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

THE ANNALS *of* STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Vol. 46, No. 4—August 2018

THE ANNALS OF STATISTICS Vol. 46, No. 4, pp. 1383–1836 August 2018

The Annals of Statistics **[ISSN 0090-5364 (print); ISSN 2168-8966 (online)],** Volume 46, Number 4, August 2018. Published bimonthly by the Institute of Mathematical Statistics, 3163 Somerset Drive, Cleveland, Ohio 44122, USA. Periodicals postage paid at Cleveland, Ohio, and at additional mailing offices.

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

The Annals of Statistics
2018, Vol. 46, No. 4, 1383–1414
https://doi.org/10.1214/17-AOS1588

LARGE COVARIANCE ESTIMATION THROUGH ELLIPTICAL FACTOR MODELS

BY JIANQING FAN[†,*,1], HAN LIU[*,2] AND WEICHEN WANG[*]

Princeton University[*] and Fudan University[†]

We propose a general Principal Orthogonal complEment Thresholding (POET) framework for large-scale covariance matrix estimation based on the approximate factor model. A set of high-level sufficient conditions for the procedure to achieve optimal rates of convergence under different matrix norms is established to better understand how POET works. Such a framework allows us to recover existing results for sub-Gaussian data in a more transparent way that only depends on the concentration properties of the sample covariance matrix. As a new theoretical contribution, for the first time, such a framework allows us to exploit conditional sparsity covariance structure for the heavy-tailed data. In particular, for the elliptical distribution, we propose a robust estimator based on the marginal and spatial Kendall's tau to satisfy these conditions. In addition, we study conditional graphical model under the same framework. The technical tools developed in this paper are of general interest to high-dimensional principal component analysis. Thorough numerical results are also provided to back up the developed theory.

REFERENCES

AGARWAL, A., NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *Ann. Statist.* **40** 1171–1197. MR2985947

AMINI, A. A. and WAINWRIGHT, M. J. (2009). High-dimensional analysis of semidefinite relaxations for sparse principal components. *Ann. Statist.* **37** 2877–2921. MR2541450

ANTONIADIS, A. and FAN, J. (2001). Regularization of wavelet approximations. *J. Amer. Statist. Assoc.* **96** 939–967. MR1946364

BAI, J. and LI, K. (2012). Statistical analysis of factor models of high dimension. *Ann. Statist.* **40** 436–465. MR3014313

BAI, J. and NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221. MR1926259

BAI, J. and NG, S. (2013). Principal components estimation and identification of static factors. *J. Econometrics* **176** 18–29. MR3067022

BELLONI, A. and CHERNOZHUKOV, V. (2011). ℓ_1-penalized quantile regression in high-dimensional sparse models. *Ann. Statist.* **39** 82–130. MR2797841

BERAN, R. (1978). An efficient and robust adaptive estimator of location. *Ann. Statist.* **6** 292–313. MR0518885

BERTHET, Q. and RIGOLLET, P. (2013a). Optimal detection of sparse principal components in high dimension. *Ann. Statist.* **41** 1780–1815. MR3127849

MSC2010 subject classifications. Primary 62H25; secondary 62H12.

Key words and phrases. Principal component analysis, approximate factor model, sub-Gaussian family, elliptical distribution, conditional graphical model, marginal and spatial Kendall's tau.

BERTHET, Q. and RIGOLLET, P. (2013b). Complexity theoretic lower bounds for sparse principal component detection. In *Proceedings of the 26th Annual Conference on Learning Theory* 1046–1066. PMLR, Princeton, NJ.

BICKEL, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10** 647–671. MR0663424

BICKEL, P. J. and LEVINA, E. (2008a). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

BICKEL, P. J. and LEVINA, E. (2008b). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969

BIRNBAUM, A., JOHNSTONE, I. M., NADLER, B. and PAUL, D. (2013). Minimax bounds for sparse PCA with noisy high-dimensional data. *Ann. Statist.* **41** 1055–1084. MR3113803

CAI, T. and LIU, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.* **106** 672–684. MR2847949

CAI, T., LIU, W. and LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973

CAI, T. T., MA, Z. and WU, Y. (2013). Sparse PCA: Optimal rates and adaptive estimation. *Ann. Statist.* **41** 3074–3110. MR3161458

CAI, T., MA, Z. and WU, Y. (2015). Optimal estimation and rank detection for sparse spiked covariance matrices. *Probab. Theory Related Fields* **161** 781–815. MR3334281

CAI, T. T., REN, Z. and ZHOU, H. H. (2013). Optimal rates of convergence for estimating Toeplitz covariance matrices. *Probab. Theory Related Fields* **156** 101–143. MR3055254

CAI, T. T., ZHANG, C.-H. and ZHOU, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38** 2118–2144. MR2676885

CAI, T. T. and ZHOU, H. H. (2012). Optimal rates of convergence for sparse covariance matrix estimation. *Ann. Statist.* **40** 2389–2420. MR3097607

CAI, T. T., LI, H., LIU, W. and XIE, J. (2013). Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika* **100** 139–156. MR3034329

CATONI, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 1148–1185. MR3052407

CHAMBERLAIN, G. and ROTHSCHILD, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica* **51** 1281–1304. MR0736050

CHANDRASEKARAN, V., PARRILO, P. A. and WILLSKY, A. S. (2012). Latent variable graphical model selection via convex optimization. *Ann. Statist.* **40** 1935–1967. MR3059067

CHANDRASEKARAN, V., SANGHAVI, S., PARRILO, P. A. and WILLSKY, A. S. (2011). Rank-sparsity incoherence for matrix decomposition. *SIAM J. Optim.* **21** 572–596. MR2817479

CHOI, K. and MARDEN, J. (1998). A multivariate version of Kendall's τ. *J. Nonparametr. Stat.* **9** 261–293. MR1649510

CHRISTENSEN, D. (2005). Fast algorithms for the calculation of Kendall's τ. *Comput. Statist.* **20** 51–62. MR2162534

ČÍŽEK, P., HÄRDLE, W. and WERON, R., eds. (2005). *Statistical Tools for Finance and Insurance*. Springer, Berlin. MR2156757

CROUX, C., OLLILA, E. and OJA, H. (2002). Sign and rank covariance matrices: Statistical properties and application to principal components analysis. In *Statistical Data Analysis Based on the L_1-Norm and Related Methods* (*Neuchâtel*, 2002) 257–269. Birkhäuser, Basel. MR2001321

DAVIS, C. and KAHAN, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.* **7** 1–46. MR0264450

DUPUIS LOZERON, E. and VICTORIA-FESER, M. P. (2010). Robust estimation of constrained covariance matrices for confirmatory factor analysis. *Comput. Statist. Data Anal.* **54** 3020–3032. MR2727731

DÜRRE, A., VOGEL, D. and TYLER, D. E. (2014). The spatial sign covariance matrix with unknown location. *J. Multivariate Anal.* **130** 107–117. MR3229526

EL KAROUI, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.* **36** 2717–2756. MR2485011

FAN, J., FAN, Y. and BARUT, E. (2014). Adaptive robust variable selection. *Ann. Statist.* **42** 324–351. MR3189488

FAN, J., FAN, Y. and LV, J. (2008). High dimensional covariance matrix estimation using a factor model. *J. Econometrics* **147** 186–197. MR2472991

FAN, J., FENG, Y. and WU, Y. (2009). Network exploration via the adaptive lasso and SCAD penalties. *Ann. Appl. Stat.* **3** 521–541. MR2750671

FAN, J., LI, Q. and WANG, Y. (2017). Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 247–265. MR3597972

FAN, J., LIAO, Y. and MINCHEVA, M. (2011). High-dimensional covariance matrix estimation in approximate factor models. *Ann. Statist.* **39** 3320–3356. MR3012410

FAN, J., LIAO, Y. and MINCHEVA, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 603–680. With 33 discussions by 57 authors and a reply by Fan, Liao and Mincheva. MR3091653

FAN, J., LIAO, Y. and WANG, W. (2016). Projected principal component analysis in factor models. *Ann. Statist.* **44** 219–254. MR3449767

FAN, J., LIU, H. and WANG, W. (2018). Supplement to "Large covariance estimation through elliptical factor models". DOI:10.1214/17-AOS1588SUPP.

FAN, J., WANG, W. and ZHONG, Y. (2016). Robust covariance estimation for approximate factor models. Available at arXiv:1602.00719.

FAN, J., WANG, W. and ZHONG, Y. (2017). An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation. Available at arXiv:1603.03516.

FAN, J., KE, Z. T., LIU, H. and XIA, L. (2015). QUADRO: A supervised dimension reduction method via Rayleigh quotient optimization. *Ann. Statist.* **43** 1498–1534. MR3357869

FAN, J., LIU, H., WANG, W. and ZHU, Z. (2016). Heterogeneity adjustment with applications to graphical model inference. Available at arXiv:1602.05455.

FANG, K. T., KOTZ, S. and NG, K. W. (1990). *Symmetric Multivariate and Related Distributions. Monographs on Statistics and Applied Probability* **36**. Chapman & Hall, London. MR1071174

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

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

HAMPEL, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.* **69** 383–393. MR0362657

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

HAN, F. and LIU, H. (2018). ECA: High dimensional elliptical component analysis in non-Gaussian distributions. *J. Amer. Statist. Assoc.* **113** 252–268. MR3803462

HAN, F. and LIU, H. (2017). Statistical analysis of latent generalized correlation matrix estimation in transelliptical distribution. *Bernoulli* **23** 23–57. MR3556765

HAN, F., LU, J. and LIU, H. (2014). Robust scatter matrix estimation for high dimensional distributions with heavy tails. Technical report.

HSU, D., KAKADE, S. M. and ZHANG, T. (2011). Robust matrix decomposition with sparse corruptions. *IEEE Trans. Inform. Theory* **57** 7221–7234. MR2883652

HSU, D. and SABATO, S. (2014). Heavy-tailed regression with a generalized median-of-means. In *Proceedings of the* 31*st International Conference on Machine Learning* (*ICML*-14) 37–45. PMLR, Bejing, China.

HUBER, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.* **35** 73–101. MR0161415

HULT, H. and LINDSKOG, F. (2002). Multivariate extremes, aggregation and dependence in elliptical distributions. *Adv. in Appl. Probab.* **34** 587–608. MR1929599

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

KENDALL, M. G. (1948). *Rank Correlation Methods*. C. Griffin, London.

KNIGHT, W. R. (1966). A computer method for calculating Kendall's tau with ungrouped data. *J. Amer. Statist. Assoc.* **61** 436–439.

KOENKER, R. (2005). *Quantile Regression*. *Econometric Society Monographs* **38**. Cambridge Univ. Press Cambridge. MR2268657

LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. MR2572459

LEVINA, E. and VERSHYNIN, R. (2012). Partial estimation of covariance matrices. *Probab. Theory Related Fields* **153** 405–419. MR2948681

LIU, H., HAN, F. and ZHANG, C. (2012). Transelliptical graphical models. In *Advances in Neural Information Processing Systems* 800–808. Curran Associates, Inc., Lake Tahoe, NV.

LIU, L., HAWKINS, D. M., GHOSH, S. and YOUNG, S. S. (2003). Robust singular value decomposition analysis of microarray data. *Proc. Natl. Acad. Sci. USA* **100** 13167–13172. MR2016727

MA, Z. (2013). Sparse principal component analysis and iterative thresholding. *Ann. Statist.* **41** 772–801. MR3099121

MARDEN, J. I. (1999). Some robust estimates of principal components. *Statist. Probab. Lett.* **43** 349–359. MR1707944

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

MITRA, R. and ZHANG, C.-H. (2014). Multivariate analysis of nonparametric estimates of large correlation matrices. Available at arXiv:1403.6195.

PANG, H., LIU, H. and VANDERBEI, R. (2014). The fastclime package for linear programming and large-scale precision matrix estimation in R. *J. Mach. Learn. Res.* **15** 489–493.

PAUL, D. and JOHNSTONE, I. M. (2012). Augmented sparse principal component analysis for high dimensional data. Available at arXiv:1202.1242.

PISON, G., ROUSSEEUW, P. J., FILZMOSER, P. and CROUX, C. (2003). Robust factor analysis. *J. Multivariate Anal.* **84** 145–172. MR1965827

POSEKANY, A., FELSENSTEIN, K. and SYKACEK, P. (2011). Biological assessment of robust noise models in microarray data analysis. *Bioinformatics* **27** 807–814.

RACHEV, S. T. (2003). *Handbook of Heavy Tailed Distributions in Finance*. *Handbooks in Finance* **1**. Elsevier, Amsterdam.

RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980. MR2836766

ROTHMAN, A. J., LEVINA, E. and ZHU, J. (2009). Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.* **104** 177–186. MR2504372

ROTHMAN, A. J., LEVINA, E. and ZHU, J. (2010). Sparse multivariate regression with covariance estimation. *J. Comput. Graph. Statist.* **19** 947–962. Supplementary materials available online. MR2791263

ROUSSEEUW, P. J. and CROUX, C. (1993). Alternatives to the median absolute deviation. *J. Amer. Statist. Assoc.* **88** 1273–1283. MR1245360

RUTTIMANN, U. E., UNSER, M., RAWLINGS, R. R., RIO, D., RAMSEY, N. F., MATTAY, V. S., HOMMER, D. W., FRANK, J. A. and WEINBERGER, D. R. (1998). Statistical analysis of functional MRI data in the wavelet domain. *IEEE Trans. Med. Imag.* **17** 142–154.

SHEN, D., SHEN, H. and MARRON, J. S. (2013). Consistency of sparse PCA in high dimension, low sample size contexts. *J. Multivariate Anal.* **115** 317–333. MR3004561

TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York. Revised and extended from the 2004 French original. Translated by Vladimir Zaiats. MR2724359

TYLER, D. E. (1982). Radial estimates and the test for sphericity. *Biometrika* **69** 429–436. MR0671982

VANDERBEI, R. J. (2008). *Linear Programming*: *Foundations and Extensions*, 3rd ed. *International Series in Operations Research & Management Science* **114**. Springer, New York. MR2363558

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

VISURI, S., KOIVUNEN, V. and OJA, H. (2000). Sign and rank covariance matrices. *J. Statist. Plann. Inference* **91** 557–575. MR1814801

VOGEL, D. and FRIED, R. (2011). Elliptical graphical modelling. *Biometrika* **98** 935–951. MR2860334

VU, V. Q. and LEI, J. (2013). Minimax sparse principal subspace estimation in high dimensions. *Ann. Statist.* **41** 2905–2947. MR3161452

WANG, W. and FAN, J. (2017). Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *Ann. Statist.* **45** 1342–1374. MR3662457

WEGKAMP, M. and ZHAO, Y. (2016). Adaptive estimation of the copula correlation matrix for semiparametric elliptical copulas. *Bernoulli* **22** 1184–1226. MR3449812

WU, Y. and LIU, Y. (2009). Variable selection in quantile regression. *Statist. Sinica* **19** 801–817. MR2514189

YUAN, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* **11** 2261–2286. MR2719856

YUAN, M. and LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824

ZHAO, T., ROEDER, K. and LIU, H. (2014). Positive semidefinite rank-based correlation matrix estimation with application to semiparametric graph estimation. *J. Comput. Graph. Statist.* **23** 895–922. MR3270703

ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.* **15** 265–286. MR2252527

ZOU, H. and YUAN, M. (2008). Composite quantile regression and the oracle model selection theory. *Ann. Statist.* **36** 1108–1126. MR2418651

The Annals of Statistics
2018, Vol. 46, No. 4, 1415–1444
https://doi.org/10.1214/17-AOS1589

CURRENT STATUS LINEAR REGRESSION

BY PIET GROENEBOOM AND KIM HENDRICKX[1]

Delft University of Technology and Hasselt University

We construct $\sqrt{n}$-consistent and asymptotically normal estimates for the finite dimensional regression parameter in the current status linear regression model, which do not require any smoothing device and are based on maximum likelihood estimates (MLEs) of the infinite dimensional parameter. We also construct estimates, again only based on these MLEs, which are arbitrarily close to efficient estimates, if the generalized Fisher information is finite. This type of efficiency is also derived under minimal conditions for estimates based on smooth nonmonotone plug-in estimates of the distribution function. Algorithms for computing the estimates and for selecting the bandwidth of the smooth estimates with a bootstrap method are provided. The connection with results in the econometric literature is also pointed out.

