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

THE ANNALS *of* STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Articles

Vol. 45, No. 5—October 2017

The Annals of Statistics
2017, Vol. 45, No. 5, 1835–1862
DOI: 10.1214/16-AOS1510

"LOCAL" VS. "GLOBAL" PARAMETERS—BREAKING THE GAUSSIAN COMPLEXITY BARRIER

BY SHAHAR MENDELSON[1]

Technion—Israel Institute of Technology

We show that if F is a convex class of functions that is L-sub-Gaussian, the error rate of learning problems generated by independent noise is equivalent to a fixed point determined by "local" covering estimates of the class (i.e., the covering number at a specific level), rather than by the Gaussian average, which takes into account the structure of F at an arbitrarily small scale. To that end, we establish new sharp upper and lower estimates on the error rate in such learning problems.

REFERENCES

[1] ANTHONY, M. and BARTLETT, P. L. (1999). *Neural Network Learning*: *Theoretical Foundations*. Cambridge Univ. Press, Cambridge. MR1741038

[2] BIRGÉ, L. and MASSART, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields* **97** 113–150. MR1240719

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

[4] DUDLEY, R. M. (1999). *Uniform Central Limit Theorems*. *Cambridge Studies in Advanced Mathematics* **63**. Cambridge Univ. Press, Cambridge. MR1720712

[5] KOLTCHINSKII, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. *Lecture Notes in Math*. **2033**. Springer, Heidelberg. MR2829871

[6] LECUÉ, G. and MENDELSON, S. (2013). Learning subgaussian classes: Upper and minimax bounds. Technical report, CNRS, Ecole polytechnique and Technion.

[7] 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

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

[9] MENDELSON, S. (2008). Obtaining fast error rates in nonconvex situations. *J. Complexity* **24** 380–397. MR2426759

[10] MENDELSON, S. (2014). Learning without concentration for general loss functions. Preprint. Available at arXiv:1410.3192.

[11] MENDELSON, S. (2015). Learning without concentration. *J. ACM* **62** Art. 21, 25. MR3367000

[12] MENDELSON, S. (2016). Upper bounds on product and multiplier empirical processes. *Stochastic Process*. *Appl*. **126** 3652–3680. MR3565471

[13] MENDELSON, S. (2017). Supplement to "'Local' vs. 'global' parameters—breaking the Gaussian complexity barrier." DOI:10.1214/16-AOS1510SUPP.

[14] MENDELSON, S., PAJOR, A. and TOMCZAK-JAEGERMANN, N. (2007). Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geom. Funct. Anal*. **17** 1248–1282. MR2373017

MSC2010 subject classifications. 62G08, 62C20, 60G15.
Key words and phrases. Error rates, Gaussian averages, covering numbers.

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

[16] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York. MR1385671

[17] YANG, Y. and BARRON, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.* **27** 1564–1599. MR1742500

The Annals of Statistics
2017, Vol. 45, No. 5, 1863–1894
DOI: 10.1214/16-AOS1511

CONFOUNDER ADJUSTMENT IN MULTIPLE HYPOTHESIS TESTING

BY JINGSHU WANG*, QINGYUAN ZHAO*, TREVOR HASTIE†,1 AND ART B. OWEN†,2

University of Pennsylvania and Stanford University†*

We consider large-scale studies in which thousands of significance tests are performed simultaneously. In some of these studies, the multiple testing procedure can be severely biased by latent confounding factors such as batch effects and unmeasured covariates that correlate with both primary variable(s) of interest (e.g., treatment variable, phenotype) and the outcome. Over the past decade, many statistical methods have been proposed to adjust for the confounders in hypothesis testing. We unify these methods in the same framework, generalize them to include multiple primary variables and multiple nuisance variables, and analyze their statistical properties. In particular, we provide theoretical guarantees for RUV-4 [Gagnon-Bartsch, Jacob and Speed (2013)] and LEAPP [*Ann. Appl. Stat.* **6** (2012) 1664–1688], which correspond to two different identification conditions in the framework: the first requires a set of "negative controls" that are known a priori to follow the null distribution; the second requires the true nonnulls to be sparse. Two different estimators which are based on RUV-4 and LEAPP are then applied to these two scenarios. We show that if the confounding factors are strong, the resulting estimators can be asymptotically as powerful as the oracle estimator which observes the latent confounding factors. For hypothesis testing, we show the asymptotic z-tests based on the estimators can control the type I error. Numerical experiments show that the false discovery rate is also controlled by the Benjamini–Hochberg procedure when the sample size is reasonably large.

REFERENCES

[1] ALTER, O., BROWN, P. O. and BOTSTEIN, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA* **97** 10101–10106.

[2] ANDERSON, T. W. and RUBIN, H. (1956). Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. V* 111–150. Univ. California Press, Berkeley and Los Angeles. MR0084943

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

[4] BAI, J. and LI, K. (2014). Theory and methods of panel data models with interactive effects. *Ann. Statist.* **42** 142–170. MR3178459

MSC2010 subject classifications. Primary 62J15; secondary 62H25.

Key words and phrases. Empirical null, surrogate variable analysis, unwanted variation, batch effect, robust regression.

[5] BAI, J. and LI, K. (2016). Maximum likelihood estimation and inference for approximate factor models of high dimension. *Rev. Econ. Stat.* **98** 298–309.
[6] BAI, J. and NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221. MR1926259
[7] BAI, J. and NG, S. (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica* **74** 1133–1150. MR2238213
[8] BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392
[9] BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245
[10] BLALOCK, E. M., GEDDES, J. W., CHEN, K. C., PORTER, N. M., MARKESBERY, W. R. and LANDFIELD, P. W. (2004). Incipient Alzheimer's disease: Microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proc. Natl. Acad. Sci. USA* **101** 2173–2178.
[11] BOLLEN, K. A. (1989). *Structural Equations with Latent Variables*. Wiley, New York. MR0996025
[12] BRYS, G., HUBERT, M. and STRUYF, A. (2004). A robust measure of skewness. *J. Comput. Graph. Statist.* **13** 996–1017. MR2109062
[13] CHANDRASEKARAN, V., PARRILO, P. A. and WILLSKY, A. S. (2012). Latent variable graphical model selection via convex optimization. *Ann. Statist.* **40** 1935–1967. MR3059067
[14] CLARKE, S. and HALL, P. (2009). Robustness of multiple testing procedures against dependence. *Ann. Statist.* **37** 332–358. MR2488354
[15] CRAIG, A., CLOAREC, O., HOLMES, E., NICHOLSON, J. K. and LINDON, J. C. (2006). Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Anal. Chem.* **78** 2262–2267.
[16] DESAI, K. H. and STOREY, J. D. (2012). Cross-dimensional inference of dependent high-dimensional data. *J. Amer. Statist. Assoc.* **107** 135–151. MR2949347
[17] DE LA FUENTE, A., BING, N., HOESCHELE, I. and MENDES, P. (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* **20** 3565–3574.
[18] EFRON, B. (2007). Correlation and large-scale simultaneous significance testing. *J. Amer. Statist. Assoc.* **102** 93–103. MR2293302
[19] EFRON, B. (2010). Correlated z-values and the accuracy of large-scale statistical estimates. *J. Amer. Statist. Assoc.* **105** 1042–1055. MR2752597
[20] FAN, J. and HAN, X. (2013). Estimation of false discovery proportion with unknown dependence. Available at arXiv:1305.7007.
[21] FAN, J., HAN, X. and GU, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *J. Amer. Statist. Assoc.* **107** 1019–1035. MR3010887
[22] FARE, T. L., COFFEY, E. M., DAI, H., HE, Y. D., KESSLER, D. A., KILIAN, K. A., KOCH, J. E., LEPROUST, E., MARTON, M. J., MEYER, M. R. et al. (2003). Effects of atmospheric ozone on microarray data quality. *Anal. Chem.* **75** 4672–4675.
[23] FISHER, R. A. (1935). *The Design of Experiments*. Oliver & Boyd, Edinburgh.
[24] FRIGUET, C., KLOAREG, M. and CAUSEUR, D. (2009). A factor model approach to multiple testing under dependence. *J. Amer. Statist. Assoc.* **104** 1406–1415. MR2750571
[25] GAGNON-BARTSCH, J., JACOB, L. and SPEED, T. P. (2013). Removing unwanted variation from high dimensional data with negative controls. Technical Report 820, Dept. Statistics, Univ. California, Berkeley, Berkeley, CA.
[26] GAGNON-BARTSCH, J. A. and SPEED, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics* **13** 539–552.

[27] GASCH, A. P., SPELLMAN, P. T., KAO, C. M., CARMEL-HAREL, O., EISEN, M. B., STORZ, G., BOTSTEIN, D. and BROWN, P. O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11** 4241–4257.

[28] GREENLAND, S., ROBINS, J. M. and PEARL, J. (1999). Confounding and collapsibility in causal inference. *Statist. Sci.* **14** 29–46.

[29] GRZEBYK, M., WILD, P. and CHOUANIÈRE, D. (2004). On identification of multi-factor models with correlated residuals. *Biometrika* **91** 141–151. MR2050465

[30] IRIZARRY, R. A., HOBBS, B., COLLIN, F., BEAZER-BARCLAY, Y. D., ANTONELLIS, K. J., SCHERF, U., SPEED, T. P. et al. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4** 249–264.

[31] JIN, J. (2012). Comment: "Estimating false discovery proportion under arbitrary covariance dependence." [MR3010887] *J. Amer. Statist. Assoc.* **107** 1042–1045. MR3010891

[32] KISH, L. (1959). Some statistical problems in research design. *Am. Sociol. Rev.* **24** 328–338.

[33] KORN, E. L., TROENDLE, J. F., MCSHANE, L. M. and SIMON, R. (2004). Controlling the number of false discoveries: Application to high-dimensional genomic data. *J. Statist. Plann. Inference* **124** 379–398. MR2080371

[34] KUROKI, M. and PEARL, J. (2014). Measurement bias and effect restoration in causal inference. *Biometrika* **101** 423–437. MR3215357

[35] LAN, W. and DU, L. (2014). A factor-adjusted multiple testing procedure with application to mutual fund selection. Available at arXiv:1407.5515.

