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

THE ANNALS *of* STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles

continued

Vol. 49, No. 6—December 2021

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

THE ANNALS *of* STATISTICS

An Official Journal of the
Institute of Mathematical Statistics

Articles—*Continued from front cover*

Vol. 49, No. 6—December 2021

THE ANNALS OF STATISTICS Vol. 49, No. 6, pp. 3043–3649 December 2021

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.

The Annals of Statistics **[ISSN 0090-5364 (print); ISSN 2168-8966 (online)],** Volume 49, Number 6, December 2021. Published bimonthly by the Institute of Mathematical Statistics, 9760 Smith Road, Waite Hill, Ohio 44094, USA. Periodicals postage paid at Cleveland, Ohio, and at additional mailing offices.

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

The Annals of Statistics
2021, Vol. 49, No. 6, 3043–3069
https://doi.org/10.1214/21-AOS2072

ESTIMATING THE NUMBER OF COMPONENTS IN FINITE MIXTURE MODELS VIA THE GROUP-SORT-FUSE PROCEDURE

BY TUDOR MANOLE[1] AND ABBAS KHALILI[2]

[1]*Department of Statistics and Data Science, Carnegie Mellon University, tmanole@andrew.cmu.edu*
[2]*Department of Mathematics and Statistics, McGill University, abbas.khalili@mcgill.ca*

Estimation of the number of components (or order) of a finite mixture model is a long standing and challenging problem in statistics. We propose the Group-Sort-Fuse (GSF) procedure—a new penalized likelihood approach for simultaneous estimation of the order and mixing measure in multidimensional finite mixture models. Unlike methods which fit and compare mixtures with varying orders using criteria involving model complexity, our approach directly penalizes a continuous function of the model parameters. More specifically, given a conservative upper bound on the order, the GSF groups and sorts mixture component parameters to fuse those which are redundant. For a wide range of finite mixture models, we show that the GSF is consistent in estimating the true mixture order and achieves the $n^{-1/2}$ convergence rate for parameter estimation up to polylogarithmic factors. The GSF is implemented for several univariate and multivariate mixture models in the R package `GroupSortFuse`. Its finite sample performance is supported by a thorough simulation study, and its application is illustrated on two real data examples.

REFERENCES

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **AC-19** 716–723. MR0423716 https://doi.org/10.1109/tac.1974.1100705

BECHTEL, Y. C., BONAITI-PELLIE, C., POISSON, N., MAGNETTE, J. and BECHTEL, P. R. (1993). A population and family study of N-acetyltransferase using caffeine urinary metabolites. *Clin. Pharmacol. Ther.* **54** 134–141.

BENAGLIA, T., CHAUVEAU, D., HUNTER, D. and YOUNG, D. (2009). mixtools: An R package for analyzing finite mixture models. *J. Stat. Softw.* **32** 1–29.

BIERNACKI, C., CELEUX, G. and GOVAERT, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.* **22** 719–725.

BOSCH-DOMÈNECH, A., MONTALVO, J. G., NAGEL, R. and SATORRA, A. (2010). A finite mixture analysis of beauty-contest data using generalized beta distributions. *Exp. Econm.* **13** 461–475.

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

CHEN, H. and CHEN, J. (2003). Tests for homogeneity in normal mixtures in the presence of a structural parameter. *Statist. Sinica* **13** 351–365. MR1977730

CHEN, H., CHEN, J. and KALBFLEISCH, J. D. (2004). Testing for a finite mixture model with two components. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 95–115. MR2035761 https://doi.org/10.1111/j.1467-9868.2004.00434.x

CHEN, J. and KALBFLEISCH, J. D. (1996). Penalized minimum-distance estimates in finite mixture models. *Canad. J. Statist.* **24** 167–175. MR1406173 https://doi.org/10.2307/3315623

CHEN, J. and KHALILI, A. (2008). Order selection in finite mixture models with a nonsmooth penalty. *J. Amer. Statist. Assoc.* **103** 1674–1683. MR2722574 https://doi.org/10.1198/016214508000001075

CHEN, J. and LI, P. (2009). Hypothesis test for normal mixture models: The EM approach. *Ann. Statist.* **37** 2523–2542. MR2543701 https://doi.org/10.1214/08-AOS651

MSC2020 subject classifications. Primary 62F10, 62F12; secondary 62H12.

Key words and phrases. Finite mixture models, maximum penalized likelihood estimation, Wasserstein distance, strong identifiability.

DACUNHA-CASTELLE, D. and GASSIAT, E. (1999). Testing the order of a model using locally conic parametrization: Population mixtures and stationary ARMA processes. *Ann. Statist.* **27** 1178–1209. MR1740115 https://doi.org/10.1214/aos/1017938921

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

DRTON, M. and PLUMMER, M. (2017). A Bayesian information criterion for singular models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 323–380. MR3611750 https://doi.org/10.1111/rssb.12187

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

FRALEY, C. and RAFTERY, A. E. (1999). MCLUST: Software for model-based cluster analysis. *J. Classification* **16** 297–306.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, New York.

FRÜHWIRTH-SCHNATTER, S. (2006). *Finite Mixture and Markov Switching Models. Springer Series in Statistics*. Springer, New York. MR2265601

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

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

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

GUHA, A., HO, N. and NGUYEN, X. (2019). On posterior contraction of parameters and interpretability in Bayesian mixture modeling. *Bernoulli* Available at arXiv:1901.05078.

HATHAWAY, R. J. (1986). A constrained EM algorithm for univariate normal mixtures. *J. Stat. Comput. Simul.* **23** 211–230.

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

HO, N. P. M. (2017). *Parameter Estimation and Multilevel Clustering with Mixture and Hierarchical Models*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)—University of Michigan. MR3809858

HO, N. and NGUYEN, X. (2016a). 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

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

HO, N., NGUYEN, X. and RITOV, Y. (2020). Robust estimation of mixing measures in finite mixture models. *Bernoulli* **26** 828–857. MR4058353 https://doi.org/10.3150/18-BEJ1087

HO, N., YANG, C.-Y. and JORDAN, M. I. (2019). Convergence rates for Gaussian mixtures of experts. arXiv preprint. Available at arXiv:1907.04377.

HOLZMANN, H., MUNK, A. and STRATMANN, B. (2004). Identifiability of finite mixtures—with applications to circular distributions. *Sankhyā* **66** 440–449. MR2108200

HUNG, Y., WANG, Y., ZARNITSYNA, V., ZHU, C. and WU, C. F. J. (2013). Hidden Markov models with applications in cell adhesion experiments. *J. Amer. Statist. Assoc.* **108** 1469–1479. MR3174722 https://doi.org/10.1080/01621459.2013.836973

ISHWARAN, H., JAMES, L. F. and SUN, J. (2001). Bayesian model selection in finite mixtures by marginal density decompositions. *J. Amer. Statist. Assoc.* **96** 1316–1332. MR1946579 https://doi.org/10.1198/016214501753382255

JAMES, L. F., PRIEBE, C. E. and MARCHETTE, D. J. (2001). Consistent estimation of mixture complexity. *Ann. Statist.* **29** 1281–1296. MR1873331 https://doi.org/10.1214/aos/1013203454

KERIBIN, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā Ser. A* **62** 49–66. MR1769735

LEROUX, B. G. (1992). Consistent estimation of a mixing distribution. *Ann. Statist.* **20** 1350–1360. MR1186253 https://doi.org/10.1214/aos/1176348772

LI, P. and CHEN, J. (2010). Testing the order of a finite mixture. *J. Amer. Statist. Assoc.* **105** 1084–1092. MR2752604 https://doi.org/10.1198/jasa.2010.tm09032

LI, P., CHEN, J. and MARRIOTT, P. (2009). Non-finite Fisher information and homogeneity: An EM approach. *Biometrika* **96** 411–426. MR2507152 https://doi.org/10.1093/biomet/asp011

LIU, X. and SHAO, Y. (2003). Asymptotics for likelihood ratio tests under loss of identifiability. *Ann. Statist.* **31** 807–832. MR1994731 https://doi.org/10.1214/aos/1056562463

MANOLE, T. and KHALILI, A. (2021). Supplement to "Estimating the Number of Components in Finite Mixture Models via the Group-Sort-Fuse Procedure." https://doi.org/10.1214/21-AOS2072SUPP

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test stastistic for the number of components in a normal mixture. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **36** 318–324.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models. Wiley Series in Probability and Statistics: Applied Probability and Statistics*. Wiley Interscience, New York. MR1789474 https://doi.org/10.1002/0471721182

Miller, J. W. and Harrison, M. T. (2018). Mixture models with a prior on the number of components. *J. Amer. Statist. Assoc.* **113** 340–356. MR3803469 https://doi.org/10.1080/01621459.2016.1255636

Morel, J. G. and Nagaraj, N. K. (1993). A finite mixture distribution for modelling multinomial extra variation. *Biometrika* **80** 363–371. MR1243510 https://doi.org/10.1093/biomet/80.2.363

Morris, T. H., Richmond, D. R. and Grimshaw, S. D. (1996). Orientation of dinosaur bones in riverine environments: Insights into sedimentary dynamics and taphonomy. In *The Continental Jurassic* 521–530. Museum of Northern Arizona, Flagstaff, AZ.

Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika* **49** 65–82. MR0143299 https://doi.org/10.1093/biomet/49.1-2.65

Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization. Applied Optimization* **87**. Kluwer Academic, Boston, MA. MR2142598 https://doi.org/10.1007/978-1-4419-8853-9

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

Nobile, A. (1994). *Bayesian Analysis of Finite Mixture Distributions*. ProQuest LLC, Pittsburgh, PA. Thesis (Ph.D.)—Carnegie Mellon University. MR2692049

Petralia, F., Rao, V. and Dunson, D. (2012). Repulsive mixtures. In *Advances in Neural Information Processing Systems* (F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger, eds.) **25** 1889–1897. Curran Associates, Red Hook, NY.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Roy. Statist. Soc. Ser. B* **59** 731–792. MR1483213 https://doi.org/10.1111/1467-9868.00095

Rousseau, J. and Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 689–710. MR2867454 https://doi.org/10.1111/j.1467-9868.2011.00781.x

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components—an alternative to reversible jump methods. *Ann. Statist.* **28** 40–74. MR1762903 https://doi.org/10.1214/aos/1016120364

Teicher, H. (1963). Identifiability of finite mixtures. *Ann. Math. Stat.* **34** 1265–1269. MR0155376 https://doi.org/10.1214/aoms/1177703862

Thompson, T. J., Smith, P. J. and Boyle, J. P. (1998). Finite mixture models with concomitant information: Assessing diagnostic criteria for diabetes. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **47** 393–404.

van de Geer, S. (2000). *Empirical Processes in M-Estimation*. Cambridge University Press, Cambridge.

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

Woo, M.-J. and Sriram, T. N. (2006). Robust estimation of mixture complexity. *J. Amer. Statist. Assoc.* **101** 1475–1486. MR2279473 https://doi.org/10.1198/016214506000000555

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

Xie, F. and Xu, Y. (2020). Bayesian repulsive Gaussian mixture model. *J. Amer. Statist. Assoc.* **115** 187–203. MR4078456 https://doi.org/10.1080/01621459.2018.1537918

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

Zhang, Y., Li, R. and Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion. *J. Amer. Statist. Assoc.* **105** 312–323. MR2656055 https://doi.org/10.1198/jasa.2009.tm08013

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

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533. MR2435443 https://doi.org/10.1214/009053607000000802

The Annals of Statistics
2021, Vol. 49, No. 6, 3070–3102
https://doi.org/10.1214/21-AOS2073

A SIMPLE MEASURE OF CONDITIONAL DEPENDENCE

BY MONA AZADKIA[*] AND SOURAV CHATTERJEE[†]

Department of Statistics, Stanford University, [*]*mazadkia@stanford.edu;* [†]*souravc@stanford.edu*

We propose a coefficient of conditional dependence between two random variables Y and Z given a set of other variables $X_1, \dots, X_p$, based on an i.i.d. sample. The coefficient has a long list of desirable properties, the most important of which is that under absolutely no distributional assumptions, it converges to a limit in [0, 1], where the limit is 0 if and only if Y and Z are conditionally independent given $X_1, \dots, X_p$, and is 1 if and only if Y is equal to a measurable function of Z given $X_1, \dots, X_p$. Moreover, it has a natural interpretation as a nonlinear generalization of the familiar partial R^2 statistic for measuring conditional dependence by regression. Using this statistic, we devise a new variable selection algorithm, called Feature Ordering by Conditional Independence (FOCI), which is model-free, has no tuning parameters, and is provably consistent under sparsity assumptions. A number of applications to synthetic and real data sets are worked out.

REFERENCES

[1] ADRAGNI, K. P. and COOK, R. D. (2009). Sufficient dimension reduction and prediction in regression. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **367** 4385–4405. MR2546393 https://doi.org/10.1098/rsta.2009.0110

[2] AMIT, Y. and GEMAN, D. (1997). Shape quantization and recognition with randomized trees. *Neural Comput.* **9** 1545–1588.

[3] AZADKIA, M., CHATTERJEE, S. and MATLOFF, N. (2020). FOCI: Feature Ordering by Conditional Independence. R Package. https://CRAN.R-project.org/package=FOCI.

[4] BATTITI, R. (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. Neural Netw.* **5** 537–550.

[5] BERGSMA, W. P. (2004). Testing conditional independence for continuous random variables.

[6] BERRETT, T. B., WANG, Y., BARBER, R. F. and SAMWORTH, R. J. (2020). The conditional permutation test for independence while controlling for confounders. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 175–197. MR4060981

[7] BREIMAN, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37** 373–384. MR1365720 https://doi.org/10.2307/1269730

[8] BREIMAN, L. (1996). Bagging predictors. *Mach. Learn.* **26** 123–140.

[9] BREIMAN, L. (2018). Random forests. *Mach. Learn.* **45** 5–32.

[10] 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

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

[12] 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

[13] CHATTERJEE, S. (2019). A new coefficient of correlation. *J. Amer. Statist. Assoc.* Available at https://arxiv.org/abs/1909.10140.

[14] CHEN, S. and DONOHO, D. (1994). Basis pursuit. In *Proc. of* 28*th Asilomar Conf. on Signals, Systems and Computers* **1** 41–44. IEEE, Los Alamitos.

[15] COCHRAN, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics* **10** 417–451. MR0067428 https://doi.org/10.2307/3001616

MSC2020 subject classifications. 62G05, 62H20.
Key words and phrases. Conditional dependence, nonparametric measures of association, variable selection.

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

[17] DETTE, H., SIBURG, K. F. and STOIMENOV, P. A. (2013). A copula-based non-parametric measure of regression dependence. *Scand. J. Stat.* **40** 21–41. MR3024030 https://doi.org/10.1111/j.1467-9469.2011.00767.x

[18] DORAN, G., MUANDET, K., ZHANG, K. and SCHÖLKOPF, B. (2014). A permutation-based kernel conditional independence test. *Uncertainty Artif. Intell.* **30** 132–141.

[19] DUA, D. and GRAFF, C. (2019). *UCI Machine Learning Repository* [*http://archive.ics.uci.edu/ml*]. Univ. California, School of Information and Computer Science, Irvine, CA.

[20] DURRETT, R. (2010). *Probability: Theory and Examples*, 4th ed. *Cambridge Series in Statistical and Probabilistic Mathematics* **31**. Cambridge Univ. Press, Cambridge. MR2722836 https://doi.org/10.1017/CBO9780511779398

[21] DVORETZKY, A., KIEFER, J. and WOLFOWITZ, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Stat.* **27** 642–669. MR0083864 https://doi.org/10.1214/aoms/1177728174

[22] EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166 https://doi.org/10.1214/009053604000000067

[23] FAN, J., FENG, Y. and XIA, L. (2020). A projection-based conditional dependence measure with applications to high-dimensional undirected graphical models. *J. Econometrics* **218** 119–139. MR4111748 https://doi.org/10.1016/j.jeconom.2019.12.016

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

[25] FREUND, Y. and SCHAPIRE, R. (1996). Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the Thirteenth International Conference*. pp. 148–156.

[26] FRIEDMAN, J. H. (1991). Multivariate adaptive regression splines. *Ann. Statist.* **19** 1–141. MR1091842 https://doi.org/10.1214/aos/1176347963

[27] FRIEDMAN, J. H., BENTLEY, J. L. and FINKEL, R. A. (1977). An algorithm for finding best matches in logarithmic expected time. *ACM Trans. Math. Software* **3** 209–226.

[28] FUKUMIZU, K., GRETTON, A., SUN, X. and SCHÖLKOPF, B. (2008). Kernel measures of conditional dependence. *Adv. Neural Inf. Process. Syst.* **20** 489–496.

[29] GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.

[30] HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics*. Springer, New York. MR1851606 https://doi.org/10.1007/978-0-387-21606-5

[31] HO, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* **20** 832–844.

[32] HUANG, T.-M. (2010). Testing conditional independence using maximal nonlinear conditional correlation. *Ann. Statist.* **38** 2047–2091. MR2676883 https://doi.org/10.1214/09-AOS770

[33] JOE, H. (1989). Relative entropy measures of multivariate dependence. *J. Amer. Statist. Assoc.* **84** 157–164. MR0999674

[34] JOSSE, J. and HOLMES, S. (2016). Measuring multivariate association and beyond. *Stat. Surv.* **10** 132–167. MR3573303 https://doi.org/10.1214/16-SS116

[35] KNUTH, D. E. (1998). *The Art of Computer Programming. Vol.* 3: *Sorting and Searching*. Addison-Wesley, Reading, MA. MR3077154

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

[37] LINTON, O. and GOZALO, P. (1996). Conditional independence restrictions: Testing and estimation. Cowles Foundation Discussion Paper, no. 1140.

[38] MANTEL, N. and HAENSZEL, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.* **22** 719–748.

[39] MASSART, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.* **18** 1269–1283. MR1062069

[40] MCDIARMID, C. (1989). On the method of bounded differences. In *Surveys in Combinatorics*, 1989 (*Norwich*, 1989). *London Mathematical Society Lecture Note Series* **141** 148–188. Cambridge Univ. Press, Cambridge. MR1036755

[41] MILLER, A. J. (1990). *Subset Selection in Regression. Monographs on Statistics and Applied Probability* **40**. CRC Press, London. MR1072361 https://doi.org/10.1007/978-1-4899-2939-6

[42] PATRA, R. K., SEN, B. and SZÉKELY, G. J. (2016). On a nonparametric notion of residual and its applications. *Statist. Probab. Lett.* **109** 208–213. MR3434980 https://doi.org/10.1016/j.spl.2015.10.011

[43] PEARL, J. (1988). *Probabilistic Reasoning in Intelligent Systems*: *Networks of Plausible Inference*. *The Morgan Kaufmann Series in Representation and Reasoning*. Morgan Kaufmann, San Mateo, CA. MR0965765

[44] PÓCZOS, B. and SCHNEIDER, J. (2012). Nonparametric estimation of conditional information and divergences. *Proc*. *Mach*. *Learn*. *Res*. **22** 914–923.

[45] RAVIKUMAR, P., LAFFERTY, J., LIU, H. and WASSERMAN, L. (2009). Sparse additive models. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **71** 1009–1030. MR2750255 https://doi.org/10.1111/j.1467-9868.2009.00718.x

[46] RUDIN, W. (1987). *Real and Complex Analysis*, 3rd ed. McGraw-Hill, New York. MR0924157

[47] RUNGE, J. (2018). Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. *Proc*. 21*st Internat*. *Conf*. *Artif*. *Intell*. *Statist*. **84** 938–947.

[48] SEN, R., SURESH, A. T., SHANMUGAM, K., DIMAKIS, A. G. and SHAKKOTTAI, S. (2017). Model-powered conditional independence test. *Adv*. *Neural Inf*. *Process*. *Syst*. **31** 2955–2965.

[49] SETH, S. and PRÍNCIPE, J. C. (2012). Conditional association. *Neural Comput*. **24** 1882–1905.

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

[51] SONG, K. (2009). Testing conditional independence via Rosenblatt transforms. *Ann*. *Statist*. **37** 4011–4045. MR2572451 https://doi.org/10.1214/09-AOS704

[52] STROBL, E. V., ZHANG, K. and VISWESWARAN, S. (2019). Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *J*. *Causal Inference* **7**.

[53] SU, L. and WHITE, H. (2007). A consistent characteristic function-based test for conditional independence. *J*. *Econometrics* **141** 807–834. MR2413488 https://doi.org/10.1016/j.jeconom.2006.11.006

[54] SU, L. and WHITE, H. (2008). A nonparametric Hellinger metric test for conditional independence. *Econometric Theory* **24** 829–864. MR2428851 https://doi.org/10.1017/S0266466608080341

[55] SU, L. and WHITE, H. (2014). Testing conditional independence via empirical likelihood. *J*. *Econometrics* **182** 27–44. MR3212759 https://doi.org/10.1016/j.jeconom.2014.04.006

[56] SZÉKELY, G. J. and RIZZO, M. L. (2014). Partial distance correlation with methods for dissimilarities. *Ann*. *Statist*. **42** 2382–2412. MR3269983 https://doi.org/10.1214/14-AOS1255

[57] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J*. *Roy*. *Statist*. *Soc*. *Ser*. *B* **58** 267–288. MR1379242

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

[59] VERGARA, J. R. and ESTÉVEZ, P. A. (2014). A review of feature selection methods based on mutual information. *Neural Comput*. *Appl*. **24** 175–186.

[60] WANG, X., PAN, W., HU, W., TIAN, Y. and ZHANG, H. (2015). Conditional distance correlation. *J*. *Amer*. *Statist*. *Assoc*. **110** 1726–1734. MR3449068 https://doi.org/10.1080/01621459.2014.993081

[61] 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

[62] YUKICH, J. E. (1998). *Probability Theory of Classical Euclidean Optimization Problems*. *Lecture Notes in Math*. **1675**. Springer, Berlin. MR1632875 https://doi.org/10.1007/BFb0093472

[63] ZHANG, K., PETERS, J., JANZING, D. and SCHÖLKOPF, B. (2011). Kernel-based conditional independence test and application in causal discovery. *Uncertainty Artif*. *Intell*. **27** 804–813.

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

[65] ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **67** 301–320. MR2137327 https://doi.org/10.1111/j.1467-9868.2005.00503.x

The Annals of Statistics
2021, Vol. 49, No. 6, 3103–3126
https://doi.org/10.1214/21-AOS2076

ONLINE INFERENCE WITH MULTI-MODAL LIKELIHOOD FUNCTIONS

BY MATHIEU GERBER[1] AND KARI HEINE[2]

[1]*School of Mathematics, University of Bristol, mathieu.gerber@bristol.ac.uk*
[2]*Department of Mathematical Sciences, University of Bath, k.m.p.heine@bath.ac.uk*

Let $(Y_t)_{t\geq 1}$ be a sequence of i.i.d. observations and $\{f_\theta, \theta \in \mathbb{R}^d\}$ be a parametric model. We introduce a new online algorithm for computing a sequence $(\hat{\theta}_t)_{t\geq 1}$, which is shown to converge almost surely to $\mathrm{argmax}_{\theta\in\mathbb{R}^d} \mathbb{E}[\log f_\theta(Y_1)]$ at rate $\mathcal{O}(\log(t)^{(1+\varepsilon)/2}t^{-1/2})$, with $\varepsilon > 0$ a user specified parameter. This convergence result is obtained under standard conditions on the statistical model and, most notably, we allow the mapping $\theta \mapsto \mathbb{E}[\log f_\theta(Y_1)]$ to be multi-modal. However, the computational cost to process each observation grows exponentially with the dimension of θ, which makes the proposed approach applicable to low or moderate dimensional problems only. We also derive a version of the estimator $\hat{\theta}_t$, which is well suited to student-t linear regression models that are popular tools for robust linear regression. As shown by experiments on simulated and real data, the corresponding estimator of the regression coefficients is, as expected, robust to the presence of outliers and thus, as a by-product, we obtain a new adaptive and robust online estimation procedure for linear regression models.

REFERENCES

DICK, J. and PILLICHSHAMMER, F. (2010). *Digital Nets and Sequences*: *Discrepancy Theory and Quasi-Monte Carlo Integration*. Cambridge Univ. Press, Cambridge. MR2683394 https://doi.org/10.1017/CBO9780511761188

EBY, W. M., TABATABAI, M. A. and BURSAC, Z. (2010). Hyperbolastic modeling of tumor growth with a combined treatment of iodoacetate and dimethylsulphoxide. *BMC Cancer* **10** 509.

FERNANDEZ, C. and STEEL, M. F. J. (1999). Multivariate Student-*t* regression models: Pitfalls and inference. *Biometrika* **86** 153–167. MR1688079 https://doi.org/10.1093/biomet/86.1.153

GERBER, M. and HEINE, K. (2021). Supplement to "Online inference with multi-modal likelihood functions." https://doi.org/10.1214/21-AOS2076SUPP

GUEMMEGNE, J., KENGWOUNG-KEUMO, J., TABATABAI, M. and SINGH, K. (2014). Modeling the dynamics of the us healthcare expenditure using a hyperbolastic function. *Adv. Appl. Stat.* **42** 95.

HE, D., SUN, D. and HE, L. (2021). Objective Bayesian analysis for the Student-*t* linear regression. *Bayesian Anal.* **16** 129–145. MR4194276 https://doi.org/10.1214/20-BA1198

KLEIJN, B. J. K. and VAN DER VAART, A. W. (2012). The Bernstein–Von-Mises theorem under misspecification. *Electron. J. Stat.* **6** 354–381. MR2988412 https://doi.org/10.1214/12-EJS675

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

LEI, Y., HU, T., LI, G. and TANG, K. (2020). Stochastic gradient descent for nonconvex learning without bounded gradient assumptions. *IEEE Trans. Neural Netw. Learn. Syst.* **31** 4394–4400. MR4162089

O'CALLAGHAN, L., MISHRA, N., MEYERSON, A., GUHA, S. and MOTWANI, R. (2002). Streaming-data algorithms for high-quality clustering. In *Data Engineering*, 2002. *Proceedings*. 18*th International Conference on* 685–694. IEEE.

PESME, S. and FLAMMARION, N. (2020). Online robust regression via sgd on the l1 loss. Preprint. Available at arXiv:2007.00399.

SHAMIR, O. and ZHANG, T. (2013). Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International Conference on Machine Learning* 71–79.

TABATABAI, M., WILLIAMS, D. K. and BURSAC, Z. (2005). Hyperbolastic growth models: Theory and application. *Theor. Biol. Med. Model.* **2** 14.

MSC2020 subject classifications. Primary 68W27, 62F10, 62F12, 62L12; secondary 90C26.
Key words and phrases. Online inference, scalable inference, multi-modal likelihood functions.

TADIĆ, V. B. (2015). Convergence and convergence rate of stochastic gradient search in the case of multiple and non-isolated extrema. *Stochastic Process*. *Appl*. **125** 1715–1755. MR3315611 https://doi.org/10.1016/j.spa.2014.11.001

TOULIS, P. and AIROLDI, E. M. (2015). Scalable estimation strategies based on stochastic approximations: Classical results and new insights. *Stat*. *Comput*. **25** 781–795. MR3360492 https://doi.org/10.1007/s11222-015-9560-y

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

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

The Annals of Statistics
2021, Vol. 49, No. 6, 3127–3152
https://doi.org/10.1214/21-AOS2077

ON FIXED-DOMAIN ASYMPTOTICS, PARAMETER ESTIMATION AND ISOTROPIC GAUSSIAN RANDOM FIELDS WITH MATÉRN COVARIANCE FUNCTIONS

BY WEI-LIEM LOH[*], SAIFEI SUN[†] AND JUN WEN[‡]

Dedicated to Louis H. Y. Chen on his eightieth birthday

Department of Statistics and Applied Probability, National University of Singapore,
[*]*stalohwl@nus.edu.sg;* [†]*saifeisun@u.nus.edu;* [‡]*jun.wen@u.nus.edu*

A method is proposed for estimating the microergodic parameters (including the smoothness parameter) of stationary Gaussian random fields on $\mathbb{R}^d$ with isotropic Matérn covariance functions using irregularly spaced data. This approach uses higher-order quadratic variations and is applied to three designs, namely stratified sampling design, randomized sampling design and deformed lattice design. Microergodic parameter estimators are constructed for each of the designs. Under mild conditions, these estimators are shown to be consistent with respect to fixed-domain asymptotics. Upper bounds to the convergence rate of the estimators are also established. A simulation study is conducted to gauge the accuracy of the proposed estimators.