REFERENCES

[1] CHEN, X., LINTON, O. and VAN KEILEGOM, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica* **71** 1591–1608. MR2000259

[2] COSSLETT, S. R. (1983). Distribution-free maximum likelihood estimator of the binary choice model. *Econometrica* **51** 765–782. MR0712369

[3] COSSLETT, S. R. (1987). Efficiency bounds for distribution-free estimators of the binary choice and the censored regression models. *Econometrica* **55** 559–585. MR0890854

[4] COSSLETT, S. R. (2007). Efficient estimation of semiparametric models by smoothed maximum likelihood. *Internat. Econom. Rev.* **48** 1245–1272. MR2375624

[5] DING, Y. and NAN, B. (2011). A sieve M-theorem for bundled parameters in semiparametric models, with application to the efficient estimation in a linear model for censored data. *Ann. Statist.* **39** 3032–3061. MR3012400

[6] DOMINITZ, J. and SHERMAN, R. P. (2005). Some convergence theory for iterative estimation procedures with an application to semiparametric estimation. *Econometric Theory* **21** 838–863. MR2189497

[7] FINKELSTEIN, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* **42** 845–854. MR0872963

[8] FINKELSTEIN, D. M. and WOLFE, R. A. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics* **41** 933–945. MR0833140

[9] GINÉ, E. and NICKL, R. (2015). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Univ. Press, New York. MR3588285

[10] GROENEBOOM, P. and HENDRICKX, K. (2018). Supplement to "Current status linear regression." DOI:10.1214/17-AOS1589SUPP.

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

MSC2010 subject classifications. Primary 62G05, 62N01; secondary 62-04.
Key words and phrases. Current status, linear regression, MLE, semiparametric model.

[12] GROENEBOOM, P., JONGBLOED, G. and WITTE, B. I. (2010). Maximum smoothed likelihood estimation and smoothed maximum likelihood estimation in the current status model. *Ann. Statist.* **38** 352–387. MR2589325

[13] GROENEBOOM, P. and WELLNER, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. *DMV Seminar* **19**. Birkhäuser, Basel. MR1180321

[14] HAN, A. K. (1987). Nonparametric analysis of a generalized regression model. The maximum rank correlation estimator. *J. Econometrics* **35** 303–316. MR0903188

[15] HÄRDLE, W., HALL, P. and ICHIMURA, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21** 157–178. MR1212171

[16] HÖFFDING, W. (1940). Maszstabinvariante Korrelationstheorie. *Schr. Math. Inst. U. Inst. Angew. Math. Univ. Berlin* **5** 181–233. Translated in: *The Collected Works of Wassily Hoeffding* (N. I. Fisher and P. K. Sen, eds.). Springer, New York, 1994. MR0004426

[17] HUANG, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Ann. Statist.* **24** 540–568. MR1394975

[18] HUANG, J. and WELLNER, J. (1993). Regression models with interval censoring. In *Proceedings of the Kolmogorov Seminar*. Euler Mathematics Institute, St. Petersburg, Russia.

[19] KLEIN, R. W. and SPADY, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica* **61** 387–421. MR1209737

[20] KOSOROK, M. R. (2008). Bootstrapping in Grenander estimator. In *Beyond Parametrics in Interdisciplinary Research*: *Festschrift in Honor of Professor Pranab K. Sen*. *Inst. Math. Stat.* (*IMS*) *Collect.* **1** 282–292. IMS, Beachwood, OH. MR2462212

[21] LI, G. and ZHANG, C.-H. (1998). Linear regression with interval censored data. *Ann. Statist.* **26** 1306–1327. MR1647661

[22] MURPHY, S. A., VAN DER VAART, A. W. and WELLNER, J. A. (1999). Current status regression. *Math. Methods Statist.* **8** 407–425. MR1735473

[23] RABINOWITZ, D., TSIATIS, A. and ARAGON, J. (1995). Regression with interval-censored data. *Biometrika* **82** 501–513. MR1366277

[24] ROSSINI, A. J. and TSIATIS, A. A. (1996). A semiparametric proportional odds regression model for the analysis of current status data. *J. Amer. Statist. Assoc.* **91** 713–721. MR1395738

[25] SEN, B., BANERJEE, M. and WOODROOFE, M. (2010). Inconsistency of bootstrap: The Grenander estimator. *Ann. Statist.* **38** 1953–1977. MR2676880

[26] SEN, B. and XU, G. (2015). Model based bootstrap methods for interval censored data. *Comput. Statist. Data Anal.* **81** 121–129. MR3257405

[27] SHEN, X. (2000). Linear regression with current status data. *J. Amer. Statist. Assoc.* **95** 842–852. MR1804443

[28] SHERMAN, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica* **61** 123–137. MR1201705

[29] SHIBOSKI, S. C. and JEWELL, N. P. (1992). Statistical analysis of the time dependence of HIV infectivity based on partner study data. *J. Amer. Statist. Assoc.* **87** 360–372.

[30] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247

The Annals of Statistics
2018, Vol. 46, No. 4, 1445–1480
https://doi.org/10.1214/17-AOS1591

JUMP FILTERING AND EFFICIENT DRIFT ESTIMATION FOR LÉVY-DRIVEN SDES

BY ARNAUD GLOTER*, DASHA LOUKIANOVA* AND HILMAR MAI†

Université d'Evry Val d'Essonne and ENSAE-ParisTech*†

The problem of drift estimation for the solution X of a stochastic differential equation with Lévy-type jumps is considered under discrete high-frequency observations with a growing observation window. An efficient and asymptotically normal estimator for the drift parameter is constructed under minimal conditions on the jump behavior and the sampling scheme. In the case of a bounded jump measure density, these conditions reduce to $n\Delta_n^{3-\varepsilon} \to 0$, where n is the number of observations and Δ_n is the maximal sampling step. This result relaxes the condition $n\Delta_n^2 \to 0$ usually required for joint estimation of drift and diffusion coefficient for SDEs with jumps. The main challenge in this estimation problem stems from the appearance of the unobserved continuous part X^c in the likelihood function. In order to construct the drift estimator, we recover this continuous part from discrete observations. More precisely, we estimate, in a nonparametric way, stochastic integrals with respect to X^c. Convergence results of independent interest are proved for these nonparametric estimators.

REFERENCES

[1] APPLEBAUM, D. (2009). *Lévy Processes and Stochastic Calculus*, 2nd ed. *Cambridge Studies in Advanced Mathematics* **116**. Cambridge Univ. Press, Cambridge. MR2512800

[2] BARNDORFF-NIELSEN, O. E. and SHEPHARD, N. (2001). Non-Gaussian Ornstein-Uhlenbeck-based models and some of their uses in financial economics. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 167–241. MR1841412

[3] BIBINGER, M. and WINKELMANN, L. (2015). Econometrics of co-jumps in high-frequency data with noise. *J. Econometrics* **184** 361–378. MR3291008

[4] CONT, R. and TANKOV, P. (2004). *Financial Modelling with Jump Processes*. Chapman & Hall/CRC, Boca Raton, FL. MR2042661

[5] DITLEVSEN, S. and GREENWOOD, P. (2013). The Morris-Lecar neuron model embeds a leaky integrate-and-fire model. *J. Math. Biol.* **67** 239–259. MR3071157

[6] FIGUEROA-LÓPEZ, J. E. and NISEN, J. (2013). Optimally thresholded realized power variations for Lévy jump diffusion models. *Stochastic Process. Appl.* **123** 2648–2677. MR3054540

[7] FLORENS-ZMIROU, D. (1989). Approximate discrete-time schemes for statistics of diffusion processes. *Statistics* **20** 547–557. MR1047222

[8] GENON-CATALOT, V. and JACOD, J. (1993). On the estimation of the diffusion coefficient for multi-dimensional diffusion processes. *Ann. Inst. Henri Poincaré Probab. Stat.* **29** 119–151. MR1204521

MSC2010 subject classifications. 60J75, 62F12, 62M05.

Key words and phrases. Lévy-driven SDE, efficient drift estimation, maximum likelihood estimation, high frequency data, ergodic properties.

[9] GLOTER, A., LOUKIANOVA, D. and MAI, H. (2018). Supplement to "Jump filtering and efficient drift estimation for Lévy-driven SDEs." DOI:10.1214/17-AOS1591SUPP.
[10] HUTTON, J. E. and NELSON, P. I. (1984). Interchanging the order of differentiation and stochastic integration. *Stochastic Process*. *Appl*. **18** 371–377. MR0770201
[11] IBRAGIMOV, I. and HAS'MINSKII, R. (2013). *Statistical Estimation*: *Asymptotic Theory*, Springer-Verlag, New York.
[12] JACOD, J. and SHIRYAEV, A. N. (2003). *Limit Theorems for Stochastic Processes*, 2nd ed. *Grundlehren der Mathematischen Wissenschaften* **288**. Springer, Berlin. MR1943877
[13] KESSLER, M. (1997). Estimation of an ergodic diffusion from discrete observations. *Scand*. *J*. *Stat*. **24** 211–229. MR1455868
[14] KOU, S. G. (2002). A jump-diffusion model for option pricing. *Manage*. *Sci*. **48** 1086–1101.
[15] KÜCHLER, U. and SØRENSEN, M. (1999). A note on limit theorems for multivariate martingales. *Bernoulli* **5** 483–493. MR1693604
[16] LOUKIANOVA, D. and LOUKIANOV, O. (2005). Uniform law of large numbers and consistency of estimators for Harris diffusions. *Statist*. *Probab*. *Lett*. **74** 347–355. MR2186479
[17] MAI, H. (2014). Efficient maximum likelihood estimation for Lévy-driven Ornstein-Uhlenbeck processes. *Bernoulli* **20** 919–957. MR3178522
[18] MANCINI, C. (2011). The speed of convergence of the threshold estimator of integrated variance. *Stochastic Process*. *Appl*. **121** 845–855. MR2770909
[19] MASUDA, H. (2007). Ergodicity and exponential β-mixing bounds for multidimensional diffusions with jumps. *Stochastic Process*. *Appl*. **117** 35–56. MR2287102
[20] MASUDA, H. (2009). Erratum to: "Ergodicity and exponential β-mixing bound for multidimensional diffusions with jumps" [*Stochastic Process. Appl.* 117 (2007) 35–56] [MR2287102]. *Stochastic Process*. *Appl*. **119** 676–678. MR2494009
[21] MASUDA, H. (2010). Approximate self-weighted LAD estimation of discretely observed ergodic Ornstein-Uhlenbeck processes. *Electron*. *J*. *Stat*. **4** 525–565. MR2660532
[22] MASUDA, H. (2013). Convergence of Gaussian quasi-likelihood random fields for ergodic Lévy driven SDE observed at high frequency. *Ann*. *Statist*. **41** 1593–1641. MR3113823
[23] MERTON, R. (1976). Option pricing when underlying stock returns are discontinuous. *J*. *Financ*. *Econ*. **3** 125–144.
[24] OGIHARA, T. and YOSHIDA, N. (2011). Quasi-likelihood analysis for the stochastic differential equation with jumps. *Stat*. *Inference Stoch*. *Process*. **14** 189–229. MR2832818
[25] REVUZ, D. and YOR, M. (1991). *Continuous Martingales and Brownian Motion*. *Grundlehren der Mathematischen Wissenschaften* **293**. Springer, Berlin. MR1083357
[26] SHIMIZU, Y. (2006). M-estimation for discretely observed ergodic diffusion processes with infinitely many jumps. *Stat*. *Inference Stoch*. *Process*. **9** 179–225. MR2249182
[27] SHIMIZU, Y. (2008). Some remarks on estimation of diffusion coefficients for jump-diffusions from finite samples. *Bull*. *Inform*. *Cybernet*. **40** 51–60. MR2484445
[28] SHIMIZU, Y. (2008). A practical inference for discretely observed jump-diffusions from finite samples. *J*. *Japan Statist*. *Soc*. **38** 391–413. MR2510946
[29] SHIMIZU, Y. and YOSHIDA, N. (2006). Estimation of parameters for diffusion processes with jumps from discrete observations. *Stat*. *Inference Stoch*. *Process*. **9** 227–277. MR2252242
[30] TRAN, N. K. (2014). LAN property for jump diffusion processes with discrete observations via Malliavin calculus Ph.D. thesis Univ. Paris 13.
[31] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247
[32] YOSHIDA, N. (1992). Estimation for diffusion processes from discrete observation. *J*. *Multivariate Anal*. **41** 220–242. MR1172898

The Annals of Statistics
2018, Vol. 46, No. 4, 1481–1512
https://doi.org/10.1214/17-AOS1592

CONSISTENCY AND CONVERGENCE RATE OF PHYLOGENETIC INFERENCE VIA REGULARIZATION[2]

BY VU DINH[*,1], LAM SI TUNG HO[†,1], MARC A. SUCHARD[†] AND FREDERICK A. MATSEN IV[*,3]

Fred Hutchinson Cancer Research Center and University of California, Los Angeles†*

It is common in phylogenetics to have some, perhaps partial, information about the overall evolutionary tree of a group of organisms and wish to find an evolutionary tree of a specific gene for those organisms. There may not be enough information in the gene sequences alone to accurately reconstruct the correct "gene tree." Although the gene tree may deviate from the "species tree" due to a variety of genetic processes, in the absence of evidence to the contrary it is parsimonious to assume that they agree. A common statistical approach in these situations is to develop a likelihood penalty to incorporate such additional information. Recent studies using simulation and empirical data suggest that a likelihood penalty quantifying concordance with a species tree can significantly improve the accuracy of gene tree reconstruction compared to using sequence data alone. However, the consistency of such an approach has not yet been established, nor have convergence rates been bounded. Because phylogenetics is a nonstandard inference problem, the standard theory does not apply. In this paper, we propose a penalized maximum likelihood estimator for gene tree reconstruction, where the penalty is the square of the Billera–Holmes–Vogtmann geodesic distance from the gene tree to the species tree. We prove that this method is consistent, and derive its convergence rate for estimating the discrete gene tree structure and continuous edge lengths (representing the amount of evolution that has occurred on that branch) simultaneously. We find that the regularized estimator is "adaptive fast converging," meaning that it can reconstruct all edges of length greater than any given threshold from gene sequences of polynomial length. Our method does not require the species tree to be known exactly; in fact, our asymptotic theory holds for any such guide tree.

REFERENCES

ÅKERBORG, Ö., SENNBLAD, B., ARVESTAD, L. and LAGERGREN, J. (2009). Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proc. Natl. Acad. Sci. USA* **106** 5714–5719.

AMENTA, N., GODWIN, M., POSTARNAKEVICH, N. and JOHN, K. S. (2007). Approximating geodesic tree distance. *Inform. Process. Lett.* **103** 61–65. MR2322069

MSC2010 subject classifications. Primary 05C05, 62F12; secondary 92B10, 92D15.

Key words and phrases. Phylogenetics, tree reconstruction, gene tree, species tree, maximum likelihood estimator, regularization.