[36] LAZAR, C., MEGANCK, S., TAMINAU, J., STEENHOFF, D., COLETTA, A., MOLTER, C., WEISS-SOLÍS, D. Y., DUQUE, R., BERSINI, H. and NOWÉ, A. (2013). Batch effect removal methods for microarray gene expression data integration: A survey. *Brief. Bioinform.* **14** 469–490.

[37] LEEK, J. T., SCHARPF, R. B., BRAVO, H. C., SIMCHA, D., LANGMEAD, B., JOHNSON, W. E., GEMAN, D., BAGGERLY, K. and IRIZARRY, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11** 733–739.

[38] LEEK, J. T. and STOREY, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3** 1724–1735.

[39] LEEK, J. T. and STOREY, J. D. (2008). A general framework for multiple testing dependence. *Proc. Natl. Acad. Sci. USA* **105** 18718–18723.

[40] LI, J. and ZHONG, P.-S. (2016). A rate optimal procedure for recovering sparse differences between high-dimensional means under dependence. *Ann. Statist.* To appear.

[41] LIN, D. W., COLEMAN, I. M., HAWLEY, S., HUANG, C. Y., DUMPIT, R., GIFFORD, D., KEZELE, P., HUNG, H., KNUDSEN, B. S., KRISTAL, A. R. et al. (2006). Influence of surgical manipulation on prostate gene expression: Implications for molecular correlates of treatment effects and disease prognosis. *J. Clin. Oncol.* **24** 3763–3770.

[42] MARONNA, R. A., MARTIN, R. D. and YOHAI, V. J. (2006). *Robust Statistics: Theory and Methods*. Wiley, Chichester. MR2238141

[43] ONATSKI, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *Rev. Econ. Stat.* **92** 1004–1016.

[44] OWEN, A. B. (2005). Variance of the number of false discoveries. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 411–426. MR2155346

[45] OWEN, A. B. and WANG, J. (2016). Bi-cross-validation for factor analysis. *Statist. Sci.* **31** 119–139. MR3458596

[46] PEARL, J. (2009). *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2548166

[47] PERRY, P. O. and PILLAI, N. S. (2013). Degrees of freedom for combining regression with factor analysis. Preprint. Available at arXiv:1310.7269.

[48] PESARAN, M. H. (2004). General diagnostic tests for cross section dependence in panels. Cambridge Working Papers in Economics No. 0435.
[49] PRICE, A. L., PATTERSON, N. J., PLENGE, R. M., WEINBLATT, M. E., SHADICK, N. A. and REICH, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38** 904–909.
[50] RANSOHOFF, D. F. (2005). Bias as a threat to the validity of cancer molecular-marker research. *Nat. Rev. Cancer* **5** 142–149.
[51] RHODES, D. R. and CHINNAIYAN, A. M. (2005). Integrative analysis of the cancer transcriptome. *Nat. Genet.* **37** S31–S37.
[52] SCHWARTZMAN, A. (2010). Comment: "Correlated z-values and the accuracy of large-scale statistical estimates." [MR2752597] *J. Amer. Statist. Assoc.* **105** 1059–1063. MR2752600
[53] SCHWARTZMAN, A., DOUGHERTY, R. F. and TAYLOR, J. E. (2008). False discovery rate analysis of brain diffusion direction maps. *Ann. Appl. Stat.* **2** 153–175. MR2415598
[54] SHE, Y. and OWEN, A. B. (2011). Outlier detection using nonconvex penalized regression. *J. Amer. Statist. Assoc.* **106** 626–639. MR2847975
[55] SINGH, D., FOX, S. M., TAL-SINGER, R., PLUMB, J., BATES, S., BROAD, P., RILEY, J. H. and CELLI, B. (2011). Induced sputum genes associated with spirometric and radiological disease severity in COPD ex-smokers. *Thorax* **66** 489–495.
[56] STOREY, J. D., TAYLOR, J. E. and SIEGMUND, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66** 187–205. MR2035766
[57] SUN, W. and CAI, T. T. (2009). Large-scale multiple testing under dependence. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **71** 393–424. MR2649603
[58] SUN, Y. (2011). On latent systemic effects in multiple hypotheses. Ph.D. thesis, Stanford University.
[59] SUN, Y., ZHANG, N. R. and OWEN, A. B. (2012). Multiple hypothesis testing adjusted for latent variables, with an application to the AGEMAP gene expression data. *Ann. Appl. Stat.* **6** 1664–1688. MR3058679
[60] TUSHER, V. G., TIBSHIRANI, R. and CHU, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98** 5116–5121.
[61] VAWTER, M. P., EVANS, S., CHOUDARY, P., TOMITA, H., MEADOR-WOODRUFF, J., MOLNAR, M., LI, J., LOPEZ, J. F., MYERS, R., COX, D. et al. (2004). Gender-specific gene expression in post-mortem human brain: Localization to sex chromosomes. *Neuropsychopharmacology* **29** 373–384.
[62] WANG, J., ZHAO, Q., HASTIE, T. and OWEN, A. B. (2017). Supplement to "Confounder adjustment in multiple hypothesis testing." DOI:10.1214/16-AOS1511SUPP.
[63] WANG, S., CUI, G. and LI, K. (2015). Factor-augmented regression models with structural change. *Econom. Lett.* **130** 124–127. MR3336182
[64] YOHAI, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Ann. Statist.* **15** 642–656. MR0888431

The Annals of Statistics
2017, Vol. 45, No. 5, 1895–1919
DOI: 10.1214/16-AOS1512

GAUSSIAN APPROXIMATION FOR HIGH DIMENSIONAL TIME SERIES

BY DANNA ZHANG AND WEI BIAO WU

University of California, San Diego and University of Chicago

We consider the problem of approximating sums of high dimensional stationary time series by Gaussian vectors, using the framework of functional dependence measure. The validity of the Gaussian approximation depends on the sample size n, the dimension p, the moment condition and the dependence of the underlying processes. We also consider an estimator for long-run covariance matrices and study its convergence properties. Our results allow constructing simultaneous confidence intervals for mean vectors of high-dimensional time series with asymptotically correct coverage probabilities. As an application, we propose a Kolmogorov–Smirnov-type statistic for testing distributions of high-dimensional time series.

REFERENCES

[1] ALEXOPOULOS, C. and GOLDSMAN, D. (2004). To batch or not to batch? *ACM Trans. Model. Comput. Simul.* **14** 76–114.

[2] ANDERSON, T. W. (1971). *The Statistical Analysis of Time Series*. Wiley, New York. MR0283939

[3] AZZALINI, A. (1981). A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika* **68** 326–328. MR0614972

[4] BRADLEY, R. C. (2007). *Introduction to Strong Mixing Conditions*. Kendrick Press.

[5] BÜHLMANN, P. (2002). Bootstraps for time series. *Statist. Sci.* **17** 52–72. MR1910074

[6] BURKHOLDER, D. L. (1973). Distribution function inequalities for martingales. *Ann. Probab.* **1** 19–42. MR0365692

[7] CAI, T. T. and JIANG, T. (2011). Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *Ann. Statist.* **39** 1496–1525. MR2850210

[8] CHEN, X., SHAO, Q. M., WU, W. B. and XU, L. (2015). Self-normalized Cramér type moderate deviations under dependence. Preprint. Available at arXiv:1409.3642.

[9] CHENG, M.-Y. and PENG, L. (2002). Regression modeling for nonparametric estimation of distribution and quantile functions. *Statist. Sinica* **12** 1043–1060. MR1947061

[10] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41** 2786–2819. MR3161448

[11] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2013). Testing many moment inequalities. Preprint. Available at arXiv:1312.7614.

[12] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2015). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probab. Theory Related Fields* **162** 47–70. MR3350040

MSC2010 subject classifications. Primary 62M10; secondary 62E17.

Key words and phrases. Gaussian approximation, high-dimensional time series, Kolmogorov–Smirnov test, long run covariance matrix, simultaneous inference.

[13] DECO, G., JIRSA, V. K. and MCINTOSH, A. R. (2013). Resting brains never rest: Computational insights into potential cognitive architectures. *Trends Neurosci*. **36** 268–274.

[14] DEDECKER, J., DOUKHAN, P., LANG, G., LEÓN, J. R., LOUHICHI, S. and PRIEUR, C. (2007). *Weak Dependence*: *With Examples and Applications*. *Lecture Notes in Statistics* **190**. Springer, New York. MR2338725

[15] 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

[16] FALK, M. (1985). Asymptotic normality of the kernel quantile estimator. *Ann*. *Statist*. **13** 428–433. MR0773180

[17] FERNIQUE, X. (1975). Regularité des trajectoires des fonctions aléatoires gaussiennes. In *École d'Été de Probabilités de Saint-Flour, IV*-1974. *Lecture Notes in Math*. **480** 1–96. Springer, Berlin. MR0413238

[18] HIPP, J. F., HAWELLEK, D. J., CORBETTA, M., SIEGEL, M. and ENGEL, A. K. (2012). Large-scale cortical correlation structure of spontaneous oscillatory activity. *Nat*. *Neurosci*. **15** 884–890.

[19] HUTCHISON, R. M., WOMELSDORF, T., ALLEN, E. A., BANDETTINI, P. A., CALHOUN, V. D., CORBETTA, M., DELLA PENNA, S., DUYN, J. H., GLOVER, G. H., GONZALEZ-CASTILLO, J., HANDWERKER, D. A., KEILHOLZ, S., KIVINIEMI, V., LEOPOLD, D. A., DE PASQUALE, F., SPORNS, O., WALTER, M. and CHANG, C. (2013). Dynamic functional connectivity: Promise, issues, and interpretations. *NeuroImage* **80** 360–378.

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

[21] JIANG, T. (2004). The asymptotic distributions of the largest entries of sample correlation matrices. *Ann*. *Appl*. *Probab*. **14** 865–880. MR2052906

[22] KOSOROK, M. R. and MA, S. (2007). Marginal asymptotics for the "large p, small n" paradigm: With applications to microarray data. *Ann*. *Statist*. **35** 1456–1486. MR2351093

[23] LAHIRI, S. N. (2003). *Resampling Methods for Dependent Data*. Springer, Berlin.

