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

THE ANNALS *of* STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles

Finding a large submatrix of a Gaussian random matrix . . . . DAVID GAMARNIK AND QUAN LI 2511
Support points . . . . SIMON MAK AND V. ROSHAN JOSEPH 2562
Debiasing the lasso: Optimal sample size for Gaussian designs ADEL JAVANMARD AND ANDREA MONTANARI 2593
Margins of discrete Bayesian networks . . . . ROBIN J. EVANS 2623
Multi-threshold accelerated failure time model . . . . JIALIANG LI AND BAISUO JIN 2657
Measuring and testing for interval quantile dependence LIPING ZHU, YAOWU ZHANG AND KAI XU 2683
Barycentric subspace analysis on manifolds . . . . XAVIER PENNEC 2711
The landscape of empirical risk for nonconvex losses SONG MEI, YU BAI AND ANDREA MONTANARI 2747
Designs with blocks of size two and applications to microarray experiments JANET GODOLPHIN 2775
Local robust estimation of the Pickands dependence function MIKAEL ESCOBAR-BACH, YURI GOEGEBEUR AND ARMELLE GUILLOU 2806
Strong identifiability and optimal minimax rates for finite mixture estimation PHILIPPE HEINRICH AND JONAS KAHN 2844
Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries STANISLAV MINSKER 2871
Causal inference in partially linear structural equation models DOMINIK ROTHENHÄUSLER, JAN ERNEST AND PETER BÜHLMANN 2904
On MSE-optimal crossover designs . . . . CHRISTOPH NEUMANN AND JOACHIM KUNERT 2939
Testing for periodicity in functional time series SIEGFRIED HÖRMANN, PIOTR KOKOSZKA AND GILLES NISOL 2960
Limiting behavior of eigenvalues in high-dimensional MANOVA via RMT ZHIDONG BAI, KWOK PUI CHOI AND YASUNORI FUJIKOSHI 2985
Two-sample Kolmogorov–Smirnov-type tests revisited: Old and new tests in terms of local levels . . . . HELMUT FINNER AND VERONIKA GONTSCHARUK 3014
Robust Gaussian stochastic process emulation MENGYANG GU, XIAOJING WANG AND JAMES O. BERGER 3038
Convergence of contrastive divergence algorithm in exponential family BAI JIANG, TUNG-YU WU, YIFAN JIN AND WING H. WONG 3067
Overcoming the limitations of phase transition by higher order analysis of regularization techniques . . . . HAOLEI WENG, ARIAN MALEKI AND LE ZHENG 3099
Optimal adaptive estimation of linear functionals under sparsity . . . . . OLIVIER COLLIER, LAËTITIA COMMINGES, ALEXANDRE B. TSYBAKOV AND NICOLAS VERZELEN 3130
High-dimensional consistency in score-based and hybrid structure learning PREETAM NANDY, ALAIN HAUSER AND MARLOES H. MAATHUIS 3151

Vol. 46, No. 6A—December 2018

THE ANNALS OF STATISTICS Vol. 46, No. 6A, pp. 2511–3183 December 2018

INSTITUTE OF MATHEMATICAL STATISTICS

(Organized September 12, 1935)

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

The Annals of Statistics **[ISSN 0090-5364 (print); ISSN 2168-8966 (online)],** Volume 46, Number 6A, December 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. 6A, 2511–2561
https://doi.org/10.1214/17-AOS1628

FINDING A LARGE SUBMATRIX OF A GAUSSIAN RANDOM MATRIX

BY DAVID GAMARNIK[1] AND QUAN LI

Massachusetts Institute of Technology

We consider the problem of finding a $k \times k$ submatrix of an $n \times n$ matrix with i.i.d. standard Gaussian entries, which has a large average entry. It was shown in [Bhamidi, Dey and Nobel (2012)] using nonconstructive methods that the largest average value of a $k \times k$ submatrix is $2(1 + o(1))\sqrt{\log n/k}$, with high probability (w.h.p.), when $k = O(\log n/\log\log n)$. In the same paper, evidence was provided that a natural greedy algorithm called the Largest Average Submatrix ($\mathcal{LAS}$) for a constant k should produce a matrix with average entry at most $(1 + o(1))\sqrt{2\log n/k}$, namely approximately $\sqrt{2}$ smaller than the global optimum, though no formal proof of this fact was provided.

In this paper, we show that the average entry of the matrix produced by the $\mathcal{LAS}$ algorithm is indeed $(1+o(1))\sqrt{2\log n/k}$ w.h.p. when k is constant and n grows. Then, by drawing an analogy with the problem of finding cliques in random graphs, we propose a simple greedy algorithm which produces a $k \times k$ matrix with asymptotically the same average value $(1 + o(1))\sqrt{2\log n/k}$ w.h.p., for $k = o(\log n)$. Since the greedy algorithm is the best known algorithm for finding cliques in random graphs, it is tempting to believe that beating the factor $\sqrt{2}$ performance gap suffered by both algorithms might be very challenging. Surprisingly, we construct a very simple algorithm which produces a $k \times k$ matrix with average value $(1 + o_k(1) + o(1))(4/3)\sqrt{2\log n/k}$ for $k = o((\log n)^{1.5})$, that is, with the asymptotic factor $4/3$ when k grows.

To get an insight into the algorithmic hardness of this problem, and motivated by methods originating in the theory of spin glasses, we conduct the so-called expected overlap analysis of matrices with average value asymptotically $(1+o(1))\alpha\sqrt{2\log n/k}$ for a fixed value $\alpha \in [1, \sqrt{2}]$. The overlap corresponds to the number of common rows and the number of common columns for pairs of matrices achieving this value (see the paper for details). We discover numerically an intriguing phase transition at $\alpha^* \triangleq 5\sqrt{2}/(3\sqrt{3}) \approx 1.3608\ldots \in [4/3, \sqrt{2}]$: when $\alpha < \alpha^*$ the space of overlaps is a continuous subset of $[0, 1]^2$, whereas $\alpha = \alpha^*$ marks the onset of discontinuity, and as a result the model exhibits the *Overlap Gap Property (OGP)* when $\alpha > \alpha^*$, appropriately defined. We conjecture that the *OGP* observed for $\alpha > \alpha^*$ also marks the onset of the algorithmic hardness—no polynomial time algorithm exists for finding matrices with average value at least $(1+o(1))\alpha\sqrt{2\log n/k}$, when $\alpha > \alpha^*$ and k is a mildly growing function of n.

MSC2010 subject classifications. 68Q87, 97K50, 60C05, 68Q25.
Key words and phrases. Random matrix, random graphs, maximum clique, submatrix detection, computational complexity, overlap gap property.

REFERENCES

[1] ACHLIOPTAS, D. and COJA-OGHLAN, A. (2008). Algorithmic barriers from phase transitions. In 2008 49*th Annual IEEE Symposium on Foundations of Computer Science* 793–802. IEEE, New York.

[2] ACHLIOPTAS, D., COJA-OGHLAN, A. and RICCI-TERSENGHI, F. (2011). On the solution-space geometry of random constraint satisfaction problems. *Random Structures Algorithms* **38** 251–268. MR2663730

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

[4] BERTHET, Q. and RIGOLLET, P. (2013). Complexity theoretic lower bounds for sparse principal component detection. In *Conference on Learning Theory* 1046–1066.

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

[6] BHAMIDI, S., DEY, P. S. and NOBEL, A. B. (2012). Energy landscape for large average submatrix detection problems in Gaussian random matrices. Preprint. Available at arXiv:1211.2284.

[7] COJA-OGHLAN, A. and EFTHYMIOU, C. (2011). On independent sets in random graphs. In *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms* 136–144. SIAM, Philadelphia.

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

[9] GAMARNIK, D. and SUDAN, M. (2014). Limits of local algorithms over sparse random graphs. In *Proceedings of the* 5*th Conference on Innovations in Theoretical Computer Science* 369–376. ACM, New York.

[10] GAMARNIK, D. and SUDAN, M. (2014). Performance of the survey propagation-guided decimation algorithm for the random NAE-K-SAT problem. Preprint. Available at arXiv:1402.0052.

[11] GAMARNIK, D. and ZADIK, I. (2017). High-dimensional regression with binary coefficients. Estimating squared error and a phase transition. Preprint. Available at arXiv:1701.04455.

[12] KARP, R. M. (1976). The probabilistic analysis of some combinatorial search algorithms. In *Algorithms and complexity*: *New directions and recent results* 1–19. MR0445898

[13] LEADBETTER, M. R., LINDGREN, G. and ROOTZÉN, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer, New York. MR0691492

[14] MADEIRA, S. C. and OLIVEIRA, A. L. (2004). Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **1** 24–45.

[15] MONTANARI, A. (2015). Finding one community in a sparse graph. *J. Stat. Phys.* **161** 273–299. MR3401018

[16] RAHMAN, M. and VIRÁG, B. (2014). Local algorithms for independent sets are half-optimal. Preprint. Available at arXiv:1402.0485.

[17] SHABALIN, A. A., WEIGMAN, V. J., PEROU, C. M. and NOBEL, A. B. (2009). Finding large average submatrices in high dimensional data. *Ann. Appl. Stat.* 985–1012. MR2750383

[18] SUN, X. and NOBEL, A. B. (2013). On the maximal size of large-average and ANOVA-fit submatrices in a Gaussian random matrix. *Bernoulli* **19** 275–294. MR3019495

The Annals of Statistics
2018, Vol. 46, No. 6A, 2562–2592
https://doi.org/10.1214/17-AOS1629

SUPPORT POINTS[1]

BY SIMON MAK AND V. ROSHAN JOSEPH

Georgia Institute of Technology

This paper introduces a new way to compact a continuous probability distribution F into a set of representative points called support points. These points are obtained by minimizing the energy distance, a statistical potential measure initially proposed by Székely and Rizzo [*InterStat* **5** (2004) 1–6] for testing goodness-of-fit. The energy distance has two appealing features. First, its distance-based structure allows us to exploit the duality between powers of the Euclidean distance and its Fourier transform for theoretical analysis. Using this duality, we show that support points converge in distribution to F, and enjoy an improved error rate to Monte Carlo for integrating a large class of functions. Second, the minimization of the energy distance can be formulated as a difference-of-convex program, which we manipulate using two algorithms to efficiently generate representative point sets. In simulation studies, support points provide improved integration performance to both Monte Carlo and a specific quasi-Monte Carlo method. Two important applications of support points are then highlighted: (a) as a way to quantify the propagation of uncertainty in expensive simulations and (b) as a method to optimally compact Markov chain Monte Carlo (MCMC) samples in Bayesian computation.

REFERENCES

[1] ASCHER, U. M. and GREIF, C. (2011). *A First Course in Numerical Methods*. *Computational Science & Engineering* **7**. SIAM, Philadelphia, PA. MR2839122

[2] BAHOURI, H., CHEMIN, J.-Y. and DANCHIN, R. (2011). *Fourier Analysis and Nonlinear Partial Differential Equations* **343**. Springer, Heidelberg. MR2768550

[3] BORODACHOV, S. V., HARDIN, D. P. and SAFF, E. B. (2014). Low complexity methods for discretizing manifolds via Riesz energy minimization. *Found. Comput. Math.* **14** 1173–1208.

[4] BOUSQUET, O. and BOTTOU, L. (2008). The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems* 161–168.

[5] CARPENTER, B., GELMAN, A., HOFFMAN, M., LEE, D., GOODRICH, B., BETANCOURT, M., BRUBAKER, M. A., GUO, J., LI, P. and RIDDELL, A. (2017). Stan: A probabilistic programming language. *J. Stat. Softw.* **76** 1–32.

[6] COX, D. R. (1957). Note on grouping. *J. Amer. Statist. Assoc.* **52** 543–547.

[7] DALENIUS, T. (1950). The problem of optimum stratification. *Scand. Actuar. J.* **1950** 203–213.

[8] DICK, J., KUO, F. Y. and SLOAN, I. H. (2013). High-dimensional integration: The quasi-Monte Carlo way. *Acta Numer.* **22** 133–288.

MSC2010 subject classifications. 62E17.
Key words and phrases. Bayesian computation, energy distance, Monte Carlo, numerical integration, quasi-Monte Carlo, representative points.

[9] DICK, J. and PILLICHSHAMMER, F. (2010). *Digital Nets and Sequences*: *Discrepancy Theory and Quasi-Monte Carlo Integration*. Cambridge Univ. Press, Cambridge. MR2683394

[10] DI NEZZA, E., PALATUCCI, G. and VALDINOCI, E. (2012). Hitchhiker's guide to the fractional Sobolev spaces. *Bulletin des Sciences Mathématiques* **136** 521–573.

[11] DRAPER, N. R. and SMITH, H. (1981). *Applied Regression Analysis*, 2nd ed. Wiley, New York. MR0610978

[12] DUTANG, C. and SAVICKY, P. (2013). `randtoolbox`: Generating and testing random numbers. R package.

[13] FANG, K.-T. (1980). The uniform design: Application of number-theoretic methods in experimental design. *Acta Math*. *Appl*. *Sin*. **3** 363–372.

[14] FANG, K.-T., LU, X., TANG, Y. and YIN, J. (2004). Constructions of uniform designs by using resolvable packings and coverings. *Discrete Math*. **274** 25–40.

[15] FANG, K.-T. and WANG, Y. (1994). *Number-Theoretic Methods in Statistics*. *Monographs on Statistics and Applied Probability* **51**. Chapman & Hall, London. MR1284470

[16] FLURY, B. A. (1990). Principal points. *Biometrika* **77** 33–41. MR1049406

[17] GELFAND, I. and SHILOV, G. (1964). *Generalized Functions*, *Vol*. *I*: *Properties and Operations*. Academic Press, New York.

[18] GENZ, A. (1984). Testing multidimensional integration routines. In *Proc*. *of International Conference on Tools*, *Methods and Languages for Scientific and Engineering Computation* 81–94. Elsevier, North-Holland.

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

[20] GHADIMI, S. and LAN, G. (2013). Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM J*. *Optim*. **23** 2341–2368. MR3134439

[21] GIROLAMI, M. and CALDERHEAD, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J*. *R*. *Stat*. *Soc*. *Ser. B*. *Stat*. *Methodol*. **73** 123–214. MR2814492

[22] GRAF, S. and LUSCHGY, H. (2000). *Foundations of Quantization for Probability Distributions*. Springer, Berlin.

[23] HICKERNELL, F. J. (1998). A generalized discrepancy and quadrature error bound. *Math*. *Comp*. **67** 299–322. MR1433265

[24] HICKERNELL, F. J. (1999). Goodness-of-fit statistics, discrepancies and robust designs. *Statist*. *Probab. Lett*. **44** 73–78. MR1706366

[25] HUNTER, J. K. and NACHTERGAELE, B. (2001). *Applied Analysis*. World Scientific, Singapore.

[26] JOE, S. and KUO, F. Y. (2003). Remark on algorithm 659: Implementing Sobol's quasirandom sequence generator. *ACM Trans*. *Math*. *Software* **29** 49–57.

[27] JOSEPH, V. R., DASGUPTA, T., TUO, R. and WU, C. F. J. (2015). Sequential exploration of complex surfaces using minimum energy designs. *Technometrics* **57** 64–74. MR3318350

[28] JOSEPH, V. R., GUL, E. and BA, S. (2015). Maximum projection designs for computer experiments. *Biometrika* **102** 371–380.

[29] KIEFER, J. (1961). On large deviations of the empiric df of vector chance variables and a law of the iterated logarithm. *Pacific J*. *Math*. **11** 649–660.

[30] KOLMOGOROV, A. (1933). Sulla determinazione empirica delle leggi di probabilita. *G*. *Ist*. *Ital*. *Attuari* **4** 1–11.

[31] KOROLYUK, V. and BOROVSKIKH, Y. V. (1989). Convergence rate for degenerate von Mises functionals. *Theory Probab*. *Appl*. **33** 125–135.

[32] KUO, F. Y. and SLOAN, I. H. (2005). Lifting the curse of dimensionality. *Notices Amer. Math*. *Soc*. **52** 1320–1328.

[33] LANGE, K. (2016). *MM Optimization Algorithms*. SIAM, Philadelphia.

[34] LINK, W. A. and EATON, M. J. (2012). On thinning of chains in MCMC. *Methods Ecol*. *Evol*. **3** 112–115.

[35] LIPP, T. and BOYD, S. (2016). Variations and extension of the convex-concave procedure. *Optim. Eng.* **17** 263–287. MR3511445
[36] LLOYD, S. (1982). Least squares quantization in PCM. *IEEE Trans. Inform. Theory* **28** 129–137.
[37] MAIRAL, J. (2013). Stochastic majorization-minimization algorithms for large-scale optimization. In *Advances in Neural Information Processing Systems* 2283–2291.
[38] MAK, S. (2017). `support`: Support Points. R package version 0.1.0.
[39] MAK, S. and JOSEPH, V. R. (2018). Minimax and minimax projection designs using clustering. *J. Comput. Graph. Statist.* **27** 166-178. MR3788310
[40] MAK, S. and JOSEPH, V. R (2018). Supplement to "Support points." DOI:10.1214/17-AOS1629SUPP.
[41] MAK, S., SUNG, C.-L., WANG, X., YEH, S.-T., CHANG, Y.-H., JOSEPH, V. R., YANG, V. and WU, C. F. J. (2017). An efficient surrogate model for emulation and physics extraction of large eddy simulations. Preprint. Available at arXiv:1611.07911.
[42] MAK, S. and XIE, Y. (2017). Uncertainty quantification and design for noisy matrix completion-a unified framework. Preprint. Available at arXiv:1706.08037.
[43] MATSUMOTO, M. and NISHIMURA, T. (1998). Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans. Model. Comput. Simul.* **8** 3–30.
[44] NICHOLS, J. A. and KUO, F. Y. (2014). Fast CBC construction of randomly shifted lattice rules achieving $\mathcal{O}(n^{-1+\delta})$ convergence for unbounded integrands over $\mathbb{R}^s$ in weighted spaces with POD weights. *J. Complexity* **30** 444–468.
[45] NIEDERREITER, H. (1992). *Random Number Generation and Quasi-Monte Carlo Methods*. *CBMS-NSF Regional Conference Series in Applied Mathematics* **63**. SIAM, Philadelphia, PA. MR1172997
[46] NUYENS, D. and COOLS, R. (2006). Fast algorithms for component-by-component construction of rank-1 lattice rules in shift-invariant reproducing kernel Hilbert spaces. *Math. Comp.* **75** 903–920. MR2196999
[47] ORTEGA, J. M. and RHEINBOLDT, W. C. (2000). *Iterative Solution of Nonlinear Equations in Several Variables*. SIAM, Philadelphia.
[48] OWEN, A. B. (1998). Scrambling Sobol' and Niederreiter–Xing points. *J. Complexity* **14** 466–489. MR1659008
[49] OWEN, A. B. and TRIBBLE, S. D. (2005). A quasi-Monte Carlo Metropolis algorithm. *Proc. Natl. Acad. Sci. USA* **102** 8844–8849.
[50] PAGÈS, G., PHAM, H. and PRINTEMS, J. (2004). Optimal quantization methods and applications to numerical problems in finance. In *Handbook of Computational and Numerical Methods in Finance* 253–297. Birkhäuser, Boston, MA. MR2083055
[51] PALEY, R. and ZYGMUND, A. (1930). On some series of functions. In *In Mathematical Proceedings of the Cambridge Philosophical Society* **26** 337–357. Cambridge Univ. Press, Cambridge.
[52] R CORE TEAM (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
[53] RESNICK, S. I. (2014). *A Probability Path*. Birkhäuser/Springer, New York. MR3135152
[54] ROSENBLATT, M. (1952). Remarks on a multivariate transformation. *Ann. Math. Stat.* **23** 470–472. MR0049525
[55] ROYDEN, H. L. and FITZPATRICK, P. (2010). *Real Analysis*. Macmillan, New York.
[56] SANTNER, T. J., WILLIAMS, B. J. and NOTZ, W. I. (2013). *The Design and Analysis of Computer Experiments*. Springer, New York. MR2160708
[57] SERFLING, R. J. (2009). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

[58] SHORACK, G. R. (2000). *Probability for Statisticians*. Springer, New York. MR1762415

[59] SLOAN, I. H., KUO, F. Y. and JOE, S. (2002). Constructing randomly shifted lattice rules in weighted Sobolev spaces. *SIAM J. Numer. Anal*. **40** 1650–1665.

[60] SOBOL', I. M. (1967). Distribution of points in a cube and approximate evaluation of integrals. *Ž. Vyčisl. Mat. i Mat. Fiz.* **7** 784–802. MR0219238

[61] SU, Y. (2000). Asymptotically optimal representative points of bivariate random vectors. *Statist. Sinica* **10** 559–576.

[62] SZÉKELY, G. J. (2003). E-statistics: The energy of statistical samples Technical Report 03-05, Dept. Mathematics and Statistics, Bowling Green State Univ, Bowling Green, OH.

[63] SZÉKELY, G. J. and RIZZO, M. L. (2004). Testing for equal distributions in high dimension. *InterStat* **5** 1–6.