ATTESON, K. (1997). The performance of neighbor-joining algorithms of phylogeny reconstruction. In *Computing and Combinatorics* (*Shanghai*, 1997). *Lecture Notes in Computer Science* **1276** 101–110. Springer, Berlin. MR1616308

BAČÁK, M. (2013). The proximal point algorithm in metric spaces. *Israel J. Math.* **194** 689–701. MR3047087

BAČÁK, M. (2014). Computing medians and means in Hadamard spaces. *SIAM J. Optim.* **24** 1542–1566. MR3264572

BANSAL, M. S., WU, Y.-C., ALM, E. J. and KELLIS, M. (2015). Improved gene tree error correction in the presence of horizontal gene transfer. *Bioinformatics* **31** 1211–1218.

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

BOUSSAU, B. and DAUBIN, V. (2010). Genomes as documents of evolutionary history. *Trends Ecol. Evol.* **25** 224–232.

BOUSSAU, B., SZÖLLŐSI, G. J., DURET, L., GOUY, M., TANNIER, E. and DAUBIN, V. (2013). Genome-scale coestimation of species and gene trees. *Genome Res.* **23** 323–330.

CHAKERIAN, J. and HOLMES, S. (2013). distory: Distance Between Phylogenetic Histories. R package version 1.4.2.

CSURÖS, M. (2002). Fast recovery of evolutionary trees with thousands of nodes. *J. Comput. Biol.* **9** 277–297.

CUCKER, F. and SMALE, S. (2002). On the mathematical foundations of learning. *Bull. Amer. Math. Soc.* (*N.S.*) **39** 1–49. MR1864085

CUONG, N. V., HO, L. S. T. and DINH, V. (2013). Generalization and robustness of batched weighted average algorithm with V-geometrically ergodic Markov data. In *Algorithmic Learning Theory*. *Lecture Notes in Computer Science* **8139** 264–278. Springer, Heidelberg. MR3133071

DASKALAKIS, C., MOSSEL, E. and ROCH, S. (2011). Phylogenies without branch bounds: Contracting the short, pruning the deep. *SIAM J. Discrete Math.* **25** 872–893. MR2817535

DAVID, L. A. and ALM, E. J. (2011). Rapid evolutionary innovation during an Archaean genetic expansion. *Nature* **469** 93–96.

ENGL, H. W., HANKE, M. and NEUBAUER, A. (1996). *Regularization of Inverse Problems*. *Mathematics and Its Applications* **375**. Kluwer Academic, Dordrecht. MR1408680

ERDŐS, P. L., STEEL, M. A., SZÉKELY, L. A. and WARNOW, T. J. (1999). A few logs suffice to build (almost) all trees. I. *Random Structures Algorithms* **14** 153–184. MR1667319

FELSENSTEIN, J. (1984). DNAML in PHYLIP 2. University of Washington, Seattle. 6.

FELSENSTEIN, J. (2004). *Inferring Phylogenies*. Sinauer Associates, Sunderland.

GRONAU, I., MORAN, S. and SNIR, S. (2012). Fast and reliable reconstruction of phylogenetic trees with indistinguishable edges. *Random Structures Algorithms* **40** 350–384. MR2900143

HELED, J. and DRUMMOND, A. J. (2010). Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.* **27** 570–580.

HENDY, M. D. and PENNY, D. (1993). Spectral analysis of phylogenetic data. *J. Classification* **10** 5–24.

HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30. MR0144363

HOERL, A. E. (1962). Application of ridge analysis to regression problems. *Chem. Eng. Prog.* **58** 54–59.

HOFMANN, B. and YAMAMOTO, M. (2010). On the interplay of source conditions and variational inequalities for nonlinear ill-posed problems. *Appl. Anal.* **89** 1705–1727. MR2683677

HOFMANN, B., KALTENBACHER, B., PÖSCHL, C. and SCHERZER, O. (2007). A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators. *Inverse Probl.* **23** 987–1010. MR2329928

HOHAGE, T. and WEIDLING, F. (2015). Verification of a variational source condition for acoustic inverse medium scattering problems. *Inverse Probl.* **31** 075006, 14. MR3366755

HOMRIGHAUSEN, D. and MCDONALD, D. (2013). The lasso, persistence, and cross-validation. In *Proceedings of the* 30*th International Conference on Machine Learning* 1031–1039.

HUSON, D. H., NETTLES, S. M. and WARNOW, T. J. (1999). Disk-covering, a fast-converging method for phylogenetic tree reconstruction. *J. Comput. Biol.* **6** 369–386.

JI, S., KOLLÁR, J. and SHIFFMAN, B. (1992). A global Łojasiewicz inequality for algebraic varieties. *Trans. Amer. Math. Soc.* **329** 813–818. MR1046016

KAZIMIERSKI, K. S. (2010). *Aspects of Regularization in Banach Spaces*. Logos Verlag, Berlin.

KIM, J. (2000). Slicing hyperdimensional oranges: The geometry of phylogenetic estimation. *Mol. Phylogenet. Evol.* **17** 58–75.

KUHNER, M. K. and FELSENSTEIN, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* **11** 459–468.

LIU, L. and PEARL, D. K. (2007). Species trees from gene trees: Reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst. Biol.* **56** 504–514.

MADDISON, W. P. (1997). Gene trees in species trees. *Syst. Biol.* **46** 523–536.

MAHMUDI, O., SJÖSTRAND, J., SENNBLAD, B. and LAGERGREN, J. (2013). Genome-wide probabilistic reconciliation analysis across vertebrates. *BMC Bioinform.* **14** 1–11.

MOSSEL, E. (2003). On the impossibility of reconstructing ancestral data and phylogenies. *J. Comput. Biol.* **10** 669–676.

MOSSEL, E. (2004). Phase transitions in phylogeny. *Trans. Amer. Math. Soc.* **356** 2379–2404. MR2048522

MOSSEL, E. (2007). Distorted metrics on trees and phylogenetic forests. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **4** 108–116.

MOULTON, V. and STEEL, M. (2004). Peeling phylogenetic 'oranges'. *Adv. in Appl. Math.* **33** 710–727. MR2095862

NYE, T. M. W. (2011). Principal components analysis in the space of phylogenetic trees. *Ann. Statist.* **39** 2716–2739. MR2906884

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.

RASMUSSEN, M. D. and KELLIS, M. (2011). A Bayesian approach for fast and accurate gene tree reconstruction. *Mol. Biol. Evol.* **28** 273–290.

ROBINSON, D. F. (1971). Comparison of labeled trees with valency three. *J. Combin. Theory Ser. B* **11** 105–119. MR0294170

ROCH, S. and SLY, A. (2017). Phase transition in the sample complexity of likelihood-based phylogeny inference. *Probab. Theory Related Fields* **169** 3–62. MR3704765

ROCH, S. and WARNOW, T. (2015). On the robustness to gene tree estimation error (or lack thereof) of coalescent-based species tree methods. *Syst. Biol.* **64** 663–676.

ROKAS, A., WILLIAMS, B. L., KING, N. and CARROLL, S. B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425** 798–804.

SCHLIEP, K. (2011). phangorn: Phylogenetic analysis in R. *Bioinformatics* **27** 592–593.

SCORNAVACCA, C., JACOX, E. and SZÖLLŐSI, G. J. (2015). Joint amalgamation of most parsimonious reconciled gene trees. *Bioinformatics* **31** 841–848.

SEMPLE, C. and STEEL, M. (2003). *Phylogenetics*. *Oxford Lecture Series in Mathematics and Its Applications* **24**. Oxford Univ. Press, Oxford. MR2060009

STEEL, M., HENDY, M. D. and PENNY, D. (1998). Reconstructing phylogenies from nucleotide pattern probabilities: A survey and some new results. *Discrete Appl. Math.* **88** 367–396. MR1658533

STEEL, M. A. and SZÉKELY, L. A. (2002). Inverting random functions. II. Explicit bounds for discrete maximum likelihood estimation, with applications. *SIAM J. Discrete Math.* **15** 562–575. MR1935839

STEEL, M. A. and SZÉKELY, L. A. (2009). Inverting random functions. III. Discrete MLE revisited. *Ann. Comb.* **13** 365–382. MR2557045

SZÖLLŐSI, G. J., TANNIER, E., DAUBIN, V. and BOUSSAU, B. (2015). The inference of gene trees with species trees. *Syst*. *Biol*. **64** e42–e62.

VAN ERVEN, T. and HARREMOËS, P. (2014). Rényi divergence and Kullback–Leibler divergence. *IEEE Trans*. *Inform*. *Theory* **60** 3797–3820. MR3225930

WANG, Q. (2004). Maximum Likelihood Estimation of Phylogenetic Tree with Evolutionary Parameters Ph.D. thesis, The Ohio State University. Available at https://etd.ohiolink.edu/!etd.send_file?accession=osu1083177084&disposition=inline.

WARNOW, T., MORET, B. M. E. and JOHN, K. S. (2001). Absolute convergence: True trees from short sequences. In *Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms* (*Washington*, *DC*, 2001) 186–195. SIAM, Philadelphia, PA. MR1958406

WU, Y.-C., RASMUSSEN, M. D., BANSAL, M. S. and KELLIS, M. (2013). TreeFix: Statistically informed gene tree error correction using species trees. *Syst*. *Biol*. **62** 110–120.

ZHANG, P. (1993). Model selection via multifold cross validation. *Ann*. *Statist*. **21** 299–313. MR1212178

The Annals of Statistics
2018, Vol. 46, No. 4, 1513–1540
https://doi.org/10.1214/17-AOS1593

PARETO QUANTILES OF UNLABELED TREE OBJECTS

BY ELA SIENKIEWICZ AND HAONAN WANG[1]

Colorado State University

In this paper, we consider a set of unlabeled tree objects with topological and geometric properties. For each data object, two curve representations are developed to characterize its topological and geometric aspects. We further define the notions of topological and geometric medians as well as quantiles based on both representations. In addition, we take a novel approach to define the Pareto medians and quantiles through a multi-objective optimization problem. In particular, we study two different objective functions which measure the topological variation and geometric variation, respectively. Analytical solutions are provided for topological and geometric medians and quantiles, and in general, for Pareto medians and quantiles, the genetic algorithm is implemented. The proposed methods are applied to analyze a data set of pyramidal neurons.

REFERENCES

[1] ANTONIOU, A. and LU, W.-S. (2007). *Practical Optimization*: *Algorithms and Engineering Applications*. Springer, New York. MR2364120

[2] ASCOLI, G. A., DONOHUE, D. E. and HALAVI, M. (2007). NeuroMorpho.Org: A central resource for neuronal morphologies. *J. Neurosci*. **27** 9247–9251.

[3] ASCOLI, G. A. and KRICHMAR, J. L. (2000). L-Neuron: A modeling tool for the efficient generation and parsimonious description of dendritic morphology. *Neurocomputing* **32** 1003–1011.

[4] ASCOLI, G. A., KRICHMAR, J. L., SCORCIONI, R., NASUTO, S. J., SENFT, S. L. and KRICHMAR, G. L. (2001). Computer generation and quantitative morphometric analysis of virtual neurons. *Anat*. *Embryol*. **204** 283–301.

[5] AYDIN, B., PATAKI, G., WANG, H., BULLITT, E. and MARRON, J. S. (2009). A principal component analysis for trees. *Ann*. *Appl*. *Stat*. **3** 1597–1615. MR2752149

[6] BANKS, D. and CONSTANTINE, G. M. (1998). Metric models for random graphs. *J. Classification* **15** 199–223. MR1665974

[7] 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

[8] CANALE, A. and DUNSON, D. B. (2011). Bayesian kernel mixtures for counts. *J. Amer. Statist*. *Assoc*. **106** 1528–1539. MR2896854

[9] CHANG, H.-W., IYER, H., BULLITT, E. and WANG, H. (2013). Generalized linear mixed models for branching probabilities of brain artery systems. *Model Assist*. *Stat*. *Appl*. **8** 121–133.

[10] CLINE, H. and HAAS, K. (2008). The regulation of dendritic arbor development and plasticity by glutamatergic synaptic input: A review of the synaptotrophic hypothesis. *J. Physiol*. **586** 1509–1517.

MSC2010 subject classifications. Primary 62G99; secondary 62P10.
Key words and phrases. Data object, genetic algorithm, multi-objective optimization, object oriented data, tree-structured data.

[11] COELLO COELLO, C. A., LAMONT, G. B. and VAN VELDHUIZEN, D. A. (2007). *Evolutionary Algorithms for Solving Multi-Objective Problems*, 2nd ed. Springer, New York. With a foreword by David E. Goldberg. MR2350880

[12] FLAJOLET, P., GAO, Z., ODLYZKO, A. and RICHMOND, B. (1993). The distribution of heights of binary trees and other simple trees. *Combin. Probab. Comput.* **2** 145–156. MR1249127

[13] FRÉCHET, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. H. Poincaré* **10** 215–310. MR0027464

[14] GOLDBERG, D. (1989). *Genetic Algorithms in Optimization, Search and Machine Learning*. Addison Wesley Publishing Company, New York.

[15] GRUTZENDLER, J., HELMIN, K., TSAI, J. and GAN, W.-B. (2007). Various dendritic abnormalities are associated with fibrillar amyloid deposits in Alzheimer's disease. *Ann. N.Y. Acad. Sci.* **1097** 30–39.

[16] HARRIS, T. E. (1952). First passage and recurrence distributions. *Trans. Amer. Math. Soc.* **73** 471–486. MR0052057

[17] HENDRICKSON, P. J., GENE, J. Y., SONG, D. and BERGER, T. W. (2016). A million-plus neuron model of the hippocampal dentate gyrus: Critical role for topography in determining spatiotemporal network dynamics. *IEEE Trans. Biomed. Eng.* **63** 199–209.

[18] JOHNSTON, D. and WU, S. M.-S. (1994). *Foundations of Cellular Neurophysiology*. MIT Press, Cambridge.

[19] KOENKER, R. and HALLOCK, K. (2001). Quantile regression. *J. Econ. Perspect.* **15** 143–156.

[20] LIU, R. Y. (1990). On a notion of data depth based on random simplices. *Ann. Statist.* **18** 405–414. MR1041400

[21] MÄKINEN, E. (1999). Generating random binary trees—a survey. *Inform. Sci.* **115** 123–136. MR1671899

[22] MARRON, J. S. and ALONSO, A. M. (2014). Overview of object oriented data analysis. *Biom. J.* **56** 732–753. MR3258083

[23] MIGLIORE, M. and SHEPHERD, G. M. (2005). An integrated approach to classifying neuronal phenotypes. *Nat. Rev., Neurosci.* **6** 810–818.

[24] PADURARIU, M., CIOBICA, A., MAVROUDIS, I., FOTIOU, D. and BALOYANNIS, S. (2012). Hippocampal neuronal loss in the CA1 and CA3 areas of Alzheimer's disease patients. *Psychiatr. Danub.* **24** 152–158.

[25] PHILLIPS, C. and WARNOW, T. J. (1996). The asymmetric median tree—a new model for building consensus trees. *Discrete Appl. Math.* **71** 311–335. MR1420306

[26] PITMAN, J. (2006). *Combinatorial Stochastic Processes: Lectures from the* 32*nd Summer School on Probability Theory Held in Saint-Flour, July* 7–24, 2002. *Lecture Notes in Math.* **1875**. Springer, Berlin. With a foreword by Jean Picard. MR2245368

[27] PYAPALI, G. K., SIK, A., PENTTONEN, M., BUZSAKI, G. and TURNER, D. A. (1998). Dendritic properties of hippocampal CA1 pyramidal neurons in the rat: Intracellular staining in vivo and in vitro. *J. Comp. Neurol.* **391** 335–352.

[28] PYAPALI, G. K. and TURNER, D. A. (1994). Denervation-induced dendritic alterations in CA1 pyramidal cells following kainic acid hippocampal lesions in rats. *Brain Res.* **652** 279–290.

[29] PYAPALI, G. K. and TURNER, D. A. (1996). Increased dendritic extent in hippocampal CA1 neurons from aged F344 rats. *Neurobiol. Aging* **17** 601–611.

