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. 1—February 2026

THE ANNALS OF STATISTICS Vol. 54, No. 1, pp. 1–541 February 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

**President:** Kavita Ramanan, Division of Applied Mathematics, Brown University, Providence, RI 02912, USA

**President-Elect:** Richard J. Samworth, Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, Cambridge, CB3 0WB, UK

**Past President:** Tony Cai, Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA 19104-6304, USA

**Executive Secretary:** Peter Hoff, Department of Statistical Science, Duke University, Durham, NC 27708-0251, USA

**Treasurer:** Jiashun Jin, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA

**Program Secretary:** Annie Qu, Department of Statistics, University of California, Irvine, Irvine, CA 92697-3425, USA

## IMS EDITORS

**The Annals of Statistics.** *Editors*: Hans-Georg Müller, Department of Statistics, University of California, Davis, Davis, CA 95616, USA. Harrison Zhou, Department of Statistics and Data Science, Yale University, New Haven, CT 06520, USA

**The Annals of Applied Statistics.** *Editor-in-Chief*: Lexin Li, Department of Biostatistics and Epidemiology, University of California, Berkeley, Berkeley, CA 94720-7360, USA

**The Annals of Probability.** *Editors*: Paul Bourgade, Courant Institute of Mathematical Sciences, New York University, New York, NY 10012-1185, USA. Julien Dubedat, Department of Mathematics, Columbia University, New York, NY 10027, USA

**The Annals of Applied Probability.** *Editors*: Jian Ding, School of Mathematical Sciences, Peking University, 100871, Bejing, China. Claudio Landim, IMPA, 22461-320, Rio de Janeiro, Brazil

**Statistical Science.** *Editor*: Lutz Dümbgen, Institute of Mathematical Statistics and Actuarial Science, University of Bern, Alpeneggstrasse 22, CH-3012 Bern, Switzerland

**The IMS Bulletin.** *Editor*: Tati Howell, bulletin@imstat.org

***The Annals of Statistics* [ISSN 0090-5364 (print); ISSN 2168-8966 (online)],** Volume 54, Number 1, February 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.

Copyright © 2026 by the Institute of Mathematical Statistics
Printed in the United States of America

*The Annals of Statistics*
2026, Vol. 54, No. 1, 1–24
https://doi.org/10.1214/25-AOS2498

# DUALIZING LE CAM'S METHOD FOR FUNCTIONAL ESTIMATION I: GENERAL THEORY

BY YURY POLYANSKIY[1,a] AND YIHONG WU[2,b]

[1]*Department of Statistics and Data Science, Yale University,* [a]*yp@mit.edu*
[2]*Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology,* [b]*yihong.wu@yale.edu*

Le Cam's method (or the two-point method) is a commonly used tool for obtaining statistical lower bound and especially popular for functional estimation problems. This work aims to explain and give conditions for the tightness of Le Cam's lower bound in functional estimation from the perspective of convex duality. Under a variety of settings, it is shown that the maximization problem that searches for the best two-point lower bound, upon dualizing, becomes a minimization problem minimizing an upper bound on the quadratic risk over a family of estimators. Since by the minimax theorem two problems have the same value, this value also characterizes (up to a universal factor) the optimal estimation rate. For estimating linear functionals of a distribution, our work strengthens prior results of Donoho–Liu (*Ann. Statist.* **19** (1991) 633–667) (for quadratic loss) by dropping the Hölderian assumption on the modulus of continuity. For exponential families, our results extend those of Juditsky–Nemirovski (*Ann. Statist.* **37** (2009) 2278–2300) by characterizing the minimax risk for the quadratic loss under weaker assumptions on the exponential family. We also provide an extension to the high-dimensional setting for estimating separable functionals.

An application of our methodology to the area of "estimating the unseens" is provided in the companion paper (Polyanskiy and Wu (2023)), resolving the optimal rates (within logarithmic factors) of the distinct elements problem and Fisher's species problem.

## REFERENCES

[1] BORWEIN, J. M. and ZHUANG, D. (1986). On Fan's minimax theorem. *Math. Program.* **34** 232–234. MR0838482 https://doi.org/10.1007/BF01580587

[2] BROWN, L. D. (1986). Fundamentals of statistical exponential families with applications in statistical decision theory. In *Lecture Notes-Monograph Series* (S. S. Gupta, ed.) **9**. Institute of Mathematical Statistics, Hayward, CA.

[3] BROWN, L. D. and LOW, M. G. (1996). A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Statist.* **24** 2524–2535. MR1425965 https://doi.org/10.1214/aos/1032181166

[4] CAI, T. T. and LOW, M. G. (2011). Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *Ann. Statist.* **39** 1012–1041. MR2816346 https://doi.org/10.1214/10-AOS849

[5] COLLIER, O., COMMINGES, L. and TSYBAKOV, A. B. (2017). Minimax estimation of linear and quadratic functionals on sparsity classes. *Ann. Statist.* **45** 923–958. MR3662444 https://doi.org/10.1214/15-AOS1432

[6] COLLIER, O., COMMINGES, L., TSYBAKOV, A. B. and VERZELEN, N. (2018). Optimal adaptive estimation of linear functionals under sparsity. *Ann. Statist.* **46** 3130–3150. MR3851767 https://doi.org/10.1214/17-AOS1653

[7] DE, A., O'DONNELL, R. and SERVEDIO, R. A. (2020). Sharp bounds for population recovery. *Theory Comput.* **16** Paper No. 6. MR4189126 https://doi.org/10.4086/toc.2020.v016a006

[8] DONOHO, D. L. and LIU, R. C. (1991). Geometrizing rates of convergence. II. *Ann. Statist.* **19** 633–667. MR1105839 https://doi.org/10.1214/aos/1176348114

---

*MSC2020 subject classifications.* Primary 62G05; secondary 62C20, 62G86, 62G07.

*Key words and phrases.* Functional estimation, minimax optimality, convex optimization and duality, modulus of continuity, linear programming.

[9] DUNFORD, N. and SCHWARTZ, J. T. (1958). *Linear Operators*. *I*. *General Theory*. *Pure and Applied Mathematics*, *Vol*. 7. Interscience Publishers, New York. MR0117523

[10] DVIR, Z., RAO, A., WIGDERSON, A. and YEHUDAYOFF, A. (2012). Restriction access. In *Proceedings of the* 3*rd Innovations in Theoretical Computer Science Conference* 19–33. ACM, New York. MR3388375

[11] EFRON, B. and THISTED, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika* **63** 435–447.

[12] FAN, K. (1953). Minimax theorems. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **39** 42–47. MR0055678 https://doi.org/10.1073/pnas.39.1.42

[13] 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.

[14] GIBBS, A. L. and SU, F. E. (2002). On choosing and bounding probability metrics. *Int*. *Stat*. *Rev*. **70** 419–435.

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

[16] GROENEBOOM, P. and JONGBLOED, G. (2014). *Nonparametric Estimation Under Shape Constraints*: *Estimators*, *Algorithms and Asymptotics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **38**. Cambridge Univ. Press, New York. MR3445293 https://doi.org/10.1017/CBO9781139020893

[17] GROENEBOOM, P. and WELLNER, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. *DMV Seminar* **19**. Birkhäuser, Basel. MR1180321 https://doi.org/10.1007/978-3-0348-8621-5

[18] HUANG, J. and WELLNER, J. A. (1997). Interval censored survival data: A review of recent progress. In *Proceedings of the First Seattle Symposium in Biostatistics* 123–169. Springer, Berlin.

[19] JANA, S., POLYANSKIY, Y. and WU, Y. (2020). Extrapolating the profile of a finite population. In *Proceedings of Conference on Learning Theory* (*COLT*). Available at arXiv:2005.10561.

[20] JUDITSKY, A. and NEMIROVSKI, A. (2020). Near-optimal recovery of linear and $N$-convex functions on unions of convex sets. *Inf*. *Inference* **9** 423–453. MR4108974 https://doi.org/10.1093/imaiai/iaz011

[21] JUDITSKY, A. and NEMIROVSKI, A. (2020). *Statistical Inference via Convex Optimization*. *Princeton Series in Applied Mathematics*. Princeton Univ. Press, Princeton, NJ. MR4264425

[22] JUDITSKY, A. B. and NEMIROVSKI, A. S. (2009). Nonparametric estimation by convex programming. *Ann*. *Statist*. **37** 2278–2300. MR2543692 https://doi.org/10.1214/08-AOS654

[23] KAPLAN, E. L. and MEIER, P. (1958). Nonparametric estimation from incomplete observations. *J*. *Amer*. *Statist*. *Assoc*. **53** 457–481. MR0093867

[24] LE CAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. *Springer Series in Statistics*. Springer, New York. MR0856411 https://doi.org/10.1007/978-1-4612-4946-7

[25] LECAM, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann*. *Statist*. **1** 38–53. MR0334381

[26] LEPSKI, O., NEMIROVSKI, A. and SPOKOINY, V. (1999). On estimation of the $L_r$ norm of a regression function. *Probab*. *Theory Related Fields* **113** 221–253. MR1670867 https://doi.org/10.1007/s004409970006

[27] LUGOSI, G. and MENDELSON, S. (2020). Risk minimization by median-of-means tournaments. *J*. *Eur*. *Math*. *Soc*. (*JEMS*) **22** 925–965. MR4055993 https://doi.org/10.4171/jems/937

[28] ORLITSKY, A., SURESH, A. T. and WU, Y. (2016). Optimal prediction of the number of unseen species. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **113** 13283–13288. MR3582444 https://doi.org/10.1073/pnas.1607774113

[29] POLYANSKIY, Y., SURESH, A. T. and WU, Y. (2017). Sample complexity of population recovery. In *Proceedings of Conference on Learning Theory* (*COLT*). Available at arXiv:1702.05574v3.

[30] POLYANSKIY, Y. and WU, Y. (2026). Supplement to "Dualizing Le Cam's method for functional estimation I: General theory." https://doi.org/10.1214/25-AOS2498SUPP

[31] POLYANSKIY, Y. and WU, Y. (2023). Dualizing Le Cam's method for functional estimation II: Applications to estimating the unseens.

[32] POLYANSKIY, Y. and WU, Y. (2025). *Information Theory*: *From Coding to Learning*. Cambridge Univ. Press, Cambridge.

[33] RASKHODNIKOVA, S., RON, D., SHPILKA, A. and SMITH, A. (2009). Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM J*. *Comput*. **39** 813–842. MR2538840 https://doi.org/10.1137/070701649

[34] 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

[35] STRASSER, H. (1985). *Mathematical Theory of Statistics*: *Statistical Experiments and Asymptotic Decision Theory*. *De Gruyter Studies in Mathematics* **7**. de Gruyter, Berlin. MR0812467 https://doi.org/10.1515/9783110850826

[36] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. Springer, New York. MR2724359 https://doi.org/10.1007/b13794

[37] VALIANT, P. (2011). Testing symmetric properties of distributions. *SIAM J. Comput.* **40** 1927–1968. MR2863200 https://doi.org/10.1137/080734066

[38] WIGDERSON, A. and YEHUDAYOFF, A. (2012). Population recovery and partial identification. In 2012 *IEEE* 53*rd Annual Symposium on Foundations of Computer Science—FOCS* 2012 390–399. IEEE Computer Soc., Los Alamitos, CA. MR3186626

[39] WOLFOWITZ, J. (1957). The minimum distance method. *Ann. Math. Statist.* **28** 75–88. MR0088126 https://doi.org/10.1214/aoms/1177707038

[40] WU, Y. and YANG, P. (2018). Sample complexity of the distinct elements problem. *Math. Stat. Learn.* **1** 37–72. MR4049450 https://doi.org/10.4171/msl/1-1-2

[41] WU, Y. and YANG, P. (2020). Optimal estimation of Gaussian mixtures via denoised method of moments. *Ann. Statist.* **48** 1981–2007. MR4134783 https://doi.org/10.1214/19-AOS1873

[42] YU, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam* 423–435. Springer, New York. MR1462963

[43] ZHANG, C.-H. (1997). Empirical Bayes and compound estimation of normal means. *Statist. Sinica* **7** 181–193.

[44] ZHANG, C.-H. (2003). Compound decision theory and empirical Bayes methods. *Ann. Statist.* **31** 379–390.

*The Annals of Statistics*
2026, Vol. 54, No. 1, 25–47
https://doi.org/10.1214/25-AOS2505

# ON THE STRUCTURAL DIMENSION OF SLICED INVERSE REGRESSION

BY DONGMING HUANG[1,a], SONGTAO TIAN[2,b] AND QIAN LIN[3,c]

[1]*Department of Statistics & Data Science, National University of Singapore,* [a]*stahd@nus.edu.sg*
[2]*Department of Mathematical Sciences, Tsinghua University,* [b]*tiansongtao.2020@tsinghua.org.cn*
[3]*Department of Statistics and Data Science, Tsinghua University,* [c]*qianlin@tsinghua.edu.cn*

In this work, we address the longstanding puzzle that Sliced Inverse Regression (SIR) often performs poorly for sufficient dimension reduction when the structural dimension $d$ (the dimension of the central space) exceeds 4. We first show that in the multiple index model $Y = f(\mathcal{P}\boldsymbol{X}) + \epsilon$ where $\boldsymbol{X}$ is a $p$-standard normal vector, $\epsilon$ is an independent noise, and $\mathcal{P}$ is a projection operator from $\mathbb{R}^p$ to $\mathbb{R}^d$, if the link function $f$ follows the law of a Gaussian process. Then with high probability, the $d$th eigenvalue $\lambda_d$ of $\mathrm{Cov}[\mathbb{E}(\boldsymbol{X}|Y)]$ satisfies $\lambda_d \le Ce^{-\theta d}$ for some positive constants $C$ and $\theta$. We then focus on the low signal regime where $\lambda_d$ can be arbitrarily small and not larger than $d^{-8.1}$, and prove that the minimax risk of estimating the central space is lower bounded by $\frac{dp}{n\lambda_d}$. Combining these two results, we provide a convincing explanation for the poor performance of SIR when $d$ is large, a phenomenon that has perplexed researchers for nearly three decades. The technical tools developed here may be of independent interest for studying other sufficient dimension reduction methods.

## REFERENCES

ARMENTANO, D., AZAÏS, J.-M., GINSBOURGER, D. and LEÓN, J. R. (2019). Conditions for the finiteness of the moments of the volume of level sets. *Electron. Commun. Probab.* **24** Paper No. 17, 8. MR3933041 https://doi.org/10.1214/19-ECP214

AZAÏS, J.-M. and LEÓN, J. R. (2020). Necessary and sufficient conditions for the finiteness of the second moment of the measure of level sets. *Electron. J. Probab.* **25** Paper No. 107, 15. MR4147520

AZAÏS, J.-M. and WSCHEBOR, M. (2009). *Level Sets and Extrema of Random Processes and Fields*. Wiley, Hoboken, NJ. MR2478201 https://doi.org/10.1002/9780470434642

BENZAQUEN, S. and CABAÑA, E. M. (1982). The expected measure of the level sets of a regular stationary Gaussian process. *Pacific J. Math.* **103** 9–16. MR0687959

CAI, T. T., MA, Z. and WU, Y. (2013). Sparse PCA: Optimal rates and adaptive estimation. *Ann. Statist.* **41** 3074–3110. MR3161458 https://doi.org/10.1214/13-AOS1178

COOK, R. D. and CRITCHLEY, F. (2000). Identifying regression outliers and mixtures graphically. *J. Amer. Statist. Assoc.* **95** 781–794. MR1803878 https://doi.org/10.2307/2669462

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

FERRÉ, L. (1998). Determining the dimension in sliced inverse regression and related methods. *J. Amer. Statist. Assoc.* **93** 132–140. MR1614604 https://doi.org/10.2307/2669610

GAO, C., MA, Z., REN, Z. and ZHOU, H. H. (2015). Minimax estimation in sparse canonical correlation analysis. *Ann. Statist.* **43** 2168–2197. MR3396982 https://doi.org/10.1214/15-AOS1332

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

HSING, T. and CARROLL, R. J. (1992). An asymptotic theory for sliced inverse regression. *Ann. Statist.* **20** 1040–1061. MR1165605 https://doi.org/10.1214/aos/1176348669

HUANG, D. (2020). Reliable and Flexible Inference for High Dimensional Data. PhD thesis, Harvard Univ. MR4257451

---

*MSC2020 subject classifications.* Primary 62J02; secondary 62C20, 62H12.

*Key words and phrases.* Central space, sufficient dimension reduction, sliced inverse regression, structural dimension, minimax lower bound, multiple-index model.

HUANG, D., TIAN, S. and LIN, Q. (2026). Supplement to "On the Structural Dimension of Sliced Inverse Regression." https://doi.org/10.1214/25-AOS2505SUPP

JIANG, B. and LIU, J. S. (2014). Variable selection for general index models via sliced inverse regression. *Ann. Statist.* **42** 1751–1786. MR3262467 https://doi.org/10.1214/14-AOS1233

LI, B. (2017). *Sufficient Dimension Reduction: Methods and Applications with R*, Chapman and Hall, London. https://doi.org/10.1201/9781315119427

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

LI, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86** 316–327. MR1137117

LI, K.-C. (1992). On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *J. Amer. Statist. Assoc.* **87** 1025–1039. MR1209564

LI, K.-C. and DUAN, N. (1989). Regression analysis under link violation. *Ann. Statist.* **17** 1009–1052. MR1015136 https://doi.org/10.1214/aos/1176347254

LI, L. (2007). Sparse sufficient dimension reduction. *Biometrika* **94** 603–613. MR2410011 https://doi.org/10.1093/biomet/asm044

LI, Y. and ZHU, L.-X. (2007). Asymptotics for sliced average variance estimation. *Ann. Statist.* **35** 41–69. MR2332268 https://doi.org/10.1214/009053606000001091

LIN, Q., LI, X., HUANG, D. and LIU, J. S. (2021). On the optimality of sliced inverse regression in high dimensions. *Ann. Statist.* **49** 1–20. MR4206667 https://doi.org/10.1214/19-AOS1813

LIN, Q., ZHAO, Z. and LIU, J. S. (2018). On consistency and sparsity for sliced inverse regression in high dimensions. *Ann. Statist.* **46** 580–610. MR3782378 https://doi.org/10.1214/17-AOS1561

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

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

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

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

TAN, K., SHI, L. and YU, Z. (2020). Sparse SIR: Optimal rates and adaptive estimation. *Ann. Statist.* **48** 64–85. MR4065153 https://doi.org/10.1214/18-AOS1791

WU, Y. and LI, L. (2011). Asymptotic properties of sufficient dimension reduction with a diverging number of predictors. *Statist. Sinica* **21** 707–730. MR2829852 https://doi.org/10.5705/ss.2011.031a

XIA, Y., TONG, H., LI, W. K. and ZHU, L.-X. (2009). An adaptive estimation of dimension reduction space. In *Exploration of a Nonlinear World: An Appreciation of Howell Tong's Contributions to Statistics* 299–346. World Scientific, Singapore.

YANG, S. S. (1977). General distribution theory of the concomitants of order statistics. *Ann. Statist.* **5** 996–1002. MR0501519

YE, Z. and WEISS, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *J. Amer. Statist. Assoc.* **98** 968–979. MR2041485 https://doi.org/10.1198/016214503000000927

YU, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam* 423–435. Springer, New York. MR1462963

YU, Z., DONG, Y. and SHAO, J. (2016). On marginal sliced inverse regression for ultrahigh dimensional model-free feature selection. *Ann. Statist.* **44** 2594–2623. MR3576555 https://doi.org/10.1214/15-AOS1424

ZHU, L., MIAO, B. and PENG, H. (2006). On sliced inverse regression with high-dimensional covariates. *J. Amer. Statist. Assoc.* **101** 630–643. MR2281245 https://doi.org/10.1198/016214505000001285

ZHU, L. X. and NG, K. W. (1995). Asymptotics of sliced inverse regression. *Statist. Sinica* **5** 727–736. MR1347616

*The Annals of Statistics*
2026, Vol. 54, No. 1, 48–73
https://doi.org/10.1214/25-AOS2524

# OPTIMAL AND EXACT RECOVERY ON THE GENERAL NONUNIFORM HYPERGRAPH STOCHASTIC BLOCK MODEL

BY IOANA DUMITRIU[a] AND HAI-XIAO WANG[b]

*Department of Mathematics, University of California,* [a]*idumitriu@ucsd.edu,* [b]*h9wang@ucsd.edu*

Consider the community detection problem in random hypergraphs under the nonuniform hypergraph stochastic block model (HSBM), where each hyperedge appears independently with some given probability depending only on the labels of its vertices. We establish, for the first time in the literature, a sharp threshold for exact recovery for the nonuniform, multiple-community case, subject to only minor constraints. One crucial point here is that, by aggregating information from all the uniform layers, we may obtain exact recovery, even in cases when this is impossible if each layer were considered alone. In addition, for the almost exact recovery regime, we prove a wide-ranging, information-theoretic lower bound on the number of misclassified vertices *for any algorithm*. The bound depends on a *generalized Chernoff–Hellinger* divergence involving model parameters. We provide two efficient algorithms (one for the minimal information case, another for the full information case) which successfully achieve exact recovery when above the threshold and attain the lowest possible mismatch ratio when the exact recovery is impossible. The theoretical analysis of our algorithms relies on the concentration and regularization of the adjacency matrix for nonuniform random hypergraphs, which could be of independent interest. We also address some open problems regarding parameter knowledge and estimation.

## REFERENCES

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

[2] ABBE, E., BANDEIRA, A. S. and HALL, G. (2016). Exact recovery in the stochastic block model. *IEEE Trans. Inf. Theory* **62** 471–487. MR3447993 https://doi.org/10.1109/TIT.2015.2490670

[3] ABBE, E., FAN, J. and WANG, K. (2022). An $\ell_p$ theory of PCA and spectral clustering. *Ann. Statist.* **50** 2359–2385. MR4474494 https://doi.org/10.1214/22-aos2196

[4] ABBE, E., FAN, J., WANG, K. and ZHONG, Y. (2020). Entrywise eigenvector analysis of random matrices with low expected rank. *Ann. Statist.* **48** 1452–1474. MR4124330 https://doi.org/10.1214/19-AOS1854

[5] ABBE, E. and SANDON, C. (2015). Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In 2015 *IEEE 56th Annual Symposium on Foundations of Computer Science—FOCS* 2015 670–688. IEEE Computer Soc., Los Alamitos, CA. MR3473334 https://doi.org/10.1109/FOCS.2015.47

[6] AGTERBERG, J. and ZHANG, A. R. (2025). Estimating higher-order mixed memberships via the $\ell_{2,\infty}$ tensor perturbation bound. *J. Amer. Statist. Assoc.* **120** 1214–1224. MR4917290 https://doi.org/10.1080/01621459.2024.2404265

[7] ALALUUSUA, K., AVRACHENKOV, K., VINAY KUMAR, B. R. and LESKELÄ, L. (2023). Multilayer hypergraph clustering using the aggregate similarity matrix. In *Algorithms and Models for the Web Graph. Lecture Notes in Computer Science* **13894** 83–98. Springer, Cham. MR4606811 https://doi.org/10.1007/978-3-031-32296-9_6

[8] ANGELINI, M. C., CALTAGIRONE, F., KRZAKALA, F. and ZDEBOROVÁ, L. (2015). Spectral detection on sparse hypergraphs. In 2015 53*rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)* 66–73. https://doi.org/10.1109/ALLERTON.2015.7446987

[9] ARIAS-CASTRO, E. and VERZELEN, N. (2014). Community detection in dense random networks. *Ann. Statist.* **42** 940–969. MR3210992 https://doi.org/10.1214/14-AOS1208

---

*MSC2020 subject classifications.* Primary 60B20; secondary 05C80.

*Key words and phrases.* Random matrices, random hypergraphs, stochastic block model, strong consistency.

[10] BÓNA, M. (2017). *A Walk Through Combinatorics*. World Scientific, Hackensack, NJ. MR3560666

[11] CHEN, S., LIU, S. and MA, Z. (2022). Global and individualized community detection in inhomogeneous multilayer networks. *Ann. Statist.* **50** 2664–2693. MR4500621 https://doi.org/10.1214/22-aos2202

[12] CHIEN, I., LIN, C.-Y. and WANG, I. (2019). On the minimax misclassification ratio of hypergraph community detection. *IEEE Trans. Inf. Theory* **65** 8095–8118. MR4038918 https://doi.org/10.1109/TIT.2019.2928301

[13] CHODROW, P., EIKMEIER, N. and HADDOCK, J. (2023). Nonbacktracking spectral clustering of nonuniform hypergraphs. *SIAM J. Math. Data Sci.* **5** 251–279. MR4580339 https://doi.org/10.1137/22M1494713

[14] COJA-OGHLAN, A. (2010). Graph partitioning via adaptive spectral techniques. *Combin. Probab. Comput.* **19** 227–284. MR2593622 https://doi.org/10.1017/S0963548309990514

[15] DUMITRIU, I. and WANG, H.-X (2026). Supplement to "Optimal and exact recovery on the general nonuniform Hypergraph Stochastic Block Model." https://doi.org/10.1214/25-AOS2524SUPP

[16] DUMITRIU, I., WANG, H.-X. and ZHU, Y. (2025). Partial recovery and weak consistency in the non-uniform hypergraph stochastic block model. *Combin. Probab. Comput.* **34** 1–51. MR4837030 https://doi.org/10.1017/s0963548324000166

[17] FEIGE, U. and OFEK, E. (2005). Spectral techniques applied to sparse random graphs. *Random Structures Algorithms* **27** 251–275. MR2155709 https://doi.org/10.1002/rsa.20089

[18] GAO, C., MA, Z., ZHANG, A. Y. and ZHOU, H. H. (2017). Achieving optimal misclassification proportion in stochastic block models. *J. Mach. Learn. Res.* **18** Paper No. 60, 45. MR3687603

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

[20] GAUDIO, J. and JOSHI, N. (2023). Community detection in the hypergraph SBM: Optimal recovery given the similarity matrix. In *The Thirty Sixth Annual Conference on Learning Theory* 469–510.

[21] GHOSHDASTIDAR, D. and DUKKIPATI, A. (2014). Consistency of spectral partitioning of uniform hypergraphs under planted partition model. *Adv. Neural Inf. Process. Syst.* **27**.

[22] GHOSHDASTIDAR, D. and DUKKIPATI, A. (2017). Consistency of spectral hypergraph partitioning under planted partition model. *Ann. Statist.* **45** 289–315. MR3611493 https://doi.org/10.1214/16-AOS1453

[23] GU, Y. and PANDEY, A. (2024). Community detection in the hypergraph stochastic block model and reconstruction on hypertrees. In *Proceedings of Thirty Seventh Conference on Learning Theory* **247** 1–38.

[24] GU, Y. and POLYANSKIY, Y. (2023). Weak recovery threshold for the Hypergraph Stochastic Block Model. In *Proceedings of Thirty Sixth Conference on Learning Theory.* **195** 885–920.

[25] HAN, R., LUO, Y., WANG, M. and ZHANG, A. R. (2022). Exact clustering in tensor block model: Statistical optimality and computational limit. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 1666–1698. MR4515554

[26] HILLAR, C. J. and LIM, L.-H. (2013). Most tensor problems are NP-hard. *J. ACM* **60** Art. 45, 39. MR3144915 https://doi.org/10.1145/2512329

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

[28] HU, J. and WANG, M. (2023). Multiway spherical clustering via degree-corrected tensor block models. *IEEE Trans. Inf. Theory* **69** 3880–3919. MR4604474 https://doi.org/10.1109/tit.2023.3239521

[29] KE, Z. T., SHI, F. and XIA, D. (2019). Community detection for hypergraph networks via regularized tensor power iteration. arXiv preprint, arXiv:1909.06503. https://doi.org/10.48550/arXiv.1909.06503

[30] KIM, C., BANDEIRA, A. S. and GOEMANS, M. X. (2018). Stochastic Block Model for Hypergraphs: Statistical limits and a semidefinite programming approach. arXiv preprint, arXiv:1807.02884.

[31] LAI, T. L. and ROBBINS, H. (1985). Asymptotically efficient adaptive allocation rules. *Adv. in Appl. Math.* **6** 4–22. MR0776826 https://doi.org/10.1016/0196-8858(85)90002-8

[32] LE, C. M., LEVINA, E. and VERSHYNIN, R. (2017). Concentration and regularization of random graphs. *Random Structures Algorithms* **51** 538–561. MR3689343 https://doi.org/10.1002/rsa.20713

[33] LEI, J. and RINALDO, A. (2015). Consistency of spectral clustering in stochastic block models. *Ann. Statist.* **43** 215–237. MR3285605 https://doi.org/10.1214/14-AOS1274

[34] MA, Z. and NANDY, S. (2023). Community detection with contextual multilayer networks. *IEEE Trans. Inf. Theory* **69** 3203–3239. MR4581286 https://doi.org/10.1109/tit.2023.3238352

[35] MOSSEL, E., NEEMAN, J. and SLY, A. (2016). Consistency thresholds for the planted bisection model. *Electron. J. Probab.* **21** Paper No. 21, 24. MR3485363 https://doi.org/10.1214/16-EJP4185

[36] NEWMAN, M. E., WATTS, D. J. and STROGATZ, S. H. (2002). Random graph models of social networks. *Proc. Natl. Acad. Sci. USA* **99** 2566–2572.

[37] NG, A. Y., JORDAN, M. I. and WEISS, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems* 849–856.

[38] PAL, S. and ZHU, Y. (2021). Community detection in the sparse hypergraph stochastic block model. *Random Structures Algorithms* **59** 407–463. MR4295569 https://doi.org/10.1002/rsa.21006

[39] STEPHAN, L. and ZHU, Y. (2024). Sparse random hypergraphs: Non-backtracking spectra and community detection. *Inf. Inference* **13** Paper No. iaae004, 70. MR4707550 https://doi.org/10.1093/imaiai/iaae004

[40] WANG, H.-X. (2023). Information-theoretic limits and strong consistency on the binary non-uniform hypergraph stochastic block model. arXiv preprint.

