. (2020). Analytical nonlinear shrinkage of large-dimensional covariance matrices. *Ann. Statist.* **48** 3043–3065. MR4152634 https://doi.org/10.1214/19-AOS1921
LEDOIT, O. and WOLF, M. (2021). Shrinkage estimation of large covariance matrices: Keep it simple, statistician? *J. Multivariate Anal.* **186** Paper No. 104796, 24. MR4308803 https://doi.org/10.1016/j.jmva.2021.104796
LEDOIT, O. and WOLF, M. (2022). Quadratic shrinkage for large covariance matrices. *Bernoulli* **28** 1519–1547. MR4411501 https://doi.org/10.3150/20-bej1315
MEYER, C. D. JR. (1973). Generalized inversion of modified matrices. *SIAM J. Appl. Math.* **24** 315–323. MR0316463 https://doi.org/10.1137/0124033
MUIRHEAD, R. J. (1982). *Aspects of Multivariate Statistical Theory*. *Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. MR0652932
PAN, G. (2014). Comparison between two types of large sample covariance matrices. *Ann. Inst. Henri Poincaré Probab. Stat.* **50** 655–677. MR3189088 https://doi.org/10.1214/12-AIHP506
PENROSE, R. (1955). A generalized inverse for matrices. *Proc. Camb. Philos. Soc.* **51** 406–413. MR0069793
RAO, C. R. and MITRA, S. K. (1972). Generalized inverse of a matrix and its applications. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (*Univ. California*, *Berkeley*, *Calif.*, 1970/1971), *Vol. I*: *Theory of Statistics* 601–620. Univ. California Press, Berkeley, CA. MR0403093
RENCHER, A. C. and CHRISTENSEN, W. F. (2012). *Methods of Multivariate Analysis*, 3rd ed. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. MR2962097 https://doi.org/10.1002/9781118391686
RUBIO, F. and MESTRE, X. (2011). Spectral convergence for a general class of random matrices. *Statist. Probab. Lett.* **81** 592–602. MR2772917 https://doi.org/10.1016/j.spl.2011.01.004
RUDIN, W. (1987). *Real and Complex Analysis*, 3rd ed. McGraw-Hill, New York. MR0924157
SHI, H., HALLIN, M., DRTON, M. and HAN, F. (2022). On universally consistent and fully distribution-free rank tests of vector independence. *Ann. Statist.* **50** 1933–1959. MR4474478 https://doi.org/10.1214/21-aos2151
SRIVASTAVA, M. S. (2003). Singular Wishart and multivariate beta distributions. *Ann. Statist.* **31** 1537–1560. MR2012825 https://doi.org/10.1214/aos/1065705118
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
WANG, G., WEI, Y. and QIAO, S. (2018). *Generalized Inverses*: *Theory and Computations*, 2nd ed. *Developments in Mathematics* **53**. Springer, Singapore. MR3793648 https://doi.org/10.1007/978-981-13-0146-9
YANG, R. and BERGER, J. O. (1994). Estimation of a covariance matrix using the reference prior. *Ann. Statist.* **22** 1195–1211. MR1311972 https://doi.org/10.1214/aos/1176325625
YAO, J., ZHENG, S. and BAI, Z. (2015). *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. *Cambridge Series in Statistical and Probabilistic Mathematics* **39**. Cambridge Univ. Press, New York. MR3468554 https://doi.org/10.1017/CBO9781107588080

The Annals of Statistics
2026, Vol. 54, No. 2, 1080–1099
https://doi.org/10.1214/25-AOS2605

ADAPTIVE BAYESIAN REGRESSION ON DATA WITH LOW INTRINSIC DIMENSIONALITY

BY TAO TANG[1], NAN WU[3], XIUYUAN CHENG[1,a] AND DAVID DUNSON[2,b]

[1]*Department of Mathematics, Duke University,* [a]*xiuyuan.cheng@duke.edu*
[2]*Department of Statistical Science, Duke University,* [b]*dunson@duke.edu*
[3]*Department of Mathematical Sciences, The University of Texas at Dallas*