[24] LARSON-PRIOR, L. J., OOSTENVELD, R., PENNA, S. D., MICHALAREAS, G., PRIOR, F., BABAJANI-FEREMI, A., SCHOFFELEN, J. M., MARZETTI, L., DE PASQUALE, F., POMPEO, F. D., STOUT, J., WOOLRICH, M., LUO, Q., BUCHOLZ, R., FRIES, P., PIZZELLA, V., ROMANI, G. L., CORBETTA, M. and SNYDER, A. Z. (2013). Adding dynamics to the human connectome project with {MEG}. *NeuroImage* **80** 190–201.

[25] LI, D., LIU, W.-D. and ROSALSKY, A. (2010). Necessary and sufficient conditions for the asymptotic distribution of the largest entry of a sample correlation matrix. *Probab. Theory Related Fields* **148** 5–35. MR2653220

[26] LI, D. and ROSALSKY, A. (2006). Some strong limit theorems for the largest entries of sample correlation matrices. *Ann*. *Appl*. *Probab*. **16** 423–447. MR2209348

[27] LIU, W. and WU, W. B. (2010). Asymptotics of spectral density estimates. *Econometric Theory* **26** 1218–1245. MR2660298

[28] LIU, W.-D., LIN, Z. and SHAO, Q.-M. (2008). The asymptotic distribution and Berry–Esseen bound of a new test for independence in high dimension with an application to stochastic optimization. *Ann*. *Appl*. *Probab*. **18** 2337–2366. MR2474539

[29] NAGAEV, S. V. (1979). Large deviations of sums of independent random variables. *Ann*. *Probab*. **7** 745–789. MR0542129

[30] NEWEY, W. K. and WEST, K. D. (1987). A simple, positive semidefinite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55** 703–708. MR0890864

[31] PINELIS, I. (1994). Optimum bounds for the distributions of martingales in Banach spaces. *Ann*. *Probab*. **22** 1679–1706. MR1331198

[32] POLITIS, D. N., ROMANO, J. P. and WOLF, M. (1999). *Subsampling*. Springer, New York. MR1707286

[33] PORTNOY, S. (1986). On the central limit theorem in $\mathbf{R}^p$ when $p \to \infty$. *Probab. Theory Related Fields* **73** 571–583. MR0863546

[34] PRIESTLEY, M. B. (1981). *Spectral Analysis and Time Series*. Academic Press, San Diego.

[35] PRIESTLEY, M. B. (1988). *Non-linear and Non-stationary Time Series Analysis*. Academic Press, San Diego.

[36] REISS, R.-D. (1981). Nonparametric estimation of smooth distribution functions. *Scand. J. Stat.* **8** 116–119. MR0623587

[37] ROSENBLATT, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Natl. Acad. Sci. USA* **42** 43–47. MR0074711

[38] ROSENBLATT, M. (1971). *Markov Processes. Structure and Asymptotic Behavior*. Springer, New York. MR0329037

[39] ROSENBLATT, M. (1985). *Stationary Sequences and Random Fields*. Springer, Berlin.

[40] TONG, H. (1990). *Nonlinear Time Series: A Dynamical System Approach*. Oxford Univ. Press, New York. MR1079320

[41] TSAY, R. S. (2005). *Analysis of Financial Time Series*, 2nd ed. Wiley, Hoboken, NJ. MR2162112

[42] WANG, J., CHENG, F. and YANG, L. (2013). Smooth simultaneous confidence bands for cumulative distribution functions. *J. Nonparametr. Stat.* **25** 395–407. MR3056092

[43] WIENER, N. (1958). *Nonlinear Problems in Random Theory*. Wiley, New York. MR0100912

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

[45] WU, W. B. (2011). Asymptotic theory for stationary processes. *Stat. Interface* **0** 1–20.

[46] XIAO, H. and WU, W. B. (2013). Asymptotic theory for maximum deviations of sample covariance matrix estimates. *Stochastic Process. Appl.* **123** 2899–2920. MR3054550

[47] YAMATO, H. (1972/73). Uniform convergence of an estimator of a distribution function. *Bull. Math. Statist.* **15** 69–78. MR0329113

[48] ZHANG, D. and WU, W. B (2016). Supplement to "Gaussian approximation for high dimensional time series." DOI:10.1214/16-AOS1512SUPP.

[49] ZHANG, X. and CHENG, G. (2014). Bootstrapping high dimensional time series. Preprint. Available at arXiv:1406.1037.

[50] ZHOU, W. (2007). Asymptotic distribution of the largest off-diagonal entry of correlation matrices. *Trans. Amer. Math. Soc.* **359** 5345–5363. MR2327033

The Annals of Statistics
2017, Vol. 45, No. 5, 1920–1950
DOI: 10.1214/16-AOS1513

DETECTION AND FEATURE SELECTION IN SPARSE MIXTURE MODELS[1]

BY NICOLAS VERZELEN AND ERY ARIAS-CASTRO

INRA and University of California, San Diego

We consider Gaussian mixture models in high dimensions, focusing on the twin tasks of detection and feature selection. Under sparsity assumptions on the difference in means, we derive minimax rates for the problems of testing and of variable selection. We find these rates to depend crucially on the knowledge of the covariance matrices and on whether the mixture is symmetric or not. We establish the performance of various procedures, including the top sparse eigenvalue of the sample covariance matrix (popular in the context of Sparse PCA), as well as new tests inspired by the normality tests of Malkovich and Afifi [*J. Amer. Statist. Assoc.* **68** (1973) 176–179].

REFERENCES

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **AC-19** 716–723. MR0423716

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

AZIZYAN, M., SINGH, A. and WASSERMAN, L. (2013). Minimax theory for high-dimensional Gaussian mixtures with sparse mean separation. *Neural Information Processing Systems* (*NIPS*).

AZIZYAN, M., SINGH, A. and WASSERMAN, L. (2015). Efficient sparse clustering of high-dimensional non-spherical Gaussian mixtures. In *AISTATS*.

BELKIN, M. and SINHA, K. (2010). Polynomial learning of distribution families. In 2010 *IEEE* 51*st Annual Symposium on Foundations of Computer Science FOCS* 2010 103–112. IEEE Computer Soc., Los Alamitos, CA. MR3024780

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

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

BICKEL, P. J. and LEVINA, E. (2004). Some theory of Fisher's linear discriminant function, "naive Bayes", and some alternatives when there are many more variables than observations. *Bernoulli* **10** 989–1010. MR2108040

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

BOUCHERON, S., BOUSQUET, O., LUGOSI, G. and MASSART, P. (2005). Moment inequalities for functions of independent random variables. *Ann. Probab.* **33** 514–560. MR2123200

BRUBAKER, S. C. and VEMPALA, S. S. (2008). Isotropic PCA and affine-invariant clustering. In *Building Bridges*. *Bolyai Soc. Math. Stud.* **19** 241–281. Springer, Berlin. MR2484643

MSC2010 subject classifications. 62H30, 62H15.

Key words and phrases. Gaussian mixture models, detection of mixtures, feature selection for mixtures, sparse mixture models, the sparse eigenvalue problem, projection tests based on moments.

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

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

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

CHAN, Y. and HALL, P. (2010). Using evidence of mixed populations to select variables for clustering very high-dimensional data. *J. Amer. Statist. Assoc.* **105** 798–809. MR2724862

CHANG, W.-C. (1983). On using principal components before separating a mixture of two multivariate normal distributions. *J. Roy. Statist. Soc. Ser. C* **32** 267–275. MR0770316

CHAUDHURI, K., DASGUPTA, S. and VATTANI, A. (1999). Learning mixtures of Gaussians using the k-means algorithm. Preprint. Available at arXiv:0912.0086.

CHEN, S. S., DONOHO, D. L. and SAUNDERS, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* **20** 33–61. MR1639094

D'ASPREMONT, A., EL GHAOUI, L., JORDAN, M. I. and LANCKRIET, G. R. G. (2007). A direct formulation for sparse pca using semidefinite programming. *SIAM Rev.* **49** 434–448.

DONOHO, D. and JIN, J. (2009). Feature selection by higher criticism thresholding achieves the optimal phase diagram. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **367** 4449–4470. MR2546396

FAN, J. and PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32** 928–961. MR2065194

FRIEDMAN, J. H. and MEULMAN, J. J. (2004). Clustering objects on subsets of attributes. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66** 815–849. MR2102467

HARDT, M. and PRICE, E. (2015). Tight bounds for learning a mixture of two Gaussians [extended abstract]. In *STOC'15—Proceedings of the* 2015 *ACM Symposium on Theory of Computing* 753–760. ACM, New York. MR3388255

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, New York. MR2722294

HSU, D. and KAKADE, S. M. (2013). Learning mixtures of spherical Gaussians: Moment methods and spectral decompositions. In *ITCS'13—Proceedings of the* 2013 *ACM Conference on Innovations in Theoretical Computer Science* 11–19. ACM, New York. MR3385380

INGSTER, Y. I., POUET, C. and TSYBAKOV, A. B. (2009). Classification of sparse high-dimensional vectors. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **367** 4427–4448. MR2546395

JI, P. and JIN, J. (2012). UPS delivers optimal phase diagram in high-dimensional variable selection. *Ann. Statist.* **40** 73–103. MR3013180

JIN, J. (2009). Impossibility of successful classification when useful features are rare and weak. *Proc. Natl. Acad. Sci. USA* **106** 8859–8864. MR2520682

JIN, J., KE, Z. T. and WANG, W. (2015). Phase transitions for high dimensional clustering and related problems. Preprint. Available at arXiv:1502.06952.

JIN, J. and WANG, W. (2014). Important feature pca for high dimensional clustering. Preprint. Available at arXiv:1407.5241.

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

KALAI, A. T., MOITRA, A. and VALIANT, G. (2012). Disentangling Gaussians. *Commun. ACM* **55** 113–120.

MALKOVICH, J. F. and AFIFI, A. (1973). On tests for multivariate normality. *J. Amer. Statist. Assoc.* **68** 176–179.