[41] YUN, S.-Y. and PROUTIERE, A. (2014). Community detection via random and adaptive sampling. In *Conference on Learning Theory* 138–175. https://doi.org/10.48550/arXiv.1402.3072

[42] YUN, S.-Y. and PROUTIERE, A. (2016). Optimal cluster recovery in the labeled stochastic block model. *Adv. Neural Inf. Process. Syst.* **29**. https://doi.org/10.48550/arXiv.1510.05956

[43] ZHANG, A. Y. (2024). Fundamental limits of spectral clustering in stochastic block models. *IEEE Trans. Inf. Theory* **70** 7320–7348. MR4810402 https://doi.org/10.1109/tit.2024.3425581

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

[45] ZHANG, Q. and TAN, V. Y. F. (2023). Exact recovery in the general hypergraph stochastic block model. *IEEE Trans. Inf. Theory* **69** 453–471. MR4544967 https://doi.org/10.1109/tit.2022.3205959

[46] ZHEN, Y. and WANG, J. (2023). Community detection in general hypergraph via graph embedding. *J. Amer. Statist. Assoc.* **118** 1620–1629. MR4646593 https://doi.org/10.1080/01621459.2021.2002157

*The Annals of Statistics*
2026, Vol. 54, No. 1, 74–92
https://doi.org/10.1214/25-AOS2535

# PSEUDO-LIKELIHOOD-BASED $M$-ESTIMATION OF RANDOM GRAPHS WITH DEPENDENT EDGES AND PARAMETER VECTORS OF INCREASING DIMENSION

BY JONATHAN R. STEWART[1,a] AND MICHAEL SCHWEINBERGER[2,b]

[1]*Department of Statistics, Florida State University,* [a]*jrstewart@fsu.edu*
[2]*Department of Statistics, The Pennsylvania State University,* [b]*mus47@psu.edu*

An important question in statistical network analysis is how to estimate models of discrete and dependent network data with intractable likelihood functions, without sacrificing computational scalability and statistical guarantees. We demonstrate that scalable estimation of random graph models with dependent edges is possible, by establishing convergence rates of pseudo-likelihood-based $M$-estimators for discrete undirected graphical models with exponential parameterizations and parameter vectors of increasing dimension in single-observation scenarios. We highlight the impact of two complex phenomena on the convergence rate: phase transitions and model near-degeneracy. The main results have possible applications to discrete and dependent network, spatial, and temporal data. To showcase convergence rates, we introduce a novel class of generalized $\beta$-models with dependent edges and parameter vectors of increasing dimension, which leverage additional structure in the form of overlapping subpopulations to control dependence. We establish convergence rates of pseudo-likelihood-based $M$-estimators for generalized $\beta$-models in dense- and sparse-graph settings.

## REFERENCES

[1] AMINI, A. A., CHEN, A., BICKEL, P. J. and LEVINA, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *Ann. Statist.* **41** 2097–2122. MR3127859 https://doi.org/10.1214/13-AOS1138

[2] BHATTACHARYA, B. B. and MUKHERJEE, S. (2018). Inference in Ising models. *Bernoulli* **24** 493–525. MR3706767 https://doi.org/10.3150/16-BEJ886

[3] BICKEL, P. J. and CHEN, A. (2009). A nonparametric view of network models and Newman-Girvan and other modularities. In *Proceedings of the National Academy of Sciences* **106** 21068–21073.

[4] BROWN, L. D. (1986). *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **9**. IMS, Hayward, CA. MR0882001

[5] CHATTERJEE, S. (2007). Estimation in spin glasses: A first step. *Ann. Statist.* **35** 1931–1946. MR2363958 https://doi.org/10.1214/009053607000000109

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

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

[8] CHAZOTTES, J.-R., COLLET, P., KÜLSKE, C. and REDIG, F. (2007). Concentration inequalities for random fields via coupling. *Probab. Theory Related Fields* **137** 201–225. MR2278456 https://doi.org/10.1007/s00440-006-0026-1

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

[10] COMETS, F. (1992). On consistency of a class of estimators for exponential families of Markov random fields on the lattice. *Ann. Statist.* **20** 455–468. MR1150354 https://doi.org/10.1214/aos/1176348532

---

*MSC2020 subject classifications.* Primary 05C80; secondary 62B05, 62F10, 91D30.

*Key words and phrases.* Markov random fields, graphical models, conditional independence, statistical exponential families, phase transitions, model near-degeneracy.

[11] DAWID, A. P. (1979). Conditional independence in statistical theory. *J. Roy. Statist. Soc. Ser. B, Methodol.* **41** 1–31. MR0535541

[12] FRANK, O. and STRAUSS, D. (1986). Markov graphs. *J. Amer. Statist. Assoc.* **81** 832–842. MR0860518

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

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

[15] HANDCOCK, M. S. (2003). Statistical models for social networks: Inference and degeneracy. In *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers* (R. Breiger, K. Carley and P. Pattison, eds.) 1–12. National Academies Press, Washington, DC.

[16] HILLAR, C. J. and WIBISONO, A. (2015). A Hadamard-type lower bound for symmetric diagonally dominant positive matrices. *Linear Algebra Appl.* **472** 135–141. MR3314371 https://doi.org/10.1016/j.laa.2015.01.037

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

[18] HOLLAND, P. W. and LEINHARDT, S. (1981). An exponential family of probability distributions for directed graphs. *J. Amer. Statist. Assoc.* **76** 33–65. MR0608176

[19] LAURITZEN, S., RINALDO, A. and SADEGHI, K. (2018). Random networks, graphical models and exchangeability. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 481–508. MR3798875 https://doi.org/10.1111/rssb.12266

[20] LINDVALL, T. (2002). *Lectures on the Coupling Method*. Dover, Mineola, NY. MR1924231

[21] MUKHERJEE, R., MUKHERJEE, S. and SEN, S. (2018). Detection thresholds for the $\beta$-model on sparse graphs. *Ann. Statist.* **46** 1288–1317. MR3798004 https://doi.org/10.1214/17-AOS1585

[22] MUKHERJEE, S. (2020). Degeneracy in sparse ERGMs with functions of degrees as sufficient statistics. *Bernoulli* **26** 1016–1043. MR4058359 https://doi.org/10.3150/19-BEJ1135

[23] ORTEGA, J. M. and RHEINBOLDT, W. C. (2000). *Iterative Solution of Nonlinear Equations in Several Variables*. *Classics in Applied Mathematics* **30**. SIAM, Philadelphia, PA. MR1744713 https://doi.org/10.1137/1.9780898719468

[24] PORTNOY, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.* **16** 356–366. MR0924876 https://doi.org/10.1214/aos/1176350710

[25] RINALDO, A., PETROVIĆ, S. and FIENBERG, S. E. (2013). Maximum likelihood estimation in the $\beta$-model. *Ann. Statist.* **41** 1085–1110. MR3113804 https://doi.org/10.1214/12-AOS1078

[26] ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915. MR2893856 https://doi.org/10.1214/11-AOS887

[27] SCHWEINBERGER, M. (2011). Instability, sensitivity, and degeneracy of discrete exponential families. *J. Amer. Statist. Assoc.* **106** 1361–1370. MR2896841 https://doi.org/10.1198/jasa.2011.tm10747

[28] SCHWEINBERGER, M. and STEWART, J. (2020). Concentration and consistency results for canonical and curved exponential-family models of random graphs. *Ann. Statist.* **48** 374–396. MR4065166 https://doi.org/10.1214/19-AOS1810

[29] STEWART, J. R and SCHWEINBERGER, M. (2026). Supplement to "Pseudo-likelihood-based $M$-estimation of random graphs with dependent edges and parameter vectors of increasing dimension." https://doi.org/10.1214/25-AOS2535SUPP

[30] VAN DEN BERG, J. and MAES, C. (1994). Disagreement percolation in the study of Markov fields. *Ann. Probab.* **22** 749–763. MR1288130

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

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

*The Annals of Statistics*
2026, Vol. 54, No. 1, 93–119
https://doi.org/10.1214/25-AOS2539

# NEAR OPTIMAL SAMPLE COMPLEXITY FOR MATRIX AND TENSOR NORMAL MODELS VIA GEODESIC CONVEXITY

BY COLE FRANKS[1,a], RAFAEL OLIVEIRA[2,b], AKSHAY RAMACHANDRAN[3,c] AND MICHAEL WALTER[4,d]

[1]*Department of Mathematics, Massachusetts Institute of Technology,* [a]*franks@mit.edu*
[2]*Cheriton School of Computer Science, University of Waterloo,* [b]*rafael@uwaterloo.ca*
[3]*Computer Science Department, University of British Columbia,* [c]*aramach@cs.ubc.ca*
[4]*Ludwig-Maximilians-Universität München,* [d]*michael.walter@lmu.de*

The matrix normal model, that is, the family of Gaussian matrix-variate distributions whose covariance matrices are the Kronecker product of two lower-dimensional factors, is frequently used to model matrix-variate data. The tensor normal model generalizes this family to Kronecker products of three or more factors. We study the estimation of the Kronecker factors of the covariance matrix in the matrix and tensor normal models.

For the above models, we show that the maximum likelihood estimator (MLE) achieves *nearly optimal nonasymptotic sample complexity* and *nearly tight error rates* in the Fisher–Rao and Thompson metrics. In contrast to prior work, our results do not rely on the factors being well conditioned or sparse, nor do we need to assume an accurate enough initial guess. For the matrix normal model, all our bounds are minimax optimal up to logarithmic factors, and for the tensor normal model our bounds for the largest factor and for overall covariance matrix are minimax optimal up to constant factors provided there are enough samples for any estimator to obtain constant Frobenius error. In the same regimes as our sample complexity bounds, we show that the flip-flop algorithm, a practical and widely used iterative procedure to compute the MLE, converges linearly with high probability.

Our main technical insight is that, given enough samples, the negative log-likelihood function is *strongly geodesically convex* in the geometry on positive-definite matrices induced by the Fisher information metric. This strong convexity is determined by the expansion of certain random quantum channels.

## REFERENCES

[1] BAČÁK, M. (2014). *Convex Analysis and Optimization in Hadamard Spaces*. *De Gruyter Series in Nonlinear Analysis and Applications* **22**. de Gruyter, Berlin. MR3241330 https://doi.org/10.1515/9783110361629

[2] BHATIA, R. (2009). *Positive Definite Matrices*. *Princeton Series in Applied Mathematics*. Princeton Univ. Press, Princeton, NJ. MR2284176

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

[4] BROWN, P. J., KENWARD, M. G. and BASSETT, E. E. (2001). Bayesian discrimination with longitudinal data. *Biostatistics* **2** 417–432.

[5] BÜRGISSER, P., FRANKS, C., GARG, A., OLIVEIRA, R., WALTER, M. and WIGDERSON, A. (2018). Efficient algorithms for tensor scaling, quantum marginals, and moment polytopes. In 59*th Annual IEEE Symposium on Foundations of Computer Science—FOCS* 2018 883–897. IEEE Computer Soc., Los Alamitos, CA. MR3899650 https://doi.org/10.1109/FOCS.2018.00088

[6] BÜRGISSER, P., FRANKS, C., GARG, A., OLIVEIRA, R., WALTER, M. and WIGDERSON, A. (2019). Towards a theory of non-commutative optimization: Geodesic 1st and 2nd order methods for moment maps and polytopes. In 2019 *IEEE* 60*th Annual Symposium on Foundations of Computer Science* 845–861. IEEE Comput. Soc. Press, Los Alamitos, CA. MR4228202 https://doi.org/10.1109/FOCS.2019.00055

---

*MSC2020 subject classifications.* Primary 62F12; secondary 62F30.

*Key words and phrases.* Covariance estimation, matrix normal model, tensor normal model, maximum likelihood estimation, geodesic convexity, operator scaling, quantum expansion.

[7] Bürgisser, P., Garg, A., Oliveira, R., Walter, M. and Wigderson, A. (2018). Alternating minimization, scaling algorithms, and the null-cone problem from invariant theory. In 9*th Innovations in Theoretical Computer Science*. *LIPIcs*. *Leibniz Int*. *Proc*. *Inform*. **94** Art. No. 24, 20 pp. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern. MR3761760

[8] Čencov, N. N. (1978). Algebraic foundation of mathematical statistics. *Math*. *Operforsch*. *Stat*., *Ser. Stat*. **9** 267–276. MR0512264 https://doi.org/10.1080/02331887808801428

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

[10] Franks, C., Oliveira, R., Ramachandran, A. and Walter, M. (2026). Supplement to "Near optimal sample complexity for matrix and tensor normal models via geodesic convexity." https://doi.org/10.1214/25-AOS2539SUPP

[11] Franks, W. C. and Moitra, A. (2020). Rigorous guarantees for Tyler's M-estimator via quantum expansion. In *Conference on Learning Theory* 1601–1632. PMLR.

[12] Garg, A., Gurvits, L., Oliveira, R. and Wigderson, A. (2020). Operator scaling: Theory and applications. *Found*. *Comput*. *Math*. **20** 223–290. MR4081171 https://doi.org/10.1007/s10208-019-09417-z

[13] Hastings, M. B. (2007). Random unitaries give quantum expanders. *Phys*. *Rev*. *A* (3) **76** 032315, 11 pp. MR2486279 https://doi.org/10.1103/PhysRevA.76.032315

[14] Kwok, T. C., Lau, L. C. and Ramachandran, A. (2019). Spectral analysis of matrix scaling and operator scaling. In 2019 *IEEE* 60*th Annual Symposium on Foundations of Computer Science* 1184–1204. IEEE Comput. Soc. Press, Los Alamitos, CA. MR4228221 https://doi.org/10.1109/FOCS.2019.00074

[15] Lyu, X., Wei Sun, W., Wang, Z., Liu, H., Yang, J. and Chengk, G. (2020). Tensor graphical model: Nonconvex optimization and statistical inference. *IEEE Trans*. *Pattern Anal*. *Mach*. *Intell*. **8** 2024–2037.

[16] 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

[17] Mitchell, T. M., Hutchinson, R., Niculescu, R. S., Pereira, F., Wang, X., Just, M. and Newman, S. (2004). Learning to decode cognitive states from brain images. *Mach*. *Learn*. **57** 145–175.

[18] Pisier, G. (2012). Grothendieck's theorem, past and present. *Bull*. *Amer*. *Math*. *Soc*. (*N*.*S*.) **49** 237–323. MR2888168 https://doi.org/10.1090/S0273-0979-2011-01348-9

[19] Pisier, G. (2014). Quantum expanders and geometry of operator spaces. *J*. *Eur*. *Math*. *Soc*. (*JEMS*) **16** 1183–1219. MR3226740 https://doi.org/10.4171/JEMS/458

[20] Skovgaard, L. T. (1984). A Riemannian geometry of the multivariate normal model. *Scand*. *J*. *Stat*. **11** 211–223. MR0793171

[21] Tsiligkaridis, T., Hero, A. O. III and Zhou, S. (2013). On convergence of Kronecker graphical lasso algorithms. *IEEE Trans*. *Signal Process*. **61** 1743–1755. MR3038387 https://doi.org/10.1109/TSP.2013.2240157

[22] Werner, K., Jansson, M. and Stoica, P. (2008). On estimation of covariance matrices with Kronecker product structure. *IEEE Trans*. *Signal Process*. **56** 478–491. MR2445531 https://doi.org/10.1109/TSP.2007.907834

[23] Wiesel, A. (2012). Geodesic convexity and covariance estimation. *IEEE Trans*. *Signal Process*. **60** 6182–6189. MR3006411 https://doi.org/10.1109/TSP.2012.2218241

[24] Xu, P., Zhang, T. and Gu, Q. (2017). Efficient algorithm for sparse tensor-variate Gaussian graphical models via gradient descent. In *Artificial Intelligence and Statistics* 923–932. PMLR.

[25] Zhou, S. (2014). Gemini: Graph estimation with matrix variate normal instances. *Ann*. *Statist*. **42** 532–562. MR3210978 https://doi.org/10.1214/13-AOS1187

*The Annals of Statistics*
2026, Vol. 54, No. 1, 120–145
https://doi.org/10.1214/25-AOS2549

# SCALABLE INFERENCE IN FUNCTIONAL LINEAR REGRESSION WITH STREAMING DATA

BY JINHAN XIE[1,a], ENZE SHI[2,b], PEIJUN SANG[3,e], ZUOFENG SHANG[4,f], BEI JIANG[2,c] AND LINGLONG KONG[2,d]

[1] *Yunnan Key Laboratory of Statistical Modeling and Data Analysis, Yunnan University,* [a]*jinhanxie@ynu.edu.cn*
[2] *Department of Mathematical and Statistical Sciences, University of Alberta,* [b]*eshi@ualberta.ca,* [c]*bei1@ualberta.ca,* [d]*lkong@ualberta.ca*
[3] *Department of Statistics and Actuarial Science, University of Waterloo,* [e]*peijun.sang@uwaterloo.ca*
[4] *Department of Mathematical Sciences, New Jersey Institute of Technology,* [f]*zshang@njit.edu*

Traditional static functional data analysis is facing new challenges due to streaming data, where data constantly flows in. A major challenge is that storing such an ever-increasing amount of data in memory is nearly infeasible. In addition, existing inferential tools in online learning are mainly developed for finite-dimensional problems, while inference methods for functional data are focused on the offline settings. In this paper, we focus on the online estimation of functional linear regression with a scalar response and a functional covariate. To tackle these issues in this setting, we develop a functional stochastic gradient descent algorithm and propose an online bootstrap resampling procedure that enables us to perform the local (and global) inference for the slope function in functional linear regression. In particular, the proposed estimation and inference procedures use only one pass over the data; thus they are easy to implement and suitable to the situation where data arrive in a streaming manner. Furthermore, we establish the convergence rate and the asymptotic distribution for the proposed slope function estimator. The proposed perturbed estimator from the bootstrap procedure is shown to enjoy the same theoretical properties, which provide the theoretical justification for our online inference tool. As far as we know, this is the first inference result on functional linear regression with streaming data. Numerical studies are conducted to demonstrate the finite-sample performance of our proposed procedure.

## REFERENCES

[1] BORDES, A., BOTTOU, L. and GALLINARI, P. (2009). SGD-QN: Careful quasi-Newton stochastic gradient descent. *J. Mach. Learn. Res.* **10** 1737–1754. MR2534877

[2] CAI, T. T. and HALL, P. (2006). Prediction in functional linear regression. *Ann. Statist.* **34** 2159–2179. MR2291496 https://doi.org/10.1214/009053606000000830

[3] CAI, T. T. and YUAN, M. (2012). Minimax and adaptive prediction for functional linear regression. *J. Amer. Statist. Assoc.* **107** 1201–1216. MR3010906 https://doi.org/10.1080/01621459.2012.716337

[4] CALANDRIELLO, D., LAZARIC, A. and VALKO, M. (2017). Efficient second-order online kernel learning with adaptive embedding. *Adv. Neural Inf. Process. Syst.* **30**.

[5] 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

[6] CHEN, X., TANG, B., FAN, J. and GUO, X. (2022). Online gradient descent algorithms for functional data learning. *J. Complexity* **70** Paper No. 101635, 14 pp. MR4388513 https://doi.org/10.1016/j.jco.2021.101635

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

---

*MSC2020 subject classifications.* Primary 62R10; secondary 62G20.

*Key words and phrases.* Online learning, reproducing kernel Hilbert space, functional data, bootstrap, stochastic gradient descent.

[8] Cucker, F. and Smale, S. (2002). On the mathematical foundations of learning. *Bull. Amer. Math. Soc.* (*N.S.*) **39** 1–49. MR1864085 https://doi.org/10.1090/S0273-0979-01-00923-5

[9] Dai, X., Müller, H.-G. and Yao, F. (2017). Optimal Bayes classifiers for functional data and density ratios. *Biometrika* **104** 545–560. MR3694582 https://doi.org/10.1093/biomet/asx024

[10] Davidson, C. I., Phalen, R. F. and Solomon, P. A. (2005). Airborne particulate matter and human health: A review. *Aerosol Sci. Technol.* **39** 737–749.

[11] Dette, H., Kokot, K. and Aue, A. (2020). Functional data analysis in the Banach space of continuous functions. *Ann. Statist.* **48** 1168–1192. MR4102692 https://doi.org/10.1214/19-AOS1842

[12] 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

[13] Dockery, D. W., Schwartz, J. and Spengler, J. D. (1992). Air pollution and daily mortality: Associations with particulates and acid aerosols. *Environ. Res.* **59** 362–373.

[14] Fang, Y. (2019). Scalable statistical inference for averaged implicit stochastic gradient descent. *Scand. J. Stat.* **46** 987–1002. MR4033800 https://doi.org/10.1111/sjos.12378

[15] 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, 21 pp. MR3899780

[16] Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis*: *Theory and Practice*. *Springer Series in Statistics*. Springer, New York. MR2229687

[17] Gaillard, P. and Gerchinovitz, S. (2015). A chaining algorithm for online nonparametric regression. In *Conference on Learning Theory* 764–796. PMLR.

[18] 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

[19] Gu, H. and Lafferty, J. (2012). Sequential nonparametric regression. In *Proceedings of the* 29*th International Coference on International Conference on Machine Learning* 387–394.

[20] Guo, X., Guo, Z.-C. and Shi, L. (2023). Capacity dependent analysis for functional online learning algorithms. *Appl. Comput. Harmon. Anal.* **67** Paper No. 101567, 30 pp. MR4612182 https://doi.org/10.1016/j.acha.2023.06.002

[21] Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.* **35** 70–91. MR2332269 https://doi.org/10.1214/009053606000000957

[22] Hazan, E., Agarwal, A. and Kale, S. (2007). Logarithmic regret algorithms for online convex optimization. *Mach. Learn.* **69** 169–192.

[23] 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

[24] Lin, N. and Xi, R. (2011). Aggregated estimating equation estimation. *Stat. Interface* **4** 73–83. MR2775250 https://doi.org/10.4310/SII.2011.v4.n1.a8

[25] Liu, M., Shang, Z. and Yang, Y. (2023). Scalable statistical inference in non-parametric least squares. Preprint. Available at arXiv:2310.00881.

[26] Luo, L. and Song, P. X.-K. (2020). Renewable estimation and incremental inference in generalized linear models with streaming data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 69–97. MR4060977

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

[28] 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

[29] Pope III, C. A. and Dockery, D. W. (2006). Health effects of fine particulate air pollution: Lines that connect. *J. Air Waste Manage. Assoc.* **56** 709–742.

[30] Quan, M. and Lin, Z. (2024). Optimal one-pass nonparametric estimation under memory constraint. *J. Amer. Statist. Assoc.* **119** 285–296. MR4713892 https://doi.org/10.1080/01621459.2022.2115374

[31] Ramprasad, P., Li, Y., Yang, Z., Wang, Z., Sun, W. W. and Cheng, G. (2023). Online bootstrap inference for policy evaluation in reinforcement learning. *J. Amer. Statist. Assoc.* **118** 2901–2914. MR4681629 https://doi.org/10.1080/01621459.2022.2096620

[32] Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2168993

[33] Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Ann. Math. Statist.* **22** 400–407. MR0042668 https://doi.org/10.1214/aoms/1177729586

[34] Ruppert, D. (1988). Stochastic approximation. Technical report, Cornell Univ. Operations Research and Industrial Engineering.

[35] Schifano, E. D., Wu, J., Wang, C., Yan, J. and Chen, M.-H. (2016). Online updating of statistical inference in the big data setting. *Technometrics* **58** 393–403. MR3520668 https://doi.org/10.1080/00401706.2016.1142900

[36] SHANG, Z. and CHENG, G. (2013). Local and global asymptotic inference in smoothing spline models. *Ann. Statist.* **41** 2608–2638. MR3161439 https://doi.org/10.1214/13-AOS1164

[37] SHANG, Z. and CHENG, G. (2015). Nonparametric inference in generalized functional linear models. *Ann. Statist.* **43** 1742–1773. MR3357877 https://doi.org/10.1214/15-AOS1322

[38] STIEB, D. M., BURNETT, R. T., SMITH-DOIRON, M., BRION, O., SHIN, H. H. and ECONOMOU, V. (2008). A new multipollutant, no-threshold air quality health index based on short-term associations observed in daily time-series analyses. *J. Air Waste Manage. Assoc.* **58** 435–450.

[39] TARRÈS, P. and YAO, Y. (2014). Online learning as stochastic approximation of regularization paths: Optimality and almost-sure convergence. *IEEE Trans. Inf. Theory* **60** 5716–5735. MR3252416 https://doi.org/10.1109/TIT.2014.2332531

[40] TOULIS, P. and AIROLDI, E. M. (2017). Asymptotic and finite-sample properties of estimators based on stochastic gradients. *Ann. Statist.* **45** 1694–1727. MR3670193 https://doi.org/10.1214/16-AOS1506

[41] XIE, J., SHI, E., SANG, P., SHANG, Z., JIANG, B. and KONG, L. (2026). Supplement to "Scalable inference in functional linear regression with streaming data." https://doi.org/10.1214/25-AOS2549SUPPA, https://doi.org/10.1214/25-AOS2549SUPPB

[42] XUE, D. and YAO, F. (2022). Dynamic penalized splines for streaming data. *Statist. Sinica* **32** 1363–1380. MR4449898 https://doi.org/10.5705/ss.202020.0167

[43] YANG, Y. and YAO, F. (2023). Online estimation for functional data. *J. Amer. Statist. Assoc.* **118** 1630–1644. MR4646594 https://doi.org/10.1080/01621459.2021.2002158

[44] YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005). Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33** 2873–2903. MR2253106 https://doi.org/10.1214/009053605000000660

[45] YUAN, M. and CAI, T. T. (2010). A reproducing kernel Hilbert space approach to functional linear regression. *Ann. Statist.* **38** 3412–3444. MR2766857 https://doi.org/10.1214/09-AOS772

[46] 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.

[47] 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

[48] ZHANG, T. and SIMON, N. (2023). An online projection estimator for nonparametric regression in reproducing kernel Hilbert spaces. *Statist. Sinica* **33** 127–148. MR4527765

[49] ZHOU, D.-X. (2003). Capacity of reproducing kernel spaces in learning theory. *IEEE Trans. Inf. Theory* **49** 1743–1752. MR1985575 https://doi.org/10.1109/TIT.2003.813564

[50] ZHOU, H., YAO, F. and ZHANG, H. (2023). Functional linear regression for discretely observed data: From ideal to reality. *Biometrika* **110** 381–393. MR4588359 https://doi.org/10.1093/biomet/asac053

*The Annals of Statistics*
2026, Vol. 54, No. 1, 146–175
https://doi.org/10.1214/25-AOS2560

# POISSON EMPIRICAL BAYES ESTIMATION: WHEN DOES $g$-MODELING BEAT $f$-MODELING IN THEORY (AND IN PRACTICE)?

BY YANDI SHEN[1,a] AND YIHONG WU[2,b]

[1] *Department of Statistics and Data Science, Carnegie Mellon University,* [a] *yandis@andrew.cmu.edu*
[2] *Department of Statistics and Data Science, Yale University,* [b] *yihong.wu@yale.edu*

Empirical Bayes (EB) is a popular framework for large-scale inference that aims to find data-driven estimators to compete with the Bayesian oracle that knows the true prior. Two principled approaches to EB estimation have emerged over the years: $f$*-modeling*, which constructs an approximate Bayes rule by estimating the marginal distribution of the data, and $g$*-modeling*, which estimates the prior from data and then applies the learned Bayes rule. For the Poisson model, the prototypical examples are the celebrated Robbins estimator and the nonparametric MLE (NPMLE), respectively. It has long been recognized in practice that the Robbins estimator, while being conceptually appealing and computationally simple, lacks robustness and can be easily derailed by "outliers" (data points that were rarely observed before), unlike the NPMLE which provides more stable and interpretable fit thanks to its Bayes form. On the other hand, not only do the existing theories shed little light on this phenomenon, but they all point to the opposite, as both methods have recently been shown optimal in terms of the *regret* (excess over the Bayes risk) for compactly supported and subexponential priors with exact logarithmic factors (Brown, Greenshtein and Ritov (2013) *J. Amer. Statist. Assoc.* **108** 741–749; Polyanskiy and Wu (2021) arXiv:2008.08244).

In this paper we provide a theoretical justification for the superiority of $g$-modeling over $f$-modeling for heavy-tailed data by considering priors with bounded $p$th moment, a class previously studied for the Gaussian model (Jiang and Zhang (2009) *Ann. Statist.* **37** 1647–1684). For the Poisson model with sample size $n$, assuming $p > 1$ (for otherwise triviality arises), we show that with mild regularization, any $g$-modeling method that is Hellinger rate-optimal in density estimation achieves a total regret $\widetilde{\Theta}(n^{\frac{3}{2p+1}})$, which is minimax optimal within logarithmic factors; in particular, the special case of NPMLE succeeds without regularization. In constrast, there exists an $f$-modeling estimator whose density estimation rate is optimal but whose EB regret is suboptimal by a polynomial factor. These results show that the proper Bayes form provides a "general recipe of success" for optimal EB estimation that applies to all $g$-modeling (but not $f$-modeling) methods. As by-products of our analysis, we also obtain: (a) the minimax Hellinger rate of estimating Poisson mixture over the moment class, (b) the characterization of the regret suboptimality of the Robbins estimator, (c) an extension to the compound setting.

## REFERENCES

[1] BARBEHENN, A. and ZHAO, S. D. (2022). A nonparametric regression approach to asymptotically optimal estimation of normal means. Preprint. Available at arXiv:2205.00336.

[2] BROWN, L. D. and GREENSHTEIN, E. (2009). Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *Ann. Statist.* **37** 1685–1704. MR2533468 https://doi.org/10.1214/08-AOS630

---

*MSC2020 subject classifications.* Primary 62C12; secondary 62H30.

*Key words and phrases.* Empirical Bayes, Poisson distribution, mixture model, density estimation, nonparametric MLE, Robbins, heavy tail, regret.

[3] Brown, L. D., Greenshtein, E. and Ritov, Y. (2013). The Poisson compound decision problem revisited. *J. Amer. Statist. Assoc.* **108** 741–749. MR3174656 https://doi.org/10.1080/01621459.2013.771582

[4] Carlin, B. P. and Louis, T. A. (2009). *Bayesian Methods for Data Analysis*, 3rd ed. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. MR2442364

[5] Casella, G. (1985). An introduction to empirical Bayes data analysis. *Amer. Statist.* **39** 83–87. MR0789118 https://doi.org/10.2307/2682801

[6] Chen, J. (2017). Consistency of the MLE under mixture models. *Statist. Sci.* **32** 47–63. MR3634306 https://doi.org/10.1214/16-STS578

[7] Chen, S. F. and Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Comput. Speech Lang.* **13** 359–394.

[8] Daniel, J. and Martin, J. H. (2009). *Speech and Language Processing*, 2nd ed. Prentice Hall, New York.

[9] Doss, N., Wu, Y., Yang, P. and Zhou, H. H. (2023). Optimal estimation of high-dimensional Gaussian location mixtures. *Ann. Statist.* **51** 62–95. MR4564849 https://doi.org/10.1214/22-aos2207

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

[11] Efron, B. (2014). Two modeling strategies for empirical Bayes estimation. *Statist. Sci.* **29** 285–301. MR3264543 https://doi.org/10.1214/13-STS455

[12] Efron, B. (2019). Bayes, oracle Bayes and empirical Bayes. *Statist. Sci.* **34** 177–201. MR3983318 https://doi.org/10.1214/18-STS674

[13] Efron, B. (2024). Empirical Bayes: Concepts and methods. In *Handbook of Bayesian, Fiducial, and Frequentist Inference* 8–34. Chapman & Hall/CRC, Boca Raton.