We study how the posterior contraction rate under a Gaussian process (GP) prior depends on the intrinsic dimension of the predictors and the smoothness of the regression function. An open question is whether a generic GP prior that does not incorporate knowledge of the intrinsic lower-dimensional structure of the predictors can attain an adaptive rate for a broad class of such structures. We show that this is indeed the case, establishing conditions under which the posterior contraction rates become adaptive to the intrinsic dimension in terms of the covering number of the data domain (the Minkowski dimension) and prove the nonparametric posterior contraction rate, up to a logarithmic factor. When the domain is a compact manifold, we prove the RKHS approximation to intrinsically defined Hölder functions on the manifold of any order of smoothness by a novel analysis, leading to the optimal adaptive posterior contraction rate. We propose an empirical Bayes prior on the kernel bandwidth using kernel affinity and k-nearest neighbor statistics, bypassing explicit estimation of the intrinsic dimension. The efficiency of the proposed Bayesian regression approach is demonstrated in various numerical experiments.

REFERENCES

[1] ARONSZAJN, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68** 337–404. MR0051437 https://doi.org/10.2307/1990404
[2] BANACH, S. (1938). Über homogene polynome in (L^2). *Studia Math.* **7** 36–44.
[3] BERLINET, A. and THOMAS-AGNAN, C. (2011). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer, Berlin.
[4] BICKEL, P. J. and LI, B. (2007). Local polynomial regression on unknown manifolds. In *Complex Datasets and Inverse Problems. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **54** 177–186. IMS, Beachwood, OH. MR2459188 https://doi.org/10.1214/074921707000000148
[5] CASTILLO, I. and EGELS, P. (2025). Posterior and variational inference for deep neural networks with heavy-tailed weights. *J. Mach. Learn. Res.* **26** 122. MR4962876
[6] CASTILLO, I., KERKYACHARIAN, G. and PICARD, D. (2014). Thomas Bayes' walk on manifolds. *Probab. Theory Related Fields* **158** 665–710. MR3176362 https://doi.org/10.1007/s00440-013-0493-0
[7] CASTILLO, I. and RANDRIANARISOA, T. (2025). Deep horseshoe Gaussian processes. *Ann. Statist.* **53** 1886–1912. MR4985253 https://doi.org/10.1214/25-AOS2522
[8] CHENG, M.-Y. and WU, H.-T. (2013). Local linear regression on manifolds and its geometric interpretation. *J. Amer. Statist. Assoc.* **108** 1421–1434. MR3174718 https://doi.org/10.1080/01621459.2013.827984
[9] CHENG, X. and WU, H.-T. (2022). Convergence of graph Laplacian with kNN self-tuned kernels. *Inf. Inference* **11** 889–957. MR4491976 https://doi.org/10.1093/imaiai/iaab019
[10] DO CARMO, M. P. and FLAHERTY FRANCIS, J. (1992). *Riemannian Geometry* 6. Springer, Berlin.
[11] DUNSON, D. B., WU, H.-T. and WU, N. (2022). Graph based Gaussian processes on restricted domains. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 414–439. MR4412992 https://doi.org/10.1111/rssb.12486
[12] FALCONER, K. (2004). *Fractal Geometry*: *Mathematical Foundations and Applications*. Wiley, New York.

MSC2020 subject classifications. Primary 62G08, 60G15, 62G20; secondary 51-08.
Key words and phrases. Adaptive rate, Gaussian process, posterior contraction rate, manifold regression, nonparametric rate.