[64] SZÉKELY, G. J. and RIZZO, M. L. (2013). Energy statistics: A class of statistics based on distances. *J. Statist. Plann. Inference* **143** 1249–1272.

[65] TAO, P. D. and AN, L. T. H. (1997). Convex analysis approach to DC programming: Theory, algorithms and applications. *Acta Math. Vietnam*. **22** 289–355.

[66] TUY, H. (1986). A general deterministic approach to global optimization via dc programming. In *FERMAT Days* 85: *Mathematics for Optimization* (J. B. Hiriart-Urruty, ed.). *North-Holland Mathematics Studies* **129** 137–162. North-Holland, Amsterdam. MR0874358

[67] TUY, H. (1995). DC optimization: Theory, methods and algorithms. In *Handbook of Global Optimization* 149–216. Springer, Berlin.

[68] WENDLAND, H. (2005). *Scattered Data Approximation*. Cambridge Univ. Press, Cambridge.

[69] WORLEY, B. A. (1987). Deterministic uncertainty analysis. Technical Report ORNL-6428, Oak Ridge National Laboratory, Oak Ridge, TN.

[70] YANG, G. (2012). The energy goodness-of-fit test for univariate stable distributions. Ph.D. thesis, Bowling Green State Univ., Bowling Green, OH. MR3093953

[71] YUILLE, A. L. and RANGARAJAN, A. (2003). The concave-convex procedure. *Neural Comput*. **15** 915–936.

[72] ZADOR, P. (1982). Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Trans. Inform. Theory* **28** 139–149.

The Annals of Statistics
2018, Vol. 46, No. 6A, 2593–2622
https://doi.org/10.1214/17-AOS1630

DEBIASING THE LASSO: OPTIMAL SAMPLE SIZE FOR GAUSSIAN DESIGNS

BY ADEL JAVANMARD[1] AND ANDREA MONTANARI[2]

University of Southern California and Stanford University

Performing statistical inference in high-dimensional models is challenging because of the lack of precise information on the distribution of high-dimensional regularized estimators.

Here, we consider linear regression in the high-dimensional regime $p \gg n$ and the Lasso estimator: we would like to perform inference on the parameter vector $\theta^* \in \mathbb{R}^p$. Important progress has been achieved in computing confidence intervals and p-values for single coordinates θ_i^*, $i \in \{1, \dots, p\}$. A key role in these new inferential methods is played by a certain debiased estimator $\widehat{\theta}^{\mathrm{d}}$. Earlier work establishes that, under suitable assumptions on the design matrix, the coordinates of $\widehat{\theta}^{\mathrm{d}}$ are asymptotically Gaussian provided the true parameters vector θ^* is s_0-sparse with $s_0 = o(\sqrt{n}/\log p)$.

The condition $s_0 = o(\sqrt{n}/\log p)$ is considerably stronger than the one for consistent estimation, namely $s_0 = o(n/\log p)$. In this paper, we consider Gaussian designs with known or unknown population covariance. When the covariance is known, we prove that the debiased estimator is asymptotically Gaussian under the nearly optimal condition $s_0 = o(n/(\log p)^2)$.

The same conclusion holds if the population covariance is unknown but can be estimated sufficiently well. For intermediate regimes, we describe the trade-off between sparsity in the coefficients θ^*, and sparsity in the inverse covariance of the design. We further discuss several applications of our results beyond high-dimensional inference. In particular, we propose a thresholded Lasso estimator that is minimax optimal up to a factor $1 + o_n(1)$ for i.i.d. Gaussian designs.

REFERENCES

[1] AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **AC-19** 716–723. System identification and time-series analysis. MR0423716

[2] BARBER, R. F. and CANDÈS, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.* **43** 2055–2085. MR3375876

[3] BAYATI, M., ERDOGDU, M. A. and MONTANARI, A. (2013). Estimating lasso risk and noise level. In *Advances in Neural Information Processing Systems* 944–952.

[4] BAYATI, M., LELARGE, M. and MONTANARI, A. (2015). Universality in polytope phase transitions and message passing algorithms. *Ann. Appl. Probab.* **25** 753–822. MR3313755

[5] BAYATI, M. and MONTANARI, A. (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. Inform. Theory* **57** 764–785. MR2810285

MSC2010 subject classifications. Primary 62J05, 62J07; secondary 62F12.

Key words and phrases. Lasso, high-dimensional regression, confidence intervals, hypothesis testing, bias and variance, sample size.

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

[7] BELLONI, A. and CHERNOZHUKOV, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19** 521–547.

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

[9] BÜHLMANN, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* **19** 1212–1242.

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

[11] BÜHLMANN, P. and VAN DE GEER, S. (2015). High-dimensional inference in misspecified linear models. *Electron. J. Stat.* **9** 1449–1473. MR3367666

[12] CAI, T. T. and GUO, Z. (2016). Accuracy assessment for high-dimensional linear regression. Available at arXiv:1603.03474.

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

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

[15] CANDÈS, E. J., ROMBERG, J. K. and TAO, T. (2006). Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.* **59** 1207–1223. MR2230846

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

[17] CHEN, M., REN, Z., ZHAO, H. and ZHOU, H. (2016). Asymptotically normal and efficient estimation of covariate-adjusted Gaussian graphical model. *J. Amer. Statist. Assoc.* **111** 394–406.

[18] CHEN, S. S. and DONOHO, D. L. (1995). Examples of basis pursuit. In *Proceedings of Wavelet Applications in Signal and Image Processing III*.

[19] CHERNOZHUKOV, V., HANSEN, C. and SPINDLER, M. (2015). Valid post-selection and post-regularization inference: An elementary, general approach. *Ann. Rev. Econ.* **7** 649–688.

[20] DEZEURE, R., BÜHLMANN, P., MEIER, L., MEINSHAUSEN, N. et al. (2015). High-dimensional inference: Confidence intervals, p-values and R-software hdi. *Statist. Sci.* **30** 533–558. MR3432840

[21] DICKER, L. H. (2012). Residual variance and the signal-to-noise ratio in high-dimensional linear models. Available at arXiv:1209.0012.

[22] DONOHO, D. L., ELAD, M. and TEMLYAKOV, V. N. (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory* **52** 6–18. MR2237332

[23] DONOHO, D. L. and HUO, X. (2001). Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Theory* **47** 2845–2862. MR1872845

[24] DONOHO, D. L., JOHNSTONE, I. and MONTANARI, A. (2013). Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising. *IEEE Trans. Inform. Theory* **59** 3396–3433. MR3061255

[25] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2009). Message passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci. USA* **106** 18914–18919.

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

[27] EFRON, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation. *J. Amer. Statist. Assoc.* **99** 619–642. MR2090899

[28] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360.

[29] FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911.

[30] FAN, J., SAMWORTH, R. and WU, Y. (2009). Ultrahigh dimensional feature selection: Beyond the linear model. *J. Mach. Learn. Res.* **10** 2013–2038.

[31] FITHIAN, W., SUN, D. and TAYLOR, J. (2014). Optimal inference after model selection. Available at arXiv:1410.2597.

[32] JANKOVÁ, J. and VAN DE GEER, S. (2015). Confidence intervals for high-dimensional inverse covariance estimation. *Electron. J. Stat.* **9** 1205–1229. MR3354336

[33] JANKOVÁ, J. and VAN DE GEER, S. (2017). Honest confidence regions and optimality in high-dimensional precision matrix estimation. *TEST* **26** 143–162.

[34] JANSON, L., FOYGEL BARBER, R. and CANDÈS, E. (2017). EigenPrism: Inference for high dimensional signal-to-noise ratios. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1037–1065. MR3689308

[35] JANSON, L., SU, W. et al. (2016). Familywise error rate control via knockoffs. *Electron. J. Stat.* **10** 960–975.

[36] JAVANMARD, A. and LEE, J. D. (2017). A Flexible Framework for Hypothesis Testing in High-dimensions. Available at arXiv:1704.07971.

[37] JAVANMARD, A. and MONTANARI, A. (2013). Nearly optimal sample size in hypothesis testing for high-dimensional regression. In 51*st Annual Allerton Conference* 1427–1434.

[38] JAVANMARD, A. and MONTANARI, A. (2014). Hypothesis testing in high-dimensional regression under the Gaussian random design model: Asymptotic theory. *IEEE Trans. Inform. Theory* **60** 6522–6554. MR3265038

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

[40] JAVANMARD, A. and MONTANARI, A. (2018). Supplement to "Debiasing the Lasso: Optimal sample size for Gaussian designs." DOI:10.1214/17-AOS1630SUPP.

[41] LOCKHART, R., TAYLOR, J., TIBSHIRANI, R. J. and TIBSHIRANI, R. (2014). A significance test for the Lasso. *Ann. Statist.* **42** 413–468. MR3210970

[42] MALLOWS, C. L. (1973). Some comments on C_p. *Technometrics* **15** 661–675.

[43] MEINSHAUSEN, N. (2014). Group bound: Confidence intervals for groups of variables in sparse high dimensional regression without assumptions on the design. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 923–945. MR3414134

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

[45] MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 417–473. MR2758523

[46] OBUCHI, T. and KABASHIMA, Y. (2016). Cross validation in LASSO and its acceleration. *J. Stat. Mech. Theory Exp.* **2016** 53304–53339.

[47] REID, S., TIBSHIRANI, R. and FRIEDMAN, J. (2013). A study of error variance estimation in Lasso regression. Available at arXiv:1311.5274.

[48] RUDELSON, M. and ZHOU, S. (2013). Reconstruction from anisotropic random measurements. *IEEE Trans. Inform. Theory* **59** 3434–3447. MR3061256

[49] STÄDLER, N., BÜHLMANN, P. and VAN DE GEER, S. (2010). ℓ_1-penalization for mixture regression models. *TEST* **19** 209–256. MR2677722

[50] SU, W., CANDES, E. (2016). SLOPE is adaptive to unknown sparsity and asymptotically minimax. *Ann. Statist.* **44** 1038–1068. MR3485953

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

[52] TAYLOR, J., LOCKHART, R., TIBSHIRANI, R. J. and TIBSHIRANI, R. (2014). Exact post-selection inference for forward stepwise and least angle regression. Available at arXiv:1401.3889.

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

[54] 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

[55] WASSERMAN, L. and ROEDER, K. (2009). High-dimensional variable selection. *Ann. Statist.* **37** 2178–2201. MR2543689

[56] 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

[57] ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701

[58] ZHANG, C.-H. and HUANG, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. MR2435448

[59] ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242.

[60] ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2007). On the "degrees of freedom" of the lasso. *Ann. Statist.* **35** 2173–2192. MR2363967

The Annals of Statistics
2018, Vol. 46, No. 6A, 2623–2656
https://doi.org/10.1214/17-AOS1631

MARGINS OF DISCRETE BAYESIAN NETWORKS

BY ROBIN J. EVANS

University of Oxford

Bayesian network models with latent variables are widely used in statistics and machine learning. In this paper, we provide a complete algebraic characterization of these models when the observed variables are discrete and no assumption is made about the state-space of the latent variables. We show that it is algebraically equivalent to the so-called nested Markov model, meaning that the two are the same up to inequality constraints on the joint probabilities. In particular, these two models have the same dimension, differing only by inequality constraints for which there is no general description. The nested Markov model is therefore the closest possible description of the latent variable model that avoids consideration of inequalities. A consequence of this is that the constraint finding algorithm of Tian and Pearl [In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence* (2002) 519–527] is complete for finding equality constraints.

Latent variable models suffer from difficulties of unidentifiable parameters and nonregular asymptotics; in contrast the nested Markov model is fully identifiable, represents a curved exponential family of known dimension, and can easily be fitted using an explicit parameterization.

REFERENCES

ČENCOV, N. N. (1982). *Statistical Decision Rules and Optimal Inference*. *Translations of Mathematical Monographs* **53**. Amer. Math. Soc., Providence, RI. Translation from the Russian edited by Lev J. Leifman. MR0645898

ALLMAN, E. S., MATIAS, C. and RHODES, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.* **37** 3099–3132. MR2549554

ANANDKUMAR, A., HSU, D., JAVANMARD, A. and KAKADE, S. (2013). Learning linear Bayesian networks with latent variables. In *Proceedings of the* 30*th International Conference on Machine Learning* **28** 249–257.

BASU, S., POLLACK, R. and ROY, M.-F. (1996). *Algorithms in Real Algebraic Geometry*, Springer, Berlin.

BISHOP, C. M. (2007). *Pattern Recognition and Machine Learning*. Springer, Berlin.

COX, D., LITTLE, J. and O'SHEA, D. (2007). *Ideals*, *Varieties*, *and Algorithms*: *An Introduction to Computational Algebraic Geometry and Commutative Algebra*, 3rd ed. Springer, New York. MR2290010

DARWICHE, A. (2009). *Modeling and Reasoning with Bayesian Networks*. Cambridge Univ. Press, Cambridge.

DAWID, A. P. (2002). Influence diagrams for causal modelling and inference. *Int. Stat. Rev.* **70** 161–189.

MSC2010 subject classifications. 62H99, 62F12.

Key words and phrases. Algebraic statistics, Bayesian network, latent variable model, nested Markov model, Verma constraint.

DRTON, M. (2009). Likelihood ratio tests and singularities. *Ann. Statist.* **37** 979–1012. MR2502658
EVANS, R. J. (2016). Graphs for margins of Bayesian networks. *Scand. J. Stat.* **43** 625–648.
EVANS, R. J. (2018). Supplement to "Margins of discrete Bayesian networks." DOI:10.1214/17-AOS1631SUPP.
EVANS, R. J. and RICHARDSON, T. S. (2010). Maximum likelihood fitting of acyclic directed mixed graphs to binary data. In *Proceedings of the* 26*th Conference on Uncertainty in Artificial Intelligence* 177–184.
EVANS, R. J. and RICHARDSON, T. S. (2014). Markovian acyclic directed mixed graphs for discrete data. *Ann. Statist.* **42** 1452–1482. MR3262457
EVANS, R. J. and RICHARDSON, T. S. (2018). Smooth identifiable supermodels of discrete DAG models with latent variables. *Bernoulli*. To appear. Available at arXiv:1511.06813.
FOX, C. J., KÄUFL, A. and DRTON, M. (2015). On the causal interpretation of acyclic mixed graphs under multivariate normality. *Linear Algebra Appl.* **473** 93–113. MR3338327
GARCIA, L. D., STILLMAN, M. and STURMFELS, B. (2005). Algebraic geometry of Bayesian networks. *J. Symbolic Comput.* **39** 331–355. MR2168286
GILL, R. D. (2014). Statistics, causality and Bell's theorem. *Statist. Sci.* **29** 512–528.
HENSON, J., LAL, R. and PUSEY, M. F. (2014). Theory-independent limits on correlations from generalized Bayesian networks. *New J. Phys.* **16** 113043.
KASS, R. E. and VOS, P. W. (1997). *Geometrical Foundations of Asymptotic Inference*. Wiley, New York. MR1461540
LAURITZEN, S. L. (1996). *Graphical Models*. *Oxford Statistical Science Series* **17**. Oxford Univ. Press, New York. MR1419991
NEYMAN, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Rocz. Nauk Rol.* **10** 1–51. In Polish; English translation by D. Dabrowska and T. Speed in *Statist. Sci.* **5** 463–472 (1990).
PEARL, J. (2009). *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge Univ. Press, Cambridge.
PROVAN, J. S. and BILLERA, L. J. (1980). Decompositions of simplicial complexes related to diameters of convex polyhedra. *Math. Oper. Res.* **5** 576–594. MR0593648
RICHARDSON, T. S. (2003). Markov properties for acyclic directed mixed graphs. *Scand. J. Stat.* **30** 145–157.
RICHARDSON, T. S., EVANS, R. J. and ROBINS, J. M. (2011). Transparent parametrizations of models for potential outcomes. In *Bayesian Statistics* 9 569–610. Oxford Univ. Press, Oxford. MR3204019
RICHARDSON, T. S. and SPIRTES, P. (2002). Ancestral graph Markov models. *Ann. Statist.* **30** 962–1030. MR1926166
RICHARDSON, T. S., EVANS, R. J., ROBINS, J. M. and SHPITSER, I. (2017). Nested Markov properties for acyclic directed mixed graphs. Available at arXiv:1701.06686.
ROBINS, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math. Model.* **7** 1393–1512. MR0877758
RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688–701.
SHPITSER, I., EVANS, R. J., RICHARDSON, T. S. and ROBINS, J. M. (2013). Sparse nested Markov models with log-linear parameters. In 29*th Conference on Uncertainty in Artificial Intelligence* 576–585.
SILVA, R. and GHAHRAMANI, Z. (2009). The hidden life of latent variables: Bayesian learning with mixed graph models. *J. Mach. Learn. Res.* **10** 1187–1238.
SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation, Prediction, and Search*, 2nd ed. MIT Press, Cambridge, MA. MR1815675

TIAN, J. and PEARL, J. (2002). On the testable implications of causal models with hidden variables. In *Proceedings of the* 18*th Conference on Uncertainty in Artificial Intelligence* 519–527.
VER STEEG, G. and GALSTYAN, A. (2011). A sequence of relaxations constraining hidden variable models. In *Proceedings of the* 27*th Conference on Uncertainty in Artificial Intelligence* 717–726.
VERMA, T. S. and PEARL, J. (1990). Equivalence and synthesis of causal models. In *Proceedings of the* 6*th Conference on Uncertainty in Artificial Intelligence* 255–268.

The Annals of Statistics
2018, Vol. 46, No. 6A, 2657–2682
https://doi.org/10.1214/17-AOS1632

MULTI-THRESHOLD ACCELERATED FAILURE TIME MODEL[1]

BY JIALIANG LI AND BAISUO JIN

National University of Singapore and University of Science and Technology of China

A two-stage procedure for simultaneously detecting multiple thresholds and achieving model selection in the segmented accelerated failure time (AFT) model is developed in this paper. In the first stage, we formulate the threshold problem as a group model selection problem so that a concave 2-norm group selection method can be applied. In the second stage, the thresholds are finalized via a refining method. We establish the strong consistency of the threshold estimates and regression coefficient estimates under some mild technical conditions. The proposed procedure performs satisfactorily in our simulation studies. Its real world applicability is demonstrated via analyzing a follicular lymphoma data.

REFERENCES

BAI, J. and PERRON, P. (2003). Computation and analysis of multiple structural change models. *J. Appl. Econometrics* **18** 1–22.

BUCKLEY, J. and JAMES, I. (1979). Linear regression with censored data. *Biometrika* **66** 429–436.

DAVE, S. S., WRIGHT, G., TAN, B., ROSENWALD, A., GASCOYNE, R. D., CHAN, W. C. et al. (2004). Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells. *N. Engl. J. Med.* **351** 2159–2169.

DAVIS, R. A., LEE, T. C. M. and RODRIGUEZ-YAM, G. A. (2006). Structural break estimation for nonstationary time series models. *J. Amer. Statist. Assoc.* **101** 223–239.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360.

FEARNHEAD, P. and VASILEIOU, D. (2009). Bayesian analysis of isochores. *J. Amer. Statist. Assoc.* **104** 132–141.

GORDON, A. D. (1981). *Classification: Methods for the Exploratory Analysis of Multivariate Data*. Chapman & Hall, New York. MR0637465

HÁJEK, J. and RÉNYI, A. (1955). Generalization of an inequality of Kolmogorov. *Acta Math. Acad. Sci. Hung.* **6** 281–283.

HANSEN, B. E. (2000). Sample splitting and threshold estimation. *Econometrica* **68** 575–603. DOI:10.1111/1468-0262.00124. MR1769379

HARCHAOUI, Z. and LÉVY-LEDUC, C. (2010). Multiple change-point estimation with a total variation penalty. *J. Amer. Statist. Assoc.* **105** 1480–1493.

HUANG, J., MA, S. and XIE, H. (2006). Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics* **62** 813–820.

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

MSC2010 subject classifications. 60K35.

Key words and phrases. Break points, MCP penalty, SCAD penalty, Stute estimator, threshold regression.

JIN, B., SHI, X. and WU, Y. (2013). A novel and fast methodology for simultaneous multiple structural break estimation and variable selection for nonstationary time series models. *Stat. Comput.* **23** 221–231. DOI:10.1007/s11222-011-9304-6. MR3016940

KALBFLEISCH, J. D. and PRENTICE, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

KOSOROK, M. R. and SONG, R. (2007). Inference under right censoring for transformation models with a change-point based on a covariate threshold. *Ann. Statist.* **35**. MR2341694

LAWLESS, J. F. (2011). *Statistical Models and Methods for Lifetime Data*. John Wiley & Sons, New York.

LIN, D. Y., WEI, L. J. and YING, Z. (1998). Accelerated failure time models for counting processes. *Biometrika* **85** 605–618. MR1665810

LUO, X., TURNBULL, B. W. and CLARK, L. C. (1997). Likelihood ratio tests for a changepoint with survival data. *Biometrika* **84** 555–565. MR1603981

PERRON, P. (2006). Dealing with structural breaks. In *Palgrave Handbook of Econometrics*, *Vol.* 1: *Econometric Theory* (K. Patterson and T. C. Mills, eds.) 278–352. Palgrave Macmillan, Basingstoke, UK.