[30] SERFLING, R. (2002). Quantile functions for multivariate analysis: Approaches and applications. *Stat. Neerl.* **56** 214–232. Special issue: Frontier Research in Theoretical Statistics, 2000 (Eindhoven). MR1916321

[31] SHEN, D., SHEN, H., BHAMIDI, S., MUÑOZ MALDONADO, Y., KIM, Y. and MARRON, J. S. (2014). Functional data analysis of tree data objects. *J. Comput. Graph. Statist.* **23** 418–438. MR3215818

[32] SIENKIEWICZ, E. (2015). Analysis of big data and structured data with application in neuroscience. Ph.D. thesis, Colorado State Univ.

[33] SIENKIEWICZ, E. and WANG, H. (2017). Supplement to "Pareto quantiles of unlabeled tree objects." DOI:10.1214/17-AOS1593SUPP.

[34] ŠIMIĆ, G., KOSTOVIĆ, I., WINBLAD, B. and BOGDANOVIĆ, N. (1997). Volume and number of neurons of the human hippocampal formation in normal aging and Alzheimer's disease. *J. Comp. Neurol.* **379** 482–494.

[35] SIVANANDAM, S. N. and DEEPA, S. N. (2008). *Introduction to Genetic Algorithms*. Springer, Berlin. MR2441025

[36] SONG, D., CHAN, R. H. M., MARMARELIS, V. Z., HAMPSON, R. E., DEADWYLER, S. A. and BERGER, T. W. (2007). Nonlinear dynamic modeling of spike train transformations for hippocampal-cortical prostheses. *IEEE Trans. Biomed. Eng.* **54** 1053–1066.

[37] VIDA, I. (2010). Morphology of hippocampal neurons. In *Hippocampal Microcircuits* 27–67. Springer, Berlin.

[38] WALTER, S. (2011). Defining quantiles for functional data. Ph.D. thesis, The Univ. Melbourne.

[39] WANG, H. and MARRON, J. S. (2007). Object oriented data analysis: Sets of trees. *Ann. Statist.* **35** 1849–1873. MR2363955

[40] WANG, Y., MARRON, J. S., AYDIN, B., LADHA, A., BULLITT, E. and WANG, H. (2012). A nonparametric regression model with tree-structured response. *J. Amer. Statist. Assoc.* **107** 1272–1285. MR3036394

[41] WEST, M. J., COLEMAN, P. D., FLOOD, D. G. and TRONCOSO, J. C. (1994). Differences in the pattern of hippocampal neuronal loss in normal ageing and Alzheimer's disease. *Lancet* **344** 769–772.

The Annals of Statistics
2018, Vol. 46, No. 4, 1541–1572
https://doi.org/10.1214/17-AOS1594

EFFICIENT AND ADAPTIVE LINEAR REGRESSION IN SEMI-SUPERVISED SETTINGS[1]

BY ABHISHEK CHAKRABORTTY[2] AND TIANXI CAI

University of Pennsylvania and Harvard University

We consider the linear regression problem under semi-supervised settings wherein the available data typically consists of: (i) a small or moderate sized "labeled" data, and (ii) a *much larger* sized "unlabeled" data. Such data arises naturally from settings where the outcome, unlike the covariates, is expensive to obtain, a frequent scenario in modern studies involving large databases like electronic medical records (EMR). Supervised estimators like the ordinary least squares (OLS) estimator utilize only the labeled data. It is often of interest to investigate if and when the unlabeled data can be exploited to improve estimation of the regression parameter in the adopted linear model.

In this paper, we propose a class of "Efficient and Adaptive Semi-Supervised Estimators" (EASE) to improve estimation efficiency. The EASE are two-step estimators adaptive to model mis-specification, leading to improved (optimal in some cases) efficiency under model mis-specification, and equal (optimal) efficiency under a linear model. This adaptive property, often unaddressed in the existing literature, is crucial for advocating "safe" use of the unlabeled data. The construction of EASE primarily involves a flexible "semi-nonparametric" imputation, including a smoothing step that works well even when the number of covariates is not small; and a follow up "refitting" step along with a cross-validation (CV) strategy both of which have useful practical as well as theoretical implications towards addressing two important issues: under-smoothing and over-fitting. We establish asymptotic results including consistency, asymptotic normality and the adaptive properties of EASE. We also provide influence function expansions and a "double" CV strategy for inference. The results are further validated through extensive simulations, followed by application to an EMR study on auto-immunity.

REFERENCES

ANDREWS, D. W. K. (1995). Nonparametric kernel estimation for semiparametric models. *Econometric Theory* **11** 560–596. MR1349935

BELKIN, M., NIYOGI, P. and SINDHWANI, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *J. Mach. Learn. Res.* **7** 2399–2434. MR2274444

CASTELLI, V. and COVER, T. M. (1995). The exponential value of labeled samples. *Pattern Recogn. Lett.* **16** 105–111.

MSC2010 subject classifications. 62F35, 62J05, 62F12, 62G08.

Key words and phrases. Semi-supervised linear regression, semiparametric inference, model misspecification, adaptive estimation, semi-nonparametric imputation.

CASTELLI, V. and COVER, T. M. (1996). The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Trans. Inform. Theory* **42** 2102–2117. MR1447517

CHAKRABORTTY, A. and CAI, T. (2018). Supplement to "Efficient and adaptive linear regression in semi-supervised settings." DOI:10.1214/17-AOS1594SUPP.

CHAPELLE, O., SCHÖLKOPF, B. and ZIEN, A. (2006). *Semi-Supervised Learning*. MIT Press, Cambridge, MA.

COOK, R. D. (1998). Principal Hessian directions revisited. *J. Amer. Statist. Assoc.* **93** 84–100. MR1614584

COOK, R. D. and LEE, H. (1999). Dimension reduction in binary response regression. *J. Amer. Statist. Assoc.* **94** 1187–1200. MR1731482

COOK, R. D. and WEISBERG, S. (1991). Discussion of "Sliced inverse regression" by K.-C. Li. *J. Amer. Statist. Assoc.* **86** 328–332.

COZMAN, F. G. and COHEN, I. (2001). Unlabeled data can degrade classification performance of generative classifiers. Technical Report No. HPL-2001-234, HP Laboratories, Palo Alto, CA, USA.

COZMAN, F. G., COHEN, I. and CIRELO, M. C. (2003). Semi-supervised learning of mixture models. In *Proceedings of the Twentieth ICML* 99–106.

DUAN, N. and LI, K.-C. (1991). Slicing regression: A link-free regression method. *Ann. Statist.* **19** 505–530. MR1105834

HANSEN, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory* **24** 726–748. MR2409261

KAWAKITA, M. and KANAMORI, T. (2013). Semi-supervised learning with density-ratio estimation. *Mach. Learn.* **91** 189–209. MR3046904

KOHANE, I. S. (2011). Using electronic health records to drive discovery in disease genomics. *Nat. Rev. Genet.* **12** 417–428.

LAFFERTY, J. D. and WASSERMAN, L. (2007). Statistical analysis of semi-supervised regression. *Adv. Neural Inf. Process. Syst.* **20** 801–808.

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

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

LIAO, K. P., CAI, T., GAINER, V. et al. (2010). Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care and Research* **62** 1120–1127.

MASRY, E. (1996). Multivariate local polynomial regression for time series: Uniform strong consistency and rates. *J. Time Series Anal.* **17** 571–599. MR1424907

NEWEY, W. K. (1994). Kernel estimation of partial means and a general variance estimator. *Econometric Theory* **10** 233–253. MR1293201

NEWEY, W. K., HSIEH, F. and ROBINS, J. (1998). Undersmoothing and bias corrected functional estimation. Technical Report No. 98-17, Dept. of Economics, MIT, USA.

NIGAM, K. P. (2001). Using unlabeled data to improve text classification. Ph.D. thesis, Carnegie Mellon University, USA. CMU-CS-01-126. MR2703067

NIGAM, K., MCCALLUM, A. K., THRUN, S. and MITCHELL, T. (2000). Text classification from labeled and unlabeled documents using EM. *Mach. Learn.* **39** 103–134.

SEEGER, M. (2002). Learning with labeled and unlabeled data. Technical Report No. EPFL-REPORT-161327, Univ. Edinburgh, UK.

SOKOLOVSKA, N., CAPPÉ, O. and YVON, F. (2008). The asymptotics of semi-supervised learning in discriminative probabilistic models. In *Proceedings of the Twenty Fifth ICML* 984–991.

ZHANG, T. and OLES, F. J. (2000). The value of unlabeled data for classification problems. In *Proceedings of the Seventeenth ICML* 1191–1198.

ZHU, X. (2005). Semi-supervised learning through graphs. Ph.D. thesis, Carnegie Mellon Univ., USA. CMU-LTI-05-192.

ZHU, X. (2008). Semi-supervised learning literature survey. Technical Report No. 1530, Computer Sciences, Univ. Wisconsin-Madison, USA.

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

The Annals of Statistics
2018, Vol. 46, No. 4, 1573–1602
https://doi.org/10.1214/17-AOS1595

CONVEXIFIED MODULARITY MAXIMIZATION FOR DEGREE-CORRECTED STOCHASTIC BLOCK MODELS

BY YUDONG CHEN[1], XIAODONG LI[2] AND JIAMING XU[3]

Cornell University, University of California, Davis and Purdue University

The stochastic block model (SBM), a popular framework for studying community detection in networks, is limited by the assumption that all nodes in the same community are statistically equivalent and have equal expected degrees. The degree-corrected stochastic block model (DCSBM) is a natural extension of SBM that allows for degree heterogeneity within communities. To find the communities under DCSBM, this paper proposes a *convexified modularity maximization* approach, which is based on a convex programming relaxation of the classical (generalized) modularity maximization formulation, followed by a novel doubly-weighted ℓ_1-norm k-medoids procedure. We establish nonasymptotic theoretical guarantees for approximate and perfect clustering, both of which build on a new *degree-corrected density gap condition*. Our approximate clustering results are insensitive to the minimum degree, and hold even in sparse regime with bounded average degrees. In the special case of SBM, our theoretical guarantees match the best-known results of computationally feasible algorithms. Numerically, we provide an efficient implementation of our algorithm, which is applied to both synthetic and real-world networks. Experiment results show that our method enjoys competitive performance compared to the state of the art in the literature.

REFERENCES

[1] ABBE, E., BANDEIRA, A. S. and HALL, G. (2016). Exact recovery in the stochastic block model. *IEEE Trans. Inform. Theory* **62** 471–487. MR3447993

[2] ADAMIC, A. and GLANCE, N. (2005). The political blogosphere and the 2004 US election: Divided they blog. In *Proceedings of the* 3*rd International Workshop on Link Discovery* 36–43. ACM, New York.

[3] AGARWAL, N., BANDEIRA, A. S., KOILIARIS, K. and KOLLA, A. (2015). Multisection in the stochastic block model using semidefinite programming. Preprint. Available at arXiv:1507.02323.

[4] AMES, B. P. W. and VAVASIS, S. A. (2014). Convex optimization for the planted k-disjoint-clique problem. *Math. Program.* **143** 299–337. MR3152071

[5] AMINI, A. A., CHEN, A., BICKEL, P. J. and LEVINA, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *Ann. Statist.* **41** 2097–2122. MR3127859

[6] AMINI, A. A. and LEVINA, E. (2014). On semidefinite relaxations for the block model. Available at arXiv:1406.5647.

MSC2010 subject classifications. 62H30, 91C20.

Key words and phrases. Community detection, modularity maximization, degree-corrected stochastic block model, convex relaxation, k-medians, social network.

[7] BANDEIRA, A. S. (2015). Random Laplacian matrices and convex relaxations. Preprint. Available at arXiv:1504.03987.

[8] BORDENAVE, C., LELARGE, M. and MASSOULIÉ, L. (2015). Non-backtracking spectrum of random graphs: Community detection and non-regular Ramanujan graphs. Preprint. Available at arXiv:1501.06087.

[9] CAI, T. T. and LI, X. (2015). Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *Ann. Statist.* **43** 1027–1059. MR3346696

[10] CHARIKAR, M., GUHA, S., TARDOS, É. and SHMOYS, D. B. (1999). A constant-factor approximation algorithm for the k-median problem (extended abstract). In *Annual ACM Symposium on Theory of Computing* (*Atlanta*, *GA*, 1999) 1–10. ACM, New York. MR1797458

[11] CHAUDHURI, K., CHUNG, F. and TSIATAS, A. (2012). Spectral clustering of graphs with general degrees in the extended planted partition model. In *Proceedings of the* 25*th Annual Conference on Learning Theory* (*COLT*) 35.1–35.23.

[12] CHEN, Y., LI, X. and XU, J. (2018). Supplement to "Convexified modularity maximization for degree-corrected stochastic block models." DOI:10.1214/17-AOS1595SUPP.

[13] CHEN, Y., SANGHAVI, S. and XU, H. (2012). Clustering sparse graphs. *Adv. Neural Inf. Process. Syst.* 2213–2221.

[14] CHEN, Y. and XU, J. (2014). Statistical-computational phase transitions in planted models: The high-dimensional setting. In *Proceedings of the* 31*st International Conference on Machine Learning* 244–252.

[15] COJA-OGHLAN, A. and LANKA, A. (2009). Finding planted partitions in random graphs with general degree distributions. *SIAM J. Discrete Math.* **23** 1682–1714. MR2570199

[16] CONDON, A. and KARP, R. M. (2001). Algorithms for graph partitioning on the planted partition model. *Random Structures Algorithms* **18** 116–140. MR1809718

[17] DASGUPTA, A., HOPCROFT, J. and MCSHERRY, F. (2004). Spectral analysis of random graphs with skewed degree distributions. In *The* 45*th IEEE FOCS* 602–610.

[18] FORTUNATO, S. (2010). Community detection in graphs. *Phys. Rep.* **486** 75–174. MR2580414

[19] FORTUNATO, S. and BARTHELEMY, M. (2007). Resolution limit in community detection. *Proc. Natl. Acad. Sci. USA* **104** 36–41.

[20] GAO, C., MA, Z., ZHANG, A. Y. and ZHOU, H. H. (2015). Achieving optimal misclassification proportion in stochastic block model. Preprint. Available at arXiv:1505.03772.

[21] GROTHENDIECK, A. (1996). Résumé de la théorie métrique des produits tensoriels topologiques. *Resen. Inst. Mat. Estat. Univ. Sao Paulo* **2** 401–480. MR1466414

[22] GUÉDON, O. and VERSHYNIN, R. (2016). Community detection in sparse networks via Grothendieck's inequality. *Probab. Theory Related Fields* **165** 1025–1049. MR3520025

[23] GULIKERS, L., LELARGE, M. and MASSOULIÉ, L. (2017). A spectral method for community detection in moderately sparse degree-corrected stochastic block models. *Adv. in Appl. Probab.* **49** 686–721. MR3694314

[24] HAJEK, B., WU, Y. and XU, J. (2016). Achieving exact cluster recovery threshold via semidefinite programming: Extensions. *IEEE Trans. Inform. Theory* **62** 5918–5937. MR3552431

[25] HAJEK, B., WU, Y. and XU, J. (2016). Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Trans. Inform. Theory* **62** 2788–2797. MR3493879

[26] HAJEK, B., WU, Y. and XU, J. (2016). Semidefinite programs for exact recovery of a hidden community. In *Proceedings of Conference on Learning Theory* (*COLT*). Available at arXiv:1602.06410.