[14] Efron, B. and Hastie, T. (2021). *Computer Age Statistical Inference—Algorithms, Evidence, and Data Science. Institute of Mathematical Statistics* (*IMS*) *Monographs* **6**. Cambridge Univ. Press, Cambridge. MR4412544

[15] Genovese, C. R. and Wasserman, L. (2000). Rates of convergence for the Gaussian mixture sieve. *Ann. Statist.* **28** 1105–1127. MR1810921 https://doi.org/10.1214/aos/1015956709

[16] Ghosal, S. and van der Vaart, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.* **35** 697–723. MR2336864 https://doi.org/10.1214/009053606000001271

[17] Ghosal, S. and van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.* **29** 1233–1263. MR1873329 https://doi.org/10.1214/aos/1013203453

[18] Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40** 237–264. MR0061330 https://doi.org/10.1093/biomet/40.3-4.237

[19] Han, Y., Jiao, J. and Weissman, T. Local moment matching: A unified methodology for symmetric functional estimation and distribution estimation under Wasserstein distance. In *Proc.* 2018 *Conference on Learning Theory* (*COLT*) 3189–3221.

[20] Han, Y., Niles-Weed, J., Shen, Y. and Wu, Y. (2025). Besting Good–Turing: Optimality of nonparametric maximum likelihood for distribution estimation. Preprint.

[21] Heckman, J. and Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* **52** 271–320. MR0735309 https://doi.org/10.2307/1911491

[22] Heinrich, P. and Kahn, J. (2018). Strong identifiability and optimal minimax rates for finite mixture estimation. *Ann. Statist.* **46** 2844–2870. MR3851757 https://doi.org/10.1214/17-AOS1641

[23] Ho, N. and Nguyen, X. (2016). Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Ann. Statist.* **44** 2726–2755. MR3576559 https://doi.org/10.1214/16-AOS1444

[24] Jana, S., Polyanskiy, Y., Teh, A. Z. and Wu, Y. (2023). Empirical Bayes via ERM and Rademacher complexities: The Poisson model. In *The Thirty Sixth Annual Conference on Learning Theory* 5199–5235. PMLR.

[25] Jana, S., Polyanskiy, Y. and Wu, Y. (2025). Optimal empirical Bayes estimation for the Poisson model via minimum-distance methods. *Inf. Inference* **14** Paper No. iaaf027, 42 pp. MR4968757 https://doi.org/10.1093/imaiai/iaaf027

[26] Jewell, N. P. (1982). Mixtures of exponential distributions. *Ann. Statist.* **10** 479–484. MR0653523

[27] Jiang, W. and Zhang, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist.* **37** 1647–1684. MR2533467 https://doi.org/10.1214/08-AOS638

[28] Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27** 887–906. MR0086464 https://doi.org/10.1214/aoms/1177728066

[29] Kim, A. K. H. (2014). Minimax bounds for estimation of normal mixtures. *Bernoulli* **20** 1802–1818. MR3263090 https://doi.org/10.3150/13-BEJ542

[30] KIM, A. K. H. and GUNTUBOYINA, A. (2022). Minimax bounds for estimating multivariate Gaussian location mixtures. *Electron. J. Stat.* **16** 1461–1484. MR4387848 https://doi.org/10.1214/21-ejs1975

[31] KOENKER, R. and MIZERA, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *J. Amer. Statist. Assoc.* **109** 674–685. MR3223742 https://doi.org/10.1080/01621459.2013.869224

[32] LAMBERT, D. and TIERNEY, L. (1984). Asymptotic properties of maximum likelihood estimates in the mixed Poisson model. *Ann. Statist.* **12** 1388–1399. MR0765931 https://doi.org/10.1214/aos/1176346799

[33] LI, J., GUPTA, S. S. and LIESE, F. (2005). Convergence rates of empirical Bayes estimation in exponential family. *J. Statist. Plann. Inference* **131** 101–115. MR2137529 https://doi.org/10.1016/j.jspi.2003.12.017

[34] LI, J. and SCHMIDT, L. (2017). Robust and proper learning for mixtures of Gaussians via systems of polynomial inequalities. In *Conference on Learning Theory* 1302–1382. PMLR.

[35] LINDSAY, B. G. (1995). *Mixture Models*: *Theory*, *Geometry*, *and Applications*. IMS.

[36] MARITZ, J. S. (1968). On the smooth empirical Bayes approach to testing of hypotheses and the compound decision problem. *Biometrika* **55** 83–100. MR0226779 https://doi.org/10.1093/biomet/55.1.83

[37] MARITZ, J. S. and LWIN, T. (1989). *Empirical Bayes Methods*, 2nd ed. *Monographs on Statistics and Applied Probability* **35**. CRC Press, London. MR1019835

[38] MORRIS, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *J. Amer. Statist. Assoc.* **78** 47–65. MR0696849

[39] ORLITSKY, A. and THEERTHA SURESH, A. (2015). Competitive distribution estimation: Why is Good-Turing good. *Adv. Neural Inf. Process. Syst.* **28** 2936–2944.

[40] PARK, H. and PARK, J. (2022). Poisson mean vector estimation with nonparametric maximum likelihood estimation and application to protein domain data. *Electron. J. Stat.* **16** 3789–3835. MR4451378 https://doi.org/10.1214/22-ejs2029

[41] PENSKY, M. (1999). Nonparametric empirical Bayes estimation via wavelets. In *Bayesian Inference in Wavelet-Based Models*. *Lect. Notes Stat.* **141** 323–340. Springer, New York. MR1699853 https://doi.org/10.1007/978-1-4612-0567-8_20

[42] PFANZAGL, J. (1988). Consistency of maximum likelihood estimators for certain nonparametric families, in particular: Mixtures. *J. Statist. Plann. Inference* **19** 137–158. MR0944202 https://doi.org/10.1016/0378-3758(88)90069-9

[43] POLYANSKIY, Y. and WU, Y. (2020). Self-regularizing property of nonparametric maximum likelihood estimator in mixture models. Preprint. Available at arXiv:2008.08244.

[44] POLYANSKIY, Y. and WU, Y. (2021). Sharp regret bounds for empirical Bayes and compound decision problems. Preprint. Available at arXiv:2109.03943.

[45] 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

[46] ROBBINS, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. I* 157–163. Univ. California Press, Berkeley, CA. MR0084919

[47] ROBBINS, H. and ZHANG, C.-H. (2000). Efficiency of the $u$, $v$ method of estimation. *Proc. Natl. Acad. Sci. USA* **97** 12976–12979. MR1795617 https://doi.org/10.1073/pnas.97.24.12976

[48] SAHA, S. and GUNTUBOYINA, A. (2020). On the nonparametric maximum likelihood estimator for Gaussian location mixture densities with application to Gaussian denoising. *Ann. Statist.* **48** 738–762. MR4102674 https://doi.org/10.1214/19-AOS1817

[49] SHEN, X. and WONG, W. H. (1994). Convergence rate of sieve estimates. *Ann. Statist.* **22** 580–615. MR1292531 https://doi.org/10.1214/aos/1176325486

[50] SHEN, Y. and WU, Y. (2026). Supplement to "Poisson empirical Bayes estimation: When does $g$-modeling beat $f$-modeling in theory (and in practice)?" https://doi.org/10.1214/25-AOS2560SUPP

[51] SIMAR, L. (1976). Maximum likelihood estimation of a compound Poisson process. *Ann. Statist.* **4** 1200–1209. MR0428582

[52] SINGH, R. S. (1979). Empirical Bayes estimation in Lebesgue-exponential families with rates near the best possible rate. *Ann. Statist.* **7** 890–902. MR0532252

[53] SZEGŐ, G. (1975). *Orthogonal Polynomials*, 4th ed. *American Mathematical Society Colloquium Publications*, *Vol. XXIII*. Amer. Math. Soc., Providence, RI. MR0372517

[54] THEERTHA SURESH, A., ORLITSKY, A., ACHARYA, J. and JAFARPOUR, A. (2014). Near-optimal-sample estimators for spherical Gaussian mixtures. *Adv. Neural Inf. Process. Syst.* **27**.

[55] THISTED, R. and EFRON, B. (1987). Did Shakespeare write a newly-discovered poem? *Biometrika* **74** 445–455. MR0909350 https://doi.org/10.1093/biomet/74.3.445

[56] VAN HOUWELINGEN, J. C. and STIJNEN, T. (1983). Monotone empirical Bayes estimators for the continuous one-parameter exponential family. *Stat. Neerl.* **37** 29–43. MR0700672 https://doi.org/10.1111/j.1467-9574.1983.tb00796.x

[57] VAN DE GEER, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.* **21** 14–44. MR1212164 https://doi.org/10.1214/aos/1176349013

[58] VAN DE GEER, S. (1996). Rates of convergence for the maximum likelihood estimator in mixture models. *J. Nonparametr. Stat.* **6** 293–310. MR1386341 https://doi.org/10.1080/10485259608832677

[59] WOLFOWITZ, J. (1953). Estimation by the minimum distance method. *Ann. Inst. Statist. Math., Tokyo* **5** 9–23. MR0058931 https://doi.org/10.1007/bf02949797

[60] WONG, W. H. and SHEN, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.* **23** 339–362. MR1332570 https://doi.org/10.1214/aos/1176324524

[61] WU, Y. and YANG, P. (2020). Optimal estimation of Gaussian mixtures via denoised method of moments. *Ann. Statist.* **48** 1981–2007. MR4134783 https://doi.org/10.1214/19-AOS1873

[62] ZHANG, C.-H. (1997). Empirical Bayes and compound estimation of normal means. *Statist. Sinica* **7** 181–193. MR1441153

[63] ZHANG, C.-H. (2003). Compound decision theory and empirical Bayes methods. *Ann. Statist.* **31** 379–390. MR1983534 https://doi.org/10.1214/aos/1051027872

[64] ZHANG, C.-H. (2005). Estimation of sums of random variables: Examples and information bounds. *Ann. Statist.* **33** 2022–2041. MR2211078 https://doi.org/10.1214/009053605000000390

[65] ZHANG, C.-H. (2005). General empirical Bayes wavelet methods and exactly adaptive minimax estimation. *Ann. Statist.* **33** 54–100. MR2157796 https://doi.org/10.1214/009053604000000995

[66] ZHANG, C.-H. (2009). Generalized maximum likelihood estimation of normal mixture densities. *Statist. Sinica* **19** 1297–1318. MR2536157

*The Annals of Statistics*
2026, Vol. 54, No. 1, 176–200
https://doi.org/10.1214/25-AOS2563
This research was funded, in whole or in part, by EPSRC, 2261074. A CC BY 4.0 license is applied to this article arising from this submission, in accordance with the grant's open access conditions.

# AVERAGE PARTIAL EFFECT ESTIMATION USING DOUBLE MACHINE LEARNING

BY HARVEY KLYNE[a] AND RAJEN D. SHAH[b]

*Statistical Laboratory, University of Cambridge,* [a]*hck33@cantab.ac.uk,* [b]*r.shah@statslab.cam.ac.uk*

Single-parameter summaries of variable effects in regression settings are desirable for ease of interpretation. However, (partially) linear models, for example, which would deliver these, may fit poorly to the data. On the other hand, an interpretable summary of the contribution of a given predictor is provided by the so-called average partial effect—the average slope of the regression function with respect to the predictor of interest. Although one can construct a doubly robust procedure for estimating this quantity, it entails estimating the derivative of the conditional mean and also the conditional score of the predictor of interest given all others, tasks which can be very challenging in moderate dimensions: in particular, popular decision tree based regression methods cannot be used.

In this work, we introduce an approach for estimating the average partial effect whose accuracy depends primarily on the estimation of certain regression functions, which may be performed by user-chosen machine learning methods that produce potentially nondifferentiable estimates. Our procedure involves resmoothing a given first-stage regression estimator to produce a differentiable version, and modelling the conditional distribution of the predictor of interest through a location–scale model. We show that with the latter assumption, surprisingly the overall error in estimating the conditional score is controlled by a sum of errors of estimating the conditional mean and conditional standard deviation, and the estimation error in a much more tractable univariate score estimation problem. Our theory makes use of a new result on the sub-Gaussianity of Lipschitz score functions that may be of independent interest. We demonstrate the attractive numerical performance of our approach in a variety of settings including ones with misspecification. An R package `drape` implementing the methodology is available on CRAN.

## REFERENCES

BACH, P., CHERNOZHUKOV, V., KURZ, M. S. and SPINDLER, M. (2021). DoubleML – an object-oriented implementation of double machine learning in R.

BERA, A. K. and NG, P. T. (1995). Tests for normality using estimated score function. *J. Stat. Comput. Simul.* **52** 273–287.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. *Johns Hopkins Series in the Mathematical Sciences*. Johns Hopkins Univ. Press, Baltimore, MD. MR1245941

BOJER, C. S. and MELDGAARD, J. P. (2021). Kaggle forecasting competitions: An overlooked learning opportunity. *Int. J. Forecast.* **37** 587–603.

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

CANNY, J. (1986). A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-8** 679–698.

ĆEVID, D., MICHEL, L., NÄF, J., BÜHLMANN, P. and MEINSHAUSEN, N. (2022). Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. *J. Mach. Learn. Res.* **23** 333. MR4577772

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

---

*MSC2020 subject classifications.* 62G20.

*Key words and phrases.* Asymptotic properties of nonparametric inference, average partial effect, average derivative effect, incremental causal effect, double/debiased machine learning, score estimation.

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

CHERNOZHUKOV, V., ESCANCIANO, J. C., ICHIMURA, H., NEWEY, W. K. and ROBINS, J. M. (2022a). Locally robust semiparametric estimation. *Econometrica* **90** 1501–1535. MR4467437 https://doi.org/10.3982/ecta16294

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

CHERNOZHUKOV, V., NEWEY, W. K., QUINTAS-MARTINEZ, V. and SYRGKANIS, V. (2021). Automatic debiased machine learning via neural nets for generalized linear regression. arXiv, arXiv:2104.14737v1.

CHERNOZHUKOV, V., NEWEY, W. K., QUINTAS-MARTINEZ, V. and SYRGKANIS, V. (2022b). Riesznet and forestriesz: Automatic debiased machine learning with neural nets and random forests. In *Proceedings of the Thirty-Ninth International Conference on Machine Learning*.

CHERNOZHUKOV, V., NEWEY, W. K. and SINGH, R. (2022a). Automatic debiased machine learning of causal and structural effects. *Econometrica* **90** 967–1027. MR4436042 https://doi.org/10.3982/ecta18515

CHERNOZHUKOV, V., NEWEY, W. K. and SINGH, R. (2022b). Debiased machine learning of global and local parameters using regularized Riesz representers. *Econom. J.* **25** 576–601. MR4565439 https://doi.org/10.1093/ectj/utac002

CHERNOZHUKOV, V., NEWEY, W. K. and SINGH, R. (2023). A simple and general debiased machine learning theorem with finite-sample guarantees. *Biometrika* **110** 257–264. MR4565455 https://doi.org/10.1093/biomet/asac033

COX, D. D. (1985). A penalty method for nonparametric estimation of the logarithmic derivative of a density function. *Ann. Inst. Statist. Math.* **37** 271–288. MR0799240 https://doi.org/10.1007/BF02481097

DA ROSA, J. C., VEIGA, A. and MEDEIROS, M. C. (2008). Tree-structured smooth transition regression models. *Comput. Statist. Data Anal.* **52** 2469–2488. MR2411951 https://doi.org/10.1016/j.csda.2007.08.018

DAI, W., TONG, T. and GENTON, M. G. (2016). Optimal estimation of derivatives in nonparametric regression. *J. Mach. Learn. Res.* **17** 164. MR3555052

FONSECA, Y., MEDEIROS, M., VASCONCELOS, G. and VEIGA, A. (2018). Boost: Boosting smooth trees for partial effect estimation in nonlinear regressions. arXiv preprint arXiv:1808.03698.

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

GYÖRFI, L., KOHLER, M., KRZYŻAK, A. and WALK, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. *Springer Series in Statistics*. Springer, New York. MR1920390 https://doi.org/10.1007/b97848

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

HASTIE, T. (2009). The elements of statistical learning: Data mining, inference, and prediction.

HINES, O., DIAZ-ORDAZ, K. and VANSTEELANDT, S. (2021). Parameterising the effect of a continuous exposure using average derivative effects. arXiv preprint arXiv:2109.13124.

HIRSHBERG, D. A. and WAGER, S. (2020). Debiased inference of average partial effects in single-index models: Comment on Wooldridge and Zhu [ 4044946 ]. *J. Bus. Econom. Statist.* **38** 19–24. MR4044947 https://doi.org/10.1080/07350015.2019.1681277

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

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

KENNEDY, E. H., MA, Z., MCHUGH, M. D. and SMALL, D. S. (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1229–1245. MR3689316 https://doi.org/10.1111/rssb.12212

KLYNE, H. and SHAH, R. D (2026). Supplement to "Average partial effect estimation using double machine learning." https://doi.org/10.1214/25-AOS2563SUPPA, https://doi.org/10.1214/25-AOS2563SUPPB

LUNDBORG, A. R., KIM, I., SHAH, R. D. and SAMWORTH, R. J. (2024). The projected covariance measure for assumption-lean variable significance testing. *Ann. Statist.* **52** 2851–2878. MR4842829 https://doi.org/10.1214/24-aos2447

NEWEY, W. K. and STOKER, T. M. (1993). Efficiency of weighted average derivative estimators and index models. *Econometrica* **61** 1199–1223. MR1234794 https://doi.org/10.2307/2951498

NG, P. T. (1994). Smoothing spline score estimation. *SIAM J. Sci. Comput.* **15** 1003–1025. MR1289149 https://doi.org/10.1137/0915061

NG, P. T. (2003). Computing Cox's smoothing spline score estimator. Northern Arizona University Working Paper.

POPOV, V., VOVK, I., GOGORYAN, V., SADEKOVA, T., KUDINOV, M. and WEI, J. (2021). Diffusion-based voice conversion with fast maximum likelihood sampling scheme. arXiv preprint arXiv:2109.13821.

POWELL, J. L., STOCK, J. H. and STOKER, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica* **57** 1403–1430. MR1035117 https://doi.org/10.2307/1913713

ROBINS, J. M. and ROTNITZKY, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *J. Amer. Statist. Assoc.* **90** 122–129. MR1325119

ROBINS, J. M. and ROTNITZKY, A. (2001). Comment on Peter J. Bickel and Jaimyoung Kwon article. *Statist. Sinica* **11** 920–936.

ROBINS, J. M., ROTNITZKY, A. and VAN DER LAAN, M. (2000). Comment on S. A. Murphy and A. W. van der Vaart article. *J. Amer. Statist. Assoc.* **95** 477–482.

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

ROTHENHÄUSLER, D. and YU, B. (2020). Incremental causal effects. arXiv, arXiv:1907.13258v4.

ROTNITZKY, A., SMUCLER, E. and ROBINS, J. M. (2021). Characterization of parameters with a mixed bias property. *Biometrika* **108** 231–238. MR4226201 https://doi.org/10.1093/biomet/asaa054

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688.

SCHARFSTEIN, D. O., ROTNITZKY, A. and ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Amer. Statist. Assoc.* **94** 1096–1120.

SCHICK, A. (1986). On asymptotically efficient estimation in semiparametric models. *Ann. Statist.* **14** 1139–1151. MR0856811 https://doi.org/10.1214/aos/1176350055

SHAH, R. D. and BÜHLMANN, P. (2023). Double-estimation-friendly inference for high-dimensional misspecified models. *Statist. Sci.* **38** 68–91. MR4535395 https://doi.org/10.1214/22-sts850

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

SONG, Y., SHEN, L., XING, L. and ERMON, S. (2021). Solving inverse problems in medical imaging with score-based generative models. arXiv preprint arXiv:2111.08005.

SONG, Y., SOHL-DICKSTEIN, J., KINGMA, D. P., KUMAR, A., ERMON, S. and POOLE, B. (2020). Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456.

SRIPERUMBUDUR, B., FUKUMIZU, K., GRETTON, A., HYVÄRINEN, A. and KUMAR, R. (2017). Density estimation in infinite dimensional exponential families. *J. Mach. Learn. Res.* **18** 57. MR3687600

STOKER, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica* **54** 1461–1481. MR0868152 https://doi.org/10.2307/1914309

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

VANSTEELANDT, S. and DUKES, O. (2022). Assumption-lean inference for generalised linear model parameters. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 657–685. MR4460549

WIBISONO, A., WU, Y. and YANG, K. Y. (2024). Optimal score estimation via empirical Bayes smoothing. arXiv preprint arXiv:2402.07747.

WOOLDRIDGE, J. M. and ZHU, Y. (2020). Inference in approximately sparse correlated random effects probit models with panel data. *J. Bus. Econom. Statist.* **38** 1–18. MR4044946 https://doi.org/10.1080/07350015.2019.1681276

WRIGHT, M. N. and ZIEGLER, A. (2017). ranger: A fast implementation of random forests for high dimensional data in c++ and r. *J. Stat. Softw.* **77**.

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

*The Annals of Statistics*
2026, Vol. 54, No. 1, 201–225
https://doi.org/10.1214/25-AOS2564

# SOLVING THE POISSON EQUATION USING COUPLED MARKOV CHAINS

BY RANDAL DOUC[1,a], PIERRE E. JACOB[2,b], ANTHONY LEE[3,c] AND DOOTIKA VATS[4,d]

[1]*SAMOVAR, Télécom SudParis, Institut Polytechnique de Paris,* [a]*randal.douc@telecom-sudparis.eu*
[2]*Department of Information Systems, Data Analytics and Operations, ESSEC Business School,* [b]*pierre.jacob@essec.edu*
[3]*School of Mathematics, University of Bristol,* [c]*anthony.lee@bristol.ac.uk*
[4]*Department of Mathematics and Statistics, IIT Kanpur,* [d]*dootika@iitk.ac.in*

This article shows how coupled Markov chains that meet exactly after a random number of iterations can be used to generate unbiased estimators of the solutions of the Poisson equation. Through this connection, we rederive known unbiased estimators of expectations with respect to the stationary distribution of a Markov chain and provide conditions for the finiteness of their moments. We further construct unbiased estimators of the asymptotic variance of Markov chain ergodic averages, and provide conditions for the finiteness of the estimators' moments of any order. If their second moment is finite, the average of independent copies of such estimators converges to the asymptotic variance at the Monte Carlo rate, comparing favorably to known rates for batch means and spectral variance estimators. The results are illustrated with numerical experiments.

## REFERENCES

[1] AGAPIOU, S., ROBERTS, G. O. and VOLLMER, S. J. (2018). Unbiased Monte Carlo: Posterior estimation for intractable/infinite-dimensional models. *Bernoulli* **24** 1726–1786. MR3757514 https://doi.org/10.3150/16-BEJ911

[2] AGARWAL, M. and VATS, D. (2022). Globally centered autocovariances in MCMC. *J. Comput. Graph. Statist.* **31** 629–638. MR4495699 https://doi.org/10.1080/10618600.2022.2037433

[3] ALEXOPOULOS, A., DELLAPORTAS, P. and TITSIAS, M. K. (2023). Variance reduction for Metropolis–Hastings samplers. *Stat. Comput.* **33** Paper No. 6, 20 pp. MR4514995 https://doi.org/10.1007/s11222-022-10183-2

[4] ANDRADÓTTIR, S., HEYMAN, D. P. and OTT, T. J. (1993). Variance reduction through smoothing and control variates for Markov chain simulations. *ACM Trans. Model. Comput. Simul.* **3** 167–189.

[5] ANDRIEU, C. and VIHOLA, M. (2015). Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. *Ann. Appl. Probab.* **25** 1030–1077. MR3313762 https://doi.org/10.1214/14-AAP1022

[6] ATCHADÉ, Y. F. (2016). Markov chain Monte Carlo confidence intervals. *Bernoulli* **22** 1808–1838. MR3474834 https://doi.org/10.3150/15-BEJ712

[7] ATCHADÉ, Y. F. and JACOB, P. E. (2024). Unbiased Markov chain Monte Carlo: What, why, and how. Preprint. Available at arXiv:2406.06851.

[8] BEDNORZ, W., ŁATUSZYŃSKI, K. and LATAŁA, R. (2008). A regeneration proof of the central limit theorem for uniformly ergodic Markov chains. *Electron. Commun. Probab.* **13** 85–98. MR2386065 https://doi.org/10.1214/ECP.v13-1354

[9] BERG, S. and SONG, H. (2023). Efficient shape-constrained inference for the autocovariance sequence from a reversible Markov chain. *Ann. Statist.* **51** 2440–2470. MR4682704 https://doi.org/10.1214/23-aos2335

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

[11] BISWAS, N., BHATTACHARYA, A., JACOB, P. E. and JOHNDROW, J. E. (2022). Coupling-based convergence assessment of some Gibbs samplers for high-dimensional Bayesian regression with shrinkage priors. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 973–996. MR4460582 https://doi.org/10.1111/rssb.12495

---

*MSC2020 subject classifications.* Primary 65C40, 65C05, 60F05; secondary 60J22, 60J05, 62F15.
*Key words and phrases.* Markov chain Monte Carlo, coupling, Poisson equation, asymptotic variance.

[12] Brockwell, A. E. and Kadane, J. B. (2005). Identification of regeneration times in MCMC simulation, with application to adaptive schemes. *J. Comput. Graph. Statist.* **14** 436–458. MR2161623 https://doi.org/10.1198/106186005X47453

[13] Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P. et al. (2017). Stan: A probabilistic programming language. *J. Stat. Softw.* **76** 1.

[14] Chakraborty, S., Bhattacharya, S. K. and Khare, K. (2022). Estimating accuracy of the MCMC variance estimator: Asymptotic normality for batch means estimators. *Statist. Probab. Lett.* **183** Paper No. 109337, 6 pp. MR4363701 https://doi.org/10.1016/j.spl.2021.109337

[15] Corenflos, A., Sutton, M. and Chopin, N. (2023). Debiasing piecewise deterministic Markov process samplers using couplings. Preprint. Available at arXiv:2306.15422.

[16] Deligiannidis, G., Jacob, P. E., Khribch, E. M. and Wang, G. (2024). On importance sampling and independent Metropolis–Hastings with an unbounded weight function. Preprint. Available at arXiv:2411.09514.

[17] Dellaportas, P. and Kontoyiannis, I. (2012). Control variates for estimation based on reversible Markov chain Monte Carlo samplers. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 133–161. MR2885843 https://doi.org/10.1111/j.1467-9868.2011.01000.x

[18] Douc, R.,Jacob, P. E.,Lee, A. and Vats, D. (2026). Supplement to “Solving the Poisson equation using coupled Markov chains.” https://doi.org/10.1214/25-AOS2564SUPPA, https://doi.org/10.1214/25-AOS2564SUPPB

[19] Douc, R., Moulines, E., Priouret, P. and Soulier, P. (2018). *Markov Chains. Springer Series in Operations Research and Financial Engineering*. Springer, Cham. MR3889011 https://doi.org/10.1007/978-3-319-97704-1

[20] Durmus, A., Moulines, É. and Saksman, E. (2020). Irreducibility and geometric ergodicity of Hamiltonian Monte Carlo. *Ann. Statist.* **48** 3545–3564. MR4185819 https://doi.org/10.1214/19-AOS1941

[21] Flegal, J. M., Hughes, J., Vats, D. and Dai, N. (2020). mcmcse: Monte Carlo standard errors for MCMC, Riverside, CA, Denver, CO, Coventry, UK, and Minneapolis, MN. R package version 1.4-1.

[22] Flegal, J. M. and Jones, G. L. (2010). Batch means and spectral variance estimators in Markov chain Monte Carlo. *Ann. Statist.* **38** 1034–1070. MR2604704 https://doi.org/10.1214/09-AOS735

[23] Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statist. Sci.* **7** 473–483.

[24] Glynn, P. W. and Infanger, A. (2022). Solving Poisson’s equation: Existence, uniqueness, martingale structure, and CLT. Preprint. Available at arXiv:2202.10404.

[25] Glynn, P. W. and Meyn, S. P. (1996). A Liapounov bound for solutions of the Poisson equation. *Ann. Probab.* **24** 916–931. MR1404536 https://doi.org/10.1214/aop/1039639370

[26] Glynn, P. W. and Rhee, C.-H. (2014). Exact estimation for Markov chain equilibrium expectations. *J. Appl. Probab.* **51A** 377–389. MR3317370 https://doi.org/10.1239/jap/1417528487

[27] Glynn, P. W. and Whitt, W. (1992). The asymptotic efficiency of simulation estimators. *Oper. Res.* **40** 505–520. MR1180030 https://doi.org/10.1287/opre.40.3.505

[28] Gupta, K. and Vats, D. (2020). Estimating Monte Carlo variance from multiple Markov chains. Preprint. Available at arXiv:2007.04229.

[29] Henderson, S. G. (1997). Variance reduction via an approximating Markov process. Ph.D. thesis, Stanford Univ.

[30] Heng, J. and Jacob, P. E. (2019). Unbiased Hamiltonian Monte Carlo with couplings. *Biometrika* **106** 287–302. MR3949304 https://doi.org/10.1093/biomet/asy074

[31] Hobert, J. P., Jones, G. L., Presnell, B. and Rosenthal, J. S. (2002). On the applicability of regenerative simulation in Markov chain Monte Carlo. *Biometrika* **89** 731–743. MR1946508 https://doi.org/10.1093/biomet/89.4.731

[32] Jacob, P. E., O’Leary, J. and Atchadé, Y. F. (2020). Unbiased Markov chain Monte Carlo methods with couplings. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 543–600. MR4112777 https://doi.org/10.1111/rssb.12336

[33] Jarner, S. F. and Hansen, E. (2000). Geometric ergodicity of Metropolis algorithms. *Stochastic Process. Appl.* **85** 341–361. MR1731030 https://doi.org/10.1016/S0304-4149(99)00082-4

[34] Johnson, V. E. (1998). A coupling-regeneration scheme for diagnosing convergence in Markov chain Monte Carlo algorithms. *J. Amer. Statist. Assoc.* **93** 238–248. MR1614640 https://doi.org/10.2307/2669620

[35] Jones, G. L. (2004). On the Markov chain central limit theorem. *Probab. Surv.* **1** 299–320. MR2068475 https://doi.org/10.1214/154957804100000051

[36] Karjalainen, J., Lee, A., Singh, S. S. and Vihola, M. (2023). Mixing time of the conditional backward sampling particle filter. Preprint. Available at arXiv:2312.17572.

[37] KELLY, L. J., RYDER, R. J. and CLARTÉ, G. (2023). Lagged couplings diagnose Markov chain Monte Carlo phylogenetic inference. *Ann. Appl. Stat.* **17** 1419–1443. MR4582719 https://doi.org/10.1214/22-aoas1676

[38] KONTOYIANNIS, I. and DELLAPORTAS, P. (2009). Notes on using control variates for estimation with reversible MCMC samplers. Preprint. Available at arXiv:0907.4160.