MALLAT, S. and ZHANG, Z. (1993). Matching pursuit with time-frequency dictionaries. *IEEE Trans. Image Process.* **41** 3397–3415.

MALLOWS, C. (1973). Some comments on cp. *Technometrics* **15** 661–675.

MARDIA, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika* **57** 519–530. MR0397994

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

MAUGIS, C. and MICHEL, B. (2011). A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM Probab. Stat*. **15** 41–68. MR2870505

PAN, W. and SHEN, X. (2007). Penalized model-based clustering with application to variable selection. *J. Mach. Learn. Res*. **8** 1145–1164.

RAFTERY, A. E. and DEAN, N. (2006). Variable selection for model-based clustering. *J. Amer. Statist. Assoc*. **101** 168–178. MR2268036

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

SRIVASTAVA, M. S. (1984). A measure of skewness and kurtosis and a graphical method for assessing multivariate normality. *Statist. Probab. Lett*. **2** 263–267. MR0777837

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

TROPP, J. A. (2004). Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory* **50** 2231–2242. MR2097044

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

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

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

VERZELEN, N. and ARIAS-CASTRO, E. (2016). Supplement to "Detection and feature selection in sparse mixture models." DOI:10.1214/16-AOS1513SUPP.

VU, V. Q. and LEI, J. (2012). Minimax rates of estimation for sparse pca in high dimensions. In *International Conference on Artificial Intelligence and Statistics* 1278–1286.

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

WANG, S. and ZHU, J. (2008). Variable selection for model-based high-dimensional clustering and its application to microarray data. *Biometrics* **64** 440–448, 666. MR2432414

WITTEN, D. M. and TIBSHIRANI, R. (2010). A framework for feature selection in clustering. *J. Amer. Statist. Assoc*. **105** 713–726. MR2724855

XIE, B., PAN, W. and SHEN, X. (2008). Penalized model-based clustering with cluster-specific diagonal covariance matrices and grouped variables. *Electron. J. Stat*. **2** 168–212. MR2386092

ZHU, J. and HASTIE, T. (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics* **5** 427–443.

The Annals of Statistics
2017, Vol. 45, No. 5, 1951–1987
DOI: 10.1214/16-AOS1515

MINIMAX ESTIMATION OF A FUNCTIONAL ON A STRUCTURED HIGH-DIMENSIONAL MODEL[1]

BY JAMES M. ROBINS*, LINGLING LI†, RAJARSHI MUKHERJEE‡, ERIC TCHETGEN TCHETGEN* AND AAD VAN DER VAART§

Harvard T. H. Chan School of Public Health, Sanofi Genzyme,† Stanford University‡ and Universiteit Leiden§*

We introduce a new method of estimation of parameters in semiparametric and nonparametric models. The method employs U-statistics that are based on higher-order influence functions of the parameter of interest, which extend ordinary linear influence functions, and represent higher derivatives of this parameter. For parameters for which the representation cannot be perfect the method often leads to a bias-variance trade-off, and results in estimators that converge at a slower than $\sqrt{n}$-rate. In a number of examples, the resulting rate can be shown to be optimal. We are particularly interested in estimating parameters in models with a nuisance parameter of high dimension or low regularity, where the parameter of interest cannot be estimated at $\sqrt{n}$-rate, but we also consider efficient $\sqrt{n}$-estimation using novel nonlinear estimators. The general approach is applied in detail to the example of estimating a mean response when the response is not always observed.

REFERENCES

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

[2] BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins Univ. Press, Baltimore, MD. MR1245941

[3] BICKEL, P. J. and RITOV, Y. (1988). Estimating integrated squared density derivatives: Sharp best order of convergence estimates. *Sankhyā Ser. A* **50** 381–393. MR1065550

[4] BIRGÉ, L. and MASSART, P. (1995). Estimation of integral functionals of a density. *Ann. Statist.* **23** 11–29. MR1331653

[5] BOLTHAUSEN, E., PERKINS, E. and VAN DER VAART, A. (2002). *Lectures on Probability Theory and Statistics* (P. Bernard, ed.). *Lecture Notes in Math.* **1781**. Springer, Berlin. MR1915443

[6] CAI, T. T. and LOW, M. G. (2005). Nonquadratic estimators of a quadratic functional. *Ann. Statist.* **33** 2930–2956. MR2253108

[7] CAI, T. T. and LOW, M. G. (2006). Optimal adaptive estimation of a quadratic functional. *Ann. Statist.* **34** 2298–2325. MR2291501

[8] DONOHO, D. L. and NUSSBAUM, M. (1990). Minimax quadratic estimation of a quadratic functional. *J. Complexity* **6** 290–323. MR1081043

[9] HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. and STAHEL, W. A. (1986). *Robust Statistics*: *The Approach Based on Influence Functions*. Wiley, New York. MR0829458

MSC2010 subject classifications. 62G05, 62G20, 62F25.

Key words and phrases. Nonlinear functional, nonparametric estimation, U-statistic, influence function, tangent space.

[10] HAS'MINSKIĬ, R. Z. and IBRAGIMOV, I. A. (1979). On the nonparametric estimation of functionals. In *Proceedings of the Second Prague Symposium on Asymptotic Statistics* (*Hradec Králové*, 1978) 41–51. North-Holland, Amsterdam. MR0571174

[11] HUBER, P. J. and RONCHETTI, E. M. (2009). *Robust Statistics*, 2nd ed. Wiley, Hoboken, NJ. MR2488795

[12] KERKYACHARIAN, G. and PICARD, D. (1996). Estimating nonquadratic functionals of a density using Haar wavelets. *Ann*. *Statist*. **24** 485–507. MR1394973

[13] KOŠEVNIK, J. A. and LEVIT, B. J. (1976). On a nonparametric analogue of the information matrix. *Teor. Verojatnost*. *i Primenen*. **21** 759–774. MR0428578

[14] LAURENT, B. (1997). Estimation of integral functionals of a density and its derivatives. *Bernoulli* **3** 181–211. MR1466306

[15] LAURENT, B. and MASSART, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Ann*. *Statist*. **28** 1302–1338. MR1805785

[16] LINDSAY, B. G. (1983). Efficiency of the conditional score in a mixture setting. *Ann*. *Statist*. **11** 486–497. MR0696061

[17] MURPHY, S. A. and VAN DER VAART, A. W. (2000). On profile likelihood. *J. Amer. Statist*. *Assoc*. **95** 449–485. MR1803168

[18] NEMIROVSKI, A. (2000). Topics in non-parametric statistics. In *Lectures on Probability Theory and Statistics* (*Saint-Flour*, 1998). *Lecture Notes in Math*. **1738** 85–277. Springer, Berlin. MR1775640

[19] PFANZAGL, J. (1982). *Contributions to a General Asymptotic Statistical Theory*. *Lecture Notes in Statistics* **13**. Springer, New York. MR0675954

[20] PFANZAGL, J. (1985). *Asymptotic Expansions for General Statistical Models*. *Lecture Notes in Statistics* **31**. Springer, Berlin. MR0810004

[21] PFANZAGL, J. (1990). *Estimation in Semiparametric Models*: *Some Recent Developments*. *Lecture Notes in Statistics* **63**. Springer, New York. MR1048589

[22] ROBINS, J. M., LI, L., MUKHERJEE, R., TCHETGEN, E. T. and VAN DER VAART, A. (2017). Supplement to "Minimax estimation of a functional on a structured high-dimensional model." DOI:10.1214/16-AOS1515SUPP.

[23] ROBINS, J., LI, L., TCHETGEN, E. and VAN DER VAART, A. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics*: *Essays in Honor of David A. Freedman*. *Inst*. *Math*. *Stat*. *Collect*. **2** 335–421. IMS, Beachwood, OH. MR2459958

[24] ROBINS, J., LI, L., TCHETGEN, E. and VAN DER VAART, A. W. (2009). Quadratic semiparametric von Mises calculus. *Metrika* **69** 227–247. MR2481922

[25] ROBINS, J., TCHETGEN TCHETGEN, E., LI, L. and VAN DER VAART, A. (2009). Semiparametric minimax rates. *Electron*. *J. Stat*. **3** 1305–1321. MR2566189

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

[27] ROTNITZKY, A. and ROBINS, J. M. (1995). Semi-parametric estimation of models for means and covariances in the presence of missing data. *Scand*. *J. Statist*. **22** 323–333. MR1363216

[28] TSYBAKOV, A. B. (2004). *Introduction à l'estimation non-paramétrique*. *Mathématiques & Applications* (*Berlin*) [*Mathematics & Applications*] **41**. Springer, Berlin. MR2013911

[29] VAN ZWET, W. R. (1984). A Berry–Esseen bound for symmetric statistics. *Z. Wahrsch*. *Verw. Gebiete* **66** 425–440. MR0751580

[30] VAN DER LAAN, M. J. and ROBINS, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer, New York. MR1958123

[31] VAN DER VAART, A. (1991). On differentiable functionals. *Ann*. *Statist*. **19** 178–204. MR1091845

[32] VAN DER VAART, A. (1996). Efficient maximum likelihood estimation in semiparametric mixture models. *Ann. Statist.* **24** 862–878. MR1394993

[33] VAN DER VAART, A. (2014). Higher order tangent spaces and influence functions. *Statist. Sci.* **29** 679–686. MR3300365

[34] VAN DER VAART, A. W. (1988). Estimating a real parameter in a class of semiparametric models. *Ann. Statist.* **16** 1450–1474. MR0964933

[35] VAN DER VAART, A. W. (1988). *Statistical Estimation in Large Parameter Spaces*. *CWI Tract* **44**. Stichting Mathematisch Centrum, Centrum voor Wiskunde en Informatica, Amsterdam. MR0927725

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

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

[38] VON MISES, R. (1947). On the asymptotic distribution of differentiable statistical functions. *Ann. Math. Stat.* **18** 309–348. MR0022330

[39] WATERMAN, R. P. and LINDSAY, B. G. (1996). Projected score methods for approximating conditional scores. *Biometrika* **83** 1–13. MR1399151

The Annals of Statistics
2017, Vol. 45, No. 5, 1988–2015
DOI: 10.1214/16-AOS1516

ASYMPTOTIC THEORY OF GENERALIZED ESTIMATING EQUATIONS BASED ON JACK-KNIFE PSEUDO-OBSERVATIONS

BY MORTEN OVERGAARD, ERIK THORLUND PARNER AND JAN PEDERSEN[1]

Aarhus University

A general asymptotic theory of estimates from estimating functions based on jack-knife pseudo-observations is established by requiring that the underlying estimator can be expressed as a smooth functional of the empirical distribution. Using results in p-variation norms, the theory is applied to important estimators from time-to-event analysis, namely the Kaplan–Meier estimator and the Aalen–Johansen estimator in a competing risks model, and the corresponding estimators of restricted mean survival and cause-specific lifetime lost. Under an assumption of completely independent censorings, this allows for estimating parameters in regression models of survival, cumulative incidences, restricted mean survival, and cause-specific lifetime lost. Considering estimators as functionals and applying results in p-variation norms is apparently an excellent way of studying the asymptotics of such estimators.