[27] HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. MR0718088

[28] JIN, J. (2015). Fast community detection by SCORE. *Ann. Statist.* **43** 57–89. MR3285600

[29] KARRER, B. and NEWMAN, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* (3) **83** 016107, 10. MR2788206

[30] KRZAKALA, F., MOORE, C., MOSSEL, E., NEEMAN, J., SLY, A., ZDEBOROVÁ, L. and ZHANG, P. (2013). Spectral redemption in clustering sparse networks. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **110** 20935–20940. MR3174850
[31] LANCICHINETTI, A. and FORTUNATO, S. (2011). Limits of modularity maximization in community detection. *Phys*. *Rev*. *E* **84**.
[32] LE, C. M., LEVINA, E. and VERSHYNIN, R. (2016). Optimization via low-rank approximation for community detection in networks. *Ann*. *Statist*. **44** 373–400. MR3449772
[33] LE, C. M. and VERSHYNIN, R. (2015). Concentration and regularization of random graphs. Preprint. Available at arXiv:1506.00669.
[34] LEI, J. and RINALDO, A. (2015). Consistency of spectral clustering in stochastic block models. *Ann*. *Statist*. **43** 215–237. MR3285605
[35] LINDENSTRAUSS, J. and PEŁCZYŃSKI, A. (1968). Absolutely summing operators in L_p-spaces and their applications. *Studia Math*. **29** 275–326. MR0231188
[36] MCSHERRY, F. (2001). Spectral partitioning of random graphs. In 42*nd IEEE Symposium on Foundations of Computer Science* (*Las Vegas*, *NV*, 2001) 529–537. IEEE Computer Soc., Los Alamitos, CA. MR1948742
[37] MONTANARI, A. and SEN, S. (2015). Semidefinite programs on sparse random graphs. Preprint. Available at arXiv:1504.05910.
[38] NEWMAN, M. E. J. (2006). Modularity and community structure in networks. *PNAS* **103** 8577–8582.
[39] OYMAK, S. and HASSIBI, B. (2011). Finding dense clusters via low rank + sparse decomposition. Preprint. Available at arXiv:1104.5186.
[40] PERRY, W. and WEIN, A. S. (2015). A semidefinite program for unbalanced multisection in the stochastic block model. Preprint. Available at arXiv:1507.05605.
[41] QIN, T. and ROHE, K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems* 3120–3128.
[42] REICHARDT, J. and BORNHOLDT, S. (2006). Statistical mechanics of community detection. *Phys*. *Rev*. *E* (3) **74** 016110, 14. MR2276596
[43] ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann*. *Statist*. **39** 1878–1915. MR2893856
[44] TRAUD, A. L., KELSIC, E. D., MUCHA, P. J. and PORTER, M. A. (2011). Comparing community structure to characteristics in online collegiate social networks. *SIAM Rev*. **53** 526–543. MR2834086
[45] TRAUD, A. L., MUCHA, P. J. and PORTER, M. A. (2012). Social structure of Facebook networks. *Phys*. *A* **391** 4165–4180.
[46] ZHANG, A. Y. and ZHOU, H. H. (2015). Minimax rates of community detection in stochastic block models. Preprint. Available at arXiv:1507.05313.
[47] ZHANG, Y., LEVINA, E. and ZHU, J. (2014). Detecting overlapping communities in networks using spectral methods. Preprint. Available at arXiv:1412.3432.
[48] ZHAO, Y., LEVINA, E. and ZHU, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *Ann*. *Statist*. **40** 2266–2292. MR3059083

The Annals of Statistics
2018, Vol. 46, No. 4, 1603–1629
https://doi.org/10.1214/17-AOS1596

NEAR-OPTIMALITY OF LINEAR RECOVERY IN GAUSSIAN OBSERVATION SCHEME UNDER $\|\cdot\|_2^2$-LOSS

BY ANATOLI JUDITSKY[1] AND ARKADI NEMIROVSKI[2]

Université Grenoble-Alpes and Georgia Institute of Technology

We consider the problem of recovering linear image Bx of a signal x known to belong to a given convex compact set $\mathcal{X}$ from indirect observation $\omega = Ax + \sigma\xi$ of x corrupted by Gaussian noise ξ. It is shown that under some assumptions on $\mathcal{X}$ (satisfied, e.g., when $\mathcal{X}$ is the intersection of K concentric ellipsoids/elliptic cylinders), an easy-to-compute linear estimate is near-optimal in terms of its worst case, over $x \in \mathcal{X}$, expected $\|\cdot\|_2^2$-loss. The main novelty here is that the result imposes no restrictions on A and B. To the best of our knowledge, preceding results on optimality of linear estimates dealt either with one-dimensional Bx (estimation of linear forms) or with the "diagonal case" where A, B are diagonal and $\mathcal{X}$ is given by a "separable" constraint like $\mathcal{X} = \{x : \sum_i a_i^2 x_i^2 \le 1\}$ or $\mathcal{X} = \{x : \max_i |a_i x_i| \le 1\}$.

REFERENCES

ARNOLD, B. F. and STAHLECKER, P. (2000). Another view of the Kuks–Olman estimator. *J. Statist. Plann. Inference* **89** 169–174. MR1794419

BEN-TAL, A., EL GHAOUI, L. and NEMIROVSKI, A. (2009). *Robust Optimization*. Princeton Univ. Press, Princeton, NJ. MR2546839

BOYD, S., EL GHAOUI, L., FERON, E. and BALAKRISHNAN, V. (1994). *Linear Matrix Inequalities in System and Control Theory*. *SIAM Studies in Applied Mathematics* **15**. SIAM, Philadelphia, PA. MR1284712

CHRISTOPEIT, N. and HELMES, K. (1996). Linear minimax estimation with ellipsoidal constraints. *Acta Appl. Math.* **43** 3–15. MR1385975

DONOHO, D. L. (1994). Statistical estimation and optimal recovery. *Ann. Statist.* **22** 238–270. MR1272082

DONOHO, D. L., LIU, R. C. and MACGIBBON, B. (1990). Minimax risk over hyperrectangles, and implications. *Ann. Statist.* **18** 1416–1437. MR1062717

DRYGAS, H. (1996). Spectral methods in linear minimax estimation. *Acta Appl. Math.* **43** 17–42. MR1385976

EFROMOVICH, S. (1999). *Nonparametric Curve Estimation: Methods, Theory, and Applications*. Springer, New York. MR1705298

EFROMOVICH, S. and PINSKER, M. (1996). Sharp-optimal and adaptive estimation for heteroscedastic nonparametric regression. *Statist. Sinica* **6** 925–942. MR1422411

GOLUBEV, Y. K., LEVIT, B. Y. and TSYBAKOV, A. B. (1996). Asymptotically efficient estimation of analytic functions in Gaussian noise. *Bernoulli* **2** 167–181. MR1410136

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

MSC2010 subject classifications. Primary 62G05, 62G08, 62J05; secondary 90C25, 90C22, 90C46.

Key words and phrases. Linear regression, linear estimation, minimax estimation.

JUDITSKY, A. and NEMIROVSKI, A. (2016). Near-optimality of linear recovery in Gaussian observation scheme under $\|\cdot\|_2$-loss. ArXiv Preprint ArXiv:1602.01355.

JUDITSKY, A. and NEMIROVSKI, A. (2018). Supplement to "Near-optimality of linear recovery in Gaussian observation scheme under $\|\cdot\|_2^2$-loss." DOI:10.1214/17-AOS1596SUPP.

KUKS, J. A. and OLMAN, W. (1971). Minimax linear estimation of regression coefficients (I). *Iswestija Akademija Nauk Estonskoj SSR* **20** 480–482.

KUKS, JA. and OLMAN, V. (1972). Linear minimax estimation of regression coefficients. *Eesti NSV Tead. Akad. Toim., Füüs. Mat.* **21** 66–72. (errata insert). MR0319297

LEPSKIĬ, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatn. Primen.* **35** 459–470. MR1091202

LIPTSER, R. S. and SHIRYAYEV, A. N. (1977). *Statistics of Random Processes. I: General Theory*. Springer, New York. MR0474486

PILZ, J. (1986). Minimax linear regression estimation with symmetric parameter restrictions. *J. Statist. Plann. Inference* **13** 297–318. MR0835614

PINSKER, M. S. (1980). Optimal filtration of square-integrable signals in Gaussian noise. *Probl. Inf. Transm.* **16** 120–133.

RAO, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. Wiley, New York. MR0346957

RAO, C. R. (1976). Estimation of parameters in a linear model. *Ann. Statist.* **4** 1023–1037. MR0420979

TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York. MR2724359

WASSERMAN, L. (2006). *All of Nonparametric Statistics*. Springer, New York. MR2172729

The Annals of Statistics
2018, Vol. 46, No. 4, 1630–1663
https://doi.org/10.1214/17-AOS1597

AN MCMC APPROACH TO EMPIRICAL BAYES INFERENCE AND BAYESIAN SENSITIVITY ANALYSIS VIA EMPIRICAL PROCESSES

BY HANI DOSS[1] AND YEONHEE PARK

University of Florida and MD Anderson Cancer Center

Consider a Bayesian situation in which we observe $Y \sim p_\theta$, where $\theta \in \Theta$, and we have a family $\{\nu_h, h \in \mathcal{H}\}$ of potential prior distributions on Θ. Let g be a real-valued function of θ, and let $I_g(h)$ be the posterior expectation of $g(\theta)$ when the prior is ν_h. We are interested in two problems: (i) selecting a particular value of h, and (ii) estimating the family of posterior expectations $\{I_g(h), h \in \mathcal{H}\}$. Let $m_y(h)$ be the marginal likelihood of the hyperparameter h: $m_y(h) = \int p_\theta(y)\nu_h(d\theta)$. The empirical Bayes estimate of h is, by definition, the value of h that maximizes $m_y(h)$. It turns out that it is typically possible to use Markov chain Monte Carlo to form point estimates for $m_y(h)$ and $I_g(h)$ for each individual h in a continuum, and also confidence intervals for $m_y(h)$ and $I_g(h)$ that are valid pointwise. However, we are interested in forming estimates, with confidence statements, of the entire families of integrals $\{m_y(h), h \in \mathcal{H}\}$ and $\{I_g(h), h \in \mathcal{H}\}$: we need estimates of the first family in order to carry out empirical Bayes inference, and we need estimates of the second family in order to do Bayesian sensitivity analysis. We establish strong consistency and functional central limit theorems for estimates of these families by using tools from empirical process theory. We give two applications, one to latent Dirichlet allocation, which is used in topic modeling, and the other is to a model for Bayesian variable selection in linear regression.

REFERENCES

ASUNCION, A., WELLING, M., SMYTH, P. and TEH, Y. W. (2009). On smoothing and inference for topic models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. *UAI* '09. AUAI Press, Arlington, Virginia, United States.

BERGER, J. O. (1994). An overview of robust Bayesian analysis. *TEST* **3** 5–124. MR1293110

BLEI, D. M., NG, A. Y. and JORDAN, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3** 993–1022.

BUTA, E. (2010). Computational approaches for empirical Bayes methods and Bayesian sensitivity analysis. Ph.D. thesis, Univ. Florida, Gainesville, FL.

DOSS, H. (2007). Bayesian model selection: Some thoughts on future directions. *Statist. Sinica* **17** 413–426. MR2408674

DOSS, H. and PARK, Y. (2018). Supplement to "An MCMC approach to empirical Bayes inference and Bayesian sensitivity analysis via empirical processes." DOI:10.1214/17-AOS1597SUPP.

DOSS, H. and TAN, A. (2014). Estimates and standard errors for ratios of normalizing constants from multiple Markov chains via regeneration. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 683–712. MR3248672

MSC2010 subject classifications. Primary 62F15, 91-08; secondary 62F12.

Key words and phrases. Donsker class, geometric ergodicity, hyperparameter selection, regenerative simulation, latent Dirichlet allocation model.

DOSS, C. R., FLEGAL, J. M., JONES, G. L. and NEATH, R. C. (2014). Markov chain Monte Carlo estimation of quantiles. *Electron. J. Stat.* **8** 2448–2478. MR3285872

FLEGAL, J. M., HARAN, M. and JONES, G. L. (2008). Markov chain Monte Carlo: Can we trust the third significant figure? *Statist. Sci.* **23** 250–260. MR2516823

FLEGAL, J. M. and JONES, G. L. (2010). Batch means and spectral variance estimators in Markov chain Monte Carlo. *Ann. Statist.* **38** 1034–1070. MR2604704

GEORGE, P. C. (2015). Latent Dirichlet Allocation: Hyperparameter selection and applications to electronic discovery. Ph.D. Thesis, Univ. Florida, Gainesville, FL.

GEORGE, E. I. and FOSTER, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87** 731–747. MR1813972

GEYER, C. J. (2011). Importance sampling, simulated tempering, and umbrella sampling. In *Handbook of Markov Chain Monte Carlo* (S. P. Brooks, A. E. Gelman, G. L. Jones and X. L. Meng, eds.) 295–311. CRC Press, Boca Raton, FL. MR2858453

GEYER, C. J. and THOMPSON, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Amer. Statist. Assoc.* **90** 909–920.

GRIFFITHS, T. L. and STEYVERS, M. (2004). Finding scientific topics. *Proc. Natl. Acad. Sci. USA* **101** 5228–5235.

HOBERT, J. P., JONES, G. L., PRESNELL, B. and ROSENTHAL, J. S. (2002). On the applicability of regenerative simulation in Markov chain Monte Carlo. *Biometrika* **89** 731–743. MR1946508

IBRAGIMOV, I. A. and LINNIK, YU. V. (1971). *Independent and Stationary Sequences of Random Variables*. Wolters-Noordhoff Publishing, Groningen. MR0322926

JONES, G. L., HARAN, M., CAFFO, B. S. and NEATH, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *J. Amer. Statist. Assoc.* **101** 1537–1547. MR2279478

KADANE, J. and WOLFSON, L. J. (1998). Experiences in elicitation. *J. R. Stat. Soc., Ser. D Stat.* **47** 3–19.

LEVENTAL, S. (1988). Uniform limit theorems for Harris recurrent Markov chains. *Probab. Theory Related Fields* **80** 101–118. MR0970473

LIANG, F., PAULO, R., MOLINA, G., CLYDE, M. A. and BERGER, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *J. Amer. Statist. Assoc.* **103** 410–423. MR2420243

MARINARI, E. and PARISI, G. (1992). Simulated tempering: A new Monte Carlo scheme. *Europhys. Lett.* **19** 451–458.

MEYN, S. P. and TWEEDIE, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer, London. MR1287609

MITCHELL, T. J. and BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83** 1023–1036. MR0997578

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

NEWTON, M. A. and RAFTERY, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *J. R. Stat. Soc., B* **56** 3–48. MR1257793

PARK, Y. (2015). A Markov chain Monte Carlo approach to empirical Bayes inference and Bayesian sensitivity analysis via empirical processes. Ph.D. thesis, Univ. Florida, Gainesville, FL.

PETRONE, S., ROUSSEAU, J. and SCRICCIOLO, C. (2014). Bayes and empirical Bayes: Do they merge? *Biometrika* **101** 285–302. MR3215348

ŘEHŮŘEK, R. and SOJKA, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC* 2010 *Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta.

ROY, V. and HOBERT, J. P. (2007). Convergence rates and asymptotic standard errors for Markov chain Monte Carlo algorithms for Bayesian probit regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 607–623. MR2370071

RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003). *Semiparametric Regression*. Cambridge Univ. Press, Cambridge. MR1998720

SMITH, M. and KOHN, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics* **75** 317–343.

STARK, P. C., RYAN, L. M., MCDONALD, J. L. and BURGE, H. A. (1997). Using meteorologic data to model and predict daily ragweed pollen levels. *Aerobiologia* **13** 177–184.

SUNG, Y. J. and GEYER, C. J. (2007). Monte Carlo likelihood inference for missing data models. *Ann. Statist.* **35** 990–1011. MR2341695

TAN, Z. (2017). Optimally adjusted mixture sampling and locally weighted histogram analysis. *J. Comput. Graph. Statist.* **26** 54–65.