PONS, O. (2003). Estimation in a Cox regression model with a change-point according to a threshold in a covariate. *Ann. Statist.* **31** 442–463.

PRENTICE, R. L. (1978). Linear rank test with right censored data. *Biometrika* **65** 167–179.

PUNTANEN, S. (2011). Projection matrices, generalized inverse matrices, and singular value decomposition by Haruo Yanai, Kei Takeuchi, Yoshio Takane. *Int. Stat. Rev.* **79** 503–504.

STUTE, W. (1993). Consistent estimation under random censorship when covariables are present. *J. Multivariate Anal.* **45** 89–103. MR1222607

STUTE, W. (1995). The central limit theorem under random censorship. *Ann. Statist.* **23** 422–439.

STUTE, W. (1996). Distributional convergence under random censorship when covariables are present. *Scand. J. Stat.* **23** 461–471.

TONG, H. (2012). *Threshold Models in Non-linear Time Series Analysis*. Springer, Berlin.

TSIATIS, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *Ann. Statist.* **18** 354–372.

XIA, X., JIANG, B., LI, J. and ZHANG, W. (2016). Low-dimensional confounder adjustment and high-dimensional penalized estimation for survival analysis. *Lifetime Data Anal.* **22** 547–569.

YANG, S., SU, C. and YU, K. (2008). A general method to the strong law of large numbers and its applications. *Statist. Probab. Lett.* **78** 794–803. MR2409544

YAO, Y.-C. and AU, S. T. (1989). Least-squares estimation of a step function. *Sankhya, Ser. A* **51** 370–381. MR1175613

YING, Z. (1993). A large sample study of rank estimation for censored regression data. *Ann. Statist.* **21** 76–99.

YU, T., LI, J. and MA, S. (2012). Adjusting confounders in ranking biomarkers: amodel-based ROC approach. *Brief. Bioinform.* **13** 513–523.

ZHANG, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942.

The Annals of Statistics
2018, Vol. 46, No. 6A, 2683–2710
https://doi.org/10.1214/17-AOS1635

MEASURING AND TESTING FOR INTERVAL QUANTILE DEPENDENCE[1]

BY LIPING ZHU*, YAOWU ZHANG† AND KAI XU†

Renmin University of China and Shanghai University of Finance and Economics†*

In this article, we introduce the notion of interval quantile independence which generalizes the notions of statistical independence and quantile independence. We suggest an index to measure and test departure from interval quantile independence. The proposed index is invariant to monotone transformations, nonnegative and equals zero if and only if the interval quantile independence holds true. We suggest a moment estimate to implement the test. The resultant estimator is root-n-consistent if the index is positive and n-consistent otherwise, leading to a consistent test of interval quantile independence. The asymptotic distribution of the moment estimator is free of parent distribution, which facilitates to decide the critical values for tests of interval quantile independence. When our proposed index is used to perform feature screening for ultrahigh dimensional data, it has the desirable sure screening property.

REFERENCES

[1] ANDERSON, T. W. and DARLING, D. A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Ann. Math. Stat.* **23** 193–212. MR0050238

[2] ANDREWS, D. W. K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica* **68** 399–405. MR1748009

[3] BERGSMA, W. and DASSIOS, A. (2014). A consistent test of independence based on a sign covariance related to Kendall's tau. *Bernoulli* **20** 1006–1028.

[4] BLUM, J. R., KIEFER, J. and ROSENBLATT, M. (1961). Distribution free tests of independence based on the sample distribution function. *Ann. Math. Stat.* **32** 485–498. MR0125690

[5] CSÖRGÖ, M. and HORVÁTH, L. (1997). *Limit Theorems in Change-Point Analysis*. Wiley, New York.

[6] DAVIES, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **64** 247–254.

[7] FAN, J., FENG, Y. and SONG, R. (2011). Nonparametric independence screening in sparse ultra-high dimensional additive models. *J. Amer. Statist. Assoc.* **106** 544–557.

[8] FAN, J., HU, T.-C. and TRUONG, Y. K. (1994). Robust non-parametric function estimation. *Scand. J. Stat.* **21** 433–446.

[9] FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911. MR2530322

[10] FAN, J. and SONG, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.* **38** 3567–3604. MR2766861

MSC2010 subject classifications. Primary 62G10, 62H20; secondary 68Q32.

Key words and phrases. Correlation, independence, quantile regression, rank test, sure screening property.

[11] HE, X., WANG, L. and HONG, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Ann. Statist.* **41** 342–369. MR3059421
[12] HELLER, R., HELLER, Y. and CORFINE, M. (2013). A consistent multivariate test of association based on ranks of distances. *Biometrika* **100** 503–510.
[13] JOE, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman & Hall, London. MR1462613
[14] KOENKER, R. (2005). *Quantile Regression*. *Econometric Society Monographs* **38**. Cambridge Univ. Press, Cambridge. MR2268657
[15] KOENKER, R. and BASSETT, G. (1978). Regression quantiles. *Econometrica* **46** 33–49.
[16] KOENKER, R. and PORTNOY, S. (1987). L-estimation for linear models. *J. Amer. Statist. Assoc.* **82** 851–857. MR0909992
[17] LI, G., LI, Y. and TSAI, C. L. (2015). Quantile correlations and quantile autoregressive modeling. *J. Amer. Statist. Assoc.* **110** 246–261.
[18] LI, G., PENG, H., ZHANG, J. and ZHU, L. (2012). Robust rank correlation based screening. *Ann. Statist.* **40** 1846–1877. MR3015046
[19] LI, R., ZHONG, W. and ZHU, L. (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc.* **107** 1129–1139.
[20] MCKEAGUE, I. W. and QIAN, M. An adaptive resampling test for detecting the presence of significant predictors. *J. Amer. Statist. Assoc.* **110** 1422–1433.
[21] PARK, T., SHAO, X. and YAO, S. (2015). Partial martingale difference correlation. *Electron. J. Stat.* **9** 1492–1517.
[22] SHAO, X. and ZHANG, J. (2014). Martingale difference correlation and its use in high-dimensional variable screening. *J. Amer. Statist. Assoc.* **109** 1302–1318. MR3265698
[23] SZEKELY, G. J. and RIZZO, M. L. (2009). Brownian distance covariance. *Ann. Appl. Stat.* **3** 1236–1265. MR2752127
[24] SZÉKELY, G. J., RIZZO, M. L. and BAKIROV, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35** 2769–2794. MR2382665
[25] WU, Y. and YIN, G. (2015). Conditional quantile screening in ultrahigh-dimensional heterogeneous data. *Biometrika* **102** 65–76. MR3335096
[26] ZHENG, Q., PENG, L. and HE, X. (2015). Globally adaptive quantile regression with ultrahigh dimensional data. *Ann. Statist.* **43** 2225–2258. MR3396984
[27] ZHU, L., ZHANG, Y. and XU, K. (2018). Supplement to “Measuring and testing for interval quantile dependence.” DOI:10.1214/17-AOS1635SUPP.
[28] ZHU, L. P., LI, L., LI, R. and ZHU, L. X. (2011). Model-free feature screening for ultrahigh dimensional data. *J. Amer. Statist. Assoc.* **106** 1464–1475.

The Annals of Statistics
2018, Vol. 46, No. 6A, 2711–2746
https://doi.org/10.1214/17-AOS1636

BARYCENTRIC SUBSPACE ANALYSIS ON MANIFOLDS

BY XAVIER PENNEC[1]

Université Côte d'Azur and INRIA

This paper investigates the generalization of Principal Component Analysis (PCA) to Riemannian manifolds. We first propose a new and general type of family of subspaces in manifolds that we call barycentric subspaces. They are implicitly defined as the locus of points which are weighted means of $k+1$ reference points. As this definition relies on points and not on tangent vectors, it can also be extended to geodesic spaces which are not Riemannian. For instance, in stratified spaces, it naturally allows principal subspaces that span several strata, which is impossible in previous generalizations of PCA. We show that barycentric subspaces locally define a submanifold of dimension k which generalizes geodesic subspaces.

Second, we rephrase PCA in Euclidean spaces as an optimization on flags of linear subspaces (a hierarchy of properly embedded linear subspaces of increasing dimension). We show that the Euclidean PCA minimizes the Accumulated Unexplained Variances by all the subspaces of the flag (AUV). Barycentric subspaces are naturally nested, allowing the construction of hierarchically nested subspaces. Optimizing the AUV criterion to optimally approximate data points with flags of affine spans in Riemannian manifolds lead to a particularly appealing generalization of PCA on manifolds called Barycentric Subspace Analysis (BSA).

REFERENCES

ABSIL, P.-A., MAHONY, R. and SEPULCHRE, R. (2008). *Optimization Algorithms on Matrix Manifolds*. MR2364186

AFSARI, B. (2011). Riemannian L^p center of mass: Existence, uniqueness, and convexity. *Proc. Amer. Math. Soc.* **139** 655–673. MR2736346

BHATTACHARYA, R. and PATRANGENARU, V. (2003). Large sample theory of intrinsic and extrinsic sample means on manifolds. I. *Ann. Statist.* **31** 1–29. MR1962498

BHATTACHARYA, R. and PATRANGENARU, V. (2005). Large sample theory of intrinsic and extrinsic sample means on manifolds. II. *Ann. Statist.* **33** 1225–1259. MR2195634

BISHOP, R. L. and O'NEILL, B. (1969). Manifolds of negative curvature. *Trans. Amer. Math. Soc.* **145** 1–49. MR251664

BREWIN, L. (2009). Riemann normal coordinate expansions using Cadabra. *Classical Quantum Gravity* **26** 175017. MR2534336

BUSEMANN, H. (1955). *The Geometry of Geodesics*. Academic Press, San Diego. MR0075623

BUSER, P. and KARCHER, H. (1981). *Gromov's Almost Flat Manifolds*. *Astérisque* **81**. Société Mathématique de France, Paris. MR0619537

BUSS, S. R. and FILLMORE, J. P. (2001). Spherical averages and applications to spherical splines and interpolation. *ACM Trans. Graph.* **20** 95–126.

MSC2010 subject classifications. Primary 60D05, 62H25; secondary 58C06, 62H11.
Key words and phrases. Manifold, Fréchet mean, barycenter, flag of subspaces, PCA.

COSTA, S. I. R., SANTOS, S. A. and STRAPASSON, J. E. (2015). Fisher information distance: A geometrical reading. *Discrete Appl. Math.* **197** 59–69. MR3398961

CROUCH, P. and LEITE, F. S. (1995). The dynamic interpolation problem: On Riemannian manifolds, Lie groups, and symmetric spaces. *J. Dyn. Control Syst.* **1** 177–202. MR1333770

CUTLER, A. and BREIMAN, L. (1994). Archetypal analysis. *Technometrics* **36** 338–347. MR1304898

DAMON, J. and MARRON, J. S. (2013). Backwards principal component analysis and principal nested relations. *J. Math. Imaging Vision* **50** 107–114. MR3233137

DRYDEN, I. L. (2005). Statistical analysis on high-dimensional spheres and shape spaces. *Ann. Statist.* **33** 1643–1665. MR2166558

ELTZNER, B., JUNG, S. and HUCKEMANN, S. (2015). Dimension reduction on polyspheres with application to skeletal representations. In *Geometric Science of Information* (F. Nielsen and F. Barbaresco, eds.). *Lecture Notes in Computer Science* **9389** 22–29. Springer, Cham. MR3442181

FERAGEN, A., OWEN, M., PETERSEN, J., WILLE, M. M. W., THOMSEN, L. H., DIRKSEN, A. and DE BRUIJNE, M. (2013). Tree-space statistics and approximations for large-scale analysis of anatomical trees. In *International Conference on Information Processing in Medical Imaging* (*IPMI* 2013), *Asilomar*, *CA*, *USA*. *Lecture Notes in Computer Science* **7917** 74–85. Springer, Berlin.

FLETCHER, P. T., LU, C., PIZER, S. M. and JOSHI, S. (2004). Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Trans. Med. Imag.* **23** 995–1005.

FRÉCHET, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. H. Poincareé* **10** 215–310. MR27464

GAY-BALMAZ, F., HOLM, D., MEIER, D., RATIU, T. and VIALARD, F.-X. (2012). Invariant higher-order variational problems. *Comm. Math. Phys.* **309** 413–458. MR2864799

GROISSER, D. (2004). Newton's method, zeroes of vector fields, and the Riemannian center of mass. *Adv. in Appl. Math.* **33** 95–135. MR2064359

HUCKEMANN, S., HOTZ, T. and MUNK, A. (2010). Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. *Statist. Sinica* **20** 1–58. MR2640651

HUCKEMANN, S. and ZIEZOLD, H. (2006). Principal component analysis for Riemannian manifolds, with an application to triangular shape spaces. *Adv. in Appl. Probab.* **38** 299–319. MR2264946

JUNG, S., DRYDEN, I. L. and MARRON, J. S. (2012). Analysis of principal nested spheres. *Biometrika* **99** 551–568. MR2966769

JUNG, S., LIU, X., MARRON, J. S. and PIZER, S. M. (2010). Generalized PCA via the backward stepwise approach in image analysis. In *Proc. of the Int. Symposium Brain, Body and Machine*. *Advances in Intelligent and Soft Computing* **83** 111–123. Springer, Berlin.

KARCHER, H. (1977). Riemannian center of mass and mollifier smoothing. *Comm. Pure Appl. Math.* **30** 509–541. MR442975

KARCHER, H. (2014). Riemannian center of mass and so called Karcher mean. Available at arXiv:1407.2087.

KENDALL, W. S. (1990). Probability, convexity, and harmonic maps with small image. I. Uniqueness and fine existence. *Proc. Lond. Math. Soc.* (3) **61** 371–406. MR1063050

LE, H. (2004). Estimation of Riemannian barycentres. *LMS J. Comput. Math.* **7** 193–200. MR2085875

LEPORÉ, N., BRUN, C., CHOU, Y.-Y., LEE, A., BARYSHEVA, M., PENNEC, X., MCMAHON, K., MEREDITH, M., DE ZUBICARAY, G., WRIGHT, M., TOGA ARTHUR, W. and THOMPSON, P. (2008). Best individual template selection from deformation tensor minimization. In *Proc. of the* 2008 *IEEE Int. Symp. ISBI* 2008, *Paris*, *France* 460–463.

LORENZI, M., AYACHE, N. and PENNEC, X. (2015). Regional flux analysis for discovering and quantifying anatomical changes: An application to the brain morphometry in Alzheimer's disease. *NeuroImage* **115** 224–234.

LORENZI, M. and PENNEC, X. (2013). Geodesics, parallel transport & one-parameter subgroups for diffeomorphic image registration. *Int*. *J*. *Comput*. *Vis*. **105** 111–127. MR3104013

MACHADO, L., SILVA LEITE, F. and KRAKOWSKI, K. (2010). Higher-order smoothing splines versus least squares problems on Riemannian manifolds. *J*. *Dyn*. *Control Syst*. **16** 121–148. MR2580471

OLVER, P. J. (2001). Geometric foundations of numerical algorithms and symmetry. *Appl*. *Algebra Engrg*. *Comm*. *Comput*. **11** 417–436. MR1828189

PENNEC, X. (2006). Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. *J*. *Math*. *Imaging Vision* **25** 127–154. MR2254442

PENNEC, X. (2015). Barycentric subspaces and affine spans in manifolds. In *Geometric Science of Information*. *Lecture Notes in Computer Science* **9389** 12–21. Springer, Cham. MR3442180

PENNEC, X. (2018a). Supplement A to "Barycentric subspace analysis on manifolds." DOI:10.1214/17-AOS1636SUPPA.

PENNEC, X. (2018b). Supplement B to "Barycentric subspace analysis on manifolds." DOI:10.1214/17-AOS1636SUPPB.

PENNEC, X. and ARSIGNY, V. (2013). Exponential barycenters of the canonical Cartan connection and invariant means on Lie groups. In Matrix information geometry 123–166. Springer, Heidelberg. MR2964451

PENNEC, X., FILLARD, P. and AYACHE, N. (2006). A Riemannian framework for tensor computing. *Int*. *J*. *Comput*. *Vis*. **66** 41–66.

SMALL, C. G. (1996). *The Statistical Theory of Shapes*. Springer, Berlin. MR1418639

SOMMER, S. (2013). Horizontal dimensionality reduction and iterated frame bundle development. In *Proceedings of the* 1*st International Conference Geometric Science of Information Held in Paris*, *August* 28–30, 2013 (*GSI* 2013) (F. Nielsen and F. Barbaresco, eds.). *Lecture Notes in Computer Science* **8085** 76–83. Springer, Heidelberg. MR3126126

SOMMER, S., LAUZE, F. and NIELSEN, M. (2014). Optimization over geodesics for exact principal geodesic analysis. *Adv*. *Comput*. *Math*. **40** 283–313. MR3194707

TIPPING, M. E. and BISHOP, C. M. (1999). Probabilistic principal component analysis. *J*. *Roy*. *Statist*. *Soc*. *Ser*. *B* **61** 611–622. MR1707864

WANG, X. and MARRON, J. S. (2008). A scale-based approach to finding effective dimensionality in manifold learning. *Electron*. *J*. *Stat*. **2** 127–148. MR2386090

WEYENBERG, G. S. (2015). Statistics in the Billera–Holmes–Vogtmann treespace. Ph.D. thesis, Univ. Kentucky. MR3450194

WILSON, R. C., HANCOCK, E. R., PEKALSKA, E. and DUIN, R. P. W. (2014). Spherical and hyperbolic embeddings of data. *IEEE Trans*. *Pattern Anal*. *Mach*. *Intell*. **36** 2255–2269.

YANG, L. (2011). Medians of probability measures in Riemannian manifolds and applications to radar target detection. Ph.D. thesis, Poitier Univ.

ZHAI, H. (2016). Principal component analysis in phylogenetic tree space. Ph.D. thesis, Univ. North Carolina at Chapel Hill, Ann Arbor, MI. MR3542232

ZHANG, M. and FLETCHER, P. T. (2013). Probabilistic principal geodesic analysis. In *Proceedings of the* 26*th International Conference on Neural Information Processing Systems* (*NIPS*'13) **1** 1178–1186.

The Annals of Statistics
2018, Vol. 46, No. 6A, 2747–2774
https://doi.org/10.1214/17-AOS1637

THE LANDSCAPE OF EMPIRICAL RISK FOR NONCONVEX LOSSES

BY SONG MEI[1], YU BAI[2] AND ANDREA MONTANARI[3]

Stanford University

Most high-dimensional estimation methods propose to minimize a cost function (empirical risk) that is a sum of losses associated to each data point (each example). In this paper, we focus on the case of nonconvex losses. Classical empirical process theory implies uniform convergence of the empirical (or sample) risk to the population risk. While under additional assumptions, uniform convergence implies consistency of the resulting M-estimator, it does not ensure that the latter can be computed efficiently.

In order to capture the complexity of computing M-estimators, we study the landscape of the empirical risk, namely its stationary points and their properties. We establish uniform convergence of the gradient and Hessian of the empirical risk to their population counterparts, as soon as the number of samples becomes larger than the number of unknown parameters (modulo logarithmic factors). Consequently, good properties of the population risk can be carried to the empirical risk, and we are able to establish one-to-one correspondence of their stationary points. We demonstrate that in several problems such as nonconvex binary classification, robust regression and Gaussian mixture model, this result implies a complete characterization of the landscape of the empirical risk, and of the convergence properties of descent algorithms.

We extend our analysis to the very high-dimensional setting in which the number of parameters exceeds the number of samples, and provides a characterization of the empirical risk landscape under a nearly information-theoretically minimal condition. Namely, if the number of samples exceeds the sparsity of the parameters vector (modulo logarithmic factors), then a suitable uniform convergence result holds. We apply this result to nonconvex binary classification and robust regression in very high-dimension.

REFERENCES

[1] AI, A., LAPANOWSKI, A., PLAN, Y. and VERSHYNIN, R. (2014). One-bit compressed sensing with non-Gaussian measurements. *Linear Algebra Appl.* **441** 222–239. MR3134344
[2] ALON, U., BARKAI, N., NOTTERMAN, D. A., GISH, K., YBARRA, S., MACK, D. and LEVINE, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* **96** 6745–6750.
[3] ANANDKUMAR, A., GE, R. and JANZAMIN, M. (2015). Learning overcomplete latent variable models through tensor methods. In *Proceedings of the Conference on Learning Theory* (*COLT*), *Paris*, *France*.

MSC2010 subject classifications. Primary 62F10, 62J02; secondary 62H30.
Key words and phrases. Nonconvex optimization, empirical risk minimization, landscape, uniform convergence.

[4] ATIA, G. K. and SALIGRAMA, V. (2012). Boolean compressed sensing and noisy group testing. *IEEE Trans. Inform. Theory* **58** 1880–1901. MR2932872

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

[6] BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford.

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

[8] CANDES, E. J. and TAO, T. (2005). Decoding by linear programming. *IEEE Trans. Inform. Theory* **51** 4203–4215.