[13] FARAHMAND, A. M., SZEPESVÁRI, C. and AUDIBERT, J.-Y. (2007). Manifold-adaptive dimension estimation. In *ICML '07: Proceedings of the 24th International Conference on Machine Learning* 265–272. https://doi.org/10.1145/1273496.12735
[14] FINOCCHIO, G. and SCHMIDT-HIEBER, J. (2023). Posterior contraction for deep Gaussian process priors. *J. Mach. Learn. Res.* **24** 66. MR4582488
[15] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007 https://doi.org/10.1214/aos/1016218228
[16] GHOSAL, S. and VAN DER VAART, A. (2007). Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.* **35** 192–223. MR2332274 https://doi.org/10.1214/009053606000001172
[17] GHOSAL, S. and VAN DER VAART, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. *Cambridge Series in Statistical and Probabilistic Mathematics* **44**. Cambridge Univ. Press, Cambridge. MR3587782 https://doi.org/10.1017/9781139029834
[18] HAMM, T. and STEINWART, I. (2021). Adaptive learning rates for support vector machines working on data with low intrinsic dimension. *Ann. Statist.* **49** 3153–3180. MR4352526 https://doi.org/10.1214/21-aos2078
[19] JIANG, S. and TOKDAR, S. T. (2021). Variable selection consistency of Gaussian process regression. *Ann. Statist.* **49** 2491–2505. MR4338372 https://doi.org/10.1214/20-aos2043
[20] KARVONEN, T. and OATES, C. J. (2023). Maximum likelihood estimation in Gaussian process regression is ill-posed. *J. Mach. Learn. Res.* **24** 120. MR4583281
[21] KI, D. and PARK, B. U. (2021). Intrinsic Hölder classes of density functions on Riemannian manifolds and lower bounds to convergence rates. *Statist. Probab. Lett.* **169** 108959. MR4163210 https://doi.org/10.1016/j.spl.2020.108959
[22] KNAPIK, B. T., SZABÓ, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2016). Bayes procedures for adaptive inference in inverse problems for the white noise model. *Probab. Theory Related Fields* **164** 771–813. MR3477780 https://doi.org/10.1007/s00440-015-0619-7
[23] KPOTUFE, S. (2011). kNN regression adapts to local intrinsic dimension. *Adv. Neural Inf. Process. Syst.* **24**.
[24] KPOTUFE, S. and GARG, V. (2013). Adaptivity to local smoothness and dimension in kernel regression. *Adv. Neural Inf. Process. Syst.* **26**.
[25] KULKARNI, S. R. and POSNER, S. E. (1995). Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Trans. Inf. Theory* **41** 1028–1039. MR1366756 https://doi.org/10.1109/18.391248
[26] LAFFERTY, J. and WASSERMAN, L. (2008). Rodeo: Sparse, greedy nonparametric regression. *Ann. Statist.* **36** 28–63. MR2387963 https://doi.org/10.1214/009053607000000811
[27] LEVINA, E. and BICKEL, P. (2004). Maximum likelihood estimation of intrinsic dimension. *Adv. Neural Inf. Process. Syst.* **17**.
[28] NENE, S. A., NAYAR, S. K., MURASE, H. et al. (1996). Columbia object image library (coil-20).
[29] PETERSEN, P. (2006). *Riemannian Geometry*, 2nd ed. *Graduate Texts in Mathematics* **171**. Springer, New York. MR2243772
[30] ROSA, P., BOROVITSKIY, S., TERENIN, A. and ROUSSEAU, J. (2024). Posterior contraction rates for Matérn Gaussian processes on Riemannian manifolds. *Adv. Neural Inf. Process. Syst.* **36**.
[31] ROSA, P. and ROUSSEAU, J. (2025). Nonparametric regression on random geometric graphs sampled from submanifolds. *J. Mach. Learn. Res.* **26** 164. MR4962918
[32] ROUSSEAU, J. and SZABO, B. (2017). Asymptotic behaviour of the empirical Bayes posteriors associated to maximum marginal likelihood estimator. *Ann. Statist.* **45** 833–865. MR3650402 https://doi.org/10.1214/16-AOS1469
[33] SCOTT, C. and NOWAK, R. D. (2006). Minimax-optimal classification with dyadic decision trees. *IEEE Trans. Inf. Theory* **52** 1335–1353. MR2241192 https://doi.org/10.1109/TIT.2006.871056
[34] SHEN, X. and WASSERMAN, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* **29** 687–714. MR1865337 https://doi.org/10.1214/aos/1009210686
[35] STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053. MR0673642
[36] SZABÓ, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2013). Empirical Bayes scaling of Gaussian priors in the white noise model. *Electron. J. Stat.* **7** 991–1018. MR3044507 https://doi.org/10.1214/13-EJS798
[37] TANG, T., WU, N., CHENG, X. and DUNSON, D. (2026). Supplement to "Adaptive Bayesian regression on data with low intrinsic dimensionality." https://doi.org/10.1214/25-AOS2605SUPP
[38] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B, Methodol.* **58** 267–288. MR1379242
[39] VAN DER VAART, A. and VAN ZANTEN, H. (2011). Information rates of nonparametric Gaussian process methods. *J. Mach. Learn. Res.* **12** 2095–2119. MR2819028