REFERENCES

ANDERSEN, P. K. (2013). Decomposition of number of life years lost according to causes of death. *Stat. Med.* **32** 5278–5285. MR3141373

ANDERSEN, P. K., HANSEN, M. G. and KLEIN, J. P. (2004). Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime Data Anal.* **10** 335–350. MR2137286

ANDERSEN, P. K., KLEIN, J. P. and ROSTHØJ, S. (2003). Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika* **90** 15–27. MR1966547

ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. and KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York. MR1198884

DUDLEY, R. M. and NORVAIŠA, R. (1999). *Differentiability of Six Operators on Nonsmooth Functions and p-Variation*. *Lecture Notes in Math.* **1703**. Springer, Berlin. MR1705318

DUDLEY, R. M. and NORVAIŠA, R. (2011). *Concrete Functional Calculus*. Springer, New York. MR2732563

GILL, R. D. (1980). *Censoring and Stochastic Integrals*. *Mathematical Centre Tracts* **124**. Mathematisch Centrum, Amsterdam. MR596815

GILL, R. D. (1989). Non- and semi-parametric maximum likelihood estimators and the von Mises method. I. *Scand. J. Stat.* **16** 97–128. MR1028971

GRAW, F., GERDS, T. A. and SCHUMACHER, M. (2009). On pseudo-values for regression analysis in competing risks models. *Lifetime Data Anal.* **15** 241–255. MR2510866

JACOBSEN, M. and MARTINUSSEN, T. (2016). A note on the large sample properties of estimators based on generalized linear models for correlated pseudo-observations. *Scand. J. Stat.* **43** 845–862.

MSC2010 subject classifications. Primary 62N02; secondary 62F12, 62J12.

Key words and phrases. Pseudo-values, functional differentiability, von Mises expansion, pseudo-observation method, p-variation, U-statistics.

JAMES, L. F. (1997). A study of a class of weighted bootstraps for censored data. *Ann. Statist.* **25** 1595–1621. MR1463566

KLEIN, J. P. (2006). Modelling competing risks in cancer studies. *Stat. Med.* **25** 1015–1034. MR2225189

KLEIN, J. P. and ANDERSEN, P. K. (2005). Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics* **61** 223–229. MR2135864

KLEIN, J. P., LOGAN, B., HARHOFF, M. and ANDERSEN, P. K. (2007). Analyzing survival curves at a fixed point in time. *Stat. Med.* **26** 4505–4519. MR2410057

OVERGAARD, M., PARNER, E. T and PEDERSEN, J. (2017). Supplement to "Asymptotic theory of generalized estimating equations based on jack-knife pseudo-observations." DOI:10.1214/16-AOS1516SUPP.

REEDS, J. A. III (1976). On the definition of Von Mises functionals. Ph.D. thesis, Harvard University, ProQuest LLC, Ann Arbor, MI. MR2940746

SØRENSEN, M. (1999). On asymptotics of estimating functions. *Braz. J. Probab. Stat.* **13** 111–136. MR1803041

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

VON MISES, R. (1947). On the asymptotic distribution of differentiable statistical functions. *Ann. Math. Stat.* **18** 309–348. MR0022330

The Annals of Statistics
2017, Vol. 45, No. 5, 2016–2045
DOI: 10.1214/16-AOS1517

BAYESIAN POISSON CALCULUS FOR LATENT FEATURE MODELING VIA GENERALIZED INDIAN BUFFET PROCESS PRIORS

BY LANCELOT F. JAMES[1]

Hong Kong University of Science and Technology

Statistical latent feature models, such as latent factor models, are models where each observation is associated with a vector of latent features. A general problem is how to select the number/types of features, and related quantities. In Bayesian statistical machine learning, one seeks (nonparametric) models where one can learn such quantities in the presence of observed data. The Indian Buffet Process (IBP), devised by Griffiths and Ghahramani (2005), generates a (sparse) latent binary matrix with columns representing a potentially unbounded number of features and where each row corresponds to an individual or object. Its generative scheme is cast in terms of customers entering sequentially an Indian Buffet restaurant and selecting previously sampled dishes as well as new dishes. Dishes correspond to latent features shared by individuals. The IBP has been applied to a wide range of statistical problems. Recent works have demonstrated the utility of generalizations to nonbinary matrices. The purpose of this work is to describe a unified mechanism for construction, Bayesian analysis, and practical sampling of broad generalizations of the IBP that generate (sparse) matrices with general entries. An adaptation of the Poisson partition calculus is employed to handle the complexities, including combinatorial aspects, of these models. Our work reveals a spike and slab characterization, and also presents a general framework for multivariate extensions. We close by highlighting a multivariate IBP with condiments, and the role of a stable-Beta Dirichlet multivariate prior.

REFERENCES

[1] ARCHAMBEAU, C., LAKSHMINARAYANAN, B. and BOUCHARD, G. (2015). Latent IBP compound Dirichlet allocation. *IEEE Trans. Pattern Anal. Mach. Intell.* **37** 321–333.

[2] ARRATIA, R., BARBOUR, A. D. and TAVARÉ, S. (2003). *Logarithmic Combinatorial Structures: A Probabilistic Approach*. European Mathematical Society (EMS), Zürich. MR2032426

[3] BARNDORFF-NIELSEN, O. E., PEDERSEN, J. and SATO, K. (2001). Multivariate subordination, self-decomposability and stability. *Adv. in Appl. Probab.* **33** 160–187. MR1825321

[4] BERTI, P., CRIMALDI, I., PRATELLI, L. and RIGO, P. (2015). Central limit theorems for an Indian buffet model with random weights. *Ann. Appl. Probab.* **25** 523–547. MR3313747

[5] BRODERICK, T., JORDAN, M. I. and PITMAN, J. (2013). Cluster and feature modeling from combinatorial stochastic processes. *Statist. Sci.* **28** 289–312. MR3135534

MSC2010 subject classifications. Primary 60C05, 60G09; secondary 60G57, 60E99.

Key words and phrases. Bayesian statistical machine learning, Indian buffet process, nonparametric latent feature models, Poisson process calculus, spike and slab priors.

[6] BRODERICK, T., MACKEY, L., PAISLEY, J. and JORDAN, M. I. (2015). Combinatorial clustering and the beta negative binomial process. *IEEE Trans. Pattern Anal. Mach. Intell.* **37** 290–306.

[7] BRODERICK, T., PITMAN, J. and JORDAN, M. I. (2013). Feature allocations, probability functions, and paintboxes. *Bayesian Anal.* **8** 801–836.

[8] BRODERICK, T., WILSON, A. C. and JORDAN, M. I. (2017). Posteriors, conjugacy, and exponential families for completely random measures. *Bernoulli*. To appear. DOI: 10.3150/16-BEJ855.

[9] CARON, F. (2012). Bayesian nonparametric models for bipartite graphs. In *Neural Information Processing Systems* (*NIPS* 2012), Lake Tahoe, CA.

[10] DOKSUM, K. (1974). Tailfree and neutral random probabilities and their posterior distributions. *Ann. Probab.* **2** 183–201. MR0373081

[11] GERSHMAN, S. J., FRAZIER, P. I. and BLEI, D. M. (2015). Distance dependent infinite latent feature models. *IEEE Trans. Pattern Anal. Mach. Intell.* **37** 334–345.

[12] GHAHRAMANI, Z., GRIFFITHS, T. L. and SOLLICH, P. (2007). Bayesian nonparametric latent feature models. In *Bayesian Statistics* 8. 201–226. Oxford Univ. Press, Oxford. MR2433194

[13] GNEDIN, A. and PITMAN, J. (2005). Regenerative composition structures. *Ann. Probab.* **33** 445–479.

[14] GÖRUR, D., JÄKEL, F. and RASMUSSEN, C. E. (2006). A choice model with infinitely many latent features. In *Proceedings of the* 23*rd International Conference on Machine Learning* 361–368. ACM, New York.

[15] GRIFFITHS, T. L. and GHAHRAMANI, Z. (2006). Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems* 18 (*NIPS*-2005).

[16] GRIFFITHS, T. L. and GHAHRAMANI, Z. (2011). The Indian buffet process: An introduction and review. *J. Mach. Learn. Res.* **12** 1185–1224.

[17] HEAUKULANI, C. and ROY, D. M. (2016). The combinatorial structure of beta negative binomial processes. *Bernoulli* **22** 2301–2324.

[18] HJORT, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.* **18** 1259–1294. MR1062708

[19] HOFF, P. D., RAFTERY, A. E. and HANDCOCK, M. S. (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* **97** 1090–1098. MR1951262

[20] ISHWARAN, H. and JAMES, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.* **96** 161–173.

[21] ISHWARAN, H. and JAMES, L. F. (2004). Computational methods for multiplicative intensity models using weighted gamma processes: Proportional hazards, marked point processes, and panel count data. *J. Amer. Statist. Assoc.* **99** 175–190. MR2054297

[22] ISHWARAN, H. and RAO, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.* **33** 730–773. MR2163158

[23] JAMES, L. F. (2002). Poisson process partition calculus with applications to exchangeable models and Bayesian nonparametrics. Unpublished manuscript. Available at arXiv:math.PR/0205093.

[24] JAMES, L. F. (2003). Bayesian calculus for gamma processes with applications to semiparametric intensity models. *Sankhyā* **65** 179–206.