[9] CHAPELLE, O., DO, C. B., TEO, C. H., LE, Q. V. and SMOLA, A. J. (2009). Tighter bounds for structured estimation. In *Advances in Neural Information Processing Systems* 281–288.

[10] CHEN, Y. and CANDES, E. (2015). Solving random quadratic systems of equations is nearly as easy as solving linear systems. In *Advances in Neural Information Processing Systems* 739–747.

[11] DASKALAKIS, C., TZAMOS, C. and ZAMPETAKIS, M. (2016). Ten steps of EM suffice for mixtures of two Gaussians. Available at arXiv:1609.00368.

[12] DONOHO, D. L. (2006). Compressed sensing. *IEEE Trans. Inform. Theory* **52** 1289–1306. MR2241189

[13] DUBROVIN, B. A., FOMENKO, A. T. and NOVIKOV, S. P. (2012). *Modern Geometry—Methods and Applications: Part II: The Geometry and Topology of Manifolds* **104**. Springer, Berlin.

[14] FISHER, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* **222** 309–368.

[15] FISHER, R. A. (1925). Theory of statistical estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society* **22** 700–725. Cambridge Univ Press, Cambridge.

[16] GE, R., HUANG, F., JIN, C. and YUAN, Y. (2015). Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory* 797–842.

[17] GUILLEMIN, V. and POLLACK, A. (2010). *Differential Topology* **370**. AMS, Providence.

[18] HUBER, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1** 799–821. MR0356373

[19] KESHAVAN, R. H., OH, S. and MONTANARI, A. (2009). Matrix completion from a few entries. In *IEEE International Symposium on Information Theory*, 2009. *ISIT* 2009 324–328. IEEE, New York.

[20] LASKA, J. N. and BARANIUK, R. G. (2012). Regime change: Bit-depth versus measurement-rate in compressive sensing. *IEEE Trans. Signal Process.* **60** 3496–3505.

[21] LASKA, J. N., WEN, Z., YIN, W. and BARANIUK, R. G. (2011). Trust, but verify: Fast and accurate signal recovery from 1-bit compressive measurements. *IEEE Trans. Signal Process.* **59** 5289–5301.

[22] LECUN, Y., BENGIO, Y. and HINTON, G. (2015). Deep learning. *Nature* **521** 436–444.

[23] LOH, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *Ann. Statist.* **45** 866–896. MR3650403

[24] LOH, P.-L. and WAINWRIGHT, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Statist.* **40** 1637–1664. MR3015038

[25] LOH, P.-L. and WAINWRIGHT, M. J. (2013). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems* 476–484.

[26] LOZANO, A. C. and MEINSHAUSEN, N. (2013). Minimum distance estimation for robust high-dimensional regression. Available at arXiv:1307.3227.

[27] MEI, S., BAI, Y. and MONTANARI, A. (2018). Supplement to "The landscape of empirical risk for nonconvex losses." DOI:10.1214/17-AOS1637SUPP.

[28] MILNOR, J. (1963). *Morse Theory* **51**. Princeton Univ. Press, Princeton.

[29] MILNOR, J. W. (1997). *Topology from the Differentiable Viewpoint*. Princeton Univ. Press, Princeton, NH.

[30] MONTANARI, A. and RICHARD, E. (2014). A statistical model for tensor PCA. In *Advances in Neural Information Processing Systems* 2897–2905.

[31] NEGAHBAN, S. N., RAVIKUMAR, P., WAINWRIGHT, M. J. and YU, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statist. Sci.* **27** 538–557. MR3025133

[32] NESTEROV, Y. (2013). *Introductory Lectures on Convex Optimization*: *A Basic Course* **87**. Springer Science & Business Media, New York.

[33] NESTEROV, Y. (2013). Gradient methods for minimizing composite functions. *Math. Program.* **140** 125–161.

[34] NGUYEN, T. and SANNER, S. (2013). Algorithms for direct 0–1 loss optimization in binary classification. In *Proceedings of the* 30*th International Conference on Machine Learning* 1085–1093.

[35] PENG, J., ZHU, J., BERGAMASCHI, A., HAN, W., NOH, D.-Y., POLLACK, J. R. and WANG, P. (2010). Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. Appl. Stat.* **4** 53–77. MR2758084

[36] PLAN, Y. and VERSHYNIN, R. (2013). One-bit compressed sensing by linear programming. *Comm. Pure Appl. Math.* **66** 1275–1297. MR3069959

[37] PLAN, Y. and VERSHYNIN, R. (2013). Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Trans. Inform. Theory* **59** 482–494.

[38] PLAN, Y., VERSHYNIN, R. and YUDOVINA, E. (2014). High-dimensional estimation with geometric constraints. Preprint. Available at arXiv:1404.3749.

[39] ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Stat.* **22** 400–407. MR0042668

[40] SERDOBOLSKII, V. I. (2013). *Multivariate Statistical Analysis*: *A High-Dimensional Approach* **41**. Springer, New York.

[41] SUN, J., QU, Q. and WRIGHT, J. (2016). A geometric analysis of phase retrieval. Available at arXiv:1602.06664.

[42] TSITSIKLIS, J. N., BERTSEKAS, D. P. and ATHANS, M. (1984). Distributed asynchronous deterministic and stochastic gradient optimization algorithms. In 1984 *American Control Conference* 484–489.

[43] VAN DE GEER, S. A. (2000). *Applications of Empirical Process Theory*. *Cambridge Series in Statistical and Probabilistic Mathematics* **6**. Cambridge Univ. Press, Cambridge. MR1739079

[44] VAPNIK, V. N. (1998). *Statistical Learning Theory*. Wiley, New York. MR1641250

[45] WU, Y. and LIU, Y. (2012). Robust truncated hinge loss support vector machines. *J. Amer. Statist. Assoc.* **102** 974–983. MR2411659

[46] XU, J., HSU, D. and MALEKI, A. (2016). Global analysis of expectation maximization for mixtures of two Gaussians. Preprint. Available at arXiv:1608.07630.

[47] YANG, Z., WANG, Z., LIU, H., ELDAR, Y. C. and ZHANG, T. (2015). Sparse nonlinear regression: Parameter estimation and asymptotic inference. Available at arXiv:1511.04514.

The Annals of Statistics
2018, Vol. 46, No. 6A, 2775–2805
https://doi.org/10.1214/17-AOS1638

DESIGNS WITH BLOCKS OF SIZE TWO AND APPLICATIONS TO MICROARRAY EXPERIMENTS

BY JANET GODOLPHIN

University of Surrey

Designs with blocks of size two have numerous applications. In experimental situations where observation loss is common, it is important for a design to be robust against breakdown. For designs with one treatment factor and a single blocking factor, with blocks of size two, conditions for connectivity and robustness are obtained using combinatorial arguments and results from graph theory. Lower bounds are given for the breakdown number in terms of design parameters. For designs with equal or near equal treatment replication, the concepts of treatment and block partitions, and of linking blocks, are used to obtain information on the number of blocks required to guarantee various levels of robustness. The results provide guidance for construction of designs with good robustness properties.

Robustness conditions are also established for row column designs in which one of the blocking factors involves blocks of size two. Such designs are particularly relevant for microarray experiments, where the high risk of observation loss makes robustness important. Disconnectivity in row column designs can be classified as three types. Techniques are given to assess design robustness according to each type, leading to lower bounds for the breakdown number. Guidance is given for robust design construction.

Cyclic designs and interwoven loop designs are shown to have good robustness properties.

REFERENCES

BAGCHI, S. and CHENG, C.-S. (1993). Some optimal designs of block size two. *J. Statist. Plann. Inference* **37** 245–253. MR1243802

BAILEY, R. A. (2007). Designs for two-colour microarray experiments. *J. Roy. Statist. Soc. Ser. C* **56** 365–394. MR2409757

BAILEY, R. A. and CAMERON, P. J. (2009). Combinatorics of optimal designs. In *Surveys in Combinatorics* 2009 (S. Huczynska, J. D. Mitchell and C. M. Roney-Dougal, eds.). *London Mathematical Society Lecture Notes* **365** 19–73. Cambridge Univ. Press, Cambridge.

BAILEY, R. A., SCHIFFL, K. and HILGERS, R.-D. (2013). A note on robustness of D-optimal block designs for two-colour microarray experiments. *J. Statist. Plann. Inference* **143** 1195–1202. MR3049621

BAKSALARY, J. K. and TABIS, Z. (1987). Conditions for the robustness of block designs against the unavailability of data. *J. Statist. Plann. Inference* **16** 49–54.

BAUER, D., HAKIMI, S. L., KAHL, N. and SCHMEICHEL, E. (2009). Sufficient degree conditions for k-edge-connectedness of a graph. *Networks* **54** 95–98.

MSC2010 subject classifications. Primary 62K10; secondary 62K05.

Key words and phrases. Breakdown number, connectivity, incomplete block design, microarray experiment, observation loss, robustness, row column design.

BHAUMIK, D. K. and WHITTINGHILL, D. C. (1991). Optimality and robustness to the unavailibility of blocks in block designs. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **53** 399–407.

BONDY, J. A. (1969). Properties of graphs with constraints on degrees. *Studia Sci. Math. Hungar.* **4** 473–475.

BONDY, J. A. and MURTY, U. S. R. (2008). *Graph Theory*. Springer, Berlin.

BUTZ, L. (1982). *Connectivity in Multifactor Designs*. Heldermann Verlag, Berlin. MR0671251

CHAI, F.-S., LIAO, C.-T. and TSAI, S.-F. (2007). Statistical designs for two-color spotted microarray experiments. *Biom. J.* **49** 259–271. MR2364336

CHARTRAND, G. (1966). A graph-theoretic approach to a communications problem. *SIAM J. Appl. Math.* **14** 778–781.

DEY, A. (1993). Robustness of block designs against missing data. *Statist. Sinica* **3** 219–231.

GHOSH, S. (1979). On robustness of designs against incomplete data. *Sankhya* **40** 204–208.

GHOSH, S. (1982). Robustness of BIBD against the unavailability of data. *J. Statist. Plann. Inference* **6** 29–32.

GODOLPHIN, J. D. (2004). Simple pilot procedures for the avoidance of disconnected experimental designs. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **53** 133–147.

GODOLPHIN, J. D. (2013). On the connectivity problem for m-way designs. *J. Stat. Theory Pract.* **7** 732–744. MR3196631

GODOLPHIN, J. (2018). Supplement to "Designs with blocks of size two and applications to microarray experiments." DOI:10.1214/17-AOS1638SUPP.

GODOLPHIN, J. D. and GODOLPHIN, E. J. (2001). On the connectivity of row-column designs. *Util. Math.* **60** 51–65.

GODOLPHIN, J. D. and GODOLPHIN, E. J. (2015). The use of treatment concurrences to assess robustness of binary block designs against the loss of whole blocks. *Aust. N. Z. J. Stat.* **57** 225–239.

GODOLPHIN, J. D. and WARREN, H. R. (2014). An efficient procedure for the avoidance of disconnected incomplete block designs. *Comput. Statist. Data Anal.* **71** 1134–1146.

GUPTA, S. (2006). Balanced factorial designs for cDNA microarray experiments. *Comm. Statist. Theory Methods* **35** 1469–1476.

JOHN, J. A. and WILLIAMS, E. R. (1995). *Cyclic and Computer Generated Designs*, 2nd ed. Chapman & Hall, London. MR1382127

KEL'MANS, A. K. (1972). Asymptotic formulas for the probability of k-connectedness of random graphs. *Theory Probab. Appl.* **17** 243–254.

KERR, M. K. (2006). Efficient 2^k factorial designs for blocks of size 2 with microarray applications. *J. Qual. Technol.* **38** 309–318.

KERR, M. K. and CHURCHILL, G. A. (2001). Experimental design for gene expression microarrays. *Biostatistics* **2** 183–201.

MAHBUB LATIF, A. H. M., BRETZ, F. and BRUNNER, E. (2009). Robustness considerations in selecting efficient two-color microarray designs. *Bioinformatics* **25** 2355–2361.

MORGAN, J. P. (2015). *Blocking with Independent Responses* (A. Dean, M. Morris, J. Stufken and D. Bingham, eds.). *Handbook of Design and Analysis of Experiments*. CRC Press, Boca Raton, FL.

NGUYEN, D. V., ARPAT, A. B., WANG, N. and CARROLL, R. J. (2002). DNA microarray experiments: Biological and technological aspects. *Biometrics* **58** 701–717.

PATERSON, L. (1983). Circuits and efficiency in incomplete block designs. *Biometrika* **70** 215–225.

SANCHEZ, P. S. and GLONEK, G. F. V. (2009). Optimal designs for 2-colour microarray experiments. *Biostatistics* **10** 561–574.

TSAI, S.-F. and LIAO, C.-T. (2013). Minimum breakdown designs in blocks of size two. *J. Statist. Plann. Inference* **143** 202–208.

WIT, E., NOBILE, A. and KHANIN, R. (2005). Near-optimal designs for dual channel microarray studies. *J. Roy. Statist. Soc. Ser. C* **54** 817–830. MR2209033

WU, C. F. J. and HAMADA, M. S. (2009). *Experiments*: *Planning*, *Analysis*, *and Optimization*, 2nd ed. Wiley, Hoboken, NJ. MR2583259

YANG, Y. J. and DRAPER, N. R. (2003). Two-level factorial and fractional factorial designs in blocks of size two. *J*. *Qual*. *Technol*. **35** 294–305.

The Annals of Statistics
2018, Vol. 46, No. 6A, 2806–2843
https://doi.org/10.1214/17-AOS1640

LOCAL ROBUST ESTIMATION OF THE PICKANDS DEPENDENCE FUNCTION[1]

BY MIKAEL ESCOBAR-BACH*, YURI GOEGEBEUR* AND ARMELLE GUILLOU†

University of Southern Denmark and Université de Strasbourg et CNRS*†

We consider the robust estimation of the Pickands dependence function in the random covariate framework. Our estimator is based on local estimation with the minimum density power divergence criterion. We provide the main asymptotic properties, in particular the convergence of the stochastic process, correctly normalized, towards a tight centered Gaussian process. The finite sample performance of our estimator is evaluated with a simulation study involving both uncontaminated and contaminated samples. The method is illustrated on a dataset of air pollution measurements.

REFERENCES

ABEGAZ, F., GIJBELS, I. and VERAVERBEKE, N. (2012). Semiparametric estimation of conditional copulas. *J. Multivariate Anal.* **110** 43–73. MR2927509

ALIPRANTIS, C. D. and BORDER, K. C. (2006). *Infinite Dimensional Analysis: A Hitchhiker's Guide*, 3rd ed. Springer, Berlin. MR2378491

BAI, S. and TAQQU, M. S. (2013). Multivariate limit theorems in the context of long-range dependence. *J. Time Series Anal.* **34** 717–743. MR3127215

BASU, A., HARRIS, I. R., HJORT, N. L. and JONES, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika* **85** 549–559. MR1665873

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

BÜCHER, A., DETTE, H. and VOLGUSHEV, S. (2011). New estimators of the Pickands dependence function and a test for extreme-value dependence. *Ann. Statist.* **39** 1963–2006. MR2893858

CAPÉRAÀ, P., FOUGÈRES, A.-L. and GENEST, C. (1997). A nonparametric estimation procedure for bivariate extreme value copulas. *Biometrika* **84** 567–577. MR1603985

DAOUIA, A., GARDES, L., GIRARD, S. and LEKINA, A. (2011). Kernel estimators of extreme level curves. *TEST* **20** 311–333. MR2834049

DEHEUVELS, P. (1991). On the limiting behavior of the Pickands estimator for bivariate extreme-value distributions. *Statist. Probab. Lett.* **12** 429–439. MR1142097

EINMAHL, U. and MASON, D. M. (1997). Gaussian approximation of local empirical processes indexed by functions. *Probab. Theory Related Fields* **107** 283–311.

ESCOBAR-BACH, M., GOEGEBEUR, Y. and GUILLOU, A. (2018). Supplement to "Local robust estimation of the Pickands dependence function." DOI:10.1214/17-AOS1640SUPP.

FILS-VILLETARD, A., GUILLOU, A. and SEGERS, J. (2008). Projection estimators of Pickands dependence functions. *Canad. J. Statist.* **36** 369–382.

GARDES, L. and GIRARD, S. (2015). Nonparametric estimation of the conditional tail copula. *J. Multivariate Anal.* **137** 1–16.

MSC2010 subject classifications. Primary 62G32, 62G05, 62G20; secondary 60F05, 60G70.

Key words and phrases. Conditional Pickands dependence function, robustness, stochastic convergence.

GIJBELS, I., OMELKA, M. and VERAVERBEKE, N. (2015). Estimation of a copula when a covariate affects only marginal distributions. *Scand*. *J*. *Stat*. **42** 1109–1126. MR3426313

GINÉ, E. and GUILLOU, A. (2001). On consistency of kernel density estimators for randomly censored data: Rates holding uniformly over adaptive intervals. *Ann*. *Inst*. *Henri Poincaré Probab*. *Stat*. **37** 503–522.

GINÉ, E. and GUILLOU, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Ann*. *Inst*. *Henri Poincaré Probab*. *Stat*. **38** 907–921.

GINÉ, E., KOLTCHINSKII, V. and ZINN, J. (2004). Weighted uniform consistency of kernel density estimators. *Ann*. *Probab*. **32** 2570–2605. MR2078551

GUDENDORF, G. and SEGERS, J. (2010). Extreme-value copulas. In *Copula Theory and Its Applications* (P. Jaworski, F. Durante, W. Härdle and W. Rychlik, eds.) 127–145. Springer, Berlin.

KOTZ, S. and NADARAJAH, S. (2000). *Extreme Value Distributions—Theory and Applications*. Imperial College Press, London.

LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation*, 2nd ed. Springer, New York.

NOLAN, D. and POLLARD, D. (1987). U-Processes: Rates of convergence. *Ann*. *Statist*. **15** 780–799. MR0888439

PICKANDS, J. III (1981). Multivariate extreme value distributions. *Bull*. *Inst*. *Internat*. *Statist*. **49** 859–878, 894–902. MR0820979

PORTIER, F. and SEGERS, J. (2017). On the weak convergence of the empirical conditional copula under a simplifying assumption. Available at arXiv:1511.06544.

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

SEVERINI, T. A. (2005). *Elements of Distribution Theory*. *Cambridge Series in Statistical and Probabilistic Mathematics* **17**. Cambridge Univ. Press, Cambridge. MR2168237

SKLAR, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ*. *Inst*. *Stat*. *Univ*. *Paris* **8** 229–231. MR0125600

STUTE, W. (1986). Conditional empirical processes. *Ann*. *Statist*. **14** 638–647. MR0840519

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

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*: *With Applications to Statistics*. Springer, New York. MR1385671

VAN DER VAART, A. W. and WELLNER, J. A. (2007). Empirical processes indexed by estimated functions. In *Asymptotics*: *Particles*, *Processes and Inverse Problems*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **55** 234–252. IMS, Beachwood, OH. MR2459942

The Annals of Statistics
2018, Vol. 46, No. 6A, 2844–2870
https://doi.org/10.1214/17-AOS1641

STRONG IDENTIFIABILITY AND OPTIMAL MINIMAX RATES FOR FINITE MIXTURE ESTIMATION[1]

BY PHILIPPE HEINRICH AND JONAS KAHN

Université Lille 1 and Université de Toulouse, CNRS

We study the rates of estimation of finite mixing distributions, that is, the parameters of the mixture. We prove that under some regularity and strong identifiability conditions, around a given mixing distribution with m_0 components, the optimal local minimax rate of estimation of a mixing distribution with m components is $n^{-1/(4(m-m_0)+2)}$. This corrects a previous paper by Chen [*Ann. Statist.* **23** (1995) 221–233].

By contrast, it turns out that there are estimators with a (nonuniform) pointwise rate of estimation of $n^{-1/2}$ for all mixing distributions with a finite number of components.

REFERENCES

BONTEMPS, D. and GADAT, S. (2014). Bayesian methods for the shape invariant model. *Electron. J. Stat.* **8** 1522–1568. MR3263130

CAILLERIE, C., CHAZAL, F., DEDECKER, J. and MICHEL, B. (2013). Deconvolution for the Wasserstein metric and geometric inference. In *Geometric Science of Information* 561–568. Springer, Berlin.

CHEN, J. H. (1995). Optimal rate of convergence for finite mixture models. *Ann. Statist.* **23** 221–233. MR1331665

DACUNHA-CASTELLE, D. and GASSIAT, E. (1997). The estimation of the order of a mixture model. *Bernoulli* **3** 279–299. MR1468306

DEDECKER, J. and MICHEL, B. (2013). Minimax rates of convergence for Wasserstein deconvolution with supersmooth errors in any dimension. *J. Multivariate Anal.* **122** 278–291.

DEELY, J. J. and KRUSE, R. L. (1968). Construction of sequences estimating the mixing distribution. *Ann. Math. Stat.* **39** 286–288.