REFERENCES

[1] ADLER, R. J. and PYKE, R. (1993). Uniform quadratic variation for Gaussian processes. *Stochastic Process. Appl.* **48** 191–209. MR1244542 https://doi.org/10.1016/0304-4149(93)90044-5

[2] ANDERES, E. and CHATTERJEE, S. (2009). Consistent estimates of deformed isotropic Gaussian random fields on the plane. *Ann. Statist.* **37** 2324–2350. MR2543694 https://doi.org/10.1214/08-AOS647

[3] ANDERES, E. B. and STEIN, M. L. (2008). Estimating deformations of isotropic Gaussian random fields on the plane. *Ann. Statist.* **36** 719–741. MR2396813 https://doi.org/10.1214/009053607000000893

[4] ANDREWS, G. E., ASKEY, R. and ROY, R. (1999). *Special Functions. Encyclopedia of Mathematics and Its Applications* **71**. Cambridge Univ. Press, Cambridge. MR1688958 https://doi.org/10.1017/CBO9781107325937

[5] BENASSI, A., COHEN, S., ISTAS, J. and JAFFARD, S. (1998). Identification of filtered white noises. *Stochastic Process. Appl.* **75** 31–49. MR1629014 https://doi.org/10.1016/S0304-4149(97)00123-3

[6] BLANKE, D. and VIAL, C. (2014). Global smoothness estimation of a Gaussian process from general sequence designs. *Electron. J. Stat.* **8** 1152–1187. MR3263116 https://doi.org/10.1214/14-EJS925

[7] CHAN, G. and WOOD, A. T. A. (2000). Increment-based estimators of fractal dimension for two-dimensional surface data. *Statist. Sinica* **10** 343–376. MR1769748

[8] CONSTANTINE, A. G. and HALL, P. (1994). Characterizing surface smoothness via estimation of effective fractal dimension. *J. Roy. Statist. Soc. Ser. B* **56** 97–113. MR1257799

[9] DU, J., ZHANG, H. and MANDREKAR, V. S. (2009). Fixed-domain asymptotic properties of tapered maximum likelihood estimators. *Ann. Statist.* **37** 3330–3361. MR2549562 https://doi.org/10.1214/08-AOS676

[10] GRENANDER, U. and MILLER, M. I. (2007). *Pattern Theory: From Representation to Inference*. Oxford Univ. Press, Oxford. MR2285439

[11] GUTTORP, P. and GNEITING, T. (2006). Studies in the history of probability and statistics. XLIX. On the Matérn correlation family. *Biometrika* **93** 989–995. MR2285084 https://doi.org/10.1093/biomet/93.4.989

[12] GUYON, X. and LEÓN, J. (1989). Convergence en loi des H-variations d'un processus gaussien stationnaire sur **R**. *Ann. Inst. Henri Poincaré Probab. Stat.* **25** 265–282. MR1023952

MSC2020 subject classifications. Primary 60M30; secondary 60M40.
Key words and phrases. Consistency, convergence rate, deformed lattice, fixed-domain asymptotics, Gaussian random field, quadratic variation, irregularly spaced data, Matérn covariance, microergodic parameter, random sampling, smoothness.

[13] HALL, P. and WOOD, A. (1993). On the performance of box-counting estimators of fractal dimension. *Biometrika* **80** 246–252. MR1225230 https://doi.org/10.1093/biomet/80.1.246

[14] IM, H. K., STEIN, M. L. and ZHU, Z. (2007). Semiparametric estimation of spectral density with irregular observations. *J. Amer. Statist. Assoc.* **102** 726–735. MR2381049 https://doi.org/10.1198/016214507000000220

[15] ISTAS, J. and LANG, G. (1997). Quadratic variations and estimation of the local Hölder index of a Gaussian process. *Ann. Inst. Henri Poincaré Probab. Stat.* **33** 407–436. MR1465796 https://doi.org/10.1016/S0246-0203(97)80099-4

[16] JOSHI, S. C. and MILLER, M. I. (2000). Landmark matching via large deformation diffeomorphisms. *IEEE Trans. Image Process.* **9** 1357–1370. MR1808275 https://doi.org/10.1109/83.855431

[17] KAUFMAN, C. G. and SHABY, B. A. (2013). The role of the range parameter for estimation and prediction in geostatistics. *Biometrika* **100** 473–484. MR3068447 https://doi.org/10.1093/biomet/ass079

[18] KENT, J. T. and WOOD, A. T. A. (1997). Estimating the fractal dimension of a locally self-similar Gaussian process by using increments. *J. Roy. Statist. Soc. Ser. B* **59** 679–699. MR1452033

[19] KESHAVARZ, H., NGUYEN, X. and SCOTT, C. (2019). Local inversion-free estimation of spatial Gaussian processes. *Electron. J. Stat.* **13** 4224–4279. MR4021265 https://doi.org/10.1214/19-EJS1592

[20] LÉVY, P. (1940). Le mouvement brownien plan. *Amer. J. Math.* **62** 487–550. MR0002734 https://doi.org/10.2307/2371467

[21] LOH, W.-L. (2015). Estimating the smoothness of a Gaussian random field from irregularly spaced data via higher-order quadratic variations. *Ann. Statist.* **43** 2766–2794. MR3405611 https://doi.org/10.1214/15-AOS1365

[22] LOH, W.-L, SUN, S. and WEN, J. (2021). Supplement to "On fixed-domain asymptotics, parameter estimation and isotropic Gaussian random fields with Matérn covariance functions." https://doi.org/10.1214/21-AOS2077SUPP

[23] MATHERON, G. (1971). *The Theory of Regionalized Variables and Its Applications*. Ecole des Mines, Fontainebleau.

[24] OLVER, P. J. (2006). On multivariate interpolation. *Stud. Appl. Math.* **116** 201–240. MR2200737 https://doi.org/10.1111/j.1467-9590.2006.00335.x

[25] STEIN, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. *Springer Series in Statistics*. Springer, New York. MR1697409 https://doi.org/10.1007/978-1-4612-1494-6

[26] WANG, D. and LOH, W.-L. (2011). On fixed-domain asymptotics and covariance tapering in Gaussian random field models. *Electron. J. Stat.* **5** 238–269. MR2792553 https://doi.org/10.1214/11-EJS607

[27] ZHANG, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *J. Amer. Statist. Assoc.* **99** 250–261. MR2054303 https://doi.org/10.1198/016214504000000241

The Annals of Statistics
2021, Vol. 49, No. 6, 3153–3180
https://doi.org/10.1214/21-AOS2078

ADAPTIVE LEARNING RATES FOR SUPPORT VECTOR MACHINES WORKING ON DATA WITH LOW INTRINSIC DIMENSION

BY THOMAS HAMM* AND INGO STEINWART†

Institute for Stochastics and Applications, University of Stuttgart, *thomas.hamm@mathematik.uni-stuttgart.de; †ingo.steinwart@mathematik.uni-stuttgart.de*

We derive improved regression and classification rates for support vector machines using Gaussian kernels under the assumption that the data has some low-dimensional intrinsic structure that is described by the box-counting dimension. Under some standard regularity assumptions for regression and classification, we prove learning rates, in which the dimension of the ambient space is replaced by the box-counting dimension of the support of the data generating distribution. In the regression case, our rates are in some cases minimax optimal up to logarithmic factors, whereas in the classification case our rates are minimax optimal up to logarithmic factors in a certain range of our assumptions and otherwise of the form of the best known rates. Furthermore, we show that a training validation approach for choosing the hyperparameters of a SVM in a data dependent way achieves the same rates adaptively, that is, without any knowledge on the data generating distribution.

REFERENCES

[1] ADAMS, R. A. and FOURNIER, J. J. F. (2003). *Sobolev Spaces*, 2nd ed. *Pure and Applied Mathematics* (*Amsterdam*) **140**. Elsevier/Academic Press, Amsterdam. MR2424078

[2] 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

[3] BERGH, J. and LÖFSTRÖM, J. (1976). *Interpolation Spaces. An Introduction. Grundlehren der Mathematischen Wissenschaften* **223** Springer, Berlin. MR0482275

[4] BEZDEK, K. (1983). Über einige Kreisüberdeckungen. *Beitr. Algebra Geom.* **14** 7–13. MR0695880

[5] BICKEL, P. J. and LI, B. (2007). Local polynomial regression on unknown manifolds. In *Complex Datasets and Inverse Problems. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **54** 177–186. IMS, Beachwood, OH. MR2459188 https://doi.org/10.1214/074921707000000148

[6] BLASCHZYK, I. and STEINWART, I. (2018). Improved classification rates under refined margin conditions. *Electron. J. Stat.* **12** 793–823. MR3770888 https://doi.org/10.1214/18-EJS1406

[7] BOUSQUET, O. (2003). New approaches to statistical learning theory. *Ann. Inst. Statist. Math.* **55** 371–389. MR2001870 https://doi.org/10.1023/A:1026303510251

[8] CARL, B. and STEPHANI, I. (1990). *Entropy, Compactness and the Approximation of Operators. Cambridge Tracts in Mathematics* **98**. Cambridge Univ. Press, Cambridge. MR1098497 https://doi.org/10.1017/CBO9780511897467

[9] DASGUPTA, S. and FREUND, Y. (2008). Random projection trees and low dimensional manifolds. In *STOC'*08 537–546. ACM, New York. MR2582911 https://doi.org/10.1145/1374376.1374452

[10] DE VITO, S., PIGA, M., MARTINOTTO, L. and DI FRANCIA, G. (2009). CO, NO2 and NOx urban pollution monitoring with on-field calibrated electronic nose by automatic Bayesian regularization. *Sens. Actuators B, Chem.* **143** 182–191. https://doi.org/10.1016/j.snb.2009.08.041

[11] EBERTS, M. and STEINWART, I. (2013). Optimal regression rates for SVMs using Gaussian kernels. *Electron. J. Stat.* **7** 1–42. MR3020412 https://doi.org/10.1214/12-EJS760

[12] FALCONER, K. (2003). *Fractal Geometry: Mathematical Foundations and Applications*, 2nd ed. Wiley, Hoboken, NJ. MR2118797 https://doi.org/10.1002/0470013850

[13] FARMER, J. D., OTT, E. and YORKE, J. A. (1983). The dimension of chaotic attractors. *Phys. D* **7** 153–180. MR0719051 https://doi.org/10.1016/0167-2789(83)90125-2

MSC2020 subject classifications. Primary 68T05; secondary 62H30, 68Q32, 62G08.

Key words and phrases. Curse of dimensionality, support vector machines, learning rates, regression, classification.

[14] FAROOQ, M. and STEINWART, I. (2019). Learning rates for kernel-based expectile regression. *Mach. Learn.* **108** 203–227. MR3918701 https://doi.org/10.1007/s10994-018-5762-9

[15] FOLLAND, G. B. (1999). *Real Analysis*: *Modern Techniques and Their Applications*, 2nd ed. *Pure and Applied Mathematics* (*New York*). Wiley, New York. MR1681462

[16] 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

[17] HANG, H. and STEINWART, I. (2018). Optimal learning with anisotropic Gaussian SVMs. Preprint. Available at arXiv:1810.02321.

[18] HÖRMANDER, L. (2003). *The Analysis of Linear Partial Differential Operators*. *I*: *Distribution Theory and Fourier Analysis*. *Classics in Mathematics*. Springer, Berlin. MR1996773 https://doi.org/10.1007/978-3-642-61497-2

[19] KANAGAWA, M., HENNIG, P., SEJDINOVIC, D. and SRIPERUMBUDUR, B. K. (2018). Gaussian processes and kernel methods: A review on connections and equivalences. Preprint. Available at arXiv:1805.08845v1.

[20] KPOTUFE, S. (2011). k-NN regression adapts to local intrinsic dimension. In *Advances in Neural Information Processing Systems* 24 (J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira and K. Q. Weinberger, eds.) 729–737. Curran Associates, Red Hook.

[21] KPOTUFE, S. and DASGUPTA, S. (2012). A tree-based regressor that adapts to intrinsic dimension. *J. Comput. System Sci.* **78** 1496–1515. MR2926146 https://doi.org/10.1016/j.jcss.2012.01.002

[22] KPOTUFE, S. and GARG, V. (2013). Adaptivity to local smoothness and dimension in kernel regression. In *Advances in Neural Information Processing Systems* 26 (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger, eds.) 3075–3083. Curran Associates, Red Hook.

[23] KÜHN, T. (2011). Covering numbers of Gaussian reproducing kernel Hilbert spaces. *J. Complexity* **27** 489–499. MR2805533 https://doi.org/10.1016/j.jco.2011.01.005

[24] KULKARNI, S. R. and POSNER, S. E. (1995). Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Trans. Inf. Theory* **41** 1028–1039. MR1366756 https://doi.org/10.1109/18.391248

[25] MAMMEN, E. and TSYBAKOV, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.* **27** 1808–1829. MR1765618 https://doi.org/10.1214/aos/1017939240

[26] MCGUINNESS, M. J. (1983). The fractal dimension of the Lorenz attractor. *Phys. Lett. A* **99** 5–9. MR0726508 https://doi.org/10.1016/0375-9601(83)90052-X

[27] MILNOR, J. (1985). On the concept of attractor. *Comm. Math. Phys.* **99** 177–195. MR0790735

[28] REED, M. and SIMON, B. (1975). *Methods of Modern Mathematical Physics. II. Fourier Analysis, Self-Adjointness*. Academic Press [Harcourt Brace Jovanovich, Publishers], New York–London. MR0493420

[29] RUELLE, D. (1979/80). Strange attractors. *Math. Intelligencer* **2** 126–137. MR0595080 https://doi.org/10.1007/BF03023053

[30] SCHOENBERG, I. J. (1938). Metric spaces and completely monotone functions. *Ann. of Math.* (2) **39** 811–841. MR1503439 https://doi.org/10.2307/1968466

[31] SCOTT, C. and NOWAK, R. D. (2006). Minimax-optimal classification with dyadic decision trees. *IEEE Trans. Inf. Theory* **52** 1335–1353. MR2241192 https://doi.org/10.1109/TIT.2006.871056

[32] SPIVAK, M. (1979). *A Comprehensive Introduction to Differential Geometry. Vol. I*, 2nd ed. Publish or Perish, Inc., Wilmington, Del. MR0532830

[33] STEINWART, I. and CHRISTMANN, A. (2008). *Support Vector Machines*. Springer, New York. MR2450103

[34] STEINWART, I. and CHRISTMANN, A. (2011). Estimating conditional quantiles with the help of the pinball loss. *Bernoulli* **17** 211–225. MR2797989 https://doi.org/10.3150/10-BEJ267

[35] STEINWART, I. and SCOVEL, C. (2007). Fast rates for support vector machines using Gaussian kernels. *Ann. Statist.* **35** 575–607. MR2336860 https://doi.org/10.1214/009053606000001226

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

[37] TRIEBEL, H. (1992). *Theory of Function Spaces. II. Monographs in Mathematics* **84**. Birkhäuser, Basel. MR1163193 https://doi.org/10.1007/978-3-0346-0419-2

[38] TSYBAKOV, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** 135–166. MR2051002 https://doi.org/10.1214/aos/1079120131

[39] URNER, R. and BEN-DAVID, S. (2013). Probabilistic Lipschitzness: A niceness assumption for deterministic labels. In *Learning Faster from Easy Data Workshop@NIPS*.

[40] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.* **37** 2655–2675. MR2541442 https://doi.org/10.1214/08-AOS678

[41] YANG, Y. and DUNSON, D. B. (2016). Bayesian manifold regression. *Ann. Statist.* **44** 876–905. MR3476620 https://doi.org/10.1214/15-AOS1390

[42] YE, G.-B. and ZHOU, D.-X. (2008). Learning and approximation by Gaussians on Riemannian manifolds. *Adv. Comput. Math.* **29** 291–310. MR2438346 https://doi.org/10.1007/s10444-007-9049-0

[43] YE, G.-B. and ZHOU, D.-X. (2009). SVM learning and L^p approximation by Gaussians on Riemannian manifolds. *Anal. Appl.* (*Singap.*) **7** 309–339. MR2542742 https://doi.org/10.1142/S0219530509001384

The Annals of Statistics
2021, Vol. 49, No. 6, 3181–3205
https://doi.org/10.1214/21-AOS2079

COMMUNITY DETECTION ON MIXTURE MULTILAYER NETWORKS VIA REGULARIZED TENSOR DECOMPOSITION

BY BING-YI JING[1,*], TING LI[2], ZHONGYUAN LYU[1,†] AND DONG XIA[1,‡]

[1]*Department of Mathematics, The Hong Kong University of Science and Technology, [*]majing@ust.hk; [†]zlyuab@connect.ust.hk; [‡]madxia@ust.hk*

[2]*Department of Applied Mathematics, The Hong Kong Polytechnic University, tingeric.li@polyu.edu.hk*

We study the problem of community detection in multilayer networks, where pairs of nodes can be related in multiple modalities. We introduce a general framework, that is, mixture multilayer stochastic block model (MMSBM), which includes many earlier models as special cases. We propose a tensor-based algorithm (TWIST) to reveal both global/local memberships of nodes, and memberships of layers. We show that the TWIST procedure can accurately detect the communities with small misclassification error as the number of nodes and/or number of layers increases. Numerical studies confirm our theoretical findings. To our best knowledge, this is the first systematic study on the mixture multilayer networks using tensor decomposition. The method is applied to two real datasets: worldwide trading networks and malaria parasite genes networks, yielding new and interesting findings.

REFERENCES

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

[2] ARROYO, J., ATHREYA, A., CAPE, J., CHEN, G., PRIEBE, C. E. and VOGELSTEIN, J. T. (2019). Inference for multiple heterogeneous networks with a common invariant subspace. Preprint. Available at arXiv:1906.10026.

[3] ATHREYA, A., PRIEBE, C. E., TANG, M., LYZINSKI, V., MARCHETTE, D. J. and SUSSMAN, D. L. (2016). A limit theorem for scaled eigenvectors of random dot product graphs. *Sankhya A* **78** 1–18. MR3494576 https://doi.org/10.1007/s13171-015-0071-x

[4] BACCO, C. D., POWER, E. A., LARREMORE, D. B. and MOORE, C. (2017). Community detection, link prediction, and layer interdependence in multilayer networks. *Phys. Rev. Lett.* **95** 042317. https://doi.org/10.1103/PhysRevE.95.042317

[5] BEN AROUS, G., MEI, S., MONTANARI, A. and NICA, M. (2019). The landscape of the spiked tensor model. *Comm. Pure Appl. Math.* **72** 2282–2330. MR4011861 https://doi.org/10.1002/cpa.21861

[6] BHATTACHARYYA, S. and CHATTERJEE, S. (2018). Spectral clustering for multiple sparse networks: I. Preprint. Available at arXiv:1805.10594.

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

[8] 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

[9] CAI, C., LI, G., CHI, Y., POOR, H. V. and CHEN, Y. (2021). Subspace estimation from unbalanced and incomplete data matrices: $\ell_{2,\infty}$ statistical guarantees. *Ann. Statist.* **49** 944–967. MR4255114 https://doi.org/10.1214/20-aos1986

[10] CAI, C., LI, G., POOR, H. V. and CHEN, Y. (2019). Nonconvex low-rank tensor completion from noisy data. In *Advances in Neural Information Processing Systems* 1861–1872.

[11] CATTELL, R. B. (1966). The scree test for the number of factors. *Multivar. Behav. Res.* **1** 245–276.

[12] CHEN, Y., FAN, J., MA, C. and WANG, K. (2019). Spectral method and regularized MLE are both optimal for top-K ranking. *Ann. Statist.* **47** 2204–2235. MR3953449 https://doi.org/10.1214/18-AOS1745

[13] DOMENICO, M. D., NICOSIA, V., ARENAS, A. and LATORA, V. (2015). Structural reducibility of multilayer networks. *Nat. Commun.* **6** 6864. https://doi.org/10.1038/ncomms7864

MSC2020 subject classifications. 62H30, 91C20.
Key words and phrases. Network community detection, multilayer network, tensor, tucker decomposition.

[14] DONG, X., FROSSARD, P., VANDERGHEYNST, P. and NEFEDOV, N. (2012). Clustering with multi-layer graphs: A spectral perspective. *IEEE Trans. Signal Process.* **60** 5820–5831. MR2990287 https://doi.org/10.1109/TSP.2012.2212886

[15] FAN, X., PENSKY, M., YU, F. and ZHANG, T. (2021). ALMA: Alternating minimization algorithm for clustering mixture multilayer network. Preprint. Available at arXiv:2102.10226.

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

[17] HAN, Q., XU, K. and AIROLDI, E. (2015). Consistent estimation of dynamic and multi-layer block models. In *International Conference on Machine Learning* 1511–1520.

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

[19] JAIN, P. and OH, S. (2014). Provable tensor factorization with missing data. In *Advances in Neural Information Processing Systems* 1431–1439.

[20] JIN, J. (2015). Fast community detection by SCORE. *Ann. Statist.* **43** 57–89. MR3285600 https://doi.org/10.1214/14-AOS1265

[21] JIN, J., KE, Z. T. and LUO, S. (2017). Estimating network memberships by simplex vertex hunting. Preprint. Available at arXiv:1708.07852.

[22] JING, B.-Y., LI, T., LYU, Z. and XIA, D. (2021). Supplement to "Community detection on mixture multilayer networks via regularized tensor decomposition." https://doi.org/10.1214/21-AOS2079SUPP

[23] JOSSE, J. and HUSSON, F. (2012). Selecting the number of components in principal component analysis using cross-validation approximations. *Comput. Statist. Data Anal.* **56** 1869–1879. MR2892383 https://doi.org/10.1016/j.csda.2011.11.012

[24] KE, Z. T., SHI, F. and XIA, D. (2019). Community detection for hypergraph networks via regularized tensor power iteration. Preprint. Available at arXiv:1909.06503.

[25] KIM, C., BANDEIRA, A. S. and GOEMANS, M. X. (2018). Stochastic block model for hypergraphs: Statistical limits and a semidefinite programming approach. Preprint. Available at arXiv:1807.02884.

[26] KIVELÄ, M., ARENAS, A., BARTHELEMY, M., GLEESON, J. P., MORENO, Y. and PORTER, M. A. (2014). Multilayer networks. *J. Complex Netw.* **2** 203–271.

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

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

[29] KUMAR, A., RAI, P. and DAUME, H. (2011). Co-regularized multi-view spectral clustering. In *Advances in Neural Information Processing Systems* 1413–1421.

[30] LARREMORE, D. B., CLAUSET, A. and BUCKEE, C. O. (2013). A network approach to analyzing highly recombinant malaria parasite genes. *PLoS Comput. Biol.* **9** e1003268. https://doi.org/10.1371/journal.pcbi.1003268

[31] LE, C. M., LEVIN, K. and LEVINA, E. (2018). Estimating a network from multiple noisy realizations. *Electron. J. Stat.* **12** 4697–4740. MR3894068 https://doi.org/10.1214/18-ejs1521

[32] 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

[33] LEI, J., CHEN, K. and LYNCH, B. (2020). Consistent community detection in multi-layer network data. *Biometrika* **107** 61–73. MR4064140 https://doi.org/10.1093/biomet/asz068

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

[35] LEI, L., LI, X. and LOU, X. (2020). Consistency of spectral clustering on hierarchical stochastic block models. Preprint. Available at arXiv:2004.14531.

[36] LEVIN, K., LODHIA, A. and LEVINA, E. (2019). Recovering low-rank structure from multiple networks with unknown edge distributions. Preprint. Available at arXiv:1906.07265.

[37] LI, J., BIEN, J. and WELLS, M. T. (2018). rTensor: An R package for multidimensional array (tensor) unfolding, multiplication, and decomposition. *J. Stat. Softw.* **87** 1–31.

[38] LI, T., LEI, L., BHATTACHARYYA, S., SARKAR, P., BICKEL, P. J. and LEVINA, E. (2018). Hierarchical community detection by recursive partitioning. Preprint. Available at arXiv:1810.01509.

[39] LYZINSKI, V., TANG, M., ATHREYA, A., PARK, Y. and PRIEBE, C. E. (2017). Community detection and classification in hierarchical stochastic blockmodels. *IEEE Trans. Netw. Sci. Eng.* **4** 13–26. MR3625952 https://doi.org/10.1109/TNSE.2016.2634322

[40] MATIAS, C. and MIELE, V. (2017). Statistical clustering of temporal networks through a dynamic stochastic block model. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1119–1141. MR3689311 https://doi.org/10.1111/rssb.12200

[41] NICKEL, M., TRESP, V. and KRIEGEL, H.-P. (2011). A three-way model for collective learning on multi-relational data. In *ICML* **11** 809–816.
[42] PAUL, S. and CHEN, Y. (2016). Consistent community detection in multi-relational data through restricted multi-layer stochastic blockmodel. *Electron. J. Stat.* **10** 3807–3870. MR3579677 https://doi.org/10.1214/16-EJS1211
[43] PAUL, S. and CHEN, Y. (2020). Spectral and matrix factorization methods for consistent community detection in multi-layer networks. *Ann. Statist.* **48** 230–250. MR4065160 https://doi.org/10.1214/18-AOS1800
[44] PAUL, S. and CHEN, Y. (2020). A random effects stochastic block model for joint community detection in multiple networks with applications to neuroimaging. *Ann. Appl. Stat.* **14** 993–1029. MR4117838 https://doi.org/10.1214/20-AOAS1339
[45] PENSKY, M. and ZHANG, T. (2019). Spectral clustering in the dynamic stochastic block model. *Electron. J. Stat.* **13** 678–709. MR3914178 https://doi.org/10.1214/19-ejs1533
[46] RICHARD, E. and MONTANARI, A. (2014). A statistical model for tensor PCA. In *Advances in Neural Information Processing Systems* 2897–2905.
[47] 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
[48] SHEEHAN, B. N. and SAAD, Y. (2007). Higher order orthogonal iteration of tensors (HOOI) and its relation to PCA and GLRAM. In *Proceedings of the* 2007 *SIAM International Conference on Data Mining* 355–365. SIAM, Philadelphia.
[49] STANLEY, N., SHAI, S., TAYLOR, D. and MUCHA, P. J. (2016). Clustering network layers with the strata multilayer stochastic block model. *IEEE Trans. Netw. Sci. Eng.* **3** 95–105. MR3515211 https://doi.org/10.1109/TNSE.2016.2537545
[50] SUN, W. W., LU, J., LIU, H. and CHENG, G. (2017). Provable sparse tensor decomposition. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 899–916. MR3641413 https://doi.org/10.1111/rssb.12190
[51] TANG, R., TANG, M., VOGELSTEIN, J. T. and PRIEBE, C. E. (2017). Robust estimation from multiple graphs under gross error contamination. Preprint. Available at arXiv:1707.03487.
[52] TANG, W., LU, Z. and DHILLON, I. S. (2009). Clustering with multiple graphs. In 2009 *Ninth IEEE International Conference on Data Mining* 1016–1021. IEEE, New York.
[53] WANG, D., YU, Y. and RINALDO, A. (2018). Optimal change point detection and localization in sparse dynamic networks. Preprint. Available at arXiv:1809.09602.
[54] WANG, M. and LI, L. (2020). Learning from binary multiway data: Probabilistic tensor decomposition and its statistical optimality. *J. Mach. Learn. Res.* **21** Paper No. 154. MR4209440
[55] XIA, D. and YUAN, M. (2019). On polynomial time methods for exact low-rank tensor completion. *Found. Comput. Math.* **19** 1265–1313. MR4029842 https://doi.org/10.1007/s10208-018-09408-6
[56] XIA, D. and YUAN, M. (2021). Statistical inferences of linear forms for noisy matrix completion. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **83** 58–77. MR4220984 https://doi.org/10.1111/rssb.12400
[57] XIA, D., YUAN, M. and ZHANG, C.-H. (2021). Statistically optimal and computationally efficient low rank tensor completion from noisy entries. *Ann. Statist.* **49** 76–99. MR4206670 https://doi.org/10.1214/20-AOS1942
[58] XIA, D. and ZHOU, F. (2019). The sup-norm perturbation of HOSVD and low rank tensor denoising. *J. Mach. Learn. Res.* **20** Paper No. 61. MR3960915
[59] YUAN, M. and ZHANG, C.-H. (2017). Incoherent tensor norms and their applications in higher order tensor completion. *IEEE Trans. Inf. Theory* **63** 6753–6766. MR3707566 https://doi.org/10.1109/TIT.2017.2724549
[60] ZHANG, A. (2019). Cross: Efficient low-rank tensor completion. *Ann. Statist.* **47** 936–964. MR3909956 https://doi.org/10.1214/18-AOS1694
[61] ZHANG, A. and HAN, R. (2019). Optimal sparse singular value decomposition for high-dimensional high-order data. *J. Amer. Statist. Assoc.* **114** 1708–1725. MR4047294 https://doi.org/10.1080/01621459.2018.1527227
[62] ZHANG, A. and XIA, D. (2018). Tensor SVD: Statistical and computational limits. *IEEE Trans. Inf. Theory* **64** 7311–7338. MR3876445 https://doi.org/10.1109/TIT.2018.2841377
[63] ZHANG, J. and CAO, J. (2017). Finding common modules in a time-varying network with application to the *Drosophila melanogaster* gene regulation network. *J. Amer. Statist. Assoc.* **112** 994–1008. MR3735355 https://doi.org/10.1080/01621459.2016.1260465
[64] ZHANG, Y. and XIA, D. (2020). Edgeworth expansions for network moments. Preprint. Available at arXiv:2004.06615.

The Annals of Statistics
2021, Vol. 49, No. 6, 3206–3227
https://doi.org/10.1214/21-AOS2080

AUGMENTED MINIMAX LINEAR ESTIMATION

BY DAVID A. HIRSHBERG[*] AND STEFAN WAGER[†]

Graduate School of Business, Stanford University, [*]*davidahirshberg@stanford.edu;* [†]*swager@stanford.edu*

Many statistical estimands can expressed as continuous linear functionals of a conditional expectation function. This includes the average treatment effect under unconfoundedness and generalizations for continuous-valued and personalized treatments. In this paper, we discuss a general approach to estimating such quantities: we begin with a simple plug-in estimator based on an estimate of the conditional expectation function, and then correct the plug-in estimator by subtracting a minimax linear estimate of its error. We show that our method is semiparametrically efficient under weak conditions and observe promising performance on both real and simulated data.

REFERENCES

ARMSTRONG, T. B. and KOLESÁR, M. (2018). Optimal inference in a class of regression models. *Econometrica* **86** 655–683. MR3783342 https://doi.org/10.3982/ECTA14434

ARMSTRONG, T. B. and KOLESÁR, M. (2021). Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness. *Econometrica* **89** 1141–1177. MR4325179 https://doi.org/10.3982/ECTA16907

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

ATHEY, S., TIBSHIRANI, J. and WAGER, S. (2019). Generalized random forests. *Ann. Statist.* **47** 1148–1178. MR3909963 https://doi.org/10.1214/18-AOS1709

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York. Reprint of the 1993 original. MR1623559

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

CAI, T. T. and LOW, M. G. (2003). A note on nonparametric estimation of linear functionals. *Ann. Statist.* 1140–1153.