[25] JAMES, L. F. (2005). Bayesian Poisson process partition calculus with an application to Bayesian Lévy moving averages. *Ann. Statist.* **33** 1771–1799.

[26] JAMES, L. F. (2006). Poisson calculus for spatial neutral to the right processes. *Ann. Statist.* **34** 416–440. MR2275248

[27] JAMES, L. F. (2014). Poisson latent feature calculus for generalized Indian buffet processes. Preprint. Available at arXiv:1411.2936.

[28] JAMES, L. F., ORBANZ, P. and TEH, Y. W. (2015). Scaled subordinators and generalizations of the Indian buffet process. Preprint. Available at arXiv:1510.07309.

[29] JAMES, L. F., ROYNETTE, B. and YOR, M. (2008). Generalized gamma convolutions, Dirichlet means, Thorin measures, with explicit examples. *Probab. Surv.* **5** 346–415. MR2476736

[30] KANG, H. and CHOI, S. (2013). Bayesian multi-subject common spatial patterns with Indian buffet process priors. In *ICASSP* 3347–3351. IEEE, New York.

[31] KIM, Y. (1999). Nonparametric Bayesian estimators for counting processes. *Ann. Statist.* **27** 562–588. MR1714717

[32] KIM, Y., JAMES, L. and WEISSBACH, R. (2012). Bayesian analysis of multistate event history data: Beta-Dirichlet process prior. *Biometrika* **99** 127–140. MR2899668

[33] KIM, Y. and LEE, J. (2001). On posterior consistency of survival models. *Ann. Statist.* **29** 666–686. MR1865336

[34] KNOWLES, D. and GHAHRAMANI, Z. (2011). Nonparametric Bayesian sparse factor models with application to gene expression modeling. *Ann. Appl. Stat.* **5** 1534–1552. MR2849785

[35] LEE, J., MÜLLER, P., SENGUPTA, S., GULUKOTA, K. and JI, Y. (2014). Bayesian inference for tumor subclones accounting for sequencing and structural variants. Preprint. Available at arXiv:1409.7158 [stat.ME].

[36] LO, A. Y. (1982). Bayesian nonparametric statistical inference for Poisson point processes. *Z. Wahrsch. Verw. Gebiete* **59** 55–66. MR0643788

[37] LO, A. Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Ann. Statist.* **12** 351–357. MR0733519

[38] LO, A. Y. and WENG, C. S. (1989). On a class of Bayesian nonparametric estimates: II. Hazard rate estimates. *Ann. Inst. Statist. Math.* **41** 227–245.

[39] MILLER, K. T., GRIFFITHS, T. L. and JORDAN, M. I. (2009). Nonparametric latent feature models for link prediction. In *Advances in Neural Information Processing Systems* (*NIPS*) 22.

[40] ORBANZ, P. and ROY, D. M. (2015). Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE Trans. Pattern Anal. Mach. Intell.* **37** 437–461.

[41] PAISLEY, J. and CARIN, L. (2009). Nonparametric factor analysis with beta process priors. In *International Conference on Machine Learning* (*ICML*), Montreal, Canada.

[42] PALLA, K., KNOWLES, D. and GHAHRAMANI, Z. (2012). An infinite latent attribute model for network data. In *ICML* 2012.

[43] PITMAN, J. (1996). Some developments of the Blackwell–MacQueen urn scheme. In *Statistics, Probability and Game Theory. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **30** 245–267. IMS, Hayward, CA. MR1481784

[44] PITMAN, J. (1997). Partition structures derived from Brownian motion and stable subordinators. *Bernoulli* **3** 79–96. MR1466546

[45] PITMAN, J. (2003). Poisson–Kingman partitions. In *Statistics and Science: A Festschrift for Terry Speed. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **40** 1–34. IMS, Beachwood, OH. MR2004330

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

[47] RAI, P. and DAUME, H. (2009). The infinite hierarchical factor regression model. In *Advances in Neural Information Processing Systems* 1321–1328.

[48] TEH, Y. W. and GÖRÜR, D. (2009). Indian buffet processes with power-law behavior. In *NIPS* 2009.

[49] THIBAUX, R. (2008). Nonparametric Bayesian Models for Machine Learning. Doctoral dissertation, Univ. California, Berkeley.

[50] THIBAUX, R. and JORDAN, M. I. (2007). Hierarchical beta processes and the Indian buffet process. In *International Conference on Artificial Intelligence and Statistics* 564–571.

[51] TITSIAS, M. K. (2008). The infinite gamma-Poisson feature model. In *Advances in Neural Information Processing Systems*.

[52] WILLIAMSON, S., WANG, C., HELLER, K. and BLEI, D. (2010). The IBP compound Dirichlet process and its application to focused topic modeling. In *The* 27*th International Conference on Machine Learning* (*ICML* 2010).

[53] ZHOU, M., HANNAH, L., DUNSON, D. and CARIN, L. (2012). Beta-negative binomial process and Poisson factor analysis. AISTATS. Preprint. Available at arXiv:1112.3605.

[54] ZHOU, M., MADRID-PADILLA, O-H. and SCOTT, J. G. (2016). Priors for random count matrices derived from a family of negative binomial processes. *J. Amer. Statist. Assoc.* **111** 1144–1156.

The Annals of Statistics
2017, Vol. 45, No. 5, 2046–2073
DOI: 10.1214/16-AOS1518

INFORMATION-REGRET COMPROMISE IN COVARIATE-ADAPTIVE TREATMENT ALLOCATION

BY ASYA METELKINA AND LUC PRONZATO

Université Côte d'Azur, CNRS, I3S

Covariate-adaptive treatment allocation is considered in the situation when a compromise must be made between information (about the dependency of the probability of success of each treatment upon influential covariates) and cost (in terms of number of subjects receiving the poorest treatment). Information is measured through a design criterion for parameter estimation, the cost is additive and is related to the success probabilities. Within the framework of approximate design theory, the determination of optimal allocations forms a compound design problem. We show that when the covariates are i.i.d. with a probability measure μ, its solution possesses some similarities with the construction of optimal design measures bounded by μ. We characterize optimal designs through an equivalence theorem and construct a covariate-adaptive sequential allocation strategy that converges to the optimum. Our new optimal designs can be used as benchmarks for other, more usual, allocation methods. A response-adaptive implementation is possible for practical applications with unknown model parameters. Several illustrative examples are provided.

REFERENCES

[1] ATKINSON, A. C. (1982). Optimum biased coin designs for sequential clinical trials with prognostic factors. *Biometrika* **69** 61–67. MR0655670

[2] ATKINSON, A. C. (1999). Optimum biased-coin designs for sequential treatment allocation with covariate information. *Stat. Med.* **18** 1741–1752.

[3] ATKINSON, A. C. (2002). The comparison of designs for sequential clinical trials with covariate information. *J. Roy. Statist. Soc. Ser. A* **165** 349–373. MR1904822

[4] ATKINSON, A. C. (2014). Selecting a biased-coin design. *Statist. Sci.* **29** 144–163. MR3201860

[5] ATKINSON, A. C. and BISWAS, A. (2005). Bayesian adaptive biased-coin designs for clinical trials with normal responses. *Biometrics* **61** 118–125. MR2135851

[6] BALDI ANTOGNINI, A. and GIOVAGNOLI, A. (2010). Compound optimal allocation for individual and collective ethics in binary clinical trials. *Biometrika* **97** 935–946. MR2746162

[7] BALDI ANTOGNINI, A. and GIOVAGNOLI, A. (2015). *Adaptive Designs for Sequential Treatment Allocation*. CRC Press, Boca Raton.

[8] BALDI ANTOGNINI, A. and ZAGORAIOU, M. (2015). On the almost sure convergence of adaptive allocation procedures. *Bernoulli* **21** 881–908. MR3338650

[9] BALL, F. G., SMITH, A. F. M. and VERDINELLI, I. (1993). Biased coin designs with a Bayesian bias. *J. Statist. Plann. Inference* **34** 403–421. MR1210443

MSC2010 subject classifications. Primary 62K05; secondary 62P10.

Key words and phrases. Optimal design, treatment allocation, bounded design measure, equivalence theorem, multi-armed bandit problem.

[10] BISWAS, A., PARK, E. and BHATTACHARYA, R. (2012). Covariate-adjusted response-adaptive designs for longitudinal treatment responses: PEMF trial revisited. *Stat*. *Methods Med*. *Res*. **21** 379–392. MR3190616

[11] CAPPÉ, O., GARIVIER, A., MAILLARD, O.-A., MUNOS, R. and STOLTZ, G. (2013). Kullback–Leibler upper confidence bounds for optimal sequential allocation. *Ann*. *Statist*. **41** 1516–1541.

[12] COOK, R. D. and WONG, W. K. (1994). On the equivalence of constrained and compound optimal designs. *J*. *Amer*. *Statist*. *Assoc*. **89** 687–692. MR1294092

[13] DRAGALIN, V. and FEDOROV, V. (2006). Adaptive designs for dose-finding based on efficacy-toxicity response. *J*. *Statist*. *Plann*. *Inference* **136** 1800–1823. MR2255597

[14] DRAGALIN, V., FEDOROV, V. and WU, Y. (2008). Adaptive designs for selecting drug combinations based on efficacy-toxicity response. *J*. *Statist*. *Plann*. *Inference* **138** 352–373.

[15] DRAGALIN, V., FEDOROV, V. and WU, Y. (2008). Two-stage design for dose-finding that accounts for both efficacy and safety. *Stat*. *Med*. **27** 5156–5176.

[16] EFRON, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika* **58** 403–417. MR0312660

[17] FEDOROV, V. V. (1989). Optimal design with bounded density: Optimization algorithms of the exchange type. *J*. *Statist*. *Plann*. *Inference* **22** 1–13. MR0996795

[18] FEDOROV, V. V. and LEONOV, S. L. (2014). *Optimal Design for Nonlinear Response Models*. CRC Press, Boca Raton, FL.

[19] GOLDENSHLUGER, A. and ZEEVI, A. (2011). A note on performance limitations in bandit problems with side information. *IEEE Trans*. *Inform*. *Theory* **57** 1707–1713.