DUDLEY, R. M. (2002). *Real Analysis and Probability*. *Cambridge Studies in Advanced Mathematics* **74**. Cambridge Univ. Press, Cambridge. MR1932358

FAN, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.* **19** 1257–1272. MR1126324

GASSIAT, E. and VAN HANDEL, R. (2013). Consistent order estimation and minimal penalties. *IEEE Trans. Inform. Theory* **59** 1115–1128. MR3015722

GENOVESE, C. R. and WASSERMAN, L. (2000). Rates of convergence for the Gaussian mixture sieve. *Ann. Statist.* **28** 1105–1127. MR1810921

GHOSAL, S. and VAN DER VAART, A. W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.* **29** 1233–1263. MR1873329

MSC2010 subject classifications. Primary 62G05; secondary 62G20.

Key words and phrases. Local asymptotic normality, convergence of experiments, maximum likelihood estimate, Wasserstein metric, mixing distribution, mixture model, rate of convergence, strong identifiability, pointwise rate, superefficiency.

HÁJEK, J. (1972). Local asymptotic minimax and admissibility in estimation. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (*Univ. California*, *Berkeley*, *Calif.*, 1970/1971), *Vol. I*: *Theory of Statistics* 175–194. Univ. California Press, Berkeley, CA.

HEINRICH, P. and KAHN, J. (2018). Supplement to "Strong identifiability and optimal minimax rates for finite mixture estimation." DOI:10.1214/17-AOS1641SUPP.

HO, N. and NGUYEN, X. (2015). Identifiability and optimal rates of convergence for parameters of multiple types in finite mixtures. Preprint. Available at arXiv:1501.02497.

HO, N. and NGUYEN, X. (2016a). On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electron*. *J*. *Stat*. **10** 271–307. MR3466183

HO, N. and NGUYEN, X. (2016b). Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Ann*. *Statist*. **44** 2726–2755. MR3576559

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

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

KUHN, M. A., FEIGELSON, E. D., GETMAN, K. V., BADDELEY, A. J., BROOS, P. S., SILLS, A., BATE, M. R., POVICH, M. S., LUHMAN, K. L., BUSK, H. A. et al. (2014). The spatial structure of young stellar clusters. I. Subclusters. *The Astrophysical Journal* **787** 107.

LE CAM, L. (1960). Locally asymptotically normal families of distributions. Certain approximations to families of distributions and their use in the theory of estimation and testing hypotheses. *Univ. California Publ*. *Statist*. **3** 37–98. MR0126903

LE CAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, New York.

LINDSAY, B. G. (1989). Moment matrices: Applications in mixtures. *Ann*. *Statist*. **17** 722–740. MR0994263

LIU, M. and HANCOCK, G. R. (2014). Unrestricted mixture models for class identification in growth mixture modeling. *Educational and Psychological Measurement* **74** 557–584.

MARTIN, R. (2012). Convergence rate for predictive recursion estimation of finite mixtures. *Statist*. *Probab. Lett*. **82** 378–384.

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

MCLACHLAN, G. and PEEL, D. (2000). *Finite Mixture Models*. Wiley, New York.

NGUYEN, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *Ann*. *Statist*. **41** 370–400. MR3059422

PEARSON, K. (1894). Contributions to the theory of mathematical evolution. *Philos*. *Trans*. *R*. *Soc*. *Lond*. *Ser. A Math*. *Phys*. *Eng*. *Sci*. **185** 71–110.

ROUSSEAU, J. and MENGERSEN, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J*. *R*. *Stat*. *Soc*. *Ser. B*. *Stat*. *Methodol*. **73** 689–710.

TEH, Y. W. (2010). Dirichlet process. In *Encyclopedia of Machine Learning* (C. Sammut and G. Webb, eds.) 280–287. Springer, New York.

TITTERINGTON, D. M., SMITH, A. F. M. and MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. Wiley, Chichester.

VAN DE GEER, S. (1996). Rates of convergence for the maximum likelihood estimator in mixture models. *J*. *Nonparametr. Stat*. **6** 293–310.

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

YANG, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika* **92** 937–950. MR2234196

ZHU, H.-T. and ZHANG, H. (2004). Hypothesis testing in mixture regression models. *J*. *R*. *Stat*. *Soc*. *Ser. B*. *Stat*. *Methodol*. **66** 3–16. MR2035755

ZHU, H. and ZHANG, H. (2006). Asymptotics for estimation and testing procedures under loss of identifiability. *J*. *Multivariate Anal*. **97** 19–45. MR2208842

The Annals of Statistics
2018, Vol. 46, No. 6A, 2871–2903
https://doi.org/10.1214/17-AOS1642

SUB-GAUSSIAN ESTIMATORS OF THE MEAN OF A RANDOM MATRIX WITH HEAVY-TAILED ENTRIES[1]

BY STANISLAV MINSKER

University of Southern California

Estimation of the covariance matrix has attracted a lot of attention of the statistical research community over the years, partially due to important applications such as principal component analysis. However, frequently used empirical covariance estimator, and its modifications, is very sensitive to the presence of outliers in the data. As P. Huber wrote [*Ann. Math. Stat.* **35** (1964) 73–101], "...This raises a question which could have been asked already by Gauss, but which was, as far as I know, only raised a few years ago (notably by Tukey): what happens if the true distribution deviates slightly from the assumed normal one? As is now well known, the sample mean then may have a catastrophically bad performance...." Motivated by Tukey's question, we develop a new estimator of the (element-wise) mean of a random matrix, which includes covariance estimation problem as a special case. Assuming that the entries of a matrix possess only finite second moment, this new estimator admits sub-Gaussian or sub-exponential concentration around the unknown mean in the operator norm. We explain the key ideas behind our construction, and discuss applications to covariance estimation and matrix completion problems.

REFERENCES

[1] AHLSWEDE, R. and WINTER, A. (2002). Strong converse for identification via quantum channels. *IEEE Trans. Inform. Theory* **48** 569–579. MR1889969

[2] ALEKSANDROV, A. B. and PELLER, V. V. (2016). Operator Lipschitz functions. *Russian Math. Surveys* **71** 605.

[3] ALON, N., MATIAS, Y. and SZEGEDY, M. (1996). The space complexity of approximating the frequency moments. In *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing* 20–29. ACM, New York.

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

[5] BROWNLEES, C., JOLY, E. and LUGOSI, G. (2015). Empirical risk minimization for heavy-tailed losses. *Ann. Statist.* **43** 2507–2536. MR3405602

[6] BUTLER, R. W., DAVIES, P. L. and JHUN, M. (1993). Asymptotics for the minimum covariance determinant estimator. *Ann. Statist.* **21** 1385–1400. MR1241271

[7] CAI, T. T., REN, Z. and ZHOU, H. H. (2016). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electron. J. Stat.* **10** 1–59.

MSC2010 subject classifications. Primary 60B20, 62G35; secondary 62H12.
Key words and phrases. Random matrix, heavy tails, concentration inequality, covariance estimation, matrix completion.

[8] 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

[9] CANDÈS, E. J., LI, X., MA, Y. and WRIGHT, J. (2011). Robust principal component analysis? *J. ACM* **58** Art. 11, 37. MR2811000

[10] CARLEN, E. (2010). Trace inequalities and quantum entropy: An introductory course. Available at http://www.mathphys.org/AZschool/material/AZ09-carlen.pdf.

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

[12] CATONI, O. (2016). PAC-Bayesian bounds for the Gram matrix and least squares regression with a random design. Preprint. Available at arXiv:1603.05229.

[13] DAVIES, L. (1992). The asymptotics of Rousseeuw's minimum volume ellipsoid estimator. *Ann. Statist.* MR1193314 1828–1843.

[14] DEVROYE, L., LERASLE, M., LUGOSI, G. and OLIVEIRA, R. I. (2015). Sub-Gaussian mean estimators. Preprint. Available at arXiv:1509.05845.

[15] FAN, J., WANG, W. and ZHONG, Y. (2016). An ℓ_∞ eigenvector perturbation bound and its application to robust covariance estimation. Preprint. Available at arXiv:1603.03516.

[16] FAN, J., WANG, W. and ZHU, Z. (2016). Robust low-rank matrix recovery. Preprint. Available at arXiv:1603.08315.

[17] GIULINI, I. (2015). PAC-Bayesian bounds for Principal Component Analysis in Hilbert spaces. Preprint. Available at arXiv:1511.06263.

[18] HSU, D. and SABATO, S. (2016). Loss minimization and parameter estimation with heavy tails. *J. Mach. Learn. Res.* **17** Paper No. 18, 40. MR3491112

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

[20] HUBER, P. J. and RONCHETTI, E. M. (2009). *Robust Statistics*, 2nd ed. Wiley, Hoboken, NJ.

[21] HUBERT, M., ROUSSEEUW, P. J. and VAN AELST, S. (2008). High-breakdown robust multivariate methods. *Statist. Sci.* **23** 92–119. MR2431867

[22] JERRUM, M. R., VALIANT, L. G. and VAZIRANI, V. V. (1986). Random generation of combinatorial structures from a uniform distribution. *Theoret. Comput. Sci.* **43** 169–188.

[23] JOLY, E., LUGOSI, G. and OLIVEIRA, R. I. (2017). On the estimation of the mean of a random vector. *Electron. J. Stat.* **11** 440–451. MR3619312

[24] KLOPP, O., LOUNICI, K. and TSYBAKOV, A. B. (2017). Robust matrix completion. *Probab. Theory Related Fields* **169** 523–564. MR3704775

[25] KOLTCHINSKII, V. and LOUNICI, K. (2016). New asymptotic results in principal component analysis. Preprint. Available at arXiv:1601.01457.

[26] KOLTCHINSKII, V. and LOUNICI, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli* **23** 110–133.

[27] KOLTCHINSKII, V., LOUNICI, K. and TSYBAKOV, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.* **39** 2302–2329. MR2906869

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

[29] LEPSKI, O. (1992). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.* **36** 682–697.

[30] LERASLE, M. and OLIVEIRA, R. I. (2011). Robust empirical mean estimators. Preprint. Available at arXiv:1112.3914.

[31] LIEB, E. H. (1973). Convex trace functions and the Wigner–Yanase–Dyson conjecture. *Adv. Math.* **11** 267–288. MR0332080

[32] LOUNICI, K. (2014). High-dimensional covariance matrix estimation with missing observations. *Bernoulli* **20** 1029–1058.

[33] LUGOSI, G. and MENDELSON, S. (2017). Sub-Gaussian estimators of the mean of a random vector. Preprint. Available at arXiv:1702.00482.

[34] MARONNA, R. A. (1976). Robust M-estimators of multivariate location and scatter. *Ann. Statist.* **4** 51–67. MR0388656

[35] MINSKER, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli* **21** 2308–2335.

[36] MINSKER, S. (2018). Supplement to "Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries." DOI:10.1214/17-AOS1642SUPP.

[37] MINSKER, S. and WEI, X. (2017). Estimation of the covariance structure of heavy-tailed distributions. Preprint. Available at arXiv:1708.00502.

[38] NEMIROVSKI, A. and YUDIN, D. (1983). *Problem Complexity and Method Efficiency in Optimization*. Wiley, New York.

[39] OLIVEIRA, R. I. (2009). Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. Preprint. Available at arXiv:0911.0600.

[40] SRIVASTAVA, N. and VERSHYNIN, R. (2013). Covariance estimation for distributions with $2+\varepsilon$ moments. *Ann. Probab.* **41** 3081–3111. MR3127875

[41] TROPP, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.* **12** 389–434. MR2946459

[42] TROPP, J. A. (2015). An introduction to matrix concentration inequalities. Preprint. Available at arXiv:1501.01571.

[43] TYLER, D. E. (1987). A distribution-free M-estimator of multivariate scatter. *Ann. Statist.* **15** 234–251. MR0885734

[44] VERSHYNIN, R. (2010). Introduction to the non-asymptotic analysis of random matrices. Preprint. Available at arXiv:1011.3027.

[45] ZHANG, T., CHENG, X. and SINGER, A. (2016). Marčenko–Pastur law for Tyler's M-estimator. *J. Multivariate Anal.* **149** 114–123. MR3507318

The Annals of Statistics
2018, Vol. 46, No. 6A, 2904–2938
https://doi.org/10.1214/17-AOS1643

CAUSAL INFERENCE IN PARTIALLY LINEAR STRUCTURAL EQUATION MODELS

BY DOMINIK ROTHENHÄUSLER[1], JAN ERNEST[1,2] AND PETER BÜHLMANN

ETH Zürich

We consider identifiability of partially linear additive structural equation models with Gaussian noise (PLSEMs) and estimation of distributionally equivalent models to a given PLSEM. Thereby, we also include robustness results for errors in the neighborhood of Gaussian distributions. Existing identifiability results in the framework of additive SEMs with Gaussian noise are limited to linear and nonlinear SEMs, which can be considered as special cases of PLSEMs with vanishing nonparametric or parametric part, respectively. We close the wide gap between these two special cases by providing a comprehensive theory of the identifiability of PLSEMs by means of (A) a graphical, (B) a transformational, (C) a functional and (D) a causal ordering characterization of PLSEMs that generate a given distribution $\mathbb{P}$. In particular, the characterizations (C) and (D) answer the fundamental question to which extent nonlinear functions in additive SEMs with Gaussian noise restrict the set of potential causal models, and hence influence the identifiability.

On the basis of the transformational characterization (B) we provide a score-based estimation procedure that outputs the graphical representation (A) of the distribution equivalence class of a given PLSEM. We derive its (high-dimensional) consistency and demonstrate its performance on simulated datasets.

REFERENCES

[1] ANDERSSON, S. A., MADIGAN, D. and PERLMAN, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *Ann. Statist.* **25** 505–541. MR1439312

[2] BÜHLMANN, P. (2013). Causal statistical inference in high dimensions. *Math. Methods Oper. Res.* **77** 357–370. MR3072790

[3] BÜHLMANN, P., PETERS, J. and ERNEST, J. (2014). CAM: Causal additive models, high-dimensional order search and penalized regression. *Ann. Statist.* **42** 2526–2556. MR3277670

[4] CASTELO, R. and KOCKA, T. (2003). On inclusion-driven learning of Bayesian networks. *J. Mach. Learn. Res.* **4** 527–574.

[5] CHICKERING, D. (2002). Optimal structure identification with greedy search. *J. Mach. Learn. Res.* **3** 507–554.

[6] CHICKERING, D. M. (1995). A transformational characterization of equivalent Bayesian network structures. In *Proceedings of the* 11*th Conference on Uncertainty in Artificial Intelligence* (*UAI*) 87–98. Morgan Kaufmann, San Francisco, CA.

MSC2010 subject classifications. Primary 62G99, 62H99; secondary 68T99.

Key words and phrases. Causal inference, distribution equivalence class, graphical model, high-dimensional consistency, partially linear structural equation model.

[7] GLASS, T. A., GOODMAN, S. N., HERNÁN, M. A. and SAMET, J. M. (2013). Causal inference in public health. *Annu. Rev. Public Health* **34** 61–75.

[8] HOYER, P., HYVARINEN, A., SCHEINES, R., SPIRTES, P., RAMSEY, J., LACERDA, G. and SHIMIZU, S. (2008). Causal discovery of linear acyclic models with arbitrary distributions. In *Proceedings of the* 24*th Annual Conference on Uncertainty in Artificial Intelligence* (*UAI*) 282–289. AUAI Press, Corvallis, OR.

[9] HOYER, P. O., JANZING, D., MOOIJ, J. M., PETERS, J. and SCHÖLKOPF, B. (2009). Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems* 21 (*NIPS*) 689–696. Curran, Red Hook, NY.

[10] KALISCH, M. and BÜHLMANN, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.* **8** 613–636.

[11] KALISCH, M., MÄCHLER, M., COLOMBO, D., MAATHUIS, M. H. and BÜHLMANN, P. (2012). Causal inference using graphical models with the R package pcalg. *J. Stat. Softw.* **47** 1–26.

[12] MEEK, C. (1995). Causal inference and causal explanation with background knowledge. In *Proceedings of the* 11*th Conference on Uncertainty in Artificial Intelligence* (*UAI*) 403–410. Morgan Kaufmann, San Francisco, CA.

[13] NANDY, P., HAUSER, A. and MAATHUIS, M. (2015). High-dimensional consistency in score-based and hybrid structure learning. arXiv:1507.02608.

[14] NOWZOHOUR, C. and BÜHLMANN, P. (2016). Score-based causal learning in additive noise models. *Statistics* **50** 471–485. MR3506653

[15] PEARL, J. (2009). *Causality*: *Models*, *Reasoning and Inference*, 2nd ed. Cambridge Univ. Press, New York.

[16] PETERS, J. and BÜHLMANN, P. (2014). Identifiability of Gaussian structural equation models with equal error variances. *Biometrika* **101** 219–228.

[17] PETERS, J., MOOIJ, J., JANZING, D. and SCHÖLKOPF, B. (2011). Identifiability of causal graphs using functional models. In *Proceedings of the* 27*th Conference on Uncertainty in Artificial Intelligence* (*UAI*) 589–598. AUAI Press, Corvallis, OR.

[18] PETERS, J., MOOIJ, J., JANZING, D. and SCHÖLKOPF, B. (2014). Causal discovery with continuous additive noise models. *J. Mach. Learn. Res.* **15** 2009–2053.

[19] RAMSEY, J., HANSON, S., HANSON, C., HALCHENKO, Y., POLDRACK, R. and GLYMOUR, C. (2010). Six problems for causal inference from fmri. *NeuroImage* **49** 1545–1558.

[20] ROTHENHÄUSLER, D., ERNEST, J. and BÜHLMANN, P. (2018). Supplement to "Causal inference in partially linear structural equation models." DOI:10.1214/17-AOS1643SUPP.

[21] SHIMIZU, S., HOYER, P., HYVÄRINEN, A. and KERMINEN, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.* **7** 2003–2030.

[22] SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation*, *Prediction*, *and Search*, 2nd ed. MIT Press.

[23] SPIRTES, P. and ZHANG, K. (2016). Causal discovery and inference: Concepts and recent methodological advances. *Applied Informatics* **3** 1–28.

[24] STATNIKOV, A., HENAFF, M., LYTKIN, N. I. and ALIFERIS, C. F. (2012). New methods for separating causes from effects in genomics data. *BMC Genomics* **13** S22.

[25] STEKHOVEN, D., MORAES, I., SVEINBJÖRNSSON, G., HENNIG, L., MAATHUIS, M. and BÜHLMANN, P. (2012). Causal stability ranking. *Bioinformatics* **28** 2819–2823.

[26] VAN DE GEER, S. (2014). On the uniform convergence of empirical norms and inner products, with application to causal inference. *Electron. J. Stat.* **8** 543–574.

[27] VAN DE GEER, S. and BÜHLMANN, P. (2013). ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *Ann. Statist.* **41** 536–567. MR3099113

[28] VERMA, T. and PEARL, J. (1990). Equivalence and synthesis of causal models. In *Proceedings of the* 6*th Conference on Uncertainty in Artificial Intelligence* (*UAI*) 220–227. AUAI Press, Corvallis, OR.

[29] WOOD, S. N. (2003). Thin plate regression splines. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 95–114. MR1959095

[30] WOOD, S. N. (2006). *Generalized Additive Models*: *An Introduction with R*. CRC Press, Boca Raton, FL.

[31] ZHANG, K. and HYVÄRINEN, A. (2009). On the identifiability of the post-nonlinear causal model. In *Proceedings of the* 25*th Conference on Uncertainty in Artificial Intelligence* (*UAI*) 647–655. AUAI Press, Corvallis, OR.

The Annals of Statistics
2018, Vol. 46, No. 6A, 2939–2959
https://doi.org/10.1214/17-AOS1644

ON MSE-OPTIMAL CROSSOVER DESIGNS[1]

BY CHRISTOPH NEUMANN AND JOACHIM KUNERT

Technische Universität Dortmund

In crossover designs, each subject receives a series of treatments one after the other. Most papers on optimal crossover designs consider an estimate which is corrected for carryover effects. We look at the estimate for direct effects of treatment, which is not corrected for carryover effects. If there are carryover effects, this estimate will be biased. We try to find a design that minimizes the mean square error, that is, the sum of the squared bias and the variance. It turns out that the designs which are optimal for the corrected estimate are highly efficient for the uncorrected estimate.

REFERENCES

AZAÏS, J.-M. and DRUILHET, P. (1997). Optimality of neighbour balanced designs when neighbour effects are neglected. *J. Statist. Plann. Inference* **64** 353–367. MR1621623

BOSE, M. and DEY, A. (2009). *Optimal Crossover Designs*. World Scientific, Hackensack, NJ.

CHÊNG, C. S. and WU, C.-F. (1980). Balanced repeated measurements designs. *Ann. Statist.* **8** 1272–1283. MR0594644

DAVID, O., MONOD, H., LORGEOU, J. and PHILIPPEAU, G. (2001). Control of interplot interference in Grain Maize. *Crop Science* **41** 406.

HORN, R. A. and JOHNSON, C. R. (2013). *Matrix Analysis*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2978290

KIEFER, J. (1975). Construction and optimality of generalized youden designs. In *A Survey of Statistical Design and Linear Models* (J. N. Srivastava, ed.) 333–366. North-Holland, Amsterdam.