TAN, A. and HOBERT, J. P. (2009). Block Gibbs sampling for Bayesian random effects models with improper priors: Convergence and regeneration. *J. Comput. Graph. Statist.* **18** 861–878. MR2598033

WALLACH, H. M., MURRAY, I., SALAKHUTDINOV, R. and MIMNO, D. (2009). Evaluation methods for topic models. In *Proceedings of the* 26*th Annual International Conference on Machine Learning*. ACM, New York.

WELLNER, J. (2005). Empirical processes: Theory and applications. Available at https://www.stat.washington.edu/people/jaw/RESEARCH/TALKS/Delft/emp-proc-delft-big.pdf.

WOLPERT, R. L. and SCHMIDLER, S. C. (2012). α-stable limit laws for harmonic mean estimators of marginal likelihoods. *Statist. Sinica* **22** 1233–1251. MR2987490

ZELLNER, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques* (P. K. Goel and A. Zellner, eds.). *Stud. Bayesian Econometrics Statist.* **6** 233–243. North-Holland, Amsterdam. MR0881437

The Annals of Statistics
2018, Vol. 46, No. 4, 1664–1692
https://doi.org/10.1214/17-AOS1598

CURVATURE AND INFERENCE FOR MAXIMUM LIKELIHOOD ESTIMATES[1]

BY BRADLEY EFRON

Stanford University

Maximum likelihood estimates are sufficient statistics in exponential families, but not in general. The theory of statistical curvature was introduced to measure the effects of MLE insufficiency in one-parameter families. Here, we analyze curvature in the more realistic venue of multiparameter families—more exactly, *curved exponential families*, a broad class of smoothly defined nonexponential family models. We show that within the set of observations giving the same value for the MLE, there is a "region of stability" outside of which the MLE is no longer even a local maximum. Accuracy of the MLE is affected by the location of the observation vector within the region of stability. Our motivating example involves "g-modeling," an empirical Bayes estimation procedure.

REFERENCES

AMARI, S. (1982). Differential geometry of curved exponential families—Curvatures and information loss. *Ann. Statist.* **10** 357–385. MR0653513

EFRON, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Statist.* **3** 1189–1242. With a discussion by C. R. Rao, D. A. Pierce, D. R. Cox, D. V. Lindley, L. LeCam, J. K. Ghosh, J. Pfanzagl, N. Keiding, A. P. Dawid, J. Reeds and with a reply by the author. MR0428531

EFRON, B. (1978). The geometry of exponential families. *Ann. Statist.* **6** 362–376. MR0471152

EFRON, B. (2010). *Large-Scale Inference. Empirical Bayes Methods for Estimation, Testing, and Prediction. Institute of Mathematical Statistics (IMS) Monographs* **1**. Cambridge Univ. Press, Cambridge. MR2724758

EFRON, B. (2016). Empirical Bayes deconvolution estimates. *Biometrika* **103** 1–20. MR3465818

EFRON, B. and HINKLEY, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika* **65** 457–487. With comments by O. Barndorff-Nielsen, A. T. James, G. K. Robinson and D. A. Sprott and a reply by the authors. MR0521817

FISHER, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. Lond. Ser. A* **222** 309–368. Available at www.jstor.org/stable/91208.

FISHER, R. A. (1925). Theory of statistical estimation. *Math. Proc. Cambridge Philos. Soc.* **22** 700–725. DOI:10.1017/S0305004100009580.

FISHER, R. A. (1934). Two new properties of mathematical likelihood. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **144** 285–307. Available at www.jstor.org/stable/2935559.

GOOD, I. J. and GASKINS, R. A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika* **58** 255–277. MR0319314

MSC2010 subject classifications. Primary 62Bxx; secondary 62Hxx.

Key words and phrases. Observed information, g-modeling, region of stability, curved exponential families, regularized MLE.

HAYASHI, M. and WATANABE, S. (2016). Information geometry approach to parameter estimation in Markov chains. *Ann. Statist.* **44** 1495–1535. MR3519931

HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67. DOI:10.1080/00401706.1970.10488634.

KASS, R. E. and VOS, P. W. (1997). *Geometrical Foundations of Asymptotic Inference*. Wiley, New York. MR1461540

LINDSEY, J. K. (1974). Construction and comparison of statistical models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **36** 418–425. MR0365794

MADSEN, L. T. (1979). The geometry of statistical model—A generalization of curvature. Technical report, Danish Medical Research Council. Statistical Research Unit Report 79-1.

RAO, C. R. (1961). Asymptotic efficiency and limiting information. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I* 531–545. Univ. California Press, Berkeley, CA. MR0133192

RAO, C. R. (1962). Efficient estimates and optimum inference procedures in large samples. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **24** 46–72. MR0293766

RAO, C. R. (1963). Criteria of estimation in large samples. *Sankhya, Ser. A* **25** 189–206. MR0175225

SCHWARTZMAN, A., DOUGHERTY, R. F. and TAYLOR, J. E. (2005). Cross-subject comparison of principal diffusion direction maps. *Magn. Reson. Med.* **53** 1423–1431. DOI:10.1002/mrm.20503.

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

TIERNEY, L. and KADANE, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *J. Amer. Statist. Assoc.* **81** 82–86. MR0830567

The Annals of Statistics
2018, Vol. 46, No. 4, 1693–1720
https://doi.org/10.1214/17-AOS1599

EMPIRICAL BAYES ESTIMATES FOR A TWO-WAY CROSS-CLASSIFIED MODEL

BY LAWRENCE D. BROWN[1], GOURAB MUKHERJEE[2] AND ASAF WEINSTEIN[1]

University of Pennsylvania, University of Southern California and Stanford University

We develop an empirical Bayes procedure for estimating the cell means in an unbalanced, two-way additive model with fixed effects. We employ a hierarchical model, which reflects exchangeability of the effects within treatment and within block but not necessarily between them, as suggested before by Lindley and Smith [*J. R. Stat. Soc., B* **34** (1972) 1–41]. The hyperparameters of this hierarchical model, instead of considered fixed, are to be substituted with data-dependent values in such a way that the point risk of the empirical Bayes estimator is small. Our method chooses the hyperparameters by minimizing an unbiased risk estimate and is shown to be asymptotically optimal for the estimation problem defined above, under suitable conditions. The usual empirical Best Linear Unbiased Predictor (BLUP) is shown to be substantially different from the proposed method in the unbalanced case and, therefore, performs suboptimally. Our estimator is implemented through a computationally tractable algorithm that is scalable to work under large designs. The case of missing cell observations is treated as well.

REFERENCES

BATES, D. M. (2010). lme4: Mixed-effects modeling with R. http://lme4.r-forge.r-project.org/book.

BROWN, L. D., MUKHERJEE, G. and WEINSTEIN, A. (2018). Supplement to "Empirical Bayes estimates for a two-way cross-classified model." DOI:10.1214/17-AOS1599SUPP.

CANDÈS, E. J., SING-LONG, C. A. and TRZASKO, J. D. (2013). Unbiased risk estimates for singular value thresholding and spectral estimators. *IEEE Trans. Signal Process.* **61** 4643–4657. MR3105401

DICKER, L. H. (2013). Optimal equivariant prediction for high-dimensional linear models with arbitrary predictor covariance. *Electron. J. Stat.* **7** 1806–1834. MR3084672

DICKER, L. H. and ERDOGDU, M. A. (2017). Flexible results for quadratic forms with applications to variance components estimation. *Ann. Statist.* **45** 386–414. MR3611496

DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1995). Wavelet shrinkage: Asymptopia? *J. R. Stat. Soc., B* **57** 301–369. MR1323344

DRAPER, N. R. and VAN NOSTRAND, R. C. (1979). Ridge regression and James–Stein estimation: Review and comments. *Technometrics* **21** 451–466. MR0555086

EFRON, B. and MORRIS, C. (1972). Empirical Bayes on vector observations: An extension of Stein's method. *Biometrika* **59** 335–347. MR0334386

MSC2010 subject classifications. Primary 62C12; secondary 62C25, 62F10, 62J07.

Key words and phrases. Shrinkage estimation, empirical Bayes, two-way ANOVA, oracle optimality, Stein's unbiased risk estimate (SURE), empirical BLUP.

EFRON, B. and MORRIS, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *J. Amer. Statist. Assoc.* **68** 117–130. MR0388597

FAHRMEIR, L., KNEIB, T., LANG, S. and MARX, B. (2013). *Regression: Models, Methods and Applications*. Springer, Heidelberg. MR3075546

GELMAN, A. (2005). Analysis of variance—why it is more important than ever. *Ann. Statist.* **33** 1–53. MR2157795

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2004). *Bayesian Data Analysis*, 2nd ed. Chapman & Hall/CRC, Boca Raton, FL. MR2027492

GHOSH, M., NICKERSON, D. M. and SEN, P. K. (1987). Sequential shrinkage estimation. *Ann. Statist.* **15** 817–829. MR0888442

GOLDSTEIN, H., BROWNE, W. and RASBASH, J. (2002). Multilevel modelling of medical data. *Stat. Med.* **21** 3291–3315.

HARVILLE, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Statist. Assoc.* **72** 320–340. MR0451550

HENDERSON, C. (1984). ANOVA, MIVQUE, REML, and ML algorithms for estimation of variances and covariances. In *Statistics: An Appraisal: Proceedings* 50*th Anniversary Conference* (H. A. David and H. T. David, eds.) 257–280. The Iowa State University Press, Ames, IA.

JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. In *Proc.* 4*th Berkeley Sympos. Math. Statist. and Prob., Vol. I* 361–379. Univ. California Press, Berkeley, CA. MR0133191

JIANG, J., NGUYEN, T. and RAO, J. S. (2011). Best predictive small area estimation. *J. Amer. Statist. Assoc.* **106** 732–745. MR2847987

JOHNSTONE, I. M. (2011). Gaussian estimation: Sequence and wavelet models. *Unpublished Manuscript*.

JOHNSTONE, I. M. and SILVERMAN, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Statist.* **32** 1594–1649. MR2089135

KOU, S. C. and YANG, J. J. (2017). Optimal shrinkage estimation in heteroscedastic hierarchical linear models. In *Big and Complex Data Analysis*. *Contrib. Stat.* 249–284. Springer, Cham. MR3674720

LI, K.-C. (1986). Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. *Ann. Statist.* **14** 1101–1112. MR0856808

LINDLEY, D. (1962). Discussion of the paper by Stein. *J. R. Stat. Soc., B* **24** 265–296.

LINDLEY, D. V. and SMITH, A. F. M. (1972). Bayes estimates for the linear model. *J. R. Stat. Soc., B* **34** 1–41. MR0415861

MASON, W. M., WONG, G. Y. and ENTWISLE, B. (1983). Contextual analysis through the multilevel linear model. *Sociol. Method.* **1984** 72–103.

MCCULLOCH, C. E. and SEARLE, S. R. (2001). *Generalized, Linear, and Mixed Models*. Wiley, New York. MR1884506

OMAN, S. D. (1982). Shrinking towards subspaces in multiple linear regression. *Technometrics* **24** 307–311. MR0687188

RASBASH, J. and GOLDSTEIN, H. (1994). Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. *J. Educ. Behav. Stat.* **19** 337–350.

ROLPH, J. E. (1976). Choosing shrinkage estimators for regression problems. *Comm. Statist. Theory Methods* **A5** 789–802. MR0436481

SCLOVE, S. L. (1968). Improved estimators for coefficients in linear regression. *J. Amer. Statist. Assoc.* **63** 596–606. MR0237057

SEARLE, S. R., CASELLA, G. and MCCULLOCH, C. E. (1992). *Variance Components*. Wiley, New York. MR1190470

STEIN, C. M. (1962). Confidence sets for the mean of a multivariate normal distribution. *J. R. Stat. Soc., B* **24** 265–296. MR0148184

TAN, Z. (2016). Steinized empirical Bayes estimation for heteroscedastic data. *Statist. Sinica* **26** 1219–1248. MR3559950

XIE, X., KOU, S. C. and BROWN, L. D. (2012). SURE estimates for a heteroscedastic hierarchical model. *J. Amer. Statist. Assoc.* **107** 1465–1479. MR3036408

ZACCARIN, S. and RIVELLINI, G. (2002). Multilevel analysis in social research: An application of a cross-classified model. *Stat. Methods Appl.* **11** 95–108.

The Annals of Statistics
2018, Vol. 46, No. 4, 1721–1741
https://doi.org/10.1214/17-AOS1600

ESTIMATING VARIANCE OF RANDOM EFFECTS TO SOLVE MULTIPLE PROBLEMS SIMULTANEOUSLY

BY MASAYO YOSHIMORI HIROSE[1] AND PARTHA LAHIRI[2]

Institute of Statistical Mathematics and University of Maryland

The two-level normal hierarchical model (NHM) has played a critical role in statistical theory for the last several decades. In this paper, we propose random effects variance estimator that simultaneously (i) improves on the estimation of the related shrinkage factors, (ii) protects empirical best linear unbiased predictors (EBLUP) [same as empirical Bayes (EB)] of the random effects from the common overshrinkage problem, (iii) avoids complex bias correction in generating strictly positive second-order unbiased mean square error (MSE) (same as integrated Bayes risk) estimator either by the Taylor series or single parametric bootstrap method. The idea of achieving multiple desirable properties in an EBLUP or EB method through a suitably devised random effects variance estimator is the first of its kind and holds promise in providing good inferences for random effects under the EBLUP or EB framework. The proposed methodology is also evaluated using a Monte Carlo simulation study and real data analysis.

REFERENCES

ARORA, V. and LAHIRI, P. (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Statist. Sinica* **7** 1053–1063. MR1488659

ARORA, V., LAHIRI, P. and MUKHERJEE, K. (1997). Empirical Bayes estimation of finite population means from complex surveys. *J. Amer. Statist. Assoc.* **92** 1555–1562. MR1615265

BELL, W. R. (1999). *Accounting for Uncertainty About Variances in Small Area Estimation. Bulletin of the International Statistical Institute* **52**. Session, Helsinki.

BELL, W. R. (2008). Examining sensitivity of small area inferences to uncertainty about sampling error variances. In *Proceedings of the American Statistical Association, Survey Research Methods Section* 327–334. CD-ROM, Amer. Statist. Assoc., Alexandria, VA.

BELL, W. R., BASEL, W. W. and MAPLES, J. J. (2015). An Overview of the U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program. In *Analysis of Poverty Data by Small Area Methods* (M. Pratesi, ed.). Wiley, New York.

BUTAR, F. B. and LAHIRI, P. (2003). On measures of uncertainty of empirical Bayes small-area estimators. *J. Statist. Plann. Inference* **112** 63–76. Model selection, model diagnostics, empirical Bayes and hierarchical Bayes, Special Issue II (Lincoln, NE, 1999). MR1961720

BUTAR, B. F. (1997). Empirical Bayes methods in survey sampling. Ph.D. thesis, Department of Mathematics and Statistics, Univ. Nebraska-Lincoln.

CARTER, G. M. and ROLPH, J. F. (1974). Empirical Bayes methods applied to estimating fire alarm probabilities. *J. Amer. Statist. Assoc.* **69** 880–885.

MSC2010 subject classifications. Primary 62C12; secondary 62F40.

Key words and phrases. Adjusted maximum likelihood method, empirical Bayes, empirical best linear unbiased prediction, linear mixed model, second-order unbiasedness.

CASAS-CORDERO, C., ENCINA, J. and LAHIRI, P. (2015). Poverty mapping for the Chilean comunas. In *Analysis of Poverty Data by Small Area Estimation* (M. Pratesi, ed.). Wiley, New York.

CHATTERJEE, S. and LAHIRI, P. (2007). A simple computational method for estimating mean squared prediction error in general small-area model. In *Proceedings of the Section on Survey Research Methods* 3486–3493. American Statistical Association, Alexandria, VA.

CITRO, C. and KALTON, G., eds. (2000) *"Small-Area Income and Poverty Estimates: Priorities for* 2000 *and Beyond," Panel on Estimates of Poverty for Small Geogruphic Area*. National Academy Press, Washington, DC.

