

THE ANNALS *of* STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles

- Nonparametric regression using deep neural networks with ReLU activation function JOHANNES SCHMIDT-HIEBER 1875
- Discussion of: "Nonparametric regression using deep neural networks with ReLU activation function" BEHROOZ GHORBANI, SONG MEI, THEODOR MISIAKIEWICZ AND ANDREA MONTANARI 1898
- Discussion of: "Nonparametric regression using deep neural networks with ReLU activation function" GITTA KUTYNIOK 1902
- Discussion of: "Nonparametric regression using deep neural networks with ReLU activation function" MICHAEL KOHLER AND SOPHIE LANGER 1906
- Discussion of: "Nonparametric regression using deep neural networks with ReLU activation function" OHAD SHAMIR 1911
- Rejoinder: "Nonparametric regression using deep neural networks with ReLU activation function" JOHANNES SCHMIDT-HIEBER 1916
- Nonclassical Berry–Esseen inequalities and accuracy of the bootstrap MAYYA ZHILOVA 1922
- On the validity of the formal Edgeworth expansion for posterior densities JOHN E. KOLASSA AND TODD A. KUFFNER 1940
- Model selection for high-dimensional linear regression with dependent observations CHING-KANG ING 1959
- Optimal estimation of Gaussian mixtures via denoised method of moments YIHONG WU AND PENGKUN YANG 1981
- Sharp instruments for classifying compliers and generalizing causal effects EDWARD H. KENNEDY, SIVARAMAN BALAKRISHNAN AND MAX G'SELL 2008
- Empirical risk minimization and complexity of dynamical models KEVIN MCGOFF AND ANDREW B. NOBEL 2031
- Adaptive estimation in structured factor models with applications to overlapping clustering XIN BING, FLORENTINA BUNEA, YANG NING AND MARTEN WEGKAMP 2055
- Partial identifiability of restricted latent class models. YUQI GU AND GONGJUN XU 2082
- Posterior concentration for Bayesian regression trees and forests VERONIKA ROČKOVÁ AND STÉPHANIE VAN DER PAS 2108
- Double-slicing assisted sufficient dimension reduction for high-dimensional censored data SHANSHAN DING, WEI QIAN AND LAN WANG 2132
- Asymptotic frequentist coverage properties of Bayesian credible sets for sieve priors JUDITH ROUSSEAU AND BOTOND SZABO 2155
- Convergence rates of variational posterior distributions FENGSHUO ZHANG AND CHAO GAO 2180
- Two-sample hypothesis testing for inhomogeneous random graphs DEBARGHYA GHOSHDASTIDAR, MAURILIO GUTZEIT, ALEXANDRA CARPENTIER AND ULRIKE VON LUXBURG 2208
- Beyond HC: More sensitive tests for rare/weak alternatives THOMAS PORTER AND MICHAEL STEWART 2230

continued

THE ANNALS *of* STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles—*Continued from front cover*

- Minimax optimal rates for Mondrian trees and forests
JAOUAD MOURTADA, STÉPHANE GAÏFFAS AND ERWAN SCORNET 2253
- Identifiability of nonparametric mixture models and Bayes optimal clustering
BRYON ARAGAM, CHEN DAN, ERIC P. XING AND PRADEEP RAVIKUMAR 2277
- A test for separability in covariance operators of random surfaces
PRAMITA BAGCHI AND HOLGER DETTE 2303
- A general approach for cure models in survival analysis
VALENTIN PATILEA AND INGRID VAN KEILEGOM 2323
- Adaptive distributed methods under communication constraints
BOTOND SZABÓ AND HARRY VAN ZANTEN 2347
- Bayesian analysis of the covariance matrix of a multivariate normal distribution with a
new class of priors JAMES O. BERGER, DONGCHU SUN
AND CHENGYUAN SONG 2381
- Extending the validity of frequency domain bootstrap methods to general stationary
processes MARCO MEYER, EFSTATHIOS PAPARODITIS
AND JENS-PETER KREISS 2404
- Minimax estimation of large precision matrices with bandable Cholesky factor
YU LIU AND ZHAO REN 2428
- Estimation and inference for precision matrices of nonstationary time series
XIUCAI DING AND ZHOU ZHOU 2455
- Isotropic covariance functions on graphs and their edges
ETHAN ANDERES, JESPER MØLLER AND JAKOB G. RASMUSSEN 2478

THE ANNALS OF STATISTICS

Vol. 48, No. 4, pp. 1875–2503 August 2020

INSTITUTE OF MATHEMATICAL STATISTICS

(Organized September 12, 1935)

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

IMS OFFICERS

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

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

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

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

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

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

IMS EDITORS

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

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

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

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

Statistical Science. *Editor:* Sonia Petrone, Department of Decision Sciences, Università Bocconi, 20100 Milano MI, Italy

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

The Annals of Statistics [ISSN 0090-5364 (print); ISSN 2168-8966 (online)], Volume 48, Number 4, August 2020. Published bimonthly by the Institute of Mathematical Statistics, 3163 Somerset Drive, Cleveland, Ohio 44122, USA. Periodicals postage paid at Cleveland, Ohio, and at additional mailing offices.

POSTMASTER: Send address changes to *The Annals of Statistics*, Institute of Mathematical Statistics, Dues and Subscriptions Office, 9650 Rockville Pike, Suite L 2310, Bethesda, Maryland 20814-3998, USA.

NONPARAMETRIC REGRESSION USING DEEP NEURAL NETWORKS WITH RELU ACTIVATION FUNCTION¹

BY JOHANNES SCHMIDT-HIEBER

Department of Applied Mathematics, University of Twente, a.j.schmidt-hieber@utwente.nl

Consider the multivariate nonparametric regression model. It is shown that estimators based on sparsely connected deep neural networks with ReLU activation function and properly chosen network architecture achieve the minimax rates of convergence (up to $\log n$ -factors) under a general composition assumption on the regression function. The framework includes many well-studied structural constraints such as (generalized) additive models. While there is a lot of flexibility in the network architecture, the tuning parameter is the sparsity of the network. Specifically, we consider large networks with number of potential network parameters exceeding the sample size. The analysis gives some insights into why multilayer feedforward neural networks perform well in practice. Interestingly, for ReLU activation function the depth (number of layers) of the neural network architectures plays an important role, and our theory suggests that for nonparametric regression, scaling the network depth with the sample size is natural. It is also shown that under the composition assumption wavelet estimators can only achieve suboptimal rates.

REFERENCES

- [1] ANTHONY, M. and BARTLETT, P. L. (1999). *Neural Network Learning: Theoretical Foundations*. Cambridge Univ. Press, Cambridge. [MR1741038](#) <https://doi.org/10.1017/CBO9780511624216>
- [2] BACH, F. (2017). Breaking the curse of dimensionality with convex neural networks. *J. Mach. Learn. Res.* **18** Art. ID 19. [MR3634886](#)
- [3] BARAUD, Y. and BIRGÉ, L. (2014). Estimating composite functions by model selection. *Ann. Inst. Henri Poincaré Probab. Stat.* **50** 285–314. [MR3161532](#) <https://doi.org/10.1214/12-AIHP516>
- [4] 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>
- [5] BARRON, A. R. (1994). Approximation and estimation bounds for artificial neural networks. *Mach. Learn.* **14** 115–133. <https://doi.org/10.1007/BF00993164>
- [6] BAUER, B. and KOHLER, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *Ann. Statist.* **47** 2261–2285. [MR3953451](#) <https://doi.org/10.1214/18-AOS1747>
- [7] BÖLCSKEI, H., GROHS, P., KUTYNIOK, G. and PETERSEN, P. (2019). Optimal approximation with sparsely connected deep neural networks. *SIAM J. Math. Data Sci.* **1** 8–45. [MR3949699](#) <https://doi.org/10.1137/18M118709X>
- [8] CANDÈS, E. J. (2002). New ties between computational harmonic analysis and approximation theory. In *Approximation Theory, X (St. Louis, MO, 2001)*. *Innov. Appl. Math.* 87–153. Vanderbilt Univ. Press, Nashville, TN. [MR1924879](#)
- [9] CHOROMANSKA, A., HENAFF, M., MATHIEU, M., AROUS, G. B. and LECUN, Y. (2015). The loss surface of multilayer networks. In *Aistats. Proceedings of Machine Learning Research* **38** 192–204.
- [10] COHEN, A., DAUBECHIES, I. and VIAL, P. (1993). Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.* **1** 54–81. [MR1256527](#) <https://doi.org/10.1006/acha.1993.1005>
- [11] CYBENKO, G. (1989). Approximation by superpositions of a sigmoidal function. *Math. Control Signals Systems* **2** 303–314. [MR1015670](#) <https://doi.org/10.1007/BF02551274>
- [12] GINÉ, E. and KOLTCHINSKII, V. (2006). Concentration inequalities and asymptotic results for ratio type empirical processes. *Ann. Probab.* **34** 1143–1216. [MR2243881](#) <https://doi.org/10.1214/009117906000000070>

MSC2020 subject classifications. 62G08.

Key words and phrases. Nonparametric regression, multilayer neural networks, ReLU activation function, minimax estimation risk, additive models, wavelets.

- [13] GOROT, X., BORDES, A. and BENGIO, Y. (2011). Deep sparse rectifier neural networks. In *Aistats. Proceedings of Machine Learning Research* **15** 315–323.
- [14] GOODFELLOW, I., BENGIO, Y. and COURVILLE, A. (2016). *Deep Learning. Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. [MR3617773](#)
- [15] 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>
- [16] HAMERS, M. and KOHLER, M. (2006). Nonasymptotic bounds on the L_2 error of neural network regression estimates. *Ann. Inst. Statist. Math.* **58** 131–151. [MR2281209](#) <https://doi.org/10.1007/s10463-005-0005-9>
- [17] HE, K. and SUN, J. (2015). Convolutional neural networks at constrained time cost. In *CVPR* 5353–5360.
- [18] HE, K., ZHANG, X., REN, S. and SUN, J. (2016). Deep residual learning for image recognition. In *CVPR* 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- [19] HORNIK, K., STINCHCOMBE, M. and WHITE, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Netw.* **2** 359–366.
- [20] HORNIK, K., STINCHCOMBE, M. and WHITE, H. (1990). Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Netw.* **3** 551–560.
- [21] HOROWITZ, J. L. and MAMMEN, E. (2007). Rate-optimal estimation for a general class of nonparametric regression models with unknown link functions. *Ann. Statist.* **35** 2589–2619. [MR2382659](#) <https://doi.org/10.1214/009053607000000415>
- [22] JUDITSKY, A. B., LEPSKI, O. V. and TSYBAKOV, A. B. (2009). Nonparametric estimation of composite functions. *Ann. Statist.* **37** 1360–1404. [MR2509077](#) <https://doi.org/10.1214/08-AOS611>
- [23] KLUSOWSKI, J. M. and BARRON, A. R. (2016). Risk bounds for high-dimensional ridge function combinations including neural networks. Preprint. Available at [arXiv:1607.01434](#).
- [24] KLUSOWSKI, J. M. and BARRON, A. R. (2016). Uniform approximation by neural networks activated by first and second order ridge splines. Preprint. Available at [arXiv:1607.07819](#).
- [25] KOHLER, M. and KRZYŻAK, A. (2005). Adaptive regression estimation with multilayer feedforward neural networks. *J. Nonparametr. Stat.* **17** 891–913. [MR2192165](#) <https://doi.org/10.1080/10485250500309608>
- [26] KOHLER, M. and KRZYŻAK, A. (2017). Nonparametric regression based on hierarchical interaction models. *IEEE Trans. Inf. Theory* **63** 1620–1630. [MR3625984](#) <https://doi.org/10.1109/TIT.2016.2634401>
- [27] KOLTCHINSKII, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.* **34** 2593–2656. [MR2329442](#) <https://doi.org/10.1214/0090536060000001019>
- [28] KRIZHEVSKY, A., SUTSKEVER, I. and HINTON, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* 25 (F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger, eds.) 1097–1105. Curran Associates, Red Hook, NY.
- [29] LESHNO, M., LIN, V. YA., PINKUS, A. and SCHOCKEN, S. (1993). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Netw.* **6** 861–867.
- [30] LIANG, S. and SRIKANT, R. (2017). Why deep neural networks for function approximation? In *ICLR* 2017.
- [31] LIAO, Q. and POGGIO, T. (2017). Theory II: Landscape of the empirical risk in deep learning. Preprint. Available at [arXiv:1703.09833](#).
- [32] MASSART, P. (2007). *Concentration Inequalities and Model Selection. Lecture Notes in Math.* **1896**. Springer, Berlin. [MR2319879](#)
- [33] MCCAFFREY, D. F. and GALLANT, A. R. (1994). Convergence rates for single hidden layer feedforward networks. *Neural Netw.* **7** 147–158. [https://doi.org/10.1016/0893-6080\(94\)90063-9](https://doi.org/10.1016/0893-6080(94)90063-9)
- [34] MHASKAR, H. N. (1993). Approximation properties of a multilayered feedforward artificial neural network. *Adv. Comput. Math.* **1** 61–80. [MR1230251](#) <https://doi.org/10.1007/BF02070821>
- [35] MHASKAR, H. N. and POGGIO, T. (2016). Deep vs. shallow networks: An approximation theory perspective. *Anal. Appl. (Singap.)* **14** 829–848. [MR3564936](#) <https://doi.org/10.1142/S0219530516400042>
- [36] MONTÚFAR, G. F. (2014). Universal approximation depth and errors of narrow belief networks with discrete units. *Neural Comput.* **26** 1386–1407. [MR3222078](#) https://doi.org/10.1162/NECO_a_00601
- [37] MONTÚFAR, G. F., PASCANU, R., CHO, K. and BENGIO, Y. (2014). On the number of linear regions of deep neural networks. In *Advances in Neural Information Processing Systems* 27 2924–2932. Curran Associates, Red Hook, NY.
- [38] PEDAMONTI, D. (2018). Comparison of non-linear activation functions for deep neural networks on MNIST classification task. Preprint. Available at [arXiv:1804.02763](#).
- [39] PINKUS, A. (1999). Approximation theory of the MLP model in neural networks. *Acta Numer.* **8** 143–195. [MR1819645](#) <https://doi.org/10.1017/S0962492900002919>

- [40] POGGIO, T., MHASKAR, H., ROSASCO, L., MIRANDA, B. and LIAO, Q. (2017). Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review. *Int. J. Autom. Comput.* **14** 503–519. <https://doi.org/10.1007/s11633-017-1054-2>
- [41] RAY, K. and SCHMIDT-HIEBER, J. (2017). A regularity class for the roots of nonnegative functions. *Ann. Mat. Pura Appl.* (4) **196** 2091–2103. [MR3714756 https://doi.org/10.1007/s10231-017-0655-2](https://doi.org/10.1007/s10231-017-0655-2)
- [42] SCARSELLI, F. and TSOI, A. C. (1998). Universal approximation using feedforward neural networks: A survey of some existing methods, and some new results. *Neural Netw.* **11** 15–37. [https://doi.org/10.1016/S0893-6080\(97\)00097-X](https://doi.org/10.1016/S0893-6080(97)00097-X)
- [43] SCHMIDT-HIEBER, J. (2020). Supplement to “Nonparametric regression using deep neural networks with ReLU activation function.” <https://doi.org/10.1214/19-AOS1875SUPP>
- [44] SRIVASTAVA, N., HINTON, G., KRIZHEVSKY, A., SUTSKEVER, I. and SALAKHUTDINOV, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15** 1929–1958. [MR3231592](https://doi.org/10.1016/j.neunet.2017.07.002)
- [45] SRIVASTAVA, R. K., GREFF, K. and SCHMIDHUBER, J. (2015). Highway networks. Preprint. Available at [arXiv:1505.00387](https://arxiv.org/abs/1505.00387).
- [46] STINCHCOMBE, M. B. (1999). Neural network approximation of continuous functionals and continuous functions on compactifications. *Neural Netw.* **12** 467–477. [https://doi.org/10.1016/s0893-6080\(98\)00108-7](https://doi.org/10.1016/s0893-6080(98)00108-7)
- [47] SUZUKI, T. (2018). Fast generalization error bound of deep learning from a kernel perspective. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research* **84** 1397–1406.
- [48] SZEGEDY, C., IOFFE, S., VANHOUCKE, V. and ALEMI, A. A. (2016). Inception-v4, Inception-ResNet and the impact of residual connections on learning. In *ICLR 2016 Workshop*.
- [49] TELGARSKY, M. (2016). Benefits of depth in neural networks. In *29th Annual Conference on Learning Theory. Proceedings of Machine Learning Research* **49** 1517–1539.
- [50] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation. Springer Series in Statistics*. Springer, New York. [MR2724359 https://doi.org/10.1007/b13794](https://doi.org/10.1007/b13794)
- [51] WASSERMAN, L. (2006). *All of Nonparametric Statistics. Springer Texts in Statistics*. Springer, New York. [MR2172729](https://doi.org/10.1016/j.neunet.2017.07.002)
- [52] YAROTSKY, D. (2017). Error bounds for approximations with deep ReLU networks. *Neural Netw.* **94** 103–114. <https://doi.org/10.1016/j.neunet.2017.07.002>

DISCUSSION OF: “NONPARAMETRIC REGRESSION USING DEEP NEURAL NETWORKS WITH RELU ACTIVATION FUNCTION”

BY BEHROOZ GHORBANI¹, SONG MEI², THEODOR MISIAKIEWICZ³ AND ANDREA MONTANARI⁴

¹*Department of Electrical Engineering, Stanford University, ghorbani@stanford.edu*

²*Institute for Computational and Mathematical Engineering, Stanford University, songmei@stanford.edu*

³*Department of Statistics, Stanford University, misiakie@stanford.edu*

⁴*Department of Electrical Engineering and Department of Statistics, Stanford University, montanari@stanford.edu*

REFERENCES

- [1] BACH, F. (2017). Breaking the curse of dimensionality with convex neural networks. *J. Mach. Learn. Res.* **18** Art. ID 19. [MR3634886](#)
- [2] BERTHET, Q. and RIGOLLET, P. (2013). Complexity theoretic lower bounds for sparse principal component detection. In *Conference on Learning Theory* 1046–1066.
- [3] CHIZAT, L. and BACH, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems* 3036–3046.
- [4] MONTANARI, A. and VENKATARAMANAN, R. (2017). Estimation of low-rank matrices via approximate message passing. *Ann. Statist.* To appear. <https://doi.org/10.1214/20-AOS1958>
- [5] DONOHO, D. L. and JOHNSTONE, I. M. (1989). Projection-based approximation and a duality with kernel methods. *Ann. Statist.* **17** 58–106. [MR0981438](#) <https://doi.org/10.1214/aos/1176347004>
- [6] GAMARNIK, D. and ILIAS, Z. (2017). High dimensional regression with binary coefficients. Estimating squared error and a phase transition. In *Conference on Learning Theory* 948–953.
- [7] GUNASEKAR, S., LEE, J., SOUDRY, D. and SREBRO, N. (2018). Characterizing implicit bias in terms of optimization geometry. Preprint. Available at [arXiv:1802.08246](#).
- [8] HAJEK, B., WU, Y. and XU, J. (2015). Computational lower bounds for community detection on random graphs. In *Conference on Learning Theory* 899–928.
- [9] JI, Z. and TELGARSKY, M. (2018). Risk and parameter convergence of logistic regression. Preprint. Available at [arXiv:1803.07300](#).
- [10] MEI, S., MONTANARI, A. and NGUYEN, P.-M. (2018). A mean field view of the landscape of two-layer neural networks. *Proc. Natl. Acad. Sci. USA* **115** E7665–E7671. [MR3845070](#) <https://doi.org/10.1073/pnas.1806579115>
- [11] MONTANARI, A. and RICHARD, E. (2014). A statistical model for tensor PCA. In *Advances in Neural Information Processing Systems* 2897–2905.
- [12] NEYSHABUR, B., TOMIOKA, R. and SREBRO, N. (2014). In search of the real inductive bias: On the role of implicit regularization in deep learning. Preprint. Available at [arXiv:1412.6614](#).
- [13] SIMONYAN, K. and ZISSERMAN, A. (2014). Very deep convolutional networks for large-scale image recognition. Preprint. Available at [arXiv:1409.1556](#).
- [14] SOUDRY, D., HOFFER, E., NACSON, M. S., GUNASEKAR, S. and SREBRO, N. (2018). The implicit bias of gradient descent on separable data. *J. Mach. Learn. Res.* **19** Art. ID 70. [MR3899772](#)

DISCUSSION OF: “NONPARAMETRIC REGRESSION USING DEEP NEURAL NETWORKS WITH RELU ACTIVATION FUNCTION”

BY GITTA KUTYNIOK

Institute of Mathematics, Technische Universität Berlin, kutyniok@math.tu-berlin.de

I would like to congratulate Johannes Schmidt–Hieber on a very interesting paper in which he considers regression functions belonging to the class of so-called compositional functions and analyzes the ability of estimators based on the multivariate nonparametric regression model of deep neural networks to achieve minimax rates of convergence.

In my discussion, I will first regard such a type of result from the general viewpoint of the theoretical foundations of deep neural networks. This will be followed by a discussion from the viewpoint of expressivity, optimization and generalization. Finally, I will consider some specific aspects of the main result.

REFERENCES

- [1] BECK, C., HORNING, F., HUTZENTHALER, M., JENTZEN, A. and KRUSE, T. (2019). Overcoming the curse of dimensionality in the numerical approximation of Allen–Cahn partial differential equations via truncated full-history recursive multilevel Picard approximations. [arXiv:1907.06729](https://arxiv.org/abs/1907.06729).
- [2] GROHS, P., HORNING, F., JENTZEN, A. and VON WURSTEMBERGER, P. (2018). A proof that artificial neural networks overcome the curse of dimensionality in the numerical approximation of Black–Scholes partial differential equations. [arXiv:1809.02362](https://arxiv.org/abs/1809.02362).
- [3] KUTYNIOK, G., PETERSEN, P., RASLAN, M. and SCHNEIDER, R. (2019). A theoretical analysis of deep neural networks and parametric PDEs. [arXiv:1904.00377](https://arxiv.org/abs/1904.00377).
- [4] MHASKAR, H. N. and POGGIO, T. (2016). Deep vs. shallow networks: An approximation theory perspective. *Anal. Appl. (Singap.)* **14** 829–848. [MR3564936 https://doi.org/10.1142/S0219530516400042](https://doi.org/10.1142/S0219530516400042)
- [5] PETERSEN, P. and VOIGTLAENDER, F. (2018). Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Netw.* **108** 296–330. <https://doi.org/10.1016/j.neunet.2018.08.019>
- [6] ROTSKOFF, G. M. and VANDEN-EIJNDEN, E. (2018). Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. [arXiv:1805.00915](https://arxiv.org/abs/1805.00915).
- [7] ZHANG, C., BENGIO, S., HARDT, M., RECHT, B. and VINYALS, O. (2017). Understanding deep learning requires rethinking generalization. In *ICLR 2017*.

DISCUSSION OF: “NONPARAMETRIC REGRESSION USING DEEP NEURAL NETWORKS WITH RELU ACTIVATION FUNCTION”

BY MICHAEL KOHLER^{*} AND SOPHIE LANGER[†]

Fachbereich Mathematik, TU Darmstadt, ^{}kohler@mathematik.tu-darmstadt.de; [†]langier@mathematik.tu-darmstadt.de*

REFERENCES

- BAUER, B. and KOHLER, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *Ann. Statist.* **47** 2261–2285. [MR3953451](#) <https://doi.org/10.1214/18-AOS1747>
- FAN, J., MA, C. and ZHONG, Y. (2019). A selective overview of deep learning. Available at [arXiv:1904.05526](#).
- HOROWITZ, J. L. and MAMMEN, E. (2007). Rate-optimal estimation for a general class of nonparametric regression models with unknown link functions. *Ann. Statist.* **35** 2589–2619. [MR2382659](#) <https://doi.org/10.1214/009053607000000415>
- KOHLER, M. and KRZYŻAK, A. (2017). Nonparametric regression based on hierarchical interaction models. *IEEE Trans. Inf. Theory* **63** 1620–1630. [MR3625984](#) <https://doi.org/10.1109/TIT.2016.2634401>
- KOHLER, M., KRZYŻAK, A. and LANGER, S. (2019). Deep learning and MARS: A connection. Available at [arXiv:1908.11140](#).
- KOHLER, M. and LANGER, S. (2019). On the rate of convergence of fully connected very deep neural network regression estimates using ReLU activation functions. Available at [arXiv:1908.11133](#).

DISCUSSION OF: “NONPARAMETRIC REGRESSION USING DEEP NEURAL NETWORKS WITH RELU ACTIVATION FUNCTION”

BY OHAD SHAMIR

Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, ohad.shamir@weizmann.ac.il

REFERENCES

- ANTHONY, M. and BARTLETT, P. L. (2009). *Neural Network Learning: Theoretical Foundations*. Cambridge Univ. Press, Cambridge. [MR1741038](#) <https://doi.org/10.1017/CBO9780511624216>
- ARORA, S., COHEN, N., HU, W. and LUO, Y. (2019). Implicit regularization in deep matrix factorization. Preprint. Available at [arXiv:1905.13655](#).
- BARTLETT, P. L., FOSTER, D. J. and TELGARSKY, M. J. (2017). Spectrally-normalized margin bounds for neural networks. In *NIPS*.
- BARTLETT, P. L., LONG, P. M., LUGOSI, G. and TSIGLER, A. (2019). Benign overfitting in linear regression. Preprint. Available at [arXiv:1906.11300](#).
- 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>
- CHOROMANSKA, A., HENAFF, M., MATHIEU, M., AROUS, G. B. and LECUN, Y. (2015). The loss surfaces of multilayer networks. In *AISTATS*.
- DAVIS, D., DRUSVYATSKIY, D., KAKADE, S. and LEE, J. D. (2020). Stochastic subgradient method converges on tame functions. *Found. Comput. Math.* **20** 119–154. [MR4056927](#) <https://doi.org/10.1007/s10208-018-09409-5>
- DZIUGAITE, G. K. and ROY, D. M. (2017). Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. Preprint. Available at [arXiv:1703.11008](#).
- GHORBANI, B., MEI, S., MISIAKIEWICZ, T. and MONTANARI, A. (2019). Limitations of lazy training of two-layers neural networks. Preprint. Available at [arXiv:1906.08899](#).
- GOLOWICH, N., RAKHLIN, A. and SHAMIR, O. (2018). Size-independent sample complexity of neural networks. In *COLT*.
- HASTIE, T., MONTANARI, A., ROSSET, S. and TIBSHIRANI, R. J. (2019). Surprises in high-dimensional ridge-less least squares interpolation. Preprint. Available at [arXiv:1903.08560](#).
- LI, Y., MA, T. and ZHANG, H. (2018). Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *COLT*.
- MEI, S., BAI, Y. and MONTANARI, A. (2018). The landscape of empirical risk for nonconvex losses. *Ann. Statist.* **46** 2747–2774. [MR3851754](#) <https://doi.org/10.1214/17-AOS1637>
- NEYSHABUR, B., TOMIOKA, R. and SREBRO, N. (2015). Norm-based capacity control in neural networks. In *COLT*.
- RAHIMI, A. and RECHT, B. (2008). Random features for large-scale kernel machines. In *NIPS*.
- SCHAPIRE, R. E., FREUND, Y., BARTLETT, P. and LEE, W. S. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *Ann. Statist.* **26** 1651–1686. [MR1673273](#) <https://doi.org/10.1214/aos/1024691352>
- SHALEV-SHWARTZ, S., SHAMIR, O. and SHAMMAH, S. (2017). Failures of gradient-based deep learning. In *ICML*.
- SHAMIR, O. (2018). Distribution-specific hardness of learning neural networks. *J. Mach. Learn. Res.* **19** Art. ID 32. [MR3862439](#)
- SOLTANOLKOTABI, M. (2017). Learning relus via gradient descent. In *NIPS*.
- SOUDRY, D., HOFFER, E., NACSON, M. S., GUNASEKAR, S. and SREBRO, N. (2018). The implicit bias of gradient descent on separable data. *J. Mach. Learn. Res.* **19** Art. ID 70. [MR3899772](#)
- YEHUDAI, G. and SHAMIR, O. (2019). On the power and limitations of random features for understanding neural networks. In *NeurIPS*.
- ZHANG, C., BENGIO, S., HARDT, M., RECHT, B. and VINYALS, O. (2016). Understanding deep learning requires rethinking generalization. Preprint. Available at [arXiv:1611.03530](#).

REJOINER: “NONPARAMETRIC REGRESSION USING DEEP NEURAL NETWORKS WITH RELU ACTIVATION FUNCTION”

BY JOHANNES SCHMIDT-HIEBER

Department of Applied Mathematics, University of Twente, a.j.schmidt-hieber@utwente.nl

REFERENCES

- [1] AHMAD, S. and SCHEINKMAN, L. (2019). How can we be so dense? The benefits of using highly sparse representations. arXiv e-prints [arXiv:1903.11257](https://arxiv.org/abs/1903.11257).
- [2] BARRON, A. R. and KLUSOWSKI, J. M. (2018). Approximation and estimation for high-dimensional deep learning networks. arXiv e-prints [arXiv:1809.03090](https://arxiv.org/abs/1809.03090).
- [3] BELKIN, M., RAKHLIN, A. and TSYBAKOV, A. B. (2019). Does data interpolation contradict statistical optimality? In *Proceedings of Machine Learning Research* (K. Chaudhuri and M. Sugiyama, eds.). *Proceedings of Machine Learning Research* **89** 1611–1619. PMLR.
- [4] ECKLE, K. and SCHMIDT-HIEBER, J. (2019). A comparison of deep networks with ReLU activation function and linear spline-type methods. *Neural Netw.* **110** 232–242. <https://doi.org/10.1016/j.neunet.2018.11.005>
- [5] FRANKLE, J. and CARBIN, M. (2018). The lottery ticket hypothesis: Finding sparse, trainable neural networks. arXiv:1803.03635.
- [6] GAIER, A. and HA, D. (2019). Weight agnostic neural networks. arXiv e-prints [arXiv:1906.04358](https://arxiv.org/abs/1906.04358).
- [7] GALE, T., ELSEN, E. and HOOKER, S. (2019). The state of sparsity in deep neural networks. arXiv e-prints [arXiv:1902.09574](https://arxiv.org/abs/1902.09574).
- [8] GOWER, R. M. and RICHTÁRIK, P. (2015). Randomized iterative methods for linear systems. *SIAM J. Matrix Anal. Appl.* **36** 1660–1690. [MR3432148 https://doi.org/10.1137/15M1025487](https://doi.org/10.1137/15M1025487)
- [9] GREEN, P. J. and SILVERMAN, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach. Monographs on Statistics and Applied Probability* **58**. CRC Press, London. [MR1270012 https://doi.org/10.1007/978-1-4899-4473-3](https://doi.org/10.1007/978-1-4899-4473-3)
- [10] GUNASEKAR, S., LEE, J., SOUDRY, D. and SREBRO, N. (2018). Characterizing implicit bias in terms of optimization geometry. arXiv e-prints [arXiv:1802.08246](https://arxiv.org/abs/1802.08246).
- [11] HAN, S., POOL, J., TRAN, J. and DALLY, W. (2015). Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems* 28 (C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama and R. Garnett, eds.) 1135–1143. Curran Associates, Red Hook, NY.
- [12] HASSIBI, B. and STORK, D. G. (1993). Second order derivatives for network pruning: Optimal brain surgeon. In *Advances in Neural Information Processing Systems* 5 (S. J. Hanson, J. D. Cowan and C. L. Giles, eds.) 164–171. Morgan Kaufmann, Burlington, MA.
- [13] HENDRYCKS, D. and DIETTERICH, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*.
- [14] KERKYCHARIAN, G. and PICARD, D. (2002). Minimax or maxisets? *Bernoulli* **8** 219–253. [MR1895892](https://doi.org/10.1017/S10835146701804316)
- [15] LECUN, Y., DENKER, J. S. and SOLLA, S. A. (1990). Optimal brain damage. In *Advances in Neural Information Processing Systems* 2 (D. S. Touretzky, ed.) 598–605. Morgan Kaufmann, Burlington, MA.
- [16] MANSEL GOWER, R. and RICHTARIK, P. (2015). Stochastic dual ascent for solving linear systems. arXiv e-prints [arXiv:1512.06890](https://arxiv.org/abs/1512.06890).
- [17] MOCANU, D. C., MOCANU, E., STONE, P., NGUYEN, P. H., GIBESCU, M. and LIOTTA, A. (2018). Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science. *Nat. Commun.* **9** 2383. <https://doi.org/10.1038/s41467-018-04316-3>
- [18] NAKADA, R. and IMAIZUMI, M. (2019). Adaptive approximation and estimation of deep neural network with intrinsic dimensionality. arXiv e-prints [arXiv:1907.02177](https://arxiv.org/abs/1907.02177).
- [19] PRABHU, A., VARMA, G. and NAMBOODIRI, A. (2018). Deep expander networks: Efficient deep networks from graph theory. In *Computer Vision—ECCV 2018* (V. Ferrari, M. Hebert, C. Sminchisescu and Y. Weiss, eds.) 20–36. Springer, Cham.
- [20] ROBINETT, R. A. and KEPNER, J. (2018). Neural network topologies for sparse training. arXiv e-prints [arXiv:1809.05242](https://arxiv.org/abs/1809.05242).
- [21] SAVARESE, P., EVRON, I., SOUDRY, D. and SREBRO, N. (2019). How do infinite width bounded norm networks look in function space?. arXiv e-prints [arXiv:1902.05040](https://arxiv.org/abs/1902.05040).

- [22] SCHMIDT-HIEBER, J. (2019). Deep ReLU network approximation of functions on a manifold. arXiv e-prints [arXiv:1908.00695](https://arxiv.org/abs/1908.00695).
- [23] SOUDRY, D., HOFFER, E., Nacson, M. S., GUNASEKAR, S. and SREBRO, N. (2018). The implicit bias of gradient descent on separable data. *J. Mach. Learn. Res.* **19** Paper No. 70, 57. [MR3899772](https://arxiv.org/abs/1806.04839)
- [24] STROHMER, T. and VERSHYNIN, R. (2009). A randomized Kaczmarz algorithm with exponential convergence. *J. Fourier Anal. Appl.* **15** 262–278. [MR2500924](https://arxiv.org/abs/0901.0711) <https://doi.org/10.1007/s00041-008-9030-4>
- [25] YAROTSKY, D. (2018). Optimal approximation of continuous functions by very deep ReLU networks. In *Proceedings of the 31st Conference on Learning Theory* (S. Bubeck, V. Perchet and P. Rigollet, eds.). *Proceedings of Machine Learning Research* **75** 639–649. PMLR.
- [26] ZHANG, C., BENGIO, S., HARDT, M., RECHT, B. and VINYALS, O. (2016). Understanding deep learning requires rethinking generalization. arXiv e-prints [arXiv:1611.03530](https://arxiv.org/abs/1611.03530).

NONCLASSICAL BERRY–ESSEEN INEQUALITIES AND ACCURACY OF THE BOOTSTRAP

BY MAYYA ZHILOVA

School of Mathematics, Georgia Institute of Technology, mzhilova@math.gatech.edu

We study accuracy of bootstrap procedures for estimation of quantiles of a smooth function of a sum of independent sub-Gaussian random vectors. We establish higher-order approximation bounds with error terms depending on a sample size and a dimension explicitly. These results lead to improvements of accuracy of a weighted bootstrap procedure for general log-likelihood ratio statistics. The key element of our proofs of the bootstrap accuracy is a multivariate higher-order Berry–Esseen inequality. We consider a problem of approximation of distributions of two sums of zero mean independent random vectors, such that summands with the same indices have equal moments up to at least the second order. The derived approximation bound is uniform on the sets of all Euclidean balls. The presented approach extends classical Berry–Esseen type inequalities to higher-order approximation bounds. The theoretical results are illustrated with numerical experiments.