KUNERT, J. (1983). Optimal design and refinement of the linear model with applications to repeated measurements designs. *Ann. Statist.* **11** 247–257. MR0684882

KUNERT, J. and SAILER, O. (2006). On nearly balanced designs for sensory trials. *Food Qual. Prefer.* **17** 219–227.

KUSHNER, H. B. (1997). Optimal repeated measurements designs: The linear optimality equations. *Ann. Statist.* **25** 2328–2344. MR1604457

KUSHNER, H. B. (1998). Optimal and efficient repeated-measurements designs for uncorrelated observations. *J. Amer. Statist. Assoc.* **93** 1176–1187. MR1649211

MACFIE, H. J. H., BRATCHELL, N., GRENHOFF, K. and VALLIS, L. V. (1989). Designs to balance the effect of order of presentation and first-order carry-over effects in hall tests. *J. Sens. Stud.* **4** 129–148.

OZAN, M. O. and STUFKEN, J. (2010). Assessing the impact of carryover effects on the variances of estimators of treatment differences in crossover designs. *Stat. Med.* **29** 2480–2485. MR2897362

SENN, S. (2002). *Cross-over Trials in Clinical Research*, 2nd ed. *Statistics in Practice*. Wiley, Chichester.

SHAH, K. R. and SINHA, B. K. (1989). *Theory of Optimal Designs*. *Lecture Notes in Statistics* **54**. Springer, New York.

MSC2010 subject classifications. Primary 62K05; secondary 62K10.
Key words and phrases. Optimal design, crossover design, MSE-optimality.

The Annals of Statistics
2018, Vol. 46, No. 6A, 2960–2984
https://doi.org/10.1214/17-AOS1645

TESTING FOR PERIODICITY IN FUNCTIONAL TIME SERIES

BY SIEGFRIED HÖRMANN[1], PIOTR KOKOSZKA[2] AND GILLES NISOL[3]

Graz University of Technology, Colorado State University and Université libre de Bruxelles

We derive several tests for the presence of a periodic component in a time series of functions. We consider both the traditional setting in which the periodic functional signal is contaminated by functional white noise, and a more general setting of a weakly dependent contaminating process. Several forms of the periodic component are considered. Our tests are motivated by the likelihood principle and fall into two broad categories, which we term multivariate and fully functional. Generally, for the functional series that motivate this research, the fully functional tests exhibit a superior balance of size and power. Asymptotic null distributions of all tests are derived and their consistency is established. Their finite sample performance is examined and compared by numerical studies and application to pollution data.

REFERENCES

AUE, A., NORINHO, D. D. and HÖRMANN, S. (2015). On the prediction of stationary functional time series. *J. Amer. Statist. Assoc.* **110** 378–392. MR3338510

BROCKWELL, P. J. and DAVIS, R. A. (1991). *Time Series: Theory and Methods*, 2nd ed. Springer, New York. MR1093459

CEROVECKI, C. and HÖRMANN, S. (2017). On the CLT for discrete Fourier transforms of functional time series. *J. Multivariate Anal.* **154** 282–295. MR3588570

CHIANI, M. (2014). Distribution of the largest eigenvalue for real Wishart and Gaussian random matrices and a simple approximation for the Tracy–Widom distribution. *J. Multivariate Anal.* **129** 69–81. MR3215980

CUEVAS, A., FEBRERO, M. and FRAIMAN, R. (2004). An ANOVA test for functional data. *Comput. Statist. Data Anal.* **47** 111–122. MR2087932

FISHER, R. A. (1929). Tests of significance in harmonic analysis. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **125** 54–59.

GROMENKO, O., KOKOSZKA, P. and REIMHERR, M. (2017). Detection of change in the spatiotemporal mean function. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 29–50.

HANNAN, E. J. (1961). Testing for a jump in the spectral function. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **23** 394–404.

HAYS, S., SHEN, H. and HUANG, J. Z. (2012). Functional dynamic factor models with application to yield curve forecasting. *Ann. Appl. Stat.* **6** 870–894. MR3012513

HÖRMANN, S., HORVÁTH, L. and REEDER, R. (2013). A functional version of the ARCH model. *Econometric Theory* **29** 267–288.

HÖRMANN, S., KIDZIŃSKI, L. and HALLIN, M. (2015). Dynamic functional principal components. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 319–348.

MSC2010 subject classifications. Primary 62M15, 62G10; secondary 60G15, 62G20.

Key words and phrases. Functional data, time series data, periodicity, spectral analysis, testing, asymptotics.

HÖRMANN, S. and KOKOSZKA, P. (2010). Weakly dependent functional data. *Ann. Statist.* **38** 1845–1884.

HÖRMANN, S., KOKOSZKA, P. and NISOL, G. (2018). Supplement to "Testing for periodicity in functional time series." DOI:10.1214/17-AOS1645SUPP.

HORVÁTH, L. and KOKOSZKA, P. (2012). *Inference for Functional Data with Applications*. Springer.

HORVÁTH, L., KOKOSZKA, P. and RICE, G. (2014). Testing stationarity of functional time series. *J. Econometrics* **179** 66–82.

HSING, T. and EUBANK, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley, Chichester. MR3379106

JENKINS, G. M. and PRIESTLEY, M. B. (1957). The spectral analysis of time-series. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **19** 1–12 (discussion 47–63). MR0092329

KLEPSCH, J., KLÜPPELBERG, C. and WEI, T. (2017). Prediction of functional ARMA processes with an application to traffic data. *Econom. Stat.* **1** 128–149. MR3669993

LIN, Z. and LIU, W. (2009). On maxima of periodograms of stationary processes. *Ann. Statist.* **37** 2676–2695. MR2541443

MACNEILL, I. B. (1974). Tests for periodic components in multiple time series. *Biometrika* **61** 57–70. MR0378317

MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1979). *Multivariate Analysis*. Academic Press, London.

PANARETOS, V. M. and TAVAKOLI, S. (2013a). Fourier analysis of stationary time series in function space. *Ann. Statist.* **41** 568–603. MR3099114

PANARETOS, V. M. and TAVAKOLI, S. (2013b). Cramér–Karhunen–Loève representation and harmonic principal component analysis of functional time series. *Stochastic Process. Appl.* **123** 2779–2807.

RAMSAY, J., HOOKER, G. and GRAVES, S. (2009). *Functional Data Analysis with R and MATLAB*. Springer.

ROSS, S. M. (2010). *Introduction to Probability Models*. Elsevier, Amsterdam. MR3307945

SCHUSTER, A. (1898). On the investigation of hidden periodicities with application to the supposed 26 day period od meteorological phenomena. *Terr. Mag.* **3** 13–41.

SHAO, X. and WU, W. B. (2007). Asymptotic spectral theory for nonlinear time series. *Ann. Statist.* **35** 1773–1801. MR2351105

STADLOBER, E., HÖRMANN, S. and PFEILER, B. (2008). Quality and performance of a PM10 daily forecasting model. *Athmospheric Environment* **42** 1098–1109.

STADLOBER, E. and PFEILER, B. (2004). Explorative Analyse der Feinstaub-Konzentrationen von Oktober 2003 bis März 2004. Technical report, TU Graz.

WALKER, G. (1914). On the criteria for the reality of relationships or periodicities. *Calcutta Ind. Met. Memo* **21**.

WU, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proc. Natl. Acad. Sci. USA* **102** 14150–14154. MR2172215

ZAMANI, A., HAGHBIN, H. and SHISHEBOR, Z. (2016). Some tests for detecting cyclic behavior in functional time series with application in climate change. Technical report, Shiraz Univ.

ZHANG, X. (2016). White noise testing and model diagnostic checking for functional time series. *J. Econometrics* **194** 76–95. MR3523520

The Annals of Statistics
2018, Vol. 46, No. 6A, 2985–3013
https://doi.org/10.1214/17-AOS1646

LIMITING BEHAVIOR OF EIGENVALUES IN HIGH-DIMENSIONAL MANOVA VIA RMT

BY ZHIDONG BAI*,[1], KWOK PUI CHOI†,[2] AND YASUNORI FUJIKOSHI‡,[3]

Northeast Normal University, National University of Singapore†*
and Hiroshima University‡

In this paper, we derive the asymptotic joint distributions of the eigenvalues under the null case and the local alternative cases in the MANOVA model and multiple discriminant analysis when both the dimension and the sample size are large. Our results are obtained by random matrix theory (RMT) without assuming normality in the populations. It is worth pointing out that the null and nonnull distributions of the eigenvalues and invariant test statistics are asymptotically robust against departure from normality in high-dimensional situations. Similar properties are pointed out for the null distributions of the invariant tests in multivariate regression model. Some new formulas in RMT are also presented.

REFERENCES

AMEMIYA, Y. (1990). A note on the limiting distribution of certain characteristic roots. *Statist. Probab. Lett.* **9** 465–470. MR1060089

ANDERSON, T. W. (1951). The asymptotic distribution of certain characteristic roots and vectors. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 1950 103–130. Univ. California Press, Berkeley, CA. MR0044075

ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley, Hoboken, NJ. MR1990662

BAI, Z. D. (1985). A note on the limiting distribution of the eigenvalues of a class of random matrices. *J. Math. Res. Exposition* **5** 113–118. MR0842111

BAI, Z., CHOI, K. P. and FUJIKOSHI, Y. (2018). Supplement to "Limiting behavior of eigenvalues in high-dimensional MANOVA via RMT." DOI:10.1214/17-AOS1646SUPP.

BAI, Z. D., LIU, H. X. and WONG, W. K. (2011). Asymptotic properties of eigenmatrices of a large sample covariance matrix. *Ann. Appl. Probab.* **21** 1994–2015. MR2884057

BAI, Z. D. and SARANADASA, H. (1996). Effect of high dimension: By an example of a two sample problem. *Statist. Sinica* **6** 311–329.

BAI, Z. D. and SILVERSTEIN, J. W. (2004). CLT for linear spectral statistics of large-dimensional sample covariance matrices. *Ann. Probab.* **32** 535–605.

BAI, Z. D. and ZHOU, W. (2008). Large sample covariance matrices without independence structures in columns. *Statist. Sinica* **5** 425–442.

BAI, Z., JIANG, D., YAO, J.-F. and ZHENG, S. (2009). Corrections to LRT on large-dimensional covariance matrix by RMT. *Ann. Statist.* **37** 3822–3840. MR2572444

CAI, T. T. and XIA, Y. (2014). High-dimensional sparse MANOVA. *J. Multivariate Anal.* **131** 174–196.

MSC2010 subject classifications. Primary 62H10; secondary 62E20.
Key words and phrases. Asymptotic distribution, eigenvalues, discriminant analysis, high-dimensional case, MANOVA, nonnormality, RMT, test statistics.

CHEN, S. X. and QIN, Y. L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.* **38** 808–835. MR2604697
CHIANI, M. (2016). Distribution of the largest root of a matrix for Roy's test in multivariate analysis of variance. *J. Multivariate Anal.* **143** 467–471.
DEMPSTER, A. P. (1958). A high dimensional two sample significance test. *Ann. Math. Statist.* **29** 995–1010. MR0112207
DEMPSTER, A. P. (1960). A significance test for the separation of two highly multivariate small samples. *Biometrics* **16** 41–50.
FUJIKOSHI, Y. (1977). Asymptotic expansions for the distributions of the latent roots in MANOVA and the canonical correlations. *J. Multivariate Anal.* **7** 386–396.
FUJIKOSHI, Y., HIMENO, T. and WAKAKI, H. (2004). Asymptotic results of a high-dimensional MANOVA test and power comparison when the dimension is large compared to the sample size. *J. Japan Statist. Soc.* **34** 19–26.
FUJIKOSHI, Y., HIMENO, T. and WAKAKI, H. (2008). Asymptotic results in canonoical discriminat analysis when the dimension is large compared to the sample. *J. Statist. Plann. Inference* **138** 3457–3466.
GLYNN, W. J. and MUIRHEAD, R. J. (1978). Inference in canonical correlation analysis. *J. Multivariate Anal.* **8** 468–478. MR0512614
HSU, P. L. (1941). On the limiting distribution of roots of a determinantal equation. *J. Lond. Math. Soc.* **16** 183–194.
HU, J. and BAI, Z. D. (2016). A review of 20 years of naive tests of significance for high-dimensional mean vectors and covariance matrices. *Sci. China Math.* **59** 2281–2300.
JOHNSTONE, I. M. (2008). Multivariate analysis and Jacobi ensembles: Largest eigenvalue, Tracy–Widom limits and rate of convergence. *Ann. Statist.* **36** 2638–2716. MR2485010
KRISHNAIAH, P. R. and CHANG, T. C. (1971). On the exact distributions of the extreme roots of the Wishart and MANOVA matrices. *J. Multivariate Anal.* **1** 108–117. MR0301860
LI, J. and CHEN, S. X. (2012). Two sample tests for high-dimensional covariance matrices. *Ann. Statist.* **40** 908–940. MR2985938
MUIRHEAD, R. J. (1978). Latent roots and matrix variates. *Ann. Statist.* **6** 5–33.
MUIRHEAD, R. J. (1982). *Aspect of Multivariate Statistical Theory*. Wiley, New York.
PAN, G. M. and ZHOU, W. (2011). Central limit theorem for Hotelling's T^2 statistic under large dimension. *Ann. Appl. Probab.* **21** 1860–1910. MR2884053
SCHOTT, J. R. (2007). Some high-dimensional tests for a one-way MANOVA. *J. Multivariate Anal.* **98** 1825–1839.
SEO, T., KANDA, T. and FUJIKOSHI, Y. (1994). Asymptotic distributions of the sample roots in MANOVA models under nonnormality. *J. Japan Statist. Soc.* **24** 133–140.
SILVERSTEIN, J. (1995). Strong convergence of empirical distribution of eigenvlaues of large dimensional random matrices. *J. Multivariate Anal.* **55** 331–339.
SRIVASTAVA, M. S. and DU, M. (2008). A test for the mean vector with fewer observations than the dimension under non-normality. *J. Multivariate Anal.* **99** 386–402.
SRIVASTAVA, M. S. and FUJIKOSHI, Y. (2006). Multivariate analysis of variance with fewer observations than the dimension. *J. Multivariate Anal.* **97** 1927–1940.
SRIVASTAVA, M. S. and KUBOKAWA, T. (2013). Tests for multivariate analysis of variance in high dimension under non-normality. *J. Multivariate Anal.* **115** 204–216.
SUGIURA, N. (1976). Asymptotic expansions of the distributions of the latent roots and latent vectors of the Wishart and multivariate F matrices. *J. Multivariate Anal.* **6** 500–525.
ULLAH, I. and JONES, B. (2015). Regularised MANOVA for high-dimensional data. *Aust. N. Z. J. Stat.* **57** 377–389. MR3405507
WAKAKI, H., FUJIKOSHI, Y. and ULYANOV, V. (2014). Asymptotic expansions of the distributions of MANOVA test statistics when the dimension is large. *Hiroshima Math. J.* **44** 247–259.

WANG, L., PENG, B. and LI, R. (2015). A high-dimensional nonparamteric multivariate test for mean vector. *J. Amer. Statist. Assoc.* **110** 1658–1669.

ZHENG, S. (2012). Central limit theorems for linear spectral statistics of large-dimensional F-matrices. *Ann. Inst. Henri Poincaré Probab. Stat.* **47** 444–476. MR2954263

The Annals of Statistics
2018, Vol. 46, No. 6A, 3014–3037
https://doi.org/10.1214/17-AOS1647

TWO-SAMPLE KOLMOGOROV–SMIRNOV-TYPE TESTS REVISITED: OLD AND NEW TESTS IN TERMS OF LOCAL LEVELS[1]

BY HELMUT FINNER* AND VERONIKA GONTSCHARUK†

Institute for Biometrics and Epidemiology, German Diabetes Center and Institute for Health Services Research and Health Economics, German Diabetes Center†*

From a multiple testing viewpoint, Kolmogorov–Smirnov (KS)-type tests are union-intersection tests which can be redefined in terms of local levels. The local level perspective offers a new viewpoint on ranges of sensitivity of KS-type tests and the design of new tests. We study the finite and asymptotic local level behavior of weighted KS tests which are either tail, intermediate or central sensitive. Furthermore, we provide new tests with approximately equal local levels and prove that the asymptotics of such tests with sample sizes m and n coincides with the asymptotics of one-sample higher criticism tests with sample size $\min(m, n)$. We compare the overall power of various tests and introduce local powers that are in line with local levels. Finally, suitably parameterized local level shape functions can be used to design new tests. We illustrate how to combine tests with different sensitivity in terms of local levels.

REFERENCES

[1] ALDOR-NOIMAN, S., BROWN, L. D., BUJA, A., ROLKE, W. and STINE, R. A. (2013). The power to see: A new graphical test of normality. *Amer. Statist.* **67** 249–260. MR3303820

[2] ALDOR-NOIMAN, S., BROWN, L. D., BUJA, A., ROLKE, W. and STINE, R. A. (2014). Correction to: "The power to see: A new graphical test of normality." [*Amer. Statist.* **67**(4) (2013) 249–260. MR3303820] *Amer. Statist.* **68** 318. MR3280623

[3] BARNARD, G. A. (1947). Significance test for 2×2 tables. *Biometrika* **34** 123–138. MR0019285

[4] BERK, R. H. and JONES, D. H. (1978). Relatively optimal combinations of test statistics. *Scand. J. Stat.* **5** 158–162. MR0509452

[5] BERK, R. H. and JONES, D. H. (1979). Goodness-of-fit test statistics that dominate the Kolmogorov statistics. *Z. Wahrsch. Verw. Gebiete* **47** 47–59. MR0521531

[6] CANNER, P. L. (1975). A simulation study of one- and two-sample Kolmogorov–Smirnov statistics with a particular weight function. *J. Amer. Statist. Assoc.* **70** 209–211.

[7] CSÖRGŐ, M., CSÖRGŐ, S., HORVÁTH, L. and MASON, D. M. (1986). Weighted empirical and quantile processes. *Ann. Probab.* **14** 31–85. MR0815960

[8] DOKSUM, K. A. and SIEVERS, G. L. (1976). Plotting with confidence: Graphical comparisons of two populations. *Biometrika* **63** 421–434. MR0443210

MSC2010 subject classifications. Primary 62G10, 62G30; secondary 62G20, 60F99.

Key words and phrases. Goodness-of-fit, higher criticism test, local levels, multiple hypotheses testing, nonparametric two-sample tests, order statistics, weighted Brownian bridge.

[9] DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. MR2065195

[10] DONOHO, D. and JIN, J. (2015). Higher criticism for large-scale inference, especially for rare and weak effects. *Statist. Sci.* **30** 1–25. MR3317751

[11] FINNER, H. and GONTSCHARUK, V. (2018). Supplement A to "Two-sample Kolmogorov-Smirnov type tests revisited: Old and new tests in terms of local levels": Proofs and computation of global levels. DOI:10.1214/17-AOS1647SUPPA.

[12] FINNER, H. and GONTSCHARUK, V. (2018). Supplement B to "Two-sample Kolmogorov-Smirnov type tests revisited: Old and new tests in terms of local levels": Animated graphics of local levels. DOI:10.1214/17-AOS1647SUPPB.