[39] LEE, A., DOUCET, A. and ŁATUSZYŃSKI, K. (2014). Perfect simulation using atomic regeneration with application to sequential Monte Carlo. Preprint. Available at arXiv:1407.5770.

[40] LEE, A., SINGH, S. S. and VIHOLA, M. (2020). Coupled conditional backward sampling particle filter. *Ann. Statist.* **48** 3066–3089. MR4152635 https://doi.org/10.1214/19-AOS1922

[41] LILLARD, L. A. and PANIS, C. W. (2000). aML user's guide and reference manual. EconWare, Los Angeles.

[42] LIU, Y., VATS, D. and FLEGAL, J. M. (2022). Batch size selection for variance estimators in MCMC. *Methodol. Comput. Appl. Probab.* **24** 65–93. MR4379481 https://doi.org/10.1007/s11009-020-09841-7

[43] MAXWELL, M. and WOODROOFE, M. (2000). Central limit theorems for additive functionals of Markov chains. *Ann. Probab.* **28** 713–724. MR1782272 https://doi.org/10.1214/aop/1019160258

[44] MENGERSEN, K. L. and TWEEDIE, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.* **24** 101–121. MR1389882 https://doi.org/10.1214/aos/1033066201

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

[46] MIDDLETON, L., DELIGIANNIDIS, G., DOUCET, A. and JACOB, P. E. (2020). Unbiased Markov chain Monte Carlo for intractable target distributions. *Electron. J. Stat.* **14** 2842–2891. MR4132645 https://doi.org/10.1214/20-EJS1727

[47] MIJATOVIĆ, A. and VOGRINC, J. (2018). On the Poisson equation for Metropolis-Hastings chains. *Bernoulli* **24** 2401–2428. MR3757533 https://doi.org/10.3150/17-BEJ932

[48] MYKLAND, P., TIERNEY, L. and YU, B. (1995). Regeneration in Markov chain samplers. *J. Amer. Statist. Assoc.* **90** 233–241. MR1325131

[49] NGUYEN, T. D., TRIPPE, B. L. and BRODERICK, T. (2022). Many processors, little time: MCMC for partitions via optimal transport couplings. In *International Conference on Artificial Intelligence and Statistics* 3483–3514. PMLR.

[50] OWEN, A. B. (2017). Statistically efficient thinning of a Markov chain sampler. *J. Comput. Graph. Statist.* **26** 738–744. MR3698682 https://doi.org/10.1080/10618600.2017.1336446

[51] PITMAN, J. W. (1976). On coupling of Markov chains. *Z. Wahrsch. Verw. Gebiete* **35** 315–322. MR0415775 https://doi.org/10.1007/BF00532957

[52] PLUMMER, M., BEST, N., COWLES, K. and VINES, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News* **6** 7–11.

[53] ROBERT, C. P. (1995). Convergence control methods for Markov chain Monte Carlo algorithms. *Statist. Sci.* **10** 231–253. MR1390517

[54] ROBERTS, G. O. and TWEEDIE, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* **2** 341–363. MR1440273 https://doi.org/10.2307/3318418

[55] RODRÍGUEZ, G. (2008). Multilevel generalized linear models. In *Handbook of Multilevel Analysis* 335–376. Springer, New York. MR2412942 https://doi.org/10.1007/978-0-387-73186-5_9

[56] VANETTI, P. and DOUCET, A. (2020). Discussion on the paper by Jacob, O'Leary, and Atchadé. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 584–585.

[57] VATS, D. and FLEGAL, J. M. (2022). Lugsail lag windows for estimating time-average covariance matrices. *Biometrika* **109** 735–750. MR4472845 https://doi.org/10.1093/biomet/asab049

[58] VATS, D., FLEGAL, J. M. and JONES, G. L. (2018). Strong consistency of multivariate spectral variance estimators in Markov chain Monte Carlo. *Bernoulli* **24** 1860–1909. MR3757517 https://doi.org/10.3150/16-BEJ914

[59] VATS, D., FLEGAL, J. M. and JONES, G. L. (2019). Multivariate output analysis for Markov chain Monte Carlo. *Biometrika* **106** 321–337. MR3949306 https://doi.org/10.1093/biomet/asz002

[60] VITTER, J. S. (1985). Random sampling with a reservoir. *ACM Trans. Math. Software* **11** 37–57. MR0793056 https://doi.org/10.1145/3147.3165

*The Annals of Statistics*
2026, Vol. 54, No. 1, 226–251
https://doi.org/10.1214/25-AOS2565

# A COMPUTATIONAL TRANSITION FOR DETECTING CORRELATED STOCHASTIC BLOCK MODELS BY LOW-DEGREE POLYNOMIALS

BY GUANYI CHEN[a], JIAN DING[b], SHUYANG GONG[c] AND ZHANGSONG LI[d]

*School of Mathematical Science, Peking University,* [a]*cgy7698@stu.pku.edu.cn,* [b]*dingjian@math.pku.edu.cn,* [c]*gongshuyang@stu.pku.edu.cn,* [d]*ramblerlzs@pku.edu.cn*

Detection of correlation in a pair of random graphs is a fundamental statistical and computational problem that has been extensively studied in recent years. In this work, we consider a pair of correlated (sparse) stochastic block models $\mathcal{S}(n, \frac{\lambda}{n}; k, \epsilon; s)$ that are subsampled from a common parent stochastic block model $\mathcal{S}(n, \frac{\lambda}{n}; k, \epsilon)$ with $k = O(1)$ symmetric communities, average degree $\lambda = O(1)$, divergence parameter $\epsilon$, and subsampling probability $s$.

For the detection problem of distinguishing this model from a pair of independent Erdős–Rényi graphs with the same edge density $\mathcal{G}(n, \frac{\lambda s}{n})$, we focus on tests based on *low-degree polynomials* of the entries of the adjacency matrices, and we determine the threshold that separates the easy and hard regimes. More precisely, we show that this class of tests can distinguish these two models if and only if $s > \min\{\sqrt{\alpha}, \frac{1}{\lambda\epsilon^2}\}$, where $\alpha \approx 0.338$ is the Otter's constant and $\frac{1}{\lambda\epsilon^2}$ is the Kesten–Stigum threshold. Combining a reduction argument in (Li (2025)), our hardness result also implies low-degree hardness for partial recovery and detection (to independent block models) when $s < \min\{\sqrt{\alpha}, \frac{1}{\lambda\epsilon^2}\}$. Finally, our proof of low-degree hardness is based on a conditional variant of the low-degree likelihood calculation.

## REFERENCES

[1] ABBE, E., BANDEIRA, A. S. and HALL, G. (2016). Exact recovery in the stochastic block model. *IEEE Trans. Inf. Theory* **62** 471–487. MR3447993 https://doi.org/10.1109/TIT.2015.2490670

[2] ABBE, E. and SANDON, C. (2015). Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In 2015 *IEEE* 56*th Annual Symposium on Foundations of Computer Science—FOCS* 2015 670–688. IEEE Computer Soc., Los Alamitos, CA. MR3473334 https://doi.org/10.1109/FOCS.2015.47

[3] ABBE, E. and SANDON, C. (2016). Achieving the KS threshold in the general stochastic block model with linearized acyclic belief propagation. In *Advances in Neural Information Processing Systems* (*NIPS*) **29** 1334–1342. Curran Associates, Red Hook.

[4] ABBE, E. and SANDON, C. (2018). Proof of the achievability conjectures for the general stochastic block model. *Comm. Pure Appl. Math.* **71** 1334–1406. MR3812075 https://doi.org/10.1002/cpa.21719

[5] BANDEIRA, A. S., ALAOUI, A. E., HOPKINS, S., SCHRAMM, T., WEIN, A. S. and ZADIK, I. (2022). The Franz-Parisi criterion and computational trade-offs in high dimensional statistics. In *Advances in Neural Information Processing Systems* (*NIPS*) **35** 33831–33844. Curran Associates, Red Hook.

[6] BANDEIRA, A. S., KUNISKY, D. and WEIN, A. S. (2020). Computational hardness of certifying bounds on constrained PCA problems. In 11*th Innovations in Theoretical Computer Science Conference*. *LIPIcs. Leibniz Int. Proc. Inform.* **151** 78. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern. MR4048181

[7] BANDEIRA, A. S., PERRY, A. and WEIN, A. S. (2018). Notes on computational-to-statistical gaps: Predictions using statistical physics. *Port. Math.* **75** 159–186. MR3892753 https://doi.org/10.4171/PM/2014

[8] BANKS, J., MOORE, C., NEEMAN, J. and NETRAPALLI, P. (2016). Information-theoretic thresholds for community detection in sparse networks. In *Proceedings of the* 29*th Conference on Learning Theory* (*COLT*) 383–416. PMLR.

[9] BARAK, B., CHOU, C. N., LEI, Z., SCHRAMM, T. and SHENG, Y. (2019). (Nearly) efficient algorithms for the graph matching problem on correlated random graphs. In *Advances in Neural Information Processing Systems* (*NIPS*) **32** 9190–9198. Curran Associates, Red Hook.

---

*MSC2020 subject classifications.* Primary 62M20; secondary 68Q87, 68Q17.

*Key words and phrases.* Random graph matching, correlated stochastic block models, computation transition, low-degree polynomials.

[10] Barak, B., Hopkins, S., Kelner, J., Kothari, P. K., Moitra, A. and Potechin, A. (2019). A nearly tight sum-of-squares lower bound for the planted clique problem. *SIAM J. Comput.* **48** 687–735. MR3945259 https://doi.org/10.1137/17M1138236

[11] Berg, A., Berg, T. and Malik, J. (2005). Shape matching and object recognition using low distortion correspondences. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (*CVPR*) **1** 26–33. IEEE.

[12] Bickel, P. J. and Chen, A. (2009). A nonparametric view of network models and Newman-Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106** 21068–21073.

[13] Bordenave, C., Lelarge, M. and Massoulié, L. (2015). Non-backtracking spectrum of random graphs: Community detection and non-regular Ramanujan graphs. In 2015 *IEEE* 56*th Annual Symposium on Foundations of Computer Science—FOCS* 2015 1347–1357. IEEE Computer Soc., Los Alamitos, CA. MR3473374 https://doi.org/10.1109/FOCS.2015.86

[14] Bozorg, M., Salehkaleybar, S. and Hashemi, M. (2019). Seedless graph matching via tail of degree distribution for correlated Erdős-Rényi graphs. arXiv preprint, arXiv:1907.06334.

[15] Brennan, M. and Bresler, G. (2020). Reducibility and statistical-computational gaps from secret leakage. In *Proceedings of the* 33*rd Conference on Learning Theory* (*COLT*) 648–847. PMLR.

[16] Brennan, M., Bresler, G. and Huleihel, W. (2018). Reducibility and computational lower bounds for problems with planted sparse structure. In *Proceedings of the* 31*st Conference on Learning Theory* (*COLT*) 48–166. PMLR.

[17] Bresler, G. and Huang, B. (2022). The algorithmic phase transition of random $k$-SAT for low degree polynomials. In 2021 *IEEE* 62*nd Annual Symposium on Foundations of Computer Science—FOCS* 2021 298–309. IEEE Computer Soc., Los Alamitos, CA. MR4399691

[18] Brito, G., Dumitriu, I., Ganguly, S., Hoffman, C. and Tran, L. V. (2016). Recovery and rigidity in a regular stochastic block model. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms* 1589–1601. ACM, New York. MR3478484 https://doi.org/10.1137/1.9781611974331.ch108

[19] Chai, S. and Rácz, M. Z. (2024). Efficient graph matching for correlated stochastic block models. In *Advances in Neural Information Processing Systems* (*NIPS*) **37** 116388–116461. Curran Associates, Red Hook.

[20] Chen, G., Ding, J., Gong, S. and Li, Z. (2026). Supplement to "A computational transition for detecting correlated stochastic block models by low-degree polynomials." https://doi.org/10.1214/25-AOS2565SUPP

[21] Chen, G., Ding, J., Gong, S. and Li, Z. (2025). Detecting correlation efficiently in stochastic block models: Breaking Otter's threshold by counting decorated trees. arXiv preprint, arXiv:2503.06464.

[22] Chin, B. and Sly, A. (2021). Optimal reconstruction of general sparse stochastic block models. arXiv preprint, arXiv:2111.00697.

[23] Coja-Oghlan, A., Gebhard, O., Hahn-Klimroth, M., Wein, A. S. and Zadik, I. (2022). Statistical and computational phase transitions in group testing. In *Proceedings of the* 35*th Conference on Learning Theory* (*COLT*) 4764–4781. PMLR.

[24] Correa, A. F. D. (2022). *Data Visualization*, *Dimensionality Reduction*, *and Data Alignment via Manifold Learning*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)—Utah State University. MR4578842

[25] Cour, T., Srinivasan, P. and Shi, J. (2006). Balanced graph matching. In *Advances in Neural Information Processing Systems* (*NIPS*) **19**. MIT Press, Cambridge.

[26] Cullina, D. and Kiyavash, N. (2016). Improved achievability and converse bounds for Erdős-Rényi graph matching. In *Proceedings of the* 2016 *ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science* 63–72. ACM, New York.

[27] Cullina, D. and Kiyavash, N. (2017). Exact alignment recovery for correlated Erdős-Rényi graphs. arXiv preprint, arXiv:1711.06783.

[28] Dai, O. E., Cullina, D., Kiyavash, N. and Grossglauser, M. (2019). Analysis of a canonical labeling algorithm for the alignment of correlated Erdős-Rényi graphs. In *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 1–25. ACM, New York.

[29] Decelle, A., Krzakala, F., Moore, C. and Zdeborová, L. (2011). Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E* **84** 066106.

[30] Dhawan, A., Mao, C. and Wein, A. S. (2025). Detection of dense subhypergraphs by low-degree polynomials. *Random Structures Algorithms* **66** e21279. MR4853082 https://doi.org/10.1002/rsa.21279

[31] Ding, J. and Du, H. (2023). Detection threshold for correlated Erdős-Rényi graphs via densest subgraph. *IEEE Trans. Inf. Theory* **69** 5289–5298. MR4624732 https://doi.org/10.1109/tit.2023.3265009

[32] Ding, J. and Du, H. (2023). Matching recovery threshold for correlated random graphs. *Ann. Statist.* **51** 1718–1743. MR4658574 https://doi.org/10.1214/23-aos2305

[33] Ding, J., Du, H. and Li, Z. (2023). Low-degree hardness of detection for correlated Erdős–Rényi graphs. *Ann. Statist.* **53** 1833–1856. MR4985251 http://dx.doi.org/10.1214/25-AOS2517

[34] Ding, J. and Li, Z. (2025). A polynomial time iterative algorithm for matching Gaussian matrices with non-vanishing correlation. *Found. Comput. Math.* **25** 1287–1344. MR4951528 https://doi.org/10.1007/s10208-024-09662-x

[35] Ding, J. and Li, Z. (2025). A polynomial-time iterative algorithm for random graph matching with non-vanishing correlation. Available at arXiv:2306.00266.

[36] Ding, J., Ma, Z., Wu, Y. and Xu, J. (2021). Efficient random graph matching via degree profiles. *Probab. Theory Related Fields* **179** 29–115. MR4221654 https://doi.org/10.1007/s00440-020-00997-4

[37] Ding, Y., Kunisky, D., Wein, A. S. and Bandeira, A. S. (2024). Subexponential-time algorithms for sparse PCA. *Found. Comput. Math.* **24** 865–914. MR4760356 https://doi.org/10.1007/s10208-023-09603-0

[38] Fan, Z., Mao, C., Wu, Y. and Xu, J. (2023). Spectral graph matching and regularized quadratic relaxations I Algorithm and Gaussian analysis. *Found. Comput. Math.* **23** 1511–1565. MR4649430 https://doi.org/10.1007/s10208-022-09570-y

[39] Fan, Z., Mao, C., Wu, Y. and Xu, J. (2023). Spectral graph matching and regularized quadratic relaxations II. *Found. Comput. Math.* **23** 1567–1617. MR4649431 https://doi.org/10.1007/s10208-022-09575-7

[40] Feizi, S., Quon, G., Medard, M., Kellis, M. and Jadbabaie, A. (2016). Spectral alignment of networks. arXiv preprint, arXiv:1602.04181.

[41] Feng, C. (2025). Strong detection threshold for correlated Erdős-Rényi graphs with constant average degree. arXiv preprint, 2506.12752.

[42] Gamarnik, D. (2021). The overlap gap property: A topological barrier to optimizing over random structures. *Proc. Natl. Acad. Sci. USA* **118**. e2108492118.

[43] Gamarnik, D., Jagannath, A. and Wein, A. S. (2024). Hardness of random optimization problems for Boolean circuits, low-degree polynomials, and Langevin dynamics. *SIAM J. Comput.* **53** 1–46. MR4704592 https://doi.org/10.1137/22M150263X

[44] Ganassali, L., Lelarge, M. and Massoulié, L. (2024). Correlation detection in trees for planted graph alignment. *Ann. Appl. Probab.* **34** 2799–2843. MR4757492 https://doi.org/10.1214/23-aap2020

[45] Ganassali, L. and Massoulié, L. (2020). From tree matching to sparse graph alignment. In *Proceedings of the* 33*rd Conference on Learning Theory* (*COLT*) 1633–1665. PMLR.

[46] Ganassali, L., Massoulié, L. and Semerjian, G. (2024). Statistical limits of correlation detection in trees. *Ann. Appl. Probab.* **34** 3701–3734. MR4783028 https://doi.org/10.1214/23-aap2048

[47] Gaudi, J., Rácz, M. Z. and Sridhar, A. (2022). Exact community recovery in correlated stochastic block models. In *Proceedings of the* 35*th Conference on Learning Theory* (*COLT*) 2183–2241. PMLR.

[48] Haghighi, A., Ng, A. and Manning, C. (2005). Robust textual inference via graph matching. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing* (*EMNLP*) 387–394.

[49] Hall, G. and Massoulié, L. (2023). Partial recovery in the graph alignment problem. *Oper. Res.* **71** 259–272. MR4585566

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

[51] Hopkins, S. (2018). *Statistical Inference and the Sum of Squares Method*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)—Cornell University. MR3864930

[52] Hopkins, S. B., Kothari, P. K., Potechin, A., Raghavendra, P., Schramm, T. and Steurer, D. (2017). The power of sum-of-squares for detecting hidden structures. In 58*th Annual IEEE Symposium on Foundations of Computer Science—FOCS* 2017 720–731. IEEE Computer Soc., Los Alamitos, CA. MR3734275 https://doi.org/10.1109/FOCS.2017.72

[53] Hopkins, S. B. and Steurer, D. (2017). Efficient Bayesian estimation from few samples: Community detection and related problems. In 58*th Annual IEEE Symposium on Foundations of Computer Science—FOCS* 2017 379–390. IEEE Computer Soc., Los Alamitos, CA. MR3734245 https://doi.org/10.1109/FOCS.2017.42

[54] Kazemi, E., Hassani, S. H. and Grossglauser, M. (2015). Growing a graph matching from a handful of seeds. *Proc. VLDB Endow.* **8** 1010–1021.

[55] Kesten, H. and Stigum, B. P. (1966). Additional limit theorems for indecomposable multidimensional Galton-Watson processes. *Ann. Math. Statist.* **37** 1463–1481. MR0200979 https://doi.org/10.1214/aoms/1177699139

[56] Kunisky, D., Moore, C. and Wein, A. S. (2024). Tensor cumulants for statistical inference on invariant distributions. In 2024 *IEEE* 65*th Annual Symposium on Foundations of Computer Science—FOCS* 2024 1007–1026. IEEE Computer Soc., Los Alamitos, CA. MR4849263

[57] KUNISKY, D., WEIN, A. S. and BANDEIRA, A. S. (2022). Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio. In *Mathematical Analysis*, *Its Applications and Computation*. *Springer Proc*. *Math*. *Stat*. **385** 1–50. Springer, Cham. MR4461037 https://doi.org/10.1007/978-3-030-97127-4_1

[58] LI, Z. (2025). Algorithmic contiguity from low-degree conjecture and applications in correlated random graphs In *Approximation*, *Randomization*, *and Combinatorial Optimization*. *Algorithms and Techniques LIPIcs*. *Leibniz Int*. *Proc*. *Inform*. **353** 18. Schloss Dagstuhl-Leibniz-Zentrumfűr Informatik. Berckley. MR4967588 https://doi.org/10.4230/LIPIcs.APPROX/RANDOM.2025.30

[59] LYZINSKI, V. (2018). Information recovery in shuffled graphs via graph matching. *IEEE Trans*. *Inf*. *Theory* **64** 3254–3273. MR3798375 https://doi.org/10.1109/TIT.2018.2808999

[60] LYZINSKI, V., FISHKIND, D. E. and PRIEBE, C. E. (2014). Seeded graph matching for correlated Erdős-Rényi graphs. *J*. *Mach*. *Learn*. *Res*. **15** 3513–3540. MR3291408

[61] LYZINSKI, V., SUSSMAN, D. L., FISHKIND, D. E., PAO, H., CHEN, L., VOGELSTEIN, J. T., PARK, Y. and PRIEBE, C. E. (2015). Spectral clustering for divide-and-conquer graph matching. *Parallel Comput*. **47** 70–87. MR3371642 https://doi.org/10.1016/j.parco.2015.03.004

[62] MAO, C., RUDELSON, M. and TIKHOMIROV, K. (2021). Random graph matching with improved noise robustness. In *Proceedings of the* 34*th Conference on Learning Theory* (*COLT*) 3296–3329. PMLR.

[63] MAO, C., RUDELSON, M. and TIKHOMIROV, K. (2023). Exact matching of random graphs with constant correlation. *Probab. Theory Related Fields* **186** 327–389. MR4586222 https://doi.org/10.1007/s00440-022-01184-3

[64] MAO, C. and WEIN, A. S. (2025). Optimal spectral recovery of a planted vector in a subspace. *Bernoulli* **31** 1114–1139. MR4863069 https://doi.org/10.3150/24-bej1763

[65] MAO, C., WU, Y., XU, J. and YU, S. H. (2023). Random graph matching at Otter's threshold via counting chandeliers. In *STOC*'23—*Proceedings of the* 55*th Annual ACM Symposium on Theory of Computing* 1345–1356. ACM, New York. MR4617472 https://doi.org/10.1145/3564246.3585156

[66] MAO, C., WU, Y., XU, J. and YU, S. H. (2024). Testing network correlation efficiently via counting trees. *Ann*. *Statist*. **52** 2483–2505. MR4841674 https://doi.org/10.1214/23-AOS2261

[67] MASSOULIÉ, L. (2014). Community detection thresholds and the weak Ramanujan property. In *STOC*'14—*Proceedings of the* 2014 *ACM Symposium on Theory of Computing* 694–703. ACM, New York. MR3238997

[68] MONTANARI, A. and WEIN, A. S. (2025). Equivalence of approximate message passing and low-degree polynomials in rank-one matrix estimation. *Probab. Theory Related Fields* **191** 181–233. MR4869255 https://doi.org/10.1007/s00440-024-01322-z

[69] MOSSEL, E., NEEMAN, J. and SLY, A. (2015). Reconstruction and estimation in the planted partition model. *Probab. Theory Related Fields* **162** 431–461. MR3383334 https://doi.org/10.1007/s00440-014-0576-6

[70] MOSSEL, E., NEEMAN, J. and SLY, A. (2018). A proof of the block model threshold conjecture. *Combinatorica* **38** 665–708. MR3876880 https://doi.org/10.1007/s00493-016-3238-8

[71] MOSSEL, E., SLY, A. and SOHN, Y. (2025). Exact phase transitions for stochastic block models and reconstruction on trees. *Ann*. *Probab*. **53** 967–1018. MR4908095 https://doi.org/10.1214/24-aop1723

[72] MOSSEL, E., SLY, A. and SOHN, Y. (2025). Weak recovery, hypothesis testing, and mutual information in stochastic block models and planted factor graphs. In *STOC*'25—*Proceedings of the* 57*th Annual ACM Symposium on Theory of Computing* 2062–2073. ACM, New York. MR4928588 https://doi.org/10.1145/3717823.3718292

[73] MOSSEL, E. and XU, J. (2020). Seeded graph matching via large neighborhood statistics. *Random Structures Algorithms* **57** 570–611. MR4144076 https://doi.org/10.1002/rsa.20934

[74] NARAYANAN, A. and SHMATIKOV, V. (2008). Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy* 111–125. IEEE.

[75] NARAYANAN, A. and SHMATIKOV, V. (2009). De-anonymizing social networks. In *IEEE Symposium on Security and Privacy* 173–187. IEEE, Los Alamitos.

[76] ONARAN, E., GARG, S. and ERKIP, E. (2016). Optimal de-anonymization in random graphs with community structure. In *IEEE* 50*th Asilomar Conference on Signals*, *Systems and Computers* 709–713. IEEE.

[77] OTTER, R. (1948). The number of trees. *Ann*. *of Math*. (2) **49** 583–599. MR0025715 https://doi.org/10.2307/1969046

[78] PEDARSANI, P. and GROSSGLAUSER, M. (2011). On the privacy of anonymized networks. In *Proceedings of the* 17*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1235–1243. ACM, New York.

[79] PICCIOLI, G., SEMERJIAN, G., SICURO, G. and ZDEBOROVÁ, L. (2022). Aligning random graphs with a sub-tree similarity message-passing algorithm. *J*. *Stat*. *Mech*. *Theory Exp*. **6** 063401. MR4483763 https://doi.org/10.1088/1742-5468/ac70d2

[80] RÁCZ, M. Z. and SRIDHAR, A. (2021). Correlated stochastic block models: Exact graph matching with applications to recovering communities. In *Advances in Neural Information Processing Systems* (*NIPS*) **34** 22259–22273. Curran Associates, Red Hook.

[81] RÁCZ, M. Z. and ZHANG, J. (2025). Harnessing multiple correlated networks for exact community recovery. In *Advances in Neural Information Processing Systems* (*NIPS*). *Curran Associates* **37**. Inc..

[82] ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915. MR2893856 https://doi.org/10.1214/11-AOS887

[83] SCHRAMM, T. and WEIN, A. S. (2022). Computational barriers to estimation from low-degree polynomials. *Ann. Statist.* **50** 1833–1858. MR4441142 https://doi.org/10.1214/22-aos2179

[84] SHIRANI, F., GARG, S. and ERKIP, E. (2017). Seeded graph matching: Efficient algorithms and theoretical guarantees. In *IEEE* 51*st Asilomar Conference on Signals*, *Systems*, *and Computers* 253–257. IEEE.

[85] SINGH, R., XU, J. and BERGER, B. (2008). Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proc. Natl. Acad. Sci. USA* **105** 12763–12768.

[86] SNIJDERS, T. A. B. and NOWICKI, K. (1997). Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *J. Classification* **14** 75–100. MR1449742 https://doi.org/10.1007/s003579900004

[87] VOGELSTEIN, J. T., CONROY, J. M., LYZINSKI, V., PODRAZIK, L. J., KRATZER, S. G., HARLEY, E. T., FISHKIND, D. E., VOGELSTEIN, R. J. and PRIEBE, C. E. (2015). Fast approximate quadratic programming for graph matching. *PLoS ONE* **10** 1–17.

[88] WEIN, A. S. (2021). Optimal low-degree hardness of maximum independent set. *Math. Stat. Learn.* **4** 221–251. MR4383734

[89] WEIN, A. S. (2025). Computational complexity of statistics: New insights from low-degree polynomials. arXiv preprint, arXiv:2506.10748.

[90] WU, Y., XU, J. and YU, S. H. (2022). Settling the sharp reconstruction thresholds of random graph matching. *IEEE Trans. Inf. Theory* **68** 5391–5417. MR4476403 https://doi.org/10.1109/tit.2022.3169005

[91] WU, Y., XU, J. and YU, S. H. (2023). Testing correlation of unlabeled random graphs. *Ann. Appl. Probab.* **33** 2519–2558. MR4612649 https://doi.org/10.1214/22-aap1786

[92] YANG, J. and CHUNG, H. W. (2023). Graph matching in correlated stochastic block models for improved graph clustering. arXiv preprint, arXiv:2309.05182.

[93] YANG, J., SHIN, D. and CHUNG, H. W. (2023). Efficient algorithms for exact graph matching on correlated stochastic block models with constant correlation. In Proceedings of the 40th International Conference on Machine Learning (ICML). PMLR.

[94] YARTSEVA, L. and GROSSGLAUSER, M. (2013). On the performance of percolation graph matching. In *Proceedings of the* 1*st ACM Conference on Online Social Networks* 119–130. ACM, New York.

*The Annals of Statistics*
2026, Vol. 54, No. 1, 252–276
https://doi.org/10.1214/25-AOS2566

# PSEUDO-LABELING FOR KERNEL RIDGE REGRESSION UNDER COVARIATE SHIFT

BY KAIZHENG WANG[a]

*Department of IEOR and Data Science Institute, Columbia University,* [a]*kaizheng.wang@columbia.edu*

We develop and analyze a principled approach to kernel ridge regression under covariate shift. The goal is to learn a regression function with small mean squared error over a target distribution, based on unlabeled data from there and labeled data that may have a different feature distribution. We propose to split the labeled data into two subsets, and conduct kernel ridge regression on them separately to obtain a collection of candidate models and an imputation model. We use the latter to fill the missing labels and then select the best candidate accordingly. Our nonasymptotic excess risk bounds demonstrate that our estimator adapts effectively to both the structure of the target distribution and the covariate shift. This adaptation is quantified through a notion of effective sample size that reflects the value of labeled source data for the target regression task. Our estimator achieves the minimax optimal error rate up to a polylogarithmic factor, and we find that using pseudo-labels for model selection does not significantly hinder performance.

## REFERENCES

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

[2] 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

[3] BLANCHARD, G. and MASSART, P. (2006). Discussion: Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.* **34** 2664–2671. MR2329460 https://doi.org/10.1214/009053606000001037

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

[5] CAI, T., GAO, R., LEE, J. and LEI, Q. (2021). A theory of label propagation for subpopulation shift. In *Proceedings of the* 38*th International Conference on Machine Learning* (M. Meila and T. Zhang, eds.). *Proceedings of Machine Learning Research* **139** 1170–1182. PMLR.

[6] CAI, T. T. and WEI, H. (2021). Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *Ann. Statist.* **49** 100–128. MR4206671 https://doi.org/10.1214/20-AOS1949

[7] CAMERON, A. C., GELBACH, J. B. and MILLER, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *Rev. Econ. Stat.* **90** 414–427.

[8] CAPONNETTO, A. and YAO, Y. (2010). Cross-validation based adaptation for regularization operators in learning theory. *Anal. Appl.* (*Singap.*) **8** 161–183. MR2646909 https://doi.org/10.1142/S0219530510001564

[9] CHENG, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *J. Amer. Statist. Assoc.* **89** 81–87.