REFERENCES

- [1] AKHIEZER, N. I. (1965). *The Classical Moment Problem and Some Related Questions in Analysis*. Translated by N. Kemmer. Hafner Publishing Co., New York. [MR0184042](#)
- [2] ARLOT, S., BLANCHARD, G. and ROQUAIN, E. (2010). Some nonasymptotic results on resampling in high dimension. I. Confidence regions. *Ann. Statist.* **38** 51–82. [MR2589316](#) <https://doi.org/10.1214/08-AOS667>
- [3] BARBE, P. and BERTAIL, P. (1995). *The Weighted Bootstrap. Lecture Notes in Statistics* **98**. Springer, New York. [MR2195545](#) <https://doi.org/10.1007/978-1-4612-2532-4>
- [4] BENTKUS, V. (2003). On the dependence of the Berry–Esseen bound on dimension. *J. Statist. Plann. Inference* **113** 385–402. [MR1965117](#) [https://doi.org/10.1016/S0378-3758\(02\)00094-0](https://doi.org/10.1016/S0378-3758(02)00094-0)
- [5] BENTKUS, V. (2004). A Lyapunov type bound in $\mathbf{R}^d$. *Theory Probab. Appl.* **49** 311–323. [MR2144310](#) <https://doi.org/10.1137/S0040585X97981123>
- [6] BHATTACHARYA, R. N. and GHOSH, J. K. (1978). On the validity of the formal Edgeworth expansion. *Ann. Statist.* **6** 434–451. [MR0471142](#)
- [7] BICKEL, P. J. and FREEDMAN, D. A. (1983). Bootstrapping regression models with many parameters. In *A Festschrift for Erich L. Lehmann*. Wadsworth Statist./Probab. Ser. 28–48. Wadsworth, Belmont, CA. [MR0689736](#)
- [8] BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge Univ. Press, Cambridge. [MR2061575](#) <https://doi.org/10.1017/CBO9780511804441>
- [9] CHATTERJEE, S. and BOSE, A. (2005). Generalized bootstrap for estimating equations. *Ann. Statist.* **33** 414–436. [MR2157808](#) <https://doi.org/10.1214/009053604000000904>
- [10] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41** 2786–2819. [MR3161448](#) <https://doi.org/10.1214/13-AOS1161>
- [11] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2017). Central limit theorems and bootstrap in high dimensions. *Ann. Probab.* **45** 2309–2352. [MR3693963](#) <https://doi.org/10.1214/16-AOP1113>
- [12] CURTO, R. E. and FIALKOW, L. A. (1991). Recursiveness, positivity, and truncated moment problems. *Houston J. Math.* **17** 603–635. [MR1147276](#)

MSC2020 subject classifications. Primary 62E17, 62F40; secondary 62F25.

Key words and phrases. Multivariate Berry–Esseen inequality, dependence on dimension, Efron’s bootstrap, weighted bootstrap, multiplier bootstrap, higher-order inference, smooth function model, likelihood-based confidence sets.

- [13] CURTO, R. E. and FIALKOW, L. A. (2008). An analogue of the Riesz–Haviland theorem for the truncated moment problem. *J. Funct. Anal.* **255** 2709–2731. MR2464189 <https://doi.org/10.1016/j.jfa.2008.09.003>
- [14] EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26. MR0515681
- [15] EFRON, B. and TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap. Monographs on Statistics and Applied Probability* **57**. CRC Press, New York. MR1270903 <https://doi.org/10.1007/978-1-4899-4541-9>
- [16] HALL, P. (1992). *The Bootstrap and Edgeworth Expansion. Springer Series in Statistics*. Springer, New York. MR1145237 <https://doi.org/10.1007/978-1-4612-4384-7>
- [17] HOLMES, S. and REINERT, G. (2004). Stein’s method for the bootstrap. In *Stein’s Method: Expository Lectures and Applications. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **46** 95–136. IMS, Beachwood, OH. MR2118605
- [18] HOLMQUIST, B. (1988). Moments and cumulants of the multivariate normal distribution. *Stoch. Anal. Appl.* **6** 273–278. MR0949679 <https://doi.org/10.1080/07362998808809148>
- [19] HOROWITZ, J. L. (2001). The bootstrap. *Handbook of Econometrics* **5** 3159–3228.
- [20] HSU, D., KAKADE, S. M. and ZHANG, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.* **17** no. 52, 6. MR2994877 <https://doi.org/10.1214/ECP.v17-2079>
- [21] IBRAGIMOV, I. A. (1966). On the accuracy of approximation by the normal distribution of distribution functions of sums of independent random variables. *Theory Probab. Appl.* **11** 559–579.
- [22] JANSSEN, A. and PAULS, T. (2003). How do bootstrap and permutation tests work? *Ann. Statist.* **31** 768–806. MR1994730 <https://doi.org/10.1214/aos/1056562462>
- [23] LIU, R. Y. (1988). Bootstrap procedures under some non-i.i.d. models. *Ann. Statist.* **16** 1696–1708. MR0964947 <https://doi.org/10.1214/aos/1176351062>
- [24] LOÈVE, M. (1950). Fundamental limit theorems of probability theory. *Ann. Math. Stat.* **21** 321–338. MR0036467 <https://doi.org/10.1214/aoms/1177729793>
- [25] 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>
- [26] MAMMEN, E. (1992). *When Does Bootstrap Work?* **77**. Springer.
- [27] MAMMEN, E. (1993). Bootstrap and wild bootstrap for high-dimensional linear models. *Ann. Statist.* **21** 255–285. MR1212176 <https://doi.org/10.1214/aos/1176349025>
- [28] PAULAUSKAS, V. I. (1975). An estimate for the remainder term in the multidimensional central limit theorem. *Lith. Math. J.* **15** 484–493.
- [29] PRÆSTGAARD, J. (1990). Bootstrap with general weights and multiplier central limit theorems. Technical Report 195, Dept. Statistics, Univ. Washington.
- [30] PRÆSTGAARD, J. and WELLNER, J. A. (1993). Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.* **21** 2053–2086. MR1245301
- [31] ROTAR’, V. I. (1977). Nonclassical estimates of the rate of convergence in the multidimensional central limit theorem. I. *Theory Probab. Appl.* **22** 755–772.
- [32] SCHWARTZ, J. T. (1980). Fast probabilistic algorithms for verification of polynomial identities. *J. Assoc. Comput. Mach.* **27** 701–717. MR0594695 <https://doi.org/10.1145/322217.322225>
- [33] SHAO, J. and TU, D. S. (1995). *The Jackknife and Bootstrap. Springer Series in Statistics*. Springer, New York. MR1351010 <https://doi.org/10.1007/978-1-4612-0795-5>
- [34] SINGH, K. (1981). On the asymptotic accuracy of Efron’s bootstrap. *Ann. Statist.* **9** 1187–1195. MR0630102
- [35] SPOKOINY, V. and ZHILOVA, M. (2015). Bootstrap confidence sets under model misspecification. *Ann. Statist.* **43** 2653–2675. MR3405607 <https://doi.org/10.1214/15-AOS1355>
- [36] UL’YANOV, V. V. (1979). On more precise convergence rate estimates in the central limit theorem. *Theory Probab. Appl.* **23** 660–663.
- [37] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes. Springer Series in Statistics*. Springer, New York. MR1385671 <https://doi.org/10.1007/978-1-4757-2545-2>
- [38] WU, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.* **14** 1261–1350. MR0868303 <https://doi.org/10.1214/aos/1176350142>
- [39] ZHAI, A. (2018). A high-dimensional CLT in W_2 distance with near optimal convergence rate. *Probab. Theory Related Fields* **170** 821–845. MR3773801 <https://doi.org/10.1007/s00440-017-0771-3>
- [40] ZHILOVA, M. (2015). Simultaneous likelihood-based bootstrap confidence sets for a large number of models. Available at [arXiv:1506.05779](https://arxiv.org/abs/1506.05779).
- [41] ZHILOVA, M. (2020). Supplement to “Nonclassical Berry–Esseen inequalities and accuracy of the bootstrap.” <https://doi.org/10.1214/18-AOS1802SUPP>.

- [42] ZIPPEL, R. (1979). Probabilistic algorithms for sparse polynomials. In *Symbolic and Algebraic Computation (EUROSAM '79, Internat. Sympos., Marseille, 1979)*. *Lecture Notes in Computer Science* **72** 216–226. Springer, Berlin. [MR0575692](#)
- [43] ZOLOTAREV, V. M. (1965). On the closeness of the distributions of two sums of independent random variables. *Theory Probab. Appl.* **10** 472–479.

ON THE VALIDITY OF THE FORMAL EDGEWORTH EXPANSION FOR POSTERIOR DENSITIES

BY JOHN E. KOLASSA¹ AND TODD A. KUFFNER²

¹*Department of Statistics and Biostatistics, Rutgers University, kolassa@stat.rutgers.edu*

²*Department of Mathematics and Statistics, Washington University in St. Louis, kuffner@wustl.edu*

We consider a fundamental open problem in parametric Bayesian theory, namely the validity of the formal Edgeworth expansion of the posterior density. While the study of valid asymptotic expansions for posterior distributions constitutes a rich literature, the validity of the formal Edgeworth expansion has not been rigorously established. Several authors have claimed connections of various posterior expansions with the classical Edgeworth expansion, or have simply assumed its validity. Our main result settles this open problem. We also prove a lemma concerning the order of posterior cumulants which is of independent interest in Bayesian parametric theory. The most relevant literature is synthesized and compared to the newly-derived Edgeworth expansions. Numerical investigations illustrate that our expansion has the behavior expected of an Edgeworth expansion, and that it has better performance than the other existing expansion which was previously claimed to be of Edgeworth type.

REFERENCES

- [1] BARNDORFF-NIELSEN, O. and COX, D. R. (1979). Edgeworth and saddle-point approximations with statistical applications. *J. Roy. Statist. Soc. Ser. B* **41** 279–312. [MR0557595](#)
- [2] BARNDORFF-NIELSEN, O. E. and COX, D. R. (1989). *Asymptotic Techniques for Use in Statistics. Monographs on Statistics and Applied Probability*. CRC Press, London. [MR1010226](#) <https://doi.org/10.1007/978-1-4899-3424-6>
- [3] BERTAIL, P. and LO, A. Y. (2002). On Johnson’s asymptotic expansion for a posterior distribution. Technical Report No. 2002-39, Institut National de la Statistique et des Etudes Economiques, Série des Documents de Travail du CREST (Centre de Recherche en Economie et Statistique).
- [4] BHATTACHARYA, R. N. and GHOSH, J. K. (1978). On the validity of the formal Edgeworth expansion. *Ann. Statist.* **6** 434–451. [MR0471142](#)
- [5] BHATTACHARYA, R. N. and RAO, R. R. (2010). *Normal Approximation and Asymptotic Expansions. Classics in Applied Mathematics* **64**. SIAM, Philadelphia, PA. [MR3396213](#) <https://doi.org/10.1137/1.9780898719895.ch1>
- [6] BICKEL, P. J. and GHOSH, J. K. (1990). A decomposition for the likelihood ratio statistic and the Bartlett correction—A Bayesian argument. *Ann. Statist.* **18** 1070–1090. [MR1062699](#) <https://doi.org/10.1214/aos/1176347740>
- [7] BILLINGSLEY, P. (1995). *Probability and Measure*, 3rd ed. *Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. [MR1324786](#)
- [8] BRILLINGER, D. R. (1969). The calculation of cumulants via conditioning. *Ann. Inst. Statist. Math.* **21** 215–218.
- [9] CHANG, I. H. and MUKERJEE, R. (2006). Probability matching property of adjusted likelihoods. *Statist. Probab. Lett.* **76** 838–842. [MR2266098](#) <https://doi.org/10.1016/j.spl.2005.10.015>
- [10] DASGUPTA, A. (2008). *Asymptotic Theory of Statistics and Probability. Springer Texts in Statistics*. Springer, New York. [MR2664452](#)
- [11] DATTA, G. S. and MUKERJEE, R. (2004). *Probability Matching Priors: Higher Order Asymptotics. Lecture Notes in Statistics* **178**. Springer, New York. [MR2053794](#) <https://doi.org/10.1007/978-1-4612-2036-7>
- [12] DAVIS, A. W. (1976). Statistical distributions in univariate and multivariate Edgeworth populations. *Biometrika* **63** 661–670. [MR0518636](#) <https://doi.org/10.1093/biomet/63.3.661>

- [13] DAVISON, A. C. (1986). Approximate predictive likelihood. *Biometrika* **73** 323–332. [MR0855892](#)
<https://doi.org/10.1093/biomet/73.2.323>
- [14] DICICCIO, T. J. and STERN, S. E. (1993). On Bartlett adjustments for approximate Bayesian inference. *Biometrika* **80** 731–740. [MR1282782](#) <https://doi.org/10.1093/biomet/80.4.731>
- [15] DICICCIO, T. J. and STERN, S. E. (1994). Frequentist and Bayesian Bartlett correction of test statistics based on adjusted profile likelihoods. *J. Roy. Statist. Soc. Ser. B* **56** 397–408. [MR1281943](#)
- [16] FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications, Vol. II*, 2nd ed. Wiley, New York. [MR0270403](#)
- [17] GHOSH, J. K. (1994). *Higher-Order Asymptotics*. IMS, Hayward.
- [18] GHOSH, J. K., DELAMPADY, M. and SAMANTA, T. (2006). *An Introduction to Bayesian Analysis: Theory and Methods*. Springer Texts in Statistics. Springer, New York. [MR2247439](#)
- [19] GHOSH, J. K. and RAMAMOORTHY, R. V. (2003). *Bayesian Nonparametrics*. Springer Series in Statistics. Springer, New York. [MR1992245](#)
- [20] GHOSH, J. K., SINHA, B. K. and JOSHI, S. N. (1982). Expansions for posterior probability and integrated Bayes risk. In *Statistical Decision Theory and Related Topics, III, Vol. 1* (West Lafayette, Ind., 1981) (S. Gupta and J. Berger, eds.) 403–456. Academic Press, New York. [MR0705299](#)
- [21] HALL, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer Series in Statistics. Springer, New York. [MR1145237](#) <https://doi.org/10.1007/978-1-4612-4384-7>
- [22] HARTIGAN, J. A. (1965). The asymptotically unbiased prior distribution. *Ann. Math. Stat.* **36** 1137–1152. [MR0176539](#) <https://doi.org/10.1214/aoms/1177699988>
- [23] JAMES, G. S. (1955). Cumulants of a transformed variate. *Biometrika* **42** 529–531. [MR0071681](#)
<https://doi.org/10.1093/biomet/42.3-4.529>
- [24] JAMES, G. S. (1958). On moments and cumulants of systems of statistics. *Sankhyā* **20** 1–30. [MR0103550](#)
- [25] JAMES, G. S. and MAYNE, A. J. (1962). Cumulants of functions of random variables. *Sankhyā, Ser. A* **24** 47–54. [MR0148183](#)
- [26] JENSEN, J. L. (1995). *Saddlepoint Approximations*. Oxford Statistical Science Series **16**. Oxford University Press, New York. [MR1354837](#)
- [27] JOHNSON, R. A. (1967). An asymptotic expansion for posterior distributions. *Ann. Math. Stat.* **38** 1899–1906. [MR0219161](#) <https://doi.org/10.1214/aoms/1177698624>
- [28] JOHNSON, R. A. (1970). Asymptotic expansions associated with posterior distributions. *Ann. Math. Stat.* **41** 851–864. [MR0263198](#) <https://doi.org/10.1214/aoms/1177696963>
- [29] KASS, R. E., TIERNEY, L. and KADANE, J. P. (1990). The validity of posterior expansions based on Laplace's method. In *Bayesian and Likelihood Methods in Statistics and Econometrics* (S. Geisser, J. S. Hodges, S. J. Press and A. Zellner, eds.) 473–488. Elsevier Science Publishers B.V., Amsterdam.
- [30] KHARROUBI, S. A. and SWEETING, T. J. (2016). Exponential tilting in Bayesian asymptotics. *Biometrika* **103** 337–349. [MR3509890](#) <https://doi.org/10.1093/biomet/asw018>
- [31] KOLASSA, J. E. (2006). *Series Approximation Methods in Statistics*, 3rd ed. *Lecture Notes in Statistics* **88**. Springer, New York. [MR1295242](#) <https://doi.org/10.1007/978-1-4757-4275-6>
- [32] LEONOV, V. P. and SIRJAEV, A. N. (1959). On a method of semi-invariants. *Theory Probab. Appl.* **4** 319–329. [MR0123345](#) <https://doi.org/10.1137/1104031>
- [33] LINDLEY, D. V. (1961). The use of prior probability distributions in statistical inference and decisions. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I* 453–468. Univ. California Press, Berkeley, CA. [MR0156437](#)
- [34] LINDLEY, D. V. (1980). Approximate Bayesian methods. *Trab. Estad. Investig. Oper.* **31** 223–237.
- [35] MCCULLAGH, P. (1987). *Tensor Methods in Statistics*. Monographs on Statistics and Applied Probability. CRC Press, London. [MR0907286](#)
- [36] MYKLAND, P. A. (1999). Bartlett identities and large deviations in likelihood theory. *Ann. Statist.* **27** 1105–1117. [MR1724043](#) <https://doi.org/10.1214/aos/1018031270>
- [37] PERICCHI, L. R., SANSÓ, B. and SMITH, A. F. M. (1993). Posterior cumulant relationships in Bayesian inference involving the exponential family. *J. Amer. Statist. Assoc.* **88** 1419–1426. [MR1245379](#)
- [38] PROSCHAN, F. (1963). Theoretical explanation of observed decreasing failure rate. *Technometrics* **5** 375–383.
- [39] REID, N. (2003). Asymptotics and the theory of inference. *Ann. Statist.* **31** 1695–1731. [MR2036388](#)
<https://doi.org/10.1214/aos/1074290325>
- [40] RULI, E., SARTORI, N. and VENTURA, L. (2014). Marginal posterior simulation via higher-order tail area approximations. *Bayesian Anal.* **9** 129–145. [MR3188302](#) <https://doi.org/10.1214/13-BA851>
- [41] SPEED, T. P. (1983). Cumulants and partition lattices. *Aust. J. Stat.* **25** 378–388. [MR0725217](#)
- [42] STUART, A. and ORD, J. K. (1994). *Kendall's Advanced Theory of Statistics, Vol. 1*, 6th ed. Wiley, New York. [MR1280717](#)

- [43] SWEETING, T. J. (1995). A framework for Bayesian and likelihood approximations in statistics. *Biometrika* **82** 1–23. [MR1332836](#) <https://doi.org/10.1093/biomet/82.1.1>
- [44] TIERNEY, L. and KADANE, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81** 82–86. [MR0830567](#)
- [45] TIERNEY, L., KASS, R. E. and KADANE, J. B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *J. Amer. Statist. Assoc.* **84** 710–716. [MR1132586](#)
- [46] 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>
- [47] VENTURA, L. and REID, N. (2014). Approximate Bayesian computation with modified log-likelihood ratios. *Metron* **72** 231–245. [MR3233151](#) <https://doi.org/10.1007/s40300-014-0041-4>
- [48] WALLACE, D. L. (1958). Asymptotic approximations to distributions. *Ann. Math. Stat.* **29** 635–654. [MR0097109](#) <https://doi.org/10.1214/aoms/1177706528>
- [49] WENG, R. C. (2003). On Stein’s identity for posterior normality. *Statist. Sinica* **13** 495–506. [MR1977739](#)
- [50] WENG, R. C. (2010). A Bayesian Edgeworth expansion by Stein’s identity. *Bayesian Anal.* **5** 741–763. [MR2740155](#) <https://doi.org/10.1214/10-BA526>
- [51] WENG, R. C. and TSAI, W.-C. (2008). Asymptotic posterior normality for multiparameter problems. *J. Statist. Plann. Inference* **138** 4068–4080. [MR2455988](#) <https://doi.org/10.1016/j.jspi.2008.03.034>
- [52] WITHERS, C. S. (1982). Second order inference for asymptotically normal random variables. *Sankhyā, Ser. B* **44** 19–27. [MR0692891](#)
- [53] WITHERS, C. S. (1984). Asymptotic expansions for distributions and quantiles with power series cumulants. *J. Roy. Statist. Soc. Ser. B* **46** 389–396. [MR0790623](#)
- [54] WOODROOFE, M. (1989). Very weak expansions for sequentially designed experiments: Linear models. *Ann. Statist.* **17** 1087–1102. [MR1015139](#) <https://doi.org/10.1214/aos/1176347257>
- [55] WOODROOFE, M. (1992). Integrable expansions for posterior distributions for one-parameter exponential families. *Statist. Sinica* **2** 91–111. [MR1152299](#)

MODEL SELECTION FOR HIGH-DIMENSIONAL LINEAR REGRESSION WITH DEPENDENT OBSERVATIONS

BY CHING-KANG ING

Institute of Statistics, National Tsing Hua University, cking@stat.nthu.edu.tw

We investigate the prediction capability of the orthogonal greedy algorithm (OGA) in high-dimensional regression models with dependent observations. The rates of convergence of the prediction error of OGA are obtained under a variety of sparsity conditions. To prevent OGA from overfitting, we introduce a high-dimensional Akaike's information criterion (HDAIC) to determine the number of OGA iterations. A key contribution of this work is to show that OGA, used in conjunction with HDAIC, can achieve the optimal convergence rate without knowledge of how sparse the underlying high-dimensional model is.

REFERENCES

- ADAMCZAK, R. and WOLFF, P. (2015). Concentration inequalities for non-Lipschitz functions with bounded derivatives of higher order. *Probab. Theory Related Fields* **162** 531–586. MR3383337 <https://doi.org/10.1007/s00440-014-0579-3>
- BASU, S. and MICHAILIDIS, G. (2015). Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.* **43** 1535–1567. MR3357870 <https://doi.org/10.1214/15-AOS1315>
- BAXTER, G. (1962). An asymptotic result for the finite predictor. *Math. Scand.* **10** 137–144. MR0149584 <https://doi.org/10.7146/math.scand.a-10520>
- BERK, K. N. (1974). Consistent autoregressive spectral estimates. *Ann. Statist.* **2** 489–502. MR0421010
- BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469 <https://doi.org/10.1214/08-AOS620>
- BÜHLMANN, P. (2006). Boosting for high-dimensional linear models. *Ann. Statist.* **34** 559–583. MR2281878 <https://doi.org/10.1214/009053606000000092>
- CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n . *Ann. Statist.* **35** 2313–2351. MR2382644 <https://doi.org/10.1214/009053606000001523>
- CHEN, J. and CHEN, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95** 759–771. MR2443189 <https://doi.org/10.1093/biomet/asn034>
- DING, J., TAROKH, V. and YANG, Y. (2018). Bridging AIC and BIC: A new criterion for autoregression. *IEEE Trans. Inform. Theory* **64** 4024–4043. MR3809725 <https://doi.org/10.1109/TIT.2017.2717599>
- FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911. MR2530322 <https://doi.org/10.1111/j.1467-9868.2008.00674.x>
- GAO, F., ING, C.-K. and YANG, Y. (2013). Metric entropy and sparse linear approximation of ℓ_q -hulls for $0 < q \leq 1$. *J. Approx. Theory* **166** 42–55. MR3003947 <https://doi.org/10.1016/j.jat.2012.10.002>
- HUANG, H.-H. and ING, C.-K. (2019). Concentration inequalities for sample higher order moments of stationary linear processes. Technical Report, Institute of Statistics, National Tsing Hua Univ.
- ING, C.-K. (2007). Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series. *Ann. Statist.* **35** 1238–1277. MR2341705 <https://doi.org/10.1214/009053606000001550>
- ING, C.-K. (2020). Supplement to “Model selection for high-dimensional linear regression with dependent observations.” <https://doi.org/10.1214/19-AOS1872SUPP>.
- ING, C.-K. and LAI, T. L. (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statist. Sinica* **21** 1473–1513. MR2895106 <https://doi.org/10.5705/ss.2010.081>
- ING, C.-K., LAI, T. L., SHEN, M., TSANG, K. and YU, S.-H. (2017). Multiple testing in regression models with applications to fault diagnosis in the big data era. *Technometrics* **59** 351–360. MR3677966 <https://doi.org/10.1080/00401706.2016.1236755>

MSC2020 subject classifications. Primary 63M30; secondary 62F07, 62F12.

Key words and phrases. Best m -term approximations, high-dimensional Akaike's information criterion, orthogonal greedy algorithm, sparsity conditions, time series.

- NEGAHBAN, S. N., RAVIKUMAR, P., WAINWRIGHT, M. J. and YU, B. (2012). A unified framework for high-dimensional analysis of M -estimators with decomposable regularizers. *Statist. Sci.* **27** 538–557. [MR3025133](https://doi.org/10.1214/12-STS400) <https://doi.org/10.1214/12-STS400>
- POURAHMADI, M. (1989). On the convergence of finite linear predictors of stationary processes. *J. Multivariate Anal.* **30** 167–180. [MR1015366](https://doi.org/10.1016/0047-259X(89)90033-X) [https://doi.org/10.1016/0047-259X\(89\)90033-X](https://doi.org/10.1016/0047-259X(89)90033-X)
- RUDELSON, M. and VERSHYNIN, R. (2013). Hanson–Wright inequality and sub-Gaussian concentration. *Electron. Commun. Probab.* **18** no. 82, 9. [MR3125258](https://doi.org/10.1214/ECP.v18-2865) <https://doi.org/10.1214/ECP.v18-2865>
- SHIBATA, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. *Ann. Statist.* **8** 147–164. [MR0557560](https://doi.org/10.1214/aos/1176345138)
- TEMLYAKOV, V. N. (1998). The best m -term approximation and greedy algorithms. *Adv. Comput. Math.* **8** 249–265. [MR1628182](https://doi.org/10.1023/A:1018900431309) <https://doi.org/10.1023/A:1018900431309>
- TEMLYAKOV, V. N. (2000). Weak greedy algorithms. *Adv. Comput. Math.* **12** 213–227. [MR1745113](https://doi.org/10.1023/A:1018917218956) <https://doi.org/10.1023/A:1018917218956>
- TEMLYAKOV, V. N. (2015). Greedy approximation in convex optimization. *Constr. Approx.* **41** 269–296. [MR3315679](https://doi.org/10.1007/s00365-014-9272-0) <https://doi.org/10.1007/s00365-014-9272-0>
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. [MR1379242](https://doi.org/10.2307/2346178)
- TROPP, J. A. (2004). Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory* **50** 2231–2242. [MR2097044](https://doi.org/10.1109/TIT.2004.834793) <https://doi.org/10.1109/TIT.2004.834793>
- WANG, H. (2009). Forward regression for ultra-high dimensional variable screening. *J. Amer. Statist. Assoc.* **104** 1512–1524. [MR2750576](https://doi.org/10.1198/jasa.2008.tm08516) <https://doi.org/10.1198/jasa.2008.tm08516>
- WANG, Z., PATERLINI, S., GAO, F. and YANG, Y. (2014). Adaptive minimax regression estimation over sparse ℓ_q -hulls. *J. Mach. Learn. Res.* **15** 1675–1711. [MR3225246](https://doi.org/10.1214/13-AOS1108)
- WU, W.-B. and WU, Y. N. (2016). Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electron. J. Stat.* **10** 352–379. [MR3466186](https://doi.org/10.1214/16-EJS1108) <https://doi.org/10.1214/16-EJS1108>
- YANG, Y. (2007). Prediction/estimation with simple linear models: Is it really that simple? *Econometric Theory* **23** 1–36. [MR2338950](https://doi.org/10.1017/S0266466607070016) <https://doi.org/10.1017/S0266466607070016>
- 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) <https://doi.org/10.1214/09-AOS729>
- ZHANG, Y. and YANG, Y. (2015). Cross-validation for selecting a model selection procedure. *J. Econometrics* **187** 95–112. [MR3347297](https://doi.org/10.1016/j.jeconom.2015.02.006) <https://doi.org/10.1016/j.jeconom.2015.02.006>
- ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. [MR2274449](https://doi.org/10.1214/06-IMB1201)

OPTIMAL ESTIMATION OF GAUSSIAN MIXTURES VIA DENOISED METHOD OF MOMENTS

BY YIHONG WU¹ AND PENGKUN YANG²

¹*Department of Statistics and Data Science, Yale University, yihong.wu@yale.edu*

²*Department of Electrical Engineering, Princeton University, pengkun@princeton.edu*

The method of moments (*Philos. Trans. R. Soc. Lond. Ser. A* **185** (1894) 71–110) is one of the most widely used methods in statistics for parameter estimation, by means of solving the system of equations that match the population and estimated moments. However, in practice and especially for the important case of mixture models, one frequently needs to contend with the difficulties of non-existence or nonuniqueness of statistically meaningful solutions, as well as the high computational cost of solving large polynomial systems. Moreover, theoretical analyses of the method of moments are mainly confined to asymptotic normality style of results established under strong assumptions.

This paper considers estimating a k -component Gaussian location mixture with a common (possibly unknown) variance parameter. To overcome the aforementioned theoretic and algorithmic hurdles, a crucial step is to denoise the moment estimates by projecting to the truncated moment space (via semidefinite programming) before solving the method of moments equations. Not only does this regularization ensure existence and uniqueness of solutions, it also yields fast solvers by means of Gauss quadrature. Furthermore, by proving new moment comparison theorems in the Wasserstein distance via polynomial interpolation and majorization techniques, we establish the statistical guarantees and adaptive optimality of the proposed procedure, as well as oracle inequality in misspecified models. These results can also be viewed as provable algorithms for generalized method of moments (*Econometrica* **50** (1982) 1029–1054) which involves nonconvex optimization and lacks theoretical guarantees.

REFERENCES

- [1] AKHIEZER, N. I. (1965). *The Classical Moment Problem and Some Related Questions in Analysis*. Hafner Publishing Co., New York. [MR0184042](#)
- [2] ANANDKUMAR, A., GE, R., HSU, D., KAKADE, S. M. and TELGARSKY, M. (2014). Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.* **15** 2773–2832. [MR3270750](#)
- [3] ATKINSON, K. E. (2008). *An Introduction to Numerical Analysis*. Wiley, New York. [MR0504339](#)
- [4] BALAKRISHNAN, S., WAINWRIGHT, M. J. and YU, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Ann. Statist.* **45** 77–120. [MR3611487](#) <https://doi.org/10.1214/16-AOS1435>
- [5] BELKIN, M. and SINHA, K. (2010). Polynomial learning of distribution families. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science—FOCS 2010* 103–112. IEEE Computer Soc., Los Alamitos, CA. [MR3024780](#)
- [6] CHAUSSEÉ, P. (2010). Computing generalized method of moments and generalized empirical likelihood with R. *J. Stat. Softw.* **34** 1–35.
- [7] CHEN, J. H. (1995). Optimal rate of convergence for finite mixture models. *Ann. Statist.* **23** 221–233. [MR1331665](#) <https://doi.org/10.1214/aos/1176324464>
- [8] DASGUPTA, S. (1999). Learning mixtures of Gaussians. In *40th Annual Symposium on Foundations of Computer Science (New York, 1999)* 634–644. IEEE Computer Soc., Los Alamitos, CA. [MR1917603](#) <https://doi.org/10.1109/SFFCS.1999.814639>

MSC2020 subject classifications. Primary 62G05; secondary 62C20.

Key words and phrases. Gaussian mixture, finite mixture model, deconvolution, method of moments, Wasserstein distance, semidefinite programming, moment space, Gauss quadrature, minimax optimality.

- [9] DASKALAKIS, C., TZAMOS, C. and ZAMPETAKIS, M. (2017). Ten steps of EM suffice for mixtures of two gaussians. In *Conference on Learning Theory* 704–710.
- [10] DEELY, J. J. and KRUSE, R. L. (1968). Construction of sequences estimating the mixing distribution. *Ann. Math. Stat.* **39** 286–288. [MR0220377](#) <https://doi.org/10.1214/aoms/1177698536>
- [11] DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. [MR0501537](#)
- [12] DIACONIS, P. (1987). Application of the method of moments in probability and statistics. In *Moments in Mathematics* (San Antonio, Tex., 1987). *Proc. Sympos. Appl. Math.* **37** 125–142. Amer. Math. Soc., Providence, RI. [MR0921087](#) <https://doi.org/10.1090/psapm/037/921087>
- [13] DIAMOND, S. and BOYD, S. (2016). CVXPY: A Python-embedded modeling language for convex optimization. *J. Mach. Learn. Res.* **17** Paper No. 83, 5. [MR3517106](#)
- [14] EDELMAN, D. (1988). Estimation of the mixing distribution for a normal mean with applications to the compound decision problem. *Ann. Statist.* **16** 1609–1622. [MR0964941](#) <https://doi.org/10.1214/aos/1176351056>
- [15] FRÜHWIRTH-SCHNATTER, S. (2006). *Finite Mixture and Markov Switching Models*. *Springer Series in Statistics*. Springer, New York. [MR2265601](#)
- [16] GAUTSCHI, W. (2004). *Orthogonal Polynomials: Computation and Approximation*. *Numerical Mathematics and Scientific Computation*. Oxford Univ. Press, New York. [MR2061539](#)
- [17] 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>
- [18] 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>
- [19] GOLUB, G. H. and WELSCH, J. H. (1969). Calculation of Gauss quadrature rules. *Math. Comp.* **23** 221–230. [MR0245201](#) <https://doi.org/10.2307/2004418>
- [20] HALL, A. R. (2005). *Generalized Method of Moments*. *Advanced Texts in Econometrics*. Oxford Univ. Press, Oxford. [MR2135106](#)
- [21] HANSEN, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50** 1029–1054. [MR0666123](#) <https://doi.org/10.2307/1912775>
- [22] HARDT, M. and PRICE, E. (2015). Tight bounds for learning a mixture of two Gaussians [extended abstract]. In *STOC’15—Proceedings of the 2015 ACM Symposium on Theory of Computing* 753–760. ACM, New York. [MR3388255](#)
- [23] 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>
- [24] HOPKINS, S. B. and LI, J. (2018). Mixture models, robustness, and sum of squares proofs. In *STOC’18—Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing* 1021–1034. ACM, New York. [MR3826314](#)
- [25] HORN, R. A. and JOHNSON, C. R. (2013). *Matrix Analysis*, 2nd ed. Cambridge Univ. Press, Cambridge. [MR2978290](#)
- [26] IBRAGIMOV, I. (2001). Estimation of analytic functions. In *State of the Art in Probability and Statistics* (Leiden, 1999). *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **36** 359–383. IMS, Beachwood, OH. [MR1836570](#) <https://doi.org/10.1214/lnms/1215090078>
- [27] KALAI, A. T., MOITRA, A. and VALIANT, G. (2010). Efficiently learning mixtures of two Gaussians. In *STOC’10—Proceedings of the 2010 ACM International Symposium on Theory of Computing* 553–562. ACM, New York. [MR2743304](#)
- [28] KARLIN, S. and SHAPLEY, L. S. (1953). Geometry of moment spaces. *Mem. Amer. Math. Soc.* **12** 93. [MR0059329](#)
- [29] KARLIS, D. and XEKALAKI, E. (2003). Choosing initial values for the EM algorithm for finite mixtures. *Comput. Statist. Data Anal.* **41** 577–590. [MR1968070](#) [https://doi.org/10.1016/S0167-9473\(02\)00177-9](https://doi.org/10.1016/S0167-9473(02)00177-9)
- [30] KIEFER, J. and WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Stat.* **27** 887–906. [MR0086464](#) <https://doi.org/10.1214/aoms/1177728066>
- [31] KIM, A. K. H. (2014). Minimax bounds for estimation of normal mixtures. *Bernoulli* **20** 1802–1818. [MR3263090](#) <https://doi.org/10.3150/13-BEJ542>
- [32] 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>
- [33] KONG, W. and VALIANT, G. (2017). Spectrum estimation from samples. *Ann. Statist.* **45** 2218–2247. [MR3718167](#) <https://doi.org/10.1214/16-AOS1525>

- [34] 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>
- [35] KRAWTCHOUK, M. (1932). Sur le problème de moments. In *ICM Proceedings* 127–128. Available at <https://www.mathunion.org/fileadmin/ICM/Proceedings/ICM1932.2/ICM1932.2.ocr.pdf>.
- [36] LAIRD, N. (1978). Nonparametric maximum likelihood estimation of a mixed distribution. *J. Amer. Statist. Assoc.* **73** 805–811. MR0521328
- [37] LASSERRE, J. B. (2010). *Moments, Positive Polynomials and Their Applications*. Imperial College Press Optimization Series 1. Imperial College Press, London. MR2589247
- [38] 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.
- [39] LINDSAY, B. G. (1981). Properties of the maximum likelihood estimator of a mixing distribution. In *Statistical Distributions in Scientific Work, Vol. 5 (Trieste, 1980)*. NATO Adv. Study Inst. Ser. C: Math. Phys. Sci. **79** 95–109. Reidel, Dordrecht. MR0656328
- [40] LINDSAY, B. G. (1989). Moment matrices: Applications in mixtures. *Ann. Statist.* **17** 722–740. MR0994263 <https://doi.org/10.1214/aos/1176347138>
- [41] LINDSAY, B. G. (1995). Mixture models: Theory, geometry and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics* i–163. JSTOR.
- [42] LU, Y. and ZHOU, H. H. (2016). Statistical and computational guarantees of Lloyd’s algorithm and its variants. Preprint. Available at [arXiv:1612.02099](https://arxiv.org/abs/1612.02099).
- [43] MENG, X.-L. and VAN DYK, D. (1997). The EM algorithm—An old folk-song sung to a fast new tune. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **59** 511–567.
- [44] MOITRA, A. and VALIANT, G. (2010). Settling the polynomial learnability of mixtures of Gaussians. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science—FOCS 2010* 93–102. IEEE Computer Soc., Los Alamitos, CA. MR3024779
- [45] NGUYEN, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *Ann. Statist.* **41** 370–400. MR3059422 <https://doi.org/10.1214/12-AOS1065>
- [46] PEARSON, K. (1894). Contributions to the mathematical theory of evolution. *Philos. Trans. R. Soc. Lond. Ser. A* **185** 71–110.
- [47] PILLA, R. S. and LINDSAY, B. G. (2001). Alternative EM methods for nonparametric finite mixture models. *Biometrika* **88** 535–550. MR1844850 <https://doi.org/10.1093/biomet/88.2.535>
- [48] REDNER, R. A. and WALKER, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **26** 195–239. MR0738930 <https://doi.org/10.1137/1026034>
- [49] SCHMÜDGEN, K. (2017). *The Moment Problem*. Graduate Texts in Mathematics **277**. Springer, Cham. MR3729411
- [50] SEIDEL, W., MOSLER, K. and ALKER, M. (2000). A cautionary note on likelihood ratio tests in mixture models. *Ann. Inst. Statist. Math.* **52** 481–487. MR1794248 <https://doi.org/10.1023/A:1004117419204>
- [51] SHOHAT, J. A. and TAMARKIN, J. D. (1943). *The Problem of Moments*. American Mathematical Society Mathematical Surveys, Vol. I. Amer. Math. Soc., New York. MR0008438
- [52] STOER, J. and BULIRSCH, R. (2002). *Introduction to Numerical Analysis*, 3rd ed. Texts in Applied Mathematics **12**. Springer, New York. MR1923481 <https://doi.org/10.1007/978-0-387-21738-3>
- [53] USPENSKY, J. V. (1937). *Introduction to Mathematical Probability*. McGraw-Hill, New York.
- [54] VAN DER VAART, A. W. (2000). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics **3**. Cambridge Univ. Press, Cambridge. MR1652247 <https://doi.org/10.1017/CBO9780511802256>
- [55] VILLANI, C. (2003). *Topics in Optimal Transportation*. Graduate Studies in Mathematics **58**. Amer. Math. Soc., Providence, RI. MR1964483 <https://doi.org/10.1007/b12016>
- [56] WOLKOWICZ, H., SAIGAL, R. and VANDENBERGHE, L., eds. (2012). *Handbook of Semidefinite Programming: Theory, Algorithms, and Applications*. International Series in Operations Research & Management Science **27**. Kluwer Academic, Boston, MA. MR1778223 <https://doi.org/10.1007/978-1-4615-4381-7>
- [57] WU, Y. and YANG, P. (2020). Supplement to “Optimal estimation of Gaussian mixtures via denoised method of moments.” <https://doi.org/10.1214/19-AOS1873SUPP>.
- [58] XU, J., HSU, D. J. and MALEKI, A. (2016). Global analysis of expectation maximization for mixtures of two Gaussians. In *Advances in Neural Information Processing Systems* 2676–2684.
- [59] XU, L. and JORDAN, M. I. (1996). On convergence properties of the EM algorithm for Gaussian mixtures. *Neural Comput.* **8** 129–151.