[13] FINNER, H. and STRASSBURGER, K. (2002). Structural properties of UMPU-tests for 2×2-tables and some applications. *J. Statist. Plann. Inference* **104** 103–120. MR1900521

[14] FREY, J. (2008). Optimal distribution-free confidence bands for a distribution function. *J. Statist. Plann. Inference* **138** 3086–3098. MR2442227

[15] GONTSCHARUK, V. and FINNER, H. (2017). Asymptotics of goodness-of-fit tests based on minimum p-value statistics. *Comm. Statist. Theory Methods* **46** 2332–2342. MR3576717

[16] GONTSCHARUK, V., LANDWEHR, S. and FINNER, H. (2015). The intermediates take it all: Asymptotics of higher criticism statistics and a powerful alternative based on equal local levels. *Biom. J.* **57** 159–180. MR3298224

[17] GONTSCHARUK, V., LANDWEHR, S. and FINNER, H. (2016). Goodness of fit tests in terms of local levels with special emphasis on higher criticism tests. *Bernoulli* **22** 1331–1363. MR3474818

[18] HODGES, J. L. (1958). The significance probability of the Smirnov two-sample test. *Ark. Mat.* **3** 469–486. MR0097136

[19] JAGER, L. and WELLNER, J. A. (2004). On the "Poisson boundaries" of the family of weighted Kolmogorov statistics. In *A Festschrift for Herman Rubin. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **45** 319–331. IMS, Beachwood, OH. MR2126907

[20] JANSSEN, A. (2000). Global power functions of goodness of fit tests. *Ann. Statist.* **28** 239–253. MR1762910

[21] KOLMOGOROV, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *G. Inst. Ital. Attuari* **4** 83–91. Translated by Q. Meneghini as On the empirical determination of a distribution function. In *Breakthroughs in Statistics* **II**. *Springer Series in Statistics* (*Perspectives in Statistics*) (S. Kotz and N. L. Johnson, eds.) 106–113. Springer, New York. MR1182795

[22] MASON, D. M. (1983). The asymptotic distribution of weighted empirical distribution functions. *Stochastic Process. Appl.* **15** 99–109. MR0694539

[23] MASON, D. M. and SCHUENEMEYER, J. H. (1983). A modified Kolmogorov–Smirnov test sensitive to tail alternatives. *Ann. Statist.* **11** 933–946. MR0707943

[24] MASON, D. M. and SCHUENEMEYER, J. H. (1992). Correction to: "A modified Kolmogorov–Smirnov test sensitive to tail alternatives." [*Ann. Statist.* **11**(3) (1983) 933–946. MR0707943] *Ann. Statist.* **20** 620–621. MR1150371

[25] PYKE, R. (1959). The supremum and infimum of the Poisson process. *Ann. Math. Stat.* **30** 568–576. MR0107315

[26] SMIRNOFF, N. (1939). Sur les écarts de la courbe de distribution empirique. *Rec. Math. N.S.* [*Mat. Sbornik*] **6**(48) 3–26. MR0001483

[27] SMIRNOV, N. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Moscow Univ. Math. Bull.* **2** 3–16. MR0002062

[28] STECK, G. P. (1969). The Smirnov two sample tests as rank tests. *Ann. Math. Stat.* **40** 1449–1466. MR0246473

[29] XU, X., DING, X. and ZHAO, S. (2009). The reduction of the average width of confidence bands for an unknown continuous distribution function. *J. Stat. Comput. Simul.* **79** 335–347. MR2522355

The Annals of Statistics
2018, Vol. 46, No. 6A, 3038–3066
https://doi.org/10.1214/17-AOS1648

ROBUST GAUSSIAN STOCHASTIC PROCESS EMULATION[1]

BY MENGYANG GU, XIAOJING WANG AND JAMES O. BERGER

Johns Hopkins University, University of Connecticut and Duke University

We consider estimation of the parameters of a Gaussian Stochastic Process (GaSP), in the context of emulation (approximation) of computer models for which the outcomes are real-valued scalars. The main focus is on estimation of the GaSP parameters through various generalized maximum likelihood methods, mostly involving finding posterior modes; this is because full Bayesian analysis in computer model emulation is typically prohibitively expensive.

The posterior modes that are studied arise from objective priors, such as the reference prior. These priors have been studied in the literature for the situation of an isotropic covariance function or under the assumption of separability in the design of inputs for model runs used in the GaSP construction. In this paper, we consider more general designs (e.g., a Latin Hypercube Design) with a class of commonly used anisotropic correlation functions, which can be written as a product of isotropic correlation functions, each having an unknown range parameter and a fixed roughness parameter. We discuss properties of the objective priors and marginal likelihoods for the parameters of the GaSP and establish the posterior propriety of the GaSP parameters, but our main focus is to demonstrate that certain parameterizations result in more robust estimation of the GaSP parameters than others, and that some parameterizations that are in common use should clearly be avoided. These results are applicable to many frequently used covariance functions, for example, power exponential, Matérn, rational quadratic and spherical covariance. We also generalize the results to the GaSP model with a nugget parameter. Both theoretical and numerical evidence is presented concerning the performance of the studied procedures.

REFERENCES

[1] AN, J. and OWEN, A. (2001). Quasi-regression. *J. Complexity* **17** 588–607. MR1881660
[2] ANDRIANAKIS, I. and CHALLENOR, P. G. (2012). The effect of the nugget on Gaussian process emulators of computer models. *Comput. Statist. Data Anal.* **56** 4215–4228. MR2957866
[3] BAYARRI, M., BERGER, J., CAFEO, J., GARCIA-DONATO, G., LIU, F., PALOMO, J., PARTHASARATHY, R., PAULO, R., SACKS, J. and WALSH, D. (2007). Computer model validation with functional output. *Ann. Statist.* **35** 1874–1906. MR2363956
[4] BAYARRI, M. J., BERGER, J. O., CALDER, E. S., DALBEY, K., LUNAGOMEZ, S., PATRA, A. K., PITMAN, E. B., SPILLER, E. T. and WOLPERT, R. L. (2009). Using statistical and computer models to quantify volcanic hazards. *Technometrics* **51** 402–413. MR2756476

MSC2010 subject classifications. 62F35, 62M20.
Key words and phrases. Anisotropic covariance, emulation, Gaussian stochastic process, objective priors, posterior propriety, robust parameter estimation.

[5] BERGER, J. O., DE OLIVEIRA, V. and SANSÓ, B. (2001). Objective Bayesian analysis of spatially correlated data. *J. Amer. Statist. Assoc.* **96** 1361–1374.

[6] DETTE, H. and PEPELYSHEV, A. (2010). Generalized latin hypercube design for computer experiments. *Technometrics* **52** 421–429.

[7] DE OLIVEIRA, V. (2007). Objective Bayesian analysis of spatial data with measurement error. *Canad. J. Statist.* **35** 283–301.

[8] DIGGLE, P. and RIBEIRO, P. (2007). *Model-Based Geostatistics*. Springer, Berlin.

[9] DIXON, L. (1978). The global optimization problem: An introduction. In *Towards Global Optimiation* **2** 1–15. North-Hollad, Amsterdam.

[10] GELFAND, A. E. (2010). *Handbook of Spatial Statistics*. CRC Press, Boca Raton, FL.

[11] GRAMACY, R. B. and LEE, H. K. (2009). Adaptive design and analysis of supercomputer experiments. *Technometrics* **51** 130–145.

[12] GU, M., BERGER, J. O. et al. (2016). Parallel partial Gaussian process emulation for computer models with massive output. *Ann. Appl. Stat.* **10** 1317–1347. MR3553226

[13] GU, M., PALOMO, J. and BERGER, J. (2016). RobustGaSP: Robust Gaussian stochastic process emulation. R package version 0.5.

[14] GU, M., WANG, X. and BERGER, J. O. (2018). Supplement to "Robust Gaussian stochastic process emulation." DOI:10.1214/17-AOS1648SUPP.

[15] HANDCOCK, M. S. and STEIN, M. L. (1993). A Bayesian analysis of kriging. *Technometrics* **35** 403–410.

[16] HANDCOCK, M. S. and WALLIS, J. R. (1994). An approach to statistical spatial-temporal modeling of meteorological fields. *J. Amer. Statist. Assoc.* **89** 368–378.

[17] HIGDON, D. (2002). Space and space–time modeling using process convolutions. In *Quantitative Methods for Current Environmental Issues* 37–56. Springer, London. MR2059819

[18] KAZIANKA, H. (2013). Objective Bayesian analysis of geometrically anisotropic spatial data. *J. Agric. Biol. Environ. Stat.* **18** 514–537. MR3142598

[19] KAZIANKA, H. and PILZ, J. (2012). Objective Bayesian analysis of spatial data with uncertain nugget and range parameters. *Canad. J. Statist.* **40** 304–327. MR2927748

[20] KENNEDY, M. C. and O'HAGAN, A. (2001). Bayesian calibration of computer models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 425–464. MR1858398

[21] LI, R. and SUDJIANTO, A. (2005). Analysis of computer experiments using penalized likelihood in Gaussian kriging models. *Technometrics* **47** 111–120.

[22] LINKLETTER, C., BINGHAM, D., HENGARTNER, N., HIGDON, D. and KENNY, Q. Y. (2006). Variable selection for Gaussian process models in computer experiments. *Technometrics* **48** 478–490.

[23] LOPES, D. (2011). Development and implementation of Bayesian computer model emulators. Ph.D. thesis, Duke Univ., Durham, NC.

[24] MORRIS, M. D., MITCHELL, T. J. and YLVISAKER, D. (1993). Bayesian design and analysis of computer experiments: Use of derivatives in surface prediction. *Technometrics* **35** 243–255.

[25] OAKLEY, J. (1999). Bayesian uncertainty analysis for complex computer codes. Ph.D. thesis, Univ. Sheffield.

[26] OAKLEY, J. and O'HAGAN, A. (2002). Bayesian inference for the uncertainty distribution of computer model outputs. *Biometrika* **89** 769–784.

[27] PACIOREK, C. J. and SCHERVISH, M. J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics* **17** 483–506.

[28] PAULO, R. (2005). Default priors for Gaussian processes. *Ann. Statist.* **33** 556–582. MR2163152

[29] PENG, C.-Y. and WU, C. J. (2014). On the choice of nugget in kriging modeling for deterministic computer experiments. *J. Comput. Graph. Statist.* **23** 151–168. MR3173765

[30] QIAN, P. Z. G., WU, H. and WU, C. F. J. (2008). Gaussian process models for computer experiments with qualitative and quantitative factors. *Technometrics* **50** 383–396.

[31] RANJAN, H. R. and KARSTEN, R. (2011). A computationally stable approach to Gaussian process interpolation of deterministic computer simulation data. *Technometrics* **53** 366–378.

[32] REN, C., SUN, D. and HE, C. (2012). Objective Bayesian analysis for a spatial model with nugget effects. *J. Statist. Plann. Inference* **142** 1933–1946. MR2903403

[33] REN, C., SUN, D. and SAHU, S. K. (2013). Objective Bayesian analysis of spatial models with separable correlation functions. *Canad. J. Statist.* **41** 488–507.

[34] ROUSTANT, O., GINSBOURGER, D. and DEVILLE, Y. (2012). DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. *J. Stat. Softw.* **51** 1–55.

[35] 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

[36] SANTNER, T. J., WILLIAMS, B. J. and NOTZ, W. I. (2003). *The Design and Analysis of Computer Experiments*. Springer, New York. MR2160708

[37] SPILLER, E. T., BAYARRI, M., BERGER, J. O., CALDER, E. S., PATRA, A. K., PITMAN, E. B. and WOLPERT, R. L. (2014). Automating emulator construction for geophysical hazard maps. *SIAM/ASA J. Uncertain. Quantif.* **2** 126–152.

[38] STEIN, M. L. (2012). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Science & Business Media, Berlin.

[39] SURJANOVIC, S. and BINGHAM, D. (2017). Virtual library of simulation experiments: Test functions and datasets. Retrieved June 26, 2017, from http://www.sfu.ca/~ssurjano.

[40] ZHANG, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *J. Amer. Statist. Assoc.* **99** 250–261.

[41] ZHANG, H. and ZIMMERMAN, D. L. (2005). Towards reconciling two asymptotic frameworks in spatial statistics. *Biometrika* **92** 921–936. MR2234195

[42] ZIMMERMAN, D. L. (1993). Another look at anisotropy in geostatistics. *Math. Geol.* **25** 453–470.

The Annals of Statistics
2018, Vol. 46, No. 6A, 3067–3098
https://doi.org/10.1214/17-AOS1649

CONVERGENCE OF CONTRASTIVE DIVERGENCE ALGORITHM IN EXPONENTIAL FAMILY[1]

BY BAI JIANG, TUNG-YU WU, YIFAN JIN AND WING H. WONG[2]

Stanford University

The Contrastive Divergence (CD) algorithm has achieved notable success in training energy-based models including Restricted Boltzmann Machines and played a key role in the emergence of deep learning. The idea of this algorithm is to approximate the intractable term in the exact gradient of the log-likelihood function by using short Markov chain Monte Carlo (MCMC) runs. The approximate gradient is computationally-cheap but biased. Whether and why the CD algorithm provides an asymptotically consistent estimate are still open questions. This paper studies the asymptotic properties of the CD algorithm in canonical exponential families, which are special cases of the energy-based model. Suppose the CD algorithm runs m MCMC transition steps at each iteration t and iteratively generates a sequence of parameter estimates $\{\theta_t\}_{t\geq 0}$ given an i.i.d. data sample $\{X_i\}_{i=1}^n \sim p_{\theta_\star}$. Under conditions which are commonly obeyed by the CD algorithm in practice, we prove the existence of some bounded m such that any limit point of the time average $\sum_{s=0}^{t-1}\theta_s/t$ as $t \to \infty$ is a consistent estimate for the true parameter $\theta_\star$. Our proof is based on the fact that $\{\theta_t\}_{t\geq 0}$ is a homogenous Markov chain conditional on the data sample $\{X_i\}_{i=1}^n$. This chain meets the Foster–Lyapunov drift criterion and converges to a random walk around the maximum likelihood estimate. The range of the random walk shrinks to zero at rate $\mathcal{O}(1/\sqrt[3]{n})$ as the sample size $n \to \infty$.

REFERENCES

[1] ACKLEY, D. H., HINTON, G. E. and SEJNOWSKI, T. J. (1985). A learning algorithm for Boltzmann machines. *Cogn. Sci.* **9** 147–169.
[2] AMIT, Y. (1996). Convergence properties of the Gibbs sampler for perturbations of Gaussians. *Ann. Statist.* **24** 122–140. MR1389883
[3] ASUNCION, A., LIU, Q., IHLER, A. and SMYTH, P. (2010). Learning with blocks: Composite likelihood and contrastive divergence. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* 33–40.
[4] ATCHADÉ, Y. F., FORT, G. and MOULINES, E. (2017). On perturbed proximal gradient algorithms. *J. Mach. Learn. Res.* **18** Paper No. 10, 33. MR3634877
[5] AZUMA, K. (1967). Weighted sums of certain dependent random variables. *Tôhoku Math. J.* (2) **19** 357–367. MR0221571
[6] BECK, A. and TEBOULLE, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2** 183–202. MR2486527
[7] BENGIO, Y. and DELALLEAU, O. (2009). Justifying and generalizing contrastive divergence. *Neural Comput.* **21** 1601–1621. MR2527797

MSC2010 subject classifications. Primary 68W99, 62F12; secondary 60J20.
Key words and phrases. Contrastive Divergence, exponential family, convergence rate.

[8] BENGIO, Y., LAMBLIN, P., POPOVICI, D., LAROCHELLE, H. (2007). Greedy layer-wise training of deep networks. In *Proceedings of the* 19*th International Conference on Neural Information Processing Systems* (*NIPS'*06) 153–160. MIT Press, Cambridge, MA.

[9] CARREIRA-PERPINAN, M. A. and HINTON, G. E. (2005). On contrastive divergence learning. In *AISTATS* **10** 33–40.

[10] COATES, A., NG, A. and LEE, H. (2011). An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (G. Gordon, D. Dunson and M. Dudik, eds.). *Proceedings of Machine Learning Research* **15** 215–223.

[11] CONWAY, J. B. (1990). *A Course in Functional Analysis*, 2nd ed. *Graduate Texts in Mathematics* **96**. Springer, New York. MR1070713

[12] DIACONIS, P. (2009). The Markov chain Monte Carlo revolution. *Bull*. *Amer. Math*. *Soc*. (*N.S.*) **46** 179–205. MR2476411

[13] HAIRER, M. and MATTINGLY, J. C. (2011). Yet another look at Harris' ergodic theorem for Markov chains. In *Seminar on Stochastic Analysis*, *Random Fields and Applications VI*. *Progress in Probability* **63** 109–117. Birkhäuser, Basel. MR2857021

[14] HE, X., ZEMEL, R. S. and CARREIRA-PERPIÑÁN, M. Á. (2004). Multiscale conditional random fields for image labeling. In *Proceedings of the* 2004 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (*CVPR* 2004) **2** 692–702. IEEE, New York.

[15] HINTON, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Comput*. **14** 1771–1800.

[16] HINTON, G. E., OSINDERO, S. and TEH, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Comput*. **18** 1527–1554.

[17] HINTON, G. E. and SALAKHUTDINOV, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science* **313** 504–507. MR2242509

[18] HINTON, G. E. and SALAKHUTDINOV, R. R. (2009). Replicated softmax: An undirected topic model. In *Advances in Neural Information Processing Systems* 1607–1614.

[19] HUNTER, D. R., HANDCOCK, M. S., BUTTS, C. T., GOODREAU, S. M. and MORRIS, M. (2008). ERGM: A package to fit, simulate and diagnose exponential-family models for networks. *J. Stat*. *Softw*. **24** 1–29 nihpa54860.

[20] HYVÄRINEN, A. (2006). Consistency of pseudolikelihood estimation of fully visible Boltzmann machines. *Neural Comput*. **18** 2283–2292.

[21] JIANG, B., WU, T.-Y., JIN, Y. and WONG, W. H. (2018). Supplement to "Convergence of contrastive divergence algorithm in exponential family." DOI:10.1214/17-AOS1649SUPP.

[22] KONTOYIANNIS, I. and MEYN, S. P. (2012). Geometric ergodicity and the spectral gap of non-reversible Markov chains. *Probab. Theory Related Fields* **154** 327–339.

[23] KRIVITSKY, P. N. (2017). Using contrastive divergence to seed Monte Carlo MLE for exponential-family random graph models. *Comput*. *Statist*. *Data Anal*. **107** 149–161. MR3575065

[24] LAROCHELLE, H. and BENGIO, Y. (2008). Classification using discriminative restricted Boltzmann machines. In *Proceedings of the* 25*th International Conference on Machine Learning* 536–543. ACM, New York.

[25] LEHMANN, E. L. and CASELLA, G. (1991). *Theory of Point Estimation*. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA. Reprint of the 1983 original. MR1138212

[26] MACKAY, D. (2001). Failures of the one-step learning algorithm. Technical report, Available at http://www.inference.phy.cam.ac.uk/mackay/abstracts/gbm.html.

[27] MENGERSEN, K. L. and TWEEDIE, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Ann*. *Statist*. **24** 101–121. MR1389882

[28] MEYN, S. P. and TWEEDIE, R. L. (1992). Stability of Markovian processes I: Criteria for discrete-time chains. *Adv. in Appl. Probab.* **24** 542–574. MR1174380

[29] MEYN, S. P. and TWEEDIE, R. L. (1993). Stability of Markovian processes III: Foster–Lyapunov criteria for continuous-time processes. *Adv. in Appl. Probab.* **25** 518–548.

[30] MOHAMED, A.-R., DAHL, G. E. and HINTON, G. (2012). Acoustic modeling using deep belief networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **20** 14–22.

[31] PARIKH, N., BOYD, S. P. et al. (2014). Proximal algorithms. *Found. Trends Optim.* **1** 127–239.

[32] RIGOLLET, P. (2015). Lecture notes in high dimensional statistics.

[33] ROBERTS, G. O. and ROSENTHAL, J. S. (1997). Geometric ergodicity and hybrid Markov chains. *Electron. Commun. Probab.* **2** 13–25. MR1448322

[34] ROBERTS, G. O. and ROSENTHAL, J. S. (2004). General state space Markov chains and MCMC algorithms. *Probab. Surv.* **1** 20–71. MR2095565

[35] ROBERTS, G. O. and TWEEDIE, R. L. (1996). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika* **83** 95–110.

[36] ROBINS, G., PATTISON, P., KALISH, Y. and LUSHER, D. (2007). An introduction to exponential random graph (p*) models for social networks. *Soc. Netw.* **29** 173–191.

[37] ROTH, S. and BLACK, M. J. (2005). Fields of experts: A framework for learning image priors. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (*CVPR* 2005) **2** 860–867. IEEE, New York.

[38] RUDOLF, D. (2011). Explicit error bounds for Markov chain Monte Carlo. ArXiv preprint. Available at arXiv:1108.3201.

[39] SALAKHUTDINOV, R., MNIH, A. and HINTON, G. (2007). Restricted Boltzmann machines for collaborative filtering. In *Proceedings of the* 24*th International Conference on Machine Learning* (*ICML*'07) 791–798. ACM, New York.

[40] SUTSKEVER, I. and TIELEMAN, T. (2010). On the convergence properties of contrastive divergence. In *International Conference on Artificial Intelligence and Statistics* 789–795.

[41] TEH, Y. W., WELLING, M., OSINDERO, S. and HINTON, G. E. (2004). Energy-based models for sparse overcomplete representations. *J. Mach. Learn. Res.* **4** 1235–1260. MR2103628

[42] VÁRNAI, C., BURKOFF, N. S. and WILD, D. L. (2013). Efficient parameter estimation of generalizable coarse-grained protein force fields using contrastive divergence: A maximum likelihood approach. *J. Chem. Theory Comput.* **9** 5718–5733.

[43] WILLIAMS, C. K. I. and AGAKOV, F. V. (2002). An analysis of contrastive divergence learning in Gaussian Boltzmann machines. Working paper, Institute for Adaptive and Neural Computation, Edinburgh.

[44] YUILLE, A. L. (2005). The convergence of contrastive divergences. In *Proceedings of the* 17*th International Conference on Neural Information Processing Systems* (*NIPS*'04) 1593–1600. MIT Press, Cambridge, MA.