[10] CORTES, C., MANSOUR, Y. and MOHRI, M. (2010). Learning bounds for importance weighting. In *Advances in Neural Information Processing Systems* (J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel and A. Culotta, eds.) **23**. Curran Associates, Red Hook.

[11] DONOHO, D. L. (1994). Statistical estimation and optimal recovery. *Ann. Statist.* **22** 238–270. MR1272082 https://doi.org/10.1214/aos/1176325367

[12] DUAN, Y., JIA, Z. and WANG, M. (2020). Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning* 2701–2709. PMLR.

---

*MSC2020 subject classifications.* Primary 62J07; secondary 62G05.
*Key words and phrases.* Covariate shift, kernel ridge regression, imputation, pseudo-labeling, transfer learning.

[13] Duan, Y., Wang, M. and Wainwright, M. J. (2024). Optimal policy evaluation using kernel-based temporal difference methods. *Ann. Statist.* **52** 1927–1952. MR4829475 https://doi.org/10.1214/24-aos2399

[14] Feng, X., He, X., Wang, C., Wang, C. and Zhang, J. (2024). Towards a unified analysis of kernel-based methods under covariate shift. *Adv. Neural Inf. Process. Syst.* **36**.

[15] Hanneke, S. and Kpotufe, S. (2019). On the value of target data in transfer learning. *Adv. Neural Inf. Process. Syst.* **32**.

[16] Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica* **47** 153–161. MR0518832 https://doi.org/10.2307/1912352

[17] Hirshberg, D. A., Maleki, A. and Zubizarreta, J. R. (2019). Minimax linear estimation of the retargeted mean. arXiv preprint. Available at arXiv:1901.10296.

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

[19] Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

[20] Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B. and Smola, A. (2006). Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems* (B. Schölkopf, J. Platt and T. Hoffman, eds.) **19**. MIT Press, Cambridge.

[21] Koltchinskii, V. and Lounici, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli* **23** 110–133. MR3556768 https://doi.org/10.3150/15-BEJ730

[22] 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

[23] Kumar, A., Ma, T. and Liang, P. (2020). Understanding self-training for gradual domain adaptation. In *Proceedings of the* 37*th International Conference on Machine Learning* (H. D. III and A. Singh, eds.). *Proceedings of Machine Learning Research* **119** 5468–5479. PMLR.

[24] Lee, D.-H. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning*. *ICML* **3** 896.

[25] Lei, Q., Hu, W. and Lee, J. (2021). Near-optimal linear regression under distribution shift. In *International Conference on Machine Learning* 6164–6174. PMLR.

[26] Lepskii, O. (1991). On a problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35** 454–466. MR1091202 https://doi.org/10.1137/1135065

[27] Li, K. C. (1982). Minimaxity of the method of regularization on stochastic processes. *Ann. Statist.* **10** 937–942. MR0663444

[28] Lin, H. and Reimherr, M. (2024). Smoothness adaptive hypothesis transfer learning. In *Proceedings of the* 41*st International Conference on Machine Learning*. *ICML*'24. Available at JMLR.org.

[29] Liu, H., Wang, J. and Long, M. (2021). Cycle self-training for domain adaptation. *Adv. Neural Inf. Process. Syst.* **34** 22968–22981.

[30] Liu, M., Zhang, Y., Liao, K. P. and Cai, T. (2023). Augmented transfer regression learning with semi-non-parametric nuisance models. *J. Mach. Learn. Res.* **24** Paper No. [293], 26. pp.; 24 pp. (appendix). MR4664730

[31] Ma, C., Pathak, R. and Wainwright, M. J. (2023). Optimally tackling covariate shift in RKHS-based nonparametric regression. *Ann. Statist.* **51** 738–761. MR4601000 https://doi.org/10.1214/23-aos2268

[32] Maity, S., Sun, Y. and Banerjee, M. (2022). Minimax optimal approaches to the label shift problem in non-parametric settings. *J. Mach. Learn. Res.* **23** Paper No. [346], 45. MR4577785

[33] Mou, W., Ding, P., Wainwright, M. J. and Bartlett, P. L. (2023). Kernel-based off-policy estimation without overlap: Instance optimality beyond semiparametric efficiency. arXiv preprint. Available at arXiv:2301.06240.

[34] Mou, W., Wainwright, M. J. and Bartlett, P. L. (2022). Off-policy estimation of linear functionals: Non-asymptotic theory for semi-parametric efficiency. arXiv preprint. Available at arXiv:2209.13075.

[35] Mousavi Kalan, M., Fabian, Z., Avestimehr, S. and Soltanolkotabi, M. (2020). Minimax lower bounds for transfer learning with linear and one-hidden layer neural networks. *Adv. Neural Inf. Process. Syst.* **33** 1959–1969.

[36] Page, S. and Grünewälder, S. (2021). The Goldenshluger–Lepski method for constrained least-squares estimators over RKHSs. *Bernoulli* **27** 2241–2266. MR4303882 https://doi.org/10.3150/20-BEJ1307

[37] Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22** 1345–1359.

[38] Pathak, R., Ma, C. and Wainwright, M. (2022). A new similarity measure for covariate shift with applications to nonparametric regression. In *International Conference on Machine Learning* 17517–17530. PMLR.

[39] Reeve, H. W. J., Cannings, T. I. and Samworth, R. J. (2021). Adaptive transfer learning. *Ann. Statist.* **49** 3618–3649. MR4352543 https://doi.org/10.1214/21-aos2102

[40] SCHMIDT-HIEBER, J. and ZAMOLODTCHIKOV, P. (2024). Local convergence rates of the nonparametric least squares estimator with applications to transfer learning. *Bernoulli* **30** 1845–1877. MR4746591 https://doi.org/10.3150/23-bej1655

[41] 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

[42] SPECKMAN, P. (1979). Minimax estimates of linear functionals in a Hilbert space. Unpublished manuscript.

[43] SUGIYAMA, M. and KAWANABE, M. (2012). *Machine Learning in Non-stationary Environments*: *Introduction to Covariate Shift Adaptation*. MIT press, Cambridge.

[44] TRIPURANENI, N., ADLAM, B. and PENNINGTON, J. (2021). Covariate shift in high-dimensional random feature regression. arXiv preprint. Available at arXiv:2111.08234.

[45] VAPNIK, V. N. (2000). *The Nature of Statistical Learning Theory*, 2nd ed. *Statistics for Engineering and Information Science*. Springer, New York. MR1719582 https://doi.org/10.1007/978-1-4757-3264-1

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

[47] 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

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

[49] WANG, H. and KIM, J. K. (2023). Statistical inference using regularized M-estimation in the reproducing kernel Hilbert space for handling missing data. *Ann. Inst. Statist. Math.* **75** 911–929. MR4655783 https://doi.org/10.1007/s10463-023-00872-8

[50] WANG, K. (2026). Supplement to "Pseudo-Labeling for Kernel Ridge Regression under Covariate Shift." https://doi.org/10.1214/25-AOS2566SUPPA, https://doi.org/10.1214/25-AOS2566SUPPB

[51] WU, J., ZOU, D., BRAVERMAN, V., GU, Q. and KAKADE, S. (2022). The power and limitation of pretraining-finetuning for linear regression under covariate shift. In *Advances in Neural Information Processing Systems* (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho and A. Oh, eds.) **35** 33041–33053. Curran Associates, Red Hook.

[52] YANG, F., ZHANG, H. R., WU, S., RÉ, C. and SU, W. J. (2025). Precise high-dimensional asymptotics for quantifying heterogeneous transfers. *J. Mach. Learn. Res.* **26** Paper No. [113], 88. MR4933084

[53] ZADROZNY, B. (2004). Learning and evaluating classifiers under sample selection bias. In *Proceedings of the Twenty-First International Conference on Machine Learning*. *ICML '04* **114** Assoc. Comput. Mach., New York, NY, USA. https://doi.org/10.1145/1015330.1015425

[54] ZHANG, T. (2005). Learning bounds for kernel regression using effective data dimensionality. *Neural Comput.* **17** 2077–2098. MR2175849 https://doi.org/10.1162/0899766054323008

*The Annals of Statistics*
2026, Vol. 54, No. 1, 277–299
https://doi.org/10.1214/25-AOS2567

# CHANGE POINT ESTIMATION FOR A STOCHASTIC HEAT EQUATION

BY MARKUS REISS[1,a], CLAUDIA STRAUCH[2,b] AND LUKAS TROTTNER[3,c]

[1]*Institut für Mathematik, Humboldt-Universität zu Berlin,* [a]*mreiss@math.hu-berlin.de*
[2]*Institute for Mathematics, Heidelberg University,* [b]*strauch@math.uni-heidelberg.de*
[3]*Institute for Stochastics and Applications, University of Stuttgart,* [c]*lukas.trottner@isa.uni-stuttgart.de*

We study a change point model based on a stochastic partial differential equation (SPDE) corresponding to the heat equation governed by the weighted Laplacian $\Delta_\vartheta = \nabla\vartheta\nabla$, where $\vartheta = \vartheta(x)$ is a space dependent diffusivity. As a basic problem, the domain $(0, 1)$ is considered with a piecewise constant diffusivity with a jump at an unknown point $\tau$. Based on local measurements of the solution in space with resolution $\delta$ over a finite time horizon, we construct a simultaneous M-estimator for the diffusivity values and the change point. The change point estimator converges with rate $\delta$, while the diffusivity constants can be recovered with convergence rate $\delta^{3/2}$. Furthermore, when the diffusivity parameters are known and the jump height vanishes as the spatial resolution tends to zero, we derive a limit theorem for the change point estimator and identify the limiting distribution. For the mathematical analysis, a precise understanding of the SPDE with discontinuous $\vartheta$, tight concentration bounds for quadratic functionals in the solution and a generalization of classical M-estimators are developed.

## REFERENCES

[1] ALTMEYER, R., BRETSCHNEIDER, T., JANÁK, J. and REISS, M. (2022). Parameter estimation in an SPDE model for cell repolarization. *SIAM/ASA J. Uncertain. Quantificat.* **10** 179–199. MR4376300 https://doi.org/10.1137/20M1373347

[2] ALTMEYER, R., CIALENCO, I. and PASEMANN, G. (2023). Parameter estimation for semilinear SPDEs from local measurements. *Bernoulli* **29** 2035–2061. MR4580906 https://doi.org/10.3150/22-bej1531

[3] ALTMEYER, R., CIALENCO, I. and REISS, M. (2023). Statistics for SPDEs. Available at https://sites.google.com/view/stats4spdes.

[4] ALTMEYER, R. and REISS, M. (2021). Nonparametric estimation for linear SPDEs from local measurements. *Ann. Appl. Probab.* **31** 1–38. MR4254472 https://doi.org/10.1214/20-aap1581

[5] ALTMEYER, R., TIEPNER, A. and WAHL, M. (2024). Optimal parameter estimation for linear SPDEs from multiple measurements. *Ann. Statist.* **52** 1307–1333. MR4804810 https://doi.org/10.1214/24-aos2364

[6] ARONSON, D. G. (1968). Non-negative solutions of linear parabolic equations. *Ann. Sc. Norm. Super. Pisa Cl. Sci.* (3) **22** 607–694. MR0435594

[7] BAKRY, D., GENTIL, I. and LEDOUX, M. (2014). *Analysis and Geometry of Markov Diffusion Operators*. *Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **348**. Springer, Cham. MR3155209 https://doi.org/10.1007/978-3-319-00227-9

[8] BHATTACHARYA, P. K. (1994). Some aspects of change-point analysis. In *Change-Point Problems* (*South Hadley*, *MA*, 1992). *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **23** 28–56. IMS, Hayward, CA. MR1477912 https://doi.org/10.1214/lnms/1215463112

[9] BHATTACHARYA, P. K. and BROCKWELL, P. J. (1976/77). The minimum of an additive process with applications to signal estimation and storage theory. *Z. Wahrsch. Verw. Gebiete* **37** 51–75. MR0426166 https://doi.org/10.1007/BF00536298

[10] BILLINGSLEY, P. (1999). *Convergence of Probability Measures*, 2nd ed. *Wiley Series in Probability and Statistics*: *Probability and Statistics*. Wiley, New York. MR1700749 https://doi.org/10.1002/9780470316962

[11] CIALENCO, I. (2018). Statistical inference for SPDEs: An overview. *Stat. Inference Stoch. Process.* **21** 309–329. MR3824970 https://doi.org/10.1007/s11203-018-9177-9

---

*MSC2020 subject classifications.* Primary 60H15, 62F12; secondary 60F05.
*Key words and phrases.* Change point detection, stochastic heat equation, local measurements.

[12] CSÖRGŐ, M. and HORVÁTH, L. (1997). *Limit Theorems in Change-Point Analysis*. *Wiley Series in Probability and Statistics*. Wiley, Chichester. MR2743035

[13] DAVIES, E. B. (1990). *Heat Kernels and Spectral Theory*. *Cambridge Tracts in Mathematics* **92**. Cambridge Univ. Press, Cambridge. MR1103113

[14] DUBEY, P. and MÜLLER, H.-G. (2020). Fréchet change-point detection. *Ann. Statist.* **48** 3312–3335. MR4185810 https://doi.org/10.1214/19-AOS1930

[15] ENGEL, K.-J. and NAGEL, R. (2000). *One-Parameter Semigroups for Linear Evolution Equations*. *Graduate Texts in Mathematics* **194**. Springer, New York. MR1721989

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

[17] ÉTORÉ, P. (2006). On random walk simulation of one-dimensional diffusion processes with discontinuous coefficients. *Electron. J. Probab.* **11** 249–275. MR2217816 https://doi.org/10.1214/EJP.v11-311

[18] FUKUSHIMA, M., OSHIMA, Y. and TAKEDA, M. (2011). *Dirichlet Forms and Symmetric Markov Processes*, extended ed. *De Gruyter Studies in Mathematics* **19**. de Gruyter, Berlin. MR2778606

[19] IBRAGIMOV, I. A. and HAS'MINSKIĬ, R. Z. (1981). *Statistical Estimation*: *Asymptotic Theory*. *Applications of Mathematics* **16**. Springer, New York. MR0620321

[20] JANÁK, J. and REISS, M. (2024). Parameter estimation for the stochastic heat equation with multiplicative noise from local measurements. *Stoch. Process. Appl.* **175** Paper No. 104385. MR4757468 https://doi.org/10.1016/j.spa.2024.104385

[21] KOSOROK, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. *Springer Series in Statistics*. Springer, New York. MR2724368 https://doi.org/10.1007/978-0-387-74978-5

[22] KUTOYANTS, Y. A. (2004). *Statistical Inference for Ergodic Diffusion Processes*. *Springer Series in Statistics*. Springer, London. MR2144185 https://doi.org/10.1007/978-1-4471-3866-2

[23] LI, J., CHEN, L., WANG, W. and WU, W. B. (2024). $\ell^2$ inference for change points in high-dimensional time series via a two-way MOSUM. *Ann. Statist.* **52** 602–627. MR4744189 https://doi.org/10.1214/24-aos2360

[24] LIPTSER, R. S. and SHIRYAYEV, A. N. (1989). *Theory of Martingales*. *Mathematics and Its Applications* (*Soviet Series*) **49**. Kluwer Academic, Dordrecht. MR1022664 https://doi.org/10.1007/978-94-009-2438-3

[25] LIU, H., GAO, C. and SAMWORTH, R. J. (2021). Minimax rates in sparse, high-dimensional change point detection. *Ann. Statist.* **49** 1081–1112. MR4255120 https://doi.org/10.1214/20-aos1994

[26] NOURDIN, I. and VIENS, F. G. (2009). Density formula and concentration inequalities with Malliavin calculus. *Electron. J. Probab.* **14** 2287–2309. MR2556018 https://doi.org/10.1214/EJP.v14-707

[27] PAGE, E. S. (1954). Continuous inspection schemes. *Biometrika* **41** 100–115. MR0088850 https://doi.org/10.1093/biomet/41.1-2.100

[28] REISS, M., STRAUCH, C. and TROTTNER, L. (2026). Supplement to "Change point estimation for a stochastic heat equation." https://doi.org/10.1214/25-AOS2567SUPP

[29] TIEPNER, A. and TROTTNER, L. (2026). Multivariate change estimation for a stochastic heat equation from local measurements. *Stoch. Process. Appl.* **192** Paper No. 104832. MR4988260 https://doi.org/10.1016/j.spa.2025.104832

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

[31] WALD, A. (1945). Sequential tests of statistical hypotheses. *Ann. Math. Stat.* **16** 117–186. MR0013275 https://doi.org/10.1214/aoms/1177731118

[32] WANG, D., YU, Y. and RINALDO, A. (2020). Univariate mean change point detection: Penalization, CUSUM and optimality. *Electron. J. Stat.* **14** 1917–1961. MR4091859 https://doi.org/10.1214/20-EJS1710

[33] WANG, D., YU, Y. and RINALDO, A. (2021). Optimal change point detection and localization in sparse dynamic networks. *Ann. Statist.* **49** 203–232. MR4206675 https://doi.org/10.1214/20-AOS1953

[34] WANG, R., ZHU, C., VOLGUSHEV, S. and SHAO, X. (2022). Inference for change points in high-dimensional data via selfnormalization. *Ann. Statist.* **50** 781–806. MR4405366 https://doi.org/10.1214/21-aos2127

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

[36] WILLIAMS, D. (1974). Path decomposition and continuity of local time for one-dimensional diffusions. I. *Proc. Lond. Math. Soc.* (3) **28** 738–768. MR0350881 https://doi.org/10.1112/plms/s3-28.4.738

[37] XU, H., WANG, D., ZHAO, Z. and YU, Y. (2024). Change-point inference in high-dimensional regression models under temporal dependence. *Ann. Statist.* **52** 999–1026. MR4784067 https://doi.org/10.1214/24-aos2380

*The Annals of Statistics*
2026, Vol. 54, No. 1, 300–323
https://doi.org/10.1214/25-AOS2568

# MULTIVARIATE ROOT-*n*-CONSISTENT SMOOTHING PARAMETER-FREE MATCHING ESTIMATORS AND ESTIMATORS OF INVERSE DENSITY WEIGHTED EXPECTATIONS

BY HAJO HOLZMANN[1,a] AND ALEXANDER MEISTER[2,b]

[1]*Mathematik und Informatik, Philipps-Universität Marburg,* [a]*holzmann@mathematik.uni-marburg.de*
[2]*Institut für Mathematik, Universität Rostock,* [b]*alexander.meister@uni-rostock.de*

Expected values weighted by the inverse of a multivariate density or, equivalently, Lebesgue integrals of regression functions with multivariate regressors occur in various areas of applications, including estimating average treatment effects, nonparametric estimators in random coefficient regression models or deconvolution estimators in Berkson errors-in-variables models. The frequently used nearest-neighbor and matching estimators suffer from bias problems in multiple dimensions. By using polynomial least squares fits on each cell of the $K$th-order Voronoi tessellation for sufficiently large $K$, we develop novel modifications of nearest-neighbor and matching estimators, which again converge at the parametric $\sqrt{n}$-rate under mild smoothness assumptions on the unknown regression function and without any smoothness conditions on the unknown density of the covariates. We stress that in contrast to competing methods for correcting for the bias of matching estimators, our estimators do not involve nonparametric function estimators and in particular do not rely on sample-size dependent smoothing parameters. We complement the upper bounds with appropriate lower bounds derived from information-theoretic arguments, which show that some smoothness of the regression function is indeed required to achieve the parametric rate. Simulations illustrate the practical feasibility of the proposed methods.

## REFERENCES

ABADIE, A. and IMBENS, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica* **74** 235–267. MR2194325 https://doi.org/10.1111/j.1468-0262.2006.00655.x

ABADIE, A. and IMBENS, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *J. Bus. Econom. Statist.* **29** 1–11. MR2789386 https://doi.org/10.1198/jbes.2009.07333

AUDIBERT, J.-Y. and TSYBAKOV, A. B. (2007). Fast learning rates for plug-in classifiers. *Ann. Statist.* **35** 608–633. MR2336861 https://doi.org/10.1214/009053606000001217

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

CHERNOZHUKOV, V., HANSEN, C., KALLUS, N., SPINDLER, M. and SYRGKANIS, V. (2024). Applied causal inference powered by ML and AI. arXiv preprint. Available at arXiv:2403.02467.

DAHLKE, S., HEUER, S., HOLZMANN, H. and TAFO, P. (2022). Statistically optimal estimation of signals in modulation spaces using Gabor frames. *IEEE Trans. Inf. Theory* **68** 4182–4200. MR4433275 https://doi.org/10.1109/tit.2022.3152734

DELAIGLE, A., HALL, P. and QIU, P. (2006). Nonparametric methods for solving the Berkson errors-in-variables problem. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 201–220. MR2188982 https://doi.org/10.1111/j.1467-9868.2006.00540.x

DEVROYE, L., GYÖRFI, L., LUGOSI, G. and WALK, H. (2017). On the measure of Voronoi cells. *J. Appl. Probab.* **54** 394–408. MR3668473 https://doi.org/10.1017/jpr.2017.7

DUNKER, F., ECKLE, K., PROKSCH, K. and SCHMIDT-HIEBER, J. (2019). Tests for qualitative features in the random coefficients model. *Electron. J. Stat.* **13** 2257–2306. MR3980958 https://doi.org/10.1214/19-EJS1570

---

*MSC2020 subject classifications.* Primary 62H12; secondary 62G05.
*Key words and phrases.* Average treatment effects, Berkson errors in variables models, bias correction, inverse density weighted expectations, matching estimators, random coefficients, Voronoi tessellation.

Gaillac, C. and Gautier, E. (2022). Adaptive estimation in the linear random coefficients model when regressors have limited variation. *Bernoulli* **28** 504–524. MR4337714 https://doi.org/10.3150/21-bej1354

Gautier, E. and Kitamura, Y. (2013). Nonparametric estimation in random coefficients binary choice models. *Econometrica* **81** 581–607. MR3043343 https://doi.org/10.3982/ECTA8675

Hoderlein, S., Holzmann, H. and Meister, A. (2017). The triangular model with random coefficients. *J. Econometrics* **201** 144–169. MR3704681 https://doi.org/10.1016/j.jeconom.2017.05.018

Hoderlein, S., Klemelä, J. and Mammen, E. (2010). Analyzing the random coefficient model nonparametrically. *Econometric Theory* **26** 804–837. MR2646480 https://doi.org/10.1017/S0266466609990119

Holzmann, H. and Meister, A. (2020). Rate-optimal nonparametric estimation for random coefficient regression models. *Bernoulli* **26** 2790–2814. MR4140529 https://doi.org/10.3150/20-BEJ1207

Holzmann, H. and Meister, A. (2026). Supplement to "Multivariate root-$n$-consistent smoothing parameter-free matching estimators and estimators of inverse density weighted expectations." https://doi.org/10.1214/25-AOS2568SUPP

Hu, A. J., Green, A. and Tibshirani, R. J. (2022). The Voronoigram: Minimax estimation of bounded variation functions from scattered data. arXiv preprint. Available at arXiv:2212.14514.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev. Econ. Stat.* **86** 4–29.

Kallus, N. (2020). Generalized optimal matching methods for causal inference. *J. Mach. Learn. Res.* **21** Paper No. 62, 54. MR4095341

Kerkyacharian, G. and Picard, D. (2004). Regression in random design and warped wavelets. *Bernoulli* **10** 1053–1105. MR2108043 https://doi.org/10.3150/bj/1106314850

Kohler, M. (2014). Optimal global rates of convergence for noiseless regression estimation problems with adaptively chosen design. *J. Multivariate Anal.* **132** 197–208. MR3266270 https://doi.org/10.1016/j.jmva.2014.08.008

Kouw, W. M. and Loog, M. (2019). An introduction to domain adaptation and transfer learning. arXiv preprint. Available at arXiv:1812.11806.

Lewbel, A. and Schennach, S. M. (2007). A simple ordered data estimator for inverse density weighted expectations. *J. Econometrics* **136** 189–211. MR2328590 https://doi.org/10.1016/j.jeconom.2005.08.005

Lin, Z., Ding, P. and Han, F. (2023). Estimation based on nearest neighbor matching: From density ratio to average treatment effect. *Econometrica* **91** 2187–2217. MR4680059 https://doi.org/10.3982/ecta20598

Masten, M. A. (2018). Random coefficients on endogenous variables in simultaneous equations models. *Rev. Econ. Stud.* **85** 1193–1250. MR3782604 https://doi.org/10.1093/restud/rdx047

Meister, A. (2009). *Deconvolution Problems in Nonparametric Statistics*. *Lecture Notes in Statistics* **193**. Springer, Berlin. MR2768576 https://doi.org/10.1007/978-3-540-87557-4

Portier, F., Truquet, L. and Yamane, I. (2023). Scalable and hyper-parameter-free non-parametric covariate shift adaptation with conditional sampling. arXiv preprint. Available at arXiv:2312.09969.

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

Sharpnack, J. (2023). On $L^2$-consistency of nearest neighbor matching. *IEEE Trans. Inf. Theory* **69** 3978–3988. MR4604477 https://doi.org/10.1109/tit.2022.3226479

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**.

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

Wang, Y. and Shah, R. D. (2024). Debiased inverse propensity score weighting for estimation of average treatment effects with high-dimensional confounders. *Ann. Statist.* **52** 1978–2003. MR4829477 https://doi.org/10.1214/24-aos2409

*The Annals of Statistics*
2026, Vol. 54, No. 1, 324–349
https://doi.org/10.1214/25-AOS2569

# BERRY–ESSEEN BOUNDS FOR DESIGN-BASED CAUSAL INFERENCE WITH POSSIBLY DIVERGING TREATMENT LEVELS AND VARYING GROUP SIZES

BY LEI SHI[1,a] AND PENG DING[2,b]

[1]*Division of Biostatistics, University of California, Berkeley,* [a]*leishi@berkeley.edu*
[2]*Department of Statistics, University of California, Berkeley,* [b]*pengdingpku@berkeley.edu*

Neyman (1923/1990) introduced the randomization model, which contains the notation of potential outcomes to define causal effects and a framework for large-sample inference based on the design of the experiment. However, the existing theory for this framework is far from complete, especially when the number of treatment levels diverges and the treatment group sizes vary. We provide a unified discussion of statistical inference under the randomization model with general treatment group sizes. We formulate the estimator in terms of a linear permutation statistic and use results based on Stein's method to derive various Berry–Esseen bounds on the linear and quadratic functions of the estimator. These new Berry–Esseen bounds serve as the basis for design-based causal inference with possibly diverging treatment levels and a diverging number of causal parameters of interest. We also fill an important gap by proposing novel variance estimators for experiments with possibly many treatment levels without replications. Equipped with the newly developed results, design-based causal inference in general settings becomes more convenient with stronger theoretical guarantees.

## REFERENCES

ABADIE, A. and IMBENS, G. W. (2008). Estimation of the conditional variance in paired experiments. *Ann. Écon. Stat.* 175–187.

ATHEY, S. and IMBENS, G. W. (2017). The econometrics of randomized experiments. In *Handbook of Economic Field Experiments* (A. Banerjee and E. Duflo, eds.) **1** 73–140. North-Holland, Amsterdam. 3.

BAUER, D. J., STERBA, S. K. and HALLFORS, D. D. (2008). Evaluating group-based interventions when control participants are ungrouped. *Multivar. Behav. Res.* **43** 210–236.

BOLTHAUSEN, E. (1984). An estimate of the remainder in a combinatorial central limit theorem. *Z. Wahrsch. Verw. Gebiete* **66** 379–386. MR0751577 https://doi.org/10.1007/BF00533704

BREIDT, F. J., OPSOMER, J. D. and SANCHEZ-BORREGO, I. (2016). Nonparametric variance estimation under fine stratification: An alternative to collapsed strata. *J. Amer. Statist. Assoc.* **111** 822–833. MR3538708 https://doi.org/10.1080/01621459.2015.1058264

CAUGHEY, D., KATSUMATA, H. and YAMAMOTO, T. (2019). Item response theory for conjoint survey experiments. Technical Report, Working Paper.

CHATTERJEE, S. and MECKES, E. (2008). Multivariate normal approximation using exchangeable pairs. *ALEA Lat. Amer. J. Probab. Math. Stat.* **4** 257–283. MR2453473

COCHRAN, W. G. (1977). *Sampling Techniques*, 3rd ed. *Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. MR0474575

COPAS, J. B. (1973). Randomization models for the matched and unmatched $2 \times 2$ tables. *Biometrika* **60** 467–476. MR0448746 https://doi.org/10.2307/2334995

DASGUPTA, T., PILLAI, N. S. and RUBIN, D. B. (2015). Causal inference from $2^K$ factorial designs by using potential outcomes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 727–753. MR3382595 https://doi.org/10.1111/rssb.12085

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.

---

*MSC2020 subject classifications.* 62K15, 62J05, 62G05.
*Key words and phrases.* Central limit theorem, permutation, potential outcome, Stein's method, randomized experiment.

El Karoui, N. and Purdom, E. (2018). Can we trust the bootstrap in high-dimensions? The case of linear models. *J. Mach. Learn. Res.* **19** 170–235. MR3862412

Espinosa, V., Dasgupta, T. and Rubin, D. B. (2016). A Bayesian perspective on the analysis of unreplicated factorial experiments using potential outcomes. *Technometrics* **58** 62–73. MR3463157 https://doi.org/10.1080/00401706.2015.1006337

Fang, X. and Röllin, A. (2015). Rates of convergence for multivariate normal approximation with applications to dense graphs and doubly indexed permutation statistics. *Bernoulli* **21** 2157–2189. MR3378463 https://doi.org/10.3150/14-BEJ639

Fogarty, C. B. (2018a). Regression-assisted inference for the average treatment effect in paired experiments. *Biometrika* **105** 994–1000. MR3877880 https://doi.org/10.1093/biomet/asy034

Fogarty, C. B. (2018b). On mitigating the analytical limitations of finely stratified experiments. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 1035–1056. MR3874309 https://doi.org/10.1111/rssb.12290

Fraser, D. A. S. (1956). A vector form of the Wald–Wolfowitz–Hoeffding theorem. *Ann. Math. Statist.* **27** 540–543. MR0078597 https://doi.org/10.1214/aoms/1177728279

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

Freedman, D. A. (2008b). On regression adjustments in experiments with several treatments. *Ann. Appl. Stat.* **2** 176–196. MR2415599 https://doi.org/10.1214/07-AOAS143

Guo, K. and Basse, G. (2023). The generalized Oaxaca–Blinder estimator. *J. Amer. Statist. Assoc.* **118** 524–536. MR4571139 https://doi.org/10.1080/01621459.2021.1941053

Hainmueller, J., Hopkins, D. J. and Yamamoto, T. (2014). Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Polit. Anal.* **22** 1–30.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Magy. Tud. Akad. Mat. Kut. Intéz. Közl.* **5** 361–374. MR0125612

Hansen, M. H., Hurwitz, W. N. and Madow, W. G. (1953a). *Sample Survey Methods and Theory. Vol. I. Methods and Applications*. Wiley, New York. MR0058171

Hansen, M. H., Hurwitz, W. N. and Madow, W. G. (1953b). *Sample Survey Methods and Theory. Vol. II. Theory*. Wiley, New York. MR0058172

Hinkelmann, K. and Kempthorne, O. (2007). *Design and Analysis of Experiments, Introduction to Experimental Design* **1**. Wiley, New York.