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

CASSEL, C. M., SÄRNDAL, C. E. and WRETMAN, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika* **63** 615–620. MR0445666 https://doi.org/10.1093/biomet/63.3.615

CHAN, K. C. G., YAM, S. C. P. and ZHANG, Z. (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 673–700. MR3506798 https://doi.org/10.1111/rssb.12129

CHEN, X., HONG, H. and TAROZZI, A. (2008). Semiparametric efficiency in GMM models with auxiliary data. *Ann. Statist.* **36** 808–843. MR2396816 https://doi.org/10.1214/009053607000000947

CHERNOZHUKOV, V., NEWEY, W. and ROBINS, J. (2018). Double/de-biased machine learning using regularized Riesz representers. Preprint. Available at arXiv:1802.08667.

CHERNOZHUKOV, V., ESCANCIANO, J. C., ICHIMURA, H. and NEWEY, W. K. (2016). Locally robust semiparametric estimation. Preprint. Available at arXiv:1608.00033.

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

MSC2020 subject classifications. 62F12.
Key words and phrases. Causal inference, convex optimization, semiparametric efficiency.

CRUMP, R. K., HOTZ, V. J., IMBENS, G. W. and MITNIK, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96** 187–199. MR2482144 https://doi.org/10.1093/biomet/asn055

D'AMOUR, A., DING, P., FELLER, A., LEI, L. and SEKHON, J. (2021). Overlap in observational studies with high-dimensional covariates. *J. Econometrics* **221** 644–654. MR4215042 https://doi.org/10.1016/j.jeconom.2019.10.014

DOMAHIDI, A., CHU, E. and BOYD, S. (2013). ECOS: An SOCP solver for embedded systems. In *European Control Conference* (*ECC*) 3071–3076.

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

DONOHO, D. L. and LIU, R. C. (1991). Geometrizing rates of convergence. III. *Ann. Statist.* **19** 668–701. MR1105839 https://doi.org/10.1214/aos/1176348114

FAN, J., IMAI, K., LIU, H., NING, Y. and YANG, X. (2016). Improving covariate balancing propensity score: A doubly robust and efficient approach. Technical Report, Princeton Univ.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

FU, A., NARASIMHAN, B., DIAMOND, S. and MILLER, J. (2017). CVXR: Disciplined Convex Optimization. R package version 0.94-4.

GRAHAM, B. S., DE XAVIER PINTO, C. C. and EGEL, D. (2012). Inverse probability tilting for moment condition model with missing data. *Rev. Econ. Stud.* **79** 1053–1079. MR2986390 https://doi.org/10.1093/restud/rdr047

GRAHAM, B. S. and PINTO, C. C. D. X. (2018). Semiparametrically efficient estimation of the average linear regression function. Technical Report, National Bureau of Economic Research.

GRAHAM, B. S., PINTO, C. C. D. X. and EGEL, D. (2016). Efficient estimation of data combination models by the method of auxiliary-to-study tilting (AST). *J. Bus. Econom. Statist.* **34** 288–301. MR3475879 https://doi.org/10.1080/07350015.2015.1038544

HAINMUELLER, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Polit. Anal.* **20** 25–46.

HIRANO, K., IMBENS, G. W. and RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71** 1161–1189. MR1995826 https://doi.org/10.1111/1468-0262.00442

HIRSHBERG, D. A., MALEKI, A. and ZUBIZARRETA, J. (2019). Minimax linear estimation of the retargeted mean. Preprint. Available at arXiv:1901.10296.

HIRSHBERG, D. A. and WAGER, S. (2021). Supplement to "Augmented minimax linear estimation." https://doi.org/10.1214/21-AOS2080SUPP

IBRAGIMOV, I. A. and KHAS'MINSKIĬ, R. Z. (1984). Nonparametric estimation of the value of a linear functional in Gaussian white noise. *Theory Probab. Appl.* **29** 18–32.

IMAI, K. and RATKOVIC, M. (2014). Covariate balancing propensity score. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 243–263. MR3153941 https://doi.org/10.1111/rssb.12027

IMBENS, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **87** 706–710. MR1789821 https://doi.org/10.1093/biomet/87.3.706

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

IMBENS, G. W., RUBIN, D. B. and SACERDOTE, B. I. (2001). Estimating the effect of unearned income on labor earnings, savings, and consumption: Evidence from a survey of lottery players. *Am. Econ. Rev.* **91** 778–794.

IMBENS, G. and WAGER, S. (2019). Optimized regression discontinuity designs. *Rev. Econ. Stat.* **101** 264–278.

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

JOHNSTONE, I. M. (2015). Gaussian estimation: Sequence and wavelet models. Manuscript.

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

KALLUS, N. (2018). Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems* 8909–8920.

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

KENNEDY, E. H. (2020). Optimal doubly robust estimation of heterogeneous causal effects. Preprint. Available at arXiv:2004.14497.

LANG, S. (1993). *Real and Functional Analysis*, 3rd ed. *Graduate Texts in Mathematics* **142**. Springer, New York. MR1216137 https://doi.org/10.1007/978-1-4612-0897-6

LECUÉ, G. and MENDELSON, S. (2018). Regularization and the small-ball method I: Sparse recovery. *Ann. Statist*. **46** 611–641. MR3782379 https://doi.org/10.1214/17-AOS1562

LI, F., MORGAN, K. L. and ZASLAVSKY, A. M. (2018). Balancing covariates via propensity score weighting. *J. Amer. Statist. Assoc*. **113** 390–400. MR3803473 https://doi.org/10.1080/01621459.2016.1260466

LUGOSI, G. and ZEGER, K. (1995). Nonparametric estimation via empirical risk minimization. *IEEE Trans. Inf. Theory* **41** 677–687. MR1331260 https://doi.org/10.1109/18.382014

NEWEY, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica* **62** 1349–1382. MR1303237 https://doi.org/10.2307/2951752

NEWEY, W. K. and ROBINS, J. R. (2018). Cross-fitting and fast remainder rates for semiparametric estimation. Preprint. Available at arXiv:1801.09138.

NIE, X. and WAGER, S. (2017). Quasi-oracle estimation of heterogeneous treatment effects. Preprint. Available at arXiv:1712.04912.

NING, Y., PENG, S. and IMAI, K. (2017). High dimensional propensity score estimation via covariate balancing.

PEYPOUQUET, J. (2015). *Convex Optimization in Normed Spaces: Theory, Methods and Examples*. *SpringerBriefs in Optimization*. Springer, Cham. MR3310025 https://doi.org/10.1007/978-3-319-13710-0

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

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

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

ROBINS, J., SUED, M., LEI-GOMEZ, Q. and ROTNITZKY, A. (2007). Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable [MR2420458]. *Statist. Sci*. **22** 544–559. MR2420460 https://doi.org/10.1214/07-STS227D

ROBINS, J., TCHETGEN TCHETGEN, E., LI, L. and VAN DER VAART, A. (2009). Semiparametric minimax rates. *Electron. J. Stat*. **3** 1305–1321. MR2566189 https://doi.org/10.1214/09-EJS479

ROBINS, J. M., LI, L., MUKHERJEE, R., TCHETGEN, E. T. and VAN DER VAART, A. (2017). Minimax estimation of a functional on a structured high-dimensional model. *Ann. Statist*. **45** 1951–1987. MR3718158 https://doi.org/10.1214/16-AOS1515

ROBINSON, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica* **56** 931–954. MR0951762 https://doi.org/10.2307/1912705

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974 https://doi.org/10.1093/biomet/70.1.41

ROSENBAUM, P. R. and RUBIN, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *J. Amer. Statist. Assoc*. **79** 516–524.

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

STONE, C. J. (1977). Consistent nonparametric regression. *Ann. Statist*. **5** 595–645. With discussion and a reply by the author. MR0443204

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

VAN DER LAAN, M. J., BENKESER, D. and CAI, W. (2019). Efficient estimation of pathwise differentiable target parameters with the undersmoothed highly adaptive lasso. Preprint. Available at arXiv:1908.05607.

VAN DER LAAN, M. J. and RUBIN, D. (2006). Targeted maximum likelihood learning. *Int. J. Biostat*. **2** Art. 11, 40. MR2306500 https://doi.org/10.2202/1557-4679.1043

VAN DER VAART, A. (1991). On differentiable functionals. *Ann. Statist*. **19** 178–204. MR1091845 https://doi.org/10.1214/aos/1176347976

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

WANG, Y. and ZUBIZARRETA, J. R. (2020). Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations. *Biometrika* **107** 93–105. MR4064142 https://doi.org/10.1093/biomet/asz050

WONG, R. K. W. and CHAN, K. C. G. (2018). Kernel-based covariate functional balancing for observational studies. *Biometrika* **105** 199–213. MR3768874 https://doi.org/10.1093/biomet/asx069

ZEILEIS, A. (2004). Econometric computing with HC and HAC covariance matrix estimators. *J. Stat. Softw*. **11** 1–17.

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

ZHAO, Q. (2019). Covariate balancing propensity score by tailored loss functions. *Ann. Statist.* **47** 965–993. MR3909957 https://doi.org/10.1214/18-AOS1698

ZHAO, Q., SMALL, D. S. and ERTEFAIE, A. (2017). Selective inference for effect modification via the lasso. Preprint. Available at arXiv:1705.08020.

ZHENG, W. and VAN DER LAAN, M. J. (2011). Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*. *Springer Ser. Statist.* 459–474. Springer, New York. MR2867139 https://doi.org/10.1007/978-1-4419-9782-1_27

ZUBIZARRETA, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *J. Amer. Statist. Assoc.* **110** 910–922. MR3420672 https://doi.org/10.1080/01621459.2015.1023805

The Annals of Statistics
2021, Vol. 49, No. 6, 3228–3254
https://doi.org/10.1214/21-AOS2081

WILKS' THEOREM FOR SEMIPARAMETRIC REGRESSIONS WITH WEAKLY DEPENDENT DATA

BY MARIE DU ROY DE CHAUMARAY[*], MATTHIEU MARBAC[†] AND VALENTIN PATILEA[‡]

Univ. Rennes, Ensai, CNRS, CREST—UMR 9194, [*]marie.du-roy-de-chaumaray@ensai.fr; [†]matthieu.marbac-lourdelle@ensai.fr; [‡]valentin.patilea@ensai.fr

The empirical likelihood inference is extended to a class of semiparametric models for stationary, weakly dependent series. A partially linear single-index regression is used for the conditional mean of the series given its past, and the present and past values of a vector of covariates. A parametric model for the conditional variance of the series is added to capture further nonlinear effects. We propose suitable moment equations which characterize the mean and variance model. We derive an empirical log-likelihood ratio which includes nonparametric estimators of several functions, and we show that this ratio behaves asymptotically as if the functions were given.

REFERENCES

[1] BOSSAERTS, P., HAFNER, C. and HÄRDLE, W. (1996). A New Method for Volatility Estimation with Applications in Foreign Exchange Rate Series. In *Finanzmarktanalyse und -prognose Mit Innovativen Quantitativen Verfahren*: *Ergebnisse des* 5. *Karlsruher Ökonometrie-Workshops* 71–83. Physica-Verlag HD, Heidelberg.

[2] BRADLEY, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surv.* **2** 107–144. MR2178042 https://doi.org/10.1214/154957805100000104

[3] BRAVO, F., ESCANCIANO, J. C. and VAN KEILEGOM, I. (2020). Two-step semiparametric empirical likelihood inference. *Ann. Statist.* **48** 1–26. MR4065150 https://doi.org/10.1214/18-AOS1788

[4] CARROLL, R. J., FAN, J., GIJBELS, I. and WAND, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92** 477–489. MR1467842 https://doi.org/10.2307/2965697

[5] CHANG, J., CHEN, S. X. and CHEN, X. (2015). High dimensional generalized empirical likelihood for moment restrictions with dependent data. *J. Econometrics* **185** 283–304. MR3300347 https://doi.org/10.1016/j.jeconom.2014.10.011

[6] CHANG, J., CHEN, S. X., TANG, C. Y. and WU, T. T. (2021). High-dimensional empirical likelihood inference. *Biometrika* **108** 127–147. MR4226194 https://doi.org/10.1093/biomet/asaa051

[7] CHANG, J., TANG, C. Y. and WU, Y. (2013). Marginal empirical likelihood and sure independence feature screening. *Ann. Statist.* **41** 2123–2148. MR3127860 https://doi.org/10.1214/13-AOS1139

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

[9] CHEN, S. X. and VAN KEILEGOM, I. (2009). A review on empirical likelihood methods for regression. *TEST* **18** 415–447. MR2566404 https://doi.org/10.1007/s11749-009-0159-5

[10] CHEN, X. and CUI, H. (2008). Empirical likelihood inference for partial linear models under martingale difference sequence. *Statist. Probab. Lett.* **78** 2895–2901. MR2516809 https://doi.org/10.1016/j.spl.2008.04.012

[11] DONG, C., GAO, J. and TJØSTHEIM, D. (2016). Estimation for single-index and partially linear single-index integrated models. *Ann. Statist.* **44** 425–453. MR3449774 https://doi.org/10.1214/15-AOS1372

[12] DU ROY DE CHAUMARAY, M., MARBAC, M. and PATILEA, V. (2021). Supplement to "Wilks' theorem for semiparametric regressions with weakly dependent data." https://doi.org/10.1214/21-AOS2081SUPP

[13] FAN, G.-L. and LIANG, H.-Y. (2010). Empirical likelihood inference for semiparametric model with linear process errors. *J. Korean Statist. Soc.* **39** 55–65. MR2752218 https://doi.org/10.1016/j.jkss.2009.04.001

MSC2020 subject classifications. Primary 62M10; secondary 62F03, 62G20.

Key words and phrases. Kernel smoothing, α-mixing, nonlinear time series, nuisance function, parametric inference, pivotal statistic.

[14] FRYZLEWICZ, P. and SUBBA RAO, S. (2011). Mixing properties of ARCH and time-varying ARCH processes. *Bernoulli* **17** 320–346. MR2797994 https://doi.org/10.3150/10-BEJ270

[15] GASPARRINI, A. (2011). Distributed lag linear and non-linear models in R: The package dlnm. *J. Stat. Softw.* **43** 1–20.

[16] HALL, P. and HEYDE, C. C. (1980). *Martingale Limit Theory and Its Application*: *Probability and Mathematical Statistics*. Academic Press, New York. MR0624435

[17] HANSEN, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory* **24** 726–748. MR2409261 https://doi.org/10.1017/S0266466608080304

[18] HÄRDLE, W., LÜTKEPOHL, H., CHEN, R., HARDLE, W. and LUTKEPOHL, H. (1997). A review of nonparametric time series analysis. *Int. Stat. Rev.* **65** 49–72. https://doi.org/10.2307/1403432

[19] HÄRDLE, W., TSYBAKOV, A. and YANG, L. (1998). Nonparametric vector autoregression. *J. Statist. Plann. Inference* **68** 221–245. MR1629575 https://doi.org/10.1016/S0378-3758(97)00143-2

[20] HJORT, N. L., MCKEAGUE, I. W. and VAN KEILEGOM, I. (2009). Extending the scope of empirical likelihood. *Ann. Statist.* **37** 1079–1111. MR2509068 https://doi.org/10.1214/07-AOS555

[21] KANAI, H., OGATA, H. and TANIGUCHI, M. (2010). Estimating function approach for CHARN models. *Metron* **68** 1–21. MR3037087 https://doi.org/10.1007/bf03263521

[22] KATO, H., TANIGUCHI, M. and HONDA, M. (2006). Statistical analysis for multiplicatively modulated nonlinear autoregressive model and its applications to electrophysiological signal analysis in humans. *IEEE Trans. Signal Process.* **54** 3414–3425. https://doi.org/10.1109/tsp.2006.877663

[23] KITAMURA, Y. (1997). Empirical likelihood methods with weakly dependent processes. *Ann. Statist.* **25** 2084–2102. MR1474084 https://doi.org/10.1214/aos/1069362388

[24] LI, G., ZHU, L., XUE, L. and FENG, S. (2010). Empirical likelihood inference in partially linear single-index models for longitudinal data. *J. Multivariate Anal.* **101** 718–732. MR2575416 https://doi.org/10.1016/j.jmva.2009.08.006

[25] LIAN, H., LIANG, H. and CARROLL, R. J. (2015). Variance function partially linear single-index models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 171–194. MR3299404 https://doi.org/10.1111/rssb.12066

[26] LIANG, H., LIU, X., LI, R. and TSAI, C.-L. (2010). Estimation and testing for partially linear single-index models. *Ann. Statist.* **38** 3811–3836. MR2766869 https://doi.org/10.1214/10-AOS835

[27] LIEBSCHER, E. (2005). Towards a unified approach for proving geometric ergodicity and mixing properties of nonlinear autoregressive processes. *J. Time Series Anal.* **26** 669–689. MR2188304 https://doi.org/10.1111/j.1467-9892.2005.00412.x

[28] LU, X. (2009). Empirical likelihood for heteroscedastic partially linear models. *J. Multivariate Anal.* **100** 387–396. MR2483427 https://doi.org/10.1016/j.jmva.2008.05.006

[29] LU, Z. and JIANG, Z. (2001). L_1 geometric ergodicity of a multivariate nonlinear AR model with an ARCH term. *Statist. Probab. Lett.* **51** 121–130. MR1808034 https://doi.org/10.1016/S0167-7152(00)00138-3

[30] MA, Y. and ZHU, L. (2013). Doubly robust and efficient estimators for heteroscedastic partially linear single-index models allowing high dimensional covariates. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 305–322. MR3021389 https://doi.org/10.1111/j.1467-9868.2012.01040.x

[31] MASRY, E. and TJOSTHEIM, D. (1995). Nonparametric estimation and identification of nonlinear ARCH time series strong convergence and asymptotic normality: Strong convergence and asymptotic normality. *Econometric Theory* **11** 258–289.

[32] MEITZ, M. and SAIKKONEN, P. (2010). A note on the geometric ergodicity of a nonlinear AR-ARCH model. *Statist. Probab. Lett.* **80** 631–638. MR2595140 https://doi.org/10.1016/j.spl.2009.12.020

[33] MOKKADEM, A. (1988). Mixing properties of ARMA processes. *Stochastic Process. Appl.* **29** 309–315. MR0958507 https://doi.org/10.1016/0304-4149(88)90045-2

[34] MOKKADEM, A. (1990). Propriétés de mélange des processus autorégressifs polynomiaux. *Ann. Inst. Henri Poincaré Probab. Stat.* **26** 219–260. MR1063750

[35] NORDMAN, D. J. and LAHIRI, S. N. (2014). A review of empirical likelihood methods for time series. *J. Statist. Plann. Inference* **155** 1–18. MR3264452 https://doi.org/10.1016/j.jspi.2013.10.001

[36] OWEN, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75** 237–249. MR0946049 https://doi.org/10.1093/biomet/75.2.237

[37] OWEN, A. B. (2001). *Empirical Likelihood*. Chapman and Hall/CRC, Boca Raton, FL.

[38] QIN, J. and LAWLESS, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* **22** 300–325. MR1272085 https://doi.org/10.1214/aos/1176325370

[39] RIO, E. (2000). *Théorie Asymptotique des Processus Aléatoires Faiblement Dépendants*. *Mathématiques & Applications* (*Berlin*) [*Mathematics & Applications*] **31**. Springer, Berlin. MR2117923

[40] SEVERINI, T. A. and WONG, W. H. (1992). Profile likelihood and conditionally parametric models. *Ann. Statist.* **20** 1768–1802. MR1193312 https://doi.org/10.1214/aos/1176348889

[41] TJØSTHEIM, D. (1990). Nonlinear time series and Markov chains. *Adv. in Appl. Probab.* **22** 587–611. MR1066965 https://doi.org/10.2307/1427459

[42] TONG, H. (1990). *Nonlinear Time Series*: *A Dynamical System Approach*. *Oxford Statistical Science Series* **6**. The Clarendon Press, Oxford University Press, New York. MR1079320

[43] WANG, Q.-H. and JING, B.-Y. (1999). Empirical likelihood for partial linear models with fixed designs. *Statist*. *Probab*. *Lett*. **41** 425–433. MR1666112 https://doi.org/10.1016/S0167-7152(98)00230-2

[44] WANG, Q.-H. and JING, B.-Y. (2003). Empirical likelihood for partial linear models. *Ann*. *Inst*. *Statist*. *Math*. **55** 585–595. MR2007800 https://doi.org/10.1007/BF02517809

[45] XIA, Y. and HÄRDLE, W. (2006). Semi-parametric estimation of partially linear single-index models. *J*. *Multivariate Anal*. **97** 1162–1184. MR2276153 https://doi.org/10.1016/j.jmva.2005.11.005

[46] XIA, Y., TONG, H. and LI, W. K. (1999). On extended partially linear single-index models. *Biometrika* **86** 831–842. MR1741980 https://doi.org/10.1093/biomet/86.4.831

[47] XUE, L. and ZHU, L. (2007). Empirical likelihood semiparametric regression analysis for longitudinal data. *Biometrika* **94** 921–937. MR2416799 https://doi.org/10.1093/biomet/asm066

[48] XUE, L.-G. and ZHU, L. (2006). Empirical likelihood for single-index models. *J*. *Multivariate Anal*. **97** 1295–1312. MR2279674 https://doi.org/10.1016/j.jmva.2005.09.004

[49] ZHU, L., LIN, L., CUI, X. and LI, G. (2010). Bias-corrected empirical likelihood in a multi-link semiparametric model. *J*. *Multivariate Anal*. **101** 850–868. MR2584904 https://doi.org/10.1016/j.jmva.2009.08.009

[50] ZHU, L. and XUE, L. (2006). Empirical likelihood confidence regions in a partially linear single-index model. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **68** 549–570. MR2278341 https://doi.org/10.1111/j.1467-9868.2006.00556.x

The Annals of Statistics
2021, Vol. 49, No. 6, 3255–3298
https://doi.org/10.1214/21-AOS2082

STATISTICAL GUARANTEES FOR BAYESIAN UNCERTAINTY QUANTIFICATION IN NONLINEAR INVERSE PROBLEMS WITH GAUSSIAN PROCESS PRIORS

BY FRANÇOIS MONARD[1], RICHARD NICKL[2,*] AND GABRIEL P. PATERNAIN[2,†]

[1]*Department of Mathematics, University of California Santa Cruz, fmonard@ucsc.edu*
[2]*Department of Pure Mathematics and Mathematical Statistics, University of Cambridge,* [*]*nickl@maths.cam.ac.uk;* [†]*g.p.paternain@dpmms.cam.ac.uk*

Bayesian inference and uncertainty quantification in a general class of nonlinear inverse regression models is considered. Analytic conditions on the regression model $\{\mathscr{G}(\theta) : \theta \in \Theta\}$ and on Gaussian process priors for θ are provided such that semiparametrically efficient inference is possible for a large class of linear functionals of θ. A general Bernstein–von Mises theorem is proved that shows that the (non-Gaussian) posterior distributions are approximated by certain Gaussian measures centred at the posterior mean. As a consequence, posterior-based credible sets are valid and optimal from a frequentist point of view. The theory is illustrated with two applications with PDEs that arise in nonlinear tomography problems: an elliptic inverse problem for a Schrödinger equation, and inversion of non-Abelian X-ray transforms. New analytical techniques are deployed to show that the relevant Fisher information operators are invertible between suitable function spaces.

REFERENCES

[1] ABRAHAM, K. and NICKL, R. (2019). On statistical Calderón problems. *Math. Stat. Learn.* **2** 165–216. MR4130599

[2] AGAPIOU, S., LARSSON, S. and STUART, A. M. (2013). Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems. *Stochastic Process. Appl.* **123** 3828–3860. MR3084161 https://doi.org/10.1016/j.spa.2013.05.001

[3] BAL, G. and REN, K. (2011). Multi-source quantitative photoacoustic tomography in a diffusive regime. *Inverse Probl.* **27** 075003. MR2806367 https://doi.org/10.1088/0266-5611/27/7/075003

[4] BAL, G. and UHLMANN, G. (2010). Inverse diffusion theory of photoacoustics. *Inverse Probl.* **26** 085010. MR2658827 https://doi.org/10.1088/0266-5611/26/8/085010

[5] BESKOS, A., GIROLAMI, M., LAN, S., FARRELL, P. E. and STUART, A. M. (2017). Geometric MCMC for infinite-dimensional inverse problems. *J. Comput. Phys.* **335** 327–351. MR3612501 https://doi.org/10.1016/j.jcp.2016.12.041

[6] CASTILLO, I. and NICKL, R. (2013). Nonparametric Bernstein-von Mises theorems in Gaussian white noise. *Ann. Statist.* **41** 1999–2028. MR3127856 https://doi.org/10.1214/13-AOS1133

[7] CASTILLO, I. and NICKL, R. (2014). On the Bernstein–von Mises phenomenon for nonparametric Bayes procedures. *Ann. Statist.* **42** 1941–1969. MR3262473 https://doi.org/10.1214/14-AOS1246

[8] CASTILLO, I. and ROCKOVA, V. (2019). Uncertainty quantification for Bayesian CART. arXiv preprint. Available at arXiv:1910.07635.

[9] CASTILLO, I. and ROUSSEAU, J. (2015). A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *Ann. Statist.* **43** 2353–2383. MR3405597 https://doi.org/10.1214/15-AOS1336

[10] CASTILLO, I. and VAN DER PAS, S. (2020). Multiscale Bayesian survival analysis. arXiv preprint. Available at arXiv:2005.02889.

[11] CHUNG, K. L. and ZHAO, Z. X. (1995). *From Brownian Motion to Schrödinger's Equation. Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **312**. Springer, Berlin. MR1329992 https://doi.org/10.1007/978-3-642-57856-4

MSC2020 subject classifications. 62F15, 65N21.
Key words and phrases. X-ray transforms, Schrödinger equation, credible sets, Bernstein–von Mises theorems.

[12] CONRAD, P. R., MARZOUK, Y. M., PILLAI, N. S. and SMITH, A. (2016). Accelerating asymptotically exact MCMC for computationally intensive models via local approximations. *J. Amer. Statist. Assoc.* **111** 1591–1607. MR3601720 https://doi.org/10.1080/01621459.2015.1096787

[13] COTTER, S. L., ROBERTS, G. O., STUART, A. M. and WHITE, D. (2013). MCMC methods for functions: Modifying old algorithms to make them faster. *Statist. Sci.* **28** 424–446. MR3135540 https://doi.org/10.1214/13-STS421

[14] DAIRBEKOV, N. S., PATERNAIN, G. P., STEFANOV, P. and UHLMANN, G. (2007). The boundary rigidity problem in the presence of a magnetic field. *Adv. Math.* **216** 535–609. MR2351370 https://doi.org/10.1016/j.aim.2007.05.014

[15] DASHTI, M. and STUART, A. (2016). The Bayesian approach to inverse problems. In *Handbook of Uncertainty Quantification* (R. Ghanem, D. Higdon and H. Owhadi, eds.) Springer, Berlin.

[16] DUDLEY, R. M. (2002). *Real Analysis and Probability*. *Cambridge Studies in Advanced Mathematics* **74**. Cambridge Univ. Press, Cambridge. MR1932358 https://doi.org/10.1017/CBO9780511755347

[17] ESKIN, G. (2004). On non-Abelian Radon transform. *Russ. J. Math. Phys.* **11** 391–408. MR2179889

[18] GHANEM, R., HIGDON, D. and OWHADI, H., eds. (2017). *Handbook of Uncertainty Quantification*. *I–III* Springer, Cham.

[19] GHOSAL, S. and VAN DER VAART, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. *Cambridge Series in Statistical and Probabilistic Mathematics* **44**. Cambridge Univ. Press, Cambridge. MR3587782 https://doi.org/10.1017/9781139029834

[20] GILBARG, D. and TRUDINGER, N. S. (1998). *Elliptic Partial Differential Equations of Second Order*. Springer, Berlin.

[21] 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

[22] GIORDANO, M. and KEKKONEN, H. (2020). Bernstein–von Mises theorems and uncertainty quantification for linear inverse problems. *SIAM/ASA J. Uncertain. Quantificat.* **8** 342–373. MR4069334 https://doi.org/10.1137/18M1226269

[23] GIORDANO, M. and NICKL, R. (2020). Consistency of Bayesian inference with Gaussian process priors in an elliptic inverse problem. *Inverse Probl.* **36** 085001. MR4151406 https://doi.org/10.1088/1361-6420/ab7d2a

[24] GRIGIS, A. and SJÖSTRAND, J. (1994). *Microlocal Analysis for Differential Operators*: *An Introduction*. *London Mathematical Society Lecture Note Series* **196**. Cambridge Univ. Press, Cambridge. MR1269107 https://doi.org/10.1017/CBO9780511721441

[25] GRUBB, G. (2015). Fractional Laplacians on domains, a development of Hörmander's theory of μ-transmission pseudodifferential operators. *Adv. Math.* **268** 478–528. MR3276603 https://doi.org/10.1016/j.aim.2014.09.018

[26] GUGUSHVILI, S., VAN DER VAART, A. and YAN, D. (2020). Bayesian linear inverse problems in regularity scales. *Ann. Inst. Henri Poincaré Probab. Stat.* **56** 2081–2107. MR4116718 https://doi.org/10.1214/19-AIHP1029

[27] HAIRER, M., STUART, A. M. and VOLLMER, S. J. (2014). Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions. *Ann. Appl. Probab.* **24** 2455–2490. MR3262508 https://doi.org/10.1214/13-AAP982

[28] HILGER, A., MANKE, I., KARDILJOV, N., OSERNBERG, M., MARKOTTER, H. and BANHART, J. (2018). Tensorial neutron tomography of threedimensional magnetic vector fields in bulk materials. *Nat. Commun.*.