The Annals of Statistics
2018, Vol. 46, No. 6A, 3099–3129
https://doi.org/10.1214/17-AOS1651

OVERCOMING THE LIMITATIONS OF PHASE TRANSITION BY HIGHER ORDER ANALYSIS OF REGULARIZATION TECHNIQUES[1]

BY HAOLEI WENG, ARIAN MALEKI AND LE ZHENG

Columbia University

We study the problem of estimating a sparse vector $\beta \in \mathbb{R}^p$ from the response variables $y = X\beta + w$, where $w \sim N(0, \sigma_w^2 I_{n\times n})$, under the following high-dimensional asymptotic regime: given a fixed number δ, $p \to \infty$, while $n/p \to \delta$. We consider the popular class of ℓ_q-regularized least squares (LQLS), a.k.a. bridge estimators, given by the optimization problem

$$\hat{\beta}(\lambda, q) \in \arg\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_q^q,$$

and characterize the almost sure limit of $\frac{1}{p}\|\hat{\beta}(\lambda, q) - \beta\|_2^2$, and call it asymptotic mean square error (AMSE). The expression we derive for this limit does not have explicit forms, and hence is not useful in comparing LQLS for different values of q, or providing information in evaluating the effect of δ or sparsity level of β. To simplify the expression, researchers have considered the ideal "error-free" regime, that is, $w = 0$, and have characterized the values of δ for which AMSE is zero. This is known as the phase transition analysis.

In this paper, we first perform the phase transition analysis of LQLS. Our results reveal some of the limitations and misleading features of the phase transition analysis. To overcome these limitations, we propose the small error analysis of LQLS. Our new analysis framework not only sheds light on the results of the phase transition analysis, but also describes when phase transition analysis is reliable, and presents a more accurate comparison among different regularizers.

REFERENCES

[1] AMELUNXEN, D., LOTZ, M., MCCOY, M. B. and TROPP, J. A. (2014). Living on the edge: Phase transitions in convex programs with random data. *Inf. Inference* **3** 224–294. MR3311453

[2] BAYATI, M. and MONTANARI, A. (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Trans. Inform. Theory* **57** 764–785. MR2810285

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

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

MSC2010 subject classifications. Primary 62J05, 62J07.

Key words and phrases. Bridge regression, phase transition, comparison of estimators, small error regime, second-order term, asymptotic mean square error, optimal tuning.

[5] BOGDAN, M., VAN DEN BERG, E., SABATTI, C., SU, W. and CANDÈS, E. J. (2015). SLOPE—adaptive variable selection via convex optimization. *Ann. Appl. Stat.* **9** 1103–1140. MR3418717

[6] BRADIC, J. (2016). Robustness in sparse high-dimensional linear models: Relative efficiency and robust approximate message passing. *Electron. J. Stat.* **10** 3894–3944. MR3581957

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

[8] BUNEA, F., TSYBAKOV, A. and WEGKAMP, M. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.* **1** 169–194. MR2312149

[9] CANDÈS, E. J. (2008). The restricted isometry property and its implications for compressed sensing. *C. R. Math. Acad. Sci. Paris* **346** 589–592. MR2412803

[10] COOLEN, A. C. C. (2005). *The Mathematical Theory of Minority Games: Statistical Mechanics of Interacting Agents*. Oxford Univ. Press, Oxford. MR2127290

[11] DONOHO, D. and MONTANARI, A. (2015). Variance breakdown of Huber (M)-estimators. $n/p \to m$. Preprint.

[12] DONOHO, D. and MONTANARI, A. (2016). High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *Probab. Theory Related Fields* **166** 935–969. MR3568043

[13] DONOHO, D. L. (2006). High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension. *Discrete Comput. Geom.* **35** 617–652. MR2225676

[14] DONOHO, D. L. (2006). For most large underdetermined systems of equations, the minimal l_1-norm near-solution approximates the sparsest near-solution. *Comm. Pure Appl. Math.* **59** 907–934. MR2222440

[15] DONOHO, D. L., GAVISH, M. and MONTANARI, A. (2013). The phase transition of matrix recovery from Gaussian measurements matches the minimax MSE of matrix denoising. *Proc. Natl. Acad. Sci. USA* **110** 8405–8410. MR3082268

[16] DONOHO, D. L., MALEKI, A. and MONTANARI, A. (2009). Message-passing algorithms for compressed sensing. *Proc. Natl. Acad. Sci. USA* **106** 18914–18919.

[17] 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

[18] 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

[19] DONOHO, D. L. and TANNER, J. (2005). Sparse nonnegative solution of underdetermined linear equations by linear programming. *Proc. Natl. Acad. Sci. USA* **102** 9446–9451. MR2168715

[20] DONOHO, D. L. and TANNER, J. (2005). Neighborliness of randomly projected simplices in high dimensions. *Proc. Natl. Acad. Sci. USA* **102** 9452–9457. MR2168716

[21] EL KAROUI, N. (2013). Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: Rigorous results. Preprint. Available at arXiv:1311.2445.

[22] EL KAROUI, N., BEAN, D., BICKEL, P., LIM, C. and YU, B. (2013). On robust regression with high-dimensional predictors. *Proc. Natl. Acad. Sci. USA* **110** 14557–14562.

[23] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

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

[25] FU, W. J. (1998). Penalized regressions: The bridge versus the lasso. *J. Comput. Graph. Statist.* **7** 397–416. MR1646710

[26] GUO, D. and VERDÚ, S. (2005). Randomly spread CDMA: Asymptotics via statistical physics. *IEEE Trans. Inform. Theory* **51** 1983–2010. MR2235278

[27] HOERL, A. and KENNARD, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

[28] HUANG, J., HOROWITZ, J. L. and MA, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* **36** 587–613. MR2396808

[29] KABASHIMA, Y., WADAYAMA, T. and TANAKA, T. (2009). A typical reconstruction limit for compressed sensing based on L_p-norm minimization. *J. Stat. Mech. Theory Exp.* **2009** L09003.

[30] KNIGHT, K. and FU, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28** 1356–1378. MR1805787

[31] KOLTCHINSKII, V. (2009). Sparsity in penalized empirical risk minimization. *Ann. Inst. Henri Poincaré Probab. Stat.* **45** 7–57. MR2500227

[32] KRZAKALA, F., MÉZARD, M., SAUSSET, F., SUN, Y. and ZDEBOROVÁ, L. (2012). Statistical-physics-based reconstruction in compressed sensing. *Phys. Rev. X* **2** 021005.

[33] MALEKI, A. (2010). Approximate message passing algorithms for compressed sensing. Ph.D. thesis, Stanford Univ., Stanford, CA.

[34] MOUSAVI, A., MALEKI, A. and BARANIUK, R. G. (2017). Consistent parameter estimation for LASSO and approximate message passing. *Ann. Statist.* **45** 2427–2454. MR3737897

[35] OYMAK, S. and HASSIBI, B. (2016). Sharp MSE bounds for proximal denoising. *Found. Comput. Math.* **16** 965–1029. MR3529131

[36] OYMAK, S., THRAMPOULIDIS, C. and HASSIBI, B. (2013). The squared-error of generalized lasso: A precise analysis. In 51*st Annual Allerton Conference on Communication, Control, and Computing* (*Allerton*) 1002–1009. IEEE, New York.

[37] RANGAN, S., GOYAL, V. and FLETCHER, A. (2009). Asymptotic analysis of map estimation via the replica method and compressed sensing. In *Advances in Neural Information Processing Systems* 1545–1553.

[38] 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

[39] STOJNIC, M. (2009). Various thresholds for ℓ_1-optimization in compressed sensing. Preprint. Available at arXiv:0907.3666.

[40] SU, W., BOGDAN, M. and CANDES, E. (2015). False discoveries occur early on the lasso path. Preprint. Available at arXiv:1511.01957.

[41] TANAKA, T. (2002). A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors. *IEEE Trans. Inform. Theory* **48** 2888–2910. MR1945581

[42] THRAMPOULIDIS, C., ABBASI, E. and HASSIBI, B. (2016). Precise error analysis of regularized M-estimators in high-dimensions. Preprint. Available at arXiv:1601.06233.

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

[44] WENG, H. and MALEKI, A. (2017). Low noise sensitivity analysis of ℓ_q-minimization in oversampled systems. Preprint. Available at arXiv:1705.03533.

[45] WENG, H., MALEKI, A. and ZHENG, L. (2018). Supplement to "Overcoming the limitations of phase transition by higher order analysis of regularization techniques." DOI:10.1214/17-AOS1651SUPP.

[46] ZHENG, L., MALEKI, A., WENG, H., WANG, X. and LONG, T. (2016). Does ℓ_p-minimization outperform ℓ_1-minimization? Preprint. Available at arXiv:1501.03704v2.

[47] ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327

The Annals of Statistics
2018, Vol. 46, No. 6A, 3130–3150
https://doi.org/10.1214/17-AOS1653

OPTIMAL ADAPTIVE ESTIMATION OF LINEAR FUNCTIONALS UNDER SPARSITY

BY OLIVIER COLLIER[*,1], LAËTITIA COMMINGES[†], ALEXANDRE B. TSYBAKOV[‡,2] AND NICOLAS VERZELEN[§]

Université Paris-Ouest[*], *Université Paris Dauphine*[†], *CREST-ENSAE*[‡], *and INRA*[§]

We consider the problem of estimation of a linear functional in the Gaussian sequence model where the unknown vector $\theta \in \mathbb{R}^d$ belongs to a class of s-sparse vectors with unknown s. We suggest an adaptive estimator achieving a nonasymptotic rate of convergence that differs from the minimax rate at most by a logarithmic factor. We also show that this optimal adaptive rate cannot be improved when s is unknown. Furthermore, we address the issue of simultaneous adaptation to s and to the variance σ^2 of the noise. We suggest an estimator that achieves the optimal adaptive rate when both s and σ^2 are unknown.

REFERENCES

CAI, T. T. and LOW, M. G. (2004). Minimax estimation of linear functionals over nonconvex parameter spaces. *Ann. Statist.* **32** 552–576. MR2060169

CAI, T. T. and LOW, M. G. (2005). On adaptive estimation of linear functionals. *Ann. Statist.* **33** 2311–2343. MR2211088

COLLIER, O., COMMINGES, L. and TSYBAKOV, A. B. (2017). Minimax estimation of linear and quadratic functionals on sparsity classes. *Ann. Statist.* **45** 923–958. MR3662444

GOLUBEV, G. K. (2004). The method of risk envelopes in the estimation of linear functionals. *Problemy Peredachi Informatsii* **40** 58–72. MR2099020

GOLUBEV, Y. and LEVIT, B. (2004). An oracle approach to adaptive estimation of linear functionals in a Gaussian model. *Math. Methods Statist.* **13** 392–408. MR2126747

IBRAGIMOV, I. A. and KHAS'MINSKII, R. Z. (1984). Nonparametric estimation of the value of a linear functional in Gaussian white noise. *Theory Probab. Appl.* **29** 19–32. MR0739497

LAURENT, B., LUDEÑA, C. and PRIEUR, C. (2008). Adaptive estimation of linear functionals by model selection. *Electron. J. Stat.* **2** 993–1020. MR2448602

LEDOUX, M. and TALAGRAND, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes. Ergebnisse der Mathematik und Ihrer Grenzgebiete* (3) [*Results in Mathematics and Related Areas* (3)] **23**. Springer, Berlin. MR1102015

PETROV, V. V. (1995). *Limit Theorems of Probability Theory: Sequences of Independent Random Variables. Oxford Studies in Probability* **4**. Oxford Univ. Press, New York. MR1353441

TSYBAKOV, A. B. (1998). Pointwise and sup-norm sharp adaptive estimation of functions on the Sobolev classes. *Ann. Statist.* **26** 2420–2469. MR1700239

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

MSC2010 subject classifications. 62J05, 62G05.

Key words and phrases. Nonasymptotic minimax estimation, adaptive estimation, linear functional, sparsity, unknown noise variance.

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

The Annals of Statistics
2018, Vol. 46, No. 6A, 3151–3183
https://doi.org/10.1214/17-AOS1654

HIGH-DIMENSIONAL CONSISTENCY IN SCORE-BASED AND HYBRID STRUCTURE LEARNING

BY PREETAM NANDY*,[1], ALAIN HAUSER† AND MARLOES H. MAATHUIS*,[1]

ETH Zürich and Bern University of Applied Sciences†*

Main approaches for learning Bayesian networks can be classified as constraint-based, score-based or hybrid methods. Although high-dimensional consistency results are available for constraint-based methods like the PC algorithm, such results have not been proved for score-based or hybrid methods, and most of the hybrid methods have not even shown to be consistent in the classical setting where the number of variables remains fixed and the sample size tends to infinity. In this paper, we show that consistency of hybrid methods based on greedy equivalence search (GES) can be achieved in the classical setting with adaptive restrictions on the search space that depend on the current state of the algorithm. Moreover, we prove consistency of GES and adaptively restricted GES (ARGES) in several sparse high-dimensional settings. ARGES scales well to sparse graphs with thousands of variables and our simulation study indicates that both GES and ARGES generally outperform the PC algorithm.

REFERENCES

ALONSO-BARBA, J. I., DELAOSSA, L., GÁMEZ, J. A. and PUERTA, J. M. (2013). Scaling up the greedy equivalence search algorithm by constraining the search space of equivalence classes. *Internat. J. Approx. Reason.* **54** 429–451. MR3041110

ANANDKUMAR, A., TAN, V. Y. F., HUANG, F. and WILLSKY, A. S. (2012). High-dimensional Gaussian graphical model selection: Walk summability and local separation criterion. *J. Mach. Learn. Res.* **13** 2293–2337. MR2973603

ANDERSSON, S. A., MADIGAN, D. and PERLMAN, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *Ann. Statist.* **25** 505–541. MR1439312

BANERJEE, O., EL GHAOUI, L. and D'ASPREMONT, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9** 485–516. MR2417243

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

CHEN, J. and CHEN, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95** 759–771. MR2443189

CHICKERING, D. M. (2002). Learning equivalence classes of Bayesian-network structures. *J. Mach. Learn. Res.* **2** 445–498. MR1929415

CHICKERING, D. M. (2003). Optimal structure identification with greedy search. *J. Mach. Learn. Res.* **3** 507–554. MR1991085

MSC2010 subject classifications. 62H12, 62-09, 62F12.

Key words and phrases. Bayesian network, directed acyclic graph (DAG), linear structural equation model (linear SEM), structure learning, greedy equivalence search (GES), score-based method, hybrid method, high-dimensional data, consistency.

CHICKERING, D. M. and MEEK, C. (2002). Finding optimal Bayesian networks. In *UAI* 2002.

CHICKERING, D. M. and MEEK, C. (2015). Selective greedy equivalence search: Finding optimal Bayesian networks using a polynomial number of score evaluations. In *UAI* 2015.

CHOW, C. and LIU, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Trans. Inform. Theory* **14** 462–467.

COLOMBO, D. and MAATHUIS, M. H. (2014). Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.* **15** 3741–3782. MR3291411

DE CAMPOS, L. M. (1998). Independency relationships and learning algorithms for singly connected networks. *J. Exp. Theor. Artif. Intell.* **10** 511–549.

DORAN, G., MUANDET, K., ZHANG, K. and SCHÖLKOPF, B. (2014). A permutation-based kernel conditional independence test. In *UAI* 2014.

FAWCETT, T. (2006). An introduction to ROC analysis. *Pattern Recogn. Lett.* **27** 861–874.

FOYGEL, R. and DRTON, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. In *NIPS* 2010.

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

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

GAO, B. and CUI, Y. (2015). Learning directed acyclic graphical structures with genetical genomics data. *Bioinformatics* **31** 3953–3960.

HA, M. J., SUN, W. and XIE, J. (2016). `PenPC`: A two-step approach to estimate the skeletons of high-dimensional directed acyclic graphs. *Biometrics* **72** 146–155. MR3500583

HARRIS, N. and DRTON, M. (2013). PC algorithm for nonparanormal graphical models. *J. Mach. Learn. Res.* **14** 3365–3383. MR3144465

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, New York. MR2722294

HAUSER, A. and BÜHLMANN, P. (2012). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *J. Mach. Learn. Res.* **13** 2409–2464. MR2973606

HUETE, J. F. and DE CAMPOS, L. M. (1993). Learning causal polytrees. In *ECSQARU* 1993.

KALISCH, M. and BÜHLMANN, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.* **8** 613–636.

KALISCH, M., MÄCHLER, M., COLOMBO, D., MAATHUIS, M. and BÜHLMANN, P. (2012). Causal inference using graphical models with the R package pcalg. *J. Stat. Softw.* **47** 1–26.

KOLLER, D. and FRIEDMAN, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, MA. MR2778120

LE, T. D., LIU, L., TSYKIN, A., GOODALL, G. J., LIU, B., SUN, B.-Y. and LI, J. (2013). Inferring microRNA-mRNA causal regulatory relationships from expression data. *Bioinformatics* **29** 765–771.

LIU, H., HAN, F., YUAN, M., LAFFERTY, J. and WASSERMAN, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.* **40** 2293–2326. MR3059084

MAATHUIS, M. H., KALISCH, M. and BÜHLMANN, P. (2009). Estimating high-dimensional intervention effects from observational data. *Ann. Statist.* **37** 3133–3164. MR2549555

MAATHUIS, M. H., COLOMBO, D., KALISCH, M. and BÜHLMANN, P. (2010). Predicting causal effects in large-scale systems from observational data. *Nat. Methods* **7** 247–248.

MEEK, C. (1995). Causal inference and causal explanation with background knowledge. In *UAI* 1995.

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

MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 417–473. MR2758523

NANDY, P., HAUSER, A. and MAATHUIS, M. H. (2018). Supplement to "High-dimensional consistency in score-based and hybrid structure learning." DOI:10.1214/17-AOS1654SUPP.

NANDY, P., MAATHUIS, M. H. and RICHARDSON, T. S. (2017). Estimating the effect of joint interventions from observational data in sparse high-dimensional settings. *Ann. Statist.* **45** 647–674. MR3650396

OUERD, M., OOMMEN, B. J. and MATWIN, S. (2004). A formal approach to using data distributions for building causal polytree structures. *Inform. Sci.* **168** 111–132. MR2117288

PEARL, J. (2009). *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2548166

PERKOVIC, E., TEXTOR, J., KALISCH, M. and MAATHUIS, M. (2015a). A complete adjustment criterion. In *UAI* 2015.

PERKOVIC, E., TEXTOR, J., KALISCH, M. and MAATHUIS, M. (2015b). Complete graphical characterization and construction of adjustment sets in markov equivalence classes of ancestral graphs. Preprint. Available at arXiv:1606.06903.

RAMSEY, J. D. (2015). Scaling up greedy equivalence search for continuous variables. Preprint. Available at arXiv:1507.07749.

RAVIKUMAR, P. K., RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2008). Model selection in Gaussian graphical models: High-dimensional consistency of ℓ_1-regularized MLE. In *NIPS* 2008.

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

REBANE, G. and PEARL, J. (1987). The recovery of causal poly-trees from statistical data. In *UAI* 1987.

SCHMIDBERGER, M., LENNERT, S. and MANSMANN, U. (2011). Conceptual aspects of large meta-analyses with publicly available microarray data: A case study in oncology. *Bioinform. Biol. Insights* **5** 13–39.

SCHMIDT, M., NICULESCU-MIZIL, A. and MURPHY, K. (2007). Learning graphical model structure using L1-regularization paths. In *AAAI* 2007.

SCHULTE, O., FRIGO, G., GREINER, R. and KHOSRAVI, H. (2010). The IMAP hybrid method for learning Gaussian Bayes nets. In *Canadian AI* 2010.

SCUTARI, M. (2010). Learning Bayesian networks with the bnlearn R package. *J. Stat. Softw.* **35** 1–22.

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation, Prediction, and Search*, 2nd ed. MIT Press, Cambridge, MA. MR1815675

SPIRTES, P., RICHARDSON, T., MEEK, C., SCHEINES, R. and GLYMOUR, C. (1998). Using path diagrams as a structural equation modeling tool. *Sociol. Methods Res.* **27** 182–225.

STEKHOVEN, D. J., MORAES, I., SVEINBJÖRNSSON, G., HENNING, L., MAATHUIS, M. H. and BÜHLMANN, P. (2012). Causal stability ranking. *Bioinformatics* **28** 2819–2823.

TSAMARDINOS, I., BROWN, L. E. and ALIFERIS, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Mach. Learn.* **65** 31–78.

UHLER, C., RASKUTTI, G., BÜHLMANN, P. and YU, B. (2013). Geometry of the faithfulness assumption in causal inference. *Ann. Statist.* **41** 436–463. MR3099109

VAN DE GEER, S. and BÜHLMANN, P. (2013). ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *Ann. Statist.* **41** 536–567. MR3099113

VERDUGO, R. A., ZELLER, T., ROTIVAL, M., WILD, P. S., MÜNZEL, T., LACKNER, K. J., WEIDMANN, H., NINIO, E., TRÉGOUËT, D.-A., CAMBIEN, F., BLANKENBERG, S. and TIRET, L. (2013). Graphical modeling of gene expression in monocytes suggests molecular mechanisms explaining increased atherosclerosis in smokers. *PLoS ONE* **8** e50888.

VERMA, T. and PEARL, J. (1990). Equivalence and synthesis of causal models. In *UAI* 1990.

ZHANG, K., PETERS, J., JANZING, D. and SCHÖLKOPF, B. (2011). Kernel-based conditional independence test and application in causal discovery. In *UAI* 2011.

ZHAO, T., LIU, H., ROEDER, K., LAFFERTY, J. and WASSERMAN, L. (2012). The huge package for high-dimensional undirected graph estimation in R. *J. Mach. Learn. Res.* **13** 1059–1062. MR2930633

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469