Hoeffding, W. (1951). A combinatorial central limit theorem. *Ann. Math. Stat.* **22** 558–566. MR0044058 https://doi.org/10.1214/aoms/1177729545

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

Kempthorne, O. (1952). *The Design and Analysis of Experiments*. Wiley, New York. MR0045368

Lei, L., Bickel, P. J. and El Karoui, N. (2018). Asymptotics for high dimensional regression $M$-estimates: Fixed design results. *Probab. Theory Related Fields* **172** 983–1079. MR3877551 https://doi.org/10.1007/s00440-017-0824-7

Lei, L. and Ding, P. (2021). Regression adjustment in completely randomized experiments with a diverging number of covariates. *Biometrika* **108** 815–828. MR4341353 https://doi.org/10.1093/biomet/asaa103

Li, X. and Ding, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *J. Amer. Statist. Assoc.* **112** 1759–1769. MR3750897 https://doi.org/10.1080/01621459.2017.1295865

Li, X., Ding, P. and Rubin, D. B. (2020). Rerandomization in $2^K$ factorial experiments. *Ann. Statist.* **48** 43–63. MR4065152 https://doi.org/10.1214/18-AOS1790

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

Lu, J. (2016). Covariate adjustment in randomization-based causal inference for $2^K$ factorial designs. *Statist. Probab. Lett.* **119** 11–20. MR3555264 https://doi.org/10.1016/j.spl.2016.07.010

Masoero, L., Vijaykumar, S., Richardson, T., McQueen, J., Rosen, I., Burdick, B., Bajari, P. and Imbens, G. (2024). Multiple randomization designs: Estimation and inference with interference. arXiv preprint. Available at arXiv:2401.01264.

Mukerjee, R., Dasgupta, T. and Rubin, D. B. (2018). Using standard tools from finite population sampling to improve causal inference for complex experiments. *J. Amer. Statist. Assoc.* **113** 868–881. MR3832233 https://doi.org/10.1080/01621459.2017.1294076

Mukerjee, R. and Wu, C. F. J. (2006). *A Modern Theory of Factorial Designs*. *Springer Series in Statistics*. Springer, New York. MR2230487

Neyman, J. (1923/1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–472. MR1092986

Robins, J. M. (1988). Confidence intervals for causal parameters. *Stat. Med.* **7** 773–785.

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

SCHEFFÉ, H. (1959). *The Analysis of Variance*. Wiley, New York. MR0116429

SHI, L. and DING, P. (2026). Supplement to "Berry–Esseen bounds for design-based causal inference with possibly diverging treatment levels and varying group sizes." https://doi.org/10.1214/25-AOS2569SUPP

SHI, L., WANG, J. and DING, P. (2025). Forward selection and post-selection inference in factorial designs. *Ann. Statist.* **53** 506–530. MR4900156 https://doi.org/10.1214/24-aos2454

WANG, Y. and LI, X. (2022). Rerandomization with diminishing covariate imbalance and diverging number of covariates. *Ann. Statist.* **50** 3439–3465. MR4524503 https://doi.org/10.1214/22-aos2235

WOLTER, K. M. (2007). *Introduction to Variance Estimation*, 2nd ed. *Statistics for Social and Behavioral Sciences*. Springer, New York. MR2288542

WU, C. F. J. and HAMADA, M. S. (2021). *Experiments—Planning, Analysis, and Optimization*. *Wiley Series in Probability and Statistics*. Wiley, Hoboken, NJ. Third edition [of 1780411]. MR4484975 https://doi.org/10.1002/9781119470007

WU, J. and DING, P. (2021). Randomization tests for weak null hypotheses in randomized experiments. *J. Amer. Statist. Assoc.* **116** 1898–1913. MR4353721 https://doi.org/10.1080/01621459.2020.1750415

ZHAO, A. and DING, P. (2022). Regression-based causal inference with factorial experiments: Estimands, model specifications and design-based properties. *Biometrika* **109** 799–815. MR4472849 https://doi.org/10.1093/biomet/asab051

ZHAO, A. and DING, P. (2023). Covariate adjustment in multi-armed, possibly factorial experiments. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **85** 1–23. MR4726956 https://doi.org/10.1093/jrsssb/qkac003

ZHIRKOV, K. (2022). Estimating and using individual marginal component effects from conjoint experiments. *Polit. Anal.* **30** 236–249.

*The Annals of Statistics*
2026, Vol. 54, No. 1, 350–382
https://doi.org/10.1214/25-AOS2570

# NEURAL NETWORKS GENERALIZE ON LOW COMPLEXITY DATA

BY SOURAV CHATTERJEE[1,a] AND TIMOTHY SUDIJONO[2,b]

[1] *Department of Statistics and Mathematics, Stanford University,* [a] *souravc@stanford.edu*
[2] *Department of Statistics, Stanford University,* [b] *tsudijon@stanford.edu*

We show that feedforward neural networks with ReLU activation generalize on low complexity data, suitably defined. Given i.i.d. data generated from a simple programming language, the minimum description length (MDL) feedforward neural network which interpolates the data generalizes with high probability. We define this simple programming language, along with a notion of description length of such networks. We provide several examples on basic computational tasks, such as checking primality of a natural number. For primality testing, our theorem shows the following and more. Suppose that we draw an i.i.d. sample of $n$ numbers uniformly at random from 1 to $N$. For each number $x_i$, let $y_i = 1$ if $x_i$ is a prime and 0 if it is not. Then the interpolating MDL network accurately answers, with probability $1 - O((\ln N)/n)$, whether a newly drawn number between 1 and $N$ is a prime or not. Note that the network is not *designed* to detect primes; minimum description learning *discovers* a network which does so. Extensions to noisy data are also discussed, suggesting that MDL neural network interpolators can demonstrate tempered overfitting.

## REFERENCES

[1] ABBE, E., BENGIO, S., LOTFI, A. and RIZK, K. (2024). Generalization on the unseen, logic reasoning and degree curriculum. *J. Mach. Learn. Res.* **25** 331. MR4829175

[2] ABBE, E., BENGIO, S., LOTFI, A., SANDON, C. and SAREMI, O. (2024). How far can transformers reason? The locality barrier and inductive scratchpad. arXiv preprint. Available at arXiv:2406.06467.

[3] ABBE, E., BOIX-ADSERA, E., BRENNAN, M. S., BRESLER, G. and NAGARAJ, D. (2021). The staircase property: How hierarchical structure can guide deep learning. *Adv. Neural Inf. Process. Syst.* **34** 26989–27002.

[4] AHUJA, K. and MANSOURI, A. (2024). On provable length and compositional generalization. arXiv preprint. Available at arXiv:2402.04875.

[5] ARORA, S., GE, R., NEYSHABUR, B. and ZHANG, Y. (2018). Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning* 254–263. PMLR.

[6] BAI, Y., CHEN, F., WANG, H., XIONG, C. and MEI, S. (2023). Transformers as statisticians: Provable in-context learning with in-context algorithm selection. arXiv preprint. Available at arXiv:2306.04637.

[7] BALCÁZAR, J. L., GAVALDÀ, R. and SIEGELMANN, H. T. (1997). Computational power of neural networks: A characterization in terms of Kolmogorov complexity. *IEEE Trans. Inf. Theory* **43** 1175–1183. MR1454945 https://doi.org/10.1109/18.605580

[8] BARRON, A., RISSANEN, J. and YU, B. (1998). The minimum description length principle in coding and modeling **44** 2743–2760. MR1658898 https://doi.org/10.1109/18.720554

[9] BARRON, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inf. Theory* **39** 930–945. MR1237720 https://doi.org/10.1109/18.256500

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

[11] BARTLETT, P. L. and MAASS, W. (2003). Vapnik-Chervonenkis dimension of neural nets. In *The Handbook of Brain Theory and Neural Networks* 1188–1192.

[12] 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

---

*MSC2020 subject classifications.* Primary 68T07, 62H30; secondary 68Q32.
*Key words and phrases.* Neural networks, supervised learning, generalization error, minimum description length, Kolmogorov complexity.

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

[14] BHATTAMISHRA, S., PATEL, A., KANADE, V. and BLUNSOM, P. (2022). Simplicity bias in transformers and their ability to learn sparse Boolean functions. arXiv preprint. Available at arXiv:2211.12316.

[15] BROWN, T., MANN, B., RYDER, N., SUBBIAH, M., KAPLAN, J. D., DHARIWAL, P., NEELAKANTAN, A., SHYAM, P., SASTRY, G. et al. (2020). Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33** 1877–1901.

[16] BRUTZKUS, A., GLOBERSON, A., MALACH, E. and SHALEV-SHWARTZ, S. (2017). SGD learns overparameterized networks that provably generalize on linearly separable data. arXiv preprint. Available at arXiv:1710.10174.

[17] CHEN, M., JIANG, H., LIAO, W. and ZHAO, T. (2022). Nonparametric regression on low-dimensional manifolds using deep ReLU networks: Function approximation and statistical recovery. *Inf. Inference* **11** 1203–1253. MR4526322 https://doi.org/10.1093/imaiai/iaac001

[18] CHEN, Y., GILROY, S., MALETTI, A., MAY, J. and KNIGHT, K. (2017). Recurrent neural networks as weighted language recognizers. arXiv preprint. Available at arXiv:1711.05408.

[19] CLARK, P., TAFJORD, O. and RICHARDSON, K. (2020). Transformers as soft reasoners over language. arXiv preprint. Available at arXiv:2002.05867.

[20] CYBENKO, G. (1989). Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems* **2** 303–314. MR1015670 https://doi.org/10.1007/BF02551274

[21] GARG, S., TSIPRAS, D., LIANG, P. and VALIANT, G. (2022). *What Can Transformers Learn in-Context. A Case Study of Simple Function Classes*.

[22] GIANNOU, A., RAJPUT, S., SOHN, J.-Y., LEE, K., LEE, J. D. and PAPAILIOPOULOS, D. (2023). Looped transformers as programmable computers. In *International Conference on Machine Learning* 11398–11442. PMLR.

[23] GOLDBLUM, M., FINZI, M., ROWAN, K. and WILSON, A. G. (2023). The no free lunch theorem, Kolmogorov complexity, and the role of inductive biases in machine learning. arXiv preprint. Available at arXiv:2304.05366.

[24] GOLDT, S., MÉZARD, M., KRZAKALA, F. and ZDEBOROVÁ, L. (2020). Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Phys. Rev. X* **10** 041044.

[25] GRAU-MOYA, J., GENEWEIN, T., HUTTER, M., ORSEAU, L., DELÉTANG, G., CATT, E., RUOSS, A., WENLIANG, L. K., MATTERN, C. et al. (2024). Learning universal predictors. arXiv preprint. Available at arXiv:2401.14953.

[26] GRÜNWALD, P. D. (2007). *The Minimum Description Length Principle*. MIT Press, Cambridge.

[27] HAREL, I., HOZA, W., VARDI, G., EVRON, I., SREBRO, N. and SOUDRY, D. (2024). Provable tempered overfitting of minimal nets and typical nets. *Adv. Neural Inf. Process. Syst.* **37** 53458–53524.

[28] HINTON, G. and VAN CAMP, D. (1993). Keeping neural networks simple by minimising the description length of weights. In *Proceedings of COLT*-93 5–13.

[29] HOCHREITER, S. and SCHMIDHUBER, J. (1997). Flat minima. *Neural Comput.* **9** 1–42.

[30] HOEFFDING, W. (1994). Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding* 409–426.

[31] HORNIK, K., STINCHCOMBE, M. and WHITE, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Netw.* **2** 359–366.

[32] KORNOWSKI, G., YEHUDAI, G. and SHAMIR, O. (2023). From tempered to benign overfitting in relu neural networks. *Adv. Neural Inf. Process. Syst.* **36** 58011–58046.

[33] LAN, N., GEYER, M., CHEMLA, E. and KATZIR, R. (2022). Minimum description length recurrent neural networks. *Trans. Assoc. Comput. Linguist.* **10** 785–799.

[34] LANDAU, E. (1909). Über die Einteilung der positiven ganzen Zahlen in vier Klassen nach der Mindestzahl der zu ihrer additiven Zusammensetzung erforderlichen Quadrate.

[35] LI, M. and VITÁNYI, P. (2008). *An Introduction to Kolmogorov Complexity and Its Applications*, 3rd ed. *Texts in Computer Science*. Springer, New York. MR2494387 https://doi.org/10.1007/978-0-387-49820-1

[36] LI, Y. and LIANG, Y. (2018). Learning overparameterized neural networks via stochastic gradient descent on structured data. *Adv. Neural Inf. Process. Syst.* **31**.

[37] LIN, L., BAI, Y. and MEI, S. (2023). Transformers as decision makers: Provable in-context reinforcement learning via supervised pretraining. arXiv preprint. Available at arXiv:2310.08566.

[38] LINDNER, D., KRAMÁR, J., FARQUHAR, S., RAHTZ, M., MCGRATH, T. and MIKULIK, V. (2024). Tracr: Compiled transformers as a laboratory for interpretability. *Adv. Neural Inf. Process. Syst.* **36**.

[39] LIU, B., ASH, J. T., GOEL, S., KRISHNAMURTHY, A. and ZHANG, C. (2022). Transformers learn shortcuts to automata. arXiv preprint. Available at arXiv:2210.10749.

[40] LIVNI, R., SHALEV-SHWARTZ, S. and SHAMIR, O. (2014). On the computational efficiency of training neural networks. *Adv. Neural Inf. Process. Syst.* **27**.

[41] MALACH, E. and SHALEV-SHWARTZ, S. (2018). A provably correct algorithm for deep learning that actually works. arXiv preprint. Available at arXiv:1803.09522.

[42] MALI, A., ORORBIA, A., KIFER, D. and GILES, L. (2023). On the computational complexity and formal hierarchy of second order recurrent neural networks. arXiv preprint. Available at arXiv:2309.14691.

[43] MALLINAR, N., SIMON, J., ABEDSOLTAN, A., PANDIT, P., BELKIN, M. and NAKKIRAN, P. (2022). Benign, tempered, or catastrophic: Toward a refined taxonomy of overfitting. *Adv. Neural Inf. Process. Syst.* **35** 1182–1195.

[44] MANOJ, N. S. and SREBRO, N. (2023). Interpolation learning with minimum description length. arXiv preprint. Available at arXiv:2302.07263.

[45] MARSAGLIA, G. (2003). Xorshift rngs. *J. Stat. Softw.* **8** 1–6.

[46] MCCULLOCH, W. S. and PITTS, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **5** 115–133. MR0010388 https://doi.org/10.1007/bf02478259

[47] MEI, S. and WU, Y. (2025). Deep networks as denoising algorithms: Sample-efficient learning of diffusion models in high-dimensional graphical models. *IEEE Trans. Inf. Theory* **71** 2930–2954. MR4882452 https://doi.org/10.1109/tit.2025.3535923

[48] MÉZARD, M. (2023). Spin glass theory and its new challenge: Structured disorder. *Indian J. Phys.* 1–12.

[49] MINGARD, C., REES, H., VALLE-PÉREZ, G. and LOUIS, A. A. (2023). Do deep neural networks have an inbuilt Occam's razor? arXiv preprint. Available at arXiv:2304.06670.

[50] MINGARD, C., SKALSE, J., VALLE-PÉREZ, G., MARTÍNEZ-RUBIO, D., MIKULIK, V. and LOUIS, A. A. (2019). Neural networks are a priori biased towards Boolean functions with low entropy. arXiv preprint. Available at arXiv:1909.11522.

[51] NAKKIRAN, P., KAPLUN, G., BANSAL, Y., YANG, T., BARAK, B. and SUTSKEVER, I. (2021). Deep double descent: Where bigger models and more data hurt. *J. Stat. Mech. Theory Exp.* **12** 124003. MR4412831 https://doi.org/10.1088/1742-5468/ac3a74

[52] NANDA, N., CHAN, L., LIEBERUM, T., SMITH, J. and STEINHARDT, J. (2023). Progress measures for grokking via mechanistic interpretability. arXiv preprint. Available at arXiv:2301.05217.

[53] PÉREZ, J., BARCELÓ, P. and MARINKOVIC, J. (2021). Attention is Turing complete. *J. Mach. Learn. Res.* **22** 75. MR4253768

[54] PÉREZ, J., MARINKOVIĆ, J. and BARCELÓ, P. (2019). On the Turing completeness of modern neural network architectures. arXiv preprint. Available at arXiv:1901.03429.

[55] RAZIN, N. (2024). Understanding deep learning via notions of rank. arXiv preprint. Available at arXiv:2408.02111.

[56] RISSANEN, J. (1983). A universal prior for integers and estimation by minimum description length. *Ann. Statist.* **11** 416–431. MR0696056 https://doi.org/10.1214/aos/1176346150

[57] SANFORD, C., HSU, D. and TELGARSKY, M. (2023). Representational strengths and limitations of transformers. arXiv preprint. Available at arXiv:2306.02896.

[58] SCHMIDHUBER, J. (1997). Discovering neural nets with low Kolmogorov complexity and high generalization capability. *Neural Netw.* **10** 857–873.

[59] SHALEV-SHWARTZ, S. and BEN-DAVID, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge Univ. Press, Cambridge.

[60] SIEGELMANN, H. T. and SONTAG, E. D. (1992). On the computational power of neural nets. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* 440–449.

[61] SOLOMONOFF, R. J. (1964). A formal theory of inductive inference. I. *Inf. Control* **7** 1–22. MR0172744

[62] SONTAG, E. D. et al. (1998). VC dimension of neural networks. *NATO ASI Ser., Ser. F: Comput. Syst. Sci.* **168** 69–96.

[63] STOGIN, J., MALI, A. and GILES, C. L. (2024). A provably stable neural network Turing machine with finite precision and time. *Inform. Sci.* **658** 120034.

[64] STROBL, L., MERRILL, W., WEISS, G., CHIANG, D. and ANGLUIN, D. (2024). What formal languages can transformers express? A survey. *Trans. Assoc. Comput. Linguist.* **12** 543–561.

[65] SVETE, A. and COTTERELL, R. (2024). Transformers can represent $n$-gram language models. arXiv preprint. Available at arXiv:2404.14994.

[66] TENEY, D., NICOLICIOIU, A., HARTMANN, V. and ABBASNEJAD, E. (2024). Neural redshift: Random networks are not random functions. arXiv preprint. Available at arXiv:2403.02241.

[67] TSIGLER, A. and BARTLETT, P. L. (2023). Benign overfitting in ridge regression. *J. Mach. Learn. Res.* **24** 123. MR4583284

[68] VALLE-PEREZ, G., CAMARGO, C. Q. and LOUIS, A. A. (2018). Deep learning generalizes because the parameter-function map is biased towards simple functions. arXiv preprint. Available at arXiv:1805.08522.

[69] VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, Ł. and POLOSUKHIN, I. (2017). Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**.

[70] WANG, Z., WEI, S., HSU, D. and LEE, J. D. (2024). Transformers provably learn sparse token selection while fully-connected nets cannot. arXiv preprint. Available at arXiv:2406.06893.

[71] WEI, C., CHEN, Y. and MA, T. (2022). Statistically meaningful approximation: A case study on approximating Turing machines with transformers. *Adv. Neural Inf. Process. Syst.* **35** 12071–12083.

[72] WEISS, G., GOLDBERG, Y. and YAHAV, E. (2021). Thinking like transformers. In *International Conference on Machine Learning* 11080–11090. PMLR.

[73] YANG, Y., MANDT, S., THEIS, L. et al. (2023). An introduction to neural data compression. *Found. Trends Comput. Graph. Vis.* **15** 113–200.

[74] ZHANG, C., BENGIO, S., HARDT, M., RECHT, B. and VINYALS, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Commun. ACM* **64** 107–115.

[75] ZHOU, H., BRADLEY, A., LITTWIN, E., RAZIN, N., SAREMI, O., SUSSKIND, J., BENGIO, S. and NAKKIRAN, P. (2023). What algorithms can transformers learn? A study in length generalization. arXiv preprint. Available at arXiv:2310.16028.

[76] ZHU, H., LI, Z. and HAYASHI, M. (2022). Nearly tight universal bounds for the binomial tail probabilities. arXiv preprint. Available at arXiv:2211.01688.

*The Annals of Statistics*
2026, Vol. 54, No. 1, 383–407
https://doi.org/10.1214/25-AOS2571

# SEMIPARAMETRIC BERNSTEIN–VON MISES PHENOMENON VIA ISOTONIZED POSTERIOR IN WICKSELL'S PROBLEM

BY FRANCESCO GILI[a], GEURT JONGBLOED[b] AND AAD VAN DER VAART[c]

*Delft Institute of Applied Mathematics (DIAM), Delft University of Technology,* [a]*F.Gili@tudelft.nl,* [b]*G.Jongbloed@tudelft.nl,* [c]*A.W.vanderVaart@tudelft.nl*

In this paper, we propose a novel Bayesian approach for nonparametric estimation in Wicksell's problem. This has important applications in astronomy for estimating the distribution of the positions of the stars in a galaxy given projected stellar positions and in materials science to determine the 3D microstructure of a material, using its 2D cross-sections. We deviate from the classical Bayesian nonparametric approach, which would place a Dirichlet Process (DP) prior on the distribution function of the unobservables, by directly placing a DP prior on the distribution function of the observables. Our method offers computational simplicity due to the conjugacy of the posterior and allows for asymptotically efficient estimation by projecting the posterior onto the $\mathbb{L}_2$ subspace of increasing, right-continuous functions. Indeed, the resulting Isotonized Inverse Posterior (IIP) satisfies a Bernstein–von Mises (BvM) phenomenon with minimax asymptotic variance $g_0(x)/2\gamma$, where $\gamma > 1/2$ reflects the degree of Hölder continuity of the true cdf at $x$. Since the IIP gives automatic uncertainty quantification, it eliminates the need to estimate $\gamma$. Our results provide the first semiparametric Bernstein–von Mises theorem for projection-based posteriors with a DP prior in inverse problems.

## REFERENCES

[1] ALONSO, J., MATIO, M. and SEN, B. (2007). Uncloaking globular clusters of the inner galaxy. *Proceedings of the International Astronomical Union* **3** 359–360. https://doi.org/10.1088/0004-6256/143/3/70
[2] ANTONIAK, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistic* **2** 1152–1174. https://doi.org/10.1214/aos/1176342871
[3] AYER, M., BRUNK, H. D., EWING, G. M., REID, W. T. and SILVERMAN, E. (1955). An empirical distribution function for sampling with incomplete information. *Ann. Math. Statist.* **26** 641–647. MR0073895 https://doi.org/10.1214/aoms/1177728423
[4] BHAUMIK, P. and GHOSAL, S. (2015). Bayesian two-step estimation in differential equation models. *Electron. J. Stat.* **9** 3124–3154. MR3453972 https://doi.org/10.1214/15-EJS1099
[5] BHAUMIK, P. and GHOSAL, S. (2017). Efficient Bayesian estimation and uncertainty quantification in ordinary differential equation models. *Bernoulli* **23** 3537–3570. MR3654815 https://doi.org/10.3150/16-BEJ856
[6] CASTILLO, I. and NICKL, R. (2013). Nonparametric Bernstein-von Mises theorems in Gaussian white noise. *Ann. Statist.* **41** 1999–2028. MR3127856 https://doi.org/10.1214/13-AOS1133
[7] CASTILLO, I. and NICKL, R. (2014). On the Bernstein-von Mises phenomenon for nonparametric Bayes procedures. *Ann. Statist.* **42** 1941–1969. MR3262473 https://doi.org/10.1214/14-AOS1246
[8] CASTILLO, I. and ROUSSEAU, J. (2015). A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *Ann. Statist.* **43** 2353–2383. MR3405597 https://doi.org/10.1214/15-AOS1336
[9] CHAKRABORTY, M. and GHOSAL, S. (2021). Coverage of credible intervals in nonparametric monotone regression. *Ann. Statist.* **49** 1011–1028. MR4255117 https://doi.org/10.1214/20-aos1989
[10] CHAKRABORTY, M. and GHOSAL, S. (2022). Rates and coverage for monotone densities using projection-posterior. *Bernoulli* **28** 1093–1119. MR4388931 https://doi.org/10.3150/21-bej1379
[11] CHAN, K. C. G. and QIN, J. (2016). Nonparametric maximum likelihood estimation for the multisample Wicksell corpuscle problem. *Biometrika* **103** 273–286. MR3509886 https://doi.org/10.1093/biomet/asw011

---

*MSC2020 subject classifications.* Primary 62F15, 62C10, 62G20, 62G05; secondary 62C20, 62E20.
*Key words and phrases.* Semiparametric estimation, Bernstein–von Mises, Bayesian nonparametrics, inverse problems, isotonic estimation, efficiency theory.

[12] Cuzzi, J. and Olson, D. (2017). Recovering 3D particle size distributions from 2D sections. *Meteoritics & Planetary Science* **52** 532–545. https://doi.org/10.1111/maps.12812

[13] Deng, H., Han, Q. and Zhang, C.-H. (2021). Confidence intervals for multiple isotonic regression and other monotone models. *Ann. Statist.* **49** 2021–2052. MR4319240 https://doi.org/10.1214/20-aos2025

[14] Efromovich, S. (2010). A sharp minimax lower bound for the nonparametric estimation of Sobolev densities of order $\frac{1}{2}$. *Statist. Probab. Lett.* **80** 77–81. MR2563888 https://doi.org/10.1016/j.spl.2009.08.011

[15] Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2** 615–629. https://doi.org/10.1214/aos/1176342752

[16] 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

[17] Gili, F., Jongbloed, G. and van der Vaart, A. (2025). Adaptive and efficient isotonic estimation in Wicksell's problem. *J. Nonparametr. Stat.* **37** 491–531. MR4948920 https://doi.org/10.1080/10485252.2024.2397680

[18] Gili, F., Jongbloed, G. and van der Vaart, A. (2026). Supplement to "Semiparametric Bernstein–von Mises phenomenon via Isotonized Posterior in Wicksell's problem." https://doi.org/10.1214/25-AOS2571SUPP

[19] Gili, F., Jongbloed, G. and van der Vaart, A. (2026). Asymptotically efficient estimation under local constraint in Wicksell's problem. *J. Statist. Plann. Inference* **240** Paper No. 106299, 20. MR4908067 https://doi.org/10.1016/j.jspi.2025.106299

[20] Golubev, G. K. and Enikeeva, F. N. (2001). Asymptotically efficient smoothing in the Wicksell problem under squared losses. *Probl. Inf. Transm.* **37** 12–21. https://doi.org/10.1023/A:1010495609901

[21] Golubev, G. K. and Levit, B. Y. (1998). Asymptotically efficient estimation in the Wicksell problem. *Ann. Statist.* **26** 2407–2419. MR1700238 https://doi.org/10.1214/aos/1024691477

[22] Groeneboom, P. and Jongbloed, G. (1995). Isotonic estimation and rates of convergence in Wicksell's problem. *The Annals of Statistic* **23** 1528–1542. https://doi.org/10.1214/aos/1176324310

[23] Groeneboom, P. and Jongbloed, G. (2014). *Nonparametric Estimation Under Shape Constraints: Estimators, Algorithms and Asymptotics. Cambridge Series in Statistical and Probabilistic Mathematics* **38**. Cambridge Univ. Press, New York. MR3445293 https://doi.org/10.1017/CBO9781139020893

[24] Hall, P. and Smith, R. L. (1988). The kernel method for unfolding sphere size distributions. *J. Comput. Phys.* **74** 409–421. MR0930210 https://doi.org/10.1016/0021-9991(88)90085-X

[25] Hjort, N. L. and Petrone, S. (2007). Nonparametric quantile inference using Dirichlet processes. In *Advances in Statistical Modeling and Inference. Ser. Biostat.* **3** 463–492. World Scientific, Hackensack, NJ. MR2416129 https://doi.org/10.1142/9789812708298_0023

[26] Jongbloed, G. (2001). Sieved maximum likelihood estimation in Wicksell's problem and related deconvolution problems. *Scand. J. Statist.* **28** 161–183. MR1844355 https://doi.org/10.1111/1467-9469.00230

[27] Kim, J. and Pollard, D. (1990). Cube root asymptotics. *Ann. Statist.* **18** 191–219. https://doi.org/10.1214/aos/1176347498

[28] Kim, Y. (2006). The Bernstein–von Mises theorem for the proportional hazard model. *Ann. Statist.* **34** 1678–1700. https://doi.org/10.1214/009053606000000533

[29] Kim, Y. and Lee, J. (2004). A Bernstein-von Mises theorem in the nonparametric right-censoring model. *Ann. Statist.* **32** 1492–1512. MR2089131 https://doi.org/10.1214/009053604000000526

[30] Lopez-Sanchez, M. A. and Llana-Fúnez, S. (2016). An extension of the saltykov method to quantify 3D grain size distributions in mylonites. *Journal of Structural Geology* **93** 149–161. https://doi.org/10.1016/j.jsg.2016.10.008

[31] Monard, F., Nickl, R. and Paternain, G. P. (2019). Efficient nonparametric Bayesian inference for $X$-ray transforms. *Ann. Statist.* **47** 1113–1147. MR3909962 https://doi.org/10.1214/18-AOS1708

[32] Nickl, R. (2020). Bernstein–von Mises theorems for statistical inverse problems I: Schrödinger equation. *J. Eur. Math. Soc.* (*JEMS*) **22** 2697–2750. MR4118619 https://doi.org/10.4171/JEMS/975

[33] Nickl, R. and Söhl, J. (2019). Bernstein-von Mises theorems for statistical inverse problems II: Compound Poisson processes. *Electron. J. Stat.* **13** 3513–3571. MR4013745 https://doi.org/10.1214/19-ejs1609

[34] Ray, K. (2017). Adaptive Bernstein–von Mises theorems in Gaussian white noise. *Ann. Statist.* **45** 2511–2536. MR3737900 https://doi.org/10.1214/16-AOS1533

[35] Ray, K. and van der Vaart, A. (2020). Semiparametric Bayesian causal inference. *Ann. Statist.* **48** 2999–3020. MR4152632 https://doi.org/10.1214/19-AOS1919

[36] Rivoirard, V. and Rousseau, J. (2012). Bernstein-von Mises theorem for linear functionals of the density. *Ann. Statist.* **40** 1489–1523. MR3015033 https://doi.org/10.1214/12-AOS1004

[37] Sen, B., Banerjee, M. and Woodroofe, M. (2010). Inconsistency of bootstrap: The Grenander estimator. *Ann. Statist.* **38** 1953–1977. MR2676880 https://doi.org/10.1214/09-AOS777

[38] SEN, B. and WOODROOFE, M. (2012). Bootstrap confidence intervals for isotonic estimators in a stereological problem. *Bernoulli* **18** 1249–1266. MR2995794 https://doi.org/10.3150/12-BEJ378

[39] STOYAN, D. and KENDALL, W. S. (1987). *Stochastic Geometry and Its Applications*. Wiley, New York.