[29] HOANG, V. H., SCHWAB, C. and STUART, A. M. (2013). Complexity analysis of accelerated MCMC methods for Bayesian inversion. *Inverse Probl.* **29** 085010. MR3084684 https://doi.org/10.1088/0266-5611/29/8/085010

[30] ILMAVIRTA, J. and MONARD, F. (2019). Integral geometry on manifolds with boundary and applications, the Radon transform: The first 100 years and beyond. *Radon Ser. Comput. Appl. Math.* **22**. Editors: R. Ramlau, O. Scherzer.

[31] KAIPIO, J. and SOMERSALO, E. (2004). *Statistical and Computational Inverse Problems*. Springer, Berlin.

[32] KAIPIO, J. P., KOLEHMAINEN, V., SOMERSALO, E. and VAUHKONEN, M. (2000). Statistical inversion and Monte Carlo sampling methods in electrical impedance tomography. *Inverse Probl.* **16** 1487–1522. MR1800606 https://doi.org/10.1088/0266-5611/16/5/321

[33] KATCHALOV, A., KURYLEV, Y. and LASSAS, M. (2001). *Inverse Boundary Spectral Problems*. *Chapman & Hall/CRC Monographs and Surveys in Pure and Applied Mathematics* **123**. CRC Press/CRC, Boca Raton, FL. MR1889089 https://doi.org/10.1201/9781420036220

[34] KEKKONEN, H., LASSAS, M. and SILTANEN, S. (2016). Posterior consistency and convergence rates for Bayesian inversion with hypoelliptic operators. *Inverse Probl*. **32** 085005. MR3535664 https://doi.org/10.1088/0266-5611/32/8/085005

[35] 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

[36] LAPLACE, P. S. (1812). *Théorie Analytique des Probabilités*. Courcier.

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

[38] MONARD, F. (2020). Functional relations, sharp mapping properties, and regularization of the X-ray transform on disks of constant curvature. *SIAM J. Math. Anal*. **52** 5675–5702. MR4175835 https://doi.org/10.1137/20M1311508

[39] MONARD, F., NICKL, R. and PATERNAIN, G. P. (2019). Efficient nonparametric Bayesian inference for X-ray transforms. *Ann. Statist*. **47** 1113–1147. MR3909962 https://doi.org/10.1214/18-AOS1708

[40] MONARD, F., NICKL, R. and PATERNAIN, G. P. (2021). Consistent inversion of noisy non-Abelian X-ray transforms. *Comm. Pure Appl. Math*. **74** 1045–1099. MR4230066 https://doi.org/10.1002/cpa.21942

[41] NICKL, R. (2020). Bernstein–von Mises theorems for statistical inverse problems I: Schrödinger equation. *J. Eur. Math. Soc*. (*JEMS*) **22** 2697–2750. MR4118619 https://doi.org/10.4171/JEMS/975

[42] NICKL, R. and RAY, K. (2020). Nonparametric statistical inference for drift vector fields of multi-dimensional diffusions. *Ann. Statist*. **48** 1383–1408. MR4124327 https://doi.org/10.1214/19-AOS1851

[43] NICKL, R. and SÖHL, J. (2019). Bernstein-von Mises theorems for statistical inverse problems II: Compound Poisson processes. *Electron. J. Stat*. **13** 3513–3571. MR4013745 https://doi.org/10.1214/19-ejs1609

[44] NICKL, R., VAN DE GEER, S. and WANG, S. (2020). Convergence rates for penalized least squares estimators in PDE constrained regression problems. *SIAM/ASA J. Uncertain. Quantificat*. **8** 374–413. MR4074017 https://doi.org/10.1137/18M1236137

[45] NICKL, R. and WANG, S. (2020). On polynomial time computation of high-dimensional posterior measures by Langevin-type algorithms. Preprint.

[46] NOVIKOV, R. (2019). Non-Abelian Radon transform and its applications. In *The Radon Transform*: *The First* 100 *Years and Beyond* (R. Ramlau and O. Scherzer, eds.) 115–128.

[47] NOVIKOV, R. G. (2002). On determination of a gauge field on $\mathbb{R}^d$ from its non-Abelian Radon transform along oriented straight lines. *J. Inst. Math. Jussieu* **1** 559–629. MR1954437 https://doi.org/10.1017/S1474748002000166

[48] PATERNAIN, G. P., SALO, M. and UHLMANN, G. Geometric Inverse Problems in 2D. In preparation.

[49] PATERNAIN, G. P., SALO, M. and UHLMANN, G. (2012). The attenuated ray transform for connections and Higgs fields. *Geom. Funct. Anal*. **22** 1460–1489. MR2989440 https://doi.org/10.1007/s00039-012-0183-6

[50] PESTOV, L. and UHLMANN, G. (2005). Two dimensional compact simple Riemannian manifolds are boundary distance rigid. *Ann. of Math*. (2) **161** 1093–1110. MR2153407 https://doi.org/10.4007/annals.2005.161.1093

[51] RAY, K. (2013). Bayesian inverse problems with non-conjugate priors. *Electron. J. Stat*. **7** 2516–2549. MR3117105 https://doi.org/10.1214/13-EJS851

[52] 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

[53] REISS, M. (2008). Asymptotic equivalence for nonparametric regression with multivariate and random design. *Ann. Statist*. **36** 1957–1982. MR2435461 https://doi.org/10.1214/07-AOS525

[54] ROININEN, L., HUTTUNEN, J. M. J. and LASANEN, S. (2014). Whittle–Matérn priors for Bayesian statistical inversion with applications in electrical impedance tomography. *Inverse Probl. Imaging* **8** 561–586. MR3209311 https://doi.org/10.3934/ipi.2014.8.561

[55] SALES, M. et al. (2018). Three dimensional polarimetric neutron tomography of magnetic fields. *Natur. Sci. Rep*.

[56] SALO, M. and UHLMANN, G. (2011). The attenuated ray transform on simple surfaces. *J. Differential Geom*. **88** 161–187. MR2819758

[57] SHARAFUTDINOV, V. A. (1994). *Integral Geometry of Tensor Fields*. *Inverse and Ill-Posed Problems Series*. VSP, Utrecht. MR1374572 https://doi.org/10.1515/9783110900095

[58] STUART, A. M. (2010). Inverse problems: A Bayesian perspective. *Acta Numer.* **19** 451–559. MR2652785 https://doi.org/10.1017/S0962492910000061

[59] TAYLOR, M. E. (1996). *Partial Differential Equations*1, *Basic Theory*. *Texts in Applied Mathematics* **23**. Springer, New York. MR1395147 https://doi.org/10.1007/978-1-4684-9320-7

[60] VAN DER VAART, A. (1991). On differentiable functionals. *Ann. Statist*. **19** 178–204. MR1091845 https://doi.org/10.1214/aos/1176347976

[61] 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
[62] 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
[63] VON MISES, R. (1931). *Wahrscheinlichkeitsrechnung*. Deuticke, Vienna.

The Annals of Statistics
2021, Vol. 49, No. 6, 3299–3323
https://doi.org/10.1214/21-AOS2084

MARGINAL SINGULARITY AND THE BENEFITS OF LABELS IN COVARIATE-SHIFT

BY SAMORY KPOTUFE[1] AND GUILLAUME MARTINET[2]

[1]*Department of Statistics, Columbia University, samory@columbia.edu*
[2]*Department of Operations Research and Financial Engineering, Princeton University, ggm2@princeton.edu*

Transfer Learning addresses common situations in Machine Leaning where little or no labeled data is available for a target prediction problem—corresponding to a distribution Q, but much labeled data is available from some related but different data distribution P. This work is concerned with the fundamental limits of transfer, that is, the limits in target performance in terms of (1) sample sizes from P and Q, and (2) differences in data distributions P, Q. In particular, we aim to address practical questions such as how much *target* data from Q is sufficient given a certain amount of related data from P, and how to optimally *sample* such target data for labeling.

We present new minimax results for transfer in *nonparametric* classification (i.e., for situations where little is known about the target classifier), under the common assumption that the marginal distributions of covariates differ between P and Q (often termed *covariate-shift*). Our results are first to concisely capture the relative benefits of source and target labeled data in these settings through information-theoretic limits. Namely, we show that the benefits of target labels are tightly controlled by a *transfer-exponent* γ that encodes how *singular* Q is locally with respect to P, and interestingly paints a more favorable picture of transfer than what might be believed from insights from previous work. In fact, while previous work rely largely on refinements of traditional metrics and divergences between distributions, and often only yield a coarse view of when transfer is possible or not, our analysis—in terms of γ—reveals a *continuum of new regimes* ranging from easy to hard transfer.

We then address the practical question of how to efficiently sample target data to label, by showing that a recently proposed semi-supervised procedure—based on k-NN classification, can be refined to adapt to unknown γ and, therefore, requests target labels only when beneficial, while achieving nearly minimax-optimal transfer rates without knowledge of distributional parameters. Of independent interest, we obtain new minimax-optimality results for vanilla k-NN classification in regimes with nonuniform marginals.

REFERENCES

[1] 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

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

[3] BEN-DAVID, S., LU, T., LUU, T. and PÁL, D. (2010). Impossibility theorems for domain adaptation. In *Artificial Intelligence and Statistics* 129–136.

[4] BEN-DAVID, S. and URNER, R. (2012). On the hardness of domain adaptation and the utility of unlabeled target samples. In *Algorithmic Learning Theory*. *Lecture Notes in Computer Science* **7568** 139–153. Springer, Heidelberg. MR3042887 https://doi.org/10.1007/978-3-642-34106-9_14

[5] BERLIND, C. and URNER, R. (2015). Active nearest neighbors in changing environments. In *International Conference on Machine Learning* 1870–1879.

[6] BLITZER, J., KAKADE, S. and FOSTER, D. (2011). Domain adaptation with coupled subspaces. In *Conference on Artificial Intelligence and Statistics* 173–181.

MSC2020 subject classifications. 62G99, 62D99, 62B10, 68T05, 68Q32.
Key words and phrases. Transfer learning, domain adaptation, covariate shift, nonparametric classification.

[7] CAI, T. T. and WEI, H. (2021). Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *Ann. Statist.* **49** 100–128. MR4206671 https://doi.org/10.1214/20-AOS1949
[8] CANNINGS, T. I., BERRETT, T. B. and SAMWORTH, R. J. (2020). Local nearest neighbour classification with applications to semi-supervised learning. *Ann. Statist.* **48** 1789–1814. MR4124344 https://doi.org/10.1214/19-AOS1868
[9] CHATTOPADHYAY, R., FAN, W., DAVIDSON, I., PANCHANATHAN, S. and YE, J. (2013). Joint transfer and batch-mode active learning. In *International Conference on Machine Learning* 253–261.
[10] CHAUDHURI, K. and DASGUPTA, S. (2014). Rates of convergence for nearest neighbor classification. In *Neural Information Processing Systems* 3437–3445.
[11] CHEN, M., WEINBERGER, K. Q. and BLITZER, J. (2011). Co-training for domain adaptation. In *Neural Information Processing Systems* 2456–2464.
[12] CORTES, C., MANSOUR, Y. and MOHRI, M. (2010). Learning bounds for importance weighting. In *Neural Information Processing Systems* 442–450.
[13] CORTES, C., MOHRI, M. and MUÑOZ MEDINA, A. (2019). Adaptation based on generalized discrepancy. *J. Mach. Learn. Res.* **20** Paper No. 1, 30. MR3911408
[14] GADAT, S., KLEIN, T. and MARTEAU, C. (2014). Classification with the NN rule in general finite dimensional spaces: Necessary and sufficient conditions. Available at arXiv:1411.0894.
[15] GERMAIN, P., HABRARD, A., LAVIOLETTE, F. and MORVANT, E. (2013). A PAC-Bayesian approach for domain adaptation with specialization to linear classifiers. In *International Conference on Machine Learning* 738–746.
[16] GOLDBERG, P. W. and JERRUM, M. R. (1995). Bounding the Vapnik–Chervonenkis dimension of concept classes parameterized by real numbers. *Mach. Learn.* **18** 131–148.
[17] GOLDENSHLUGER, A. and NEMIROVSKI, A. (1997). On spatially adaptive estimation of nonparametric regression. *Math. Methods Statist.* **6** 135–170. MR1466625
[18] 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
[19] HOFFMAN, J., MOHRI, M. and ZHANG, N. (2017). Multiple-source adaptation for regression problems. Available at arXiv:1711.05037.
[20] HUANG, J., GRETTON, A., BORGWARDT, K. M., SCHÖLKOPF, B. and SMOLA, A. J. (2007). Correcting sample selection bias by unlabeled data. In *Neural Information Processing Systems* 601–608.
[21] HUANG, L., JIANG, S. H.-C., LI, J. and WU, X. (2018). ϵ-coresets for clustering (with outliers) in doubling metrics. In 59*th Annual IEEE Symposium on Foundations of Computer Science—FOCS* 2018 814–825. IEEE Comput. Soc., Los Alamitos, CA. MR3899644 https://doi.org/10.1109/FOCS.2018.00082
[22] KPOTUFE, S. (2017). Lipschitz density-ratios, structured data, and data-driven tuning. In *Artificial Intelligence and Statistics* 1320–1328.
[23] KPOTUFE, S. and MARTINET, G. (2018). Marginal singularity, and the benefits of labels in covariate-shift. In *Conference on Learning Theory* 1882–1886.
[24] KPOTUFE, S. and MARTINET, G. (2021). Supplement to "Marginal singularity and the benefits of labels in covariate-shift." https://doi.org/10.1214/21-AOS2084SUPP
[25] KULKARNI, S. R. and POSNER, S. E. (1995). Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Trans. Inf. Theory* **41** 1028–1039. MR1366756 https://doi.org/10.1109/18.391248
[26] KUZBORSKIJ, I. and ORABONA, F. (2013). Stability and hypothesis transfer learning. In *International Conference on Machine Learning* 942–950.
[27] LEPSKI, O. V., MAMMEN, E. and SPOKOINY, V. G. (1997). Optimal spatial adaptation to inhomogeneous smoothness: An approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.* **25** 929–947. MR1447734 https://doi.org/10.1214/aos/1069362731
[28] LIPTON, Z. C., WANG, Y.-X. and SMOLA, A. (2018). Detecting and correcting for label shift with black box predictors. Available at arXiv:1802.03916.
[29] MANSOUR, Y., MOHRI, M. and ROSTAMIZADEH, A. (2009). Domain adaptation: Learning bounds and algorithms. Preprint. Available at arXiv:0902.3430.
[30] MANSOUR, Y., MOHRI, M. and ROSTAMIZADEH, A. (2009). Multiple source adaptation and the Rényi divergence. In *Uncertainty in Artificial Intelligence* 367–374.
[31] MOHRI, M. and MUÑOZ MEDINA, A. (2012). New analysis and algorithm for learning with drifting distributions. In *Algorithmic Learning Theory*. *Lecture Notes in Computer Science* **7568** 124–138. Springer, Heidelberg. MR3042886 https://doi.org/10.1007/978-3-642-34106-9_13
[32] QUIONERO-CANDELA, J., SUGIYAMA, M., SCHWAIGHOFER, A. and LAWRENCE, N. D. (2009). *Dataset Shift in Machine Learning*. MIT Press, Cambridge, MA.

[33] SAHA, A., RAI, P., DAUMÉ, H., VENKATASUBRAMANIAN, S. and DUVALL, S. L. (2011). Active supervised domain adaptation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 97–112. Springer, Berlin.

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

[35] SCOTT, C. (2019). A generalized Neyman–Pearson criterion for optimal domain adaption. In *Algorithmic Learning Theory* 2019. *Proc. Mach. Learn. Res.* (*PMLR*) **98** [738–761], 24. Proceedings of Machine Learning Research PMLR. MR3932867

[36] SHALEV-SHWARTZ, S. and BEN-DAVID, S. (2014). *Understanding Machine Learning*: *From Theory to Algorithms*. Cambridge Univ. Press, Cambridge.

[37] SUGIYAMA, M., NAKAJIMA, S., KASHIMA, H., BUENAU, P. V. and KAWANABE, M. (2008). Direct importance estimation with model selection and its application to covariate shift adaptation. In *Neural Information Processing Systems* 1433–1440.

[38] SUGIYAMA, M., SUZUKI, T. and KANAMORI, T. (2012). *Density Ratio Estimation in Machine Learning*. Cambridge Univ. Press, Cambridge. With a foreword by Thomas G. Dietterich. MR2895762 https://doi.org/10.1017/CBO9781139035613

[39] 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

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

[41] YANG, L., HANNEKE, S. and CARBONELL, J. (2013). A theory of transfer learning with applications to active learning. *Mach. Learn.* **90** 161–189. MR3015741 https://doi.org/10.1007/s10994-012-5310-y

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

The Annals of Statistics
2021, Vol. 49, No. 6, 3324–3357
https://doi.org/10.1214/21-AOS2085

OPTIMAL LINEAR DISCRIMINATORS FOR THE DISCRETE CHOICE MODEL IN GROWING DIMENSIONS

BY DEBARGHYA MUKHERJEE[*], MOULINATH BANERJEE[†] AND YA'ACOV RITOV[‡]

Department of Statistics, University of Michigan, [*]*mdeb@umich.edu;* [†]*moulib@umich.edu;* [‡]*yritov@umich.edu*

Manski's celebrated maximum score estimator for the discrete choice model, which is an optimal linear discriminator, has been the focus of much investigation in both the econometrics and statistics literatures, but its behavior under growing dimension scenarios largely remains unknown. This paper addresses that gap. Two different cases are considered: p grows with n but at a slow rate, that is, $p/n \to 0$; and $p \gg n$ (fast growth). In the binary response model, we recast Manski's score estimation as empirical risk minimization for a classification problem, and derive the ℓ_2 rate of convergence of the score estimator under a new *transition condition* in terms of a margin parameter that calibrates the level of difficulty of the estimation problem. We also establish upper and lower bounds for the minimax ℓ_2 error in the binary choice model that differ by a logarithmic factor, and construct a minimax-optimal estimator in the slow growth regime. Some extensions to the multinomial choice model are also considered.

REFERENCES

[1] ABRAMOVICH, F. and GRINSHTEIN, V. (2019). High-dimensional classification by sparse logistic regression. *IEEE Trans. Inf. Theory* **65** 3068–3079. MR3951383 https://doi.org/10.1109/TIT.2018.2884963
[2] ABREVAYA, J. and HUANG, J. (2005). On the bootstrap of the maximum score estimator. *Econometrica* **73** 1175–1204. MR2149245 https://doi.org/10.1111/j.1468-0262.2005.00613.x
[3] ARLOT, S. and BARTLETT, P. L. (2011). Margin-adaptive model selection in statistical learning. *Bernoulli* **17** 687–713. MR2787611 https://doi.org/10.3150/10-BEJ288
[4] BAJARI, P., FOX, J. T. and RYAN, S. P. (2008). Evaluating wireless carrier consolidation using semiparametric demand estimation. *Quant. Mark. Econ.* **6** 299.
[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
[6] BRIESCH, R. A., CHINTAGUNTA, P. K. and MATZKIN, R. L. (2002). Semiparametric estimation of brand choice behavior. *J. Amer. Statist. Assoc.* **97** 973–982. MR1973860 https://doi.org/10.1198/016214502388618762
[7] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer Series in Statistics*. Springer, Heidelberg. MR2807761 https://doi.org/10.1007/978-3-642-20192-9
[8] CHERNOFF, H. (1964). Estimation of the mode. *Ann. Inst. Statist. Math.* **16** 31–41. MR0172382 https://doi.org/10.1007/BF02868560
[9] DEBARGHYA, M., BANERJEE, M. and RITOV, Y. (2019). Non-standard asymptotics in high dimensions: Manski's maximum score estimator revisited. Preprint. Available at arXiv:1903.10063.
[10] FENG, H., NING, Y. and ZHAO, J. (2019). Nonregular and minimax estimation of individualized thresholds in high dimension with binary responses. Preprint. Available at arXiv:1905.10888.
[11] FOX, J. T. (2007). Semiparametric estimation of multinomial discrete-choice models using a subset of choices. *Rand J. Econ.* **38** 1002–1019.
[12] FOX, J. T. and BAJARI, P. (2013). Measuring the efficiency of an fcc spectrum auction. *Am. Econ. J. Microecon.* **5** 100–146.
[13] GREENSHTEIN, E. and RITOV, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10** 971–988. MR2108039 https://doi.org/10.3150/bj/1106314846

MSC2020 subject classifications. 62P20, 62G05.
Key words and phrases. High dimensional statistics, maximum score estimator, nonstandard asymptotics.

[14] HOROWITZ, J. L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica* **60** 505–531. MR1162997 https://doi.org/10.2307/2951582

[15] HUANG, J., MA, S. and ZHANG, C.-H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statist. Sinica* **18** 1603–1618. MR2469326

[16] KIM, J. and POLLARD, D. (1990). Cube root asymptotics. *Ann. Statist.* **18** 191–219. MR1041391 https://doi.org/10.1214/aos/1176347498

[17] KLEIN, R. W. and SPADY, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica* **61** 387–421. MR1209737 https://doi.org/10.2307/2951556

[18] KOLTCHINSKII, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.* **34** 2593–2656. MR2329442 https://doi.org/10.1214/009053606000001019

[19] LEE, L. F. (1995). Semiparametric maximum likelihood estimation of polychotomous and sequential choice models. *J. Econometrics* **65** 381–428. MR1323323 https://doi.org/10.1016/0304-4076(93)01591-9

[20] MAMMEN, E. and TSYBAKOV, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.* **27** 1808–1829. MR1765618 https://doi.org/10.1214/aos/1017939240

[21] MANSKI, C. F. (1975). Maximum score estimation of the stochastic utility model of choice. *J. Econometrics* **3** 205–228. MR0436905 https://doi.org/10.1016/0304-4076(75)90032-9

[22] MANSKI, C. F. (1985). Semiparametric analysis of discrete response. Asymptotic properties of the maximum score estimator. *J. Econometrics* **27** 313–333. MR0788628 https://doi.org/10.1016/0304-4076(85)90009-0

[23] MANSKI, C. F. and THOMPSON, T. S. (1986). Operational characteristics of maximum score estimation. *J. Econometrics* **32** 85–108. MR0853046 https://doi.org/10.1016/0304-4076(86)90013-8

[24] MASSART, P. (2007). *Concentration Inequalities and Model Selection. Lecture Notes in Math.* **1896**. Springer, Berlin. MR2319879

[25] MASSART, P. and NÉDÉLEC, É. (2006). Risk bounds for statistical learning. *Ann. Statist.* **34** 2326–2366. MR2291502 https://doi.org/10.1214/009053606000000786

[26] MIOLANE, L. and MONTANARI, A. (2018). The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. Preprint. Available at arXiv:1811.01212.

[27] MUKHERJEE, D., BANERJEE, M. and RITOV, Y. (2021). Supplement to "Optimal linear discriminators for the discrete choice model in growing dimensions." https://doi.org/10.1214/21-AOS2085SUPP

[28] PATRA, R. K., SEIJO, E. and SEN, B. (2018). A consistent bootstrap procedure for the maximum score estimator. *J. Econometrics* **205** 488–507. MR3813528 https://doi.org/10.1016/j.jeconom.2018.04.001

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

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

[31] SEN, B., BANERJEE, M. and WOODROOFE, M. (2010). Inconsistency of bootstrap: The Grenander estimator. *Ann. Statist.* **38** 1953–1977. MR2676880 https://doi.org/10.1214/09-AOS777

[32] SEO, M. H. and OTSU, T. (2015). Asymptotics for maximum score method under general conditions Technical report, LSE Working Paper.

[33] SEO, M. H. and OTSU, T. (2018). Local M-estimation with discontinuous criterion for dependent and limited observations. *Ann. Statist.* **46** 344–369. MR3766955 https://doi.org/10.1214/17-AOS1552

[34] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[35] TSYBAKOV, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** 135–166. MR2051002 https://doi.org/10.1214/aos/1079120131

[36] 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

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

[38] VAPNIK, V. and CHERVONENKIS, A. (1974). Theory of Pattern Recognition.

[39] WEI, F. and HUANG, J. (2010). Consistent group selection in high-dimensional linear regression. *Bernoulli* **16** 1369–1384. MR2759183 https://doi.org/10.3150/10-BEJ252

[40] YAN, J. and YOO, H. I. (2019). Semiparametric estimation of the random utility model with rank-ordered choice data. *J. Econometrics* **211** 414–438. MR3980130 https://doi.org/10.1016/j.jeconom.2019.03.003

[41] 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

[42] 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

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

The Annals of Statistics
2021, Vol. 49, No. 6, 3358–3382
https://doi.org/10.1214/21-AOS2087

EXTREME CONDITIONAL EXPECTILE ESTIMATION IN HEAVY-TAILED HETEROSCEDASTIC REGRESSION MODELS

BY STÉPHANE GIRARD[1], GILLES STUPFLER[2] AND ANTOINE USSEGLIO-CARLEVE[3]

[1]*Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK,* Stephane.Girard@inria.fr

[2]*Univ Rennes, Ensai, CNRS, CREST—UMR 9194,* gilles.stupfler@ensai.fr

[3]*Toulouse School of Economics, University of Toulouse Capitole, France,* antoine.usseglio-carleve@tse-fr.eu

Expectiles define a least squares analogue of quantiles. They have been the focus of a substantial quantity of research in the context of actuarial and financial risk assessment over the last decade. The behaviour and estimation of unconditional extreme expectiles using independent and identically distributed heavy-tailed observations have been investigated in a recent series of papers. We build here a general theory for the estimation of extreme conditional expectiles in heteroscedastic regression models with heavy-tailed noise; our approach is supported by general results of independent interest on residual-based extreme value estimators in heavy-tailed regression models, and is intended to cope with covariates having a large but fixed dimension. We demonstrate how our results can be applied to a wide class of important examples, among which are linear models, single-index models as well as ARMA and GARCH time series models. Our estimators are showcased on a numerical simulation study and on real sets of actuarial and financial data.

REFERENCES

[1] ABDOUS, B. and RÉMILLARD, B. (1995). Relating quantiles and expectiles under weighted-symmetry. *Ann. Inst. Statist. Math.* **47** 371–384. MR1345429 https://doi.org/10.1007/BF00773468

[2] AHMAD, A. A., DEME, E. H., DIOP, A., GIRARD, S. and USSEGLIO-CARLEVE, A. (2020). Estimation of extreme quantiles from heavy-tailed distributions in a location-dispersion regression model. *Electron. J. Stat.* **14** 4421–4456. MR4194267 https://doi.org/10.1214/20-EJS1779

[3] BELLINI, F. and DI BERNARDINO, E. (2017). Risk management with expectiles. *Eur. J. Finance* **23** 487–506.

[4] BROCKWELL, P. J. and DAVIS, R. A. (1991). *Time Series: Theory and Methods*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR1093459 https://doi.org/10.1007/978-1-4419-0320-4

[5] BÜCHER, A., SEGERS, J. and VOLGUSHEV, S. (2014). When uniform weak convergence fails: Empirical processes for dependence functions and residuals via epi- and hypographs. *Ann. Statist.* **42** 1598–1634. MR3262462 https://doi.org/10.1214/14-AOS1237

[6] CHERNOZHUKOV, V. (2005). Extremal quantile regression. *Ann. Statist.* **33** 806–839. MR2163160 https://doi.org/10.1214/009053604000001165

[7] DAOUIA, A., GARDES, L., GIRARD, S. and LEKINA, A. (2011). Kernel estimators of extreme level curves. *TEST* **20** 311–333. MR2834049 https://doi.org/10.1007/s11749-010-0196-0

[8] DAOUIA, A., GIJBELS, I. and STUPFLER, G. (2019). Extremiles: A new perspective on asymmetric least squares. *J. Amer. Statist. Assoc.* **114** 1366–1381. MR4011785 https://doi.org/10.1080/01621459.2018.1498348

[9] DAOUIA, A., GIRARD, S. and STUPFLER, G. (2018). Estimation of tail risk based on extreme expectiles. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 263–292. MR3763692 https://doi.org/10.1111/rssb.12254

[10] DAOUIA, A., GIRARD, S. and STUPFLER, G. (2019). Extreme M-quantiles as risk measures: From L^1 to L^p optimization. *Bernoulli* **25** 264–309. MR3892320 https://doi.org/10.3150/17-bej987

[11] DAOUIA, A., GIRARD, S. and STUPFLER, G. (2020). Tail expectile process and risk assessment. *Bernoulli* **26** 531–556. MR4036043 https://doi.org/10.3150/19-BEJ1137

MSC2020 subject classifications. Primary 62G32; secondary 62G08, 62G20, 62G30.

Key words and phrases. Expectiles, extreme value analysis, heavy-tailed distribution, heteroscedasticity, regression models, residual-based estimators, single-index model, tail empirical process of residuals.