DAS, K., JIANG, J. and RAO, J. N. K. (2004). Mean squared error of empirical predictor. *Ann. Statist.* **32** 818–840. MR2060179

DATTA, G. S. and LAHIRI, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statist. Sinica* **10** 613–627. MR1769758

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26. MR0515681

EFRON, B. and MORRIS, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *J. Amer. Statist. Assoc.* **68** 117–130. MR0388597

EFRON, B. and MORRIS, C. (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.* **70** 311–319.

FAY, R. E. III and HERRIOT, R. A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *J. Amer. Statist. Assoc.* **74** 269–277. MR0548019

HA, N. S., LAHIRI, P. and PARSONS, V. (2014). Methods and results for small area estimation using smoking data from the 2008 National Health Interview Survey. *Stat. Med.* **33** 3932–3945. MR3260671

HALL, P. and MAITI, T. (2006). On parametric bootstrap methods for small area prediction. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 221–238. MR2188983

JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I* 361–379. Univ. California Press, Berkeley, CA. MR0133191

JIANG, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. Springer, New York. MR2308058

JIANG, J., LAHIRI, P. and NGUYEN, T. (2016). A unified Monte-Carlo jackknife for small area estimation after model selection. Preprint. Available at arXiv:1602.05238.

LAIRD, N. M. and LOUIS, T. A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *J. Amer. Statist. Assoc.* **82** 739–757. MR0909979

LI, H. and LAHIRI, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. *J. Multivariate Anal.* **101** 882–892. MR2584906

LIU, B., LAHIRI, P. and KALTON, G. (2014). Hierarchical Bayes modeling of survey-weighted small area proportions. *Surv. Methodol.* **40** 1–13.

MOHADJER, L., RAO, J. N. K., LIU, B., KRENZKE, T. and VAN DE KERCKHOVE, W. (2012). Hierarchical Bayes small area estimates of adult literacy using unmatched sampling and linking models. *J. Indian Soc. Agricultural Statist.* **66** 55–63, 232–233. MR2953459

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

MORRIS, C. and TANG, R. (2011). Estimating random effects via adjustment for density maximization. *Statist. Sci.* **26** 271–287. MR2858514

OTTO, M. C. and BELL, W. R. (1995). Sampling error modeling of poverty and income statistics for states. In *Proceedings of the Section on Government Statistics* 160–165. American Statistical Association, Alexandria, VA.

PFEFFERMANN, D. and CORREA, S. (2012). Empirical bootstrap bias correction and estimation of prediction mean square error in small area estimation. *Biometrika* **99** 457–472. MR2931265

PFEFFERMANN, D. and GLICKMAN, H. (2004). Mean square error approximation in small area estimation by use of parametric and nonparametric bootstrap. In *Proceedings of the Section on Survey Research Methods* 4167–4178. American Statistical Association, Alexandria, VA.

PRASAD, N. G. N. and RAO, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *J. Amer. Statist. Assoc.* **85** 163–171. MR1137362

RAO, J. N. K. and MOLINA, I. (2015). *Small Area Estimation*, 2nd ed. Wiley, Hoboken, NJ. MR3380626

YOSHIMORI, M. and LAHIRI, P. (2014). A new adjusted maximum likelihood method for the Fay–Herriot small area model. *J. Multivariate Anal.* **124** 281–294. MR3147326

YOU, Y. and CHAPMAN, B. (2006). Small area estimation using area level models and estimated sampling variances. *Surv. Methodol.* **32** 97–103.

The Annals of Statistics
2018, Vol. 46, No. 4, 1742–1778
https://doi.org/10.1214/17-AOS1601

OPTIMAL SHRINKAGE OF EIGENVALUES IN THE SPIKED COVARIANCE MODEL[1]

BY DAVID DONOHO*, MATAN GAVISH†,[2] AND IAIN JOHNSTONE*

Stanford University and Hebrew University†*

To the memory of Charles M. Stein, 1920–2016

We show that in a common high-dimensional covariance model, the choice of loss function has a profound effect on optimal estimation.

In an asymptotic framework based on the spiked covariance model and use of orthogonally invariant estimators, we show that optimal estimation of the population covariance matrix boils down to design of an optimal shrinker η that acts elementwise on the sample eigenvalues. Indeed, to each loss function there corresponds a unique admissible eigenvalue shrinker η^* dominating all other shrinkers. The shape of the optimal shrinker is determined by the choice of loss function and, crucially, by inconsistency of both eigenvalues *and* eigenvectors of the sample covariance matrix.

Details of these phenomena and closed form formulas for the optimal eigenvalue shrinkers are worked out for a menagerie of 26 loss functions for covariance estimation found in the literature, including the Stein, Entropy, Divergence, Fréchet, Bhattacharya/Matusita, Frobenius Norm, Operator Norm, Nuclear Norm and Condition Number losses.

REFERENCES

[1] BAI, Z. and SILVERSTEIN, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2567175

[2] BAI, Z. and YAO, J. (2008). Central limit theorems for eigenvalues in a spiked population model. *Ann. Inst. Henri Poincaré Probab. Stat.* **44** 447–474. MR2451053

[3] BAIK, J., BEN AROUS, G. and PÉCHÉ, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.* **33** 1643–1697. MR2165575

[4] BAIK, J. and SILVERSTEIN, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *J. Multivariate Anal.* **97** 1382–1408. MR2279680

[5] BENAYCH-GEORGES, F., GUIONNET, A. and MAIDA, M. (2011). Fluctuations of the extreme eigenvalues of finite rank deformations of random matrices. *Electron. J. Probab.* **16** 1621–1662. MR2835249

[6] BENAYCH-GEORGES, F. and NADAKUDITI, R. R. (2011). The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Adv. Math.* **227** 494–521. MR2782201

[7] BERGER, J. (1982). Estimation in continuous exponential families: Bayesian estimation subject to risk restrictions and inadmissibility results. In *Statistical Decision Theory and Related*

MSC2010 subject classifications. Primary 62C20, 62H25; secondary 90C25, 90C22.

Key words and phrases. Covariance estimation, optimal shrinkage, Stein loss, entropy loss, divergence loss, Fréchet distance, Bhattacharya/Matusita affinity, condition number loss, high-dimensional ssymptotics, spiked covariance.

Topics, *III*, *Vol*. 1 (*West Lafayette*, *Ind*., 1981) 109–141. Academic Press, New York. MR0705285

[8] BHATIA, R. (1997). *Matrix Analysis*. *Graduate Texts in Mathematics* **169**. Springer, New York. MR1477662

[9] BROWN, L. D. and GREENSHTEIN, E. (2009). Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *Ann*. *Statist*. **37** 1685–1704. MR2533468

[10] CACOULLOS, T. and OLKIN, I. (1965). On the bias of functions of characteristic roots of a random matrix. *Biometrika* **52** 87–94. MR0207116

[11] CHEN, Y., WIESEL, A., ELDAR, Y. C. and HERO, A. O. (2010). Shrinkage algorithms for MMSE covariance estimation. *IEEE Trans*. *Signal Process*. **58** 5016–5029. MR2722661

[12] DANIELS, M. J. and KASS, R. E. (2001). Shrinkage estimators for covariance matrices. *Biometrics* **57** 1173–1184. MR1950425

[13] DEY, D. K. and SRINIVASAN, C. (1985). Estimation of a covariance matrix under Stein's loss. *Ann*. *Statist*. **13** 1581–1591. MR0811511

[14] DONOHO, D. and GAVISH, M. (2014). Minimax risk of matrix denoising by singular value thresholding. *Ann*. *Statist*. **42** 2413–2440. MR3269984

[15] DONOHO, D., GAVISH, M. and JOHNSTONE, I. (2018). Supplement to "Optimal shrinkage of eigenvalues in the spiked covariance model." DOI:10.1214/17-AOS1601SUPP.

[16] DONOHO, D. L., GAVISH, M. and JOHNSTONE, I. M. (2016). Code supplement to "Optimal shrinkage of eigenvalues in the spiked covariance model". Available at http://purl.stanford.edu/xy031gt1574.

[17] DONOHO, D. L. and JOHNSTONE, I. M. (1994). Minimax risk over l_p-balls for l_q-error. *Probab*. *Theory Related Fields* **99** 277–303. MR1278886

[18] DOWSON, D. C. and LANDAU, B. V. (1982). The Fréchet distance between multivariate normal distributions. *J*. *Multivariate Anal*. **12** 450–455. MR0666017

[19] 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

[20] EFRON, B. and MORRIS, C. (1976). Multivariate empirical Bayes and estimation of covariance matrices. *Ann*. *Statist*. **4** 22–32. MR0394960

[21] EL KAROUI, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann*. *Statist*. **36** 2717–2756. MR2485011

[22] EL KAROUI, N. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Ann*. *Statist*. **36** 2757–2790. MR2485012

[23] FAN, J., FAN, Y. and LV, J. (2008). High dimensional covariance matrix estimation using a factor model. *J*. *Econometrics* **147** 186–197. MR2472991

[24] FÖRSTNER, W. and MOONEN, B. (1999). A metric for covariance matrices. *Quo Vadis Geodesia* 113–128.

[25] GAVISH, M. and DONOHO, D. L. (2014). The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Trans*. *Inform*. *Theory* **60** 5040–5053. MR3245370

[26] GAVISH, M. and DONOHO, D. L. (2017). Optimal shrinkage of singular values. *IEEE Trans*. *Inform*. *Theory* **63** 2137–2152. MR3626861

[27] GEMAN, S. (1980). A limit theorem for the norm of random matrices. *Ann*. *Probab*. **8** 252–261. MR0566592

[28] GUPTA, A. K. and OFORI-NYARKO, S. (1995). Improved minimax estimators of normal covariance and precision matrices. *Statistics* **26** 19–25. MR1314180

[29] HAFF, L. R. (1979). Estimation of the inverse covariance matrix: Random mixtures of the inverse Wishart matrix and the identity. *Ann*. *Statist*. **7** 1264–1276. MR0550149

[30] HAFF, L. R. (1979). An identity for the Wishart distribution with applications. *J*. *Multivariate Anal*. **9** 531–544. MR0556910

[31] HAFF, L. R. (1980). Empirical Bayes estimation of the multivariate normal covariance matrix. *Ann. Statist.* **8** 586–597. MR0568722

[32] HUANG, J. Z., LIU, N., POURAHMADI, M. and LIU, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93** 85–98. MR2277742

[33] JAMES, A. T. (1954). Normal multivariate analysis and the orthogonal group. *Ann. Math. Stat.* **25** 40–75. MR0060779

[34] JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. In *Proc.* 4*th Berkeley Sympos. Math. Statist. and Prob., Vol. I* 361–379. Univ. California Press, Berkeley, CA. MR0133191

[35] JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. MR1863961

[36] JOHNSTONE, I. M. (2018). Tail sums of Wishart and GUE eigenvalues beyond the bulk edge. *Australian and New Zealand Journal of Statistics*. To appear. Available at ArXiv:1704.06398.

[37] KAILATH, T. (1967). The divergence and Bhattacharyya distance measures in signal selection. *IEEE Trans. Commun. Technol.* **15** 52–60.

[38] KONNO, Y. (1991). On estimation of a matrix of normal means with unknown covariance matrix. *J. Multivariate Anal.* **36** 44–55. MR1094267

[39] KRISHNAMOORTHY, K. and GUPTA, A. K. (1989). Improved minimax estimation of a normal precision matrix. *Canad. J. Statist.* **17** 91–102.

[40] KRISHNAMOORTHY, K. and GUPTA, A. K. (1989). Improved minimax estimation of a normal precision matrix. *Canad. J. Statist.* **17** 91–102. MR1014094

[41] KRITCHMAN, S. and NADLER, B. (2009). Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory. *IEEE Trans. Signal Process.* **57** 3930–3941. MR2683143

[42] KUBOKAWA, T. (1989). Improved estimation of a covariance matrix under quadratic loss. *Statist. Probab. Lett.* **8** 69–71. MR1006425

[43] KUBOKAWA, T. and KONNO, Y. (1990). Estimating the covariance matrix and the generalized variance under a symmetric loss. *Ann. Inst. Statist. Math.* **42** 331–343. MR1064792

[44] LEDOIT, O. and PÉCHÉ, S. (2011). Eigenvectors of some large sample covariance matrix ensembles. *Probab. Theory Related Fields* **151** 233–264. MR2834718

[45] LEDOIT, O. and WOLF, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.* **88** 365–411. MR2026339

[46] LEDOIT, O. and WOLF, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Ann. Statist.* **40** 1024–1060. MR2985942

[47] LEDOUX, M. (2001). *The Concentration of Measure Phenomenon. Mathematical Surveys and Monographs* **89**. Amer. Math. Soc., Providence, RI. MR1849347

[48] LENGLET, C., ROUSSON, M., DERICHE, R. and FAUGERAS, O. (2006). Statistics on the manifold of multivariate normal distributions: Theory and application to diffusion tensor MRI processing. *J. Math. Imaging Vision* **25** 423–444. MR2283616

[49] LIN, S. P. and PERLMAN, M. D. (1985). A Monte Carlo comparison of four estimators of a covariance matrix. In *Multivariate Analysis VI* (*Pittsburgh*, *PA*, 1983) 411–429. North-Holland, Amsterdam. MR0822310

[50] LOH, W.-L. (1991). Estimating covariance matrices. *Ann. Statist.* **19** 283–296. MR1091851

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

[52] MATUSITA, K. (1967). On the notion of affinity of several distributions and some of its applications. *Ann. Inst. Statist. Math.* **19** 181–192. MR0216649

[53] MUIRHEAD, R. J. (1987). Developments in eigenvalue estimation. In *Advances in Multivariate Statistical Analysis. Theory Decis. Lib. Ser. B*: *Math. Statist. Methods* 277–288. Reidel, Dordrecht. MR0920436

[54] NIST Digital Library of Mathematical Functions. Available at http://dlmf.nist.gov/, Release 1.0.9 of 2014-08-29. Online companion to [56].

[55] OLKIN, I. and PUKELSHEIM, F. (1982). The distance between two random vectors with given dispersion matrices. *Linear Algebra Appl*. **48** 257–263. MR0683223

[56] OLVER, F. W. J., LOZIER, D. W., BOISVERT, R. F. and CLARK, C. W., eds. (2010). *NIST Handbook of Mathematical Functions*. Cambridge University Press, New York, NY. Print companion to [54].

[57] PAL, N. (1993). Estimating the normal dispersion matrix and the precision matrix from a decision-theoretic point of view: A review. *Statist*. *Papers* **34** 1–26. MR1221520

[58] PASSEMIER, D. and YAO, J. (2013). Variance estimation and goodness-of-fit test in a high-dimensional strict factor model. Available at ArXiv:1308.3890.

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

[60] SELLIAH, J. B. (1964). *Estimation and Testing Problems in a Wishart Distribution*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)—Stanford Univ. MR2614178

[61] SHABALIN, A. A. and NOBEL, A. B. (2013). Reconstruction of a low-rank matrix in the presence of Gaussian noise. *J. Multivariate Anal*. **118** 67–76. MR3054091

[62] SHARMA, D. and KRISHNAMOORTHY, K. (1985). Empirical Bayes estimators of normal covariance matrix. *Sankhya*, *Ser. A* **47** 247–254. MR0844026

[63] SINHA, B. K. and GHOSH, M. (1987). Inadmissibility of the best equivariant estimators of the variance-covariance matrix, the precision matrix, and the generalized variance under entropy loss. *Statist*. *Decisions* **5** 201–227. MR0905238

[64] STEIN, C. (1956). Some problems in multivariate analysis. Technical Report, Department of Statistics, Stanford Univ., Available at http://statistics.stanford.edu/~ckirby/techreports/ONR/CHE%20ONR%2006.pdf.