SHARP INSTRUMENTS FOR CLASSIFYING COMPLIERS AND GENERALIZING CAUSAL EFFECTS

BY EDWARD H. KENNEDY^{*}, SIVARAMAN BALAKRISHNAN^{**} AND MAX G’SSELL[†]

Department of Statistics and Data Science, Carnegie Mellon University, ^{*}edward@stat.cmu.edu; ^{**}siva@stat.cmu.edu;
[†]mgsell@stat.cmu.edu

It is well known that, without restricting treatment effect heterogeneity, instrumental variable (IV) methods only identify “local” effects among compliers, that is, those subjects who take treatment only when encouraged by the IV. Local effects are controversial since they seem to only apply to an unidentified subgroup; this has led many to denounce these effects as having little policy relevance. However, we show that such pessimism is not always warranted: it can be possible to accurately predict who compliers are, and obtain tight bounds on more generalizable effects in identifiable subgroups. We propose methods for doing so and study estimation error and asymptotic properties, showing that these tasks can sometimes be accomplished even with very weak IVs. We go on to introduce a new measure of IV quality called “sharpness,” which reflects the variation in compliance explained by covariates, and captures how well one can identify compliers and obtain tight bounds on identifiable subgroup effects. We develop an estimator of sharpness and show that it is asymptotically efficient under weak conditions. Finally, we explore finite-sample properties via simulation, and apply the methods to study canvassing effects on voter turnout. We propose that sharpness should be presented alongside strength to assess IV quality.

REFERENCES

- [1] ABADIE, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *J. Econometrics* **113** 231–263. [MR1960380](#) [https://doi.org/10.1016/S0304-4076\(02\)00201-4](https://doi.org/10.1016/S0304-4076(02)00201-4)
- [2] ANGRIST, J. D., IMBENS, G. W. and RUBIN, D. B. (1996). Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* **91** 444–455.
- [3] ARONOW, P. M. and CARNEGIE, A. (2013). Beyond LATE: Estimation of the average treatment effect with an instrumental variable. *Polit. Anal.* **21** 492–506.
- [4] 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>
- [5] BAIocchi, M., CHENG, J. and SMALL, D. S. (2014). Instrumental variable methods for causal inference. *Stat. Med.* **33** 2297–2340. [MR3257582](#) <https://doi.org/10.1002/sim.6128>
- [6] BALKE, A. and PEARL, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *J. Amer. Statist. Assoc.* **92** 1171–1176.
- [7] 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](#)
- [8] BOUND, J., JAEGER, D. A. and BAKER, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *J. Amer. Statist. Assoc.* **90** 443–450.
- [9] CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWWEY, W. and ROBINS, J. (2016). Double machine learning for treatment and causal parameters. ArXiv preprint. Available at [arXiv:1608.00060](https://arxiv.org/abs/1608.00060).
- [10] COVER, T. and HART, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory* **13** 21–27.

MSC2020 subject classifications. 62G05, 62H30.

Key words and phrases. Causal inference, instrumental variable, noncompliance, observational study, generalizability.

- [11] DEATON, A. (2010). Instruments, randomization, and learning about development. *J. Econ. Lit.* **48** 424–455.
- [12] DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition. Applications of Mathematics (New York)* **31**. Springer, New York. MR1383093 <https://doi.org/10.1007/978-1-4612-0711-5>
- [13] DING, P. and LU, J. (2017). Principal stratification analysis using principal scores. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 757–777. MR3641406 <https://doi.org/10.1111/rssb.12191>
- [14] FELLER, A., MEALLI, F. and MIRATRIX, L. (2017). Principal score methods: Assumptions, extensions, and practical considerations. *J. Educ. Behav. Stat.* **42** 726–758.
- [15] FLUSS, R., FARAGGI, D. and REISER, B. (2005). Estimation of the Youden index and its associated cutoff point. *Biom. J.* **47** 458–472. MR2190469 <https://doi.org/10.1002/bimj.200410135>
- [16] FOLLMANN, D. A. (2000). On the effect of treatment among would-be treatment compliers: An analysis of the multiple risk factor intervention trial. *J. Amer. Statist. Assoc.* **95** 1101–1109. MR1821718 <https://doi.org/10.2307/2669746>
- [17] GREEN, D. P., GERBER, A. S. and NICKERSON, D. W. (2003). Getting out the vote in local elections: Results from six door-to-door canvassing experiments. *J. Polit.* **65** 1083–1096.
- [18] HERNÁN, M. A. and ROBINS, J. M. (2006). Instruments for causal inference: An epidemiologist’s dream? *Epidemiology* **17** 360–372.
- [19] IMBENS, G. W. (2010). Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). *J. Econ. Lit.* **48** 399–423.
- [20] IMBENS, G. W. (2014). Instrumental variables: An econometrician’s perspective. *Statist. Sci.* **29** 323–358. MR3264545 <https://doi.org/10.1214/14-STS480>
- [21] IMBENS, G. W. and ANGRIST, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62** 467–475.
- [22] IMBENS, G. W. and MANSKI, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica* **72** 1845–1857. MR2095534 <https://doi.org/10.1111/j.1468-0262.2004.00555.x>
- [23] IMBENS, G. W. and RUBIN, D. B. (1997). Estimating outcome distributions for compliers in instrumental variables models **64** 555–574. MR1485828 <https://doi.org/10.2307/2971731>
- [24] JO, B. and STUART, E. A. (2009). On the use of propensity scores in principal causal effect estimation. *Stat. Med.* **28** 2857–2875. MR2750169 <https://doi.org/10.1002/sim.3669>
- [25] JOFFE, M. M., TEN HAVE, T. R. and BRENSINGER, C. (2003). The compliance score as a regressor in randomized trials. *Biostatistics* **4** 327–340.
- [26] KENNEDY, E. H., BALAKRISHNAN, S. and G’SELL, M. (2020). Supplement to “Sharp instruments for classifying compliers and generalizing causal effects.” <https://doi.org/10.1214/19-AOS1874SUPP>.
- [27] KENNEDY, E. H., HARRIS, S. and KEELE, L. J. (2019). Survivor-complier effects in the presence of selection on treatment, with application to a study of prompt ICU admission. *J. Amer. Statist. Assoc.* **114** 93–104. MR3941240 <https://doi.org/10.1080/01621459.2018.1469990>
- [28] KENNEDY, E. H., LORCH, S. and SMALL, D. S. (2019). Robust causal inference with continuous instruments using the local instrumental variable curve. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 121–143. MR3904782
- [29] 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>
- [30] LUEDTKE, A. R. and VAN DER LAAN, M. J. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Ann. Statist.* **44** 713–742. MR3476615 <https://doi.org/10.1214/15-AOS1384>
- [31] MANSKI, C. F. (1990). Nonparametric bounds on treatment effects. *Am. Econ. Rev.* **80** 319–323.
- [32] MARCUS, S. M., STUART, E. A., WANG, P., SHADISH, W. R. and STEINER, P. M. (2012). Estimating the causal effect of randomization versus treatment preference in a doubly randomized preference trial. *Psychol. Methods* **17** 244.
- [33] PEARL, J. (2009). *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2548166 <https://doi.org/10.1017/CBO9780511803161>
- [34] ROBINS, J. LI, L., TCHETGEN TCHETGEN, E., and VAN DER VAART, A. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. *Probability and Statistics: Essays in Honor of David A. Freedman* 335–421.
- [35] ROBINS, J. M. (1989). The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In *Health Service Research Methodology: A Focus on AIDS* 113–159.
- [36] ROBINS, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Comm. Statist. Theory Methods* **23** 2379–2412. MR1293185 <https://doi.org/10.1080/03610929408831393>

- [37] ROBINS, J. M. and GREENLAND, S. (1996). Comment: Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* **91** 456–458.
- [38] RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688.
- [39] SCHISTERMAN, E. F., PERKINS, N. J., LIU, A. and BONDELL, H. (2005). Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. *Epidemiology* **16** 73–81.
- [40] STAIGER, D. and STOCK, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica* **65** 557–586. [MR1445622](#) <https://doi.org/10.2307/2171753>
- [41] STUART, E. A. and JO, B. (2015). Assessing the sensitivity of methods for estimating principal causal effects. *Stat. Methods Med. Res.* **24** 657–674. [MR3428422](#) <https://doi.org/10.1177/0962280211421840>
- [42] SWANSON, S. A. and HERNÁN, M. A. (2014). Think globally, act globally: An epidemiologist’s perspective on instrumental variable estimation [discussion of MR3264545]. *Statist. Sci.* **29** 371–374. [MR3264549](#) <https://doi.org/10.1214/14-STS491>
- [43] TAN, Z. (2010). Marginal and nested structural models using instrumental variables. *J. Amer. Statist. Assoc.* **105** 157–169. [MR2757199](#) <https://doi.org/10.1198/jasa.2009.tm08299>
- [44] TCHETGEN TCHETGEN, E. J. and VANSTEELENDT, S. (2013). Alternative identification and inference for the effect of treatment on the treated with an instrumental variable. In *Harvard University Biostatistics Working Paper Series, Paper 166*.
- [45] TSIATIS, A. A. (2006). *Semiparametric Theory and Missing Data. Springer Series in Statistics*. Springer, New York. [MR2233926](#)
- [46] VAN DER LAAN, M. J. and LUEDTKE, A. R. (2013). Targeted learning of an optimal dynamic treatment, and statistical inference for its mean outcome. *UC Berkeley Division of Biostatistics Working Paper Series* **329** 1–96.
- [47] VAN DER LAAN, M. J. and ROBINS, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer, New York.
- [48] VAN DER VAART, A. (2002). Semiparametric statistics. In *Lectures on Probability Theory and Statistics (Saint-Flour, 1999). Lecture Notes in Math.* **1781** 331–457. Springer, Berlin. [MR1915446](#)
- [49] WRIGHT, S. (1934). The method of path coefficients. *Ann. Math. Stat.* **5** 161–215.
- [50] ZHENG, W. and VAN DER LAAN, M. J. (2010). Asymptotic theory for cross-validated targeted maximum likelihood estimation. *UC Berkeley Division of Biostatistics Working Paper Series* **273** 1–58.

EMPIRICAL RISK MINIMIZATION AND COMPLEXITY OF DYNAMICAL MODELS

BY KEVIN MCGOFF¹ AND ANDREW B. NOBEL²

¹*Department of Mathematics and Statistics, University of North Carolina at Charlotte, kmcgoff1@unc.cc.edu*

²*Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, nobel@email.unc.edu*

A dynamical model consists of a continuous self-map $T : \mathcal{X} \rightarrow \mathcal{X}$ of a compact state space $\mathcal{X}$ and a continuous observation function $f : \mathcal{X} \rightarrow \mathbb{R}$. This paper considers the fitting of a parametrized family of dynamical models to an observed real-valued stochastic process using empirical risk minimization. The limiting behavior of the minimum risk parameters is studied in a general setting. We establish a general convergence theorem for minimum risk estimators and ergodic observations. We then study conditions under which empirical risk minimization can effectively separate signal from noise in an additive observational noise model. The key condition in the latter results is that the family of dynamical models has limited complexity, which is quantified through a notion of entropy for families of infinite sequences that connects covering number based entropies with topological entropy studied in dynamical systems. We establish close connections between entropy and limiting average mean widths for stationary processes, and discuss several examples of dynamical models.

REFERENCES

- [1] ADAMS, T. M. and NOBEL, A. B. (2010). Uniform convergence of Vapnik–Chervonenkis classes under ergodic sampling. *Ann. Probab.* **38** 1345–1367. [MR2663629](#) <https://doi.org/10.1214/09-AOP511>
- [2] ADLER, R. L., KONHEIM, A. G. and MCANDREW, M. H. (1965). Topological entropy. *Trans. Amer. Math. Soc.* **114** 309–319. [MR0175106](#) <https://doi.org/10.2307/1994177>
- [3] AGARWAL, A. and DUCHI, J. C. (2013). The generalization ability of online algorithms for dependent data. *IEEE Trans. Inform. Theory* **59** 573–587. [MR3008166](#) <https://doi.org/10.1109/TIT.2012.2212414>
- [4] ALQUIER, P. and WINTENBERGER, O. (2012). Model selection for weakly dependent time series forecasting. *Bernoulli* **18** 883–913. [MR2948906](#) <https://doi.org/10.3150/11-BEJ359>
- [5] ARCONES, M. A. and YU, B. (1994). Central limit theorems for empirical and U -processes of stationary mixing sequences. *J. Theoret. Probab.* **7** 47–71. [MR1256391](#) <https://doi.org/10.1007/BF02213360>
- [6] BARREIRA, L. and PESIN, Y. (2007). *Nonuniform Hyperbolicity: Dynamics of Systems with Nonzero Lyapunov Exponents. Encyclopedia of Mathematics and Its Applications* **115**. Cambridge Univ. Press, Cambridge. [MR2348606](#) <https://doi.org/10.1017/CBO9781107326026>
- [7] BERTSEKAS, D. P. and SHREVE, S. E. (1996). *Stochastic Optimal Control: The Discrete-Time Case*. Athena Scientific.
- [8] BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford Univ. Press, Oxford. With a foreword by Michel Ledoux. [MR3185193](#) <https://doi.org/10.1093/acprof:oso/9780199535255.001.0001>
- [9] BOWEN, R. and CHAZOTTES, J.-R. (1975). *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms*, Vol. 470. Springer.
- [10] BRACKLEY, C. A., EBENHÖH, O., GREBOGI, C., KURTHS, J., DE MOURA, A., ROMANO, M. C. and THIEL, M. (2010). Introduction to focus issue: Dynamics in systems biology. *Chaos* **20** 045101. [https://doi.org/10.1063/1.3530126](#)
- [11] BRIN, M. and STUCK, G. (2002). *Introduction to Dynamical Systems*. Cambridge Univ. Press, Cambridge. [MR1963683](#) <https://doi.org/10.1017/CBO9780511755316>
- [12] DAMANIK, D., EMBREE, M. and GORODETSKI, A. (2015). Spectral properties of Schrödinger operators arising in the study of quasicrystals. In *Mathematics of Aperiodic Order. Progr. Math.* **309** 307–370. Birkhäuser/Springer, Basel. [MR3381485](#) https://doi.org/10.1007/978-3-0348-0903-0_9

- [13] DAUBE, C. and MCGOFF, K. (2018). On the dynamics of gene regulatory networks with sparsity constraints. Unpublished manuscript.
- [14] DE LA RUE, T. (2006). An introduction to joinings in ergodic theory. *Discrete Contin. Dyn. Syst.* **15** 121–142. [MR2191388](#) <https://doi.org/10.3934/dcds.2006.15.121>
- [15] DEAN, S., MANIA, H., MATNI, N., RECHT, B. and TU, S. (2017). On the sample complexity of the linear quadratic regulator. arXiv preprint, [arXiv:1710.01688](#).
- [16] DENKER, M., GRILLENBERGER, C. and SIGMUND, K. (1976). *Ergodic Theory on Compact Spaces. Lecture Notes in Mathematics* **527**. Springer, Berlin. [MR0457675](#)
- [17] FURSTENBERG, H. (1967). Disjointness in ergodic theory, minimal sets, and a problem in Diophantine approximation. *Math. Syst. Theory* **1** 1–49. [MR0213508](#) <https://doi.org/10.1007/BF01692494>
- [18] GLASNER, E. (2003). *Ergodic Theory via Joinings. Mathematical Surveys and Monographs* **101**. Amer. Math. Soc., Providence, RI. [MR1958753](#) <https://doi.org/10.1090/surv/101>
- [19] GRAY, R. M., NEUHOFF, D. L. and SHIELDS, P. C. (1975). A generalization of Ornstein’s $\bar{d}$ distance with applications to information theory. *Ann. Probab.* **3** 315–328. [MR0368127](#) <https://doi.org/10.1214/aop/1176996402>
- [20] HANG, H. and STEINWART, I. (2014). Fast learning from α -mixing observations. *J. Multivariate Anal.* **127** 184–199. [MR3188886](#) <https://doi.org/10.1016/j.jmva.2014.02.012>
- [21] HANG, H. and STEINWART, I. (2017). A Bernstein-type inequality for some mixing processes and dynamical systems with an application to learning. *Ann. Statist.* **45** 708–743. [MR3650398](#) <https://doi.org/10.1214/16-AOS1465>
- [22] HANG, H., STEINWART, I., FENG, Y. and SUYKENS, J. A. K. (2018). Kernel density estimation for dynamical systems. *J. Mach. Learn. Res.* **19** Paper No. 35, 49. [MR3862442](#)
- [23] HASSELL, M. P., LAWTON, J. H. and MAY, R. M. (1976). Patterns of dynamical behaviour in single-species populations. *J. Anim. Ecol.* 471–486.
- [24] HOFBAUER, F. (1980). The topological entropy of the transformation $x \mapsto ax(1 - x)$. *Monatsh. Math.* **90** 117–141. [MR0595319](#) <https://doi.org/10.1007/BF01303262>
- [25] WONG, K. C., LI, Z. and TEWARI, A. (2016). Regularized estimation in high dimensional time series under mixing conditions. arXiv preprint, [arXiv:1602.04265](#).
- [26] KATOK, A. (2007). Fifty years of entropy in dynamics: 1958–2007. *J. Mod. Dyn.* **1** 545–596. [MR2342699](#) <https://doi.org/10.3934/jmd.2007.1.545>
- [27] KATOK, A. and HASSELBLATT, B. (1995). *Introduction to the Modern Theory of Dynamical Systems. Encyclopedia of Mathematics and Its Applications* **54**. Cambridge Univ. Press, Cambridge. With a supplementary chapter by Katok and Leonardo Mendoza. [MR1326374](#) <https://doi.org/10.1017/CBO9780511809187>
- [28] KONTOROVICH, L. (2012). Statistical estimation with bounded memory. *Stat. Comput.* **22** 1155–1164. [MR2950092](#) <https://doi.org/10.1007/s11222-011-9293-5>
- [29] KUZNETSOV, V. and MOHRI, M. (2015). Learning theory and algorithms for forecasting non-stationary time series. In *Advances in Neural Information Processing Systems* 541–549.
- [30] KUZNETSOV, V. and MOHRI, M. (2016). Time series prediction and online learning. In *29th Annual Conference on Learning Theory* 1190–1213.
- [31] KUZNETSOV, V. and MOHRI, M. (2017). Generalization bounds for non-stationary mixing processes. *Mach. Learn.* **106** 93–117. [MR3592151](#) <https://doi.org/10.1007/s10994-016-5588-2>
- [32] LALLEY, S. P. (1999). Beneath the noise, chaos. *Ann. Statist.* **27** 461–479. [MR1714721](#) <https://doi.org/10.1214/aos/1018031203>
- [33] LALLEY, S. P. and NOBEL, A. B. (2006). Denoising deterministic time series. *Dyn. Partial Differ. Equ.* **3** 259–279. [MR2271730](#) <https://doi.org/10.4310/DPDE.2006.v3.n4.a1>
- [34] LEDOUX, M. and TALAGRAND, M. (2011). *Probability in Banach Spaces: Isoperimetry and Processes. Classics in Mathematics*. Springer, Berlin. Reprint of the 1991 edition. [MR2814399](#)
- [35] LETHAM, B., LETHAM, P. A., RUDIN, C. and BROWNE, E. P. (2016). Prediction uncertainty and optimal experimental design for learning dynamical systems. *Chaos* **26** 063110, 12. [MR3511270](#) <https://doi.org/10.1063/1.4953795>
- [36] LEV, N., PELED, R. and PERES, Y. (2015). Separating signal from noise. *Proc. Lond. Math. Soc.* (3) **110** 883–931. [MR3335290](#) <https://doi.org/10.1112/plms/pdu057>
- [37] LEVIN, S. A., CARPENTER, S. R., GODFRAY, H. C. J., KINZIG, A. P., LOREAU, M., LOSOS, J. B., WALKER, B. and WILCOVE, D. S. (2009). *The Princeton Guide to Ecology*. Princeton Univ. Press.
- [38] LIND, D. and MARCUS, B. (1995). *An Introduction to Symbolic Dynamics and Coding*. Cambridge Univ. Press, Cambridge. [MR1369092](#) <https://doi.org/10.1017/CBO9780511626302>
- [39] LIVERANI, C. (1996). Central limit theorem for deterministic systems. In *International Conference on Dynamical Systems (Montevideo, 1995)*. *Pitman Res. Notes Math. Ser.* **362** 56–75. Longman, Harlow. [MR1460797](#)

- [40] MALLET-PARET, J. and SMITH, H. L. (1990). The Poincaré–Bendixson theorem for monotone cyclic feedback systems. *J. Dynam. Differential Equations* **2** 367–421. [MR1073471](#) <https://doi.org/10.1007/BF01054041>
- [41] MARBACH, D., COSTELLO, J. C., KÜFFNER, R., VEGA, N. M., PRILL, R. J., CAMACHO, D. M., ALLISON, K. R., DREAM5 CONSORTIUM, KELLIS, M. et al. (2012). Wisdom of crowds for robust gene network inference. *Nat. Methods* **9** 796–804. <https://doi.org/10.1038/nmeth.2016>
- [42] MASUR, H. and TABACHNIKOV, S. (2002). Rational billiards and flat structures. In *Handbook of Dynamical Systems*, Vol. 1A 1015–1089. North-Holland, Amsterdam. [MR1928530](#) [https://doi.org/10.1016/S1874-575X\(02\)80015-7](https://doi.org/10.1016/S1874-575X(02)80015-7)
- [43] MAY, R. M. (1974). Biological populations with nonoverlapping generations: Stable points, stable cycles, and chaos. *Science* **186** 645–647.
- [44] MCGOFF, K., MUKHERJEE, S., NOBEL, A. and PILLAI, N. (2015). Consistency of maximum likelihood estimation for some dynamical systems. *Ann. Statist.* **43** 1–29. [MR3285598](#) <https://doi.org/10.1214/14-AOS1259>
- [45] MCGOFF, K., MUKHERJEE, S. and PILLAI, N. (2015). Statistical inference for dynamical systems: A review. *Stat. Surv.* **9** 209–252. [MR3422438](#) <https://doi.org/10.1214/15-SS111>
- [46] MCGOFF, K. and NOBEL, A. (2016). Optimal tracking for dynamical systems. [arXiv:1601.05033](#).
- [47] MCGOFF, K. and NOBEL, A. B. (2020). Supplement to “Empirical risk minimization and complexity of dynamical models.” <https://doi.org/10.1214/19-AOS1876SUPP>.
- [48] MCGOFF, K. A., GUO, X., DECKARD, A., KELLIHER, C. M., LEMAN, A. R., FRANCEY, L. J., HOGENESCH, J. B., HAASE, S. B. and HARER, J. L. (2016). The local edge machine: Inference of dynamic models of gene regulation. *Genome Biol.* **17** 214. <https://doi.org/10.1186/s13059-016-1076-z>
- [49] MEIR, R. (2000). Nonparametric time series prediction through adaptive model selection. *Mach. Learn.* **39** 5–34.
- [50] MISCHAIKOW, K., MROZEK, M. and SZYMCAK, A. (2001). Chaos in the Lorenz equations: A computer assisted proof. III. Classical parameter values *J. Differential Equations* **169** 17–56. Special issue in celebration of Jack K. Hale’s 70th birthday, Part 3 (Atlanta, GA/Lisbon, 1998). [MR1808460](#) <https://doi.org/10.1006/jdeq.2000.3894>
- [51] MODHA, D. S. and MASRY, E. (1998). Memory-universal prediction of stationary random processes. *IEEE Trans. Inform. Theory* **44** 117–133. [MR1486652](#) <https://doi.org/10.1109/18.650998>
- [52] MOHRI, M. and ROSTAMIZADEH, A. (2010). Stability bounds for stationary ϕ -mixing and β -mixing processes. *J. Mach. Learn. Res.* **11** 789–814. [MR2600630](#)
- [53] MORVAI, G. and WEISS, B. (2005). On classifying processes. *Bernoulli* **11** 523–532. [MR2146893](#) <https://doi.org/10.3150/bj/1120591187>
- [54] MORVAI, G. and WEISS, B. (2005). Prediction for discrete time series. *Probab. Theory Related Fields* **132** 1–12. [MR2136864](#) <https://doi.org/10.1007/s00440-004-0386-3>
- [55] NOBEL, A. B. (2006). Hypothesis testing for families of ergodic processes. *Bernoulli* **12** 251–269. [MR2218555](#) <https://doi.org/10.3150/bj/1145993974>
- [56] ORNSTEIN, D. (1970). Bernoulli shifts with the same entropy are isomorphic. *Adv. Math.* **4** 337–352. [MR0257322](#) [https://doi.org/10.1016/0001-8708\(70\)90029-0](https://doi.org/10.1016/0001-8708(70)90029-0)
- [57] ORNSTEIN, D. S. (1973). An application of ergodic theory to probability theory. *Ann. Probab.* **1** 43–65. [MR0348831](#) <https://doi.org/10.1098/rsnr.1972.0008>
- [58] ORNSTEIN, D. S. (1974). *Ergodic Theory, Randomness, and Dynamical Systems*. Yale Mathematical Monographs **5**. Yale Univ. Press, New Haven, Conn.–London. James K. Whittemore Lectures in Mathematics given at Yale University. [MR0447525](#)
- [59] ORNSTEIN, D. S. and WEISS, B. (1990). How sampling reveals a process. *Ann. Probab.* **18** 905–930. [MR1062052](#)
- [60] POLLARD, D. and RADCHENKO, P. (2006). Nonlinear least-squares estimation. *J. Multivariate Anal.* **97** 548–562. [MR2234037](#) <https://doi.org/10.1016/j.jmva.2005.04.002>
- [61] QUEFFÉLEC, M. (2010). *Substitution Dynamical Systems—Spectral Analysis*, 2nd ed. *Lecture Notes in Math.* **1294**. Springer, Berlin. [MR2590264](#) <https://doi.org/10.1007/978-3-642-11212-6>
- [62] RAKHLIN, A. and SRIDHARAN, K. (2017). On equivalence of martingale tail bounds and deterministic regret inequalities. In *Proceedings of the 2017 Conference on Learning Theory. Proceedings of Machine Learning Research* **65** (S. Kale and O. Shamir, eds.) 1704–1722. PMLR, Amsterdam, Netherlands.
- [63] RAKHLIN, A., SRIDHARAN, K. and TEWARI, A. (2015). Sequential complexities and uniform martingale laws of large numbers. *Probab. Theory Related Fields* **161** 111–153. [MR3304748](#) <https://doi.org/10.1007/s00440-013-0545-5>
- [64] ROBINSON, E. A. JR. (1996). The dynamical theory of tilings and quasicrystallography. In *Ergodic Theory of $\mathbb{Z}^d$ Actions* (Warwick, 1993–1994). *London Mathematical Society Lecture Note Series* **228** 451–473. Cambridge Univ. Press, Cambridge. [MR1411233](#) <https://doi.org/10.1017/CBO9780511662812.020>

- [65] RUELLE, D. (2004). *Thermodynamic Formalism: The Mathematical Structures of Equilibrium Statistical Mechanics*, 2nd ed. *Cambridge Mathematical Library*. Cambridge Univ. Press, Cambridge. [MR2129258](#) <https://doi.org/10.1017/CBO9780511617546>
- [66] SHALIZI, C. and KONTOROVICH, A. (2013). Predictive PAC learning and process decompositions. In *Advances in Neural Information Processing Systems* 1619–1627.
- [67] SHECHTMAN, D., BLECH, I., GRATIAS, D. and CAHN, J. W. (1984). Metallic phase with long-range orientational order and no translational symmetry. *Phys. Rev. Lett.* **53** 1951.
- [68] SOLOMYAK, B. (1998). Spectrum of dynamical systems arising from Delone sets. In *Quasicrystals and Discrete Geometry (Toronto, ON, 1995)*. *Fields Inst. Monogr.* **10** 265–275. Amer. Math. Soc., Providence, RI. [MR1636783](#)
- [69] STEINWART, I. and ANGHEL, M. (2009). Consistency of support vector machines for forecasting the evolution of an unknown ergodic dynamical system from observations with unknown noise. *Ann. Statist.* **37** 841–875. [MR2502653](#) <https://doi.org/10.1214/07-AOS562>
- [70] TU, S., BO CZAR, R., PACKARD, A. and RECHT, B. (2017). Non-asymptotic analysis of robust control from coarse-grained identification. arXiv preprint, [arXiv:1707.04791](#).
- [71] TURCHIN, P. (2003). *Complex Population Dynamics: A Theoretical/Empirical Synthesis*. *Monographs in Population Biology* **35**. Princeton Univ. Press, Princeton, NJ. [MR2080584](#)
- [72] VAN DE GEER, S. (1990). Estimating a regression function. *Ann. Statist.* **18** 907–924. [MR1056343](#) <https://doi.org/10.1214/aos/1176347632>
- [73] 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>
- [74] VISWANATH, D., LIANG, X. and SERKH, K. (2013). Metric entropy and the optimal prediction of chaotic signals. *SIAM J. Appl. Dyn. Syst.* **12** 1085–1113. [MR3071149](#) <https://doi.org/10.1137/110824772>
- [75] WALTERS, P. (1982). *An Introduction to Ergodic Theory*. *Graduate Texts in Mathematics* **79**. Springer, New York-Berlin. [MR0648108](#)
- [76] WU, C.-F. (1981). Asymptotic theory of nonlinear least squares estimation. *Ann. Statist.* **9** 501–513. [MR0615427](#)
- [77] YU, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences. *Ann. Probab.* **22** 94–116. [MR1258867](#)
- [78] ZIMIN, A. and LAMPERT, C. (2017). Learning theory for conditional risk minimization. In *Artificial Intelligence and Statistics* 213–222.

ADAPTIVE ESTIMATION IN STRUCTURED FACTOR MODELS WITH APPLICATIONS TO OVERLAPPING CLUSTERING

BY XIN BING^{1,*}, FLORENTINA BUNEA^{1,**}, YANG NING^{1,†} AND MARTEN WEGKAMP²

¹Department of Statistics and Data Science, Cornell University, *xb43@cornell.edu; **fb238@cornell.edu;

†yn265@cornell.edu

²Department of Mathematics & Department of Statistics and Data Science, Cornell University, mhw73@cornell.edu

This work introduces a novel estimation method, called LOVE, of the entries and structure of a loading matrix A in a latent factor model $X = AZ + E$, for an observable random vector $X \in \mathbb{R}^p$, with correlated unobservable factors $Z \in \mathbb{R}^K$, with K unknown, and uncorrelated noise E . Each row of A is scaled, and allowed to be sparse. In order to identify the loading matrix A , we require the existence of pure variables, which are components of X that are associated, via A , with one and only one latent factor. Despite the fact that the number of factors K , the number of the pure variables and their location are all unknown, we only require a mild condition on the covariance matrix of Z , and a minimum of only two pure variables per latent factor to show that A is uniquely defined, up to signed permutations. Our proofs for model identifiability are constructive, and lead to our novel estimation method of the number of factors and of the set of pure variables, from a sample of size n of observations on X . This is the first step of our LOVE algorithm, which is optimization-free, and has low computational complexity of order p^2 . The second step of LOVE is an easily implementable linear program that estimates A . We prove that the resulting estimator is near minimax rate optimal for A , with respect to the $\|\cdot\|_{\infty,q}$ loss, for $q \geq 1$, up to logarithmic factors in p , and that it can be minimax-rate optimal in many cases of interest.

The model structure is motivated by the problem of overlapping variable clustering, ubiquitous in data science. We define the population level clusters as groups of those components of X that are associated, via the matrix A , with the same unobservable latent factor, and multifactor association is allowed. Clusters are respectively anchored by the pure variables, and form overlapping subgroups of the p -dimensional random vector X . The Latent model approach to OVERlapping clustering is reflected in the name of our algorithm, LOVE.

The third step of LOVE estimates the clusters from the support of the columns of the estimated A . We guarantee cluster recovery with zero false positive proportion, and with false negative proportion control. The practical relevance of LOVE is illustrated through the analysis of a RNA-seq data set, devoted to determining the functional annotation of genes with unknown function.

REFERENCES

- ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley Series in Probability and Statistics. Wiley Interscience, Hoboken, NJ. [MR1990662](#)
- ANDERSON, T. W. and AMEMIYA, Y. (1988). The asymptotic normal distribution of estimators in factor analysis under general conditions. *Ann. Statist.* **16** 759–771. [MR0947576](#) <https://doi.org/10.1214/aos/1176350834>
- ANDERSON, T. W. and RUBIN, H. (1956). Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, Vol. V* 111–150. Univ. California Press, Berkeley and Los Angeles. [MR0084943](#)

MSC2020 subject classifications. Primary 62H25, 62H30.

Key words and phrases. Overlapping clustering, latent model, identification, high-dimensional estimation, minimax estimation, pure variables, group recovery, support recovery, sparse loading matrix, matrix factorization, adaptive estimation.