[12] DAVISON, A. C. and SMITH, R. L. (1990). Models for exceedances over high thresholds. *J. Roy. Statist. Soc. Ser. B* **52** 393–442. With discussion and a reply by the authors. MR1086795

[13] DE HAAN, L. and FERREIRA, A. (2006). *Extreme Value Theory*: *An Introduction*. *Springer Series in Operations Research and Financial Engineering*. Springer, New York. MR2234156 https://doi.org/10.1007/0-387-34471-3

[14] DEKKERS, A. L. M., EINMAHL, J. H. J. and DE HAAN, L. (1989). A moment estimator for the index of an extreme-value distribution. *Ann. Statist.* **17** 1833–1855. MR1026315 https://doi.org/10.1214/aos/1176347397

[15] DREES, H. (2003). Extreme quantile estimation for dependent data, with applications to finance. *Bernoulli* **9** 617–657. MR1996273 https://doi.org/10.3150/bj/1066223272

[16] FRANCQ, C. and ZAKOÏAN, J.-M. (2010). *GARCH Models*: *Structure*, *Statistical Inference and Financial Applications*. Wiley, Chichester. MR3186556 https://doi.org/10.1002/9780470670057

[17] GARDES, L. (2018). Tail dimension reduction for extreme quantile estimation. *Extremes* **21** 57–95. MR3764611 https://doi.org/10.1007/s10687-017-0300-x

[18] GARDES, L. and STUPFLER, G. (2014). Estimation of the conditional tail index using a smoothed local Hill estimator. *Extremes* **17** 45–75. MR3179970 https://doi.org/10.1007/s10687-013-0174-5

[19] GIRARD, S., STUPFLER, G. and USSEGLIO-CARLEVE, A. (2021). Supplement to "Extreme conditional expectile estimation in heavy-tailed heteroscedastic regression models." https://doi.org/10.1214/21-AOS2087SUPP

[20] GIRARD, S., STUPFLER, G. and USSEGLIO-CARLEVE, A. (2021). On automatic bias reduction for extreme expectile estimation. https://hal.archives-ouvertes.fr/hal-03086048.

[21] GOMES, M. I., BRILHANTE, M. F. and PESTANA, D. (2016). New reduced-bias estimators of a positive extreme value index. *Comm. Statist. Simulation Comput.* **45** 833–862. MR3473918 https://doi.org/10.1080/03610918.2013.875567

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

[23] HE, F., CHENG, Y. and TONG, T. (2016). Estimation of high conditional quantiles using the Hill estimator of the tail index. *J. Statist. Plann. Inference* **176** 64–77. MR3500055 https://doi.org/10.1016/j.jspi.2016.03.003

[24] HILL, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist.* **3** 1163–1174. MR0378204

[25] HILL, J. B. (2015). Tail index estimation for a filtered dependent time series. *Statist. Sinica* **25** 609–629. MR3379091

[26] HOGA, Y. (2019). Confidence intervals for conditional tail risk measures in ARMA–GARCH models. *J. Bus. Econom. Statist.* **37** 613–624. MR4016158 https://doi.org/10.1080/07350015.2017.1401543

[27] HOLZMANN, H. and KLAR, B. (2016). Expectile asymptotics. *Electron. J. Stat.* **10** 2355–2371. MR3544290 https://doi.org/10.1214/16-EJS1173

[28] HOROWITZ, J. L. (2009). *Semiparametric and Nonparametric Methods in Econometrics*. *Springer Series in Statistics*. Springer, New York. MR2535631 https://doi.org/10.1007/978-0-387-92870-8

[29] HORVÁTH, L. and LIESE, F. (2004). L_p-estimators in ARCH models. *J. Statist. Plann. Inference* **119** 277–309. MR2019642 https://doi.org/10.1016/S0378-3758(02)00488-3

[30] KRÄTSCHMER, V. and ZÄHLE, H. (2017). Statistical inference for expectile-based risk measures. *Scand. J. Stat.* **44** 425–454. MR3658521 https://doi.org/10.1111/sjos.12259

[31] KUAN, C.-M., YEH, J.-H. and HSU, Y.-C. (2009). Assessing value at risk with CARE, the conditional autoregressive expectile models. *J. Econometrics* **150** 261–270. MR2535521 https://doi.org/10.1016/j.jeconom.2008.12.002

[32] KUCHIBHOTLA, A. K. and PATRA, R. K. (2020). Efficient estimation in single index models through smoothing splines. *Bernoulli* **26** 1587–1618. MR4058379 https://doi.org/10.3150/19-BEJ1183

[33] KYUNG-JOON, C. and SCHUCANY, W. R. (1998). Nonparametric kernel regression estimation near endpoints. *J. Statist. Plann. Inference* **66** 289–304. MR1614481 https://doi.org/10.1016/S0378-3758(97)00082-7

[34] LEWBEL, A. and LINTON, O. (2002). Nonparametric censored and truncated regression. *Econometrica* **70** 765–779. MR1913830 https://doi.org/10.1111/1468-0262.00304

[35] MARTINS-FILHO, C., YAO, F. and TORERO, M. (2018). Nonparametric estimation of conditional value-at-risk and expected shortfall based on extreme value theory. *Econometric Theory* **34** 23–67. MR3739235 https://doi.org/10.1017/S0266466616000517

[36] MCKENZIE, M. D. (1997). ARCH modelling of Australian bilateral exchange rate data. *Appl. Financ. Econ.* **7** 147–164.

[37] MCNEIL, A. J. and FREY, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach. *J. Empir. Finance* **7** 271–300.

[38] NEWEY, W. K. and POWELL, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica* **55** 819–847. MR0906565 https://doi.org/10.2307/1911031

[39] NEWEY, W. K. and POWELL, J. L. (1990). Efficient estimation of linear and type I censored regression models under conditional quantile restrictions. *Econometric Theory* **6** 295–317. MR1085576 https://doi.org/10.1017/S0266466600005284

[40] POWELL, J. L. (1986). Symmetrically trimmed least squares estimation for Tobit models. *Econometrica* **54** 1435–1460. MR0868151 https://doi.org/10.2307/1914308

[41] ROHRBECK, C., EASTOE, E. F., FRIGESSI, A. and TAWN, J. A. (2018). Extreme value modelling of water-related insurance claims. *Ann. Appl. Stat.* **12** 246–282. MR3773393 https://doi.org/10.1214/17-AOAS1081

[42] SEGERS, J. (2001). Residual estimators. *J. Statist. Plann. Inference* **98** 15–27. MR1860222 https://doi.org/10.1016/S0378-3758(00)00321-9

[43] STUTE, W. and WANG, J.-L. (2008). The central limit theorem under random truncation. *Bernoulli* **14** 604–622. MR2537804 https://doi.org/10.3150/07-BEJ116

[44] TAYLOR, J. W. (2008). Estimating value at risk and expected shortfall using expectiles. *J. Financ. Econom.* **6** 231–252.

[45] TOBIN, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica* **26** 24–36. MR0090462 https://doi.org/10.2307/1907382

[46] TSAY, R. S. (1984). Regression models with time series errors. *J. Amer. Statist. Assoc.* **79** 118–124. MR0742860

[47] VAN KEILEGOM, I. and WANG, L. (2010). Semiparametric modeling and estimation of heteroscedasticity in regression analysis of cross-sectional data. *Electron. J. Stat.* **4** 133–160. MR2645480 https://doi.org/10.1214/09-EJS547

[48] VELTHOEN, J., CAI, J.-J., JONGBLOED, G. and SCHMEITS, M. (2019). Improving precipitation forecasts using extreme quantile regression. *Extremes* **22** 599–622. MR4031851 https://doi.org/10.1007/s10687-019-00355-1

[49] WANG, H. J. and LI, D. (2013). Estimation of extreme conditional quantiles through power transformation. *J. Amer. Statist. Assoc.* **108** 1062–1074. MR3174684 https://doi.org/10.1080/01621459.2013.820134

[50] WANG, H. J., LI, D. and HE, X. (2012). Estimation of high conditional quantiles for heavy-tailed distributions. *J. Amer. Statist. Assoc.* **107** 1453–1464. MR3036407 https://doi.org/10.1080/01621459.2012.716382

[51] WEISSMAN, I. (1978). Estimation of parameters and large quantiles based on the k largest observations. *J. Amer. Statist. Assoc.* **73** 812–815. MR0521329

[52] WOODROOFE, M. (1985). Estimating a distribution function with truncated data. *Ann. Statist.* **13** 163–177. MR0773160 https://doi.org/10.1214/aos/1176346584

[53] ZHU, K. and LING, S. (2011). Global self-weighted and local quasi-maximum exponential likelihood estimators for ARMA–GARCH/IGARCH models. *Ann. Statist.* **39** 2131–2163. MR2893864 https://doi.org/10.1214/11-AOS895

[54] ZHU, L., DONG, Y. and LI, R. (2013). Semiparametric estimation of conditional heteroscedasticity via single-index modeling. *Statist. Sinica* **23** 1235–1255. MR3114712

[55] ZIEGEL, J. F. (2016). Coherence and elicitability. *Math. Finance* **26** 901–918. MR3551510 https://doi.org/10.1111/mafi.12080

The Annals of Statistics
2021, Vol. 49, No. 6, 3383–3407
https://doi.org/10.1214/21-AOS2088

ASYMPTOTIC PROPERTIES OF PENALIZED SPLINE ESTIMATORS IN CONCAVE EXTENDED LINEAR MODELS: RATES OF CONVERGENCE

BY JIANHUA Z. HUANG[1] AND YA SU[2]

[1]*Department of Statistics, Texas A&M University, jianhua@stat.tamu.edu*
[2]*Department of Statistical Sciences and Operations Research, Virginia Commonwealth University, suyaf@vcu.edu*

This paper develops a general theory on rates of convergence of penalized spline estimators for function estimation when the likelihood functional is concave in candidate functions, where the likelihood is interpreted in a broad sense that includes conditional likelihood, quasi-likelihood and pseudo-likelihood. The theory allows all feasible combinations of the spline degree, the penalty order and the smoothness of the unknown functions. According to this theory, the asymptotic behaviors of the penalized spline estimators depends on interplay between the spline knot number and the penalty parameter. The general theory is applied to obtain results in a variety of contexts, including regression, generalized regression such as logistic regression and Poisson regression, density estimation, conditional hazard function estimation for censored data, quantile regression, diffusion function estimation for a diffusion type process and estimation of spectral density function of a stationary time series. For multidimensional function estimation, the theory (presented in the Supplementary Material) covers both penalized tensor product splines and penalized bivariate splines on triangulations.

REFERENCES

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

CLAESKENS, G., KRIVOBOKOVA, T. and OPSOMER, J. D. (2009). Asymptotic properties of penalized spline estimators. *Biometrika* **96** 529–544. MR2538755 https://doi.org/10.1093/biomet/asp035

DEVORE, R. A. and LORENTZ, G. G. (1993). *Constructive Approximation*. *Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **303**. Springer, Berlin. MR1261635 https://doi.org/10.1007/978-3-662-02888-9

EILERS, P. H. C. and MARX, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statist*. *Sci*. **11** 89–121. MR1435485 https://doi.org/10.1214/ss/1038425655

GU, C. (2013). *Smoothing Spline ANOVA Models*, 2nd ed. *Springer Series in Statistics* **297**. Springer, New York. MR3025869 https://doi.org/10.1007/978-1-4614-5369-7

HALL, P. and OPSOMER, J. D. (2005). Theory for penalised spline regression. *Biometrika* **92** 105–118. MR2158613 https://doi.org/10.1093/biomet/92.1.105

HANSEN, M. H. (1994). Extended Linear Models, Multivariate Splines, and ANOVA. PhD Thesis, Univ. California, Berkeley.

HOLLAND, A. D. (2017). Penalized spline estimation in the partially linear model. *J. Multivariate Anal*. **153** 211–235. MR3578847 https://doi.org/10.1016/j.jmva.2016.10.001

HUANG, J. Z. (1998a). Projection estimation in multiple regression with application to functional ANOVA models. *Ann*. *Statist*. **26** 242–272. MR1611780 https://doi.org/10.1214/aos/1030563984

HUANG, J. Z. (1998b). Functional ANOVA models for generalized regression. *J. Multivariate Anal*. **67** 49–71. MR1659096 https://doi.org/10.1006/jmva.1998.1753

HUANG, J. Z. (2001). Concave extended linear modeling: A theoretical synthesis. *Statist*. *Sinica* **11** 173–197. MR1820005

HUANG, J. Z. (2003). Asymptotics for polynomial spline regression under weak conditions. *Statist*. *Probab. Lett*. **65** 207–216. MR2018032 https://doi.org/10.1016/j.spl.2003.09.003

MSC2020 subject classifications. Primary 62G20; secondary 62G05, 62G07, 62G08.
Key words and phrases. Basis expansion, multivariate splines, nonparametric regression, polynomial splines, smoothing splines.

HUANG, J. Z. and SU, Y. (2021). Supplement to "Asymptotic properties of penalized spline estimators in concave extended linear models: Rates of convergence." https://doi.org/10.1214/21-AOS2088SUPP

KAUERMANN, G., KRIVOBOKOVA, T. and FAHRMEIR, L. (2009). Some asymptotic results on generalized penalized spline smoothing. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 487–503. MR2649606 https://doi.org/10.1111/j.1467-9868.2008.00691.x

KNIGHT, K. (1998). Limiting distributions for L_1 regression estimators under general conditions. *Ann. Statist.* **26** 755–770. MR1626024 https://doi.org/10.1214/aos/1028144858

LAI, M.-J. and WANG, L. (2013). Bivariate penalized splines for regression. *Statist. Sinica* **23** 1399–1417. MR3114719

LI, Y. and RUPPERT, D. (2008). On the asymptotics of penalized splines. *Biometrika* **95** 415–436. MR2521591 https://doi.org/10.1093/biomet/asn010

NIELSEN, J. P. and LINTON, O. B. (1995). Kernel estimation in a nonparametric marker dependent hazard model. *Ann. Statist.* **23** 1735–1748. MR1370305 https://doi.org/10.1214/aos/1176324321

O'SULLIVAN, F. (1986). A statistical perspective on ill-posed inverse problems. *Statist. Sci.* **1** 502–527. MR0874480

O'SULLIVAN, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM J. Sci. Statist. Comput.* **9** 363–379. MR0930052 https://doi.org/10.1137/0909024

RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003). *Semiparametric Regression. Cambridge Series in Statistical and Probabilistic Mathematics* **12**. Cambridge Univ. Press, Cambridge. MR1998720 https://doi.org/10.1017/CBO9780511755453

SCHUMAKER, L. L. (1981). *Spline Functions: Basic Theory*. Wiley, New York. MR0606200

SCHWARZ, K. and KRIVOBOKOVA, T. (2016). A unified framework for spline estimators. *Biometrika* **103** 121–131. MR3465825 https://doi.org/10.1093/biomet/asv070

SILVERMAN, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.* **10** 795–810. MR0663433

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

STONE, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14** 590–606. MR0840516 https://doi.org/10.1214/aos/1176349940

STONE, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Statist.* **22** 118–184. MR1272079 https://doi.org/10.1214/aos/1176325361

UTRERAS, F. (1981). Optimal smoothing of noisy data using spline functions. *SIAM J. Sci. Statist. Comput.* **2** 349–362. MR0632905 https://doi.org/10.1137/0902028

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

WANG, X., SHEN, J. and RUPPERT, D. (2011). On the asymptotics of penalized spline smoothing. *Electron. J. Stat.* **5** 1–17. MR2763795 https://doi.org/10.1214/10-EJS593

WEINBERGER, H. F. (1974). *Variational Methods for Eigenvalue Approximation*. SIAM, Philadelphia, PA. MR0400004

XIAO, L. (2019a). Asymptotic theory of penalized splines. *Electron. J. Stat.* **13** 747–794. MR3925516 https://doi.org/10.1214/19-ejs1541

XIAO, L. (2019b). Asymptotics of bivariate penalised splines. *J. Nonparametr. Stat.* **31** 289–314. MR3941214 https://doi.org/10.1080/10485252.2018.1563295

XIAO, L., LI, Y. and RUPPERT, D. (2013). Fast bivariate P-splines: The sandwich smoother. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 577–599. MR3065480 https://doi.org/10.1111/rssb.12007

The Annals of Statistics
2021, Vol. 49, No. 6, 3408–3433
https://doi.org/10.1214/21-AOS2089

OPTIMAL ADAPTIVITY OF SIGNED-POLYGON STATISTICS FOR NETWORK TESTING

BY JIASHUN JIN[1,*], ZHENG TRACY KE[2] AND SHENGMING LUO[1,†]

[1]*Department of Statistics & Data Science, Carnegie Mellon University,* [*]*jiashun@stat.cmu.edu;* [†]*shengmil@andrew.cmu.edu*
[2]*Department of Statistics, Harvard University, zke@fas.harvard.edu*

Given a symmetric social network, we are interested in testing whether it has only one community or multiple communities. The desired tests should (a) accommodate severe degree heterogeneity, (b) accommodate mixed memberships, (c) have a tractable null distribution and (d) adapt automatically to different levels of sparsity, and achieve the optimal phase diagram. How to find such a test is a challenging problem.

We propose the Signed Polygon as a class of new tests. Fixing $m \geq 3$, for each m-gon in the network, define a score using the centered adjacency matrix. The sum of such scores is then the mth order Signed Polygon statistic. The Signed Triangle (SgnT) and the Signed Quadrilateral (SgnQ) are special examples of the Signed Polygon.

We show that both the SgnT and SgnQ tests satisfy (a)–(d), and especially, they work well for both very sparse and less sparse networks. Our proposed tests compare favorably with existing tests. For example, the EZ and GC tests behave unsatisfactorily in the less sparse case and do not achieve the optimal phase diagram. Also, many existing tests do not allow for severe heterogeneity or mixed memberships, and they behave unsatisfactorily in our settings.

The analysis of the SgnT and SgnQ tests is delicate and extremely tedious, and the main reason is that we need a unified proof that covers a wide range of sparsity levels and a wide range of degree heterogeneity. For lower bound theory, we use a phase transition framework, which includes the standard minimax argument, but is more informative. The proof uses classical theorems on matrix scaling.

REFERENCES

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

[2] BANERJEE, D. (2018). Contiguity and non-reconstruction results for planted partition models: The dense case. *Electron. J. Probab.* **23** Paper No. 18, 28. MR3771755 https://doi.org/10.1214/17-EJP128

[3] BANERJEE, D. and MA, Z. (2017). Optimal hypothesis testing for stochastic block models with growing degrees. Available at arXiv:1705.05305.

[4] BANKS, J., MOORE, C., NEEMAN, J. and NETRAPALLI, P. (2016). Information-theoretic thresholds for community detection in sparse networks. In *COLT* 383–416.

[5] BÉJAR, J., ÁLVAREZ, S., GARCÍA, D., GÓMEZ, I., OLIVA, L., TEJEDA, A. and VÁZQUEZ-SALCEDA, J. (2016). Discovery of spatio-temporal patterns from location-based social networks. *J. Exp. Theor. Artif. Intell.* **28** 313–329.

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

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

[8] 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

[9] BRUALDI, R. A. (1974). The DAD theorem for arbitrary row sums. *Proc. Amer. Math. Soc.* **45** 189–194. MR0354737 https://doi.org/10.2307/2040060

MSC2020 subject classifications. Primary 62H30, 91C20; secondary 62P25.
Key words and phrases. Asymptotic normality, DCBM, DCMM, lower bound, matrix scaling, optimal phase diagram, phase transition, SBM, signed quadrilateral, signed triangle, Sinkhorn's theorem, sparsity.

[10] 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

[11] CHATTERJEE, S. (2015). Matrix estimation by universal singular value thresholding. *Ann. Statist.* **43** 177–214. MR3285604 https://doi.org/10.1214/14-AOS1272

[12] DALL'AMICO, L., COUILLET, R. and TREMBLAY, N. (2019). Revisiting the Bethe-Hessian: Improved community detection in sparse heterogeneous graphs. In *Adv. Neural. Inf. Process. Syst.* 4037–4047. Curran Associates, Red Hook.

[13] 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

[14] DU, H. and YANG, S. J. (2011). Discovering collaborative cyber attack patterns using social network analysis. In *Proceedings of the* 4*th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction. SBP'*11 129–136. Springer, Berlin.

[15] FU, Y.-H., HUANG, C.-Y. and SUN, C.-T. (2015). Using global diversity and local topology features to identify influential network spreaders. *Phys. A* **433** 344–355. https://doi.org/10.1016/j.physa.2015.03.042

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

[17] GIRVAN, M. and NEWMAN, M. E. J. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **99** 7821–7826. MR1908073 https://doi.org/10.1073/pnas.122653799

[18] GULIKERS, L., LELARGE, M. and MASSOULIÉ, L. (2017). Non-backtracking spectrum of degree-corrected stochastic block models. In 8*th Innovations in Theoretical Computer Science Conference. LIPIcs. Leibniz Int. Proc. Inform.* **67** Art. No. 44, 27. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern. MR3754968

[19] GULIKERS, L., LELARGE, M. and MASSOULIÉ, L. (2018). An impossibility result for reconstruction in the degree-corrected stochastic block model. *Ann. Appl. Probab.* **28** 3002–3027. MR3847979 https://doi.org/10.1214/18-AAP1381

[20] 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

[21] HU, J., QIN, H., YAN, T., ZHANG, J. and ZHU, J. (2017). Using maximum entry-wise deviation to test the goodness-of-fit for stochastic block models. Available at arXiv:1703.06558.

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

[23] JIN, J. (2015). Fast community detection by SCORE. *Ann. Statist.* **43** 57–89. MR3285600 https://doi.org/10.1214/14-AOS1265

[24] JIN, J., KE, Z. T. and LUO, S. (2017). Estimating network memberships by simplex vertices hunting. Available at arXiv:1708.07852.

[25] JIN, J., KE, Z. T. and LUO, S. (2018). Network global testing by counting graphlets. In *Proceedings of the* 35*th International Conference on Machine Learning, ICML* 2018, *Stockholmsmässan, Stockholm, Sweden, July* 10–15, 2018 2338–2346.

[26] JIN, J., KE, Z. T. and LUO, S. (2021). Supplement to "Optimal adaptivity of signed-polygon statistics for network testing." https://doi.org/10.1214/21-AOS2089SUPP

[27] JIN, J., LI, W. and WANG, T. (2019). A spectral approach for network global testing. Manuscript.

[28] JOHNSON, C. R. and REAMS, R. (2009). Scaling of symmetric matrices by positive diagonal congruence. *Linear Multilinear Algebra* **57** 123–140. MR2492099 https://doi.org/10.1080/03081080600872327

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

[30] KE, Z. T. (2019). Comparison of different network testing procedures. Manuscript.

[31] LE, C. M. and LEVINA, E. (2015). Estimating the number of communities in networks by spectral methods. Available at arXiv:1507.00827.

[32] 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

[33] LI, T., LEI, L., BHATTACHARYYA, S., VAN DEN BERGE, K. SARKAR, P., BICKEL, P. J. and LEVINA, E. (2020). Hierarchical community detection by recursive partitioning. *J. Amer. Statist. Assoc.* https://doi.org/10.1080/01621459.2020.1833888

[34] MA, Z. and WU, Y. (2015). Computational barriers in minimax submatrix detection. *Ann. Statist.* **43** 1089–1116. MR3346698 https://doi.org/10.1214/14-AOS1300

[35] MARSHALL, A. W. and OLKIN, I. (1968). Scaling of matrices to achieve specified row and column sums. *Numer. Math.* **12** 83–90. MR0238875 https://doi.org/10.1007/BF02170999

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

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

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

[39] RADICCHI, F., CASTELLANO, C., CECCONI, F., LORETO, V. and PARISI, D. (2004). Defining and identifying communities in networks. *Proc. Natl. Acad. Sci. USA* **101** 2658–2663.

[40] SALDAÑA, D. F., YU, Y. and FENG, Y. (2017). How many communities are there? *J. Comput. Graph. Statist.* **26** 171–181. MR3610418 https://doi.org/10.1080/10618600.2015.1096790

[41] SINKHORN, R. (1974). Diagonal equivalence to matrices with prescribed row and column sums. II. *Proc. Amer. Math. Soc.* **45** 195–198. MR0357434 https://doi.org/10.2307/2040061

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

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

[44] ZHAO, Y., LEVINA, E. and ZHU, J. (2011). Community extraction for social networks. *Proc. Natl. Acad. Sci. USA* **108** 7321–7326.

The Annals of Statistics
2021, Vol. 49, No. 6, 3434–3459
https://doi.org/10.1214/21-AOS2090

ANALYSIS OF GENERALIZED BREGMAN SURROGATE ALGORITHMS FOR NONSMOOTH NONCONVEX STATISTICAL LEARNING

BY YIYUAN SHE*, ZHIFENG WANG† AND JIUWU JIN‡

Department of Statistics, Florida State University, *yshe@stat.fsu.edu; †wzfmath@gmail.com; ‡jj17g@my.fsu.edu

Modern statistical applications often involve minimizing an objective function that may be nonsmooth and/or nonconvex. This paper focuses on a broad Bregman-surrogate algorithm framework including the local linear approximation, mirror descent, iterative thresholding, DC programming and many others as particular instances. The recharacterization via generalized Bregman functions enables us to construct suitable error measures and establish global convergence rates for nonconvex and nonsmooth objectives in possibly high dimensions. For sparse learning problems with a composite objective, under some regularity conditions, the obtained estimators as the surrogate's fixed points, though not necessarily local minimizers, enjoy provable statistical guarantees, and the sequence of iterates can be shown to approach the statistical truth within the desired accuracy geometrically fast. The paper also studies how to design adaptive momentum based accelerations without assuming convexity or smoothness by carefully controlling stepsize and relaxation parameters.

REFERENCES

[1] AGARWAL, A., NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Fast global convergence of gradient methods for high-dimensional statistical recovery. *Ann. Statist.* **40** 2452–2482. MR3097609 https://doi.org/10.1214/12-AOS1032

[2] AN, L. T. H. and TAO, P. D. (2005). The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. *Ann. Oper. Res.* **133** 23–46. MR2119311 https://doi.org/10.1007/s10479-004-5022-1

[3] BANERJEE, A., MERUGU, S., DHILLON, I. S. and GHOSH, J. (2005). Clustering with Bregman divergences. *J. Mach. Learn. Res.* **6** 1705–1749. MR2249870

[4] BECK, A. and TEBOULLE, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2** 183–202. MR2486527 https://doi.org/10.1137/080716542

[5] BEN-TAL, A. and NEMIROVSKI, A. (2013). Optimization III: Convex analysis, nonlinear programming theory, standard nonlinear programming algorithms. Lecture Notes.

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

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

[8] BRÈGMAN, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. Math. Phys.* **7** 200–217. MR0215617

[9] BUNEA, F., TSYBAKOV, A. and WEGKAMP, M. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.* **1** 169–194. MR2312149 https://doi.org/10.1214/07-EJS008

[10] 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

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

[12] CHEN, G. and TEBOULLE, M. (1993). Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM J. Optim.* **3** 538–543. MR1230155 https://doi.org/10.1137/0803026

MSC2020 subject classifications. 90C26, 49J52, 68Q25.

Key words and phrases. Nonconvex optimization, nonsmooth optimization, MM algorithms, Bregman divergence, statistical algorithmic analysis, momentum-based acceleration.

[13] CICHOCKI, A., ICHI AMARI, S., ZDUNEK, R., KOMPASS, R., HORI, G. and HE, Z. (2006). Extended SMART algorithms for non-negative matrix factorization. In *ICAISC* (L. Rutkowski, R. Tadeusiewicz, L. A. Zadeh and J. M. Zurada, eds.). *Lecture Notes in Computer Science* **4029** 548–562. Springer.

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

[15] DUCHI, J., HAZAN, E. and SINGER, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* **12** 2121–2159. MR2825422

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

[17] FAN, J., XUE, L. and ZOU, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Ann. Statist.* **42** 819–849. MR3210988 https://doi.org/10.1214/13-AOS1198

[18] FRANK, I. E. and FRIEDMAN, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35** 109–135.

[19] GASSO, G., RAKOTOMAMONJY, A. and CANU, S. (2009). Recovering sparse signals with a certain family of nonconvex penalties and DC programming. *IEEE Trans. Signal Process.* **57** 4686–4698. MR2722328 https://doi.org/10.1109/TSP.2009.2026004

[20] GHADIMI, S. and LAN, G. (2016). Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Math. Program.* **156** 59–99. MR3459195 https://doi.org/10.1007/s10107-015-0871-8

[21] HUBER, P. J. (1981). *Robust Statistics. Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. MR0606374

[22] HUNTER, D. R. and LANGE, K. (2004). A tutorial on MM algorithms. *Amer. Statist.* **58** 30–37. MR2055509 https://doi.org/10.1198/0003130042836

[23] HUNTER, D. R. and LI, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* **33** 1617–1642. MR2166557 https://doi.org/10.1214/009053605000000200

[24] JØRGENSEN, B. (1987). Exponential dispersion models. *J. Roy. Statist. Soc. Ser. B* **49** 127–162. With discussion and a reply by the author. MR0905186

[25] KOLTCHINSKII, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems. Lecture Notes in Math.* **2033**. Springer, Heidelberg. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School]. MR2829871 https://doi.org/10.1007/978-3-642-22147-7

[26] KRICHENE, W., BAYEN, A. and BARTLETT, P. L. (2015). Accelerated mirror descent in continuous and discrete time. In *Advances in Neural Information Processing Systems* (C. Cortes, N. Lawrence, D. Lee, M. Sugiyama and R. Garnett, eds.) **28**. Curran Associates.

[27] LANGE, K. and ZHOU, H. (2014). MM algorithms for geometric and signomial programming. *Math. Program.* **143** 339–356. MR3152072 https://doi.org/10.1007/s10107-012-0612-1

[28] LEE, D. D. and SEUNG, H. S. (1999). Learning the parts of objects by nonnegative matrix factorization. *Nature* **401** 788–791.

[29] LOH, P.-L. and WAINWRIGHT, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *J. Mach. Learn. Res.* **16** 559–616. MR3335800

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

[31] NEMIROVSKY, A. S. and YUDIN, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization. A Wiley-Interscience Publication*. Wiley, New York. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics. MR0702836

[32] NESTEROV, Y. (2004). *Introductory Lectures on Convex Optimization: A Basic Course. Applied Optimization* **87**. Kluwer Academic, Boston, MA. MR2142598 https://doi.org/10.1007/978-1-4419-8853-9

[33] NESTEROV, YU. E. (1983). A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Sov. Math., Dokl.* **27** 372–376. MR0701288

[34] NESTEROV, YU. E. (1988). An approach to constructing optimal methods for minimization of smooth convex functions. *Èkon. Mat. Metody* **24** 509–517. MR0968064

[35] PAN, W., SHEN, X. and LIU, B. (2013). Cluster analysis: Unsupervised learning via supervised learning with a non-convex penalty. *J. Mach. Learn. Res.* **14** 1865–1889. MR3104498

[36] ROCKAFELLAR, R. T. (1970). *Convex Analysis. Princeton Mathematical Series* **28**. Princeton Univ. Press, Princeton, NJ. MR0274683

[37] SCHMIDT, M. (2010). Graphical model structure learning with ℓ_1-regularization. Ph.D. thesis, Univ. British Columbia.