[40] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.* **36** 1435–1463. MR2418663 https://doi.org/10.1214/009053607000000613

[41] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.* **37** 2655–2675. MR2541442 https://doi.org/10.1214/08-AOS678

[42] WEINBERGER, S. (1994). *The Topological Classification of Stratified Spaces*. *Chicago Lectures in Mathematics*. Univ. Chicago Press, Chicago, IL. MR1308714

[43] YANG, Y. and DUNSON, D. B. (2016). Bayesian manifold regression. *Ann. Statist.* **44** 876–905. MR3476620 https://doi.org/10.1214/15-AOS1390

[44] YANG, Y. and TOKDAR, S. T. (2015). Minimax-optimal nonparametric regression in high dimensions. *Ann. Statist.* **43** 652–674. MR3319139 https://doi.org/10.1214/14-AOS1289

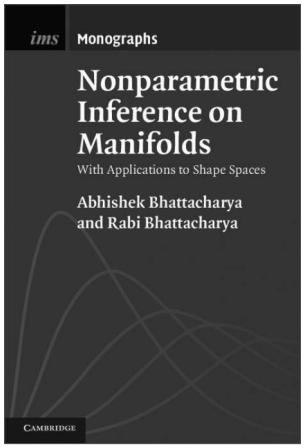
[45] YE, G.-B. and ZHOU, D.-X. (2008). Learning and approximation by Gaussians on Riemannian manifolds. *Adv. Comput. Math.* **29** 291–310. MR2438346 https://doi.org/10.1007/s10444-007-9049-0

[46] YE, G.-B. and ZHOU, D.-X. (2009). SVM learning and L^p approximation by Gaussians on Riemannian manifolds. *Anal. Appl. (Singap.)* **7** 309–339. MR2542742 https://doi.org/10.1142/S0219530509001384

[47] ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327 https://doi.org/10.1111/j.1467-9868.2005.00503.x

The Institute of Mathematical Statistics presents

IMS MONOGRAPHS



Nonparametric Inference on Manifolds

With Applications to Shape Spaces

Abhishek Bhattacharya, Rabi Bhattacharya

This book introduces in a systematic manner a general nonparametric theory of statistics on manifolds, with emphasis on manifolds of shapes. The theory has important and varied applications in medical diagnostics, image analysis, and machine vision. An early chapter of examples establishes the effectiveness of the new methods and demonstrates how they outperform their parametric counterparts. Inference is developed for both intrinsic and extrinsic Fréchet means of probability distributions on manifolds, then applied to shape spaces defined as orbits of landmarks under a Lie group of transformations—in particular, similarity, reflection similarity, affine and projective transformations. In addition, nonparametric Bayesian theory is adapted and extended to manifolds for the purposes of density estimation, regression and classification. Ideal for statisticians who analyze manifold data and wish to develop their own methodology, this book is also of interest to probabilists, mathematicians, computer scientists and morphometricians with mathematical training.

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

Hardback price US$51.00 (non-member price $85.00)

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.