- ARORA, S., GE, R., HALPERN, Y., MIMNO, D. M., MOITRA, A., SONTAG, D., WU, Y. and ZHU, M. (2013). A practical algorithm for topic modeling with provable guarantees. In *ICML* (2) 280–288.
- ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S. et al. (2000). Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25** 25–29.
- BAI, J. and LI, K. (2012). Statistical analysis of factor models of high dimension. *Ann. Statist.* **40** 436–465. [MR3014313 https://doi.org/10.1214/11-AOS966](https://doi.org/10.1214/11-AOS966)
- BAI, J. and NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221. [MR1926259 https://doi.org/10.1111/1468-0262.00273](https://doi.org/10.1111/1468-0262.00273)
- BEKKER, P. A. and TEN BERGE, J. M. F. (1997). Generic global identification in factor analysis. *Linear Algebra Appl.* **264** 255–263. [MR1465870 https://doi.org/10.1016/S0024-3795\(96\)00363-1](https://doi.org/10.1016/S0024-3795(96)00363-1)
- BELLONI, A., ROSENBAUM, M. and TSYBAKOV, A. B. (2017). Linear and conic programming estimators in high dimensional errors-in-variables models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 939–956. [MR3641415 https://doi.org/10.1111/rssb.12196](https://doi.org/10.1111/rssb.12196)
- BEST, M. G., SOL, N., KOOI, I., TANNOUN, J., WESTERMAN, B. A., RUSTENBURG, F., SCHELLEN, P., VERSCHUEREN, H., POST, E. et al. (2015). RNA-seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics. *Cancer Cell* **28** 666–676.
- BEZDEK, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York. With a foreword by L. A. Zadeh, Advanced Applications in Pattern Recognition. [MR0631231 https://doi.org/10.1007/978-1-4613-1264-1](https://doi.org/10.1007/978-1-4613-1264-1)
- BHATTACHARYA, A. and DUNSON, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98** 291–306. [MR2806429 https://doi.org/10.1093/biomet/asr013](https://doi.org/10.1093/biomet/asr013)
- BIEN, J., BUNEA, F. and XIAO, L. (2016). Convex banding of the covariance matrix. *J. Amer. Statist. Assoc.* **111** 834–845. [MR3538709 https://doi.org/10.1080/01621459.2015.1058265](https://doi.org/10.1080/01621459.2015.1058265)
- BING, X., BUNEA, F. and WEGKAMP, M. H. (2018). A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics. ArXiv E-prints [arXiv:1805.06837](https://arxiv.org/abs/1805.06837).
- BING, X. and WEGKAMP, M. H. (2018). Adaptive estimation of the rank of the coefficient matrix in high dimensional multivariate response regression models. [arXiv:1704.02381](https://arxiv.org/abs/1704.02381).
- BING, X., BUNEA, F., NING, Y. and WEGKAMP, M. (2020). Supplement to “Adaptive estimation in structured factor models with applications to overlapping clustering.” <https://doi.org/10.1214/19-AOS1877SUPP>.
- BITTORF, V., RECHT, B., RE, C. and TROPP, J. A. (2012). Factoring nonnegative matrices with linear programs. [arXiv:1206.1270](https://arxiv.org/abs/1206.1270).
- BOLLEN, K. A. (1989). *Structural Equations with Latent Variables*. *Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. Wiley, New York. A Wiley-Interscience Publication. [MR0996025 https://doi.org/10.1002/9781118619179](https://doi.org/10.1002/9781118619179)
- BUNEA, F., GIRAUD, C. and LUO, X. (2016). Minimax optimal variable clustering in G-models via cord. ArXiv Preprint [arXiv:1508.01939](https://arxiv.org/abs/1508.01939).
- BUNEA, F., SHE, Y. and WEGKAMP, M. H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann. Statist.* **39** 1282–1309. [MR2816355 https://doi.org/10.1214/11-AOS876](https://doi.org/10.1214/11-AOS876)
- BUNEA, F., GIRAUD, C., ROYER, M. and VERZELEN, N. (2016). PECOK: A convex optimization approach to variable clustering. ArXiv Preprint [arXiv:1606.05100](https://arxiv.org/abs/1606.05100).
- BUNEA, F., GIRAUD, C., LUO, X., ROYER, M. and VERZELEN, N. (2018). Model assisted variable clustering: Minimax-optimal recovery and algorithms. ArXiv E-prints [arXiv:1508.01939](https://arxiv.org/abs/1508.01939). *Ann. Statist.* To appear.
- CAI, T., LIU, W. and LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. [MR2847973 https://doi.org/10.1198/jasa.2011.tm10155](https://doi.org/10.1198/jasa.2011.tm10155)
- CAI, T. T., LIU, W. and ZHOU, H. H. (2016). Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *Ann. Statist.* **44** 455–488. [MR3476606 https://doi.org/10.1214/13-AOS1171](https://doi.org/10.1214/13-AOS1171)
- CANDÈS, E. J., LI, X., MA, Y. and WRIGHT, J. (2011). Robust principal component analysis? *J. ACM* **58** Art. 11, 37. [MR2811000 https://doi.org/10.1145/1970392.1970395](https://doi.org/10.1145/1970392.1970395)
- CARVALHO, C. M., CHANG, J., LUCAS, J. E., NEVINS, J. R., WANG, Q. and WEST, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *J. Amer. Statist. Assoc.* **103** 1438–1456. [MR2655722 https://doi.org/10.1198/016214508000000869](https://doi.org/10.1198/016214508000000869)
- CHANDRASEKARAN, V., PARRILO, P. A. and WILLSKY, A. S. (2012). Latent variable graphical model selection via convex optimization. *Ann. Statist.* **40** 1935–1967. [MR3059067 https://doi.org/10.1214/11-AOS949](https://doi.org/10.1214/11-AOS949)
- CHANDRASEKARAN, V., SANGHAVI, S., PARRILO, P. A. and WILLSKY, A. S. (2011). Rank-sparsity incoherence for matrix decomposition. *SIAM J. Optim.* **21** 572–596. [MR2817479 https://doi.org/10.1137/090761793](https://doi.org/10.1137/090761793)
- CRADDOCK, R. C., JAMES, G. A., HOLTZHEIMER, P. E., HU, X. P. and MAYBERG, H. S. (2012). A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Hum. Brain Mapp.* **33** 1914–1928.
- CRADDOCK, R. C., JBABDI, S., YAN, C.-G., VOGELSTEIN, J. T., CASTELLANOS, F. X., DI MARTINO, A., KELLY, C., HEBERLEIN, K., COLCOMBE, S. et al. (2013). Imaging human connectomes at the macroscale. *Nat. Methods* **10** 524–539.

- DONOHO, D. and STODDEN, V. (2004). When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems* 16 (S. Thrun, L. K. Saul and P. B. Schölkopf, eds.) 1141–1148. MIT Press.
- EVERITT, B. S. (1984). *An Introduction to Latent Variable Models. Monographs on Statistics and Applied Probability*. CRC Press, London; distributed by Methuen, Inc., New York. MR0769300 <https://doi.org/10.1007/978-94-009-5564-6>
- FAN, J., LIAO, Y. and MINCHEVA, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 603–680. With 33 discussions by 57 authors and a reply by Fan, Liao and Mincheva. MR3091653 <https://doi.org/10.1111/rssb.12016>
- FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.
- GAUTIER, E. and TSYBAKOV, A. B. (2011). High-dimensional instrumental variables regression and confidence sets. ArXiv Preprint [arXiv:1105.2454v4](https://arxiv.org/abs/1105.2454v4).
- GEWEKE, J. and ZHOU, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *Rev. Financ. Stud.* **9** 557–587.
- HSU, D., KAKADE, S. M. and ZHANG, T. (2011). Robust matrix decomposition with sparse corruptions. *IEEE Trans. Inform. Theory* **57** 7221–7234. MR2883652 <https://doi.org/10.1109/TIT.2011.2158250>
- IZENMAN, A. J. (2008). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning. Springer Texts in Statistics*. Springer, New York. MR2445017 <https://doi.org/10.1007/978-0-387-78189-1>
- JIANG, D., TANG, C. and ZHANG, A. (2004). Cluster analysis for gene expression data: A survey. *IEEE Trans. Knowl. Data Eng.* **16** 1370–1386.
- KOOPMANS, T. C. and REIERSØL, O. (1950). The identification of structural characteristics. *Ann. Math. Stat.* **21** 165–181. MR0039967 <https://doi.org/10.1214/aoms/1177729837>
- KRISHNAPURAM, R., JOSHI, A., NASRAOUI, O. and YI, L. (2001). Low-complexity fuzzy relational clustering algorithms for web mining. *IEEE Trans. Fuzzy Syst.* **9** 595–607.
- LAWLEY, D. N. and MAXWELL, A. E. (1971). *Factor Analysis as a Statistical Method*, 2nd ed. American Elsevier Publishing Co., Inc., New York. MR0343471
- LEDERMANN, W. (1937). On the rank of the reduced correlational matrix in multiple-factor analysis. *Psychometrika* **2** 85–93.
- MCDONALD, R. P. (1999). *Test Theory: A Unified Treatment*. Taylor and Francis.
- 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>
- RUBIN, D. B. and THAYER, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika* **47** 69–76. MR0668505 <https://doi.org/10.1007/BF02293851>
- SHAPIRO, A. (1982). Rank-reducibility of a symmetric matrix and sampling theory of minimum trace factor analysis. *Psychometrika* **47** 187–199. MR0667012 <https://doi.org/10.1007/BF02296274>
- SHAPIRO, A. (1985). Identifiability of factor analysis: Some results and open problems. *Linear Algebra Appl.* **70** 1–7. MR0808527 [https://doi.org/10.1016/0024-3795\(85\)90038-2](https://doi.org/10.1016/0024-3795(85)90038-2)
- WEGKAMP, M. and ZHAO, Y. (2016). Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. *Bernoulli* **22** 1184–1226. MR3449812 <https://doi.org/10.3150/14-BEJ690>
- WIWIE, C., BAUMBACH, J. and RÖTTGER, R. (2015). Comparing the performance of biomedical clustering methods. *Nat. Methods* **12** 1033–1038.
- YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824 <https://doi.org/10.1093/biomet/asm018>

PARTIAL IDENTIFIABILITY OF RESTRICTED LATENT CLASS MODELS

BY YUQI GU^{*} AND GONGJUN XU^{**}

Department of Statistics, University of Michigan, ^{*}yuqigu@umich.edu; ^{**}gongjun@umich.edu

Latent class models have wide applications in social and biological sciences. In many applications, prespecified restrictions are imposed on the parameter space of latent class models, through a design matrix, to reflect practitioners' assumptions about how the observed responses depend on subjects' latent traits. Though widely used in various fields, such restricted latent class models suffer from nonidentifiability due to their discreteness nature and complex structure of restrictions. This work addresses the fundamental identifiability issue of restricted latent class models by developing a general framework for strict and partial identifiability of the model parameters. Under correct model specification, the developed identifiability conditions only depend on the design matrix and are easily checkable, which provide useful practical guidelines for designing statistically valid diagnostic tests. Furthermore, the new theoretical framework is applied to establish, for the first time, identifiability of several designs from cognitive diagnosis applications.

REFERENCES

- [1] AGRESTI, A. (2002). *Categorical Data Analysis*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, New York. [MR1914507](#) <https://doi.org/10.1002/0471249688>
- [2] ALLMAN, E. S., MATIAS, C. and RHODES, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.* **37** 3099–3132. [MR2549554](#) <https://doi.org/10.1214/09-AOS689>
- [3] CASELLA, G. and BERGER, R. L. (1990). *Statistical Inference. The Wadsworth & Brooks/Cole Statistics/Probability Series*. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA. [MR1051420](#)
- [4] CHEN, J. and DE LA TORRE, J. (2018). Introducing the general polytomous diagnosis modeling framework. *Front. Psychol.* **9** 1474. <https://doi.org/10.3389/fpsyg.2018.01474>
- [5] CHEN, Y., CULPEPPER, S. A., CHEN, Y. and DOUGLAS, J. (2018). Bayesian estimation of the DINA Q matrix. *Psychometrika* **83** 89–108. [MR3767014](#) <https://doi.org/10.1007/s11336-017-9579-4>
- [6] CHEN, Y., CULPEPPER, S. A., WANG, S. and DOUGLAS, J. (2018). A hidden Markov model for learning trajectories in cognitive diagnosis with application to spatial rotation skills. *Appl. Psychol. Meas.* **42** 5–23.
- [7] CHEN, Y., LIU, J., XU, G. and YING, Z. (2015). Statistical analysis of Q -matrix based diagnostic classification models. *J. Amer. Statist. Assoc.* **110** 850–866. [MR3367269](#) <https://doi.org/10.1080/01621459.2014.934827>
- [8] CHIU, C.-Y., DOUGLAS, J. A. and LI, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika* **74** 633–665. [MR2565331](#) <https://doi.org/10.1007/s11336-009-9125-0>
- [9] CHOI, K. M., LEE, Y.-S. and PARK, Y. S. (2015). What CDM can tell about what students have learned: An analysis of TIMSS eighth grade mathematics. *Eurasia J. Math., Sci. Technol. Educ.* **11** 1563–1577.
- [10] CULPEPPER, S. A. (2015). Bayesian estimation of the DINA model with Gibbs sampling. *J. Educ. Behav. Stat.* **40** 454–476.
- [11] DE LA TORRE, J. (2011). The generalized DINA model framework. *Psychometrika* **76** 179–199. [MR2788881](#) <https://doi.org/10.1007/s11336-011-9207-7>
- [12] DE LA TORRE, J. and DOUGLAS, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika* **69** 333–353. [MR2272454](#) <https://doi.org/10.1007/BF02295640>
- [13] DE LA TORRE, J., VAN DER ARK, L. A. and ROSSI, G. (2018). Analysis of clinical data from a cognitive diagnosis modeling framework. *Meas. Eval. Couns. Dev.* **51** 281–296.

- [14] DECARLO, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q -matrix. *Appl. Psychol. Meas.* **35** 8–26.
- [15] DiBELLO, L. V., STOUT, W. F. and ROUSSOS, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. *Cogn. Diagn. Assess.* 361–389.
- [16] ELMORE, R., HALL, P. and NEEMAN, A. (2005). An application of classical invariant theory to identifiability in nonparametric mixtures. *Ann. Inst. Fourier (Grenoble)* **55** 1–28. [MR2141286](#)
- [17] GOODFELLOW, I., BENGIO, Y. and COURVILLE, A. (2016). *Deep Learning. Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. [MR3617773](#)
- [18] GOODMAN, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61** 215–231. [MR0370936](#) <https://doi.org/10.1093/biomet/61.2.215>
- [19] GU, Y. and XU, G. (2019). The sufficient and necessary condition for the identifiability and estimability of the DINA model. *Psychometrika* **84** 468–483. [MR3947373](#) <https://doi.org/10.1007/s11336-018-9619-8>
- [20] GU, Y. and XU, G. (2020). Supplement to “Partial identifiability of restricted latent class models.” <https://doi.org/10.1214/19-AOS1878SUPP>.
- [21] GYLLENBERG, M., KOSKI, T., REILINK, E. and VERLAAN, M. (1994). Nonuniqueness in probabilistic numerical identification of bacteria. *J. Appl. Probab.* **31** 542–548. [MR1274807](#) <https://doi.org/10.2307/3215044>
- [22] HENSON, R. A., TEMPLIN, J. L. and WILLSE, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika* **74** 191–210. [MR2507377](#) <https://doi.org/10.1007/s11336-008-9089-5>
- [23] JAEGER, J., TATSUOKA, C., BERNS, S. M. and VARADI, F. (2006). Distinguishing neurocognitive functions in schizophrenia using partially ordered classification models. *Schizophr. Bull.* **32** 679–691.
- [24] JUNKER, B. W. and SIJTSMA, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Meas.* **25** 258–272. [MR1842982](#) <https://doi.org/10.1177/01466210122032064>
- [25] LEE, Y.-S., PARK, Y. S. and TAYLAN, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the US national sample using the TIMSS 2007. *Int. J. Test.* **11** 144–177.
- [26] LEIGHTON, J. P., GIERL, M. J. and HUNKA, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka’s rule-space approach. *J. Educ. Meas.* **41** 205–237.
- [27] MA, W. and DE LA TORRE, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *Br. J. Math. Stat. Psychol.* **69** 253–275. <https://doi.org/10.1111/bmsp.12070>
- [28] MARIS, E. (1999). Estimating multiple classification latent class models. *Psychometrika* **64** 187–212. [MR1700708](#) <https://doi.org/10.1007/BF02294535>
- [29] MARIS, G. and BECHGER, T. M. (2009). Equivalent diagnostic classification models. *Measurement* **7** 41–46.
- [30] MCHUGH, R. B. (1956). Efficient estimation and local identification in latent class analysis. *Psychometrika* **21** 331–347. [MR0082427](#) <https://doi.org/10.1007/BF02296300>
- [31] RUPP, A. A., TEMPLIN, J. and HENSON, R. A. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. Guilford Press.
- [32] TATSUOKA, C. (2009). Diagnostic models as partially ordered sets. *Measurement* **7** 49–53.
- [33] TATSUOKA, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *J. Educ. Meas.* **20** 345–354.
- [34] TEICHER, H. (1967). Identifiability of mixtures of product measures. *Ann. Math. Stat.* **38** 1300–1302. [MR0216635](#) <https://doi.org/10.1214/aoms/1177698805>
- [35] TEMPLIN, J. and BRADSHAW, L. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika* **79** 317–339. [MR3255122](#) <https://doi.org/10.1007/s11336-013-9362-0>
- [36] TEMPLIN, J. L. and HENSON, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychol. Methods* **11** 287–305. <https://doi.org/10.1037/1082-989X.11.3.287>
- [37] 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>
- [38] VON DAVIER, M. (2005). A general diagnostic model applied to language testing data. *ETS Research Report Series* **2005**.
- [39] VON DAVIER, M. (2008). A general diagnostic model applied to language testing data. *Br. J. Math. Stat. Psychol.* **61** 287–307. [MR2649038](#) <https://doi.org/10.1348/000711007X193957>
- [40] VON DAVIER, M. (2014). The log-linear cognitive diagnostic model (LCDM) as a special case of the general diagnostic model (GDM). *ETS Res. Rep. Ser.* **2014** 1–13.

- [41] VON DAVIER, M. (2014). The DINA model as a constrained general diagnostic model: Two variants of a model equivalency. *Br. J. Math. Stat. Psychol.* **67** 49–71. [MR3158657](#) <https://doi.org/10.1111/bmsp.12003>
- [42] VON DAVIER, M. and HABERMAN, S. J. (2014). Hierarchical diagnostic classification models morphing into unidimensional ‘diagnostic’ classification models—a commentary. *Psychometrika* **79** 340–346. [MR3255123](#) <https://doi.org/10.1007/s11336-013-9363-z>
- [43] WANG, S., YANG, Y., CULPEPPER, S. A. and DOUGLAS, J. A. (2018). Tracking skill acquisition with cognitive diagnosis models: A higher-order, hidden Markov model with covariates. *J. Educ. Behav. Stat.* **43** 57–87.
- [44] WU, Z., DELORIA-KNOLL, M. and ZEGER, S. L. (2017). Nested partially latent class models for dependent binary data; estimating disease etiology. *Biostatistics* **18** 200–213. [MR3825117](#) <https://doi.org/10.1093/biostatistics/kxw037>
- [45] XU, G. (2017). Identifiability of restricted latent class models with binary responses. *Ann. Statist.* **45** 675–707. [MR3650397](#) <https://doi.org/10.1214/16-AOS1464>
- [46] XU, G. and SHANG, Z. (2018). Identifying latent structures in restricted latent class models. *J. Amer. Statist. Assoc.* **113** 1284–1295. [MR3862357](#) <https://doi.org/10.1080/01621459.2017.1340889>
- [47] XU, G. and ZHANG, S. (2016). Identifiability of diagnostic classification models. *Psychometrika* **81** 625–649. [MR3535051](#) <https://doi.org/10.1007/s11336-015-9471-z>
- [48] YAMAGUCHI, K. and OKADA, K. (2018). Comparison among cognitive diagnostic models for the TIMSS 2007 fourth grade mathematics assessment. *PLoS ONE* **12** e0188691.
- [49] ZHANG, S. S., DECARLO, L. T. and YING, Z. (2013). Non-identifiability, equivalence classes, and attribute-specific classification in Q -matrix based cognitive diagnosis models. ArXiv Preprint. Available at [arXiv:1303.0426](https://arxiv.org/abs/1303.0426).

POSTERIOR CONCENTRATION FOR BAYESIAN REGRESSION TREES AND FORESTS

BY VERONIKA ROČKOVÁ¹ AND STÉPHANIE VAN DER PAS²

¹Booth School of Business, University of Chicago, veronika.rockova@chicagobooth.edu

²Mathematics Institute, Leiden University, svdpas@math.leidenuniv.nl

Since their inception in the 1980s, regression trees have been one of the more widely used nonparametric prediction methods. Tree-structured methods yield a histogram reconstruction of the regression surface, where the bins correspond to terminal nodes of recursive partitioning. Trees are powerful, yet susceptible to overfitting. Strategies against overfitting have traditionally relied on pruning greedily grown trees. The Bayesian framework offers an alternative remedy against overfitting through priors. Roughly speaking, a good prior charges smaller trees where overfitting does not occur. While the consistency of random histograms, trees and their ensembles has been studied quite extensively, the theoretical understanding of the Bayesian counterparts has been missing. In this paper, we take a step toward understanding why/when do Bayesian trees and forests not overfit. To address this question, we study the speed at which the posterior concentrates around the true smooth regression function. We propose a *spike-and-tree* variant of the popular Bayesian CART prior and establish new theoretical results showing that regression trees (and forests) (a) are capable of recovering smooth regression surfaces (with smoothness not exceeding one), achieving optimal rates up to a log factor, (b) can adapt to the unknown level of smoothness and (c) can perform effective dimension reduction when $p > n$. These results provide a piece of missing theoretical evidence explaining why Bayesian trees (and additive variants thereof) have worked so well in practice.

REFERENCES

- [1] ANDERSON, T. W. (1966). Some nonparametric multivariate procedures based on statistically equivalent blocks. In *Multivariate Analysis (Proc. Internat. Sympos., Dayton, Ohio, 1965)* 5–27. Academic Press, New York. [MR0214250](#)
- [2] BENTLEY, J. L. (1975). Multidimensional binary search trees used for associative searching. *Commun. ACM* **18** 509–517.
- [3] BHATTACHARYA, A., PATI, D. and DUNSON, D. (2014). Anisotropic function estimation using multi-bandwidth Gaussian processes. *Ann. Statist.* **42** 352–381. [MR3189489](#) <https://doi.org/10.1214/13-AOS1192>
- [4] BIAU, G. (2012). Analysis of a random forests model. *J. Mach. Learn. Res.* **13** 1063–1095. [MR2930634](#)
- [5] BIAU, G., DEVROYE, L. and LUGOSI, G. (2008). Consistency of random forests and other averaging classifiers. *J. Mach. Learn. Res.* **9** 2015–2033. [MR2447310](#)
- [6] BIAU, G. and SCORNET, E. (2016). A random forest guided tour. *TEST* **25** 197–227. [MR3493512](#) <https://doi.org/10.1007/s11749-016-0481-7>
- [7] BLEICH, J., KAPELNER, A., GEORGE, E. I. and JENSEN, S. T. (2014). Variable selection for BART: An application to gene regulation. *Ann. Appl. Stat.* **8** 1750–1781. [MR3271352](#) <https://doi.org/10.1214/14-AOAS755>
- [8] BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.
- [9] BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. Wadsworth Statistics/Probability Series. Wadsworth, Belmont, CA. [MR0726392](#)
- [10] CASTILLO, I. (2017). Pólya tree posterior distributions on densities. *Ann. Inst. Henri Poincaré Probab. Stat.* **53** 2074–2102. [MR3729648](#) <https://doi.org/10.1214/16-AIHP784>

MSC2020 subject classifications. 62G08, 62G20.

Key words and phrases. Additive regression, asymptotic minimaxity, BART, Bayesian CART, posterior concentration, recursive partitioning, regression trees.

- [11] CASTILLO, I. and ROUSSEAU, J. (2015). A Bernstein–von Mises theorem for smooth functionals in semi-parametric models. *Ann. Statist.* **43** 2353–2383. [MR3405597](#) <https://doi.org/10.1214/15-AOS1336>
- [12] CASTILLO, I., SCHMIDT-HIEBER, J. and VAN DER VAART, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.* **43** 1986–2018. [MR3375874](#) <https://doi.org/10.1214/15-AOS1334>
- [13] CASTILLO, I. and VAN DER VAART, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Ann. Statist.* **40** 2069–2101. [MR3059077](#) <https://doi.org/10.1214/12-AOS1029>
- [14] CHIPMAN, H., GEORGE, E. I. and MCCULLOCH, R. E. (1997). Bayesian CART model search. *J. Amer. Statist. Assoc.* **93** 935–960.
- [15] CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4** 266–298. [MR2758172](#) <https://doi.org/10.1214/09-AOAS285>
- [16] CORAM, M. and LALLEY, S. P. (2006). Consistency of Bayes estimators of a binary regression function. *Ann. Statist.* **34** 1233–1269. [MR2278357](#) <https://doi.org/10.1214/009053606000000236>
- [17] DE JONGE, R. and VAN ZANTEN, H. (2013). Semiparametric Bernstein–von Mises for the error standard deviation. *Electron. J. Stat.* **7** 217–243. [MR3020419](#) <https://doi.org/10.1214/13-EJS768>
- [18] DENISON, D. G. T., MALLICK, B. K. and SMITH, A. F. M. (1998). A Bayesian CART algorithm. *Biometrika* **85** 363–377. [MR1649118](#) <https://doi.org/10.1093/biomet/85.2.363>
- [19] DEVROYE, L. and GYÖRFI, L. (1985). Distribution free exponential bound for the L_1 error of partitioning-estimates of a regression function. In *Probability and Statistical Decision Theory, Vol. A (Bad Tatzmannsdorf, 1983)* 67–76. Reidel, Dordrecht. [MR0851019](#)
- [20] DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition. Applications of Mathematics (New York)* **31**. Springer, New York. [MR1383093](#) <https://doi.org/10.1007/978-1-4612-0711-5>
- [21] DONOHO, D. L. (1997). CART and best-ortho-basis: A connection. *Ann. Statist.* **25** 1870–1911. [MR1474073](#) <https://doi.org/10.1214/aos/1069362377>
- [22] ENGEL, J. (1994). A simple wavelet approach to nonparametric regression from recursive partitioning schemes. *J. Multivariate Anal.* **49** 242–254. [MR1276437](#) <https://doi.org/10.1006/jmva.1994.1024>
- [23] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. [MR1790007](#) <https://doi.org/10.1214/aos/1016218228>
- [24] GHOSAL, S. and VAN DER VAART, A. (2007). Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.* **35** 192–223. [MR2332274](#) <https://doi.org/10.1214/0090536060000001172>
- [25] GORDON, L. and OLSHEN, R. A. (1980). Consistent nonparametric regression from recursive partitioning schemes. *J. Multivariate Anal.* **10** 611–627. [MR0599694](#) [https://doi.org/10.1016/0047-259X\(80\)90074-3](https://doi.org/10.1016/0047-259X(80)90074-3)
- [26] GORDON, L. and OLSHEN, R. A. (1984). Almost surely consistent nonparametric regression from recursive partitioning schemes. *J. Multivariate Anal.* **15** 147–163. [MR0763592](#) [https://doi.org/10.1016/0047-259X\(84\)90022-8](https://doi.org/10.1016/0047-259X(84)90022-8)
- [27] KPOTUFE, S. (2010). *The Curse of Dimension in Nonparametric Regression*. ProQuest LLC, Ann Arbor, MI. Ph.D. thesis, University of California, San Diego. [MR2794782](#)
- [28] LIAN, H. (2007). Consistency of Bayesian estimation of a step function. *Statist. Probab. Lett.* **77** 19–24. [MR2339014](#) <https://doi.org/10.1016/j.spl.2006.05.007>
- [29] LINERO, A. R. (2018). Bayesian regression trees for high-dimensional prediction and variable selection. *J. Amer. Statist. Assoc.* **113** 626–636. [MR3832214](#) <https://doi.org/10.1080/01621459.2016.1264957>
- [30] LINERO, A. R. and YANG, Y. (2018). Bayesian regression tree ensembles that adapt to smoothness and sparsity. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 1087–1110. Available at [arXiv:1707.09461](#). [MR3874311](#) <https://doi.org/10.1111/rssb.12293>
- [31] LIU, L. and WONG, W. H. (2015). Multivariate density estimation via adaptive partitioning (ii): Posterior concentration. Preprint. Available at [arXiv:1508.04812](#).
- [32] LIU, Y., ROČKOVÁ, V. and WANG, Y. (2018). ABC variable selection with Bayesian forests. Preprint. Available at [arXiv:1806.02304](#).
- [33] LU, L., JIANG, H. and WONG, W. H. (2013). Multivariate density estimation by Bayesian sequential partitioning. *J. Amer. Statist. Assoc.* **108** 1402–1410. [MR3174716](#) <https://doi.org/10.1080/01621459.2013.813389>
- [34] NOBEL, A. (1996). Histogram regression estimation using data-dependent partitions. *Ann. Statist.* **24** 1084–1105. [MR1401839](#) <https://doi.org/10.1214/aos/1032526958>
- [35] VAN DER PAS, S. and ROČKOVÁ, V. (2017). Bayesian dyadic trees and histograms for regression. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* 2086–2096.
- [36] ROČKOVÁ, V. and VAN DER PAS, S. (2020). Supplement to “Posterior concentration for Bayesian regression trees and forests.” <https://doi.org/10.1214/19-AOS1879SUPP>.

- [37] ROČKOVÁ, V. and SAHA, E. (2019). On theory for BART. *Proc. Mach. Learn. Res.* **89** 2839–2848.
- [38] SCORNET, E., BIAU, G. and VERT, J.-P. (2015). Consistency of random forests. *Ann. Statist.* **43** 1716–1741. [MR3357876](#) <https://doi.org/10.1214/15-AOS1321>
- [39] SCRICCILOLO, C. (2007). On rates of convergence for Bayesian density estimation. *Scand. J. Stat.* **34** 626–642. [MR2368802](#) <https://doi.org/10.1111/j.1467-9469.2006.00540.x>
- [40] SHEN, X. and WASSERMAN, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* **29** 687–714. [MR1865337](#) <https://doi.org/10.1214/aos/1009210686>
- [41] STONE, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595–645. [MR0443204](#)
- [42] STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053. [MR0673642](#)
- [43] STONE, C. J. (1985). An asymptotically optimal histogram selection rule. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983)*. Wadsworth Statist./Probab. Ser. 513–520. Wadsworth, Belmont, CA. [MR0822050](#)
- [44] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.* **36** 1435–1463. [MR2418663](#) <https://doi.org/10.1214/0090536070000000613>
- [45] VERMA, N., KPOTUFE, S. and DASGUPTA, S. (2009). Which spatial partition trees are adaptive to intrinsic dimension? In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* 565–574. AUAI Press.
- [46] WAGER, S. and WALTHER, G. (2015). Adaptive concentration of regression trees, with application to random forests. Preprint. Available at [arXiv:1503.06388](#).
- [47] YANG, Y. and TOKDAR, S. T. (2015). Minimax-optimal nonparametric regression in high dimensions. *Ann. Statist.* **43** 652–674. [MR3319139](#) <https://doi.org/10.1214/14-AOS1289>
- [48] YOO, W. W. and GHOSAL, S. (2016). Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. *Ann. Statist.* **44** 1069–1102. [MR3485954](#) <https://doi.org/10.1214/15-AOS1398>

DOUBLE-SLICING ASSISTED SUFFICIENT DIMENSION REDUCTION FOR HIGH-DIMENSIONAL CENSORED DATA

BY SHANSHAN DING^{1,*}, WEI QIAN^{1,**} AND LAN WANG²

¹Department of Applied Economics and Statistics, University of Delaware, *sding@udel.edu; **weiqian@udel.edu

²Department of Management Science, University of Miami, lanwang@mbs.miami.edu

This paper provides a unified framework and an efficient algorithm for analyzing high-dimensional survival data under weak modeling assumptions. In particular, it imposes neither parametric distributional assumption nor linear regression assumption. It only assumes that the survival time T depends on a high-dimensional covariate vector $\mathbf{X}$ through low-dimensional linear combinations of covariates $\Gamma^T \mathbf{X}$. The censoring time is allowed to be conditionally independent of the survival time given the covariates. This general framework includes many popular parametric and semiparametric survival regression models as special cases. The proposed algorithm produces a number of practically useful outputs with theoretical guarantees, including a consistent estimate of the sufficient dimension reduction subspace of $T \mid \mathbf{X}$, a uniformly consistent Kaplan–Meier-type estimator of the conditional distribution function of T and a consistent estimator of the conditional quantile survival time. Our asymptotic results significantly extend the classical theory of sufficient dimension reduction for censored data (particularly that of Li, Wang and Chen in *Ann. Statist.* **27** (1999) 1–23) and the celebrated nonparametric Kaplan–Meier estimator to the setting where the number of covariates p diverges exponentially fast with the sample size n . We demonstrate the promising performance of the proposed new estimators through simulations and a real data example.

REFERENCES

- BERAN, R. (1981). Nonparametric regression with randomly censored survival data. Technical report, Univ. California, Berkeley, CA.
- BICKEL, P. J. and LEVINA, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. [MR2485008 https://doi.org/10.1214/08-AOS600](https://doi.org/10.1214/08-AOS600)
- BRADIC, J., FAN, J. and JIANG, J. (2011). Regularization for Cox’s proportional hazards model with NP-dimensionality. *Ann. Statist.* **39** 3092–3120. [MR3012402 https://doi.org/10.1214/11-AOS911](https://doi.org/10.1214/11-AOS911)
- BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA. [MR0726392](https://doi.org/10.1214/11-AOS911)
- BURA, E., DUARTE, S. and FORZANI, L. (2016). Sufficient reductions in regressions with exponential family inverse predictors. *J. Amer. Statist. Assoc.* **111** 1313–1329. [MR3561952 https://doi.org/10.1080/01621459.2015.1093944](https://doi.org/10.1080/01621459.2015.1093944)
- CHAI, H., ZHANG, Q., HUANG, J. and MA, S. (2019). Inference for low-dimensional covariates in a high-dimensional accelerated failure time model. *Statist. Sinica* **29** 877–894. [MR3931392](https://doi.org/10.1214/11-AOS911)
- CHEN, X., COOK, R. D. and ZOU, C. (2015). Diagnostic studies in sufficient dimension reduction. *Biometrika* **102** 545–558. [MR3394274 https://doi.org/10.1093/biomet/asv016](https://doi.org/10.1093/biomet/asv016)
- CHENG, W., SU, Y. and XU, F. (2013). CHD1L: A novel oncogene. *Molecular Cancer* **12** 170.
- CHENG, S. C., WEI, L. J. and YING, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82** 835–845. [MR1380818 https://doi.org/10.1093/biomet/82.4.835](https://doi.org/10.1093/biomet/82.4.835)
- COOK, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics*. Wiley Series in Probability and Statistics: Probability and Statistics. Wiley, New York. A Wiley-Interscience Publication. [MR1645673 https://doi.org/10.1002/9780470316931](https://doi.org/10.1002/9780470316931)

MSC2020 subject classifications. Primary 62H20, 62J07; secondary 62N02.

Key words and phrases. Central subspace, sufficient dimension reduction, variable selection, censored data, ultrahigh dimension, nonparametric estimation.