[38] SHE, Y. (2009). Thresholding-based iterative selection procedures for model selection and shrinkage. *Electron. J. Stat.* **3** 384–415. MR2501318 https://doi.org/10.1214/08-EJS348

[39] SHE, Y. (2012). An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. *Comput. Statist. Data Anal.* **56** 2976–2990. MR2929353 https://doi.org/10.1016/j.csda.2011.11.013

[40] SHE, Y. (2016). On the finite-sample analysis of Θ-estimators. *Electron. J. Stat.* **10** 1874–1895. MR3522663 https://doi.org/10.1214/15-EJS1100

[41] SHE, Y., HE, Y. and WU, D. (2014). Learning topology and dynamics of large recurrent neural networks. *IEEE Trans. Signal Process.* **62** 5881–5891. MR3281530 https://doi.org/10.1109/TSP.2014.2358956

[42] SHE, Y., WANG, J., LI, H. and WU, D. (2013). Group iterative spectrum thresholding for super-resolution sparse spectral selection. *IEEE Trans. Signal Process.* **61** 6371–6386. MR3148325 https://doi.org/10.1109/TSP.2013.2281303

[43] SHE, Y., WANG, Z. and JIN, J. (2021). Supplement to "Analysis of generalized Bregman surrogate algorithms for nonsmooth nonconvex statistical learning." https://doi.org/10.1214/21-AOS2090SUPP

[44] TAO, P. D. and SOUAD, E. B. (1986). Algorithms for solving a class of nonconvex optimization problems. Methods of subgradients. In *FERMAT Days* 85: *Mathematics for Optimization* (*Toulouse*, 1985). *North-Holland Math. Stud.* **129** 249–271. North-Holland, Amsterdam. MR0874369 https://doi.org/10.1016/S0304-0208(08)72402-2

[45] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[46] TSENG, P. (2008). On accelerated proximal gradient methods for convex-concave optimization. Dept. Mathematics, Univ. Washington.

[47] TSYBAKOV, A. B. (2008). *Introduction to Nonparametric Estimation*. Springer, New York. MR2724359

[48] VAN DE GEER, S. A. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **3** 1360–1392. MR2576316 https://doi.org/10.1214/09-EJS506

[49] 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

[50] WAINWRIGHT, M. J. and JORDAN, M. I. (2008). Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* **1** 1–305.

[51] WANG, Z., LIU, H. and ZHANG, T. (2014). Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Ann. Statist.* **42** 2164–2201. MR3269977 https://doi.org/10.1214/14-AOS1238

[52] ZHANG, C., JIANG, Y. and CHAI, Y. (2010). Penalized Bregman divergence for large-dimensional regression and classification. *Biometrika* **97** 551–566. With supplementary data available online. MR2672483 https://doi.org/10.1093/biomet/asq033

[53] 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

[54] ZHANG, T. (2010). Analysis of multi-stage convex relaxation for sparse regularization. *J. Mach. Learn. Res.* **11** 1081–1107. MR2629825

[55] ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533. MR2435443 https://doi.org/10.1214/009053607000000802

The Annals of Statistics
2021, Vol. 49, No. 6, 3460–3481
https://doi.org/10.1214/21-AOS2092

AN ASYMPTOTIC TEST FOR CONSTANCY OF THE VARIANCE UNDER SHORT-RANGE DEPENDENCE

BY SARA K. SCHMIDT[1,*], MAX WORNOWIZKI[2,‡], ROLAND FRIED[2,§] AND HEROLD DEHLING[1,†]

[1]*Department of Mathematics, Ruhr-Universität Bochum,* [*]*sara.schmidt@ruhr-uni-bochum.de;* [†]*herold.dehling@ruhr-uni-bochum.de*

[2]*Department of Statistics, TU Dortmund University,* [‡]*wornowizki@statistik.tu-dortmund.de;* [§]*fried@statistik.tu-dortmund.de*

We present a novel approach to test for heteroscedasticity of a nonstationary time series that is based on Gini's mean difference of logarithmic local sample variances. In order to analyse the large sample behaviour of our test statistic, we establish new limit theorems for U-statistics of dependent triangular arrays. We derive the asymptotic distribution of the test statistic under the null hypothesis of a constant variance and show that the test is consistent against a large class of alternatives, including multiple structural breaks in the variance. Our test is applicable even in the case of nonstationary processes, assuming a locally stationary mean function. The performance of the test and its comparatively low computation time are illustrated in an extensive simulation study. As an application, we analyse Google Trends data, monitoring the relative search interest for the topic "global warming."

REFERENCES

[1] ABRAHAM, B. and WEI, W. W. S. (1984). Inferences about the parameters of a time series model with changing variance. *Metrika* **31** 183–194. MR0754960 https://doi.org/10.1007/BF01915199

[2] ANDEREGG, W. R. L. and GOLDSMITH, G. R. (2014). Public interest in climate change over the past decade and the effects of the 'climategate' media event. *Environ. Res. Lett.* **9** 54005.

[3] ANDREOU, E. and GHYSELS, E. (2002). Detecting multiple breaks in financial market volatility dynamics. *J. Appl. Econometrics* **17** 579–600.

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

[5] BAUFAYS, P. and RASSON, J.-P. (1985). Variance changes in autoregressive model. In *Time Series Analysis: Theory and Practice*, 7 (*Toronto*, *Ont.*, 1983) 119–128. North-Holland, Amsterdam. MR0787816

[6] BLOOMFIELD, P., HURD, H. L. and LUND, R. B. (1994). Periodic correlation in stratospheric ozone data. *J. Time Series Anal.* **15** 127–150. MR1263886 https://doi.org/10.1111/j.1467-9892.1994.tb00181.x

[7] BOROVKOVA, S., BURTON, R. and DEHLING, H. (2001). Limit theorems for functionals of mixing processes with applications to U-statistics and dimension estimation. *Trans. Amer. Math. Soc.* **353** 4261–4318. MR1851171 https://doi.org/10.1090/S0002-9947-01-02819-7

[8] BRADLEY, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surv.* **2** 107–144. MR2178042 https://doi.org/10.1214/154957805100000104

[9] BRADLEY, R. C. (2007). *Introduction to Strong Mixing Conditions. Vol.* 2. Kendrick Press, Heber City, UT. MR2325295

[10] BURIVALOVA, Z., BUTLER, R. A. and WILCOVE, D. S. (2018). Analyzing Google search data to debunk myths about the public's interest in conservation. *Front. Ecol. Environ.* **16** 509–514.

[11] CARLSTEIN, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *Ann. Statist.* **14** 1171–1179. MR0856813 https://doi.org/10.1214/aos/1176350057

[12] CHEN, J. and GUPTA, A. K. (1997). Testing and locating variance changepoints with application to stock prices. *J. Amer. Statist. Assoc.* **92** 739–747. MR1467863 https://doi.org/10.2307/2965722

MSC2020 subject classifications. Primary 62G10; secondary 62M10, 60F05.

Key words and phrases. Change-point analysis, tests for heteroscedasticity, U-statistics of triangular arrays, short-range dependence.

[13] DAHLHAUS, R. (1997). Fitting time series models to nonstationary processes. *Ann. Statist.* **25** 1–37. MR1429916 https://doi.org/10.1214/aos/1034276620

[14] DAVIS, R. A., HUANG, D. W. and YAO, Y.-C. (1995). Testing for a change in the parameter values and order of an autoregressive model. *Ann. Statist.* **23** 282–304. MR1331669 https://doi.org/10.1214/aos/1176324468

[15] DETTE, H., WU, W. and ZHOU, Z. (2015). Change point analysis of second order characteristics in non-stationary time series. Available at arXiv:1503.08610.

[16] DETTE, H., WU, W. and ZHOU, Z. (2019). Change point analysis of correlation in non-stationary time series. *Statist. Sinica* **29** 611–643. MR3931381

[17] EICHINGER, B. and KIRCH, C. (2018). A MOSUM procedure for the estimation of multiple random change points. *Bernoulli* **24** 526–564. MR3706768 https://doi.org/10.3150/16-BEJ887

[18] GALEANO, P. and PEÑA, D. (2007). Covariance changes detection in multivariate time series. *J. Statist. Plann. Inference* **137** 194–211. MR2292851 https://doi.org/10.1016/j.jspi.2005.09.003

[19] GAO, Z., SHANG, Z., DU, P. and ROBERTSON, J. L. (2019). Variance change point detection under a smoothly-changing mean trend with application to liver procurement. *J. Amer. Statist. Assoc.* **114** 773–781. MR3963179 https://doi.org/10.1080/01621459.2018.1442341

[20] GERSTENBERGER, C. and VOGEL, D. (2015). On the efficiency of Gini's mean difference. *Stat. Methods Appl.* **24** 569–596. MR3421674 https://doi.org/10.1007/s10260-015-0315-x

[21] GERSTENBERGER, C., VOGEL, D. and WENDLER, M. (2020). Tests for scale changes based on pairwise differences. *J. Amer. Statist. Assoc.* **115** 1336–1348. MR4143469 https://doi.org/10.1080/01621459.2019.1629938

[22] GOMBAY, E., HORVÁTH, L. and HUŠKOVÁ, M. (1996). Estimators and tests for change in variances. *Statist. Decisions* **14** 145–159. MR1406521

[23] INCLÁN, C. and TIAO, G. C. (1994). Use of cumulative sums of squares for retrospective detection of changes of variance. *J. Amer. Statist. Assoc.* **89** 913–923. MR1294735

[24] LEE, S. and PARK, S. (2001). The cusum of squares test for scale changes in infinite order moving average processes. *Scand. J. Stat.* **28** 625–644. MR1876504 https://doi.org/10.1111/1467-9469.00259

[25] LINDNER, A. M. (2009). Stationarity, mixing, distributional properties and moments of GARCH(p, q)-processes. In *Handbook of Financial Time Series* 43–69. Springer, Berlin.

[26] MOKKADEM, A. (1988). Mixing properties of ARMA processes. *Stochastic Process. Appl.* **29** 309–315. MR0958507 https://doi.org/10.1016/0304-4149(88)90045-2

[27] R CORE TEAM (2019). *R: a Language and Environment for Statistical Computing*. Vienna, Austria.

[28] SCHMIDT, S. K., WORNOWIZKI, M., FRIED, R. and DEHLING, H. (2021). Supplement to "An asymptotic test for constancy of the variance under short-range dependence." https://doi.org/10.1214/21-AOS2092SUPPA, https://doi.org/10.1214/21-AOS2092SUPPB

[29] WICHERN, D. W., MILLER, R. B. and HSU, D.-A. (1976). Changes of variance in first-order autoregressive time series models—with an application. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **25** 248–256.

[30] WIED, D., ARNOLD, M., BISSANTZ, N. and ZIGGEL, D. (2012). A new fluctuation test for constant variances with applications to finance. *Metrika* **75** 1111–1127. MR2989322 https://doi.org/10.1007/s00184-011-0371-7

[31] WORNOWIZKI, M., FRIED, R. and MEINTANIS, S. G. (2017). Fourier methods for analyzing piecewise constant volatilities. *AStA Adv. Stat. Anal.* **101** 289–308. MR3679347 https://doi.org/10.1007/s10182-017-0288-1

[32] WU, W. B. and ZHAO, Z. (2007). Inference of trends in time series. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 391–410. MR2323759 https://doi.org/10.1111/j.1467-9868.2007.00594.x

The Annals of Statistics
2021, Vol. 49, No. 6, 3482–3509
https://doi.org/10.1214/21-AOS2093

UNCERTAINTY QUANTIFICATION FOR BAYESIAN CART

BY ISMAËL CASTILLO[1] AND VERONIKA ROČKOVÁ[2]

[1]*Laboratoire de Probabilités, Statistique et Modélisation, Institut Universitaire de France, Sorbonne Université,* *ismael.castillo@upmc.fr*

[2]*Booth School of Business, University of Chicago,* *veronika.rockova@chicagobooth.edu*

This work affords new insights into Bayesian CART in the context of structured wavelet shrinkage. The main thrust is to develop a formal inferential framework for Bayesian tree-based regression. We reframe Bayesian CART as a *g-type* prior which departs from the typical wavelet product priors by harnessing correlation induced by the tree topology. The practically used Bayesian CART priors are shown to attain adaptive near rate-minimax posterior concentration in the *supremum norm* in regression models. For the fundamental goal of uncertainty quantification, we construct *adaptive* confidence bands for the regression function with uniform coverage under self-similarity. In addition, we show that tree-posteriors enable optimal inference in the form of efficient confidence sets for smooth functionals of the regression function.

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/sjos.12002

[2] BARANIUK, R. G., CEVHER, V., DUARTE, M. F. and HEGDE, C. (2010). Model-based compressive sensing. *IEEE Trans. Inf. Theory* **56** 1982–2001. MR2654489 https://doi.org/10.1109/TIT.2010.2040894

[3] BARBIERI, M. M. and BERGER, J. O. (2004). Optimal predictive model selection. *Ann. Statist.* **32** 870–897. MR2065192 https://doi.org/10.1214/009053604000000238

[4] BARBIERI, M. M., BERGER, J. O., GEORGE, E. I. and ROČKOVÁ, V. (2018). The median probability model and correlated variables. Preprint. Available at arXiv:1807.08336.

[5] BAYARRI, M. J., BERGER, J. O., FORTE, A. and GARCÍA-DONATO, G. (2012). Criteria for Bayesian model choice with application to variable selection. *Ann. Statist.* **40** 1550–1577. MR3015035 https://doi.org/10.1214/12-AOS1013

[6] BIAU, G. (2012). Analysis of a random forests model. *J. Mach. Learn. Res.* **13** 1063–1095. MR2930634

[7] BLANCHARD, G., SCHÄFER, C. and ROZENHOLC, Y. (2004). Oracle bounds and exact algorithm for dyadic classification trees. In *Learning Theory. Lecture Notes in Computer Science* **3120** 378–392. Springer, Berlin. MR2177922 https://doi.org/10.1007/978-3-540-27819-1_26

[8] 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

[9] 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

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

[11] 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

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

[13] 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

MSC2020 subject classifications. 62G20, 62G15.

Key words and phrases. Bayesian CART, posterior concentration, recursive partitioning, regression trees, nonparametric Bernstein–von Mises theorem.

[14] CASTILLO, I. (2014). On Bayesian supremum norm contraction rates. *Ann. Statist.* **42** 2058–2091. MR3262477 https://doi.org/10.1214/14-AOS1253

[15] 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

[16] CASTILLO, I. and MISMER, R. (2021). Spike and slab Pólya tree posterior densities: Adaptive inference. *Ann. Inst. Henri Poincaré Probab. Stat.* **57** 1521–1548. MR4291462 https://doi.org/10.1214/20-aihp1132

[17] 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

[18] 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

[19] CASTILLO, I. and ROČKOVÁ, V. (2021). Supplement to "Uncertainty quantification for Bayesian CART." https://doi.org/10.1214/21-AOS2093SUPP

[20] CHIPMAN, H., GEORGE, E. I. and MCCULLOCH, R. E. (1997). Bayesian CART model search. *J. Amer. Statist. Assoc.* **93** 935–960.

[21] CHIPMAN, H., GEORGE, E. I., MCCULLOCH, R. E. and SHIVELY, T. (2016). High-dimensional nonparametric monotone function estimation using BART. Preprint. Available at arXiv:1612.01619.

[22] 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

[23] CHIPMAN, H. A., KOLACZYK, E. D. and MCCULLOCH, R. E. (1997). Adaptive Bayesian wavelet shrinkage. *J. Amer. Statist. Assoc.* **92** 1413–1421.

[24] 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

[25] 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

[26] DONOHO, D. L. (1997). CART and best-ortho-basis: A connection. *Ann. Statist.* **25** 1870–1911. MR1474073 https://doi.org/10.1214/aos/1069362377

[27] 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

[28] FERNÁNDEZ, C., LEY, E. and STEEL, M. F. J. (2001). Benchmark priors for Bayesian model averaging. *J. Econometrics* **100** 381–427. MR1820410 https://doi.org/10.1016/S0304-4076(00)00076-2

[29] FOSTER, D. P. and GEORGE, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22** 1947–1975. MR1329177 https://doi.org/10.1214/aos/1176325766

[30] FRYZLEWICZ, P. (2007). Unbalanced Haar technique for nonparametric function estimation. *J. Amer. Statist. Assoc.* **102** 1318–1327. MR2412552 https://doi.org/10.1198/016214507000000860

[31] GEY, S. and NÉDÉLEC, E. (2005). Model selection for CART regression trees. *IEEE Trans. Inf. Theory* **51** 658–670. MR2236074 https://doi.org/10.1109/TIT.2004.840903

[32] 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

[33] GHOSAL, S. and VAN DER VAART, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. *Cambridge Series in Statistical and Probabilistic Mathematics* **44**. Cambridge Univ. Press, Cambridge. MR3587782 https://doi.org/10.1017/9781139029834

[34] GINÉ, E. and NICKL, R. (2011). Rates of contraction for posterior distributions in L^r-metrics, $1 \le r \le \infty$. *Ann. Statist.* **39** 2883–2911. MR3012395 https://doi.org/10.1214/11-AOS924

[35] 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

[36] GIRARDI, M. and SWELDENS, W. (1997). A new class of unbalanced Haar wavelets that form an unconditional basis for L_p on general measure spaces. *J. Fourier Anal. Appl.* **3** 457–474. MR1468375 https://doi.org/10.1007/BF02649107

[37] HAHN, P. R., MURRAY, J. S. and CARVALHO, C. (2017). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. Preprint. Available at arXiv:1706.09523.

[38] HILL, J. L. (2011). Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Statist.* **20** 217–240. MR2816546 https://doi.org/10.1198/jcgs.2010.08162

[39] 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

[40] KASS, R. E. and WASSERMAN, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Amer. Statist. Assoc.* **90** 928–934. MR1354008

[41] 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

[42] 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. MR3874311 https://doi.org/10.1111/rssb.12293

[43] LIU, Y., ROČKOVÁ, V. and WANG, Y. (2018). ABC variable selection with Bayesian forests. Preprint. Available at arXiv:1806.02304.

[44] MENTCH, L. and HOOKER, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *J. Mach. Learn. Res.* **17** Paper No. 26, 41 pp. MR3491120

[45] MURRAY, J. (2017). Log-linear Bayesian additive regression trees for categorical and count responses. Preprint. Available at arXiv:1706.09523.

[46] NAULET, Z. (2018). Adaptive Bayesian density estimation in sup-norm. Preprint. Available at arXiv:1805.05816.

[47] NICKL, R. and RAY, K. (2020). Nonparametric statistical inference for drift vector fields of multi-dimensional diffusions. *Ann. Statist.* **48** 1383–1408. MR4124327 https://doi.org/10.1214/19-AOS1851

[48] NICKL, R. and SÖHL, J. (2019). Bernstein–von Mises theorems for statistical inverse problems II: Compound Poisson processes. *Electron. J. Stat.* **13** 3513–3571. MR4013745 https://doi.org/10.1214/19-ejs1609

[49] 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

[50] ROČKOVÁ, V. and SAHA, E. (2019). On theory for BART. In *Proceedings of Machine Learning Research*: 22*nd International Conference on Artificial Intelligence and Statistics* **89** 2839–2848.

[51] ROČKOVÁ, V. and VAN DER PAS, S. (2020). Posterior concentration for Bayesian regression trees and forests. *Ann. Statist.* **48** 2108–2131. MR4134788 https://doi.org/10.1214/19-AOS1879

[52] 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

[53] SCOTT, C. and NOWAK, R. D. (2006). Minimax-optimal classification with dyadic decision trees. *IEEE Trans. Inf. Theory* **52** 1335–1353. MR2241192 https://doi.org/10.1109/TIT.2006.871056

[54] SCRICCIOLO, C. (2014). Adaptive Bayesian density estimation in L^p-metrics with Pitman–Yor or normalized inverse-Gaussian process kernel mixtures. *Bayesian Anal.* **9** 475–520. MR3217004 https://doi.org/10.1214/14-BA863

[55] 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

[56] VAN DER PAS, S. and ROČKOVÁ, V. (2017). Bayesian dyadic trees and histograms for regression. In *Proceedings of the* 31*st International Conference on Neural Information Processing Systems* 2086–2096.

[57] 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

[58] WAGER, S. and WALTHER, G. (2015). Adaptive concentration of regression trees, with application to random forests. Preprint. Available at arXiv:1503.06388.

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

[60] WONG, W. H. and MA, L. (2010). Optional Pólya tree and Bayesian inference. *Ann. Statist.* **38** 1433–1459. MR2662348 https://doi.org/10.1214/09-AOS755

[61] YOO, W., RIVOIRARD, R. and ROUSSEAU, J. (2017). Adaptive supremum norm posterior contraction: Wavelet spike-and-slab and anisotropic Besov spaces. Preprint. Available at arXiv:1708.01909.

[62] YOO, W. and VAN DER VAART, A. (2017). The Bayes Lepski's method and credible bands through volume of tubular neighborhoods. Preprint. Available at arXiv:1711.06926.

[63] 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

[64] ZELLNER, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques*. *Stud. Bayesian Econometrics Statist.* **6** 233–243. North-Holland, Amsterdam. MR0881437

The Annals of Statistics
2021, Vol. 49, No. 6, 3510–3533
https://doi.org/10.1214/21-AOS2095

TOTAL VARIATION REGULARIZED FRÉCHET REGRESSION FOR METRIC-SPACE VALUED DATA

BY ZHENHUA LIN[1] AND HANS-GEORG MÜLLER[2]

[1]*Department of Statistics and Data Science, National University of Singapore, linz@nus.edu.sg*
[2]*Department of Statistics, University of California, Davis, hgmueller@ucdavis.edu*

Non-Euclidean data that are indexed with a scalar predictor such as time are increasingly encountered in data applications, while statistical methodology and theory for such random objects are not well developed yet. To address the need for new methodology in this area, we develop a total variation regularization technique for nonparametric Fréchet regression, which refers to a regression setting where a response residing in a metric space is paired with a scalar predictor and the target is a conditional Fréchet mean. Specifically, we seek to approximate an unknown metric-space valued function by an estimator that minimizes the Fréchet version of least squares and at the same time has small total variation, appropriately defined for metric-space valued objects. We show that the resulting estimator is representable by a piece-wise constant function and establish the minimax convergence rate of the proposed estimator for metric data objects that reside in Hadamard spaces. We illustrate the numerical performance of the proposed method for both simulated and real data, including metric spaces of symmetric positive-definite matrices with the affine-invariant distance, of probability distributions on the real line with the Wasserstein distance, and of phylogenetic trees with the Billera–Holmes–Vogtmann metric.

REFERENCES

AFSARI, B. (2011). Riemannian L^p center of mass: Existence, uniqueness, and convexity. *Proc. Amer. Math. Soc.* **139** 655–673. MR2736346 https://doi.org/10.1090/S0002-9939-2010-10541-5

ALQUIER, P., COTTET, V. and LECUÉ, G. (2019). Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions. *Ann. Statist.* **47** 2117–2144. MR3953446 https://doi.org/10.1214/18-AOS1742

ARSIGNY, V., FILLARD, P., PENNEC, X. and AYACHE, N. (2006). Log-Euclidean metrics for fast and simple calculus on diffusion tensors. *Magn. Reson. Med.* **56** 411–421.

ARSIGNY, V., FILLARD, P., PENNEC, X. and AYACHE, N. (2007). Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM J. Matrix Anal. Appl.* **29** 328–347. MR2288028 https://doi.org/10.1137/050637996

BAČÁK, M. (2015). Convergence of nonlinear semigroups under nonpositive curvature. *Trans. Amer. Math. Soc.* **367** 3929–3953. MR3324915 https://doi.org/10.1090/S0002-9947-2015-06087-5

BERGMANN, R. and WEINMANN, A. (2016). A second-order TV-type approach for inpainting and denoising higher dimensional combined cyclic and vector space data. *J. Math. Imaging Vision* **55** 401–427. MR3489791 https://doi.org/10.1007/s10851-015-0627-3

BERGMANN, R., LAUS, F., STEIDL, G. and WEINMANN, A. (2014). Second order differences of cyclic data and applications in variational denoising. *SIAM J. Imaging Sci.* **7** 2916–2953. MR3293451 https://doi.org/10.1137/140969993

BHATTACHARYA, R. and PATRANGENARU, V. (2003). Large sample theory of intrinsic and extrinsic sample means on manifolds. I. *Ann. Statist.* **31** 1–29. MR1962498 https://doi.org/10.1214/aos/1046294456

BILLERA, L. J., HOLMES, S. P. and VOGTMANN, K. (2001). Geometry of the space of phylogenetic trees. *Adv. in Appl. Math.* **27** 733–767. MR1867931 https://doi.org/10.1006/aama.2001.0759

MSC2020 subject classifications. Primary 62R20; secondary 62R30.
Key words and phrases. Brain imaging, Fréchet mean, Hadamard space, Alexandrov space, non-Euclidean data, symmetric positive-definite matrix, random objects, Wasserstein metric, phylogenetic tree.

BRIDSON, M. R. (1991). Geodesics and curvature in metric simplicial complexes. Ph.D. thesis, Cornell University. MR2686786

BRIDSON, M. R. and HAEFLIGER, A. (1999). *Metric Spaces of Non-positive Curvature*. *Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **319**. Springer, Berlin. MR1744486 https://doi.org/10.1007/978-3-662-12494-9

BURAGO, D., BURAGO, Y. and IVANOV, S. (2001). *A Course in Metric Geometry*. *Graduate Studies in Mathematics* **33**. Amer. Math. Soc., Providence, RI. MR1835418 https://doi.org/10.1090/gsm/033

CHAMBOLLE, A., CASELLES, V., CREMERS, D., NOVAGA, M. and POCK, T. (2010). An introduction to total variation for image analysis. In *Theoretical Foundations and Numerical Methods for Sparse Recovery*. *Radon Ser. Comput. Appl. Math.* **9** 263–340. de Gruyter, Berlin. MR2731599 https://doi.org/10.1515/9783110226157.263

CHANG, T. (1989). Spherical regression with errors in variables. *Ann. Statist.* **17** 293–306. MR0981451 https://doi.org/10.1214/aos/1176347017

CHINOT, G., LECUÉ, G. and LERASLE, M. (2020). Robust statistical learning with Lipschitz and convex loss functions. *Probab. Theory Related Fields* **176** 897–940. MR4087486 https://doi.org/10.1007/s00440-019-00931-3

CORNEA, E., ZHU, H., KIM, P. and IBRAHIM, J. G. (2017). Regression models on Riemannian symmetric spaces. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 463–482. MR3611755 https://doi.org/10.1111/rssb.12169

DAI, X. and MÜLLER, H.-G. (2018). Principal component analysis for functional data on Riemannian manifolds and spheres. *Ann. Statist.* **46** 3334–3361. MR3852654 https://doi.org/10.1214/17-AOS1660

DAVIS, B. C., FLETCHER, P. T., BULLITT, E. and JOSHI, S. (2010). Population shape regression from random design data. *Int. J. Comput. Vis.* **90** 255–266.

DESIKAN, R. S., SÉGONNE, F., FISCHL, B., QUINN, B. T., DICKERSON, B. C., BLACKER, D., BUCKNER, R. L., DALE, A. M., MAGUIRE, R. P. et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* **31** 968–980.

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

DRYDEN, I. L., KOLOYDENKO, A. and ZHOU, D. (2009). Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *Ann. Appl. Stat.* **3** 1102–1123. MR2750388 https://doi.org/10.1214/09-AOAS249

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

DUBEY, P. and MÜLLER, H.-G. (2020b). Functional models for time-varying random objects. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 275–327. MR4084166

ELTZNER, B. and HUCKEMANN, S. F. (2019). A smeary central limit theorem for manifolds with application to high-dimensional spheres. *Ann. Statist.* **47** 3360–3381. MR4025745 https://doi.org/10.1214/18-AOS1781

ESSEN, D. C. V., SMITH, S. M., BARCH, D. M., BEHRENS, T. E. J., YACOUB, E., UGURBIL, K. and WU-MINN HCP CONSORTIUM (2013). The Wu–Minn human connectome project: An overview. *NeuroImage* **80** 62–79.

FANG, B., GUNTUBOYINA, A. and SEN, B. (2021). Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and Hardy–Krause variation. *Ann. Statist.* **49** 769–792. MR4255107 https://doi.org/10.1214/20-aos1977

FARAWAY, J. J. (2014). Regression for non-Euclidean data using distance matrices. *J. Appl. Stat.* **41** 2342–2357. MR3256391 https://doi.org/10.1080/02664763.2014.909794

FILLARD, P., ARSIGNY, V., AYACHE, N. and PENNEC, X. (2005). A Riemannian framework for the processing of tensor-valued images. In *International Workshop on Deep Structure, Singularities, and Computer Vision* 112–123.

FISHER, N. I. (1995). *Statistical Analysis of Circular Data*. Cambridge Univ. Press, Cambridge. MR1251957 https://doi.org/10.1017/CBO9780511564345

FLETCHER, P. T. (2013). Geodesic regression and the theory of least squares on Riemannian manifolds. *Int. J. Comput. Vis.* **105** 171–185. MR3104017 https://doi.org/10.1007/s11263-012-0591-y

FLETCHER, T. and JOSHI, S. (2007). Riemannian geometry for the statistical analysis of diffusion tensor data. *Signal Process.* **87** 250–262.