[40] VAN DER VAART, A. W. and WELLNER, J. A. (2023). *Weak Convergence and Empirical Processes—with Applications to Statistics*. *Springer Series in Statistics*. Springer, Cham. MR4628026 https://doi.org/10.1007/978-3-031-29040-4

[41] WANG, K. and GHOSAL, S. (2023). Coverage of credible intervals in Bayesian multivariate isotonic regression. *Ann. Statist.* **51** 1376–1400. MR4630953 https://doi.org/10.1214/23-aos2298

[42] WANG, K. and GHOSAL, S. (2023). Posterior contraction and testing for multivariate isotonic regression. *Electron. J. Stat.* **17** 798–822. MR4554662 https://doi.org/10.1214/23-ejs2115

[43] WATSON, G. S. (1971). Estimating functionals of particle size distributions. *Biometrika* **58** 483–490. MR0312639 https://doi.org/10.1093/biomet/58.3.483

[44] WICKSELL, S. D. (1925). The corpuscle problem. *Biometrika* **17** 84–99. https://doi.org/10.1093/biomet/17.1-2.84

*The Annals of Statistics*
2026, Vol. 54, No. 1, 408–441
https://doi.org/10.1214/25-AOS2572

# OPTIMAL CONVEX $M$-ESTIMATION VIA SCORE MATCHING

BY OLIVER Y. FENG[1,3,a], YU-CHUN KAO[2,b], MIN XU[2,c] AND RICHARD J. SAMWORTH[3,d]

[1]*Department of Mathematical Sciences, University of Bath,* [a]*of402@bath.ac.uk*
[2]*Department of Statistics, Rutgers University,* [b]*yk495@scarletmail.rutgers.edu,* [c]*mx76@stat.rutgers.edu*
[3]*Statistical Laboratory, University of Cambridge,* [d]*r.samworth@statslab.cam.ac.uk*

In the context of linear regression, we construct a data-driven convex loss function with respect to which empirical risk minimisation yields optimal asymptotic variance in the downstream estimation of the regression coefficients. At the population level, the negative derivative of the optimal convex loss is the best decreasing approximation of the derivative of the log-density of the noise distribution. This motivates a fitting process via a nonparametric extension of score matching, corresponding to a log-concave projection of the noise distribution with respect to the Fisher divergence. At the sample level, our semiparametric estimator is computationally efficient, and we prove that it attains the minimal asymptotic covariance among all convex $M$-estimators. As an example of a non-log-concave setting, the optimal convex loss function for Cauchy errors is Huber-like, and our procedure yields asymptotic efficiency greater than 0.87 relative to the maximum likelihood estimator of the regression coefficients that uses oracle knowledge of this error distribution. In this sense, we provide robustness and facilitate computation without sacrificing much statistical efficiency. Numerical experiments using our accompanying `R` package `asm` confirm the practical merits of our proposal.

## REFERENCES

AMARI, S. and NAGAOKA, H. (2000). *Methods of Information Geometry*. *Translations of Mathematical Monographs* **191**. Amer. Math. Soc., Providence, RI. MR1800071 https://doi.org/10.1090/mmono/191

ARCONES, M. A. (1998). Asymptotic theory for $M$-estimators over a convex kernel. *Econometric Theory* **14** 387–422. MR1650029 https://doi.org/10.1017/S0266466698144018

BARBER, R. F. and SAMWORTH, R. J. (2021). Local continuity of log-concave projection, with applications to estimation under model misspecification. *Bernoulli* **27** 2437–2472. MR4303890 https://doi.org/10.3150/20-BEJ1316

BARRON, J. T. (2019). A general and adaptive robust loss function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 4331–4339.

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.

BENTON, J., SHI, Y., DE BORTOLI, V., DELIGIANNIDIS, G. and DOUCET, A. (2024). From denoising diffusions to denoising Markov models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **86** 286–301. MR4896938 https://doi.org/10.1093/jrsssb/qkae005

BERAN, R. (1978). An efficient and robust adaptive estimator of location. *Ann. Statist.* **6** 292–313. MR0518885

BETANCOURT, M., BYRNE, S., LIVINGSTONE, S. and GIROLAMI, M. (2017). The geometric foundations of Hamiltonian Monte Carlo. *Bernoulli* **23** 2257–2298. MR3648031 https://doi.org/10.3150/16-BEJ810

BICKEL, P. J. (1975). One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc.* **70** 428–434. MR0386168

BICKEL, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10** 647–671. MR0663424

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

BRUNEL, V.-E. (2023). Geodesically convex $M$-estimation in metric spaces. In *The Thirty Sixth Annual Conference on Learning Theory* 2188–2210. PMLR.

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

---

*MSC2020 subject classifications.* Primary 62G35; secondary 62J05.
*Key words and phrases.* $M$-estimation, robust statistics, log-concavity, antitonic, score matching.

CELENTANO, M. and MONTANARI, A. (2022). Fundamental barriers to high-dimensional regression with convex penalties. *Ann. Statist.* **50** 170–196. MR4382013 https://doi.org/10.1214/21-aos2100

CHENG, X., CHATTERJI, N. S., BARTLETT, P. L. and JORDAN, M. I. (2018). Underdamped Langevin MCMC: A non-asymptotic analysis. In *Conference on Learning Theory* 300–323. PMLR.

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

CHINOT, G., LECUÉ, G. and LERASLE, M. (2020). Robust statistical learning with Lipschitz and convex loss functions. *Probab. Theory Related Fields* **176** 897–940. MR4087486 https://doi.org/10.1007/s00440-019-00931-3

COOK, R. D. (1977). Detection of influential observation in linear regression. *Technometrics* **19** 15–18. MR0436478 https://doi.org/10.2307/1268249

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

COX, D. D. (1985). A penalty method for nonparametric estimation of the logarithmic derivative of a density function. *Ann. Inst. Statist. Math.* **37** 271–288. MR0799240 https://doi.org/10.1007/BF02481097

CULE, M., SAMWORTH, R. and STEWART, M. (2010). Maximum likelihood estimation of a multi-dimensional log-concave density. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 545–607. MR2758237 https://doi.org/10.1111/j.1467-9868.2010.00753.x

DALALYAN, A. S., GOLUBEV, G. K. and TSYBAKOV, A. B. (2006). Penalized maximum likelihood and semiparametric second-order efficiency. *Ann. Statist.* **34** 169–201. MR2275239 https://doi.org/10.1214/009053605000000895

DERENSKI, J., FAN, Y., JAMES, G. and XU, M. (2023). An empirical Bayes shrinkage method for functional data. Submitted.

DE BORTOLI, V., MATHIEU, E., HUTCHINSON, M., THORNTON, J., TEH, Y. W. and DOUCET, A. (2022). Riemannian score-based generative modelling. *Adv. Neural Inf. Process. Syst.* **35** 2406–2422.

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

DONOHO, D. L. and MONTANARI, A. (2015). Variance breakdown of Huber $M$-estimators: $n/p \in (1, \infty)$. arXiv Preprint. Available at arXiv:1503.02106.

DOSS, C. R. and WELLNER, J. A. (2019). Univariate log-concave density estimation with symmetry or modal constraints. *Electron. J. Stat.* **13** 2391–2461. MR3983344 https://doi.org/10.1214/19-EJS1574

DÜMBGEN, L., SAMWORTH, R. and SCHUHMACHER, D. (2011). Approximation by log-concave distributions, with applications to regression. *Ann. Statist.* **39** 702–730. MR2816336 https://doi.org/10.1214/10-AOS853

DÜMBGEN, L., SAMWORTH, R. J. and SCHUHMACHER, D. (2013). Stochastic search for semiparametric linear regression models. In *From Probability to Statistics and Back: High-Dimensional Models and Processes. Inst. Math. Stat. (IMS) Collect.* **9** 78–90. IMS, Beachwood, OH. MR3186750 https://doi.org/10.1214/12-IMSCOLL907

EFRON, B. (2011). Tweedie's formula and selection bias. *J. Amer. Statist. Assoc.* **106** 1602–1614. MR2896860 https://doi.org/10.1198/jasa.2011.tm11181

EGGERMONT, P. P. B. and LARICCIA, V. N. (2000). Maximum likelihood estimation of smooth monotone and unimodal densities. *Ann. Statist.* **28** 922–947. MR1792794 https://doi.org/10.1214/aos/1015952005

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

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.

FARAWAY, J. J. (1992). Smoothing in adaptive estimation. *Ann. Statist.* **20** 414–427. MR1150352 https://doi.org/10.1214/aos/1176348530

FENG, O. Y., KAO, Y.-C., XU, M. and SAMWORTH, R. J. (2026). Supplement to "Optimal convex $M$-estimation via score matching." https://doi.org/10.1214/25-AOS2572SUPP

GHOSH, S., IGNATIADIS, N., KOEHLER, F. and LEE, A. (2025). Stein's unbiased risk estimate and Hyvärinen's score matching. arXiv Preprint. Available at arXiv:2502.20123.

GROENEBOOM, P. and JONGBLOED, G. (2014). *Nonparametric Estimation Under Shape Constraints: Estimators, Algorithms and Asymptotics. Cambridge Series in Statistical and Probabilistic Mathematics* **38**. Cambridge Univ. Press, New York. MR3445293 https://doi.org/10.1017/CBO9781139020893

GUPTA, S., LEE, J. C. H. and PRICE, E. (2023). Finite-sample symmetric mean estimation with Fisher information rate. In *The Thirty Sixth Annual Conference on Learning Theory* 4777–4830. PMLR.

HAMPEL, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **69** 383–393. MR0362657

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

HANSEN, B. E. (2022). A modern Gauss–Markov theorem. *Econometrica* **90** 1283–1294. MR4436052 https://doi.org/10.3982/ecta19255

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

HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

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

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

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

HYVÄRINEN, A. (2005). Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.* **6** 695–709. MR2249836

HYVÄRINEN, A. (2007). Some extensions of score matching. *Comput. Statist. Data Anal.* **51** 2499–2512. MR2338984 https://doi.org/10.1016/j.csda.2006.09.003

JANKOVÁ, J., SHAH, R. D., BÜHLMANN, P. and SAMWORTH, R. J. (2020). Goodness-of-fit testing in high dimensional generalized linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 773–795. MR4112784 https://doi.org/10.1111/rssb.12371

JIN, K. (1990). *Empirical Smoothing Parameter Selection in Adaptive Estimation*. University of California, Berkeley.

JOHNSON, O. (2004). *Information Theory and the Central Limit Theorem*. Imperial College Press, London. MR2109042 https://doi.org/10.1142/9781860945373

JOHNSON, O. and BARRON, A. (2004). Fisher information inequalities and the central limit theorem. *Probab. Theory Related Fields* **129** 391–409. MR2128239 https://doi.org/10.1007/s00440-004-0344-0

JOLICOEUR-MARTINEAU, A., PICHÉ-TAILLEFER, R., DES COMBES, R. T. and MITLIAGKAS, I. (2020). Adversarial score matching and improved sampling for image generation. arXiv Preprint. Available at arXiv:2009.05475.

JONES, M. C. (1992). Estimating densities, quantiles, quantile densities and density quantiles. *Ann. Inst. Statist. Math.* **44** 721–727.

KAO, Y.-C., XU, M., FENG, O. Y. and SAMWORTH, R. J. (2024). asm: Optimal convex $M$-estimation for linear regression via antitonic score matching. R package version 0.2.4. Available at https://CRAN.R-project.org/package=asm.

KAO, Y.-C., XU, M. and ZHANG, C.-H. (2024). Choosing the $p$ in $L_p$ loss: Adaptive rates for symmetric mean estimation. In *The Thirty Seventh Annual Conference on Learning Theory* 2795–2839. PMLR.

KOEHLER, F., HECKETT, A. and RISTESKI, A. (2022). Statistical efficiency of score matching: The view from isoperimetry. arXiv Preprint. Available at arXiv:2210.00726.

LAHA, N. (2021). Adaptive estimation in symmetric location model under log-concavity constraint. *Electron. J. Stat.* **15** 2939–3014. MR4280163 https://doi.org/10.1214/21-ejs1852

LEDERER, J. and OESTING, M. (2023). Extremes in high dimensions: Methods and scalable algorithms. arXiv Preprint. Available at arXiv:2303.04258.

LEI, L. and WOOLDRIDGE, J. (2022). What estimators are unbiased for linear models? arXiv Preprint. Available at arXiv:2212.14185.

LERASLE, M. (2019). Selected topics on robust statistical learning theory. arXiv Preprint. Available at arXiv:1908.10761.

LEY, C. and SWAN, Y. (2013). Stein's density approach and information inequalities. *Electron. Commun. Probab.* **18** no. 7. MR3019670 https://doi.org/10.1214/ECP.v18-2578

LI, G., WEI, Y., CHEN, Y. and CHI, Y. (2024). Towards non-asymptotic convergence for diffusion-based generative models. In *The Twelfth International Conference on Learning Representations*.

LOH, P.-L. (2021). Scale calibration for high-dimensional robust regression. *Electron. J. Stat.* **15** 5933–5994. MR4355701 https://doi.org/10.1214/21-ejs1936

LYU, S. (2012). Interpretation and generalization of score matching. arXiv Preprint. Available at arXiv:1205.2629.

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

MAMMEN, E. and PARK, B. U. (1997). Optimal smoothing in adaptive location estimation. *J. Statist. Plann. Inference* **58** 333–348. MR1450020 https://doi.org/10.1016/S0378-3758(96)00085-7

MARDIA, K. V., KENT, J. T. and LAHA, A. K. (2016). Score matching estimators for directional distributions. arXiv Preprint. Available at arXiv:1604.08470.

MARONNA, R. A. and YOHAI, V. J. (1981). Asymptotic behavior of general $M$-estimates for regression and scale with random carriers. *Z. Wahrsch. Verw. Gebiete* **58** 7–20. MR0635268 https://doi.org/10.1007/BF00536192

NOCEDAL, J. and WRIGHT, S. J. (2006). *Numerical Optimization*, 2nd ed. *Springer Series in Operations Research and Financial Engineering*. Springer, New York. MR2244940

PARISI, G. (1981). Correlation functions and computer simulations. *Nuclear Phys. B* **180** 378–384. MR0615176 https://doi.org/10.1016/0550-3213(81)90056-0

PARZEN, E. (1979). Nonparametric statistical data modeling. *J. Amer. Statist. Assoc.* **74** 105–131. MR0529528

PORTNOY, S. (1985). Asymptotic behavior of $M$ estimators of $p$ regression parameters when $p^2/n$ is large. II. Normal approximation. *Ann. Statist.* **13** 1403–1417. MR0811499 https://doi.org/10.1214/aos/1176349744

PÖTSCHER, B. M. and PREINERSTORFER, D. (2024). A comment on: “A modern Gauss–Markov theorem”. *Econometrica* **92** 913–924. MR4766938 https://doi.org/10.3982/ecta20819

ROBERTS, G. O. and TWEEDIE, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* **2** 341–363. MR1440273 https://doi.org/10.2307/3318418

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

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

SAMWORTH, R. and JOHNSON, O. (2004). Convergence of the empirical process in Mallows distance, with an application to bootstrap performance. arXiv Preprint. Available at arXiv:math/0406603.

SAMWORTH, R. J. and SHAH, R. D. (2025). *Modern Statistical Methods and Theory*. Cambridge Univ. Press, Cambridge.

SCHICK, A. (1986). On asymptotically efficient estimation in semiparametric models. *Ann. Statist.* **14** 1139–1151. MR0856811 https://doi.org/10.1214/aos/1176350055

SILVERMAN, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.* **10** 795–810. MR0663433

SILVERMAN, B. W. (1986). *Density Estimation*. CRC Press, Boca Raton.

SONG, Y. and ERMON, S. (2019). Generative modeling by estimating gradients of the data distribution. *Adv. Neural Inf. Process. Syst.* **32** 11895–11907.

SONG, Y., GARG, S., SHI, J. and ERMON, S. (2020). Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in Artificial Intelligence* 574–584.

SONG, Y. and KINGMA, D. P. (2021). How to train your energy-based models. arXiv Preprint. Available at arXiv:2101.03288.

SONG, Y., SOHL-DICKSTEIN, J., KINGMA, D. P., KUMAR, A., ERMON, S. and POOLE, B. (2021). Score-based generative modeling through stochastic differential equations. In *The Ninth International Conference on Learning Representations*.

SRIPERUMBUDUR, B., FUKUMIZU, K., GRETTON, A., HYVÄRINEN, A. and KUMAR, R. (2017). Density estimation in infinite dimensional exponential families. *J. Mach. Learn. Res.* **18** Paper No. 57. MR3687600

STEIN, C. (1956a). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. I* 197–206. Univ. California Press, Berkeley, CA. MR0084922

STEIN, C. (1956b). Efficient nonparametric testing and estimation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. I* 187–195. Univ. California Press, Berkeley, CA. MR0084921

STONE, C. J. (1975). Adaptive maximum likelihood estimators of a location parameter. *Ann. Statist.* **3** 267–284. MR0362669

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

VAN DER VAART, A. W. and WELLNER, J. A. (2021). Stein 1956: Efficient nonparametric testing and estimation. *Ann. Statist.* **49** 1836–1849. MR4319232 https://doi.org/10.1214/21-aos2056

VAN EEDEN, C. (1970). Efficiency-robust estimation of location. *Ann. Math. Statist.* **41** 172–181. MR0263194 https://doi.org/10.1214/aoms/1177697197

VINCENT, P. (2011). A connection between score matching and denoising autoencoders. *Neural Comput.* **23** 1661–1674. MR2839543 https://doi.org/10.1162/NECO_a_00142

YANG, X. and WANG, T. (2024). Multiple-output composite quantile regression through an optimal transport lens. In *The Thirty Seventh Annual Conference on Learning Theory* **247** 5076–5122. PMLR.

YANG, Y., MARTIN, R. and BONDELL, H. (2019). Variational approximations using Fisher divergence. arXiv Preprint. Available at arXiv:1905.05284.

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

YOUNG, E. H. and SHAH, R. D. (2024). Sandwich boosting for accurate estimation in partially linear models for grouped data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **86** 1286–1311. MR4825003 https://doi.org/10.1093/jrsssb/qkae032

YU, M., GUPTA, V. and KOLAR, M. (2020). Simultaneous inference for pairwise graphical models with generalized score matching. *J. Mach. Learn. Res.* **21** Paper No. 91. MR4119159

YU, S., DRTON, M. and SHOJAIE, A. (2022). Generalized score matching for general domains. *Inf. Inference* **11** 739–780. MR4474347 https://doi.org/10.1093/imaiai/iaaa041

ZOU, H. and YUAN, M. (2008). Composite quantile regression and the oracle model selection theory. *Ann. Statist.* **36** 1108–1126. MR2418651 https://doi.org/10.1214/07-AOS507

*The Annals of Statistics*
2026, Vol. 54, No. 1, 442–465
https://doi.org/10.1214/25-AOS2573

# COMMUNICATION-EFFICIENT AND DISTRIBUTED-ORACLE ESTIMATION FOR HIGH-DIMENSIONAL QUANTILE REGRESSION

BY SONGSHAN YANG[1,a], YIFAN GU[2,b], HANFANG YANG[2,c] AND XUMING HE[3,d]

[1] *Center for Applied Statistics and Institute of Statistics and Big Data, Renmin University of China,* [a]*yangss@ruc.edu.cn*
[2] *Center for Applied Statistics and School of Statistics, Renmin University of China,* [b]*guyifan@ruc.edu.cn,* [c]*hyang@ruc.edu.cn*
[3] *Department of Statistics and Data Science, Washington University in St. Louis,* [d]*hex@wustl.edu*

In this article, we present a novel communication-efficient estimator for distributed high-dimensional quantile regression with folded-concave penalties. An **I**terative **M**ulti-**S**tep **A**lgorithm (IMSA) is employed to tackle the nonconvex challenge of the objective function, taking into account both the statistical accuracy and the communication constraints. We demonstrate that the proposed IMSA estimators share similar properties with the global folded-concave penalized estimator. To establish the theoretical results, we introduce a new concept called the *distributed-oracle estimator*. We prove that the proposed estimator converges to the distributed-oracle estimator with high probability. Compared to the $\ell_1$-penalized method, the IMSA estimator possesses a faster rate of convergence and requires milder conditions to achieve support recovery. Furthermore, we extend our framework to facilitate distributed inference for the preconceived low-dimensional components within the high-dimensional model. We derive the limiting distribution of the corresponding test statistic under the null hypothesis and the local alternatives. In addition, a new feature-splitting algorithm is devised to accommodate the high-dimensional data within the distributed system. Extensive numerical studies demonstrate the effectiveness and validity of our proposed estimation and inference methods. A real example is also presented for illustration.

## REFERENCES

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

BELLONI, A. and CHERNOZHUKOV, V. (2011). $\ell_1$-penalized quantile regression in high-dimensional sparse models. *Ann. Statist.* **39** 82–130. MR2797841 https://doi.org/10.1214/10-AOS827

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

CHEN, X., LIU, W., MAO, X. and YANG, Z. (2020). Distributed high-dimensional regression under a quantile loss function. *J. Mach. Learn. Res.* **21** Paper No. 182. MR4209468

CHENG, C., FENG, X., HUANG, J. and LIU, X. (2022). Regularized projection score estimation of treatment effects in high-dimensional quantile regression. *Statist. Sinica* **32** 23–41. MR4359621 https://doi.org/10.5705/ss.202019.0247

DI, F., WANG, L. and LIAN, H. (2022). Communication-efficient estimation and inference for high-dimensional quantile regression based on smoothed decorrelated score. *Stat. Med.* **41** 5084–5101. MR4499632 https://doi.org/10.1002/sim.9555

FAN, J., GUO, Y. and WANG, K. (2023). Communication-efficient accurate statistical estimation. *J. Amer. Statist. Assoc.* **118** 1000–1010. MR4595472 https://doi.org/10.1080/01621459.2021.1969238

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

---

*MSC2020 subject classifications.* Primary 62F05, 62F12; secondary 62J07.
*Key words and phrases.* Communication efficiency, distributed statistical inference, partial penalization, SCAD.

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

FAN, J., LIU, H., SUN, Q. and ZHANG, T. (2018). I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *Ann. Statist.* **46** 814–841. MR3782385 https://doi.org/10.1214/17-AOS1568

FAN, J., XUE, L. and ZOU, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Ann. Statist.* **42** 819–849. MR3210988 https://doi.org/10.1214/13-AOS1198

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

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

HALL, P. and SHEATHER, S. J. (1988). On the distribution of a Studentized quantile. *J. Roy. Statist. Soc. Ser. B, Methodol.* **50** 381–391. MR0970974

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

JIANG, R. and YU, K. (2021). Smoothing quantile regression for a distributed system. *Neurocomputing* **466** 311–326.

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

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

KOENKER, R., CHERNOZHUKOV, V., HE, X. and PENG, L. (2017). *Handbook of Quantile Regression*. CRC press, Boca Raton.

LEE, E. R., NOH, H. and PARK, B. U. (2014). Model selection via Bayesian information criterion for quantile regression models. *J. Amer. Statist. Assoc.* **109** 216–229. MR3180558 https://doi.org/10.1080/01621459.2013.836975

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

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

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

POWELL, J. L. (1991). Estimation of monotonic regression models under quantile restrictions. In *Nonparametric and Semiparametric Methods in Econometrics and Statistics* (*Durham, NC*, 1988) (W. Barnett, J. Powell and G. Tauchen, eds.). *Internat. Sympos. Econom. Theory Econometrics* 357–384. Cambridge Univ. Press, Cambridge. MR1174980

SHAMIR, O., SREBRO, N. and ZHANG, T. (2014). Communication-efficient distributed optimization using an approximate Newton-type method. In *Proceedings of the* 31*st International Conference on Machine Learning* **32** 1000–1008.

SHI, C., SONG, R., CHEN, Z. and LI, R. (2019). Linear hypothesis testing for high dimensional generalized linear models. *Ann. Statist.* **47** 2671–2703. MR3988769 https://doi.org/10.1214/18-AOS1761

SUN, Q. and ZHANG, H. (2021). Targeted inference involving high-dimensional data using nuisance penalized regression. *J. Amer. Statist. Assoc.* **116** 1472–1486. MR4309286 https://doi.org/10.1080/01621459.2020.1737079

TAN, K. M., BATTEY, H. and ZHOU, W.-X. (2022). Communication-constrained distributed quantile regression with optimal statistical guarantees. *J. Mach. Learn. Res.* **23** Paper No. [272]. MR4577711

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

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

VOLGUSHEV, S., CHAO, S.-K. and CHENG, G. (2019). Distributed inference for quantile regression processes. *Ann. Statist.* **47** 1634–1662. MR3911125 https://doi.org/10.1214/18-AOS1730

WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using $\ell_1$-constrained quadratic programming (Lasso). *IEEE Trans. Inf. Theory* **55** 2183–2202. MR2729873 https://doi.org/10.1109/TIT.2009.2016018

WANG, J., KOLAR, M., SREBRO, N. and ZHANG, T. (2017). Efficient distributed learning with sparsity. In *Proceedings of the* 34*th International Conference on Machine Learning* **70** 3636–3645.

WANG, L. and HE, X. (2024). Analysis of global and local optima of regularized quantile regression in high dimensions: A subgradient approach. *Econometric Theory* **40** 233–277. MR4713052 https://doi.org/10.1017/s0266466622000421

WANG, L., WU, Y. and LI, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *J. Amer. Statist. Assoc.* **107** 214–222. MR2949353 https://doi.org/10.1080/01621459.2012.656014

WEN, J., YANG, S., WANG, C. D., JIANG, Y. and LI, R. (2025). Feature-splitting algorithms for ultrahigh dimensional quantile regression. *J. Econometrics* **249** Paper No. 105426. MR4905994 https://doi.org/10.1016/j.jeconom.2023.01.028

WU, S., HUANG, D. and WANG, H. (2023). Quasi-Newton updating for large-scale distributed learning. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **85** 1326–1354. MR4718543 https://doi.org/10.1093/jrsssb/qkad059

YANG, S., GU, Y., YANG, H. and HE, X. (2026). Supplement to "Communication-efficient and distributed-oracle estimation for high-dimensional quantile regression." https://doi.org/10.1214/25-AOS2573SUPP

YU, L., LIN, N. and WANG, L. (2017). A parallel algorithm for large-scale nonconvex penalized quantile regression. *J. Comput. Graph. Statist.* **26** 935–939. MR3765357 https://doi.org/10.1080/10618600.2017.1328366

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

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

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

ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533. MR2435443 https://doi.org/10.1214/009053607000000802

*The Annals of Statistics*
2026, Vol. 54, No. 1, 466–489
https://doi.org/10.1214/25-AOS2575

# OPTIMALITY OF APPROXIMATE MESSAGE PASSING FOR SPIKED MATRIX MODELS WITH ROTATIONALLY INVARIANT NOISE

BY RISHABH DUDEJA[1,a], SONGBIN LIU[2,b] AND JUNJIE MA[2,c]

[1] *Department of Statistics, University of Wisconsin, Madison,* [a]*rdudeja@wisc.edu*
[2] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences,* [b]*liusongbin@lsec.cc.ac.cn,*
[c]*majunjie@lsec.cc.ac.cn*

We study the problem of estimating a rank-one signal matrix from a noisy observed matrix corrupted by additive rotationally invariant noise. We develop a new class of approximate message passing algorithms for this problem and provide a simple and concise characterization of their dynamics in the high-dimensional limit. At each iteration, these algorithms leverage prior knowledge about the noise structure by applying a nonlinear matrix denoiser to the eigenvalues of the observed matrix, and utilize prior information regarding the signal structure by applying a nonlinear iterate denoiser to the previous iterates generated by the algorithm. We derive the optimal choices for both the matrix and iterate denoisers and demonstrate that the resulting algorithm achieves the lowest possible asymptotic estimation error among a broad class of iterative algorithms under a fixed iteration budget.

## REFERENCES

[1] ABBE, E. (2018). Community detection and stochastic block models: Recent developments. *J. Mach. Learn. Res.* **18** Paper No. 177, 86. MR3827065

[2] 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

[3] 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

[4] BARBIER, J., CAMILLI, F., MONDELLI, M. and SÁENZ, M. (2023). Fundamental limits in structured principal component analysis and how to reach them. *Proc. Natl. Acad. Sci. USA* **120** Paper No. e2302028120, 7. MR4637889 https://doi.org/10.1073/pnas.2302028120

[5] BARBIER, J., CAMILLI, F., XU, Y. and MONDELLI, M. (2025). Information limits and Thouless–Anderson–Palmer equations for spiked matrix models with structured noise. *Phys. Rev. Res.* **7** 013081.

[6] BARBIER, J., DIA, M., MACRIS, N., KRZAKALA, F., LESIEUR, T. and ZDEBOROVÁ, L. (2016). Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula. *Adv. Neural Inf. Process. Syst.* **29**.

[7] BARBIER, J. and MACRIS, N. (2019). The adaptive interpolation method for proving replica formulas. Applications to the Curie–Weiss and Wigner spike models. *J. Phys. A: Math. Theor.* **52** 294002, 26. MR3983946 https://doi.org/10.1088/1751-8121/ab2735

[8] 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

[9] BEHNE, J. K. and REEVES, G. (2022). Fundamental limits for rank-one matrix estimation with groupwise heteroskedasticity. In *International Conference on Artificial Intelligence and Statistics* 8650–8672. PMLR.

[10] 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

---

*MSC2020 subject classifications.* Primary 62H99; secondary 62F15, 62E20, 68Q17.
*Key words and phrases.* Statistical-computational gap, low-rank matrix estimation, approximate message passing.

[11] BOLTHAUSEN, E. (2014). An iterative construction of solutions of the TAP equations for the Sherrington–Kirkpatrick model. *Comm. Math. Phys.* **325** 333–366. MR3147441 https://doi.org/10.1007/s00220-013-1862-3

[12] BUN, J., ALLEZ, R., BOUCHAUD, J.-P. and POTTERS, M. (2016). Rotational invariant estimator for general noisy matrices. *IEEE Trans. Inf. Theory* **62** 7475–7490. MR3599095 https://doi.org/10.1109/TIT.2016.2616132

[13] CELENTANO, M., FAN, Z. and MEI, S. (2023). Local convexity of the TAP free energy and AMP convergence for $\mathbb{Z}_2$-synchronization. *Ann. Statist.* **51** 519–546. MR4600991 https://doi.org/10.1214/23-aos2257

[14] CELENTANO, M., MONTANARI, A. and WU, Y. (2020). The estimation error of general first order methods. In *Conference on Learning Theory* 1078–1141. PMLR.