- COOK, R. D. (2003). Dimension reduction and graphical exploration in regression including survival analysis. *Stat. Med.* **22** 1399–1413.
- COOK, R. D., FORZANI, L. and ROTHMAN, A. J. (2012). Estimating sufficient reductions of the predictors in abundant high-dimensional regressions. *Ann. Statist.* **40** 353–384. MR3014310 <https://doi.org/10.1214/11-AOS962>
- COOK, R. D. and NI, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *J. Amer. Statist. Assoc.* **100** 410–428. MR2160547 <https://doi.org/10.1198/016214504000001501>
- COX, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika* **62** 441–444. MR0378189 <https://doi.org/10.1093/biomet/62.2.441>
- DING, S., QIAN, W. and WANG, L. (2020). Supplement to “Double-slicing assisted sufficient dimension reduction for high-dimensional censored data.” <https://doi.org/10.1214/19-AOS1880SUPP>.
- DU, P., MA, S. and LIANG, H. (2010). Penalized variable selection procedure for Cox models with semiparametric relative risk. *Ann. Statist.* **38** 2092–2117. MR2676884 <https://doi.org/10.1214/09-AOS780>
- FAN, J. and GIJBELS, I. (1994). Censored regression: Local linear approximations and their applications. *J. Amer. Statist. Assoc.* **89** 560–570. MR1294083
- FAN, J. and LI, R. (2002). Variable selection for Cox’s proportional hazards model and frailty model. *Ann. Statist.* **30** 74–99. MR1892656 <https://doi.org/10.1214/aos/1015362185>
- FANG, E. X., NING, Y. and LIU, H. (2017). Testing and confidence intervals for high dimensional proportional hazards models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1415–1437. MR3731669 <https://doi.org/10.1111/rssb.12224>
- FRIEDMAN, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **29** 1189–1232. MR1873328 <https://doi.org/10.1214/aos/1013203451>
- GERDS, T. A. and SCHUMACHER, M. (2007). Efron-type measures of prediction error for survival analysis. *Biometrics* **63** 1283–1287, 1316. MR2414608 <https://doi.org/10.1111/j.1541-0420.2007.00832.x>
- GONZALEZ-MANTEIGA, W. and CADARSO-SUAREZ, C. (1994). Asymptotic properties of a generalized Kaplan–Meier estimator with some applications. *J. Nonparametr. Stat.* **4** 65–78. MR1366364 <https://doi.org/10.1080/10485259408832601>
- HALL, P. and LI, K.-C. (1993). On almost linearity of low-dimensional projections from high-dimensional data. *Ann. Statist.* **21** 867–889. MR1232523 <https://doi.org/10.1214/aos/1176349155>
- HÄRDLE, W., HALL, P. and ICHIMURA, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21** 157–178. MR1212171 <https://doi.org/10.1214/aos/1176349020>
- HEAGERTY, P. J., LUMLEY, T. and PEPE, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56** 337–344.
- HSING, T. and REN, H. (2009). An RKHS formulation of the inverse regression dimension-reduction problem. *Ann. Statist.* **37** 726–755. MR2502649 <https://doi.org/10.1214/07-AOS589>
- HUANG, J., SUN, T., YING, Z., YU, Y. and ZHANG, C.-H. (2013). Oracle inequalities for the LASSO in the Cox model. *Ann. Statist.* **41** 1142–1165. MR3113806 <https://doi.org/10.1214/13-AOS1098>
- JOHNSON, B. A. (2009). On lasso for censored data. *Electron. J. Stat.* **3** 485–506. MR2507457 <https://doi.org/10.1214/08-EJS322>
- KONG, E. and XIA, Y. (2014). An adaptive composite quantile approach to dimension reduction. *Ann. Statist.* **42** 1657–1688. MR3262464 <https://doi.org/10.1214/14-AOS1242>
- LI, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86** 316–342. MR1137117
- LI, L. (2005). Survival prediction of diffuse large-B-cell lymphoma based on both clinical and gene expression information. *Bioinformatics* **22** 466–471.
- LI, L. (2007). Sparse sufficient dimension reduction. *Biometrika* **94** 603–613. MR2410011 <https://doi.org/10.1093/biomet/asm044>
- LI, B. (2018). *Sufficient Dimension Reduction: Methods and Applications with R. Monographs on Statistics and Applied Probability* **161**. CRC Press, Boca Raton, FL. MR3838449 <https://doi.org/10.1201/9781315119427>
- LI, B. and DONG, Y. (2009). Dimension reduction for nonelliptically distributed predictors. *Ann. Statist.* **37** 1272–1298. MR2509074 <https://doi.org/10.1214/08-AOS598>
- LI, L. and LI, H. (2004). Dimension reduction methods for microarrays with application to censored survival data. *Bioinformatics* **20** 3406–3412.
- 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., WANG, J.-L. and CHEN, C.-H. (1999). Dimension reduction for censored regression data. *Ann. Statist.* **27** 1–23. MR1701098 https://doi.org/10.1007/3-540-48294-6_14
- LI, L. and YIN, X. (2008). Sliced inverse regression with regularizations. *Biometrics* **64** 124–131, 323. MR2422826 <https://doi.org/10.1111/j.1541-0420.2007.00836.x>

- 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>
- LOPEZ, O. (2011). Nonparametric estimation of the multivariate distribution function in a censored regression model with applications. *Comm. Statist. Theory Methods* **40** 2639–2660. MR2860770 <https://doi.org/10.1080/03610926.2010.489175>
- LU, W., GOLDBERG, Y. and FINE, J. P. (2012). On the robustness of the adaptive lasso to model misspecification. *Biometrika* **99** 717–731. MR2966780 <https://doi.org/10.1093/biomet/ass027>
- LU, W. and LI, L. (2011). Sufficient dimension reduction for censored regression. *Biometrics* **67** 513–523. MR2829020 <https://doi.org/10.1111/j.1541-0420.2010.01490.x>
- 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>
- NADKARNI, N. V., ZHAO, Y. and KOSOROK, M. R. (2011). Inverse regression estimation for censored data. *J. Amer. Statist. Assoc.* **106** 178–190. MR2816712 <https://doi.org/10.1198/jasa.2011.tm08250>
- QIAN, W., DING, S. and COOK, R. D. (2018). Sparse minimum discrepancy approach to sufficient dimension reduction with simultaneous variable selection in ultrahigh dimension. *J. Amer. Statist. Assoc.* To appear.
- ROTHMAN, A. J., LEVINA, E. and ZHU, J. (2009). Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.* **104** 177–186. MR2504372 <https://doi.org/10.1198/jasa.2009.0101>
- SANTINI, R., PIETROBONO, S., PANDOLFI, S., MONTAGNANI, V., D'AMICO, M., PENACHIONI, J., VINCI, M., BORGOGNONI, L. and STECCA, B. (2014). SOX2 regulates self-renewal and tumorigenicity of human melanoma-initiating cells. *Oncogene* **33** 4697–4708.
- SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* **39** 1–13.
- SUN, Q., ZHU, R., WANG, T. and ZENG, D. (2019). Counting process-based dimension reduction methods for censored outcomes. *Biometrika* **106** 181–196. MR3912390 <https://doi.org/10.1093/biomet/asy064>
- TAN, K. M., WANG, Z., LIU, H. and ZHANG, T. (2018). Sparse generalized eigenvalue problem: Optimal statistical rates via truncated Rayleigh flow. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 1057–1086. MR3874310 <https://doi.org/10.1111/rssb.12291>
- TIBSHIRANI, R. (1997). The lasso method for variable selection in the Cox model. *Stat. Med.* **16** 385–395.
- WANG, J.-L., XUE, L., ZHU, L. and CHONG, Y. S. (2010). Estimation for a partial-linear single-index model. *Ann. Statist.* **38** 246–274. MR2589322 <https://doi.org/10.1214/09-AOS712>
- WANG, T., CHEN, M., ZHAO, H. and ZHU, L. (2018). Estimating a sparse reduction for general regression in high dimensions. *Stat. Comput.* **28** 33–46. MR3741635 <https://doi.org/10.1007/s11222-016-9714-6>
- XIA, Y., ZHANG, D. and XU, J. (2010). Dimension reduction and semiparametric estimation of survival models. *J. Amer. Statist. Assoc.* **105** 278–290. MR2656052 <https://doi.org/10.1198/jasa.2009.tm09372>
- YANG, Y. (2007). Consistency of cross validation for comparing regression procedures. *Ann. Statist.* **35** 2450–2473. MR2382654 <https://doi.org/10.1214/009053607000000514>
- 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>
- YIN, X. and HILAFU, H. (2015). Sequential sufficient dimension reduction for large p , small n problems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 879–892. MR3382601 <https://doi.org/10.1111/rssb.12093>
- YIN, X. and LI, B. (2011). Sufficient dimension reduction based on an ensemble of minimum average variance estimators. *Ann. Statist.* **39** 3392–3416. MR3012413 <https://doi.org/10.1214/11-AOS950>
- 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>
- YU, Z., ZHU, L., PENG, H. and ZHU, L. (2013). Dimension reduction and predictor selection in semiparametric models. *Biometrika* **100** 641–654. MR3094442 <https://doi.org/10.1093/biomet/ast005>
- YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67. MR2212574 <https://doi.org/10.1111/j.1467-9868.2005.00532.x>
- ZENG, D. (2004). Estimating marginal survival function by adjusting for dependent censoring using many covariates. *Ann. Statist.* **32** 1533–1555. MR2089133 <https://doi.org/10.1214/009053604000000508>
- ZHANG, C.-H. and HUANG, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. MR2435448 <https://doi.org/10.1214/07-AOS520>
- ZHANG, H. H. and LU, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika* **94** 691–703. MR2410017 <https://doi.org/10.1093/biomet/asm037>
- ZHAO, G., MA, Y. and LU, W. (2017). Efficient estimation for dimension reduction with censored data. arXiv preprint, [arXiv:1710.05377](https://arxiv.org/abs/1710.05377).
- 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.-P., LI, L., LI, R. and ZHU, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc.* **106** 1464–1475. MR2896849 <https://doi.org/10.1198/jasa.2011.tm10563>

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

ASYMPTOTIC FREQUENTIST COVERAGE PROPERTIES OF BAYESIAN CREDIBLE SETS FOR SIEVE PRIORS

BY JUDITH ROUSSEAU¹ AND BOTOND SZABO²

¹*Statistics Department, Oxford University, rousseau@ceremade.dauphine.fr*

²*Mathematical Institute, Leiden University, b.t.szabo@math.leidenuniv.nl*

We investigate the frequentist coverage properties of (certain) Bayesian credible sets in a general, adaptive, nonparametric framework. It is well known that the construction of adaptive and honest confidence sets is not possible in general. To overcome this problem (in context of sieve type of priors), we introduce an extra assumption on the functional parameters, the so-called “general polished tail” condition. We then show that under standard assumptions, both the hierarchical and empirical Bayes methods, result in honest confidence sets for sieve type of priors in general settings and we characterize their size. We apply the derived abstract results to various examples, including the nonparametric regression model, density estimation using exponential families of priors, density estimation using histogram priors and the nonparametric classification model, for which we show that their size is near minimax adaptive with respect to the considered specific pseudometrics.

REFERENCES

- [1] ARBEL, J., GAYRAUD, G. and ROUSSEAU, J. (2013). Bayesian optimal adaptive estimation using a sieve prior. *Scand. J. Stat.* **40** 549–570. [MR3091697](#) <https://doi.org/10.1002/sjso.12002>
- [2] BELITSER, E. (2017). On coverage and local radial rates of credible sets. *Ann. Statist.* **45** 1124–1151. [MR3662450](#) <https://doi.org/10.1214/16-AOS1477>
- [3] BELITSER, E. and NURUSHEV, N. (2015). Needles and straw in a haystack: Empirical Bayes confidence for possibly sparse sequences. Preprint. Available at [arXiv:1511.01803](https://arxiv.org/abs/1511.01803).
- [4] BULL, A. D. (2012). Honest adaptive confidence bands and self-similar functions. *Electron. J. Stat.* **6** 1490–1516. [MR2988456](#) <https://doi.org/10.1214/12-EJS720>
- [5] BULL, A. D. and NICKL, R. (2013). Adaptive confidence sets in L^2 . *Probab. Theory Related Fields* **156** 889–919. [MR3078289](#) <https://doi.org/10.1007/s00440-012-0446-z>
- [6] CAI, T. T. and LOW, M. G. (2004). An adaptation theory for nonparametric confidence intervals. *Ann. Statist.* **32** 1805–1840. [MR2102494](#) <https://doi.org/10.1214/009053604000000049>
- [7] CARPENTIER, A. (2013). Honest and adaptive confidence sets in L_p . *Electron. J. Stat.* **7** 2875–2923. [MR3148371](#) <https://doi.org/10.1214/13-EJS867>
- [8] CARPENTIER, A. and NICKL, R. (2015). On signal detection and confidence sets for low rank inference problems. *Electron. J. Stat.* **9** 2675–2688. [MR3432430](#) <https://doi.org/10.1214/15-EJS1087>
- [9] CASTILLO, I. (2008). Lower bounds for posterior rates with Gaussian process priors. *Electron. J. Stat.* **2** 1281–1299. [MR2471287](#) <https://doi.org/10.1214/08-EJS273>
- [10] CASTILLO, I. and MISMER, R. (2018). Empirical Bayes analysis of spike and slab posterior distributions. *Electron. J. Stat.* **12** 3953–4001. [MR3885271](#) <https://doi.org/10.1214/18-EJS1494>
- [11] 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>
- [12] 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>
- [13] CASTILLO, I. and ROUSSEAU, J. (2015). A Bernstein–von Mises theorem for smooth functionals in semi-parametric models. *Ann. Statist.* **43** 2353–2383. [MR3405597](#) <https://doi.org/10.1214/15-AOS1336>
- [14] CASTILLO, I. and SZABO, B. (2020). Spike and slab empirical Bayes sparse credible sets. *Bernoulli* **26** 127–158. <https://doi.org/10.3150/19-BEJ1119>

MSC2020 subject classifications. Primary 62G20, 62G05; secondary 62G08, 62G07.

Key words and phrases. Uncertainty quantification, coverage, posterior contraction rates, adaptation, empirical Bayes, hierarchical Bayes, nonparametric regression, density estimation, classification, sieve prior.

- [15] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2014). Anti-concentration and honest, adaptive confidence bands. *Ann. Statist.* **42** 1787–1818. MR3262468 <https://doi.org/10.1214/14-AOS1235>
- [16] COX, D. D. (1993). An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.* **21** 903–923. MR1232525 <https://doi.org/10.1214/aos/1176349157>
- [17] FREEDMAN, D. (1999). On the Bernstein–von Mises theorem with infinite-dimensional parameters. *Ann. Statist.* **27** 1119–1140. MR1740119 <https://doi.org/10.1214/aos/1017938917>
- [18] GAO, C. and ZHOU, H. H. (2016). Rate exact Bayesian adaptation with modified block priors. *Ann. Statist.* **44** 318–345. MR3449770 <https://doi.org/10.1214/15-AOS1368>
- [19] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007 <https://doi.org/10.1214/aos/1016218228>
- [20] GHOSAL, S. and VAN DER VAART, A. (2007). Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.* **35** 192–223. MR2332274 <https://doi.org/10.1214/009053606000001172>
- [21] GINÉ, E. and NICKL, R. (2010). Confidence bands in density estimation. *Ann. Statist.* **38** 1122–1170. MR2604707 <https://doi.org/10.1214/09-AOS738>
- [22] GINÉ, E. and NICKL, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge Univ. Press, New York. MR3588285 <https://doi.org/10.1017/CBO9781107337862>
- [23] HADJI, A. and SZABO, B. (2019). Can we trust bayesian uncertainty quantification from Gaussian process priors with squared exponential covariance kernel? Preprint. Available at [arXiv:1904.01383](https://arxiv.org/abs/1904.01383).
- [24] HOFFMANN, M. and NICKL, R. (2011). On adaptive inference and confidence bands. *Ann. Statist.* **39** 2383–2409. MR2906872 <https://doi.org/10.1214/11-AOS903>
- [25] HOFFMANN, M., ROUSSEAU, J. and SCHMIDT-HIEBER, J. (2015). On adaptive posterior concentration rates. *Ann. Statist.* **43** 2259–2295. MR3396985 <https://doi.org/10.1214/15-AOS1341>
- [26] KIRICHENKO, A. and VAN ZANTEN, H. (2017). Estimating a smooth function on a large graph by Bayesian Laplacian regularisation. *Electron. J. Stat.* **11** 891–915. MR3629018 <https://doi.org/10.1214/17-EJS1253>
- [27] KNAPIK, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2011). Bayesian inverse problems with Gaussian priors. *Ann. Statist.* **39** 2626–2657. MR2906881 <https://doi.org/10.1214/11-AOS920>
- [28] LEAHU, H. (2011). On the Bernstein–von Mises phenomenon in the Gaussian white noise model. *Electron. J. Stat.* **5** 373–404. MR2802048 <https://doi.org/10.1214/11-EJS611>
- [29] LOW, M. G. (1997). On nonparametric confidence intervals. *Ann. Statist.* **25** 2547–2554. MR1604412 <https://doi.org/10.1214/aos/1030741084>
- [30] MCVINISH, R., ROUSSEAU, J. and MENGENSEN, K. (2009). Bayesian goodness of fit testing with mixtures of triangular distributions. *Scand. J. Stat.* **36** 337–354. MR2528988 <https://doi.org/10.1111/j.1467-9469.2008.00620.x>
- [31] MUKHERJEE, R. and SEN, S. (2018). Optimal adaptive inference in random design binary regression. *Bernoulli* **24** 699–739. MR3706774 <https://doi.org/10.3150/16-BEJ893>
- [32] NICKL, R. and SZABÓ, B. (2016). A sharp adaptive confidence ball for self-similar functions. *Stochastic Process. Appl.* **126** 3913–3934. MR3565485 <https://doi.org/10.1016/j.spa.2016.04.017>
- [33] NICKL, R. and VAN DE GEER, S. (2013). Confidence sets in sparse regression. *Ann. Statist.* **41** 2852–2876. MR3161450 <https://doi.org/10.1214/13-AOS1170>
- [34] PETRONE, S., ROUSSEAU, J. and SCRICCILO, C. (2014). Bayes and empirical Bayes: Do they merge? *Biometrika* **101** 285–302. MR3215348 <https://doi.org/10.1093/biomet/ast067>
- [35] PICARD, D. and TRIBOULEY, K. (2000). Adaptive confidence interval for pointwise curve estimation. *Ann. Statist.* **28** 298–335. MR1762913 <https://doi.org/10.1214/aos/1016120374>
- [36] 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>
- [37] 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>
- [38] RIVOIRARD, V. and ROUSSEAU, J. (2012). Posterior concentration rates for infinite dimensional exponential families. *Bayesian Anal.* **7** 311–333. MR2934953 <https://doi.org/10.1214/12-BA710>
- [39] ROBINS, J. and VAN DER VAART, A. (2006). Adaptive nonparametric confidence sets. *Ann. Statist.* **34** 229–253. MR2275241 <https://doi.org/10.1214/0090536050000000877>
- [40] ROUSSEAU, J. (2007). Approximating interval hypothesis: p -values and Bayes factors. In *Bayesian Statistics 8* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford Sci. Publ. 417–452. Oxford Univ. Press, Oxford. MR2433202
- [41] ROUSSEAU, J. and SZABO, B. (2017). Asymptotic behaviour of the empirical Bayes posteriors associated to maximum marginal likelihood estimator. *Ann. Statist.* **45** 833–865. MR3650402 <https://doi.org/10.1214/16-AOS1469>

- [42] ROUSSEAU, J. and SZABO, B. (2020). Supplement to “Asymptotic frequentist coverage properties of Bayesian credible sets for sieve priors.” <https://doi.org/10.1214/19-AOS1881SUPP>.
- [43] RUDELSON, M. (1999). Random vectors in the isotropic position. *J. Funct. Anal.* **164** 60–72. MR1694526 <https://doi.org/10.1006/jfan.1998.3384>
- [44] SCRICCILO, C. (2007). On rates of convergence for Bayesian density estimation. *Scand. J. Stat.* **34** 626–642. MR2368802 <https://doi.org/10.1111/j.1467-9469.2006.00540.x>
- [45] SERRA, P. and KRIVOBOKOVA, T. (2017). Adaptive empirical Bayesian smoothing splines. *Bayesian Anal.* **12** 219–238. MR3597573 <https://doi.org/10.1214/16-BA997>
- [46] SNIEKERS, S. and VAN DER VAART, A. (2015). Adaptive Bayesian credible sets in regression with a Gaussian process prior. *Electron. J. Stat.* **9** 2475–2527. MR3425364 <https://doi.org/10.1214/15-EJS1078>
- [47] SÖHL, J. and TRABS, M. (2016). Adaptive confidence bands for Markov chains and diffusions: Estimating the invariant measure and the drift. *ESAIM Probab. Stat.* **20** 432–462. MR3581829 <https://doi.org/10.1051/ps/2016017>
- [48] SZABÓ, B. (2015). On Bayesian based adaptive confidence sets for linear functionals. In *Bayesian Statistics from Methods to Models and Applications. Springer Proc. Math. Stat.* **126** 91–105. Springer, Cham. MR3374423 https://doi.org/10.1007/978-3-319-16238-6_8
- [49] SZABÓ, B., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *Ann. Statist.* **43** 1391–1428. MR3357861 <https://doi.org/10.1214/14-AOS1270>
- [50] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation. Springer Series in Statistics.* Springer, New York. MR2724359 <https://doi.org/10.1007/b13794>
- [51] VAN DER PAS, S., SZABÓ, B. and VAN DER VAART, A. (2017). Uncertainty quantification for the horseshoe (with discussion). *Bayesian Anal.* **12** 1221–1274. MR3724985 <https://doi.org/10.1214/17-BA1065>
- [52] 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>
- [53] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.* **36** 1435–1463. MR2418663 <https://doi.org/10.1214/0090536070000000613>
- [54] VERDINELLI, I. and WASSERMAN, L. (1998). Bayesian goodness-of-fit testing using infinite-dimensional exponential families. *Ann. Statist.* **26** 1215–1241. MR1647645 <https://doi.org/10.1214/aos/1024691240>
- [55] WEIMIN YOO, W. and VAN DER VAART, A. W. (2017). The Bayes Lepski’s method and credible bands through volume of tubular neighborhoods. ArXiv e-prints.
- [56] YANG, Y. (1999). Minimax nonparametric classification. I. Rates of convergence. *IEEE Trans. Inform. Theory* **45** 2271–2284. MR1725115 <https://doi.org/10.1109/18.796368>
- [57] YANG, Y. (1999). Minimax nonparametric classification. II. Model selection for adaptation. *IEEE Trans. Inform. Theory* **45** 2285–2292. MR1725116 <https://doi.org/10.1109/18.796369>
- [58] YOO, W. W. and GHOSAL, S. (2016). Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. *Ann. Statist.* **44** 1069–1102. MR3485954 <https://doi.org/10.1214/15-AOS1398>

CONVERGENCE RATES OF VARIATIONAL POSTERIOR DISTRIBUTIONS

BY FENGSHUO ZHANG^{*} AND CHAO GAO[†]

Department of Statistics, University of Chicago, ^{}fengshuo@galton.uchicago.edu; [†]chaogao@galton.uchicago.edu*

We study convergence rates of variational posterior distributions for non-parametric and high-dimensional inference. We formulate general conditions on prior, likelihood and variational class that characterize the convergence rates. Under similar “prior mass and testing” conditions considered in the literature, the rate is found to be the sum of two terms. The first term stands for the convergence rate of the true posterior distribution, and the second term is contributed by the variational approximation error. For a class of priors that admit the structure of a mixture of product measures, we propose a novel prior mass condition, under which the variational approximation error of the mean-field class is dominated by convergence rate of the true posterior. We demonstrate the applicability of our general results for various models, prior distributions and variational classes by deriving convergence rates of the corresponding variational posteriors.

REFERENCES

- [1] ALQUIER, P. and RIDGWAY, J. (2017). Concentration of tempered posteriors and of their variational approximations. Preprint. Available at [arXiv:1706.09293](https://arxiv.org/abs/1706.09293).
- [2] BARRON, A., SCHERVISH, M. J. and WASSERMAN, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* **27** 536–561. [MR1714718 https://doi.org/10.1214/aos/1018031206](https://doi.org/10.1214/aos/1018031206)
- [3] BARRON, A. R. (1988). The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Department of Statistics, University of Illinois.
- [4] BICKEL, P., CHOI, D., CHANG, X. and ZHANG, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Ann. Statist.* **41** 1922–1943. [MR3127853 https://doi.org/10.1214/13-AOS1124](https://doi.org/10.1214/13-AOS1124)
- [5] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. [MR2533469 https://doi.org/10.1214/08-AOS620](https://doi.org/10.1214/08-AOS620)
- [6] BLEI, D. M., KUCUKELBIR, A. and MCAULIFFE, J. D. (2017). Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* **112** 859–877. [MR3671776 https://doi.org/10.1080/01621459.2017.1285773](https://doi.org/10.1080/01621459.2017.1285773)
- [7] BLEI, D. M., NG, A. Y. and JORDAN, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3** 993–1022.
- [8] CARBONETTO, P. and STEPHENS, M. (2012). Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Anal.* **7** 73–107. [MR2896713 https://doi.org/10.1214/12-BA703](https://doi.org/10.1214/12-BA703)
- [9] CASTILLO, I. (2008). Lower bounds for posterior rates with Gaussian process priors. *Electron. J. Stat.* **2** 1281–1299. [MR2471287 https://doi.org/10.1214/08-EJS273](https://doi.org/10.1214/08-EJS273)
- [10] CASTILLO, I. (2014). On Bayesian supremum norm contraction rates. *Ann. Statist.* **42** 2058–2091. [MR3262477 https://doi.org/10.1214/14-AOS1253](https://doi.org/10.1214/14-AOS1253)
- [11] CASTILLO, I., SCHMIDT-HIEBER, J. and VAN DER VAART, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.* **43** 1986–2018. [MR3375874 https://doi.org/10.1214/15-AOS1334](https://doi.org/10.1214/15-AOS1334)
- [12] CHÉRIEF-ABDELLATIF, B.-E. and ALQUIER, P. (2018). Consistency of variational Bayes inference for estimation and model selection in mixtures. *Electron. J. Stat.* **12** 2995–3035. [MR3855643 https://doi.org/10.1214/18-EJS1475](https://doi.org/10.1214/18-EJS1475)

MSC2020 subject classifications. Primary 62C10; secondary 62F15.

Key words and phrases. Posterior contraction, mean-field variational inference, density estimation, Gaussian sequence model, piecewise constant model, empirical Bayes.

- [13] FRIEDRICH, F., KEMPE, A., LIEBSCHER, V. and WINKLER, G. (2008). Complexity penalized M -estimation: Fast computation. *J. Comput. Graph. Statist.* **17** 201–224. [MR2424802](#) <https://doi.org/10.1198/106186008X285591>
- [14] GAO, C., HAN, F. and ZHANG, C.-H. (2017). Minimax risk bounds for piecewise constant models. Preprint. Available at [arXiv:1705.06386](#).
- [15] GAO, C., VAN DER VAART, A. W. and ZHOU, H. H. (2015). A general framework for bayes structured linear models. Preprint. Available at [arXiv:1506.02174](#).
- [16] GAO, C. and ZHOU, H. H. (2016). Rate exact Bayesian adaptation with modified block priors. *Ann. Statist.* **44** 318–345. [MR3449770](#) <https://doi.org/10.1214/15-AOS1368>
- [17] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. [MR1790007](#) <https://doi.org/10.1214/aos/1016218228>
- [18] GHOSAL, S. and VAN DER VAART, A. (2007). Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.* **35** 192–223. [MR2332274](#) <https://doi.org/10.1214/009053606000001172>
- [19] HOFFMANN, M., ROUSSEAU, J. and SCHMIDT-HIEBER, J. (2015). On adaptive posterior concentration rates. *Ann. Statist.* **43** 2259–2295. [MR3396985](#) <https://doi.org/10.1214/15-AOS1341>
- [20] JOHNSTONE, I. M. (2011). Gaussian estimation: Sequence and wavelet models. Manuscript.
- [21] KRUIJER, W., ROUSSEAU, J. and VAN DER VAART, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electron. J. Stat.* **4** 1225–1257. [MR2735885](#) <https://doi.org/10.1214/10-EJS584>
- [22] LAFFERTY, J. D. and BLEI, D. M. (2006). Correlated topic models. In *Advances in Neural Information Processing Systems* 147–154.
- [23] LECAM, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** 38–53. [MR0334381](#)
- [24] MAUGIS-RABUSSEAU, C. and MICHEL, B. (2013). Adaptive density estimation for clustering with Gaussian mixtures. *ESAIM Probab. Stat.* **17** 698–724. [MR3126158](#) <https://doi.org/10.1051/ps/2012018>
- [25] PATI, D., BHATTACHARYA, A. and YANG, Y. (2018). On statistical optimality of variational Bayes. In *International Conference on Artificial Intelligence and Statistics* 1579–1588.
- [26] RAJ, A., STEPHENS, M. and PRITCHARD, J. K. (2014). faststructure: Variational inference of population structure in large snp data sets. *Genetics* **197** 573–589.
- [27] RIVOIRARD, V. and ROUSSEAU, J. (2012). Posterior concentration rates for infinite dimensional exponential families. *Bayesian Anal.* **7** 311–333. [MR2934953](#) <https://doi.org/10.1214/12-BA710>
- [28] ROUSSEAU, J. and SZABO, B. (2017). Asymptotic behaviour of the empirical Bayes posteriors associated to maximum marginal likelihood estimator. *Ann. Statist.* **45** 833–865. [MR3650402](#) <https://doi.org/10.1214/16-AOS1469>
- [29] SCHWARTZ, L. (1965). On Bayes procedures. *Z. Wahrsch. Verw. Gebiete* **4** 10–26. [MR0184378](#) <https://doi.org/10.1007/BF00535479>
- [30] SHEN, X. and WASSERMAN, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* **29** 687–714. [MR1865337](#) <https://doi.org/10.1214/aos/1009210686>
- [31] SUDDERTH, E. B. and JORDAN, M. I. (2009). Shared segmentation of natural scenes using dependent Pitman–Yor processes. In *Advances in Neural Information Processing Systems* 1585–1592.
- [32] VAN ERVEN, T. and HARREMOËS, P. (2014). Rényi divergence and Kullback–Leibler divergence. *IEEE Trans. Inform. Theory* **60** 3797–3820. [MR3225930](#) <https://doi.org/10.1109/TIT.2014.2320500>
- [33] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.* **36** 1435–1463. [MR2418663](#) <https://doi.org/10.1214/009053607000000613>
- [34] WALKER, S. and HJORT, N. L. (2001). On Bayesian consistency. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 811–821. [MR1872068](#) <https://doi.org/10.1111/1467-9868.00314>
- [35] WALKER, S. G., LIJOI, A. and PRÜNSTER, I. (2007). On rates of convergence for posterior distributions in infinite-dimensional models. *Ann. Statist.* **35** 738–746. [MR2336866](#) <https://doi.org/10.1214/009053606000001361>
- [36] WANG, Y. and BLEI, D. M. (2019). Frequentist consistency of variational Bayes. *J. Amer. Statist. Assoc.* **114** 1147–1161. [MR4011769](#) <https://doi.org/10.1080/01621459.2018.1473776>
- [37] YANG, Y., PATI, D. and BHATTACHARYA, A. (2017). α -variational inference with statistical guarantees. Preprint. Available at [arXiv:1710.03266](#).
- [38] ZHANG, A. Y. and ZHOU, H. H. (2017). Theoretical and computational guarantees of mean field variational inference for community detection. Preprint. Available at [arXiv:1710.11268](#).
- [39] ZHANG, F. and GAO, C. (2020). Supplement to “Convergence rates of variational posterior distributions.” <https://doi.org/10.1214/19-AOS1883SUPP>.
- [40] ZHANG, T. (2006). From ϵ -entropy to KL-entropy: Analysis of minimum information complexity density estimation. *Ann. Statist.* **34** 2180–2210. [MR2291497](#) <https://doi.org/10.1214/009053606000000704>

TWO-SAMPLE HYPOTHESIS TESTING FOR INHOMOGENEOUS RANDOM GRAPHS

BY DEBARGHYA GHOSHDASTIDAR^{1,*}, MAURILIO GUTZEIT^{2,†},
ALEXANDRA CARPENTIER^{2,‡} AND ULRIKE VON LUXBURG^{1,**}

¹Department of Computer Science, University of Tübingen, *debarghya.ghoshdastidar@uni-tuebingen.de;
**luxburg@informatik.uni-tuebingen.de

²Faculty of Mathematics, Otto von Guericke University Magdeburg, †maurilio.gutzeit@ovgu.de;
‡alexandra.carpentier@ovgu.de

The study of networks leads to a wide range of high-dimensional inference problems. In many practical applications, one needs to draw inference from one or few large sparse networks. The present paper studies hypothesis testing of graphs in this high-dimensional regime, where the goal is to test between two populations of inhomogeneous random graphs defined on the same set of n vertices. The size of each population m is much smaller than n , and can even be a constant as small as 1. The critical question in this context is whether the problem is solvable for small m .

We answer this question from a minimax testing perspective. Let P , Q be the population adjacencies of two sparse inhomogeneous random graph models, and d be a suitably defined distance function. Given a population of m graphs from each model, we derive minimax separation rates for the problem of testing $P = Q$ against $d(P, Q) > \rho$. We observe that if m is small, then the minimax separation is too large for some popular choices of d , including total variation distance between corresponding distributions. This implies that some models that are widely separated in d cannot be distinguished for small m , and hence, the testing problem is generally not solvable in these cases.

We also show that if $m > 1$, then the minimax separation is relatively small if d is the Frobenius norm or operator norm distance between P and Q . For $m = 1$, only the latter distance provides small minimax separation. Thus, for these distances, the problem is solvable for small m . We also present near-optimal two-sample tests in both cases, where tests are adaptive with respect to sparsity level of the graphs.