FRISTON, K. J. (2011). Functional and effective connectivity: A review. *Brain Connect.* **1** 13–36.

GREEN, M. F., HORAN, W. P. and LEE, J. (2015). Social cognition in schizophrenia. *Nat. Rev. Neurosci.* **16** 620–631.

HEIN, M. (2009). Robust nonparametric regression with metric-space valued output. In *Advances in Neural Information Processing Systems* 718–726.

HINKLE, J., FLETCHER, P. T. and JOSHI, S. (2014). Intrinsic polynomials for regression on Riemannian manifolds. *J. Math. Imaging Vision* **50** 32–52. MR3233133 https://doi.org/10.1007/s10851-013-0489-5

HOTZ, T., HUCKEMANN, S., LE, H., MARRON, J. S., MATTINGLY, J. C., MILLER, E., NOLEN, J., OWEN, M., PATRANGENARU, V. et al. (2013). Sticky central limit theorems on open books. *Ann. Appl. Probab.* **23** 2238–2258. MR3127934 https://doi.org/10.1214/12-AAP899

HÜTTER, J.-C. and RIGOLLET, P. (2016). Optimal rates for total variation denoising. In 29*th Annual Conference on Learning Theory* (V. Feldman, A. Rakhlin and O. Shamir, eds.). *Proceedings of Machine Learning Research* **49** 1115–1146. PMLR, Columbia University, New York, New York, USA.

KIM, S.-J., KOH, K., BOYD, S. and GORINEVSKY, D. (2009). l_1 trend filtering. *SIAM Rev.* **51** 339–360. MR2505584 https://doi.org/10.1137/070690274

KLOECKNER, B. (2010). A geometric study of Wasserstein spaces: Euclidean spaces. *Ann. Sc. Norm. Super. Pisa Cl. Sci.* (5) **9** 297–323. MR2731158

KOLAR, M. and XING, E. P. (2012). Estimating networks with jumps. *Electron. J. Stat.* **6** 2069–2106. MR3020257 https://doi.org/10.1214/12-EJS739

LEDOUX, M. and TALAGRAND, M. (2011). *Probability in Banach Spaces: Isoperimetry and Processes*. *Classics in Mathematics*. Springer, Berlin. MR2814399

LANG, S. (1995). *Differential and Riemannian Manifolds*. Springer, New York.

LELLMANN, J., STREKALOVSKIY, E., KOETTER, S. and CREMERS, D. (2013). Total variation regularization for functions with values in a manifold. In 2013 *IEEE International Conference on Computer Vision* 2944–2951. IEEE, New York.

LIN, Z. (2019). Riemannian geometry of symmetric positive definite matrices via Cholesky decomposition. *SIAM J. Matrix Anal. Appl.* **40** 1353–1370. MR4032859 https://doi.org/10.1137/18M1221084

LIN, Z. and MÜLLER, H.-G. (2021). Supplement to "Total variation regularized Fréchet regression for metric-space valued data." https://doi.org/10.1214/21-AOS2095SUPP

MAMMEN, E. and VAN DE GEER, S. (1997). Locally adaptive regression splines. *Ann. Statist.* **25** 387–413. MR1429931 https://doi.org/10.1214/aos/1034276635

MOAKHER, M. (2005). A differential geometric approach to the geometric mean of symmetric positive-definite matrices. *SIAM J. Matrix Anal. Appl.* **26** 735–747. MR2137480 https://doi.org/10.1137/S0895479803436937

ORTELLI, F. and VAN DE GEER, S. (2018). On the total variation regularized estimator over a class of tree graphs. *Electron. J. Stat.* **12** 4517–4570. MR3892703 https://doi.org/10.1214/18-ejs1519

OWEN, M. and PROVAN, J. S. (2011). A fast algorithm for computing geodesic distances in tree space. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **8** 2–13.

PATRANGENARU, V. and ELLINGSON, L. (2015). *Nonparametric Statistics on Manifolds and Their Applications to Object Data Analysis*. CRC Press, Boca Raton, FL. MR3444169

PELLETIER, B. (2006). Non-parametric regression estimation on closed Riemannian manifolds. *J. Nonparametr. Stat.* **18** 57–67. MR2214065 https://doi.org/10.1080/10485250500504828

PENNEC, X. (2018). Barycentric subspace analysis on manifolds. *Ann. Statist.* **46** 2711–2746. MR3851753 https://doi.org/10.1214/17-AOS1636

PENNEC, X., FILLARD, P. and AYACHE, N. (2006). A Riemannian framework for tensor computing. *Int. J. Comput. Vis.* **66** 41–66.

PETERSEN, A. and MÜLLER, H.-G. (2016). Fréchet integration and adaptive metric selection for interpretable covariances of multivariate functional data. *Biometrika* **103** 103–120. MR3465824 https://doi.org/10.1093/biomet/asv054

PETERSEN, A. and MÜLLER, H.-G. (2019). Fréchet regression for random objects with Euclidean predictors. *Ann. Statist.* **47** 691–719. MR3909947 https://doi.org/10.1214/17-AOS1624

PETRUNIN, A. and TUSCHMANN, W. (1999). Diffeomorphism finiteness, positive pinching, and second homotopy. *Geom. Funct. Anal.* **9** 736–774. MR1719602 https://doi.org/10.1007/s000390050101

RATHI, Y., TANNENBAUM, A. and MICHAILOVICH, O. (2007). Segmenting images on the tensor manifold. In 2007 *IEEE Conference on Computer Vision and Pattern Recognition* 1–8. IEEE, New York.

RUDIN, L. I., OSHER, S. and FATEMI, E. (1992). Nonlinear total variation based noise removal algorithms. *Phys. D, Nonlinear Phenom.* **60** 259–268. MR3363401 https://doi.org/10.1016/0167-2789(92)90242-F

SACKS, J. and YLVISAKER, D. (1970). Designs for regression problems with correlated errors. III. *Ann. Math. Stat.* **41** 2057–2074. MR0270530 https://doi.org/10.1214/aoms/1177696705

SADHANALA, V., WANG, Y.-X. and TIBSHIRANI, R. J. (2016). Total variation classes beyond 1d: Minimax rates, and the limitations of linear smoothers. In *Neural Information Processing Systems* 3521–3529.

SHI, X., STYNER, M., LIEBERMAN, J., IBRAHIM, J. G., LIN, W. and ZHU, H. (2009). Intrinsic regression models for manifold-valued data. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI* **12** 192–199.

STEINKE, F., HEIN, M. and SCHÖLKOPF, B. (2010). Nonparametric regression between general Riemannian manifolds. *SIAM J. Imaging Sci.* **3** 527–563. MR2736019 https://doi.org/10.1137/080744189

STRONG, D. and CHAN, T. (2003). Edge-preserving and scale-dependent properties of total variation regularization. *Inverse Probl.* **19** S165–S187. MR2036526 https://doi.org/10.1088/0266-5611/19/6/059

STURM, K.-T. (2003). Probability measures on metric spaces of nonpositive curvature. In *Heat Kernels and Analysis on Manifolds*, *Graphs*, *and Metric Spaces* (*Paris*, 2002). *Contemp*. *Math*. **338** 357–390. Amer. Math. Soc., Providence, RI. MR2039961 https://doi.org/10.1090/conm/338/06080

TIBSHIRANI, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *Ann*. *Statist*. **42** 285–323. MR3189487 https://doi.org/10.1214/13-AOS1189

TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **67** 91–108. MR2136641 https://doi.org/10.1111/j.1467-9868.2005.00490.x

VAN DE GEER, S. (1990). Estimating a regression function. *Ann*. *Statist*. **18** 907–924. MR1056343 https://doi.org/10.1214/aos/1176347632

VAN DE GEER, S. (2001). Least squares estimation with complexity penalties. *Math*. *Methods Statist*. **10** 355–374. MR1867165

WANG, X., ZHU, H. and ADNI (2017). Generalized scalar-on-image regression models via total variation. *J*. *Amer*. *Statist*. *Assoc*. **112** 1156–1168. MR3735367 https://doi.org/10.1080/01621459.2016.1194846

WANG, Y.-X., SHARPNACK, J., SMOLA, A. J. and TIBSHIRANI, R. J. (2016). Trend filtering on graphs. *J*. *Mach*. *Learn*. *Res*. **17** Paper No. 105, 41 pp. MR3543511

WEINMANN, A., DEMARET, L. and STORATH, M. (2014). Total variation regularization for manifold-valued data. *SIAM J*. *Imaging Sci*. **7** 2226–2257. MR3278838 https://doi.org/10.1137/130951075

YUAN, Y., ZHU, H., LIN, W. and MARRON, J. S. (2012). Local polynomial regression for symmetric positive definite matrices. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **74** 697–719. MR2965956 https://doi.org/10.1111/j.1467-9868.2011.01022.x

ZHOU, D., DRYDEN, I. L., KOLOYDENKO, A. A., AUDENAERT, K. M. R. and BAI, L. (2016). Regularisation, interpolation and visualisation of diffusion tensor images using non-Euclidean statistics. *J*. *Appl*. *Stat*. **43** 943–978. MR3457097 https://doi.org/10.1080/02664763.2015.1080671

ZHU, H., FAN, J. and KONG, L. (2014). Spatially varying coefficient model for neuroimaging data with jump discontinuities. *J*. *Amer*. *Statist*. *Assoc*. **109** 1084–1098. MR3265682 https://doi.org/10.1080/01621459.2014.881742

The Annals of Statistics
2021, Vol. 49, No. 6, 3534–3558
https://doi.org/10.1214/21-AOS2096

POSTERIOR ANALYSIS OF n IN THE BINOMIAL (n, p) PROBLEM WITH BOTH PARAMETERS UNKNOWN—WITH APPLICATIONS TO QUANTITATIVE NANOSCOPY

BY JOHANNES SCHMIDT-HIEBER[1], LAURA FEE SCHNEIDER[2,*], THOMAS STAUDT[2,4,†], ANDREA KRAJINA[2,‡], TIMO ASPELMEIER[2,§] AND AXEL MUNK[2,3,4,¶]

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

[2]*Institute for Mathematical Stochastics, University of Göttingen , [*]laura-fee.schneider@mathematik.uni-goettingen.de; [†]thomas.staudt@uni-goettingen.de; [‡]andrea.krajina@mathematik.uni-goettingen.de; [§]timo.aspelmeier@mathematik.uni-goettingen.de; [¶]munk@math.uni-goettingen.de*

[3]*Max Planck Institute for Biophysical Chemistry*

[4]*Cluster of Excellence "Multiscale Bioimaging: from Molecular Machines to Networks of Excitable Cells" (MBExC), University of Göttingen, University Medical Center*

Estimation of the population size n from k i.i.d. binomial observations with unknown success probability p is relevant to a multitude of applications and has a long history. Without additional prior information this is a notoriously difficult task when p becomes small, and the Bayesian approach becomes particularly useful. For a large class of priors, we establish posterior contraction and a Bernstein-von Mises type theorem in a setting where $p \to 0$ and $n \to \infty$ as $k \to \infty$. Furthermore, we suggest a new class of Bayesian estimators for n and provide a comprehensive simulation study in which we investigate their performance. To showcase the advantages of a Bayesian approach on real data, we also benchmark our estimators in a novel application from super-resolution microscopy.

REFERENCES

ASPELMEIER, T., EGNER, A. and MUNK, A. (2015). Modern statistical challenges in high-resolution fluorescence microscopy. *Annu. Rev. Stat. Appl.* **2** 163–202.

BASU, S. and EBRAHIMI, N. (2001). Bayesian capture-recapture methods for error detection and estimation of population size: Heterogeneity and dependence. *Biometrika* **88** 269–279. MR1841274 https://doi.org/10.1093/biomet/88.1.269

BERGER, J. O., BERNARDO, J. M. and SUN, D. (2012). Objective priors for discrete parameter spaces. *J. Amer. Statist. Assoc.* **107** 636–648. MR2980073 https://doi.org/10.1080/01621459.2012.682538

BETZIG, E., PATTERSON, G. H., SOUGRAT, R., LINDWASSER, O. W., OLENYCH, S., BONIFACINO, J. S., DAVIDSON, M. W., LIPPINCOTT-SCHWARTZ, J. and HESS, H. F. (2006). Imaging intracellular fluorescent proteins at nanometer resolution. *Science* **313** 1642–1645.

BLUMENTHAL, S. and DAHIYA, R. C. (1981). Estimating the binomial parameter n. *J. Amer. Statist. Assoc.* **76** 903–909. MR0650902

BOCHKINA, N. A. and GREEN, P. J. (2014). The Bernstein-von Mises theorem and nonregular models. *Ann. Statist.* **42** 1850–1878. MR3262470 https://doi.org/10.1214/14-AOS1239

BOUCHERON, S. and GASSIAT, E. (2009). A Bernstein-von Mises theorem for discrete probability distributions. *Electron. J. Stat.* **3** 114–148. MR2471588 https://doi.org/10.1214/08-EJS262

CARROLL, R. J. and LOMBARD, F. (1985). A note on N estimators for the binomial distribution. *J. Amer. Statist. Assoc.* **80** 423–426. MR0792743

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

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

MSC2020 subject classifications. Primary 62G05; secondary 62F15, 62F12, 62P10, 62P35.

Key words and phrases. Bayesian estimation, posterior contraction, Bernstein-von Mises theorem, binomial distribution, beta-binomial likelihood, quantitative cell imaging.

CHOIRAT, C. and SERI, R. (2012). Estimation in discrete parameter models. *Statist. Sci.* **27** 278–293. MR2963996 https://doi.org/10.1214/11-STS371

DASGUPTA, A. and RUBIN, H. (2005). Estimation of binomial parameters when both n, p are unknown. *J. Statist. Plann. Inference* **130** 391–404. MR2128016 https://doi.org/10.1016/j.jspi.2004.02.019

DEGROOT, M. H. and SCHERVISH, M. J. (2012). *Probability and Statistics*. Pearson Education, Upper Saddle River, NJ.

DRAPER, N. and GUTTMAN, I. (1971). Bayesian estimation of the binomial parameter. *Technometrics* **13** 667–673.

FISHER, R. A. (1941). The negative binomial distribution. *Annu. Eugen.* **11** 182–187. MR0006689

FÖLLING, J., BOSSI, M., BOCK, H., MEDDA, R., WURM, C. A., HEIN, B., JAKOBS, S., EGGELING, C. and HELL, S. W. (2008). Fluorescence nanoscopy by ground-state depletion and single-molecule return. *Nat. Methods* **5** 943–945.

GAO, C., VAN DER VAART, A. W. and ZHOU, H. H. (2020). A general framework for Bayes structured linear models. *Ann. Statist.* **48** 2848–2878. MR4152123 https://doi.org/10.1214/19-AOS1909

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

GHOSAL, S. and VAN DER VAART, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. *Cambridge Series in Statistical and Probabilistic Mathematics* **44**. Cambridge Univ. Press, Cambridge. MR3587782 https://doi.org/10.1017/9781139029834

GÜNEL, E. and CHILKO, D. (1989). Estimation of parameter n of the binomial distribution. *Comm. Statist. Simulation Comput.* **18** 537–551. MR1016224 https://doi.org/10.1080/03610918908812775

HALDANE, J. B. S. (1941). The fitting of binomial distributions. *Annu. Eugen.* **11** 179–181. MR0006688

HALL, P. (1994). On the erratic behavior of estimators of N in the binomial N, p distribution. *J. Amer. Statist. Assoc.* **89** 344–352. MR1266304

HAMEDANI, G. G. and WALTER, G. G. (1988). Bayes estimation of the binomial parameter n. *Comm. Statist. Theory Methods* **17** 1829–1843. MR0945788 https://doi.org/10.1080/03610928808829716

HELL, S. W. (2009). Microscopy and its focal switch. *Nat. Methods* **6** 24–32. https://doi.org/10.1038/nmeth.1291

HELL, S. W. (2015). Nobel lecture: Nanoscopy with freely propagating light. *Rev. Modern Phys.* **87** 1169–1181.

HESS, S. T., GIRIRAJAN, T. P. and MASON, M. D. (2006). Ultra-high resolution imaging by fluorescence photoactivation localization microscopy. *Biophys. J.* **91** 4258–72.

HOEL, P. G. (1947). Discriminating between binomial distributions. *Ann. Math. Stat.* **18** 556–564. MR0023035 https://doi.org/10.1214/aoms/1177730346

KAHN, W. D. (1987). A cautionary note for Bayesian estimation of the binomial parameter n. *Amer. Statist.* **41** 38–40. MR0882767 https://doi.org/10.2307/2684316

KARATHANASIS, C., FRICKE, F., HUMMER, G. and HELLEMANN, M. (2017). Molecule counts in localization microscopy with organic fluorophores. *ChemPhysChem* **18** 942–948.

KEMP, A. W. (1997). Characterizations of a discrete normal distribution. *J. Statist. Plann. Inference* **63** 223–229. MR1491581 https://doi.org/10.1016/S0378-3758(97)00020-7

LEE, S.-H., SHIN, J. Y., LEE, A. and BUSTAMANTE, C. (2012). Counting single photoavtiatable fluorescent molecules by photoactivated localization microscopy (palm). *Proc. Natl. Acad. Sci. USA* **109** 17436–17441.

LEHMANN, E. and CASELLA, G. (1996). *Theory of Point Estimation*, 2nd ed. Springer, Berlin.

LINK, W. A. (2013). A cautionary note on the discrete uniform prior for the binomial n. *Ecology* **94** 2173–2179.

OLKIN, I., PETKAU, A. J. and ZIDEK, J. V. (1980). A comparison of n estimators for the binomial distribution. *J. Amer. Statist. Assoc.* **76** 637–642.

OTIS, D. L., BURNHAM, K. P., WHITE, G. C. and ANDERSON, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildl. Monogr.* **64** 1–135.

RAFTERY, A. E. (1988). Inference for the binomial N parameter: A hierachical Bayes approach. *Biometrika* **75** 223–228.

ROLLINS, G. C., SHIN, J. Y., BUSTAMANTE, C. and PRESSÉ, S. (2015). Stochastic approach to the molecular counting problem in superresolution microscopy. *Proc. Natl. Acad. Sci. USA* **112** E110–8.

ROYLE, J. A. (2004). N-mixture models for estimating population size from spatially replicated counts. *Biometrics* **60** 108–115. MR2043625 https://doi.org/10.1111/j.0006-341X.2004.00142.x

RUST, M. J., BATES, M. and ZHUANG, X. (2006). Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nat. Methods* **3** 793–796.

SCHMIDT-HIEBER, J., SCHNEIDER, L. F., STAUDT, T., KRAJINA, A., ASPELMEIER, T. and MUNK, A. (2021). Supplement to "Posterior analysis of n in the binomial (n, p) problem with both parameters unknown—with applications to quantitative nanoscopy." https://doi.org/10.1214/21-AOS2096SUPPA, https://doi.org/10.1214/21-AOS2096SUPPB

SCHMIED, J. J., RAAB, M., FORTHMANN, C., PIBIRI, E., WÜNSCH, B., DAMMEYER, T. and TINNEFELD, P. (2014). Dna origami-based standards for quantitative fluoresence microscopy. *Nat. Protoc.* **9** 1367–1391.

SCHNEIDER, L. F., STAUDT, T. and MUNK, A. (2018). Posterior consistency in the binomial model with unknown parameters: A numerical study. In *International Conference on Bayesian Statistics in Action* 35–42. Springer, Berlin.

SCHWARTZ, L. (1965). On Bayes procedures. *Z. Wahrsch. Verw. Gebiete* **4** 10–26. MR0184378 https://doi.org/10.1007/BF00535479

SMITH, P. J. (1988). Bayesian methods for multiple capture-recapture surveys. *Biometrics* **44** 1177–1189.

STAUDT, T., ASPELMEIER, T., LAITENBERGER, O., GEISLER, C., EGNER, A. and MUNK, A. (2020). Statistical molecule counting in super-resolution fluorescence microscopy: Towards quantitative nanoscopy. *Statist. Sci.* **35** 92–111. MR4071360 https://doi.org/10.1214/19-STS753

STUDENT (1919). An explanation of deviations from Poisson's law in practice. *Biometrika* **12** 211–215.

SZABŁOWSKI, P. (2001). Discrete normal distribution and its relationship with Jacobi theta functions. *Statist. Probab. Lett.* **52** 289–299. MR1838217 https://doi.org/10.1016/S0167-7152(00)00223-6

TA, H., KELLER, J., HALTMEIER, M., SAKA, S. K., SCHMIED, J., OPAZO, F., TINNEFELD, P., MUNK, A. and HELL, S. W. (2015). Mapping molecules in scanning far-field fluorescence nanoscopy. *Nat. Commun.* **6** 1–7.

TANCREDI, A., STEORTS, R. and LISEO, B. (2020). A unified framework for de-duplication and population size estimation (with discussion). *Bayesian Anal.* **15** 633–682. MR4122517 https://doi.org/10.1214/19-BA1146

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

VILLA, C. and WALKER, S. G. (2014). A cautionary note on the discrete uniform prior for the binomial n: Comment. *Ecology* **95** 2674–2677.

VILLA, C. and WALKER, S. G. (2015). An objective approach to prior mass functions for discrete parameter spaces. *J. Amer. Statist. Assoc.* **110** 1072–1082. MR3420685 https://doi.org/10.1080/01621459.2014.946319

WANG, X., HE, C. Z. and SUN, D. (2007). Bayesian population estimation for small sample capture-recapture data using noninformative priors. *J. Statist. Plann. Inference* **137** 1099–1118. MR2301466 https://doi.org/10.1016/j.jspi.2006.03.004

The Annals of Statistics
2021, Vol. 49, No. 6, 3559–3582
https://doi.org/10.1214/21-AOS2097

MULTISCALE BAYESIAN SURVIVAL ANALYSIS

BY ISMAËL CASTILLO[1] AND STÉPHANIE VAN DER PAS[2]

[1]*Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université & Institut Universitaire de France,* *ismael.castillo@upmc.fr*

[2]*Department of Epidemiology and Data Science, Amsterdam UMC, Vrije Universiteit Amsterdam,* *s.l.vanderpas@amsterdamumc.nl*

We consider Bayesian nonparametric inference in the right-censoring survival model, where modeling is made at the level of the hazard rate. We derive posterior limiting distributions for linear functionals of the hazard, and then for 'many' functionals simultaneously in appropriate multiscale spaces. As an application, we derive Bernstein–von Mises theorems for the cumulative hazard and survival functions, which lead to asymptotically efficient confidence bands for these quantities. Further, we show optimal posterior contraction rates for the hazard in terms of the supremum norm. In medical studies, a popular approach is to model hazards a priori as random histograms with possibly dependent heights. This and more general classes of arbitrarily smooth prior distributions are considered as applications of our theory. A sampler is provided for possibly dependent histogram posteriors. Its finite sample properties are investigated on both simulated and real data experiments.

REFERENCES

[1] AALEN, O. O. (1975). Statistical Inference for a Family of Counting Processes. PhD thesis, Univ. California, Berkeley, CA.

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

[3] ARJAS, E. and GASBARRA, D. (1994). Nonparametric Bayesian inference from right censored survival data, using the Gibbs sampler. *Statist*. *Sinica* **4** 505–524. MR1309427

[4] ASLANIDOU, H., DEY, D. K. and SINHA, D. (1998). Bayesian analysis of multivariate survival data using Monte Carlo methods. *Canad*. *J*. *Statist*. **26** 33–48. MR1624365 https://doi.org/10.2307/3315671

[5] BERRY, S. M., BERRY, D. A., NATARAJAN, K., LIN, C.-S., HENNEKENS, C. H. and BELDER, R. (2004). Bayesian survival analysis with nonproportional hazards: Metanalysis of combination pravastatin-aspirin. *J*. *Amer. Statist*. *Assoc*. **99** 36–44. MR2061886 https://doi.org/10.1198/016214504000000052

[6] BORGAN, Ø. and LIESTØL, K. (1990). A note on confidence intervals and bands for the survival function based on transformations. *Scand*. *J*. *Stat*. **17** 35–41. MR1062844

[7] BOU-HAMAD, I., LAROCQUE, D. and BEN-AMEUR, H. (2011). A review of survival trees. *Stat*. *Surv.* **5** 44–71. MR3018509 https://doi.org/10.1214/09-SS047

[8] CASTILLO, I. (2012). A semiparametric Bernstein–von Mises theorem for Gaussian process priors. *Probab. Theory Related Fields* **152** 53–99. MR2875753 https://doi.org/10.1007/s00440-010-0316-5

[9] CASTILLO, I. (2014). On Bayesian supremum norm contraction rates. *Ann*. *Statist*. **42** 2058–2091. MR3262477 https://doi.org/10.1214/14-AOS1253

[10] 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

[11] 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

[12] CASTILLO, I. and ROUSSEAU, J. (2015). A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *Ann*. *Statist*. **43** 2353–2383. MR3405597 https://doi.org/10.1214/15-AOS1336

MSC2020 subject classifications. 62G20, 62G15.

Key words and phrases. Frequentist analysis of Bayesian procedures, survival analysis, nonparametric Bernstein–von Mises theorem, supremum norm contraction rate.

[13] CASTILLO, I. and ROUSSEAU, J. (2015). Supplement to "A Bernstein–von Mises theorem for smooth functionals in semiparametric models." *Ann. Statist.* **43** 2353–2383. MR3405597 https://doi.org/10.1214/15-AOS1336

[14] CASTILLO, I. and VAN DER PAS, S. (2021). Supplement to "Multiscale Bayesian survival analysis." https://doi.org/10.1214/21-AOS2097SUPP

[15] 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

[16] DE BLASI, P. and HJORT, N. L. (2009). The Bernstein-von Mises theorem in semiparametric competing risks models. *J. Statist. Plann. Inference* **139** 2316–2328. MR2507993 https://doi.org/10.1016/j.jspi.2008.10.018

[17] DE BLASI, P., PECCATI, G. and PRÜNSTER, I. (2009). Asymptotics for posterior hazards. *Ann. Statist.* **37** 1906–1945. MR2533475 https://doi.org/10.1214/08-AOS631

[18] DONNET, S., RIVOIRARD, V., ROUSSEAU, J. and SCRICCIOLO, C. (2017). Posterior concentration rates for counting processes with Aalen multiplicative intensities. *Bayesian Anal.* **12** 53–87. MR3597567 https://doi.org/10.1214/15-BA986

[19] FAHRMEIR, L. and LANG, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **50** 201–220. MR1833273 https://doi.org/10.1111/1467-9876.00229

[20] FRIEDMAN, M. (1982). Piecewise exponential models for survival data with covariates. *Ann. Statist.* **10** 101–113. MR0642722

[21] GAMERMAN, D. (1991). Dynamic Bayesian models for survival data. *J. R. Stat. Soc. Ser. C. Appl. Stat.* 63–79.

[22] 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

[23] GHOSAL, S. and VAN DER VAART, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. *Cambridge Series in Statistical and Probabilistic Mathematics* **44**. Cambridge Univ. Press, Cambridge. MR3587782 https://doi.org/10.1017/9781139029834

[24] HALL, W. J. and WELLNER, J. A. (1980). Confidence bands for a survival curve from censored data. *Biometrika* **67** 133–143. MR0570515 https://doi.org/10.1093/biomet/67.1.133

[25] HJORT, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.* **18** 1259–1294. MR1062708 https://doi.org/10.1214/aos/1176347749

[26] HOLFORD, T. R. (1980). The analysis of rates and of survivorship using log-linear models. *Biometrics* **36** 299–305.

[27] IBRAHIM, J. G., CHEN, M.-H. and SINHA, D. (2013). *Bayesian Survival Analysis*. Springer, New York.

[28] KIM, Y. (2006). The Bernstein-von Mises theorem for the proportional hazard model. *Ann. Statist.* **34** 1678–1700. MR2283713 https://doi.org/10.1214/009053606000000533

[29] KIM, Y. and LEE, J. (2001). On posterior consistency of survival models. *Ann. Statist.* **29** 666–686. MR1865336 https://doi.org/10.1214/aos/1009210685

[30] KIM, Y. and LEE, J. (2004). A Bernstein-von Mises theorem in the nonparametric right-censoring model. *Ann. Statist.* **32** 1492–1512. MR2089131 https://doi.org/10.1214/009053604000000526

[31] KLEIN, J. P., VAN HOUWELINGEN, H. C., IBRAHIM, J. G. and SCHEIKE, T. H., eds. (2016). *Handbook of Survival Analysis*. *Chapman & Hall/CRC Handbooks of Modern Statistical Methods*. CRC Press, Boca Raton, FL.

[32] LAIRD, N. and OLIVIER, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *J. Amer. Statist. Assoc.* **76** 231–240. MR0624329

[33] LI, W. and GHOSAL, S. (2021). Posterior contraction and credible regions for level sets. *Electron. J. Stat.* **15** 2647–2689. MR4280159 https://doi.org/10.1214/21-ejs1846

[34] LOPRINZI, C. L., LAURIE, J. A., WIEAND, H. S., KROOK, J. E., NOVOTNY, P. J., KUGLER, J. W., BARTEL, J., LAW, M., BATEMAN, M. et al. (1994). Prospective evaluation of prognostic variables from patient-completed questionnaires. North central cancer treatment group. *J. Clin. Oncol.* **12** 601–607.

[35] NAIR, V. N. (1984). Confidence bands for survival functions with censored data: A comparative study. *Technometrics* **26** 265–275.

[36] NELSON, W. (1969). Hazard plotting for incomplete failure data. *J. Qual. Technol.* **1** 27–52.