[20] GOLDENSHLUGER, A. and ZEEVI, A. (2013). A linear response bandit problem. *Stoch*. *Syst*. **3** 230–261. MR3353472

[21] HU, F. and ROSENBERGER, W. F. (2006). *The Theory of Response-Adaptive Randomization in Clinical Trials*. Wiley, Hoboken, NJ. MR2245329

[22] HU, F. and ZHANG, L.-X. (2004). Asymptotic properties of doubly adaptive biased coin designs for multitreatment clinical trials. *Ann*. *Statist*. **32** 268–301. MR2051008

[23] HU, J., ZHU, H. and HU, F. (2015). A unified family of covariate-adjusted response-adaptive designs based on efficiency and ethics. *J*. *Amer*. *Statist*. *Assoc*. **110** 357–367. MR3338508

[24] HU, Y. and HU, F. (2012). Asymptotic properties of covariate-adaptive randomization. *Ann*. *Statist*. **40** 1794–1815. MR3015044

[25] LAI, T. L. and ROBBINS, H. (1985). Asymptotically efficient adaptive allocation rules. *Adv. in Appl*. *Math*. **6** 4–22.

[26] METELKINA, A. and PRONZATO, L. (2017). Supplement to "Information-regret compromise in covariate-adaptive treatment allocation." DOI:10.1214/16-AOS1518SUPP.

[27] POCOCK, S. J. and SIMON, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* **31** 103–115.

[28] PRONZATO, L. (2006). On the sequential construction of optimum bounded designs. *J*. *Statist*. *Plann*. *Inference* **136** 2783–2804.

[29] PRONZATO, L. (2010). Penalized optimal designs for dose-finding. *J*. *Statist*. *Plann*. *Inference* **140** 283–296. MR2568139

[30] PUKELSHEIM, F. (1993). *Optimal Experimental Design*. Wiley, New York.

[31] RABIE, H. and FLOURNOY, N. (2013). Optimal designs for contingent response models with application to toxicity-efficacy studies. *J*. *Statist*. *Plann*. *Inference* **143** 1371–1379. MR3055754

[32] ROSENBERGER, W. F. and LACHIN, J. M. (2002). *Randomization in Clinical Trials*. *Theory and Practice*. Wiley, New York. MR1914364

[33] ROSENBERGER, W. F., SVERDLOV, O. and HU, F. (2012). Adaptive randomization for clinical trials. *J*. *Biopharm*. *Statist*. **22** 719–736. MR2931067

[34] SAHM, M. and SCHWABE, R. (2001). A note on optimal bounded designs. In *Optimum Design* 2000 (*Cardiff*). *Nonconvex Optim*. *Appl*. **51** 131–140. Kluwer Academic, Dordrecht. MR1829425

[35] WIENS, D. P. (2009). Robust discrimination designs. *J*. *R*. *Stat*. *Soc*., *B* **71** 805–829.

[36] WU, C. F. J. and WYNN, H. P. (1978). The convergence of general step–length algorithms for regular optimum design criteria. *Ann*. *Statist*. **6** 1273–1285. MR0523762

[37] WYNN, H. P. (1982). Optimum submeasures with applications to finite population sampling. In *Statistical Decision Theory and Related Topics III* (S. S. Gupta and J. O. Berger, eds.). *Proc*. 3*rd Purdue Symp*. **2** 485–495. Academic Press, New York.

[38] ZHANG, L.-X. and HU, F. (2009). A new family of covariate-adjusted response-adaptive designs and their properties. *Appl*. *Math*. *J*. *Chinese Univ*. **24** 1–13.

[39] ZHANG, L.-X., HU, F., CHEUNG, S. H. and CHAN, W. S. (2007). Asymptotic properties of covariate-adjusted response-adaptive designs. *Ann*. *Statist*. **35** 1166–1182. MR2341702

The Annals of Statistics
2017, Vol. 45, No. 5, 2074–2101
DOI: 10.1214/16-AOS1519

SPARSE CCA: ADAPTIVE ESTIMATION AND COMPUTATIONAL BARRIERS

BY CHAO GAO[1], ZONGMING MA[2] AND HARRISON H. ZHOU[1]

University of Chicago, University of Pennsylvania and Yale University

Canonical correlation analysis is a classical technique for exploring the relationship between two sets of variables. It has important applications in analyzing high dimensional datasets originated from genomics, imaging and other fields. This paper considers adaptive minimax and computationally tractable estimation of leading sparse canonical coefficient vectors in high dimensions. Under a Gaussian canonical pair model, we first establish separate minimax estimation rates for canonical coefficient vectors of each set of random variables under no structural assumption on marginal covariance matrices. Second, we propose a computationally feasible estimator to attain the optimal rates adaptively under an additional sample size condition. Finally, we show that a sample size condition of this kind is needed for any randomized polynomial-time estimator to be consistent, assuming hardness of certain instances of the planted clique detection problem. As a byproduct, we obtain the first computational lower bounds for sparse PCA under the Gaussian single spiked covariance model.

REFERENCES

[1] ANDERSON, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

[2] ARORA, S. and BARAK, B. (2009). *Computational Complexity*: *A Modern Approach*. Cambridge Univ. Press, Cambridge. MR2500087

[3] AVANTS, B. B., COOK, P. A., UNGAR, L., GEE, J. C. and GROSSMAN, M. (2010). Dementia induces correlated reductions in white matter integrity and cortical thickness: A multivariate neuroimaging study with sparse canonical correlation analysis. *NeuroImage* **50** 1004–1016.

[4] BAO, Z., HU, J., PAN, G. and ZHOU, W. (2014). Canonical correlation coefficients of high-dimensional normal vectors: Finite rank case. Preprint. Available at arXiv:1407.7194.

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

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

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

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

MSC2010 subject classifications. Primary 62H12; secondary 62C20.

Key words and phrases. Convex programming, group-Lasso, minimax rates, computational complexity, planted clique, sparse CCA (SCCA), sparse PCA (SPCA).

[9] BLUM, L., CUCKER, F., SHUB, M. and SMALE, S. (2012). *Complexity and Real Computation*. Springer, Berlin.

[10] BOYD, S., PARIKH, N., CHU, E., PELEATO, B. and ECKSTEIN, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Faund. Trends Mach. Learn.* **3** 1–122.

[11] BUNEA, F., LEDERER, J. and SHE, Y. (2014). The group square-root lasso: Theoretical properties and fast algorithms. *IEEE Trans. Inform. Theory* **60** 1313–1325.

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

[13] CHEN, M., GAO, C., REN, Z. and ZHOU, H. H. (2013). Sparse CCA via precision adjusted iterative thresholding. Preprint. Available at arXiv:1311.6186.

[14] DOUGLAS, J. JR. and RACHFORD, H. H. JR. (1956). On the numerical solution of heat conduction problems in two and three space variables. *Trans. Amer. Math. Soc.* **82** 421–439. MR0084194

[15] FELDMAN, V., GRIGORESCU, E., REYZIN, L., VEMPALA, S. S. and XIAO, Y. (2013). Statistical algorithms and a lower bound for detecting planted cliques. In *STOC'*13—*Proceedings of the* 2013 *ACM Symposium on Theory of Computing* 655–664. ACM, New York. MR3210827

[16] GAO, C., MA, Z., REN, Z. and ZHOU, H. H. (2015). Minimax estimation in sparse canonical correlation analysis. *Ann. Statist.* **43** 2168–2197. MR3396982

[17] GAO, C., MA, Z. and ZHOU, H. H. (2017). Supplement to "Sparse CCA: Adaptive Estimation and Computational Barriers." DOI:10.1214/16-AOS1519SUPP.

[18] HAJEK, B., WU, Y. and XU, J. (2014). Computational lower bounds for community detection on random graphs. Preprint. Available at arXiv:1406.6625.

[19] HARDOON, D. R. and SHAWE-TAYLOR, J. (2011). Sparse canonical correlation analysis. *Mach. Learn.* **83** 331–353. MR3108214

[20] HOTELLING, H. (1936). Relations between two sets of variates. *Biometrika* **28** 321–377.

[21] JOHNSTONE, I. M. (2008). Multivariate analysis and Jacobi ensembles: Largest eigenvalue, Tracy–Widom limits and rates of convergence. *Ann. Statist.* **36** 2638–2716. MR2485010

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

[23] LÊ CAO, K.-A., MARTIN, P. G. P., ROBERT-GRANIÉ, C. and BESSE, P. (2009). Sparse canonical methods for biological data integration: Application to a cross-platform study. *BMC Bioinformatics* **10** 34.

[24] LOUNICI, K., PONTIL, M., VAN DE GEER, S. and TSYBAKOV, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* **39** 2164–2204. MR2893865

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

[26] MA, Z. and WU, Y. (2013). Computational barriers in minimax submatrix detection. Preprint. Available at arXiv:1309.5914.

[27] MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1979). *Multivariate Analysis*. Academic Press, London. MR0560319

[28] 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

[29] PARKHOMENKO, E., TRITCHLER, D. and BEYENE, J. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Stat. Appl. Genet. Mol. Biol.* **8** Art. 1, 36. MR2471148

[30] ROSSMAN, B. (2010). Average-case complexity of detecting cliques. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

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

[32] VU, V. Q., CHO, J., LEI, J. and ROHE, K. (2013). Fantope projection and selection: A near-optimal convex relaxation of sparse pca. In *Advances in Neural Information Processing Systems* 2670–2678.

[33] WAAIJENBORG, S. and ZWINDERMAN, A. H. (2009). Sparse canonical correlation analysis for identifying, connecting and completing gene-expression networks. *BMC Bioinformatics* **10** 315.

[34] WANG, T., BERTHET, Q. and SAMWORTH, R. J. (2014). Statistical and computational trade-offs in estimation of sparse principal components. Preprint. Available at arXiv:1408.5369.

[35] WATSON, G. A. (1993). On matrix approximation problems with Ky Fan k norms. *Numer. Algorithms* **5** 263–272.

[36] WIESEL, A., KLIGER, M. and HERO, A. O., III (2008). A greedy approach to sparse canonical correlation analysis. Preprint. Available at arXiv:0801.2748.