REFERENCES

- 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* **29**.
ALBERT, R. and BARABÁSI, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Modern Phys.* **74** 47–97. [MR1895096](#) <https://doi.org/10.1103/RevModPhys.74.47>
ARIAS-CASTRO, E., BUBECK, S. and LUGOSI, G. (2015). Detecting positive correlations in a multivariate sample. *Bernoulli* **21** 209–241. [MR3322317](#) <https://doi.org/10.3150/13-BEJ565>
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>
BAI, Z. and SARANADASA, H. (1996). Effect of high dimension: By an example of a two sample problem. *Statist. Sinica* **6** 311–329. [MR1399305](#)
BANDEIRA, A. S. and VAN HANDEL, R. (2016). Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *Ann. Probab.* **44** 2479–2506. [MR3531673](#) <https://doi.org/10.1214/15-AOP1025>
BARAUD, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli* **8** 577–606. [MR1935648](#)
BERGER, N., BORGS, C., CHAYES, J. T. and SABERI, A. (2005). On the spread of viruses on the Internet. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms* 301–310. ACM, New York. [MR2298278](#)

- BERTHET, Q. and RIGOLLET, P. (2013). Optimal detection of sparse principal components in high dimension. *Ann. Statist.* **41** 1780–1815. [MR3127849](#) <https://doi.org/10.1214/13-AOS1127>
- BICKEL, P. J. and SARKAR, P. (2016). Hypothesis testing for automated community detection in networks. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 253–273. [MR3453655](#) <https://doi.org/10.1111/rssb.12117>
- BOLLOBÁS, B., JANSON, S. and RIORDAN, O. (2007). The phase transition in inhomogeneous random graphs. *Random Structures Algorithms* **31** 3–122. [MR2337396](#) <https://doi.org/10.1002/rsa.20168>
- BRESLER, G. and NAGARAJ, D. (2018). Optimal single sample tests for structured versus unstructured network data. In *Conference on Learning Theory*. *PMLR* **75** 1657–1690.
- BUBECK, S., DING, J., ELDAN, R. and RÁCZ, M. Z. (2016). Testing for high-dimensional geometry in random graphs. *Random Structures Algorithms* **49** 503–532. [MR3545825](#) <https://doi.org/10.1002/rsa.20633>
- CAI, T. T., LIU, W. and XIA, Y. (2014). Two-sample test of high dimensional means under dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 349–372. [MR3164870](#) <https://doi.org/10.1111/rssb.12034>
- CARPENTIER, A. and NICKL, R. (2015). On signal detection and confidence sets for low rank inference problems. *Electron. J. Stat.* **9** 2675–2688. [MR3432430](#) <https://doi.org/10.1214/15-EJS1087>
- CHAN, S.-O., DIAKONIKOLAS, I., VALIANT, G. and VALIANT, P. (2014). Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms* 1193–1203. ACM, New York. [MR3376448](#) <https://doi.org/10.1137/1.9781611973402.88>
- CHEN, S. X. and QIN, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.* **38** 808–835. [MR2604697](#) <https://doi.org/10.1214/09-AOS716>
- DASKALAKIS, C., DIKKALA, N. and KAMATH, G. (2018). Testing Ising models. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms* 1989–2007. SIAM, Philadelphia, PA. [MR3775918](#) <https://doi.org/10.1137/1.9781611975031.130>
- 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.
- DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. [MR2065195](#) <https://doi.org/10.1214/009053604000000265>
- DONOHO, D. and JIN, J. (2015). Higher criticism for large-scale inference, especially for rare and weak effects. *Statist. Sci.* **30** 1–25. [MR3317751](#) <https://doi.org/10.1214/14-STS506>
- GAO, C. and LAFFERTY, J. (2017). Testing network structure using relations between small subgraph probabilities. Preprint. Available at [arXiv:1704.06742](#).
- GHOSHDASTIDAR, D. and VON LUXBURG, U. (2018). Practical methods for graph two-sample testing. In *Advances in Neural Information Processing Systems* **31**.
- GHOSHDASTIDAR, D., GUTZEIT, M., CARPENTIER, A. and VON LUXBURG, U. (2017). Two-sample tests for large random graphs using network statistics. In *Conference on Learning Theory*. *PMLR* **65** 954–977.
- GHOSHDASTIDAR, D., GUTZEIT, M., CARPENTIER, A. and VON LUXBURG, U. (2020). Supplement to “Two-sample hypothesis testing for inhomogeneous random graphs.” <https://doi.org/10.1214/19-AOS1884SUPP>.
- GINESTET, C. E., LI, J., BALACHANDRAN, P., ROSENBERG, S. and KOLACZYK, E. D. (2017). Hypothesis testing for network data in functional neuroimaging. *Ann. Appl. Stat.* **11** 725–750. [MR3693544](#) <https://doi.org/10.1214/16-AOAS1015>
- GRETTON, A., BORGHARDT, K. M., RASCH, M. J., SCHÖLKOPF, B. and SMOLA, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.* **13** 723–773. [MR2913716](#)
- HYDUKE, D. R., LEWIS, N. E. and PALSSON, B. (2013). Analysis of omics data with genome-scale models of metabolism. *Mol. Biosyst.* **9** 167–174.
- INGSTER, YU. I. and SUSLINA, I. A. (2003). *Nonparametric Goodness-of-Fit Testing Under Gaussian Models. Lecture Notes in Statistics* **169**. Springer, New York. [MR1991446](#) <https://doi.org/10.1007/978-0-387-21580-8>
- KLOPP, O., TSYBAKOV, A. B. and VERZELEN, N. (2017). Oracle inequalities for network models and sparse graphon estimation. *Ann. Statist.* **45** 316–354. [MR3611494](#) <https://doi.org/10.1214/16-AOS1454>
- KONDOR, R. and PAN, H. (2016). The multiscale Laplacian graph kernel. In *Advances in Neural Information Processing Systems*.
- LANDMAN, B. A., HUANG, A. J., GIFFORD, A., VIKRAM, D. S., LIM, I. A., FARRELL, J. A., BOGOVIC, J. A., HUA, J., CHEN, M. et al. (2011). Multi-parametric neuroimaging reproducibility: A 3-T resource study. *NeuroImage* **54** 2854–2866.
- 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>
- LEDOIT, O. and WOLF, M. (2002). Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Ann. Statist.* **30** 1081–1102. [MR1926169](#) <https://doi.org/10.1214/aos/1031689018>
- LEI, J. (2016). A goodness-of-fit test for stochastic block models. *Ann. Statist.* **44** 401–424. [MR3449773](#) <https://doi.org/10.1214/15-AOS1370>

- LOVÁSZ, L. (2012). *Large Networks and Graph Limits*. American Mathematical Society Colloquium Publications **60**. Am. Math. Soc., Providence, RI. [MR3012035](#) <https://doi.org/10.1090/coll/060>
- LU, L. and PENG, X. (2013). Spectra of edge-independent random graphs. *Electron. J. Combin.* **20** Paper 27. [MR3158266](#)
- MUKHERJEE, R., PILLAI, N. S. and LIN, X. (2015). Hypothesis testing for high-dimensional sparse binary regression. *Ann. Statist.* **43** 352–381. [MR3311863](#) <https://doi.org/10.1214/14-AOS1279>
- OLIVEIRA, R. I. (2009). Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. Preprint. Available at [arXiv:0911.0600](#).
- RAMDAS, A., REDDI, S., POZOS, B., SINGH, A. and WASSERMAN, L. (2015). Adaptivity and computation-statistics tradeoffs for kernel and distance based high dimensional two sample testing. Preprint. Available at [arXiv:1508.00655](#).
- RYABKO, D. (2017). Hypotheses testing on infinite random graphs. In *International Conference on Algorithmic Learning Theory. Proc. Mach. Learn. Res. (PMLR)* **76** 1–12. Proceedings of Machine Learning Research PMLR. [MR3776703](#)
- SHERVASHIDZE, N., SCHWEITZER, P., VAN LEEUWEN, E. J., MEHLHORN, K. and BORGWARDT, K. M. (2011). Weisfeiler–Lehman graph kernels. *J. Mach. Learn. Res.* **12** 2539–2561. [MR2845672](#)
- STAM, C. J., JONES, B. F., NOLTE, G., BREAKSPEAR, M. and SCHELTENS, P. (2007). Small-world networks and functional connectivity in Alzheimer’s disease. *Cereb. Cortex* **17** 92–99.
- SUN, X. and NOBEL, A. B. (2008). On the size and recovery of submatrices of ones in a random binary matrix. *J. Mach. Learn. Res.* **9** 2431–2453. [MR2460888](#)
- TANG, M., ATHREYA, A., SUSSMAN, D. L., LYZINSKI, V., PARK, Y. and PRIEBE, C. E. (2017a). A semi-parametric two-sample hypothesis testing problem for random graphs. *J. Comput. Graph. Statist.* **26** 344–354. [MR3640191](#) <https://doi.org/10.1080/10618600.2016.1193505>
- TANG, M., ATHREYA, A., SUSSMAN, D. L., LYZINSKI, V. and PRIEBE, C. E. (2017b). A nonparametric two-sample hypothesis testing problem for random graphs. *Bernoulli* **23** 1599–1630. [MR3624872](#) <https://doi.org/10.3150/15-BEJ789>
- TROPP, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.* **12** 389–434. [MR2946459](#) <https://doi.org/10.1007/s10208-011-9099-z>
- VERZELEN, N. and ARIAS-CASTRO, E. (2015). Community detection in sparse random networks. *Ann. Appl. Probab.* **25** 3465–3510. [MR3404642](#) <https://doi.org/10.1214/14-AAP1080>
- VERZELEN, N. and ARIAS-CASTRO, E. (2017). Detection and feature selection in sparse mixture models. *Ann. Statist.* **45** 1920–1950. [MR3718157](#) <https://doi.org/10.1214/16-AOS1513>
- VISHWANATHAN, S. V. N., SCHRAUDOLPH, N. N., KONDOR, R. and BORGWARDT, K. M. (2010). Graph kernels. *J. Mach. Learn. Res.* **11** 1201–1242. [MR2645450](#) <https://doi.org/10.1093/chemse/bjq147>

BEYOND HC: MORE SENSITIVE TESTS FOR RARE/WEAK ALTERNATIVES

BY THOMAS PORTER¹ AND MICHAEL STEWART²

In memory of Peter Hall

¹*School of Mathematics and Statistics, University of Melbourne, thomas.porter@outlook.com.au*

²*School of Mathematics and Statistics F07, University of Sydney, michael.stewart@sydney.edu.au*

Higher criticism (HC) is a popular method for large-scale inference problems based on identifying unusually high proportions of small p -values. It has been shown to enjoy a lower-order optimality property in a simple normal location mixture model which is shared by the ‘tailor-made’ parametric generalised likelihood ratio test (GLRT) for the same model; however, HC has also been shown to perform well outside this ‘narrow’ model.

We develop a higher-order framework for analysing the power of these and similar procedures, which reveals the perhaps unsurprising fact that the GLRT enjoys an edge in power over HC for the normal location mixture model. We also identify a similar parametric mixture model to which HC is similarly ‘tailor-made’ and show that the situation is (at least partly) reversed there. We also show that in the normal location mixture model a procedure based on the empirical moment-generating function enjoys the same local power properties as the GLRT and may be recommended as an easy to implement (and interpret), complementary procedure to HC. Some other practical advice regarding the implementation of these procedures is provided. Finally, we provide some simulation results to help interpret our theoretical findings.

REFERENCES

- ARIAS-CASTRO, E. and WANG, M. (2017). Distribution-free tests for sparse heterogeneous mixtures. *TEST* **26** 71–94. [MR3613606](#) <https://doi.org/10.1007/s11749-016-0499-x>
- BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. [MR1325392](#)
- BERK, R. H. and JONES, D. H. (1979). Goodness-of-fit test statistics that dominate the Kolmogorov statistics. *Z. Wahrsch. Verw. Gebiete* **47** 47–59. [MR0521531](#) <https://doi.org/10.1007/BF00533250>
- BICKEL, P. and CHERNOFF, H. (1993). Asymptotic distribution of the likelihood ratio statistic in a prototypical non regular problem. In *Statistics and Probability: A Raghu Raj Bahadur Festschrift* 83–96. Wiley, New York.
- BICKEL, P. J. and LEVINA, E. (2004). Some theory of Fisher’s linear discriminant function, ‘naive Bayes’, and some alternatives when there are many more variables than observations. *Bernoulli* **10** 989–1010. [MR2108040](#) <https://doi.org/10.3150/bj/1106314847>
- CAI, T. T., JENG, X. J. and JIN, J. (2011). Optimal detection of heterogeneous and heteroscedastic mixtures. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 629–662. [MR2867452](#) <https://doi.org/10.1111/j.1467-9868.2011.00778.x>
- CAI, T. T. and WU, Y. (2014). Optimal detection of sparse mixtures against a given null distribution. *IEEE Trans. Inform. Theory* **60** 2217–2232. [MR3181520](#) <https://doi.org/10.1109/TIT.2014.2304295>
- CHERNOFF, H. (1954). On the distribution of the likelihood ratio. *Ann. Math. Stat.* **25** 573–578. [MR0065087](#) <https://doi.org/10.1214/aoms/1177728725>
- CSÖRGŐ, M., CSÖRGŐ, S., HORVÁTH, L. and MASON, D. M. (1986). Weighted empirical and quantile processes. *Ann. Probab.* **14** 31–85. [MR0815960](#)
- DARLING, D. A. and ERDÖS, P. (1956). A limit theorem for the maximum of normalized sums of independent random variables. *Duke Math. J.* **23** 143–155. [MR0074712](#)

MSC2020 subject classifications. Primary 62F03; secondary 62G30, 62G32, 62F05.

Key words and phrases. Higher criticism, sparse normal means, phi-divergence, multiple comparisons, mixture model.

- DONOHO, D. (2017). 50 years of data science. *J. Comput. Graph. Statist.* **26** 745–766. [MR3765335](https://doi.org/10.1080/10618600.2017.1384734) <https://doi.org/10.1080/10618600.2017.1384734>
- DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. [MR2065195](https://doi.org/10.1214/009053604000000265) <https://doi.org/10.1214/009053604000000265>
- DONOHO, D. and JIN, J. (2008). Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proc. Natl. Acad. Sci. USA* **105** 14790–14795.
- DONOHO, D. and JIN, J. (2015). Higher criticism for large-scale inference, especially for rare and weak effects. *Statist. Sci.* **30** 1–25. [MR3317751](https://doi.org/10.1214/14-STSS06) <https://doi.org/10.1214/14-STSS06>
- GONTSCHARUK, V. and FINNER, H. (2017). Asymptotics of goodness-of-fit tests based on minimum p -value statistics. *Comm. Statist. Theory Methods* **46** 2332–2342. [MR3576717](https://doi.org/10.1080/03610926.2015.1041985) <https://doi.org/10.1080/03610926.2015.1041985>
- GONTSCHARUK, V., LANDWEHR, S. and FINNER, H. (2015). The intermediates take it all: Asymptotics of higher criticism statistics and a powerful alternative based on equal local levels. *Biom. J.* **57** 159–180. [MR3298224](https://doi.org/10.1002/bimj.201300255) <https://doi.org/10.1002/bimj.201300255>
- HALL, P. (1991). On convergence rates of suprema. *Probab. Theory Related Fields* **89** 447–455. [MR1118558](https://doi.org/10.1007/BF01199788) <https://doi.org/10.1007/BF01199788>
- HALL, P. and JIN, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist.* **38** 1686–1732. [MR2662357](https://doi.org/10.1214/09-AOS764) <https://doi.org/10.1214/09-AOS764>
- HARTIGAN, J. A. (1985). A failure of likelihood asymptotics for normal mixtures. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. II* (Berkeley, Calif., 1983). Wadsworth Statist./Probab. Ser. 807–810. Wadsworth, Belmont, CA. [MR0822066](https://doi.org/10.1080/03610926.2015.1041985)
- INGSTER, Y. I. (1997). Some problems of hypothesis testing leading to infinitely divisible distributions. *Math. Methods Statist.* **6** 47–69. [MR1456646](https://doi.org/10.1080/03610926.2015.1041985)
- INGSTER, Y. I. (2001). Adaptive detection of a signal of growing dimension. I. *Math. Methods Statist.* **10** 395–421. [MR1887340](https://doi.org/10.1080/03610926.2015.1041985)
- INGSTER, Y. I. (2002). Adaptive detection of a signal of growing dimension. II. *Math. Methods Statist.* **11** 37–68. [MR1900973](https://doi.org/10.1080/03610926.2015.1041985)
- JAESCHKE, D. (1979). The asymptotic distribution of the supremum of the standardized empirical distribution function on subintervals. *Ann. Statist.* **7** 108–115. [MR0515687](https://doi.org/10.1080/03610926.2015.1041985)
- JAGER, L. and WELLNER, J. A. (2007). Goodness-of-fit tests via phi-divergences. *Ann. Statist.* **35** 2018–2053. [MR2363962](https://doi.org/10.1214/0009053607000000244) <https://doi.org/10.1214/0009053607000000244>
- JIN, J. and KE, Z. T. (2016). Rare and weak effects in large-scale inference: Methods and phase diagrams. *Statist. Sinica* **26** 1–34. [MR3468343](https://doi.org/10.1080/03610926.2015.1041985)
- LAURENT, B., MARTEAU, C. and MAUGIS-RABUSSEAU, C. (2016). Non-asymptotic detection of two-component mixtures with unknown means. *Bernoulli* **22** 242–274. [MR3449782](https://doi.org/10.3150/14-BEJ657) <https://doi.org/10.3150/14-BEJ657>
- LEADBETTER, M. R. and ROOTZÉN, H. (1988). Extremal theory for stochastic processes. *Ann. Probab.* **16** 431–478. [MR0929071](https://doi.org/10.1080/03610926.2015.1041985)
- LI, J. and SIEGMUND, D. (2015). Higher criticism: p -values and criticism. *Ann. Statist.* **43** 1323–1350. [MR3346705](https://doi.org/10.1214/15-AOS1312) <https://doi.org/10.1214/15-AOS1312>
- LIU, X. and SHAO, Y. (2004). Asymptotics for the likelihood ratio test in a two-component normal mixture model. *J. Statist. Plann. Inference* **123** 61–81. [MR2058122](https://doi.org/10.1016/S0378-3758(03)00138-1) [https://doi.org/10.1016/S0378-3758\(03\)00138-1](https://doi.org/10.1016/S0378-3758(03)00138-1)
- MOSCOVICH, A., NADLER, B. and SPIEGELMAN, C. (2016). On the exact Berk–Jones statistics and their p -value calculation. *Electron. J. Stat.* **10** 2329–2354. [MR3544289](https://doi.org/10.1214/16-EJS1172) <https://doi.org/10.1214/16-EJS1172>
- PORTER, T. and STEWART, M. (2020). Supplement to “Beyond HC: More sensitive tests for rare/weak alternatives.” <https://doi.org/10.1214/19-AOS1885SUPP>.
- SHORACK, G. R. and WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. Wiley, New York. [MR0838963](https://doi.org/10.1080/03610926.2015.1041985)
- STEPANOVA, N. and PAVLENKO, T. (2018). Goodness-of-fit tests based on sup-functionals of weighted empirical processes. *Teor. Veroyatn. Primen.* **63** 358–388. [MR3796493](https://doi.org/10.4213/tvp5160) <https://doi.org/10.4213/tvp5160>
- TUKEY, J. W. (1976). T13 N: The higher criticism. Technical report.
- TUKEY, J. W. (1989). Higher criticism for individual significances in several tables or parts of tables. Technical report.
- TUKEY, J. W. (1994). The problem of multiple comparisons. In *The Collected Works of John W. Tukey. Vol. VIII* lxii+475+i10. CRC Press, New York.
- WALTHER, G. (2013). The average likelihood ratio for large-scale multiple testing and detecting sparse mixtures. In *From Probability to Statistics and Back: High-Dimensional Models and Processes*. Inst. Math. Stat. (IMS) Collect. **9** 317–326. IMS, Beachwood, OH. [MR3202643](https://doi.org/10.1214/12-IMSCOLL923) <https://doi.org/10.1214/12-IMSCOLL923>

WELLNER, J. A. and KOLTCHINSKII, V. (2003). A note on the asymptotic distribution of Berk–Jones type statistics under the null hypothesis. In *High Dimensional Probability, III* (Sandjberg, 2002). *Progress in Probability* **55** 321–332. Birkhäuser, Basel. [MR2033896](#)

MINIMAX OPTIMAL RATES FOR MONDRIAN TREES AND FORESTS

BY JAOUAD MOURTADA^{1,*}, STÉPHANE GAÏFFAS² AND ERWAN SCORNET^{1,**}

¹CMAP, École polytechnique, *jaouad.mourtada@polytechnique.edu; **erwan.scornet@polytechnique.edu

²LPMA—Université Paris Diderot, stephane.gaiffas@lpsm.paris

Introduced by Breiman (*Mach. Learn.* **45** (2001) 5–32), Random Forests are widely used classification and regression algorithms. While being initially designed as batch algorithms, several variants have been proposed to handle online learning. One particular instance of such forests is the *Mondrian forest* (In *Adv. Neural Inf. Process. Syst.* (2014) 3140–3148; In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)* (2016)), whose trees are built using the so-called Mondrian process, therefore allowing to easily update their construction in a streaming fashion. In this paper we provide a thorough theoretical study of Mondrian forests in a batch learning setting, based on new results about Mondrian partitions. Our results include consistency and convergence rates for Mondrian trees and forests, that turn out to be minimax optimal on the set of s -Hölder function with $s \in (0, 1]$ (for trees and forests) and $s \in (1, 2]$ (for forests only), assuming a proper tuning of their complexity parameter in both cases. Furthermore, we prove that an adaptive procedure (to the unknown $s \in (0, 2]$) can be constructed by combining Mondrian forests with a standard model aggregation algorithm. These results are the first demonstrating that some particular random forests achieve minimax rates *in arbitrary dimension*. Owing to their remarkably simple distributional properties, which lead to minimax rates, Mondrian trees are a promising basis for more sophisticated yet theoretically sound random forests variants.

REFERENCES

- [1] ARLOT, S. (2008). V-fold cross-validation improved: V-fold penalization. ArXiv preprint. Available at [arXiv:0802.0566](https://arxiv.org/abs/0802.0566).
- [2] ARLOT, S. and GENUER, R. (2014). Analysis of purely random forests bias. ArXiv preprint. Available at [arXiv:1407.3939](https://arxiv.org/abs/1407.3939).
- [3] ATHEY, S., TIBSHIRANI, J. and WAGER, S. (2019). Generalized random forests. *Ann. Statist.* **47** 1148–1178. [MR3909963 https://doi.org/10.1214/18-AOS1709](https://doi.org/10.1214/18-AOS1709)
- [4] AUDIBERT, J.-Y. (2008). Progressive mixture rules are deviation suboptimal. In *Adv. Neural Inf. Process. Syst.* **20** 41–48.
- [5] BIAU, G. (2012). Analysis of a random forests model. *J. Mach. Learn. Res.* **13** 1063–1095. [MR2930634](https://doi.org/10.1214/12-AOS1063)
- [6] BIAU, G., DEVROYE, L. and LUGOSI, G. (2008). Consistency of random forests and other averaging classifiers. *J. Mach. Learn. Res.* **9** 2015–2033. [MR2447310](https://doi.org/10.1214/08-AOS1063)
- [7] BIAU, G. and SCORNET, E. (2016). A random forest guided tour. *TEST* **25** 197–227. [MR3493512 https://doi.org/10.1007/s11749-016-0481-7](https://doi.org/10.1007/s11749-016-0481-7)
- [8] BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.
- [9] BREIMAN, L. (2010). Some infinity theory for predictor ensembles. Technical Report 577, Statistics Department, Univ. California Berkeley.
- [10] CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4** 266–298. [MR2758172 https://doi.org/10.1214/09-AOAS285](https://doi.org/10.1214/09-AOAS285)
- [11] CLÉMENÇON, S., DEPECKER, M. and VAYATIS, N. (2013). Ranking forests. *J. Mach. Learn. Res.* **14** 39–73. [MR3033325](https://doi.org/10.1214/12-AOS1063)
- [12] CUI, Y., ZHU, R., ZHOU, M. and KOSOROK, M. (2017). Some asymptotic results of survival tree and forest models. ArXiv preprint. Available at [arXiv:1707.09631](https://arxiv.org/abs/1707.09631).

- [13] DENIL, M., MATHESON, D. and DE FREITAS, N. (2013). Consistency of online random forests. In *Proceedings of the 30th Annual International Conference on Machine Learning (ICML)* 1256–1264.
- [14] DENIL, M., MATHESON, D. and DE FREITAS, N. (2014). Narrowing the gap: Random forests in theory and in practice. In *Proceedings of the 31st Annual International Conference on Machine Learning (ICML)* 665–673.
- [15] DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition. Applications of Mathematics (New York)* **31**. Springer, New York. MR1383093 <https://doi.org/10.1007/978-1-4612-0711-5>
- [16] DOMINGOS, P. and HULTEN, G. (2000). Mining high-speed data streams. In *Proceedings of the Conference on Knowledge Discovery and Data Mining (KDD)* 71–80.
- [17] FERNÁNDEZ-DELGADO, M., CERNADAS, E., BARRO, S. and AMORIM, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *J. Mach. Learn. Res.* **15** 3133–3181. MR3277155
- [18] GENUER, R. (2012). Variance reduction in purely random forests. *J. Nonparametr. Stat.* **24** 543–562. MR2968888 <https://doi.org/10.1080/10485252.2012.677843>
- [19] GEURTS, P., ERNST, D. and WEHENKEL, L. (2006). Extremely randomized trees. *Mach. Learn.* **63** 3–42.
- [20] 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>
- [21] ISHWARAN, H., KOGALUR, U. B., BLACKSTONE, E. H. and LAUER, M. S. (2008). Random survival forests. *Ann. Appl. Stat.* **2** 841–860. MR2516796 <https://doi.org/10.1214/08-AOAS169>
- [22] KLUSOWSKI, J. M. (2018). Complete analysis of a random forest model. ArXiv preprint. Available at [arXiv:1805.02587](https://arxiv.org/abs/1805.02587).
- [23] LAKSHMINARAYANAN, B., ROY, D. M. and TEH, Y. W. (2014). Mondrian forests: Efficient online random forests. In *Adv. Neural Inf. Process. Syst.* **27** 3140–3148.
- [24] LAKSHMINARAYANAN, B., ROY, D. M. and TEH, Y. W. (2016). Mondrian forests for large-scale regression when uncertainty matters. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*.
- [25] MEINSHAUSEN, N. (2006). Quantile regression forests. *J. Mach. Learn. Res.* **7** 983–999. MR2274394
- [26] MENTCH, L. and HOOKER, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *J. Mach. Learn. Res.* **17** 26. MR3491120
- [27] MENZE, B. H., KELM, B. M., SPLITTHOFF, D. N., KOETHE, U. and HAMPRECHT, F. A. (2011). On oblique random forests. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 453–469. Springer.
- [28] MOURTADA, J., GAÏFFAS, S. and SCORNET, E. (2017). Universal consistency and minimax rates for online Mondrian forests. In *Adv. Neural Inf. Process. Syst.* **30** 3759–3768.
- [29] MOURTADA, J., GAÏFFAS, S. and SCORNET, E. (2020). Supplement to “Minimax optimal rates for Mondrian trees and forests.” <https://doi.org/10.1214/19-AOS1886SUPP>.
- [30] NEMIROVSKI, A. (2000). Topics in non-parametric statistics. In *Lectures on Probability Theory and Statistics (Saint-Flour, 1998). Lecture Notes in Math.* **1738** 85–277. Springer, Berlin. MR1775640
- [31] ORBANZ, P. and ROY, D. M. (2015). Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE Trans. Pattern Anal. Mach. Intell.* **37** 437–461.
- [32] ROY, D. M. (2011). Computability, inference and modeling in probabilistic programming. Ph.D. thesis, Massachusetts Institute of Technology.
- [33] ROY, D. M. and TEH, Y. W. (2009). The Mondrian process. In *Adv. Neural Inf. Process. Syst.* **21** 1377–1384.
- [34] SAFFARI, A., LEISTNER, C., SANTNER, J., GODEC, M. and BISCHOF, H. (2009). On-line random forests. In *3rd IEEE ICCV Workshop on On-Line Computer Vision*.
- [35] SCORNET, E., BIAU, G. and VERT, J.-P. (2015). Consistency of random forests. *Ann. Statist.* **43** 1716–1741. MR3357876 <https://doi.org/10.1214/15-AOS1321>
- [36] STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053. MR0673642
- [37] TADDY, M. A., GRAMACY, R. B. and POLSON, N. G. (2011). Dynamic trees for learning and design. *J. Amer. Statist. Assoc.* **106** 109–123. MR2816706 <https://doi.org/10.1198/jasa.2011.ap09769>
- [38] WAGER, S. and ATHEY, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* **113** 1228–1242. MR3862353 <https://doi.org/10.1080/01621459.2017.1319839>
- [39] WAGER, S. and WALTHER, G. (2015). Adaptive concentration of regression trees, with application to random forests. ArXiv preprint. Available at [arXiv:1503.06388](https://arxiv.org/abs/1503.06388).
- [40] WASSERMAN, L. (2006). *All of Nonparametric Statistics. Springer Texts in Statistics*. Springer, New York. MR2172729

- [41] YANG, Y. (1999). Minimax nonparametric classification. I. Rates of convergence. *IEEE Trans. Inform. Theory* **45** 2271–2284. [MR1725115](#) <https://doi.org/10.1109/18.796368>

IDENTIFIABILITY OF NONPARAMETRIC MIXTURE MODELS AND BAYES OPTIMAL CLUSTERING

BY BRYON ARAGAM¹, CHEN DAN^{2,*}, ERIC P. XING^{2,**} AND PRADEEP RAVIKUMAR^{2,†}

¹Booth School of Business, University of Chicago, bryon@chicagobooth.edu

²Machine Learning Department, School of Computer Science, Carnegie Mellon University, ^{*}cdan@andrew.cmu.edu; ^{**}epxing@cs.cmu.edu; [†]pradeep@cs.cmu.edu

Motivated by problems in data clustering, we establish general conditions under which families of nonparametric mixture models are identifiable by introducing a novel framework involving clustering overfitted *parametric* (i.e., misspecified) mixture models. These identifiability conditions generalize existing conditions in the literature and are flexible enough to include, for example, mixtures of infinite Gaussian mixtures. In contrast to the recent literature, we allow for general nonparametric mixture components and instead impose regularity assumptions on the underlying mixing measure. As our primary application we apply these results to partition-based clustering, generalizing the notion of a Bayes optimal partition from classical parametric model-based clustering to nonparametric settings. Furthermore, this framework is constructive, so that it yields a practical algorithm for learning identified mixtures, which is illustrated through several examples on real data. The key conceptual device in the analysis is the convex, metric geometry of probability measures on metric spaces and its connection to the Wasserstein convergence of mixing measures. The result is a flexible framework for nonparametric clustering with formal consistency guarantees.

REFERENCES

- [1] ACHLIOPTAS, D. and MCSHERRY, F. (2005). On spectral learning of mixtures of distributions. In *International Conference on Computational Learning Theory* 458–469. Springer, Berlin.
- [2] AHMAD, K. E. and AL-HUSSAINI, E. K. (1982). Remarks on the nonidentifiability of mixtures of distributions. *Ann. Inst. Statist. Math.* **34** 543–544. [MR0695073 https://doi.org/10.1007/BF02481052](https://doi.org/10.1007/BF02481052)
- [3] AIROLDI, E. M., COSTA, T. B. and CHAN, S. H. (2013). Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems* 692–700.
- [4] ALLMAN, E. S., MATIAS, C. and RHODES, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.* **37** 3099–3132. [MR2549554 https://doi.org/10.1214/09-AOS689](https://doi.org/10.1214/09-AOS689)
- [5] ANANDKUMAR, A., HSU, D., JANZAMIN, M. and KAKADE, S. (2015). When are overcomplete topic models identifiable? Uniqueness of tensor Tucker decompositions with structured sparsity. *J. Mach. Learn. Res.* **16** 2643–2694. [MR3450520](https://doi.org/10.1214/15-AOS689)
- [6] ANANDKUMAR, A., HSU, D., JAVANMARD, A. and KAKADE, S. (2013). Learning linear Bayesian networks with latent variables. In *Proceedings of the 30th International Conference on Machine Learning* 249–257.
- [7] ARAGAM, B., DAN, C., XING, E. P. and RAVIKUMAR, P. (2020). Supplement to “Identifiability of nonparametric mixture models and Bayes optimal clustering.” <https://doi.org/10.1214/19-AOS1887SUPP>.
- [8] ARORA, S. and KANNAN, R. (2005). Learning mixtures of separated nonspherical Gaussians. *Ann. Appl. Probab.* **15** 69–92. [MR2115036 https://doi.org/10.1214/1050516040000000512](https://doi.org/10.1214/1050516040000000512)
- [9] BARNDORFF-NIELSEN, O. (1965). Identifiability of mixtures of exponential families. *J. Math. Anal. Appl.* **12** 115–121. [MR0183058 https://doi.org/10.1016/0022-247X\(65\)90059-4](https://doi.org/10.1016/0022-247X(65)90059-4)
- [10] BASU, A., SHIOYA, H. and PARK, C. (2011). *Statistical Inference: The Minimum Distance Approach. Monographs on Statistics and Applied Probability* **120**. CRC Press, Boca Raton, FL. [MR2830561](https://doi.org/10.1214/1050516040000000512)

MSC2020 subject classifications. Primary 62G05, 62H30; secondary 62H12.

Key words and phrases. Mixture models, nonparametric statistics, identifiability, clustering, Bayes optimal partition.

- [11] BERAN, R. (1977). Minimum Hellinger distance estimates for parametric models. *Ann. Statist.* **5** 445–463. [MR0448700](#)
- [12] BOCK, H. H. (1996). Probabilistic models in cluster analysis. *Comput. Statist. Data Anal.* **23** 5–28.
- [13] BONHOMME, S., JOCHMANS, K. and ROBIN, J.-M. (2016). Estimating multivariate latent-structure models. *Ann. Statist.* **44** 540–563. [MR3476609](#) <https://doi.org/10.1214/15-AOS1376>
- [14] BONHOMME, S., JOCHMANS, K. and ROBIN, J.-M. (2016). Non-parametric estimation of finite mixtures from repeated measurements. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 211–229. [MR3453653](#) <https://doi.org/10.1111/rssb.12110>
- [15] BORDES, L., KOJADINOVIC, I. and VANDEKERKHOVE, P. (2013). Semiparametric estimation of a two-component mixture of linear regressions in which one component is known. *Electron. J. Stat.* **7** 2603–2644. [MR3121625](#) <https://doi.org/10.1214/13-EJS858>
- [16] BORDES, L., MOTTELET, S. and VANDEKERKHOVE, P. (2006). Semiparametric estimation of a two-component mixture model. *Ann. Statist.* **34** 1204–1232. [MR2278356](#) <https://doi.org/10.1214/009053606000000353>
- [17] BRUNI, C. and KOCH, G. (1985). Identifiability of continuous mixtures of unknown Gaussian distributions. *Ann. Probab.* **13** 1341–1357. [MR0806230](#)
- [18] CHACÓN, J. E. (2015). A population background for nonparametric density-based clustering. *Statist. Sci.* **30** 518–532. [MR3432839](#) <https://doi.org/10.1214/15-STS526>
- [19] CHAUDHURI, K. and DASGUPTA, S. (2010). Rates of convergence for the cluster tree. In *Advances in Neural Information Processing Systems* 343–351.
- [20] CHEN, J. H. (1995). Optimal rate of convergence for finite mixture models. *Ann. Statist.* **23** 221–233. [MR1331665](#) <https://doi.org/10.1214/aos/1176324464>
- [21] CHEN, Y.-C., GENOVESE, C. R. and WASSERMAN, L. (2016). A comprehensive approach to mode clustering. *Electron. J. Stat.* **10** 210–241. [MR3466181](#) <https://doi.org/10.1214/15-EJS1102>
- [22] CSISZÁR, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.* **2** 299–318. [MR0219345](#)
- [23] CUEVAS, A. and GONZÁLEZ MANTEIGA, W. (1991). Data-driven smoothing based on convexity properties. In *Nonparametric Functional Estimation and Related Topics (Spetses, 1990)*. NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. **335** 225–240. Kluwer Academic, Dordrecht. [MR1154331](#)
- [24] D’HAULTFÈUILLE, X. and FÉVRIER, P. (2015). Identification of mixture models using support variations. *J. Econometrics* **189** 70–82. [MR3397353](#) <https://doi.org/10.1016/j.jeconom.2015.06.022>
- [25] DASGUPTA, S. (1999). Learning mixtures of Gaussians. In *40th Annual Symposium on Foundations of Computer Science (New York, 1999)* 634–644. IEEE Computer Soc., Los Alamitos, CA. [MR1917603](#) <https://doi.org/10.1109/SFFCS.1999.814639>
- [26] DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition. Applications of Mathematics (New York)* **31**. Springer, New York. [MR1383093](#) <https://doi.org/10.1007/978-1-4612-0711-5>
- [27] ELDRIDGE, J., BELKIN, M. and WANG, Y. (2016). Graphons, mergeons, and so on! In *Advances in Neural Information Processing Systems* 2307–2315.
- [28] FRALEY, C. and RAFTERY, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.* **97** 611–631. [MR1951635](#) <https://doi.org/10.1198/016214502760047131>
- [29] GASSIAT, E. and ROUSSEAU, J. (2013). Non parametric finite translation mixtures with dependent regime. Preprint. Available at [arXiv:1302.2345](#).
- [30] 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>
- [31] 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>
- [32] HALL, P., NEEMAN, A., PAKYARI, R. and ELMORE, R. (2005). Nonparametric inference in multivariate mixtures. *Biometrika* **92** 667–678. [MR2202653](#) <https://doi.org/10.1093/biomet/92.3.667>
- [33] HALL, P. and ZHOU, X.-H. (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Ann. Statist.* **31** 201–224. [MR1962504](#) <https://doi.org/10.1214/aos/1046294462>
- [34] HARTIGAN, J. A. (1975). *Clustering Algorithms*. Wiley, New York. [MR0405726](#)
- [35] HARTIGAN, J. A. (1981). Consistency of single linkage for high-density clusters. *J. Amer. Statist. Assoc.* **76** 388–394. [MR0624340](#)
- [36] 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>

- [37] HETTMANSPERGER, T. P. and THOMAS, H. (2000). Almost nonparametric inference for repeated measures in mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 811–825. MR1796294 <https://doi.org/10.1111/1467-9868.00266>
- [38] HO, N. and NGUYEN, X. (2016). On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electron. J. Stat.* **10** 271–307. MR3466183 <https://doi.org/10.1214/16-EJS1105>
- [39] HO, N. and NGUYEN, X. (2019). Singularity structures and impacts on parameter estimation in finite mixtures of distributions. *SIAM J. Math. Data Sci.* **1** 730–758. MR4016133 <https://doi.org/10.1137/18M122947X>
- [40] 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](https://doi.org/10.1016/0378-8733(83)90021-7)
- [41] HOLZMANN, H., MUNK, A. and GNEITING, T. (2006). Identifiability of finite mixtures of elliptical distributions. *Scand. J. Stat.* **33** 753–763. MR2300914 <https://doi.org/10.1111/j.1467-9469.2006.00505.x>
- [42] HUNTER, D. R., WANG, S. and HETTMANSPERGER, T. P. (2007). Inference for mixtures of symmetric distributions. *Ann. Statist.* **35** 224–251. MR2332275 <https://doi.org/10.1214/009053606000001118>
- [43] HUNTER, D. R. and YOUNG, D. S. (2012). Semiparametric mixtures of regressions. *J. Nonparametr. Stat.* **24** 19–38. MR2885823 <https://doi.org/10.1080/10485252.2011.608430>
- [44] JACOBS, R. A., JORDAN, M. I., NOWLAN, S. J. and HINTON, G. E. (1991). Adaptive mixtures of local experts. *Neural Comput.* **3** 79–87.
- [45] JOCHMANS, K., HENRY, M. and SALANIÉ, B. (2017). Inference on two-component mixtures under tail restrictions. *Econometric Theory* **33** 610–635. MR3637969 <https://doi.org/10.1017/S0266466616000098>
- [46] KANNAN, R., SALMASIAN, H. and VEMPALA, S. (2008). The spectral method for general mixture models. *SIAM J. Comput.* **38** 1141–1156. MR2421081 <https://doi.org/10.1137/S0097539704445925>
- [47] LEVINE, M., HUNTER, D. R. and CHAUVEAU, D. (2011). Maximum smoothed likelihood for multivariate mixtures. *Biometrika* **98** 403–416. MR2806437 <https://doi.org/10.1093/biomet/asq079>
- [48] LI, J. and SCHMIDT, L. (2015). A nearly optimal and agnostic algorithm for properly learning a mixture of k Gaussians, for any constant k . Preprint. Available at [arXiv:1506.01367](https://arxiv.org/abs/1506.01367).
- [49] LINDSAY, B. G. (1995). Mixture models: Theory, geometry and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*. JSTOR.
- [50] LLOYD, S. P. (1982). Least squares quantization in PCM. *IEEE Trans. Inform. Theory* **28** 129–137. MR0651807 <https://doi.org/10.1109/TIT.1982.1056489>
- [51] MACQUEEN, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66)* 281–297. Univ. California Press, Berkeley, CA. MR0214227
- [52] MIXON, D. G., VILLAR, S. and WARD, R. (2017). Clustering subgaussian mixtures by semidefinite programming. *Inf. Inference* **6** 389–415. MR3764529 <https://doi.org/10.1093/imaiai/iax001>
- [53] NG, A. Y., JORDAN, M. I. and WEISS, Y. (2001). On spectral clustering: Analysis and an algorithm. In *NIPS* **14** 849–856.
- [54] NGUYEN, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *Ann. Statist.* **41** 370–400. MR3059422 <https://doi.org/10.1214/12-AOS1065>
- [55] PARTHASARATHY, K. R. (1967). *Probability Measures on Metric Spaces. Probability and Mathematical Statistics, No. 3*. Academic Press, New York. MR0226684
- [56] PRIEBE, C. E. (1994). Adaptive mixtures. *J. Amer. Statist. Assoc.* **89** 796–806. MR1294725
- [57] RINALDO, A. and WASSERMAN, L. (2010). Generalized density clustering. *Ann. Statist.* **38** 2678–2722. MR2722453 <https://doi.org/10.1214/10-AOS797>
- [58] RITTER, G. (2015). *Robust Cluster Analysis and Variable Selection. Monographs on Statistics and Applied Probability* **137**. CRC Press, Boca Raton, FL. MR3241693
- [59] 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>
- [60] SCHIEBINGER, G., WAINWRIGHT, M. J. and YU, B. (2015). The geometry of kernelized spectral clustering. *Ann. Statist.* **43** 819–846. MR3325711 <https://doi.org/10.1214/14-AOS1283>
- [61] SHI, T., BELKIN, M. and YU, B. (2009). Data spectroscopy: Eigenspaces of convolution operators and clustering. *Ann. Statist.* **37** 3960–3984. MR2572449 <https://doi.org/10.1214/09-AOS700>
- [62] SRIPERUMBUDUR, B. and STEINWART, I. (2012). Consistency and rates for clustering with dbscan. In *Artificial Intelligence and Statistics* 1090–1098.
- [63] STEINHAUS, H. (1956). Sur la division des corps matériels en parties. *Bull. Acad. Polon. Sci. Cl. III.* **4** 801–804. MR0090073
- [64] STEINWART, I. (2011). Adaptive density level set clustering. In *International Conference on Learning Theory* 703–738.