[37] NICKL, R. (2020). Bernstein–von Mises theorems for statistical inverse problems I: Schrödinger equation. *J. Eur. Math. Soc.* (*JEMS*) **22** 2697–2750. MR4118619 https://doi.org/10.4171/JEMS/975

[38] NICKL, R. and RAY, K. (2020). Nonparametric statistical inference for drift vector fields of multi-dimensional diffusions. *Ann. Statist.* **48** 1383–1408. MR4124327 https://doi.org/10.1214/19-AOS1851

[39] NICKL, R. and SÖHL, J. (2019). Bernstein-von Mises theorems for statistical inverse problems II: Compound Poisson processes. *Electron. J. Stat.* **13** 3513–3571. MR4013745 https://doi.org/10.1214/19-ejs1609

[40] SAHU, S. K., DEY, D. K., ASLANIDOU, H. and SINHA, D. (1997). A Weibull regression model with gamma frailties for multivariate survival data. *Lifetime Data Anal.* **3** 123–137.

[41] SINHA, D. (1993). Semiparametric Bayesian analysis of multiple event time data. *J. Amer. Statist. Assoc.* **88** 979–983.

[42] VAN DER PAS, S. and CASTILLO, I. (2020). BayesSurvival: Bayesian survival analysis for right censored data. R package version 0.1.0.

[43] 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

[44] YOUNES, N. and LACHIN, J. (1997). Link-based models for survival data with interval and continuous time censoring. *Biometrics* **53** 1199–1211. MR1614362 https://doi.org/10.2307/2533490

The Annals of Statistics
2021, Vol. 49, No. 6, 3583–3617
https://doi.org/10.1214/21-AOS2098

ARE DEVIATIONS IN A GRADUALLY VARYING MEAN RELEVANT? A TESTING APPROACH BASED ON SUP-NORM ESTIMATORS

BY AXEL BÜCHER[1], HOLGER DETTE[2,*] AND FLORIAN HEINRICHS[2,†]

[1]*Mathematisches Institut, Heinrich-Heine-Universität Düsseldorf, axel.buecher@hhu.de*
[2]*Fakultät für Mathematik, Ruhr-Universität Bochum, *holger.dette@rub.de; †florian.heinrichs@rub.de*

Classical change point analysis aims at (1) detecting abrupt changes in the mean of a possibly nonstationary time series and at (2) identifying regions where the mean exhibits a piecewise constant behavior. In many applications however, it is more reasonable to assume that the mean changes gradually in a smooth way. Those gradual changes may either be nonrelevant (i.e., small), or relevant for a specific problem at hand, and the present paper presents statistical methodology to detect the latter. More precisely, we consider the common nonparametric regression model $X_i = \mu(i/n) + \varepsilon_i$ with centered errors and propose a test for the null hypothesis that the maximum absolute deviation of the regression function μ from a functional $g(\mu)$ (such as the value $\mu(0)$ or the integral $\int_0^1 \mu(t)\,\mathrm{d}t$) is smaller than a given threshold on a given interval $[x_0, x_1] \subseteq [0, 1]$. A test for this type of hypotheses is developed using an appropriate estimator, say $\hat{d}_{\infty,n}$, for the maximum deviation $d_\infty = \sup_{t\in[x_0,x_1]} |\mu(t) - g(\mu)|$. We derive the limiting distribution of an appropriately standardized version of $\hat{d}_{\infty,n}$, where the standardization depends on the Lebesgue measure of the set of extremal points of the function $\mu(\cdot) - g(\mu)$. A refined procedure based on an estimate of this set is developed and its consistency is proved. The results are illustrated by means of a simulation study and a data example.

REFERENCES

AITCHISON, J. (1964). Confidence-region tests. *J. Roy. Statist. Soc. Ser. B* **26** 462–476. MR0174119

ASTON, J. A. D. and KIRCH, C. (2012). Evaluating stationarity via change-point alternatives with applications to fMRI data. *Ann. Appl. Stat.* **6** 1906–1948. MR3058688 https://doi.org/10.1214/12-AOAS565

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

BARANOWSKI, R., CHEN, Y. and FRYZLEWICZ, P. (2019). Narrowest-over-threshold detection of multiple change points and change-point-like features. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 649–672. MR3961502 https://doi.org/10.1111/rssb.12322

BÜCHER, A., DETTE, H. and HEINRICHS, F. (2021). Supplement to "Are deviations in a gradually varying mean relevant? A testing approach based on sup-norm estimators." https://doi.org/10.1214/21-AOS2098SUPP

CAO, H., LIU, W. and ZHOU, Z. (2018). Simultaneous nonparametric regression analysis of sparse longitudinal data. *Bernoulli* **24** 3013–3038. MR3779709 https://doi.org/10.3150/17-BEJ952

CHAKRABORTI, S. and GRAHAM, M. A. (2019). Nonparametric (distribution-free) control charts: An updated overview and some results. *Qual. Eng.* **31** 523–544. https://doi.org/10.1080/08982112.2018.1549330

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

ČÍŽEK, P. and SADIKOĞLU, S. (2020). Robust nonparametric regression: A review. *Wiley Interdiscip. Rev.: Comput. Stat.* **12** e1492, 16 pp. MR4098476 https://doi.org/10.1002/wics.1492

COLLINS, D., DELLA-MARTA, P., PLUMMER, N. and TREWIN, B. (2000). Trends in annual frequencies of extreme temperature events in Australia. *Aust. Meteorol. Mag.* **49** 277–292.

MSC2020 subject classifications. 62M10, 62G08.
Key words and phrases. Relevant change point analysis, gradual changes, maximum deviation, local-linear estimator, Gumbel distribution, Gaussian approximation.

DETTE, H., ECKLE, T. and VETTER, M. (2020). Multiscale change point detection for dependent data. *Scand. J. Stat*. **47** 1243–1274. MR4178193 https://doi.org/10.1111/sjos.12465

DETTE, H. and WU, W. (2019). Detecting relevant changes in the mean of nonstationary processes—A mass excess approach. *Ann. Statist*. **47** 3578–3608. MR4025752 https://doi.org/10.1214/19-AOS1811

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

FRICK, K., MUNK, A. and SIELING, H. (2014). Multiscale change point inference. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **76** 495–580. MR3210728 https://doi.org/10.1111/rssb.12047

FRYZLEWICZ, P. (2018). Tail-greedy bottom-up data decompositions and fast multiple change-point detection. *Ann. Statist*. **46** 3390–3421. MR3852656 https://doi.org/10.1214/17-AOS1662

GAO, J., GIJBELS, I. and VAN BELLEGEM, S. (2008). Nonparametric simultaneous testing for structural breaks. *J. Econometrics* **143** 123–142. MR2384436 https://doi.org/10.1016/j.jeconom.2007.08.009

GAO, Z., SHANG, Z., DU, P. and ROBERTSON, J. L. (2019). Variance change point detection under a smoothly-changing mean trend with application to liver procurement. *J. Amer. Statist. Assoc*. **114** 773–781. MR3963179 https://doi.org/10.1080/01621459.2018.1442341

GIJBELS, I. and GODERNIAUX, A.-C. (2004). Bandwidth selection for changepoint estimation in nonparametric regression. *Technometrics* **46** 76–86. MR2043389 https://doi.org/10.1198/004017004000000130

GIJBELS, I., LAMBERT, A. and QIU, P. (2007). Jump-preserving regression and smoothing using local linear fitting: A compromise. *Ann. Inst. Statist. Math*. **59** 235–272. MR2394168 https://doi.org/10.1007/s10463-006-0045-9

HANSEN, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory* **24** 726–748. MR2409261 https://doi.org/10.1017/S0266466608080304

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

HORVÁTH, L., KOKOSZKA, P. and STEINEBACH, J. (1999). Testing for changes in multivariate dependent observations with an application to temperature changes. *J. Multivariate Anal*. **68** 96–119. MR1668911 https://doi.org/10.1006/jmva.1998.1780

HOTZ, T., SCHÜTTE, O. M., SIELING, H., POLUPANOW, T., DIEDERICHSEN, U., STEINEM, C. and MUNK, A. (2013). Idealizing ion channel recordings by jump segmentation and statistical multiresolution analysis. *IEEE Trans. Nanobiosci*. **12** 376–386.

JANDHYALA, V., FOTOPOULOS, S., MACNEILL, I. and LIU, P. (2013). Inference for single and multiple change-points in time series. *J. Time Series Anal*. **34** 423–446. MR3070866 https://doi.org/10.1111/jtsa.12035

JOHNSTON, G. J. (1982). Probabilities of maximal deviations for nonparametric regression function estimates. *J. Multivariate Anal*. **12** 402–414. MR0666014 https://doi.org/10.1016/0047-259X(82)90074-4

KARL, T. R., KNIGHT, R. W. and PLUMMER, N. (1995). Trends in high-frequency climate variability in the twentieth century. *Nature* **377** 217.

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

LI, D., LIU, W., WANG, Q. and WU, W. B. (2017). Simultaneous confidence bands in nonlinear regression models with nonstationarity. *Statist. Sinica* **27** 1385–1400. MR3699710

LOADER, C. R. (1996). Change point estimation using nonparametric regression. *Ann. Statist*. **24** 1667–1678. MR1416655 https://doi.org/10.1214/aos/1032298290

MUDELSEE, M. (2000). Ramp function regression: A tool for quantifying climate transitions. *Comput. Geosci*. **26** 293–307. https://doi.org/10.1016/S0098-3004(99)00141-7

MÜLLER, H.-G. (1992). Change-points in nonparametric regression analysis. *Ann. Statist*. **20** 737–761. MR1165590 https://doi.org/10.1214/aos/1176348654

PITERBARG, V. I. (2012). *Asymptotic Methods in the Theory of Gaussian Processes and Fields*. *Translations of Mathematical Monographs* **148**. Amer. Math. Soc., Providence, RI. MR1361884 https://doi.org/10.1090/mmono/148

PRIESTLEY, M. B. and SUBBA RAO, T. (1969). A test for non-stationarity of time-series. *J. Roy. Statist. Soc. Ser. B* **31** 140–149. MR0269062

PROKSCH, K. (2016). On confidence bands for multivariate nonparametric regression. *Ann. Inst. Statist. Math*. **68** 209–236. MR3440220 https://doi.org/10.1007/s10463-014-0494-5

QIU, P. (2005). *Image Processing and Jump Regression Analysis*. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR2111430 https://doi.org/10.1002/0471733156

SCHUCANY, W. R. and SOMMERS, J. P. (1977). Improvement of kernel type density estimators. *J. Amer. Statist. Assoc*. **72** 420–423. MR0448691

SHARMA, S., SWAYNE, D. A. and OBIMBO, C. (2016). Trend analysis and change point techniques: A survey. *Energy Ecol. Environ.* **1** 123–130. https://doi.org/10.1007/s40974-016-0011-1

SU, L. and XIAO, Z. (2008). Testing structural change in time-series nonparametric regression models. *Stat. Interface* **1** 347–366. MR2476751 https://doi.org/10.4310/SII.2008.v1.n2.a12

TRUONG, C., OUDRE, L. and VAYATIS, N. (2020). Selective review of offline change point detection methods. *Signal Process.* **167** 107299.

VAN DER VAART, A. and WELLNER, J. (1996). *Weak Convergence and Empirical Processes. Springer Series in Statistics* **1**. Springer Science+Business Media, New York.

VOGT, M. and DETTE, H. (2015). Detecting gradual changes in locally stationary processes. *Ann. Statist.* **43** 713–740. MR3319141 https://doi.org/10.1214/14-AOS1297

WELLEK, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority*, 2nd ed. CRC Press, Boca Raton, FL. MR2676002 https://doi.org/10.1201/EBK1439808184

WOLFE, D. A. and SCHECHTMAN, E. (1984). Nonparametric statistical procedures for the changepoint problem. *J. Statist. Plann. Inference* **9** 389–396. MR0748267 https://doi.org/10.1016/0378-3758(84)90013-2

WOODALL, W. H. and MONTGOMERY, D. C. (2014). Some current directions in the theory and application of statistical process monitoring. *J. Qual. Technol.* **46** 78–94. https://doi.org/10.1080/00224065.2014.11917955

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

WU, J. S. and CHU, C. K. (1993). Kernel-type estimators of jump points and values of a regression function. *Ann. Statist.* **21** 1545–1566. MR1241278 https://doi.org/10.1214/aos/1176349271

WU, W. B. and ZHAO, Z. (2007). Inference of trends in time series. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 391–410. MR2323759 https://doi.org/10.1111/j.1467-9868.2007.00594.x

XIA, Y. (1998). Bias-corrected confidence bands in nonparametric regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 797–811. MR1649488 https://doi.org/10.1111/1467-9868.00155

ZHAO, Z. and WU, W. B. (2008). Confidence bands in nonparametric time series regression. *Ann. Statist.* **36** 1854–1878. MR2435458 https://doi.org/10.1214/07-AOS533

ZHOU, Z. (2010). Nonparametric inference of quantile curves for nonstationary time series. *Ann. Statist.* **38** 2187–2217. MR2676887 https://doi.org/10.1214/09-AOS769

ZHOU, Z. (2013). Heteroscedasticity and autocorrelation robust structural change detection. *J. Amer. Statist. Assoc.* **108** 726–740. MR3174655 https://doi.org/10.1080/01621459.2013.787184

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

ZOU, C., QIU, P. and HAWKINS, D. (2009). Nonparametric control chart for monitoring profiles using change point formulation and adaptive smoothing. *Statist. Sinica* **19** 1337–1357. MR2536159

ZOU, C., TSUNG, F. and WANG, Z. (2008). Monitoring profiles based on nonparametric regression methods. *Technometrics* **50** 512–526. MR2477862 https://doi.org/10.1198/004017008000000433

The Annals of Statistics
2021, Vol. 49, No. 6, 3618–3649
https://doi.org/10.1214/21-AOS2102

ADAPTIVE TRANSFER LEARNING

BY HENRY W. J. REEVE[1], TIMOTHY I. CANNINGS[2] AND RICHARD J. SAMWORTH[3]

[1]*School of Mathematics, University of Bristol, henry.reeve@bristol.ac.uk*
[2]*School of Mathematics, University of Edinburgh, timothy.cannings@ed.ac.uk*
[3]*Statistical Laboratory, University of Cambridge, r.samworth@statslab.cam.ac.uk*

In transfer learning, we wish to make inference about a target population when we have access to data both from the distribution itself, and from a different but related source distribution. We introduce a flexible framework for transfer learning in the context of binary classification, allowing for covariate-dependent relationships between the source and target distributions that are not required to preserve the Bayes decision boundary. Our main contributions are to derive the minimax optimal rates of convergence (up to polylogarithmic factors) in this problem, and show that the optimal rate can be achieved by an algorithm that adapts to key aspects of the unknown transfer relationship, as well as the smoothness and tail parameters of our distributional classes. This optimal rate turns out to have several regimes, depending on the interplay between the relative sample sizes and the strength of the transfer relationship, and our algorithm achieves optimality by careful, decision tree-based calibration of local nearest-neighbour procedures.

REFERENCES

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

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

BEN-DAVID, S., LU, T., LUU, T. and PÁL, D. (2010b). Impossibility theorems for domain adaptation. In *International Conference on Artificial Intelligence and Statistics* 129–136.

BERNSTEIN, S. (1924). On a modification of Chebyshev's inequality and of the error formula of Laplace. *Ann. Sci. Inst. Sav. Ukraine*, *Sect. Math* **1** 38–49.

BIAU, G. and DEVROYE, L. (2013). Cellular tree classifiers. *Electron. J. Stat.* **7** 1875–1912. MR3084675 https://doi.org/10.1214/13-EJS829

BIAU, G. and DEVROYE, L. (2015). *Lectures on the Nearest Neighbor Method. Springer Series in the Data Sciences*. Springer, Cham. MR3445317 https://doi.org/10.1007/978-3-319-25388-6

BLANCHARD, G., FLASKA, M., HANDY, G., POZZI, S. and SCOTT, C. (2016). Classification with asymmetric label noise: Consistency and maximal denoising. *Electron. J. Stat.* **10** 2780–2824. MR3549019 https://doi.org/10.1214/16-EJS1193

BLANCHARD, G., DESHMUKH, A. A., DOGAN, U., LEE, G. and SCOTT, C. (2017). Domain generalization by marginal transfer learning. Preprint. Available at arXiv:1711.07910.

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

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

CANDELA, J. Q., SUGIYAMA, M., SCHWAIGHOFER, A. and LAWRENCE, N. D. (2009). *Dataset Shift in Machine Learning*. MIT Press, Cambridge.

CANNINGS, T. I., BERRETT, T. B. and SAMWORTH, R. J. (2020). Local nearest neighbour classification with applications to semi-supervised learning. *Ann. Statist.* **48** 1789–1814. MR4124344 https://doi.org/10.1214/19-AOS1868

MSC2020 subject classifications. Primary 62G99; Secondary 62C20, 62G20.
Key words and phrases. Transfer learning, classification, decision trees, nearest neighbours, nonparametric, minimax.

CANNINGS, T. I., FAN, Y. and SAMWORTH, R. J. (2020). Classification with imperfect training labels. *Biometrika* **107** 311–330. MR4108933 https://doi.org/10.1093/biomet/asaa011

CARUANA, R. (1997). Multitask learning. *Mach. Learn.* **28** 41–75.

CHAUDHURI, K. and DASGUPTA, S. (2014). Rates of convergence for nearest neighbor classification. In *Advances in Neural Information Processing Systems* 3437–3445.

CHRISTIANSEN, R., PFISTER, N., JAKOBSEN, M. E., GNECCO, N. and PETERS, J. (2020). The difficult task of distribution generalization in nonlinear models. Preprint. Available at arXiv:2006.07433.

CORTES, C., MOHRI, M. and MUÑOZ MEDINA, A. (2019). Adaptation based on generalized discrepancy. *J. Mach. Learn. Res.* **20** 1–30. MR3911408

COVER, T. and HART, P. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **13** 21–27.

EVANS, L. C. and GARIEPY, R. F. (2015). *Measure Theory and Fine Properties of Functions*, Revised ed. *Textbooks in Mathematics*. CRC Press, Boca Raton, FL. MR3409135

FIX, E. and HODGES, J. L. (1951). *Discriminatory Analysis: Nonparametric Discrimination, Consistency Properties*. USAF School of Aviation Medicine.

FRÉNAY, B. and VERLEYSEN, M. (2014). Classification in the presence of label noise: A survey. *IEEE Trans. Neural Netw. Learn. Syst.* **25** 845–869.

GADAT, S., KLEIN, T. and MARTEAU, C. (2016). Classification in general finite dimensional spaces with the k-nearest neighbor rule. *Ann. Statist.* **44** 982–1009. MR3485951 https://doi.org/10.1214/15-AOS1395

GERMAIN, P., LACASSE, A., LAVIOLETTE, F., MARCHAND, M. and ROY, J.-F. (2015). Risk bounds for the majority vote: From a PAC-Bayesian analysis to a learning algorithm. *J. Mach. Learn. Res.* **16** 787–860. MR3361305

GRETTON, A., SMOLA, A., HUANG, J., SCHMITTFULL, M., BORGWARDT, K. and SCHÖLKOPF, B. (2009). Covariate shift by kernel mean matching. *Dataset Shift in Machine Learning* **3** 5.

HALL, P., PARK, B. U. and SAMWORTH, R. J. (2008). Choice of neighbor order in nearest-neighbor classification. *Ann. Statist.* **36** 2135–2152. MR2458182 https://doi.org/10.1214/07-AOS537

HANNEKE, S. and KPOTUFE, S. (2019). On the value of target data in transfer learning. In *Advances in Neural Information Processing Systems* 9867–9877.

HARDY, G. H. and LITTLEWOOD, J. E. (1930). A maximal theorem with function-theoretic applications. *Acta Math.* **54** 81–116. MR1555303 https://doi.org/10.1007/BF02547518

KIM, A. K. H. (2020). Obtaining minimax lower bounds: A review. *J. Korean Statist. Soc.* **49** 673–701. MR4147680 https://doi.org/10.1007/s42952-019-00027-7

KPOTUFE, S. and MARTINET, G. (2018). Marginal singularity, and the benefits of labels in covariate-shift. In *Proceedings of the* 31*st Conference on Learning Theory* **75** 1882–1886.

KULKARNI, S. R. and POSNER, S. E. (1995). Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Trans. Inf. Theory* **41** 1028–1039. MR1366756 https://doi.org/10.1109/18.391248

LEDRAPPIER, F. and YOUNG, L.-S. (1985). The metric entropy of diffeomorphisms. I. Characterization of measures satisfying Pesin's entropy formula. *Ann. of Math.* (2) **122** 509–539. MR0819556 https://doi.org/10.2307/1971328

LIPTON, Z. C., WANG, Y.-X. and SMOLA, A. (2018). Detecting and correcting for label shift with black box predictors. In *Proceedings of the* 35*th International Conference on Machine Learning* **80**.

MAITY, S., SUN, Y. and BANERJEE, M. (2020). Minimax optimal approaches to the label shift problem. Preprint. Available at arXiv:2003.10443.

MAMMEN, E. and TSYBAKOV, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.* **27** 1808–1829. MR1765618 https://doi.org/10.1214/aos/1017939240

MANSOUR, Y., MOHRI, M. and ROSTAMIZADEH, A. (2009). Domain adaptation: Learning bounds and algorithms. In *Proceedings of the* 22*nd Annual Conference on Learning Theory*.

MAURER, A., PONTIL, M. and ROMERA-PAREDES, B. (2016). The benefit of multitask representation learning. *J. Mach. Learn. Res.* **17** Paper No. 81. MR3517104

MCDIARMID, C. (1998). Concentration. In *Probabilistic Methods for Algorithmic Discrete Mathematics*. *Algorithms Combin.* **16** 195–248. Springer, Berlin. MR1678578 https://doi.org/10.1007/978-3-662-12788-9_6

MCSHANE, E. J. (1934). Extension of range of functions. *Bull. Amer. Math. Soc.* **40** 837–842. MR1562984 https://doi.org/10.1090/S0002-9904-1934-05978-0

MENON, A. K., VAN ROOYEN, B. and NATARAJAN, N. (2018). Learning from binary labels with instance-dependent noise. *Mach. Learn.* **107** 1561–1595. MR3835279 https://doi.org/10.1007/s10994-018-5715-3

MOHRI, M. and MUÑOZ MEDINA, A. (2012). New analysis and algorithm for learning with drifting distributions. In *Algorithmic Learning Theory*. *Lecture Notes in Computer Science* **7568** 124–138. Springer, Heidelberg. MR3042886 https://doi.org/10.1007/978-3-642-34106-9_13

PAN, S. J. and YANG, Q. (2009). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22** 1345–1359.

POLONIK, W. (1995). Measuring mass concentrations and estimating density contour clusters—an excess mass approach. *Ann. Statist.* **23** 855–881. MR1345204 https://doi.org/10.1214/aos/1176324626

REEVE, H. W. J. and BROWN, G. (2017). Minimax rates for cost-sensitive learning on manifolds with approximate nearest neighbours. In *International Conference on Algorithmic Learning Theory*. *Proc. Mach. Learn. Res.* (*PMLR*) **76** 11–56. Proceedings of Machine Learning Research PMLR. MR3776687

REEVE, H. W., CANNINGS, T. I. and SAMWORTH, R. J. (2021). Supplement to "Adaptive transfer learning." https://doi.org/10.1214/21-AOS2102SUPP

REEVE, H. W. J. and KABÁN, A. (2019a). Classification with unknown class-conditional label noise on non-compact feature spaces. In *Proceedings of the* 32*nd Annual Conference on Learning Theory* **99** 1–28.

REEVE, H. W. J. and KABÁN, A. (2019b). Fast rates for a kNN classifier robust to unknown asymmetric label noise. In *Proceedings of the* 36*th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.) **97** 5401–5409.

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

SCOTT, C. (2019). A generalized Neyman–Pearson criterion for optimal domain adaption. In *Algorithmic Learning Theory* 2019. *Proc. Mach. Learn. Res.* (*PMLR*) **98** 1–24. Proceedings of Machine Learning Research PMLR. MR3932867

SCOTT, C. and NOWAK, R. D. (2006). Minimax-optimal classification with dyadic decision trees. *IEEE Trans. Inf. Theory* **52** 1335–1353. MR2241192 https://doi.org/10.1109/TIT.2006.871056

SCOTT, C. and ZHANG, J. (2019). Learning from multiple corrupted sources, with application to learning from label proportions. Preprint. Available at arXiv:1910.04665.

SINHA, A., NAMKOONG, H. and DUCHI, J. C. (2018). Certifying some distributional robustness with principled adversarial training. In *Sixth International Conference on Learning Representations* (*ICLR* 2018).

STONE, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595–645. MR0443204

STORKEY, A. (2009). When training and test sets are different: Characterizing learning transfer. In *Dataset Shift in Machine Learning* 3–28. MIT Press, Cambridge.

SUGIYAMA, M., SUZUKI, T. and KANAMORI, T. (2012). *Density Ratio Estimation in Machine Learning*. Cambridge Univ. Press, Cambridge. MR2895762 https://doi.org/10.1017/CBO9781139035613

TSYBAKOV, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** 135–166. MR2051002 https://doi.org/10.1214/aos/1079120131

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

WAGER, S. and WALTHER, G. (2015). Adaptive concentration of regression trees, with application to random forests. Preprint. Available at arXiv:1503.06388.

WEICHWALD, S. and PETERS, J. (2021). Distributional robustness as a guiding principle for causality in cognitive neuroscience. *J. Cogn. Neurosci.* **33** 226–247.

WEISS, K., KHOSHGOFTAAR, T. M. and WANG, D. (2016). A survey of transfer learning. *Journal of Big Data* **3** 9.

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

ZHANG, K., SCHÖLKOPF, B., MUANDET, K. and WANG, Z. (2013). Domain adaptation under target and conditional shift. In *Proceedings of the* 30*th International Conference on Machine Learning* **28** 819–827.

THE ANNALS
of
STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

VOLUME 49

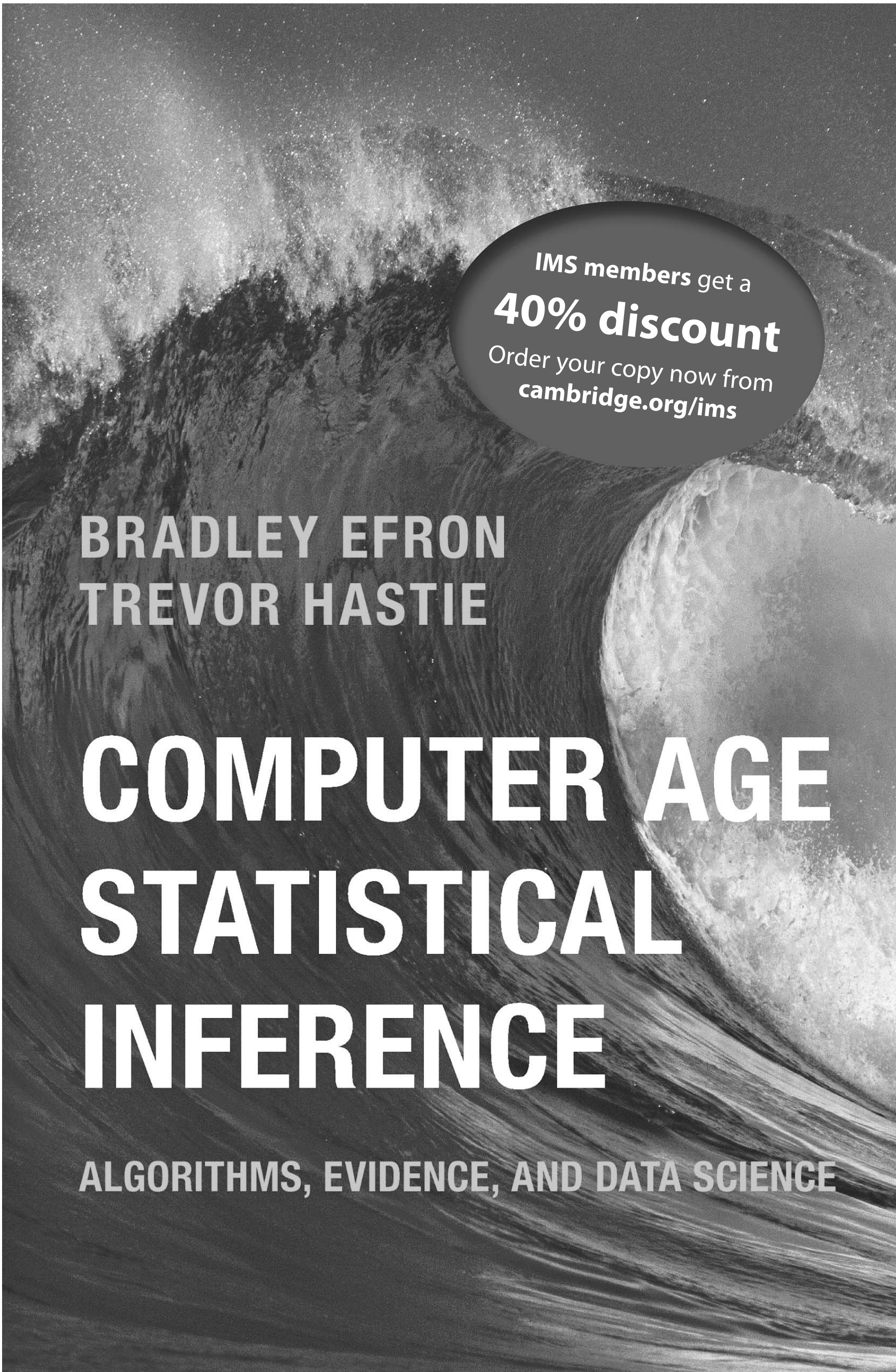
2021

CONTENTS OF VOLUME 49

Articles

Special Section of Memorial Articles for Willem van Zwet

Correction notes


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