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

[66] SUN, D. and SUN, X. (2005). Estimation of the multivariate normal precision and covariance matrices in a star-shape model. *Ann*. *Inst*. *Statist*. *Math*. **57** 455–484. MR2206534

[67] TRACY, C. A. and WIDOM, H. (1996). On orthogonal and symplectic matrix ensembles. *Comm*. *Math*. *Phys*. **177** 727–754. MR1385083

[68] VAN DER VAART, H. R. (1961). On certain characteristics of the distribution of the latent roots of asymmetric random matrix under general conditions. *Ann*. *Math*. *Stat*. **32** 864–873. MR0130749

[69] WON, J.-H., LIM, J., KIM, S.-J. and RAJARATNAM, B. (2013). Condition-number-regularized covariance estimation. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **75** 427–450. MR3065474

[70] YANG, R. and BERGER, J. O. (1994). Estimation of a covariance matrix using the reference prior. *Ann*. *Statist*. **22** 1195–1211. MR1311972

The Annals of Statistics
2018, Vol. 46, No. 4, 1779–1806
https://doi.org/10.1214/17-AOS1603

A BAYESIAN APPROACH TO THE SELECTION OF TWO-LEVEL MULTI-STRATUM FACTORIAL DESIGNS

BY MING-CHUNG CHANG AND CHING-SHUI CHENG

Academia Sinica and University of California, Berkeley

In a multi-stratum factorial experiment, there are multiple error terms (strata) with different variances that arise from complicated structures of the experimental units. For unstructured experimental units, minimum aberration is a popular criterion for choosing regular fractional factorial designs. One difficulty in extending this criterion to multi-stratum factorial designs is that the formulation of a word length pattern based on which minimum aberration is defined requires an order of desirability among the relevant words, but a natural order is often lacking. Furthermore, a criterion based only on word length patterns does not account for the different stratum variances. Mitchell, Morris and Ylvisaker [*Statist. Sinica* **5** (1995) 559–573] proposed a framework for Bayesian factorial designs. A Gaussian process is used as the prior for the treatment effects, from which a prior distribution of the factorial effects is induced. This approach is applied to study optimal and efficient multi-stratum factorial designs. Good surrogates for the Bayesian criteria that can be related to word length and generalized word length patterns for regular and nonregular designs, respectively, are derived. A tool is developed for eliminating inferior designs and reducing the designs that need to be considered without requiring any knowledge of stratum variances. Numerical examples are used to illustrate the theory in several settings.

REFERENCES

[1] AI, M., KANG, L. and JOSEPH, V. R. (2009). Bayesian optimal blocking of factorial designs. *J. Statist. Plann. Inference* **139** 3319–3328. MR2535203

[2] BAILEY, R. A. (2008). *Design of Comparative Experiments. Cambridge Series in Statistical and Probabilistic Mathematics* **25**. Cambridge Univ. Press, Cambridge. MR2422352

[3] CHEN, H. and CHENG, C.-S. (1999). Theory of optimal blocking of 2^{n-m} designs. *Ann. Statist.* **27** 1948–1973. MR1765624

[4] CHENG, C. S. (2014). *Theory of Factorial Design: Single- and Multi-Stratum Experiments*. CRC Press, Boca Raton, FL.

[5] CHENG, C.-S., STEINBERG, D. M. and SUN, D. X. (1999). Minimum aberration and model robustness for two-level fractional factorial designs. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **61** 85–93. MR1664104

[6] CHENG, C.-S. and TSAI, P.-W. (2009). Optimal two-level regular fractional factorial block and split-plot designs. *Biometrika* **96** 83–93. MR2482136

[7] CHENG, C.-S. and TSAI, P.-W. (2011). Multistratum fractional factorial designs. *Statist. Sinica* **21** 1001–1021. MR2817010

MSC2010 subject classifications. 62K05, 62K15.

Key words and phrases. A-criterion, block structure, D-criterion, Gaussian process, (M.S)-criterion, minimum aberration, word length pattern.

[8] CHENG, S.-W., LI, W. and YE, K. Q. (2004). Blocked nonregular two-level factorial designs. *Technometrics* **46** 269–279. MR2082497

[9] CHENG, S.-W. and WU, C. F. J. (2002). Choice of optimal blocking schemes in two-level and three-level designs. *Technometrics* **44** 269–277. MR1940091

[10] ECCLESTON, J. A. and HEDAYAT, A. (1974). On the theory of connected designs: Characterization and optimality. *Ann. Statist.* **2** 1238–1255. MR0362672

[11] FRIES, A. and HUNTER, W. G. (1980). Minimum aberration 2^{k-p} designs. *Technometrics* **22** 601–608. MR0596803

[12] JOSEPH, V. R. (2006). A Bayesian approach to the design and analysis of fractionated experiments. *Technometrics* **48** 219–229. MR2277676

[13] JOSEPH, V. R., AI, M. and WU, C. F. J. (2009). Bayesian-inspired minimum aberration two- and four-level designs. *Biometrika* **96** 95–106. MR2482137

[14] KANG, L. and JOSEPH, V. R. (2009). Bayesian optimal single arrays for robust parameter design. *Technometrics* **51** 250–261. MR2751071

[15] KERR, M. K. (2001). Bayesian optimal fractional factorials. *Statist. Sinica* **11** 605–630. MR1863153

[16] MITCHELL, T. J., MORRIS, M. D. and YLVISAKER, D. (1995). Two-level fractional factorials and Bayesian prediction. *Statist. Sinica* **5** 559–573. MR1347607

[17] SACKS, J., WELCH, W. J., MITCHELL, T. J. and WYNN, H. P. (1989). Design and analysis of computer experiments. *Statist. Sci.* **4** 409–435. MR1041765

[18] SUN, D. X. (1993). *Estimation Capacity and Related Topics in Experimental Designs*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)—University of Waterloo (Canada). MR2690433

[19] SUN, D. X., WU, C. F. J. and CHEN, Y. (1997). Optimal blocking schemes for 2^n and 2^{n-p} designs. *Technometrics* **39** 298–307. MR1462588

[20] TANG, B. and DENG, L.-Y. (1999). Minimum G_2-aberration for nonregular fractional factorial designs. *Ann. Statist.* **27** 1914–1926. MR1765622

The Annals of Statistics
2018, Vol. 46, No. 4, 1807–1836
https://doi.org/10.1214/17-AOS1604

ACCURACY ASSESSMENT FOR HIGH-DIMENSIONAL LINEAR REGRESSION[1]

BY T. TONY CAI AND ZIJIAN GUO

University of Pennsylvania and Rutgers University

This paper considers point and interval estimation of the ℓ_q loss of an estimator in high-dimensional linear regression with random design. We establish the minimax rate for estimating the ℓ_q loss and the minimax expected length of confidence intervals for the ℓ_q loss of rate-optimal estimators of the regression vector, including commonly used estimators such as Lasso, scaled Lasso, square-root Lasso and Dantzig Selector. Adaptivity of confidence intervals for the ℓ_q loss is also studied. Both the setting of the known identity design covariance matrix and known noise level and the setting of unknown design covariance matrix and unknown noise level are studied. The results reveal interesting and significant differences between estimating the ℓ_2 loss and ℓ_q loss with $1 \le q < 2$ as well as between the two settings.

New technical tools are developed to establish rate sharp lower bounds for the minimax estimation error and the expected length of minimax and adaptive confidence intervals for the ℓ_q loss. A significant difference between loss estimation and the traditional parameter estimation is that for loss estimation the constraint is on the performance of the estimator of the regression vector, but the lower bounds are on the difficulty of estimating its ℓ_q loss. The technical tools developed in this paper can also be of independent interest.

REFERENCES

[1] ARIAS-CASTRO, E., CANDÈS, E. J. and PLAN, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *Ann. Statist.* **39** 2533–2556. MR2906877

[2] BAYATI, M. and MONTANARI, A. (2012). The LASSO risk for Gaussian matrices. *IEEE Trans. Inform. Theory* **58** 1997–2017. MR2951312

[3] BELLONI, A., CHERNOZHUKOV, V. and WANG, L. (2011). Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika* **98** 791–806. MR2860324

[4] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

[5] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Heidelberg. MR2807761

[6] CAI, T. T. and GUO, Z. (2018). Supplement to "Accuracy assessment for high-dimensional linear regression." DOI:10.1214/17-AOS1604SUPP.

[7] CAI, T. T. and GUO, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Ann. Statist.* **45** 615–646. MR3650395

MSC2010 subject classifications. Primary 62G15; secondary 62C20, 62H35.

Key words and phrases. Accuracy assessment, adaptivity, confidence interval, high-dimensional linear regression, loss estimation, minimax lower bound, minimaxity, sparsity.

[8] CAI, T. T., LOW, M. and MA, Z. (2014). Adaptive confidence bands for nonparametric regression functions. *J. Amer. Statist. Assoc.* **109** 1054–1070. MR3265680

[9] CAI, T. T. and LOW, M. G. (2004). An adaptation theory for nonparametric confidence intervals. *Ann. Statist.* **32** 1805–1840. MR2102494

[10] CAI, T. T. and LOW, M. G. (2006). Adaptive confidence balls. *Ann. Statist.* **34** 202–228. MR2275240

[11] CAI, T. T. and ZHOU, H. H. (2009). A data-driven block thresholding approach to wavelet estimation. *Ann. Statist.* **37** 569–595. MR2502643

[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

[13] CHERNOZHUKOV, V., HANSEN, C. and SPINDLER, M. (2015). Post-selection and post-regularization inference in linear models with many controls and instruments. Preprint. Available at arXiv:1501.03185.

[14] CHERNOZHUKOV, V., HANSEN, C. and SPINDLER, M. (2015). Valid post-selection and post-regularization inference: An elementary, general approach. Preprint. Available at arXiv:1501.03430.

[15] DONOHO, D. L. and JOHNSTONE, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90** 1200–1224. MR1379464

[16] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2011). The noise-sensitivity phase transition in compressed sensing. *IEEE Trans. Inform. Theory* **57** 6920–6941. MR2882271

[17] GUO, Z., WANG, W., CAI, T. T. and LI, H. (2016). Optimal estimation of co-heritability in high-dimensional linear models. Preprint. Available at arXiv:1605.07244.

[18] HOFFMANN, M. and NICKL, R. (2011). On adaptive inference and confidence bands. *Ann. Statist.* **39** 2383–2409. MR2906872

[19] INGSTER, Y. I., TSYBAKOV, A. B. and VERZELEN, N. (2010). Detection boundary in sparse regression. *Electron. J. Stat.* **4** 1476–1526. MR2747131

[20] JANSON, L., BARBER, R. F. and CANDÈS, E. (2015). Eigenprism: Inference for high-dimensional signal-to-noise ratios. Preprint. Available at arXiv:1505.02097.

[21] LI, K.-C. (1985). From Stein's unbiased risk estimates to the method of generalized cross validation. *Ann. Statist.* **13** 1352–1377. MR0811497

[22] NICKL, R. and VAN DE GEER, S. (2013). Confidence sets in sparse regression. *Ann. Statist.* **41** 2852–2876. MR3161450

[23] RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory* **57** 6976–6994. MR2882274

[24] ROBINS, J. and VAN DER VAART, A. (2006). Adaptive nonparametric confidence sets. *Ann. Statist.* **34** 229–253. MR2275241

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

[26] SUN, T. and ZHANG, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898. MR2999166

[27] THRAMPOULIDIS, C., PANAHI, A. and HASSIBI, B. (2015). Asymptotically exact error analysis for the generalized ℓ_2^2-lasso. Preprint. Available at arXiv:1502.06287.

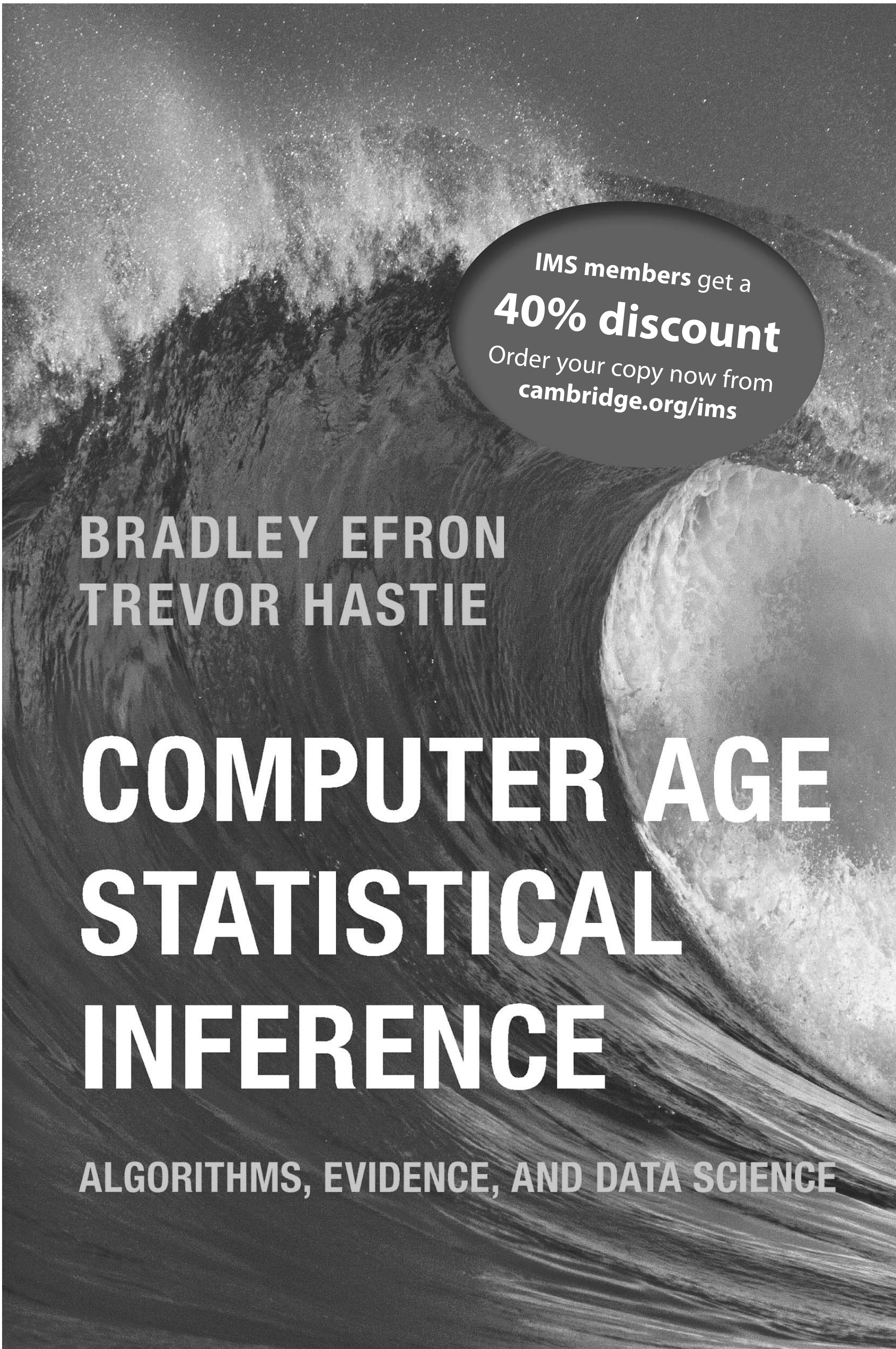
[28] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[29] 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

[30] VERZELEN, N. (2012). Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electron. J. Stat.* **6** 38–90. MR2879672

[31] YE, F. and ZHANG, C.-H. (2010). Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *J. Mach. Learn. Res.* **11** 3519–3540. MR2756192

[32] YI, F. and ZOU, H. (2013). SURE-tuned tapering estimation of large covariance matrices. *Comput. Statist. Data Anal.* **58** 339–351. MR2997947


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