- [65] STEINWART, I. (2015). Fully adaptive density-based clustering. *Ann. Statist.* **43** 2132–2167. [MR3396981](#)
<https://doi.org/10.1214/15-AOS1331>
- [66] SWEETING, T. J. (1986). On a converse to Scheffé’s theorem. *Ann. Statist.* **14** 1252–1256. [MR0856821](#)
<https://doi.org/10.1214/aos/1176350065>
- [67] TANG, J., MENG, Z., NGUYEN, X., MEI, Q. and ZHANG, M. (2014). Understanding the limiting factors of topic modeling via posterior contraction analysis. In *International Conference on Machine Learning* 190–198.
- [68] TEICHER, H. (1960). On the mixture of distributions. *Ann. Math. Stat.* **31** 55–73. [MR0121825](#)
<https://doi.org/10.1214/aoms/1177705987>
- [69] TEICHER, H. (1961). Identifiability of mixtures. *Ann. Math. Stat.* **32** 244–248. [MR0120677](#)
<https://doi.org/10.1214/aoms/1177705155>
- [70] TEICHER, H. (1963). Identifiability of finite mixtures. *Ann. Math. Stat.* **34** 1265–1269. [MR0155376](#)
<https://doi.org/10.1214/aoms/1177703862>
- [71] TEICHER, H. (1967). Identifiability of mixtures of product measures. *Ann. Math. Stat.* **38** 1300–1302. [MR0216635](#) <https://doi.org/10.1214/aoms/1177698805>
- [72] THOMANN, P., STEINWART, I. and SCHMID, N. (2015). Towards an axiomatic approach to hierarchical clustering of measures. *J. Mach. Learn. Res.* **16** 1949–2002. [MR3417804](#)
- [73] TITTERINGTON, D. M., SMITH, A. F. M. and MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. *Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. Wiley, Chichester. [MR0838090](#)
- [74] ULTSCH, A. (2005). Clustering with SOM: U*C. In *Proc. Workshop on Self-Organizing Maps* 75–82, Paris, France.
- [75] WOLFE, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivar. Behav. Res.* **5** 329–350.
- [76] YAKOWITZ, S. J. and SPRAGINS, J. D. (1968). On the identifiability of finite mixtures. *Ann. Math. Stat.* **39** 209–214. [MR0224204](#) <https://doi.org/10.1214/aoms/1177698520>
- [77] YAN, D., HUANG, L. and JORDAN, M. I. (2009). Fast approximate spectral clustering. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 907–916. ACM, New York.

A TEST FOR SEPARABILITY IN COVARIANCE OPERATORS OF RANDOM SURFACES

BY PRAMITA BAGCHI¹ AND HOLGER DETTE²

¹*Department of Statistics, George Mason University*

²*Fakultät für Mathematik, Ruhr-Universität Bochum, holger.dette@rub.de*

The assumption of separability is a simplifying and very popular assumption in the analysis of spatiotemporal or hypersurface data structures. It is often made in situations where the covariance structure cannot be easily estimated, for example, because of a small sample size or because of computational storage problems. In this paper we propose a new and very simple test to validate this assumption. Our approach is based on a measure of separability which is zero in the case of separability and positive otherwise. We derive the asymptotic distribution of a corresponding estimate under the null hypothesis and the alternative and develop an asymptotic and a bootstrap test which are very easy to implement. In particular, our approach does neither require projections on subspaces generated by the eigenfunctions of the covariance operator nor distributional assumptions as recently used by (*Ann. Statist.* **45** (2017) 1431–1461) and (*Biometrika* **104** 425–437) to construct tests for separability. We investigate the finite sample performance by means of a simulation study and also provide a comparison with the currently available methodology. Finally, the new procedure is illustrated analyzing a data example.

REFERENCES

- [1] ASTON, J. A. D., PIGOLI, D. and TAVAKOLI, S. (2017). Tests for separability in nonparametric covariance operators of random surfaces. *Ann. Statist.* **45** 1431–1461. [MR3670184](#) <https://doi.org/10.1214/16-AOS1495>
- [2] BAGCHI, P. and DETTE, H. (2020). Supplement to “A test for separability in covariance operators of random surfaces.” <https://doi.org/10.1214/19-AOS1888SUPP>.
- [3] BAR-HEN, A., BEL, L. and CHEDDADI, R. (2008). Spatio-temporal functional regression on paleoecological data. In *Functional and Operatorial Statistics. Contrib. Statist.* 53–56. Physica-Verlag/Springer, Heidelberg. [MR2490328](#)
- [4] CONSTANTINOU, P., KOKOSZKA, P. and REIMHERR, M. (2017). Testing separability of space-time functional processes. *Biometrika* **104** 425–437. [MR3698263](#) <https://doi.org/10.1093/biomet/asx013>
- [5] CONSTANTINOU, P., KOKOSZKA, P. and REIMHERR, M. (2018). Testing separability of functional time series. *J. Time Series Anal.* **39** 731–747. [MR3849524](#) <https://doi.org/10.1111/jtsa.12302>
- [6] DAUXOIS, J., POUSSE, A. and ROMAIN, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *J. Multivariate Anal.* **12** 136–154. [MR0650934](#) [https://doi.org/10.1016/0047-259X\(82\)90088-4](https://doi.org/10.1016/0047-259X(82)90088-4)
- [7] DUNFORD, N. and SCHWARTZ, J. T. (1988). *Linear Operators, Vols. I, II, III*. Wiley Classics Library. Wiley, New York.
- [8] FUENTES, M. (2006). Testing for separability of spatial-temporal covariance functions. *J. Statist. Plann. Inference* **136** 447–466. [MR2211349](#) <https://doi.org/10.1016/j.jspi.2004.07.004>
- [9] GENTON, M. G. (2007). Separable approximations of space-time covariance matrices. *Environmetrics* **18** 681–695. [MR2408938](#) <https://doi.org/10.1002/env.854>
- [10] GNEITING, T. (2002). Nonseparable, stationary covariance functions for space-time data. *J. Amer. Statist. Assoc.* **97** 590–600. [MR1941475](#) <https://doi.org/10.1198/016214502760047113>

MSC2020 subject classifications. 62G10, 62G20.

Key words and phrases. Functional data, minimum distance, separability, space-time processes, surface data structures.

- [11] GNEITING, T., GENTON, M. G. and GUTTORP, P. (2007). Geostatistical space-time models, stationarity, separability and full symmetry. In *Statistical Methods for Spatio-Temporal Systems* (B. Finkenstadt, L. Held and V. Isham, eds.). *Monographs on Statistics and Applied Probability* **107**. CRC Press/CRC, Boca Raton, FL. [MR2307967](#)
- [12] GOHBERG, I., GOLDBERG, S. and KAASHOEK, M. A. (1990). *Classes of Linear Operators, Vol. I. Operator Theory: Advances and Applications* **49**. Birkhäuser, Basel.
- [13] 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>
- [14] HUIZENGA, H. M., DE MUNCK, J. C., WALDORF, L. J. and GRASMAN, R. P. P. (2002). Spatiotemporal EEG/MEG source analysis based on a parametric noise covariance model. *IEEE Trans. Biomed. Eng.* **49** 533–539.
- [15] LU, N. and ZIMMERMAN, D. L. (2005). The likelihood ratio test for a separable covariance matrix. *Statist. Probab. Lett.* **73** 449–457. [MR2187860](#) <https://doi.org/10.1016/j.spl.2005.04.020>
- [16] LYNCH, B. and CHEN, K. (2018). A test of weak separability for multi-way functional data, with application to brain connectivity studies. *Biometrika* **105** 815–831. [MR3877867](#) <https://doi.org/10.1093/biomet/asv048>
- [17] MITCHELL, M. W., GENTON, M. G. and GUMPERTZ, M. L. (2005). Testing for separability of space-time covariances. *Environmetrics* **16** 819–831. [MR2216653](#) <https://doi.org/10.1002/env.737>
- [18] MITCHELL, M. W., GENTON, M. G. and GUMPERTZ, M. L. (2006). A likelihood ratio test for separability of covariances. *J. Multivariate Anal.* **97** 1025–1043. [MR2276147](#) <https://doi.org/10.1016/j.jmva.2005.07.005>
- [19] RABINER, L. R. and SCHAFER, R. W. (1978). *Digital Processing of Speech Signals* 100. Prentice-Hall, Englewood Cliffs, NJ.
- [20] ROUGIER, J. (2017). A representation theorem for stochastic processes with separable covariance functions, and its implications for emulation. Preprint. Available at [arXiv:1702.05599](#).
- [21] SKUP, M. (2010). Longitudinal fMRI analysis: A review of methods. *Stat. Interface* **3** 235–252. [MR2659514](#) <https://doi.org/10.4310/SII.2010.v3.n2.a10>
- [22] TAVAKOLI, S. (2016). covsep: Tests for determining if the covariance structure of 2-dimensional data is separable. R package version 1.0.0.
- [23] VAN DER VAART, A. W. (2000). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. [MR1652247](#) <https://doi.org/10.1017/CBO9780511802256>
- [24] WEIDMANN, J. (1980). *Linear Operators in Hilbert Spaces*. *Graduate Texts in Mathematics* **68**. Springer, New York–Berlin. [MR0566954](#)
- [25] WORSLEY, K. J., MARRETT, S., NEELIN, P., VANDAL, A. C., FRISTON, K. J. and EVANS, A. C. (1996). A unified statistical approach for determining significant signals in images of cerebral activation. *Hum. Brain Mapp.* **4** 58–73.

A GENERAL APPROACH FOR CURE MODELS IN SURVIVAL ANALYSIS

BY VALENTIN PATILEA¹ AND INGRID VAN KEILEGOM²

¹CREST (UMR 9194), Ensai, Univ Rennes, patilea@ensai.fr

²ORSTAT, KU Leuven, ingrid.vankeilegom@kuleuven.be

In survival analysis it often happens that some subjects under study do not experience the event of interest; they are considered to be “cured.” The population is thus a mixture of two subpopulations, one of cured subjects and one of “susceptible” subjects. We propose a novel approach to estimate a mixture cure model when covariates are present and the lifetime is subject to random right censoring. We work with a parametric model for the cure proportion, while the conditional survival function of the uncured subjects is unspecified. The approach is based on an inversion which allows us to write the survival function as a function of the distribution of the observable variables. This leads to a very general class of models which allows a flexible and rich modeling of the conditional survival function. We show the identifiability of the proposed model as well as the consistency and the asymptotic normality of the model parameters. We also consider in more detail the case where kernel estimators are used for the nonparametric part of the model. The new estimators are compared with the estimators from a Cox mixture cure model via simulations. Finally, we apply the new model on a medical data set.

REFERENCES

- AKRITAS, M. G. and VAN KEILEGOM, I. (2001). Non-parametric estimation of the residual distribution. *Scand. J. Stat.* **28** 549–567. [MR1858417](#) <https://doi.org/10.1111/1467-9469.00254>
- AMICO, M. and VAN KEILEGOM, I. (2018). Cure models in survival analysis. *Annu. Rev. Stat. Appl.* **5** 311–345. [MR3774750](#) <https://doi.org/10.1146/annurev-statistics-031017-100101>
- AMICO, M., VAN KEILEGOM, I. and LEGRAND, C. (2019). The single-index/Cox mixture cure model. *Biometrics* **75** 452–462.
- BOAG, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J. Roy. Statist. Soc.—Series B* **11** 15–53.
- CHEN, X., LINTON, O. and VAN KEILEGOM, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica* **71** 1591–1608. [MR2000259](#) <https://doi.org/10.1111/1468-0262.00461>
- FANG, H.-B., LI, G. and SUN, J. (2005). Maximum likelihood estimation in a semiparametric logistic/proportional-hazards mixture model. *Scand. J. Statist.* **32** 59–75. [MR2136802](#) <https://doi.org/10.1111/j.1467-9469.2005.00415.x>
- FAREWELL, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* **38** 1041–1046.
- GILL, R. D. (1994). *Lectures on Survival Analysis. Lectures on Probability Theory: Ecole D’été de Probabilités de Saint-Flour XXII. Lecture Notes in Math.* **1581**. Springer, Berlin.
- GILL, R. D. and JOHANSEN, S. (1990). A survey of product-integration with a view toward application in survival analysis. *Ann. Statist.* **18** 1501–1555. [MR1074422](#) <https://doi.org/10.1214/aos/1176347865>
- KUK, A. Y. C. and CHEN, C.-H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika* **79**, 531–541.
- LI, Q., LIN, J. and RACINE, J. S. (2013). Optimal bandwidth selection for nonparametric conditional distribution and quantile functions. *J. Bus. Econom. Statist.* **31** 57–65. [MR3030798](#) <https://doi.org/10.1080/07350015.2012.738955>
- LOPEZ, O. (2011). Nonparametric estimation of the multivariate distribution function in a censored regression model with applications. *Comm. Statist. Theory Methods* **40** 2639–2660. [MR2860770](#) <https://doi.org/10.1080/03610926.2010.489175>

MSC2020 subject classifications. Primary 62N01, 62N02; secondary 62F12, 62G05.

Key words and phrases. Asymptotic normality, bootstrap, kernel smoothing, logistic regression, mixture cure model, semiparametric model.

- LÓPEZ-CHEDA, A., CAO, R., JÁCOME, M. A. and VAN KEILEGOM, I. (2017). Nonparametric incidence estimation and bootstrap bandwidth selection in mixture cure models. *Comput. Statist. Data Anal.* **105** 144–165. [MR3552194 https://doi.org/10.1016/j.csda.2016.08.002](https://doi.org/10.1016/j.csda.2016.08.002)
- LU, W. (2008). Maximum likelihood estimation in the proportional hazards cure model. *Ann. Inst. Statist. Math.* **60** 545–574. [MR2434411 https://doi.org/10.1007/s10463-007-0120-x](https://doi.org/10.1007/s10463-007-0120-x)
- MEEKER, W. Q. JR. (1987). Limited failure population life tests: Application to integrated circuit reliability. *Technometrics* **29** 51–65. [MR0876885 https://doi.org/10.2307/1269883](https://doi.org/10.2307/1269883)
- OTHUS, M., LI, Y. and TIWARI, R. C. (2009). A class of semiparametric mixture cure survival models with dependent censoring. *J. Amer. Statist. Assoc.* **104** 1241–1250. [MR2750247 https://doi.org/10.1198/jasa.2009.tm08033](https://doi.org/10.1198/jasa.2009.tm08033)
- PATILEA, V. and VAN KEILEGOM, I. (2020). Supplement to “A general approach for cure models in survival analysis.” <https://doi.org/10.1214/19-AOS1889SUPP>.
- PENG, Y. and TAYLOR, J. M. G. (2014). Cure models In *Handbook of Survival Analysis* (Klein, J. P., van Houwelingen, H. C., Ibrahim, J. G. and Scheike, T. H., eds.). *Chapman and Hall/CRC Handbooks of Modern Statistical Methods*. CRC Press, Boca Raton, FL. [MR3287588](https://doi.org/10.1080/10816132.2014.900000)
- SCHMIDT, P. and WITTE, A. D. (1989). Predicting criminal recidivism using split population survival time models. *J. Econometrics* **40** 141–159.
- STONE, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595–645. [MR0443204](https://doi.org/10.1214/aos/1176330304)
- SY, J. P. and TAYLOR, J. M. G. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics* **56** 227–236. [MR1767631 https://doi.org/10.1111/j.0006-341X.2000.00227.x](https://doi.org/10.1111/j.0006-341X.2000.00227.x)
- TAYLOR, J. M. G. (1995). Semi-parametric estimation in failure time mixture models. *Biometrics* **51** 899–907.
- 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](https://doi.org/10.1017/CBO9780511802256)
- WANG, Y., KLIJN, J. G. M., SIEUWERTS, A. M., LOOK, M. P., YANG, F., TALANTOV, D., TIMMERMANS, M., MEIJET-VAN GELDER, M. E. M., YU, J. et al. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet* **365** 671–679.
- XU, J. and PENG, Y. (2014). Nonparametric cure rate estimation with covariates. *Canad. J. Statist.* **42** 1–17. [MR3181580 https://doi.org/10.1002/cjs.11197](https://doi.org/10.1002/cjs.11197)

ADAPTIVE DISTRIBUTED METHODS UNDER COMMUNICATION CONSTRAINTS

BY BOTOND SZABÓ¹ AND HARRY VAN ZANTEN²

¹Mathematical Institute, Leiden University, b.t.szabo@math.leidenuniv.nl

²Department of Mathematics, Faculty of Sciences, VU Amsterdam, j.h.van.zanten@vu.nl

We study estimation methods under communication constraints in a distributed version of the nonparametric random design regression model. We derive minimax lower bounds and exhibit methods that attain those bounds. Moreover, we show that adaptive estimation is possible in this setting.

REFERENCES

- [1] 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>
- [2] CARPENTIER, A. (2013). Honest and adaptive confidence sets in L_p . *Electron. J. Stat.* **7** 2875–2923. [MR3148371](#) <https://doi.org/10.1214/13-EJS867>
- [3] CHAGNY, G. (2013). Penalization versus Goldenshluger–Lepski strategies in warped bases regression. *ESAIM Probab. Stat.* **17** 328–358. [MR3066383](#) <https://doi.org/10.1051/ps/2011165>
- [4] DEISENROTH, M. P. and NG, J. W. (2015). Distributed Gaussian processes. ArXiv E-prints.
- [5] DUCHI, J. C. and WAINWRIGHT, M. J. (2013). Distance-based and continuum Fano inequalities with applications to statistical estimation. ArXiv E-prints.
- [6] GERSHGORIN, S. A. (1931). Über die abgrenzung der eigenwerte einer matrix. *Bulletin de L'Académie des Sciences de L'URSS. Classe des Sciences Mathématiques et na* **6** 749–754.
- [7] GINÉ, E. and NICKL, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics **40**. Cambridge Univ. Press, New York. [MR3588285](#) <https://doi.org/10.1017/CBO9781107337862>
- [8] GUHANIYOGI, R., LI, C., SAVITSKY, T. D. and SRIVASTAVA, S. (2017). A divide-and-conquer Bayesian approach to large-scale kriging. ArXiv E-prints.
- [9] HÄRDLE, W., KERKYACHARIAN, G., PICARD, D. and TSYBAKOV, A. (2012). *Wavelets, Approximation, and Statistical Applications. Lecture Notes in Statistics*. Springer, New York. [MR1618204](#) <https://doi.org/10.1007/978-1-4612-2222-4>
- [10] KLEINER, A., TALWALKAR, A., SARKAR, P. and JORDAN, M. I. (2014). A scalable bootstrap for massive data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 795–816. [MR3248677](#) <https://doi.org/10.1111/rssb.12050>
- [11] LACOUR, C. (2008). Adaptive estimation of the transition density of a particular hidden Markov chain. *J. Multivariate Anal.* **99** 787–814. [MR2405092](#) <https://doi.org/10.1016/j.jmva.2007.04.006>
- [12] LEE, J. D., LIU, Q., SUN, Y. and TAYLOR, J. E. (2017). Communication-efficient sparse regression. *J. Mach. Learn. Res.* **18** Paper No. 5, 30. [MR3625709](#)
- [13] MASSART, P. (2007). *Concentration Inequalities and Model Selection. Lecture Notes in Math.* **1896**. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard. [MR2319879](#)
- [14] ROSENBLATT, J. D. and NADLER, B. (2016). On the optimality of averaging in distributed statistical learning. *Inf. Inference* **5** 379–404. [MR3609865](#) <https://doi.org/10.1093/imaiai/iaw013>
- [15] SCOTT, S. L., BLOCKER, A. W., BONASSI, F. V., CHIPMAN, H., GEORGE, E. and MCCULLOCH, R. (2013). Bayes and big data: The consensus Monte Carlo algorithm. In *EFaBBayes 250 Conference* **16**.
- [16] SHANG, Z. and CHENG, G. (2015). A Bayesian splitotoc theory for nonparametric models. ArXiv E-prints.
- [17] SRIVASTAVA, S., CEVHER, V., TRAN-DINH, Q., DUNSON and WASP, D. B. (2015). Scalable Bayes via barycenters of subset posteriors. In *AISTATS*.

MSC2020 subject classifications. Primary 62G20, 62G08; secondary 62B10.

Key words and phrases. Distributed computation, minimax rates, adaptation, nonparametric regression, communication constraints.

- [18] SZABÓ, B. and VAN ZANTEN, H. (2020). Supplement to “Adaptive distributed methods under communication constraints.” <https://doi.org/10.1214/19-AOS1890SUPP>.
- [19] SZABÓ, B. and VAN ZANTEN, H. (2019). An asymptotic analysis of distributed nonparametric methods. *J. Mach. Learn. Res.* **20** Paper No. 87, 30. [MR3960941](#)
- [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>
- [21] WANG, J., CHEN, J. and WU, X. (2010). On the sum rate of Gaussian multiterminal source coding: New proofs and results. *IEEE Trans. Inform. Theory* **56** 3946–3960. [MR2752477](#) <https://doi.org/10.1109/TIT.2010.2050960>
- [22] YANG, Y. and BARRON, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.* **27** 1564–1599. [MR1742500](#) <https://doi.org/10.1214/aos/1017939142>
- [23] ZHANG, Y., DUCHI, J., JORDAN, M. I. and WAINWRIGHT, M. J. (2013). Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *Advances in Neural Information Processing Systems* 2328–2336.
- [24] ZHANG, Y., WAINWRIGHT, M. J. and DUCHI, J. C. (2012). Communication-efficient algorithms for statistical optimization. In *Advances in Neural Information Processing Systems* 25 1502–1510 (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.). Curran Associates, Inc.
- [25] ZHU, Y. and LAFFERTY, J. (2018). Quantized minimax estimation over Sobolev ellipsoids. *Inf. Inference* **7** 31–82. [MR3801518](#) <https://doi.org/10.1093/imaiai/iax007>
- [26] ZHU, Y. and LAFFERTY, J. (2018). Distributed nonparametric regression under communication constraints. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholm, Sweden, July 10–15, 2018* 6004–6012.

BAYESIAN ANALYSIS OF THE COVARIANCE MATRIX OF A MULTIVARIATE NORMAL DISTRIBUTION WITH A NEW CLASS OF PRIORS

BY JAMES O. BERGER¹, DONGCHU SUN² AND CHENGYUAN SONG^{*3}

¹*Department of Statistical Science, Duke University, berger@stat.duke.edu*

²*Department of Statistics, University of Nebraska-Lincoln, sund9@unl.edu*

³*School of Statistics, East China Normal University, songchengyuanchina@163.com*

Bayesian analysis for the covariance matrix of a multivariate normal distribution has received a lot of attention in the last two decades. In this paper, we propose a new class of priors for the covariance matrix, including both inverse Wishart and reference priors as special cases. The main motivation for the new class is to have available priors—both subjective and objective—that do not “force eigenvalues apart,” which is a criticism of inverse Wishart and Jeffreys priors. Extensive comparison of these “shrinkage priors” with inverse Wishart and Jeffreys priors is undertaken, with the new priors seeming to have considerably better performance. A number of curious facts about the new priors are also observed, such as that the posterior distribution will be proper with just three vector observations from the multivariate normal distribution—regardless of the dimension of the covariance matrix—and that useful inference about features of the covariance matrix can be possible. Finally, a new MCMC algorithm is developed for this class of priors and is shown to be computationally effective for matrices of up to 100 dimensions.

REFERENCES

- [1] ANDERSON, T. W., OLKIN, I. and UNDERHILL, L. G. (1987). Generation of random orthogonal matrices. *SIAM J. Sci. Statist. Comput.* **8** 625–629. [MR0892309](#) <https://doi.org/10.1137/0908055>
- [2] BERGER, J. O., STRAWDERMAN, W. and TANG, D. (2005). Posterior propriety and admissibility of hyperpriors in normal hierarchical models. *Ann. Statist.* **33** 606–646. [MR2163154](#) <https://doi.org/10.1214/009053605000000075>
- [3] BERGER, J. O. and SUN, D. (2008). Objective priors for the bivariate normal model. *Ann. Statist.* **36** 963–982. [MR2396821](#) <https://doi.org/10.1214/07-AOS501>
- [4] BERGER, J. O., SUN, D. and SONG, C. (2020). An objective prior for hyperparameters in normal hierarchical models. *J. Multivariate Anal.* To appear.
- [5] BERGER, J. O., SUN, D. and SONG, C. (2020). Supplement to “Bayesian analysis of the covariance matrix of a multivariate normal distribution with a new class of priors.” <https://doi.org/10.1214/19-AOS1891SUPP>.
- [6] CHEN, M.-H. and SCHMEISER, B. (1993). Performance of the Gibbs, hit-and-run, and Metropolis samplers. *J. Comput. Graph. Statist.* **2** 251–272. [MR1272394](#) <https://doi.org/10.2307/1390645>
- [7] DANIELS, M. J. and KASS, R. E. (1999). Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. *J. Amer. Statist. Assoc.* **94** 1254–1263. [MR1731487](#) <https://doi.org/10.2307/2669939>
- [8] DANIELS, M. J. and KASS, R. E. (2001). Shrinkage estimators for covariance matrices. *Biometrics* **57** 1173–1184. [MR1950425](#) <https://doi.org/10.1111/j.0006-341X.2001.01173.x>
- [9] DEY, D. K. and SRINIVASAN, C. (1985). Estimation of a covariance matrix under Stein’s loss. *Ann. Statist.* **13** 1581–1591. [MR0811511](#) <https://doi.org/10.1214/aos/1176349756>
- [10] EATON, M. L. and OLKIN, I. (1987). Best equivariant estimators of a Cholesky decomposition. *Ann. Statist.* **15** 1639–1650. [MR0913579](#) <https://doi.org/10.1214/aos/1176350615>
- [11] EGUCHI, N., SAITO, R., SAEKI, T., NAKATSUKA, Y., BELIKOV, D. and MAKSYUTOV, S. (2010). A priori covariance estimation for CO₂ and CH₄ retrievals. *J. Geophys. Res.* **115** Art. ID D10215.

- [12] FARRELL, R. H. (1985). *Multivariate Calculation: Use of the Continuous Groups*. Springer Series in Statistics. Springer, New York. [MR0770934](#) <https://doi.org/10.1007/978-1-4613-8528-8>
- [13] FREI, M. and KUNSCH, H. R. (2012). Sequential state and observation noise covariance estimation using combined ensemble Kalman and particle filters. *Mon. Weather Rev.* **140** 1476–1495.
- [14] GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.
- [15] GUILLOT, D., RAJARATNAM, B. and EMILE-GEAY, J. (2015). Statistical paleoclimate reconstructions via Markov random fields. *Ann. Appl. Stat.* **9** 324–352. [MR3341118](#) <https://doi.org/10.1214/14-AOAS794>
- [16] HAFF, L. R. (1979). Estimation of the inverse covariance matrix: Random mixtures of the inverse Wishart matrix and the identity. *Ann. Statist.* **7** 1264–1276. [MR0550149](#)
- [17] HAFF, L. R. (1991). The variational form of certain Bayes estimators. *Ann. Statist.* **19** 1163–1190. [MR1126320](#) <https://doi.org/10.1214/aos/1176348244>
- [18] HAMIMECHE, S. and LEWIS, A. (2009). Properties and use of CMB power spectrum likelihoods. *Phys. Rev. D* **79** Art. ID 83012.
- [19] HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109. [MR3363437](#) <https://doi.org/10.1093/biomet/57.1.97>
- [20] HOFF, P. D. (2009). A hierarchical eigenmodel for pooled covariance estimation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 971–992. [MR2750253](#) <https://doi.org/10.1111/j.1467-9868.2009.00716.x>
- [21] HOFF, P. D. (2009). Simulation of the matrix Bingham–von Mises–Fisher distribution, with applications to multivariate and relational data. *J. Comput. Graph. Statist.* **18** 438–456. [MR2749840](#) <https://doi.org/10.1198/jcgs.2009.07177>
- [22] JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Clarendon Press, Oxford. [MR0187257](#)
- [23] LEDOIT, O. and WOLF, M. (2004). Honey, I shrunk the sample covariance matrix. *J. Portf. Manag.* **4** 110–119.
- [24] LEDOIT, O. and WOLF, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.* **88** 365–411. [MR2026339](#) [https://doi.org/10.1016/S0047-259X\(03\)00096-4](https://doi.org/10.1016/S0047-259X(03)00096-4)
- [25] LIN, S. P. and PERLMAN, M. D. (1985). A Monte Carlo comparison of four estimators of a covariance matrix. In *Multivariate Analysis VI* (Pittsburgh, Pa., 1983) 411–429. North-Holland, Amsterdam. [MR0822310](#)
- [26] POPE, A. C. and SZAPUDI, I. (2005). Shrinkage estimation of the power spectrum covariance matrix. *Mon. Not. R. Astron. Soc.* **389** 766–774.
- [27] POURAHMADI, M. (2011). Covariance estimation: The GLM and regularization perspectives. *Statist. Sci.* **26** 369–387. [MR2917961](#) <https://doi.org/10.1214/11-STS358>
- [28] PRESS, S. J. (2012). *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference*. Courier Corporation, North Chelmsford, MA.
- [29] RAJARATNAM, B., MASSAM, H. and CARVALHO, C. M. (2008). Flexible covariance estimation in graphical Gaussian models. *Ann. Statist.* **36** 2818–2849. [MR2485014](#) <https://doi.org/10.1214/08-AOS619>
- [30] SCHÄFER, J. and STRIMMER, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* **4** Art. ID 32. [MR2183942](#) <https://doi.org/10.2202/1544-6115.1175>
- [31] SINHA, B. K. and GHOSH, M. (1987). Inadmissibility of the best equivariant estimators of the variance–covariance matrix, the precision matrix, and the generalized variance under entropy loss. *Statist. Decisions* **5** 201–227. [MR0905238](#)
- [32] STEIN, C. (1956). Some problems in multivariate analysis. Part I. Technical Report 6, Dept. Statistics, Stanford Univ.
- [33] STEIN, C. (1975). Estimation of a covariance matrix, Rietz Lecture. In *39th Annual Meeting IMS, Atlanta, GA*.
- [34] SUN, D. and BERGER, J. O. (2007). Objective Bayesian analysis for the multivariate normal model. In *Bayesian Statistics 8*. Oxford Sci. Publ. 525–562. Oxford Univ. Press, Oxford. [MR2433206](#)
- [35] YANG, R. and BERGER, J. O. (1994). Estimation of a covariance matrix using the reference prior. *Ann. Statist.* **22** 1195–1211. [MR1311972](#) <https://doi.org/10.1214/aos/1176325625>

EXTENDING THE VALIDITY OF FREQUENCY DOMAIN BOOTSTRAP METHODS TO GENERAL STATIONARY PROCESSES

BY MARCO MEYER^{1,*}, EFSTATHIOS PAPARODITIS² AND JENS-PETER KREISS^{1,**}

¹*Inst. f. Math. Stochastik, Technische Universität Braunschweig, *marco.meyer@tu-bs.de; **j.kreiss@tu-bs.de*

²*Department of Mathematics and Statistics, University of Cyprus, stathisp@ucy.ac.cy*

Existing frequency domain methods for bootstrapping time series have a limited range. Essentially, these procedures cover the case of linear time series with independent innovations, and some even require the time series to be Gaussian. In this paper we propose a new frequency domain bootstrap method—the hybrid periodogram bootstrap (HPB)—which is consistent for a much wider range of stationary, even nonlinear, processes and which can be applied to a large class of periodogram-based statistics. The HPB is designed to combine desirable features of different frequency domain techniques while overcoming their respective limitations. It is capable to imitate the weak dependence structure of the periodogram by invoking the concept of convolved subsampling in a novel way that is tailor-made for periodograms. We show consistency for the HPB procedure for a general class of stationary time series, ranging clearly beyond linear processes, and for spectral means and ratio statistics on which we mainly focus. The finite sample performance of the new bootstrap procedure is illustrated via simulations.