[15] CONSORTIUM, G. P. (2015). A global reference for human genetic variation. *Nature* **526** 68.

[16] CONSORTIUM, I. H. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature* **467** 52.

[17] DESHPANDE, Y., ABBE, E. and MONTANARI, A. (2017). Asymptotic mutual information for the balanced binary stochastic block model. *Inf. Inference* **6** 125–170. MR3671474 https://doi.org/10.1093/imaiai/iaw017

[18] DESHPANDE, Y. and MONTANARI, A. (2014). Information-theoretically optimal sparse PCA. In 2014 *IEEE International Symposium on Information Theory* 2197–2201. IEEE Press, New York.

[19] 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. With electronic supplementary materials available online. MR2546388 https://doi.org/10.1098/rsta.2009.0152

[20] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2009). Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci. USA* **106** 18914–18919.

[21] DUDEJA, R. and BAKHSHIZADEH, M. (2022). Universality of linearized message passing for phase retrieval with structured sensing matrices. *IEEE Trans. Inf. Theory* **68** 7545–7574. MR4524656 https://doi.org/10.1109/tit.2022.3182018

[22] DUDEJA, R., LIU, S. and MA, J. (2026). Supplement to "Optimality of Approximate Message Passing for Spiked Matrix Models with Rotationally Invariant Noise." https://doi.org/10.1214/25-AOS2575SUPP

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

[24] 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

[25] EL ALAOUI, A. and KRZAKALA, F. (2018). Estimation in the spiked Wigner model: A short proof of the replica formula. In 2018 *IEEE International Symposium on Information Theory* (*ISIT*) 1874–1878. IEEE Press, New York.

[26] EL ALAOUI, A., KRZAKALA, F. and JORDAN, M. (2020). Fundamental limits of detection in the spiked Wigner model. *Ann. Statist.* **48** 863–885. MR4102679 https://doi.org/10.1214/19-AOS1826

[27] FAN, Z. (2022). Approximate message passing algorithms for rotationally invariant matrices. *Ann. Statist.* **50** 197–224. MR4382014 https://doi.org/10.1214/21-aos2101

[28] FAN, Z., MEI, S. and MONTANARI, A. (2021). TAP free energy, spin glasses and variational inference. *Ann. Probab.* **49** 1–45. MR4203332 https://doi.org/10.1214/20-AOP1443

[29] 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.

[30] FÉRAL, D. and PÉCHÉ, S. (2007). The largest eigenvalue of rank one deformation of large Wigner matrices. *Comm. Math. Phys.* **272** 185–228. MR2291807 https://doi.org/10.1007/s00220-007-0209-3

[31] FLETCHER, A. K. and RANGAN, S. (2018). Iterative reconstruction of rank-one matrices in noise. *Inf. Inference* **7** 531–562. MR3858334 https://doi.org/10.1093/imaiai/iax014

[32] HENRICI, P. (1993). *Applied and Computational Complex Analysis, Volume* 3: *Discrete Fourier Analysis, Cauchy Integrals, Construction of Conformal Maps, Univalent Functions* **41**. Wiley, New York.

[33] 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

[34] 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

[35] KABASHIMA, Y. (2003). A CDMA multiuser detection algorithm on the basis of belief propagation. *J. Phys. A: Math. Gen.* **36** 11111–11121. MR2025247 https://doi.org/10.1088/0305-4470/36/43/030

[36] KABASHIMA, Y., KRZAKALA, F., MÉZARD, M., SAKATA, A. and ZDEBOROVÁ, L. (2016). Phase transitions and sample complexity in Bayes-optimal matrix factorization. *IEEE Trans. Inf. Theory* **62** 4228–4265. MR3515748 https://doi.org/10.1109/TIT.2016.2556702

[37] KNOWLES, A. and YIN, J. (2013). The isotropic semicircle law and deformation of Wigner matrices. *Comm. Pure Appl. Math.* **66** 1663–1749. MR3103909 https://doi.org/10.1002/cpa.21450

[38] KRZAKALA, F., XU, J. and ZDEBOROVÁ, L. (2016). Mutual information in rank-one matrix estimation. In 2016 *IEEE Information Theory Workshop* (*ITW*) 71–75. IEEE Press, New York.

[39] 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

[40] 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

[41] LELARGE, M. and MIOLANE, L. (2019). Fundamental limits of symmetric low-rank matrix estimation. *Probab. Theory Related Fields* **173** 859–929. MR3936148 https://doi.org/10.1007/s00440-018-0845-x

[42] LESIEUR, T., KRZAKALA, F. and ZDEBOROVÁ, L. (2015). MMSE of probabilistic low-rank matrix estimation: Universality with respect to the output channel. In 2015 53*rd Annual Allerton Conference on Communication*, *Control*, *and Computing* (*Allerton*) 680–687. IEEE Press, New York.

[43] LESIEUR, T., KRZAKALA, F. and ZDEBOROVÁ, L. (2015). Phase transitions in sparse PCA. In 2015 *IEEE International Symposium on Information Theory* (*ISIT*) 1635–1639. IEEE Press, New York.

[44] LI, G., FAN, W. and WEI, Y. (2023). Approximate message passing from random initialization with applications to $\mathbb{Z}_2$ synchronization. *Proc. Natl. Acad. Sci. USA* **120** Paper No. e2302930120, 7. MR4637851

[45] LI, G. and WEI, Y. (2022). A non-asymptotic framework for approximate message passing in spiked models. arXiv preprint. Available at arXiv:2208.03313.

[46] LI, Y. and SUR, P. (2023). Spectrum-aware adjustment: a new debiasing framework with applications to principal components regression. arXiv preprint. Available at arXiv:2309.07810.

[47] LOLAS, P. and YING, L. (2021). Shrinkage Estimation of Functions of Large Noisy Symmetric Matrices. arXiv preprint. Available at arXiv:2106.05183.

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

[49] MIOLANE, L. (2017). Fundamental limits of low-rank matrix estimation: the non-symmetric case. arXiv preprint. Available at arXiv:1702.00473.

[50] MONDELLI, M. and VENKATARAMANAN, R. (2021). PCA initialization for approximate message passing in rotationally invariant models. *Adv. Neural Inf. Process. Syst.* **34** 29616–29629.

[51] MONTANARI, A. and VENKATARAMANAN, R. (2021). Estimation of low-rank matrices via approximate message passing. *Ann. Statist.* **49** 321–345. MR4206680 https://doi.org/10.1214/20-AOS1958

[52] MONTANARI, A. and WEIN, A. S. (2025). Equivalence of approximate message passing and low-degree polynomials in rank-one matrix estimation. *Probab. Theory Related Fields* **191** 181–233. MR4869255 https://doi.org/10.1007/s00440-024-01322-z

[53] MONTANARI, A. and WU, Y. (2024). Statistically optimal firstorder algorithms: A proof via orthogonalization. *Inf. Inference* **13** Paper No. iaae027, 24. MR4806510 https://doi.org/10.1093/imaiai/iaae027

[54] OPPER, M., ÇAKMAK, B. and WINTHER, O. (2016). A theory of solving TAP equations for Ising models with general invariant random matrices. *J. Phys. A*: *Math. Theor.* **49** 114002, 24. MR3462332 https://doi.org/10.1088/1751-8113/49/11/114002

[55] OPPER, M. and WINTHER, O. (2001). Adaptive and self-averaging Thouless–Anderson–Palmer mean-field theory for probabilistic modeling. *Phys. Rev. E* **64** 056131.

[56] OYMAK, S. and HASSIBI, B. (2014). A case for orthogonal measurements in linear inverse problems. In 2014 *IEEE International Symposium on Information Theory* 3175–3179. IEEE Press, New York.

[57] PARKER, J. T., SCHNITER, P. and CEVHER, V. (2014). Bilinear generalized approximate message passing—Part I: Derivation. *IEEE Trans. Signal Process.* **62** 5839–5853. MR3281527 https://doi.org/10.1109/TSP.2014.2357776

[58] PASTUR, L. and SHCHERBINA, M. (2011). *Eigenvalue Distribution of Large Random Matrices*. *Mathematical Surveys and Monographs* **171**. Amer. Math. Soc., Providence, RI. MR2808038 https://doi.org/10.1090/surv/171

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

[60] PÉCHÉ, S. (2006). The largest eigenvalue of small rank perturbations of Hermitian random matrices. *Probab. Theory Related Fields* **134** 127–173. MR2221787 https://doi.org/10.1007/s00440-005-0466-z

[61] PERRY, A., WEIN, A. S., BANDEIRA, A. S. and MOITRA, A. (2018). Message-passing algorithms for synchronization problems over compact groups. *Comm. Pure Appl. Math.* **71** 2275–2322. MR3862091 https://doi.org/10.1002/cpa.21750

[62] PERRY, A., WEIN, A. S., BANDEIRA, A. S. and MOITRA, A. (2018). Optimality and sub-optimality of PCA I: Spiked random matrix models. *Ann. Statist.* **46** 2416–2451. MR3845022 https://doi.org/10.1214/17-AOS1625

[63] POURKAMALI, F. and MACRIS, N. (2023). Rectangular rotational invariant estimator for general additive noise matrices. In 2023 *IEEE International Symposium on Information Theory* (*ISIT*) 2081–2086. IEEE Press, New York.

[64] 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

[65] RUSH, C. and VENKATARAMANAN, R. (2018). Finite sample analysis of approximate message passing algorithms. *IEEE Trans. Inf. Theory* **64** 7264–7286. MR3876443 https://doi.org/10.1109/TIT.2018.2816681

[66] SCHRAMM, T. and WEIN, A. S. (2022). Computational barriers to estimation from low-degree polynomials. *Ann. Statist.* **50** 1833–1858. MR4441142 https://doi.org/10.1214/22-aos2179

[67] SEMERJIAN, G. (2024). Matrix denoising: Bayes-optimal estimators via low-degree polynomials. *J. Stat. Phys.* **191** Paper No. 139, 56. MR4813692 https://doi.org/10.1007/s10955-024-03359-9

[68] SINGER, A. (2011). Angular synchronization by eigenvectors and semidefinite programming. *Appl. Comput. Harmon. Anal.* **30** 20–36. MR2737931 https://doi.org/10.1016/j.acha.2010.02.001

[69] 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

[70] 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.

[71] 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

[72] ZHONG, X., SU, C. and FAN, Z. (2022). Empirical Bayes PCA in high dimensions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 853–878. MR4460578

[73] ZHONG, X., WANG, T. and FAN, Z. (2024). Approximate message passing for orthogonally invariant ensembles: Multivariate non-linearities and spectral initialization. *Inf. Inference* **13** Paper No. iaae024, 22. MR4797647 https://doi.org/10.1093/imaiai/iaae024

[74] ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.* **15** 265–286. MR2252527 https://doi.org/10.1198/106186006X113430

*The Annals of Statistics*
2026, Vol. 54, No. 1, 490–515
https://doi.org/10.1214/25-AOS2576

# INFORMATION THEORETIC LIMITS OF ROBUST SUB-GAUSSIAN MEAN ESTIMATION UNDER STAR-SHAPED CONSTRAINTS

BY AKSHAY PRASADAN[1,a] AND MATEY NEYKOV[2,b]

[1] *Department of Statistics and Data Science, Carnegie Mellon University,* [a]*aprasada@andrew.cmu.edu*
[2] *Department of Statistics and Data Science, Northwestern University,* [b]*mneykov@northwestern.edu*

We obtain the minimax rate for a mean location model with a bounded star-shaped set $K \subseteq \mathbb{R}^n$ constraint on the mean, in an adversarially corrupted data setting with Gaussian noise. We assume an unknown fraction $\epsilon \leq 1/2 - \kappa$ for some fixed $\kappa \in (0, 1/2]$ of $N$ observations are arbitrarily corrupted. We obtain a minimax risk up to proportionality constants under the squared $\ell_2$ loss of $\max(\eta^{*2}, \sigma^2\epsilon^2) \wedge d^2$ with

$$\eta^* = \sup\left\{\eta \geq 0 : \frac{N\eta^2}{\sigma^2} \leq \log \mathcal{M}_K^{\text{loc}}(\eta, c)\right\},$$

where $\log \mathcal{M}_K^{\text{loc}}(\eta, c)$ denotes the local entropy of the set $K$, $d$ is the diameter of $K$, $\sigma^2$ is the variance and $c$ is some sufficiently large absolute constant. A variant of our algorithm achieves the same rate for settings with known or symmetric sub-Gaussian noise, with a smaller breakdown point, still of constant order. We further study the case of unknown sub-Gaussian noise and show that the rate is slightly slower: $\max(\eta^{*2}, \sigma^2\epsilon^2 \log(1/\epsilon)) \wedge d^2$. We generalize our results to the case when $K$ is star-shaped but unbounded.

## REFERENCES

[1] AAMARI, E. and LEVRARD, C. (2018). Stability and minimax optimality of tangential Delaunay complexes for manifold reconstruction. *Discrete Comput. Geom.* **59** 923–971. MR3802310 https://doi.org/10.1007/s00454-017-9962-z

[2] ABDALLA, P. and ZHIVOTOVSKIY, N. (2024). Covariance estimation: Optimal dimension-free guarantees for adversarial corruption and heavy tails. *J. Eur. Math. Soc.*

[3] BALAKRISHNAN, S., DU, S. S., LI, J. and SINGH, A. (2017). Computationally efficient robust sparse estimation in high dimensions. In *Conference on Learning Theory* 169–212. PMLR.

[4] BATENI, A.-H. and DALALYAN, A. S. (2020). Confidence regions and minimax rates in outlier-robust estimation on the probability simplex. *Electron. J. Stat.* **14** 2653–2677. MR4124558 https://doi.org/10.1214/20-EJS1731

[5] BATENI, A.-H., MINASYAN, A. and DALALYAN, A. S. (2025). Nearly minimax robust estimator of the mean vector by iterative spectral dimension reduction. *Bernoulli* **31** 2699–2722. MR4931347 https://doi.org/10.3150/24-BEJ1822

[6] BIRGÉ, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete* **65** 181–237. MR0722129 https://doi.org/10.1007/BF00532480

[7] CHEN, M., GAO, C. and REN, Z. (2016). A general decision theory for Huber's $\epsilon$-contamination model. *Electron. J. Stat.* **10** 3752–3774. MR3579675 https://doi.org/10.1214/16-EJS1216

[8] CHEN, M., GAO, C. and REN, Z. (2018). Robust covariance and scatter matrix estimation under Huber's contamination model. *Ann. Statist.* **46** 1932–1960. MR3845006 https://doi.org/10.1214/17-AOS1607

[9] CHENG, Y., DIAKONIKOLAS, I., GE, R., GUPTA, S., KANE, D. and SOLTANOLKOTABI, M. (2022). Outlier-robust sparse estimation via non-convex optimization. *Adv. Neural Inf. Process. Syst.* **35** 7318–7327.

[10] DALALYAN, A. S. and MINASYAN, A. (2022). All-in-one robust estimator of the Gaussian mean. *Ann. Statist.* **50** 1193–1219. MR4404933 https://doi.org/10.1214/21-aos2145

[11] DEPERSIN, J. and LECUÉ, G. (2022). Robust sub-Gaussian estimation of a mean vector in nearly linear time. *Ann. Statist.* **50** 511–536. MR4382026 https://doi.org/10.1214/21-aos2118

---

*MSC2020 subject classifications.* Primary 62F35, 62F30; secondary 62B10, 62C20.

*Key words and phrases.* Robustness, adversarial corruption, star-shaped, convex, shape constraints, sub-Gaussian noise.

[12] DIAKONIKOLAS, I., HOPKINS, S. B., PENSIA, A. and TIEGEL, S. (2025). SoS certifiability of subgaussian distributions and its algorithmic applications. In *STOC'25—Proceedings of the 57th Annual ACM Symposium on Theory of Computing* 1689–1700. ACM, New York. MR4928554 https://doi.org/10.1145/3717823.3718183

[13] DIAKONIKOLAS, I., KAMATH, G., KANE, D., LI, J., MOITRA, A. and STEWART, A. (2019). Robust estimators in high-dimensions without the computational intractability. *SIAM J. Comput.* **48** 742–864. MR3945261 https://doi.org/10.1137/17M1126680

[14] DIAKONIKOLAS, I., KAMATH, G., KANE, D. M., LI, J., MOITRA, A. and STEWART, A. (2017). Being robust (in high dimensions) can be practical. In *International Conference on Machine Learning* 999–1008. PMLR.

[15] DIAKONIKOLAS, I., KANE, D., KARMALKAR, S., PRICE, E. and STEWART, A. (2019). Outlier-robust high-dimensional sparse estimation via iterative filtering. *Adv. Neural Inf. Process. Syst.* **32**.

[16] DIAKONIKOLAS, I. and KANE, D. M. (2023). *Algorithmic High-Dimensional Robust Statistics*. Cambridge Univ. Press, Cambridge.

[17] DIAKONIKOLAS, I., KANE, D. M., KARMALKAR, S., PENSIA, A. and PITTAS, T. (2022). Robust sparse mean estimation via sum of squares. In *Conference on Learning Theory* 4703–4763. PMLR.

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

[19] HUBER, P. J. and RONCHETTI, E. M. (2011). *Robust Statistics*. Wiley, New York.

[20] KOTEKAL, S. and GAO, C. (2025). Optimal estimation of the null distribution in large-scale inference. *IEEE Trans. Inf. Theory* **71** 2075–2103. MR4867929 https://doi.org/10.1109/tit.2025.3529457

[21] LAI, K. A., RAO, A. B. and VEMPALA, S. (2016). Agnostic estimation of mean and covariance. In *57th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2016* 665–674. IEEE Computer Soc., Los Alamitos, CA. MR3631029

[22] LECAM, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** 38–53. MR0334381

[23] LUGOSI, G. and MENDELSON, S. (2021). Robust multivariate mean estimation: The optimality of trimmed mean. *Ann. Statist.* **49** 393–410. MR4206683 https://doi.org/10.1214/20-AOS1961

[24] MINASYAN, A. and ZHIVOTOVSKIY, N. (2025). Statistically optimal robust mean and covariance estimation for anisotropic Gaussians. *Math. Stat. Learn.* **8** 33–69. MR4948548 https://doi.org/10.4171/msl/48

[25] MINSKER, S. (2025). Uniform bounds for robust mean estimators. *Stoch. Process. Appl.* **190** 104724. MR4933834 https://doi.org/10.1016/j.spa.2025.104724

[26] NEYKOV, M. (2023). On the minimax rate of the Gaussian sequence model under bounded convex constraints. *IEEE Trans. Inf. Theory* **69** 1244–1260. MR4564653 https://doi.org/10.1109/tit.2022.3213141

[27] NOVIKOV, G., STEURER, D. and TIEGEL, S. (2023). Robust mean estimation without moments for symmetric distributions. In *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt and S. Levine, eds.) **36** 34371–34409. Curran Associates, Red Hook.

[28] PETROV, V. V. (1976). *Sums of Independent Random Variables*. Springer, Berlin.

[29] POLYANSKIY, Y. and WU, Y. (2023). Information Theory: from Coding to Learning.

[30] PRASAD, A., BALAKRISHNAN, S. and RAVIKUMAR, P. (2020). A robust univariate mean estimator is all you need. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics* (S. Chiappa and R. Calandra, eds.). *Proceedings of Machine Learning Research* **108** 4034–4044. PMLR.

[31] PRASADAN, A. and NEYKOV, M. (2025). Some facts about the optimality of the LSE in the Gaussian sequence model with convex constraint. *IEEE Trans. Inf. Theory* **71** 8928–8958. MR4978539 https://doi.org/10.1109/tit.2025.3605342

[32] PRASADAN, A. and NEYKOV, M. (2026). Supplement to "Information theoretic limits of robust sub-Gaussian mean estimation under star-shaped constraints." https://doi.org/10.1214/25-AOS2576SUPP

[33] STAHEL, W. A., HAMPEL, F. R., ROUSSEEUW, P. J. and RONCHETTI, E. M. (2011). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.

*The Annals of Statistics*
2026, Vol. 54, No. 1, 516–541
https://doi.org/10.1214/25-AOS2577
This research was funded, in whole or in part, by ESPRC, EP/V049968/1. A CC BY 4.0 license is applied to this article arising from this submission, in accordance with the grant's open access conditions.

# CONFOUNDER SELECTION VIA ITERATIVE GRAPH EXPANSION

BY F. RICHARD GUO[1,a] AND QINGYUAN ZHAO[2,b]

[1]*Department of Statistics, University of Michigan,* [a]*ricguo@umich.edu*
[2]*Statistical Laboratory, University of Cambridge,* [b]*qyzhao@statslab.cam.ac.uk*

Confounder selection, namely choosing a set of covariates to control for confounding between a treatment and an outcome, is arguably the most important step in the design of an observational study. Previous methods, such as Pearl's back-door criterion, typically require prespecifying a causal graph, which can often be difficult in practice. We propose an interactive procedure for confounder selection that does not require prespecifying the graph or the set of observed variables. This procedure iteratively expands the causal graph by finding what we call "primary adjustment sets" for a pair of possibly confounded variables. This can be viewed as inverting a sequence of marginalizations of the underlying causal graph. Structural information in the form of primary adjustment sets is elicited from the user, bit by bit, until either a set of covariates is found to control for confounding or it can be determined that no such set exists. Other information, such as the causal relations between confounders, is not required by the procedure. We show that if the user correctly specifies the primary adjustment sets in every step, our procedure is both sound and complete.

## REFERENCES

DING, P. and MIRATRIX, L. W. (2015). To adjust or not to adjust? Sensitivity analysis of $M$-bias and butterfly-bias. *J. Causal Inference* **3** 41–57. MR4289426 https://doi.org/10.1515/jci-2013-0021

ENGLE, R. F., HENDRY, D. F. and RICHARD, J.-F. (1983). Exogeneity. *Econometrica* **51** 277–304. MR0688727 https://doi.org/10.2307/1911990

EVANS, R. (2019). Markov properties for mixed graphical models. In *Handbook of Graphical Models*. *Chapman & Hall/CRC Handb. Mod. Stat. Methods* 39–60. CRC Press, Boca Raton, FL. MR3887991

EVANS, R. J. and RICHARDSON, T. S. (2014). Markovian acyclic directed mixed graphs for discrete data. *Ann. Statist.* **42** 1452–1482. MR3262457 https://doi.org/10.1214/14-AOS1206

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

GREENLAND, S. and ROBINS, J. M. (2009). Identifiability, exchangeability and confounding revisited. *Epidemiol. Perspect. Innov.* **6** 1–9.

GUO, F. R., LUNDBORG, A. R. and ZHAO, Q. (2022). Confounder selection: Objectives and approaches. arXiv Preprint. Available at arXiv:2208.13871.

GUO, F. R. and ZHAO, Q. (2026). Supplement to "Confounder selection via iterative graph expansion." https://doi.org/10.1214/25-AOS2577SUPP

HABER, N. A., WOOD, M. E., WIETEN, S. and BRESKIN, A. (2022). DAG With Omitted Objects Displayed (DAGWOOD): A framework for revealing causal assumptions in DAGs. *Ann. Epidemiol.* **68** 64–71.

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

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

KOSTER, J. T. A. (1999). On the validity of the Markov interpretation of path diagrams of Gaussian structural equations systems with correlated errors. *Scand. J. Stat.* **26** 413–431. MR1712043 https://doi.org/10.1111/1467-9469.00157

PEARL, J. (1988). *Probabilistic Reasoning in Intelligent Systems*: *Networks of Plausible Inference*. *The Morgan Kaufmann Series in Representation and Reasoning*. Morgan Kaufmann, San Mateo, CA. MR0965765

---

*MSC2020 subject classifications.* Primary 62A09; secondary 62D20.
*Key words and phrases.* Causal inference, observational study, back-door criterion, background knowledge, adjustment, marginalization.

PEARL, J. (1993). Comment: Graphical models, causality and intervention. *Statist. Sci.* **8** 266–269.

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

PEARL, J. and VERMA, T. S. (1991). A theory of inferred causation. In *Principles of Knowledge Representation and Reasoning* (*Cambridge, MA*, 1991). *Morgan Kaufmann Ser. Represent. Reason.* 441–452. Morgan Kaufmann, San Mateo, CA. MR1142173

PERKOVIĆ, E., TEXTOR, J., KALISCH, M. and MAATHUIS, M. H. (2015). A complete generalized adjustment criterion. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence. UAI*'15 682–691. AUAI Press, Arlington, VA, USA.

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. S. and ROBINS, J. M. (2013). Single World Intervention Graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. Technical Report No. 128, Center for the Statistics and the Social Sciences, Univ. Washington Series.

RICHARDSON, T. S., ROBINS, J. M. and WANG, L. (2018). Discussion of "Data-driven confounder selection via Markov and Bayesian networks" by Häggström. *Biometrics* **74** 403–406.

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

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

RUBIN, D. B. (2008). For objective causal inference, design trumps analysis. *Ann. Appl. Stat.* **2** 808–804. MR2516795 https://doi.org/10.1214/08-AOAS187

SHACHTER, R. D. (1998). Bayes-Ball: Rational pastime (for determining irrelevance and requisite information in belief networks and influence diagrams). In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence. UAI*'98 480–487. Morgan Kaufmann Inc., San Francisco, CA, USA.

SHPITSER, I., RICHARDSON, T. S. and ROBINS, J. M. (2022). Multivariate counterfactual systems and causal graphical models. In *Probabilistic and Causal Inference: The Works of Judea Pearl* 813–852. Association for Computing Machinery, New York, NY, USA.

SHPITSER, I., VANDERWEELE, T. J. and ROBINS, J. M. (2010). On the validity of covariate adjustment for estimating causal effects. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence. UAI*'10 527–536. AUAI Press, Arlington, VA, USA.

SHRIER, I. and PLATT, R. W. (2008). Reducing bias through directed acyclic graphs. *BMC Med. Res. Methodol.* **8** 1–15.

SMUCLER, E. and ROTNITZKY, A. (2022). A note on efficient minimum cost adjustment sets in causal graphical models. *J. Causal Inference* **10** 174–189. MR4452948 https://doi.org/10.1515/jci-2022-0015

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (1993). *Causation, Prediction, and Search. Lecture Notes in Statistics* **81**. Springer, New York. MR1227558 https://doi.org/10.1007/978-1-4612-2748-9

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

VANDERWEELE, T. J. (2019). Principles of confounder selection. *Eur. J. Epidemiol.* **34** 211–219.

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

VANDERWEELE, T. J. and SHPITSER, I. (2013). On the definition of a confounder. *Ann. Statist.* **41** 196–220. MR3059415 https://doi.org/10.1214/12-AOS1058

VAN DER ZANDER, B., LIŚKIEWICZ, M. and TEXTOR, J. (2019). Separators and adjustment sets in causal graphs: Complete criteria and an algorithmic framework. *Artificial Intelligence* **270** 1–40. MR3901676 https://doi.org/10.1016/j.artint.2018.12.006

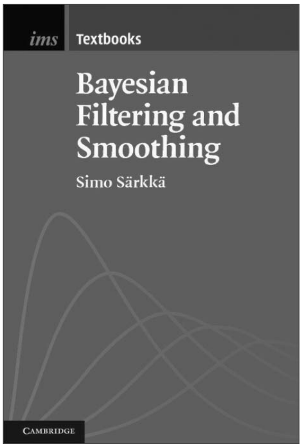
VERMA, T. S. and PEARL, J. (1990). Equivalence and synthesis of causal models. In *Proceedings of the* 6*th Conference on Uncertainty in Artificial Intelligence* (*UAI*-1990) 220–227.

ZHAO, Q. (2024). A matrix algebra for graphical statistical models. arXiv Preprint. Available at arXiv:2407.15744.

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

*The Institute of Mathematical Statistics presents*

# IMS TEXTBOOKS



ims Textbooks

Bayesian Filtering and Smoothing

Simo Särkkä

CAMBRIDGE

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

**Hardback US$59.40**
**Paperback US$22.19**

## *Bayesian Filtering and Smoothing*

Simo Särkkä

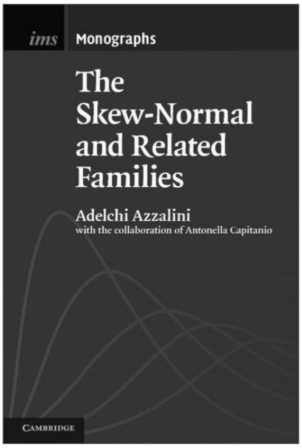
Filtering and smoothing methods are used to produce an accurate estimate of the state of a time-varying system based on multiple observational inputs (data). Interest in these methods has exploded in recent years, with numerous applications emerging in fields such as navigation, aerospace engineering, telecommunications and medicine. This compact, informal introduction for graduate students and advanced undergraduates presents the current state-of-the-art filtering and smoothing methods in a unified Bayesian framework. Readers learn what non-linear Kalman filters and particle filters are, how they are related, and their relative advantages and disadvantages. They also discover how state-of-the-art Bayesian parameter estimation methods can be combined with state-of-the-art filtering and smoothing algorithms. The book's practical and algorithmic approach assumes only modest mathematical prerequisites. Examples include MATLAB computations, and the numerous end-of-chapter exercises include computational assignments. MATLAB/GNU Octave source code is available for download at www.cambridge.org/sarkka, promoting hands-on work with the methods.

---

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.

*The Institute of Mathematical Statistics presents*

# *IMS MONOGRAPHS*



ims Monographs

The Skew-Normal and Related Families

Adelchi Azzalini

with the collaboration of Antonella Capitanio

CAMBRIDGE

## ***The Skew-Normal and Related Families***

Adelchi Azzalini
in collaboration with Antonella Capitanio

Interest in the skew-normal and related families of distributions has grown enormously over recent years, as theory has advanced, challenges of data have grown, and computational tools have made substantial progress. This comprehensive treatment, blending theory and practice, will be the standard resource for statisticians and applied researchers. Assuming only basic knowledge of (non-measure-theoretic) probability and statistical inference, the book is accessible to the wide range of researchers who use statistical modelling techniques. Guiding readers through the main concepts and results, it covers both the probability and the statistics sides of the subject, in the univariate and multivariate settings. The theoretical development is complemented by numerous illustrations and applications to a range of fields including quantitative finance, medical statistics, environmental risk studies, and industrial and business efficiency.

The author's freely available R package sn, available from CRAN, equips readers to put the methods into action with their own data.

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

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

---

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