REFERENCES

- AUESTAD, B. and TJØSTHEIM, D. (1990). Identification of nonlinear time series: First order characterization and order determination. *Biometrika* **77** 669–687. [MR1086681 https://doi.org/10.1093/biomet/77.4.669](https://doi.org/10.1093/biomet/77.4.669)
- BARTLETT, M. S. (1948). Smoothing periodograms from time-series with continuous spectra. *Nature* **161** 686–687.
- BARTLETT, M. S. (1950). Periodogram analysis and continuous spectra. *Biometrika* **37** 1–16. [MR0035934 https://doi.org/10.1093/biomet/37.1-2.1](https://doi.org/10.1093/biomet/37.1-2.1)
- BILLINGSLEY, P. (1995). *Probability and Measure*. Wiley, New York.
- BROCKWELL, P. J. and DAVIS, R. A. (1991). *Time Series: Theory and Methods*, 2nd ed. Springer, New York.
- DAHLHAUS, R. (1985). On a spectral density estimate obtained by averaging periodograms. *J. Appl. Probab.* **22** 598–610. [MR0799283 https://doi.org/10.1017/s0021900200029351](https://doi.org/10.1017/s0021900200029351)
- DAHLHAUS, R. and JANAS, D. (1996). A frequency domain bootstrap for ratio statistics in time series analysis. *Ann. Statist.* **24** 1934–1963. [MR1421155 https://doi.org/10.1214/aos/1069362304](https://doi.org/10.1214/aos/1069362304)
- EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26. [MR0515681 https://doi.org/10.1214/aos/1176348515](https://doi.org/10.1214/aos/1176348515)
- FRANKE, J. and HÄRDLE, W. (1992). On bootstrapping kernel spectral estimates. *Ann. Statist.* **20** 121–145. [MR1150337 https://doi.org/10.1214/aos/1176348515](https://doi.org/10.1214/aos/1176348515)
- HURVICH, C. M. and ZEGER, S. L. (1987). Frequency domain bootstrap methods for time series. Preprint, Department of Statistics and Operations Research, New York University.
- JANAS, D. and DAHLHAUS, R. (1994). A frequency domain bootstrap for time series. In *Proc. 26th Symposium of the Interface, Computationally Intensive Statistical Methods* (J. Sall and A. Lehmann, eds.) 423–425. Interface Foundation of North America, Fairfax Station, VA.
- KEENAN, D. M. (1987). Limiting behavior of functionals of higher-order sample cumulant spectra. *Ann. Statist.* **15** 134–151. [MR0885728 https://doi.org/10.1214/aos/1176350257](https://doi.org/10.1214/aos/1176350257)
- KREISS, J.-P. and PAPARODITIS, E. (2003). Autoregressive-aided periodogram bootstrap for time series. *Ann. Statist.* **31** 1923–1955. [MR2036395 https://doi.org/10.1214/aos/1074290332](https://doi.org/10.1214/aos/1074290332)
- KREISS, J.-P. and PAPARODITIS, E. (2012). The hybrid wild bootstrap for time series. *J. Amer. Statist. Assoc.* **107** 1073–1084. [MR3010895 https://doi.org/10.1080/01621459.2012.695664](https://doi.org/10.1080/01621459.2012.695664)
- KREISS, J.-P., PAPARODITIS, E. and POLITIS, D. N. (2011). On the range of validity of the autoregressive sieve bootstrap. *Ann. Statist.* **39** 2103–2130. [MR2893863 https://doi.org/10.1214/11-AOS900](https://doi.org/10.1214/11-AOS900)

- MEYER, M., PAPARODITIS, E. and KREISS, J.-P. (2020). Supplement to “Extending the validity of frequency domain bootstrap methods to general stationary processes.” <https://doi.org/10.1214/19-AOS1892SUPP>.
- PAPARODITIS, E. and POLITIS, D. N. (1999). The local bootstrap for periodogram statistics. *J. Time Series Anal.* **20** 193–222. [MR1701054 https://doi.org/10.1111/1467-9892.00133](https://doi.org/10.1111/1467-9892.00133)
- PRIESTLEY, M. B. (1981). *Spectral Analysis and Time Series. Vol. 1. Probability and Mathematical Statistics*. Academic Press, London. [MR0628735](https://doi.org/10.1007/978-1-4612-5156-9)
- ROSENBLATT, M. (1985). *Stationary Sequences and Random Fields*. Birkhäuser, Boston, MA. [MR0885090 https://doi.org/10.1007/978-1-4612-5156-9](https://doi.org/10.1007/978-1-4612-5156-9)
- SHAO, X. (2009). Confidence intervals for spectral mean and ratio statistics. *Biometrika* **96** 107–117. [MR2482138 https://doi.org/10.1093/biomet/asn067](https://doi.org/10.1093/biomet/asn067)
- SHAO, X. and WU, W. B. (2007). Asymptotic spectral theory for nonlinear time series. *Ann. Statist.* **35** 1773–1801. [MR2351105 https://doi.org/10.1214/009053606000001479](https://doi.org/10.1214/009053606000001479)
- TANIGUCHI, M. (1982). On estimation of the integrals of the fourth order cumulant spectral density. *Biometrika* **69** 117–122. [MR0655676 https://doi.org/10.1093/biomet/69.1.117](https://doi.org/10.1093/biomet/69.1.117)
- TEWES, J., POLITIS, D. N. and NORDMAN, D. J. (2019). Convolved subsampling estimation with applications to block bootstrap. *Ann. Statist.* **47** 468–496. [MR3909939 https://doi.org/10.1214/18-AOS1695](https://doi.org/10.1214/18-AOS1695)
- WELCH, P. D. (1967). The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Trans. Audio Electroacoust.* **15** 70–73.

MINIMAX ESTIMATION OF LARGE PRECISION MATRICES WITH BANDABLE CHOLESKY FACTOR

BY YU LIU* AND ZHAO REN**

*Department of Statistics, University of Pittsburgh, *yul125@pitt.edu; **zren@pitt.edu*

The last decade has witnessed significant methodological and theoretical advances in estimating large precision matrices. In particular, there are scientific applications such as longitudinal data, meteorology and spectroscopy in which the ordering of the variables can be interpreted through a bandable structure on the Cholesky factor of the precision matrix. However, the minimax theory has still been largely unknown, as opposed to the well established minimax results over the corresponding bandable covariance matrices. In this paper we focus on two commonly used types of parameter spaces and develop the optimal rates of convergence under both the operator norm and the Frobenius norm. A striking phenomenon is found. Two types of parameter spaces are fundamentally different under the operator norm but enjoy the same rate optimality under the Frobenius norm which is in sharp contrast to the equivalence of corresponding two types of bandable covariance matrices under both norms. This fundamental difference is established by carefully constructing the corresponding minimax lower bounds. Two new estimation procedures are developed. For the operator norm our optimal procedure is based on a novel local cropping estimator, targeting on all principle submatrices of the precision matrix, while for the Frobenius norm our optimal procedure relies on a delicate regression-based thresholding rule. Lepski's method is considered to achieve optimal adaptation. We further establish rate optimality in the nonparanormal model. Numerical studies are carried out to confirm our theoretical findings.

REFERENCES

- [1] ASSOUD, P. (1983). Deux remarques sur l'estimation. *C. R. Acad. Sci. Paris Sér. I Math.* **296** 1021–1024. [MR0777600](#)
- [2] BANERJEE, O., EL GHAOU, L. and D'ASPREMONT, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9** 485–516. [MR2417243](#)
- [3] BANERJEE, S. and GHOSAL, S. (2014). Posterior convergence rates for estimating large precision matrices using graphical models. *Electron. J. Stat.* **8** 2111–2137. [MR3273620](#) <https://doi.org/10.1214/14-EJS945>
- [4] BARBER, R. F. and KOLAR, M. (2018). ROCKET: Robust confidence intervals via Kendall's tau for transelliptical graphical models. *Ann. Statist.* **46** 3422–3450. [MR3852657](#) <https://doi.org/10.1214/17-AOS1663>
- [5] 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](#)
- [6] BICKEL, P. J. and LEVINA, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. [MR2485008](#) <https://doi.org/10.1214/08-AOS600>
- [7] 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>
- [8] BIEN, J., BUNEA, F. and XIAO, L. (2016). Convex banding of the covariance matrix. *J. Amer. Statist. Assoc.* **111** 834–845. [MR3538709](#) <https://doi.org/10.1080/01621459.2015.1058265>

MSC2020 subject classifications. Primary 62H12; secondary 62F12, 62C20, 62G09.

Key words and phrases. Optimal rate of convergence, precision matrix, local cropping, Cholesky factor, minimax lower bound, thresholding, operator norm, Frobenius norm and adaptive estimation.

- [9] CAI, T. and LIU, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.* **106** 672–684. [MR2847949](#) <https://doi.org/10.1198/jasa.2011.tm10560>
- [10] CAI, T., LIU, W. and LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. [MR2847973](#) <https://doi.org/10.1198/jasa.2011.tm10155>
- [11] CAI, T. T. (2012). Minimax and adaptive inference in nonparametric function estimation. *Statist. Sci.* **27** 31–50. [MR2953494](#) <https://doi.org/10.1214/11-STS355>
- [12] CAI, T. T., LIU, W. and ZHOU, H. H. (2016). Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *Ann. Statist.* **44** 455–488. [MR3476606](#) <https://doi.org/10.1214/13-AOS1171>
- [13] CAI, T. T., REN, Z. and ZHOU, H. H. (2013). Optimal rates of convergence for estimating Toeplitz covariance matrices. *Probab. Theory Related Fields* **156** 101–143. [MR3055254](#) <https://doi.org/10.1007/s00440-012-0422-7>
- [14] CAI, T. T., REN, Z. and ZHOU, H. H. (2016). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electron. J. Stat.* **10** 1–59. [MR3466172](#) <https://doi.org/10.1214/15-EJS1081>
- [15] CAI, T. T., ZHANG, C.-H. and ZHOU, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38** 2118–2144. [MR2676885](#) <https://doi.org/10.1214/09-AOS752>
- [16] CAI, T. T. and ZHOU, H. H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *Ann. Statist.* **40** 2389–2420. [MR3097607](#) <https://doi.org/10.1214/12-AOS998>
- [17] D’ASPROMONT, A., BANERJEE, O. and EL GHAOU, L. (2008). First-order methods for sparse covariance selection. *SIAM J. Matrix Anal. Appl.* **30** 56–66. [MR2399568](#) <https://doi.org/10.1137/060670985>
- [18] DELYON, B. and JUDITSKY, A. (1996). On minimax wavelet estimators. *Appl. Comput. Harmon. Anal.* **3** 215–228. [MR1400080](#) <https://doi.org/10.1006/acha.1996.0017>
- [19] EISEN, M. B., SPELLMAN, P. T., BROWN, P. O. and BOTSTEIN, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95** 14863–14868.
- [20] EL KAROUI, N. (2003). On the largest eigenvalue of Wishart matrices with identity covariance when n, p and p/n tend to infinity. Preprint. Available at [arXiv:math/0309355](#).
- [21] EL KAROUI, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.* **36** 2717–2756. [MR2485011](#) <https://doi.org/10.1214/07-AOS559>
- [22] FAN, J., XUE, L. and ZOU, H. (2016). Multitask quantile regression under the transnormal model. *J. Amer. Statist. Assoc.* **111** 1726–1735. [MR3601731](#) <https://doi.org/10.1080/01621459.2015.1113973>
- [23] FURRER, R. and BENGTTSSON, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *J. Multivariate Anal.* **98** 227–255. [MR2301751](#) <https://doi.org/10.1016/j.jmva.2006.08.003>
- [24] HAMILL, T. M., WHITAKER, J. S. and SNYDER, C. (2001). Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Weather Rev.* **129** 2776–2790.
- [25] HEYER, M. H. and SCHLOERB, F. P. (1997). Application of principal component analysis to large-scale spectral line imaging studies of the interstellar medium. *Astrophys. J.* **475** 173.
- [26] HU, A. and NEGAHBAN, S. (2017). Minimax estimation of bandable precision matrices. In *Advances in Neural Information Processing Systems* 4893–4901.
- [27] HUANG, J. Z., LIU, N., POURAHMADI, M. and LIU, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93** 85–98. [MR2277742](#) <https://doi.org/10.1093/biomet/93.1.85>
- [28] 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>
- [29] 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>
- [30] KE, Y., MINSKER, S., REN, Z., SUN, Q. and ZHOU, W.-X. (2019). User-friendly covariance estimation for heavy-tailed distributions. *Statist. Sci.*. To appear.
- [31] KENDALL, M. G. (1938). A new measure of rank correlation. *Biometrika* **30** 81–93.
- [32] LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. [MR2572459](#) <https://doi.org/10.1214/09-AOS720>
- [33] LEDOIT, O. and WOLF, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *J. Empir. Finance* **10** 603–621.
- [34] LEE, K. and LEE, J. (2017). Estimating large precision matrices via modified Cholesky decomposition. Preprint. Available at [arXiv:1707.01143](#).
- [35] LEPSKIĬ, O. V. (1991). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Teor. Veroyatn. Primen.* **36** 645–659. [MR1147167](#) <https://doi.org/10.1137/1136085>

- [36] LEVINA, E., ROTHMAN, A. and ZHU, J. (2008). Sparse estimation of large covariance matrices via a nested Lasso penalty. *Ann. Appl. Stat.* **2** 245–263. [MR2415602](#) <https://doi.org/10.1214/07-AOAS139>
- [37] LIU, H., LAFFERTY, J. and WASSERMAN, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.* **10** 2295–2328. [MR2563983](#)
- [38] LIU, Y. and REN, Z. (2020). Supplement to “Minimax estimation of large precision matrices with bandable Cholesky factor.” <https://doi.org/10.1214/19-AOS1893SUPP>.
- [39] 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>
- [40] MINSKER, S. (2018). Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *Ann. Statist.* **46** 2871–2903. [MR3851758](#) <https://doi.org/10.1214/17-AOS1642>
- [41] MITRA, R. and ZHANG, C.-H. (2014). Multivariate analysis of nonparametric estimates of large correlation matrices. Preprint. Available at [arXiv:1403.6195](#).
- [42] PADMANABHAN, N., WHITE, M., ZHOU, H. H. and O’CONNELL, R. (2016). Estimating sparse precision matrices. *Mon. Not. R. Astron. Soc.* **460** 1567–1576.
- [43] PAUL, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17** 1617–1642. [MR2399865](#)
- [44] RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1 -penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980. [MR2836766](#) <https://doi.org/10.1214/11-EJS631>
- [45] REN, Z., SUN, T., ZHANG, C.-H. and ZHOU, H. H. (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *Ann. Statist.* **43** 991–1026. [MR3346695](#) <https://doi.org/10.1214/14-AOS1286>
- [46] ROTHMAN, A. J., BICKEL, P. J., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Stat.* **2** 494–515. [MR2417391](#) <https://doi.org/10.1214/08-EJS176>
- [47] SAULIS, L. and STATULEVIČIUS, V. A. (1991). *Limit Theorems for Large Deviations. Mathematics and Its Applications (Soviet Series)* **73**. Kluwer Academic, Dordrecht. [MR1171883](#) <https://doi.org/10.1007/978-94-011-3530-6>
- [48] SCHÄFER, J. and STRIMMER, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* **4** Art. 32, 28. [MR2183942](#) <https://doi.org/10.2202/1544-6115.1175>
- [49] SPEARMAN, C. (1904). Spearman’s rank correlation coefficient. *Am. J. Psychol.* **15** 72–101.
- [50] SUN, T. and ZHANG, C.-H. (2013). Sparse matrix inversion with scaled lasso. *J. Mach. Learn. Res.* **14** 3385–3418. [MR3144466](#)
- [51] WACHTER, K. W. (1976). Probability plotting points for principal components. In *Ninth Interface Symposium Computer Science and Statistics* 299–308. Prindle, Weber and Schmidt, Boston, MA.
- [52] WACHTER, K. W. (1978). The strong limits of random matrix spectra for sample matrices of independent elements. *Ann. Probab.* **6** 1–18. [MR0467894](#) <https://doi.org/10.1214/aop/1176995607>
- [53] WU, W. B. and POURAHMADI, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* **90** 831–844. [MR2024760](#) <https://doi.org/10.1093/biomet/90.4.831>
- [54] WU, W. B. and POURAHMADI, M. (2009). Banding sample autocovariance matrices of stationary processes. *Statist. Sinica* **19** 1755–1768. [MR2589209](#)
- [55] XIAO, L. and BUNEA, F. (2014). On the theoretic and practical merits of the banding estimator for large covariance matrices. Preprint. Available at [arXiv:1402.0844](#).
- [56] YU, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam* 423–435. Springer, New York. [MR1462963](#)
- [57] YUAN, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* **11** 2261–2286. [MR2719856](#)
- [58] YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. [MR2367824](#) <https://doi.org/10.1093/biomet/asm018>

ESTIMATION AND INFERENCE FOR PRECISION MATRICES OF NONSTATIONARY TIME SERIES

BY XIUCAI DING^{*} AND ZHOU ZHOU^{**}

Department of Statistics, University of Toronto, ^{}xiucai.ding@mail.utoronto.ca; ^{**}zhou@utstat.toronto.edu*

We consider the estimation of and inference on precision matrices of a rich class of univariate locally stationary linear and nonlinear time series, assuming that only one realization of the time series is observed. Using a Cholesky decomposition technique, we show that the precision matrices can be directly estimated via a series of least squares linear regressions with smoothly time-varying coefficients. The method of sieves is utilized for the estimation and is shown to be optimally adaptive in terms of estimation accuracy and efficient in terms of computational complexity. We establish an asymptotic theory for a class of $\mathcal{L}^2$ tests based on the nonparametric sieve estimators. The latter are used for testing whether the precision matrices are diagonal or banded. A Gaussian approximation result is established for a wide class of quadratic forms of nonstationary and possibly nonlinear processes of diverging dimensions which is of interest by itself.

REFERENCES

- [1] BELLONI, A., CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *J. Econometrics* **186** 345–366. [MR3343791](#) <https://doi.org/10.1016/j.jeconom.2015.02.014>
- [2] 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>
- [3] BROCKWELL, P. J. and DAVIS, R. A. (1987). *Time Series: Theory and Methods*. Springer Series in Statistics. Springer, New York. [MR0868859](#) <https://doi.org/10.1007/978-1-4899-0004-3>
- [4] CAI, T. T., REN, Z. and ZHOU, H. H. (2013). Optimal rates of convergence for estimating Toeplitz covariance matrices. *Probab. Theory Related Fields* **156** 101–143. [MR3055254](#) <https://doi.org/10.1007/s00440-012-0422-7>
- [5] CAI, T. T. and ZHOU, H. H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *Ann. Statist.* **40** 2389–2420. [MR3097607](#) <https://doi.org/10.1214/12-AOS998>
- [6] CHEN, X. (2007). Large sample sieve estimation of semi-nonparametric models. In *Handbook of Econometrics, Vol. 6B* (J. J. Heckman and E. E. Leamer, eds.). Chapter 76. [https://doi.org/10.1016/S1573-4412\(07\)06076-X](#)
- [7] CHEN, X. and CHRISTENSEN, T. M. (2015). Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *J. Econometrics* **188** 447–465. [MR3383220](#) <https://doi.org/10.1016/j.jeconom.2015.03.010>
- [8] CHEN, X., XU, M. and WU, W. B. (2013). Covariance and precision matrix estimation for high-dimensional time series. *Ann. Statist.* **41** 2994–3021. [MR3161455](#) <https://doi.org/10.1214/13-AOS1182>
- [9] DEMKO, S., MOSS, W. F. and SMITH, P. W. (1984). Decay rates for inverses of band matrices. *Math. Comp.* **43** 491–499. [MR0758197](#) <https://doi.org/10.2307/2008290>
- [10] DING, X. (2017). Asymptotics of empirical eigen-structure for high dimensional sample covariance matrices of general form. Available at [arXiv:1708.06296](#).
- [11] DING, X. and YANG, F. (2018). A necessary and sufficient condition for edge universality at the largest singular values of covariance matrices. *Ann. Appl. Probab.* **28** 1679–1738. [MR3809475](#) <https://doi.org/10.1214/17-AAP1341>
- [12] DING, X. and ZHOU, Z. (2020). Supplement to “Estimation and inference for precision matrices of nonstationary time series.” <https://doi.org/10.1214/19-AOS1894SUPP>.

MSC2020 subject classifications. 62M10, 62G05, 62G10.

Key words and phrases. Nonstationary time series, precision matrices, Cholesky decomposition, sieve estimation, high-dimensional Gaussian approximation, random matrices, white noise and bandedness tests.

- [13] ERDŐS, L., YAU, H.-T. and YIN, J. (2012). Rigidity of eigenvalues of generalized Wigner matrices. *Adv. Math.* **229** 1435–1515. [MR2871147](#) <https://doi.org/10.1016/j.aim.2011.12.010>
- [14] FAN, J., LIAO, Y. and LIU, H. (2016). An overview of the estimation of large covariance and precision matrices. *Econom. J.* **19** C1–C32. [MR3501529](#) <https://doi.org/10.1111/ectj.12061>
- [15] HAMILTON, J. D. (1994). *Time Series Analysis*. Princeton Univ. Press, Princeton, NJ. [MR1278033](#)
- [16] HANSEN, B. E. (2014). Nonparametric sieve regression: Least squares, averaging least squares, and cross-validation. In *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics* 215–248. Oxford Univ. Press, Oxford. [MR3306927](#)
- [17] KNOWLES, A. and YIN, J. (2017). Anisotropic local laws for random matrices. *Probab. Theory Related Fields* **169** 257–352. [MR3704770](#) <https://doi.org/10.1007/s00440-016-0730-4>
- [18] LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. [MR2572459](#) <https://doi.org/10.1214/09-AOS720>
- [19] MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models. Monographs on Statistics and Applied Probability*. CRC Press, London. [MR3223057](#) <https://doi.org/10.1007/978-1-4899-3242-6>
- [20] MCMURRY, T. L. and POLITIS, D. N. (2015). High-dimensional autocovariance matrices and optimal linear prediction. *Electron. J. Stat.* **9** 753–788. [MR3331856](#) <https://doi.org/10.1214/15-EJS1000>
- [21] PAUL, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17** 1617–1642. [MR2399865](#)
- [22] POURAHMADI, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika* **86** 677–690. [MR1723786](#) <https://doi.org/10.1093/biomet/86.3.677>
- [23] RAN, R.-S. and HUANG, T.-Z. (2009). An inversion algorithm for a banded matrix. *Comput. Math. Appl.* **58** 1699–1710. [MR2557547](#) <https://doi.org/10.1016/j.camwa.2009.07.069>
- [24] RÖLLIN, A. (2013). Stein’s method in high dimensions with applications. *Ann. Inst. Henri Poincaré Probab. Stat.* **49** 529–549. [MR3088380](#) <https://doi.org/10.1214/11-aihp473>
- [25] STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053. [MR0673642](#)
- [26] TAO, T. and VU, V. (2011). Random matrices: Universality of local eigenvalue statistics. *Acta Math.* **206** 127–204. [MR2784665](#) <https://doi.org/10.1007/s11511-011-0061-3>
- [27] TRETTER, C. (2008). *Spectral Theory of Block Operator Matrices and Applications*. Imperial College Press, London. [MR2463978](#) <https://doi.org/10.1142/9781848161122>
- [28] TROPP, J. (2015). *An Introduction to Matrix Concentration Inequalities. Foundations and Trends in Machine Learning*. Now Publishers Inc.
- [29] WADE, W. (2014). *Introduction to Analysis*, 4th ed. Prentice Hall, Upper Saddle River, NJ.
- [30] WU, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proc. Natl. Acad. Sci. USA* **102** 14150–14154. [MR2172215](#) <https://doi.org/10.1073/pnas.0506715102>
- [31] WU, W. B. and POURAHMADI, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika* **90** 831–844. [MR2024760](#) <https://doi.org/10.1093/biomet/90.4.831>
- [32] XIAO, H. and WU, W. B. (2012). Covariance matrix estimation for stationary time series. *Ann. Statist.* **40** 466–493. [MR3014314](#) <https://doi.org/10.1214/11-AOS967>
- [33] XU, M., ZHANG, D. and WU, W. (2015). L^2 asymptotics for high-dimensional data. Available at [arXiv:1405.7244](https://arxiv.org/abs/1405.7244).
- [34] YUAN, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* **11** 2261–2286. [MR2719856](#)
- [35] YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. [MR2367824](#) <https://doi.org/10.1093/biomet/asm018>
- [36] ZHANG, C. and ZHANG, T. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *Statist. Sci.* **27** 576–593.
- [37] ZHANG, X. and CHENG, G. (2018). Gaussian approximation for high dimensional vector under physical dependence. *Bernoulli* **24** 2640–2675. [MR3779697](#) <https://doi.org/10.3150/17-BEJ939>
- [38] ZHOU, Z. (2013). Inference for non-stationary time-series autoregression. *J. Time Series Anal.* **34** 508–516. [MR3070872](#) <https://doi.org/10.1111/jtsa.12028>
- [39] ZHOU, Z. and WU, W. B. (2009). Local linear quantile estimation for nonstationary time series. *Ann. Statist.* **37** 2696–2729. [MR2541444](#) <https://doi.org/10.1214/08-AOS636>
- [40] ZHOU, Z. and WU, W. B. (2010). Simultaneous inference of linear models with time varying coefficients. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 513–531. [MR2758526](#) <https://doi.org/10.1111/j.1467-9868.2010.00743.x>

ISOTROPIC COVARIANCE FUNCTIONS ON GRAPHS AND THEIR EDGES

BY ETHAN ANDERES¹, JESPER MØLLER^{2,*} AND JAKOB G. RASMUSSEN^{2,**}

¹*Department of Statistics, University of California at Davis, anderes@ucdavis.edu*

²*Department of Mathematical Sciences, Aalborg University, *jm@math.aau.dk; **jgr@math.aau.dk*

We develop parametric classes of covariance functions on linear networks and their extension to graphs with Euclidean edges, that is, graphs with edges viewed as line segments or more general sets with a coordinate system allowing us to consider points on the graph which are vertices or points on an edge. Our covariance functions are defined on the vertices and edge points of these graphs and are isotropic in the sense that they depend only on the geodesic distance or on a new metric called the resistance metric (which extends the classical resistance metric developed in electrical network theory on the vertices of a graph to the continuum of edge points). We discuss the advantages of using the resistance metric in comparison with the geodesic metric as well as the restrictions these metrics impose on the investigated covariance functions. In particular, many of the commonly used isotropic covariance functions in the spatial statistics literature (the power exponential, Matérn, generalized Cauchy and Dagum classes) are shown to be valid with respect to the resistance metric for any graph with Euclidean edges, whilst they are only valid with respect to the geodesic metric in more special cases.

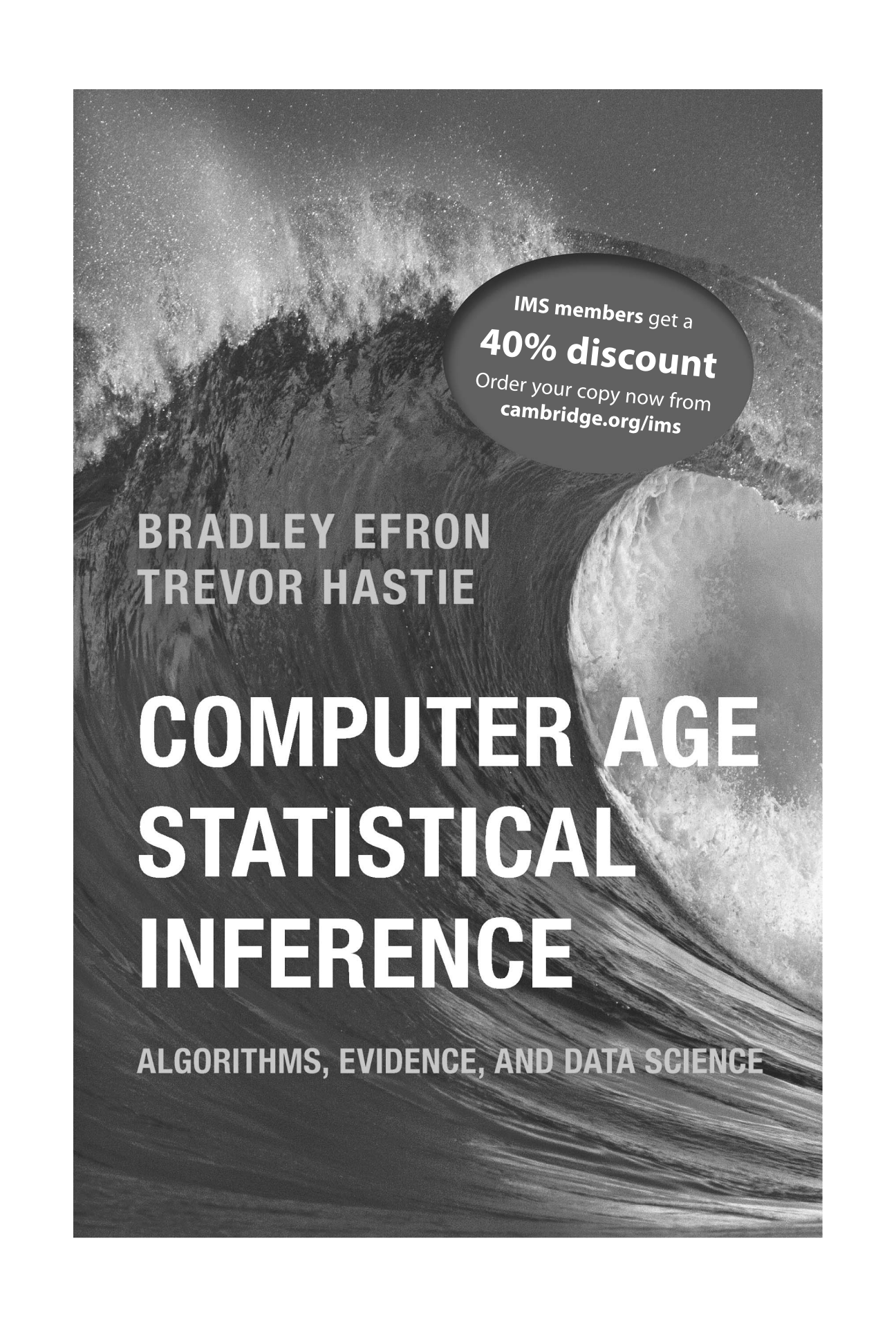
REFERENCES

- ABRAMOWITZ, M. and STEGUN, I. A. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York.
- BADDELEY, A. J., RUBAK, E. and TURNER, R. (2015). *Spatial Point Patterns: Methodology and Applications with R. Interdisciplinary Statistics*. CRC Press/CRC, Boca Raton, FL.
- BADDELEY, A., NAIR, G., RAKSHIT, S. and MCSWIGGAN, G. (2017). “Stationary” point processes are uncommon on linear networks. *Stat* **6** 68–78. [MR3613182](#) <https://doi.org/10.1002/sta4.135>
- BERENS, H. and XU, Y. (1997). l -1 summability of multiple Fourier integrals and positivity. *Math. Proc. Cambridge Philos. Soc.* **122** 149–172. [MR1443593](#) <https://doi.org/10.1017/S0305004196001521>
- BERG, C. (2008). Stieltjes–Pick–Bernstein–Schoenberg and their connection to complete monotonicity. In *Positive Definite Functions. From Schoenberg to Space–Time Challenges* (J. Mateu and E. Porcu, eds.) 15–45. Univ. Jaume I, Castellon.
- BERG, C., MATEU, J. and PORCU, E. (2008). The Dagum family of isotropic correlation functions. *Bernoulli* **14** 1134–1149. [MR2543589](#) <https://doi.org/10.3150/08-BEJ139>
- CAMBANIS, S., KEENER, R. and SIMONS, G. (1983). On α -symmetric multivariate distributions. *J. Multivariate Anal.* **13** 213–233. [MR0705548](#) [https://doi.org/10.1016/0047-259X\(83\)90023-4](https://doi.org/10.1016/0047-259X(83)90023-4)
- CRESSIE, N. and MAJURE, J. J. (1997). Spatio-temporal statistical modeling of livestock waste in streams. *J. Agric. Biol. Environ. Stat.* **2** 24–47. [MR1812249](#) <https://doi.org/10.2307/1400639>
- CRESSIE, N., FREY, J., HARCH, B. and SMITH, M. (2006). Spatial prediction on a river network. *J. Agric. Biol. Environ. Stat.* **11** 127–150.
- DEZA, M. M. and LAURENT, M. (1997). *Geometry of Cuts and Metrics. Algorithms and Combinatorics* **15**. Springer, Berlin. [MR1460488](#) <https://doi.org/10.1007/978-3-642-04295-9>
- GARDNER, B., SULLIVAN, P. and LEMBO, A. JR (2003). Predicting stream temperatures: Geostatistical model comparison using alternative distance metrics. *Canadian Journal of Fisheries and Aquatic Sciences* **60** 344–351.
- GNETING, T. (1998a). On α -symmetric multivariate characteristic functions. *J. Multivariate Anal.* **64** 131–147. [MR1621926](#) <https://doi.org/10.1006/jmva.1997.1713>

MSC2020 subject classifications. Primary 62M99; secondary 62M30, 60K99.

Key words and phrases. Geodesic metric, linear network, parametric classes of covariance functions, reproducing kernel Hilbert space, resistance metric, restricted covariance function properties.

- GNEITING, T. (1998b). Simple tests for the validity of correlation function models on the circle. *Statist. Probab. Lett.* **39** 119–122. [MR1652540](#) [https://doi.org/10.1016/S0167-7152\(98\)00042-X](https://doi.org/10.1016/S0167-7152(98)00042-X)
- GNEITING, T. (2013). Strictly and non-strictly positive definite functions on spheres. *Bernoulli* **19** 1327–1349. [MR3102554](#) <https://doi.org/10.3150/12-BEJSP06>
- JAYASUMANA, S., HARTLEY, R., SALZMANN, M., LI, H. and HARANDI, M. (2013). Kernel methods on the Riemannian manifold of symmetric positive definite matrices. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on* 73–80. IEEE Press, New York.
- JORGENSEN, P. E. T. and PEARSE, E. P. J. (2010). A Hilbert space approach to effective resistance metric. *Complex Anal. Oper. Theory* **4** 975–1013. [MR2735315](#) <https://doi.org/10.1007/s11785-009-0041-1>
- KIGAMI, J. (2003). Harmonic analysis for resistance forms. *J. Funct. Anal.* **204** 399–444. [MR2017320](#) [https://doi.org/10.1016/S0022-1236\(02\)00149-0](https://doi.org/10.1016/S0022-1236(02)00149-0)
- KLEIN, D. J. and RANDIĆ, M. (1993). Resistance distance. *J. Math. Chem.* **12** 81–95. [MR1219566](#) <https://doi.org/10.1007/BF01164627>
- LYONS, R. and PERES, Y. (2016). *Probability on Trees and Networks. Cambridge Series in Statistical and Probabilistic Mathematics* **42**. Cambridge Univ. Press, New York. [MR3616205](#) <https://doi.org/10.1017/9781316672815>
- MILLER, K. S. and SAMKO, S. G. (2001). Completely monotonic functions. *Integral Transforms Spec. Funct.* **12** 389–402. [MR1872377](#) <https://doi.org/10.1080/10652460108819360>
- MØLLER, J., SYVERSVEEN, A. R. and WAAGEPETERSEN, R. P. (1998). Log Gaussian Cox processes. *Scand. J. Stat.* **25** 451–482. [MR1650019](#) <https://doi.org/10.1111/1467-9469.00115>
- MØLLER, J. and WAAGEPETERSEN, R. P. (2004). *Statistical Inference and Simulation for Spatial Point Processes. Monographs on Statistics and Applied Probability* **100**. CRC Press/CRC, Boca Raton, FL. [MR2004226](#)
- OKABE, A. and SUGIHARA, K. (2012). *Spatial Analysis Along Networks: Statistical and Computational Methods*. Wiley, Chichester.
- SAITOH, S. (1997). *Integral Transforms, Reproducing Kernels and Their Applications. Pitman Research Notes in Mathematics Series* **369**. Longman, Harlow. [MR1478165](#)
- SCHOENBERG, I. J. (1935). Remarks to Maurice Fréchet's article “Sur la définition axiomatique d'une classe d'espace distanciés vectoriellement applicable sur l'espace de Hilbert” [MR1503246]. *Ann. of Math.* (2) **36** 724–732. [MR1503248](#) <https://doi.org/10.2307/1968654>
- SCHOENBERG, I. J. (1938a). Metric spaces and positive definite functions. *Trans. Amer. Math. Soc.* **44** 522–536. [MR1501980](#) <https://doi.org/10.2307/1989894>
- SCHOENBERG, I. J. (1938b). Metric spaces and completely monotone functions. *Ann. of Math.* (2) **39** 811–841. [MR1503439](#) <https://doi.org/10.2307/1968466>
- SPIEGEL, M. R. (1968). *Mathematical Handbook of Formulas and Tables*. McGraw/Hill, New York.
- VER HOEF, J. M. and PETERSON, E. E. (2010). A moving average approach for spatial statistical models of stream networks. *J. Amer. Statist. Assoc.* **105** 6–18. [MR2757185](#) <https://doi.org/10.1198/jasa.2009.ap08248>
- VER HOEF, J. M., PETERSON, E. and THEOBALD, D. (2006). Spatial statistical models that use flow and stream distance. *Environ. Ecol. Stat.* **13** 449–464. [MR2297373](#) <https://doi.org/10.1007/s10651-006-0022-8>
- VON NEUMANN, J. and SCHOENBERG, I. J. (1941). Fourier integrals and metric geometry. *Trans. Amer. Math. Soc.* **50** 226–251. [MR0004644](#) <https://doi.org/10.2307/1989957>
- 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>
- WELLS, J. H. and WILLIAMS, L. R. (1975). *Embeddings and Extensions in Analysis*. Springer, New York. [MR0461107](#)
- YAGLOM, A. M. (1987). *Correlation Theory of Stationary and Related Random Functions. Vol. II. Springer Series in Statistics*. Springer, New York. [MR0915557](#)



IMS members get a
40% discount

Order your copy now from
cambridge.org/ims

BRADLEY EFRON
TREVOR HASTIE

COMPUTER AGE STATISTICAL INFERENCE

ALGORITHMS, EVIDENCE, AND DATA SCIENCE