[37] WITTEN, D. M., TIBSHIRANI, R. and HASTIE, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10** 515–534.

[38] YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 49–67.

[39] ZHANG, Y., WAINWRIGHT, M. J. and JORDAN, M. I. (2014). Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. Preprint. Available at arxiv:1402.1918.

The Annals of Statistics
2017, Vol. 45, No. 5, 2102–2132
DOI: 10.1214/16-AOS1520

OPTIMAL DESIGNS FOR DOSE RESPONSE CURVES WITH COMMON PARAMETERS[1]

BY CHRYSTEL FELLER*, KIRSTEN SCHORNING†, HOLGER DETTE†, GEORGINA BERMANN* AND BJÖRN BORNKAMP*

Novartis Pharma AG and Ruhr-Universität Bochum†*

A common problem in Phase II clinical trials is the comparison of dose response curves corresponding to different treatment groups. If the effect of the dose level is described by parametric regression models and the treatments differ in the administration frequency (but not in the sort of drug), a reasonable assumption is that the regression models for the different treatments share common parameters.

This paper develops optimal design theory for the comparison of different regression models with common parameters. We derive upper bounds on the number of support points of admissible designs, and explicit expressions for D-optimal designs are derived for frequently used dose response models with a common location parameter. If the location and scale parameter in the different models coincide, minimally supported designs are determined and sufficient conditions for their optimality in the class of all designs derived. The results are illustrated in a dose-finding study comparing monthly and weekly administration.

REFERENCES

BORNKAMP, B., BRETZ, F., DMITRIENKO, A., ENAS, G., GAYDOS, B., HSU, C.-H., KÖNIG, F., KRAMS, M., LIU, Q., NEUENSCHWANDER, B., PARKE, T., PINHEIRO, J. C., ROY, A., SAX, R. and SHEN, F. (2007). Innovative approaches for designing and analyzing adaptive dose-ranging trials. *J. Biopharm. Statist.* **17** 965–995.

BRETZ, F., DETTE, H. and PINHEIRO, J. (2010). Practical considerations for optimal designs in clinical dose finding studies. *Stat. Med.* **29** 731–742. MR2752038

BRETZ, F., PINHEIRO, J. C. and BRANSON, M. (2005). Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics* **61** 738–748. MR2196162

CHALONER, K. and VERDINELLI, I. (1995). Bayesian experimental design: A review. *Statist. Sci.* **10** 273–304. MR1390519

CHERNOFF, H. (1953). Locally optimal designs for estimating parameters. *Ann. Math. Stat.* **24** 586–602.

COOK, R. D. and WONG, W. K. (1994). On the equivalence of constrained and compound optimal designs *J. Amer. Statist. Assoc.* **89** 687–692.

DETTE, H. (1990). A generalization of D- and D_1-optimal designs in polynomial regression. *Ann. Statist.* **18** 1784–1805. MR1074435

DETTE, H. (1997). Designing experiments with respect to "standardized" optimality criteria. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **59** 97–110. MR1436556

MSC2010 subject classifications. Primary 62K05; secondary 62F03.

Key words and phrases. Admissible design, different treatment groups, D-optimal design, Loewner ordering, nonlinear regression.

DETTE, H. and MELAS, V. B. (2011). A note on the de la Garza phenomenon for locally optimal designs. *Ann. Statist.* **39** 1266–1281. MR2816354

DETTE, H., BRETZ, F., PEPELYSHEV, A. and PINHEIRO, J. (2008). Optimal designs for dose-finding studies. *J. Amer. Statist. Assoc.* **103** 1225–1237. MR2462895

DETTE, H., KISS, C., BEVANDA, M. and BRETZ, F. (2010). Optimal designs for the EMAX, log-linear and exponential models. *Biometrika* **97** 513–518.

DRAGALIN, V., FEDOROV, V. V. and WU, Y. (2008). Two-stage design for dose-finding that accounts for both efficacy and safety. *Stat. Med.* **27** 5156–5176. MR2516748

DRAGALIN, V., HSUAN, F. and PADMANABHAN, S. K. (2007). Adaptive designs for dose-finding studies based on the sigmoid emax model. *J. Biopharm. Statist.* **17** 1051–1070.

FANG, X. and HEDAYAT, A. S. (2008). Locally D-optimal designs based on a class of composed models resulted from blending $E_{\max}$ and one-compartment models. *Ann. Statist.* **36** 428–444. MR2387978

FELLER, C., SCHORNING, K., DETTE, H., BERMANN, G. and BORNKAMP, B. (2017). Supplement to "Optimal designs for dose response curves with common parameters." DOI:10.1214/16-AOS1520SUPP.

GABRIELSSON, J. and WEINER, D. (2007). *Pharmacokinetic and Pharmacodynamic Data Analysis: Concepts and Applications*, 4th ed. Swedish Pharmaceutical Press, Stockholm.

GILBERT, P. B. (2010). Some design issues in phase 2B vs phase 3 prevention trials for testing efficacy of products or concepts. *Stat. Med.* **29** 1061–1071. MR2756810

GRIEVE, A. P. and KRAMS, M. (2005). ASTIN: A Bayesian adaptive dose-response trial in acute stroke. *Clin. Trials* **2** 340–351.

HARVILLE, D. A. (1997). *Matrix Algebra from a Statistician's Perspective*. Springer, New York. MR1467237

KARLIN, S. and STUDDEN, W. J. (1966). *Tchebycheff Systems: With Applications in Analysis and Statistics. Pure and Applied Mathematics, Vol. XV*. Wiley, New York. MR0204922

KIEFER, J. (1974). General equivalence theory for optimum designs (approximate theory) *Ann. Statist.* **2** 849–879. MR0356386

PRONZATO, L. and WALTER, E. (1985). Robust experiment design via stochastic approximation. *Math. Biosci.* **75** 103–120. MR0800967

PUKELSHEIM, F. (2006). *Optimal Design of Experiments*. SIAM, Philadelphia, PA.

SACKS, L. V., SHAMSUDDIN, H. H., YASINSKAYA, Y. I., BOURI, K., LANTHIER, M. L. and SHERMAN, R. E. (2014). Scientific and regulatory reasons for delay and denial of FDA approval of initial applications for new drugs, 2000–2012. *J. Am. Med. Assoc.* **311** 378–384.

SILVEY, S. D. (1980). *Optimal Design*. Chapman & Hall, London. MR0606742

THOMAS, N. (2006). Hypothesis testing and Bayesian estimation using a sigmoid $E_{\max}$ model applied to sparse dose-response designs. *J. Biopharm. Statist.* **16** 657–677. MR2252313

THOMAS, N., SWEENEY, K. and SOMAYAJI, V. (2014). Meta-analysis of clinical dose-response in a large drug development portfolio. *Stat. Biopharm. Res.* **6** 302–317.

TSAI, M.-H. and ZEN, M.-M. (2004). Criterion-robust optimal designs for model discrimination and parameter estimation: Multivariate polynomial regression case. *Statist. Sinica* **14** 591–601. MR2059298

YANG, M. (2010). On the de la Garza phenomenon. *Ann. Statist.* **38** 2499–2524. MR2676896

YANG, M. and STUFKEN, J. (2012). Identifying locally optimal designs for nonlinear models: A simple extension with profound consequences. *Ann. Statist.* **40** 1665–1681. MR3015039

ZEN, M. M. and TSAI, M. H. (2004). Criterion-robust optimal designs for model discrimination and parameter estimation in Fourier regression models. *J. Statist. Plann. Inference* **124** 475–487.

The Annals of Statistics
2017, Vol. 45, No. 5, 2133–2150
DOI: 10.1214/16-AOS1521

FALSE DISCOVERIES OCCUR EARLY ON THE LASSO PATH

BY WEIJIE SU[1], MAŁGORZATA BOGDAN[2] AND EMMANUEL CANDÈS[3]

University of Pennsylvania, University of Wroclaw and Stanford University

In regression settings where explanatory variables have very low correlations and there are relatively few effects, each of large magnitude, we expect the Lasso to find the important variables with few errors, if any. This paper shows that in a regime of linear sparsity—meaning that the fraction of variables with a nonvanishing effect tends to a constant, however small—this cannot really be the case, even when the design variables are stochastically independent. We demonstrate that true features and null features are always interspersed on the Lasso path, and that this phenomenon occurs no matter how strong the effect sizes are. We derive a sharp asymptotic trade-off between false and true positive rates or, equivalently, between measures of type I and type II errors along the Lasso path. This trade-off states that if we ever want to achieve a type II error (false negative rate) under a critical value, then anywhere on the Lasso path the type I error (false positive rate) will need to exceed a given threshold so that we can never have both errors at a low level at the same time. Our analysis uses tools from approximate message passing (AMP) theory as well as novel elements to deal with a possibly adaptive selection of the Lasso regularizing parameter.

REFERENCES

[1] BARBER, R. F. and CANDÈS, E. J. (2016). A knockoff filter for high-dimensional selective inference. Preprint. Available at arXiv:1602.03574.

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

[3] 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

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

[5] BERK, R., LAWRENCE, B., BUJA, A., ZHANG, K. and ZHAO, L. (2013). Valid post-selection inference. *Ann. Statist.* **41** 802–837. MR3099122

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

[7] BÜHLMANN, P. (2011). Invited discussion on "Regression shrinkage and selection via the Lasso: A retrospective (R. Tibshirani)". *J. R. Stat. Soc. Ser. B.* **73** 277–279.

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

MSC2010 subject classifications. Primary 62F03; secondary 62J07, 62J05.

Key words and phrases. Lasso, Lasso path, false discovery rate, false negative rate, power, approximate message passing (AMP), adaptive selection of parameters.

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

[10] DESHPANDE, Y., ABBE, E. and MONTANARI, A. (2015). Asymptotic mutual information for the two-groups stochastic block model. Preprint. Available at arXiv:1507.08685.

[11] DESHPANDE, Y. and MONTANARI, A. (2014). Information-theoretically optimal sparse PCA. In *IEEE International Symposium on Information Theory* (*ISIT*) 2197–2201.

[12] DONOHO, D. and TANNER, J. (2009). Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **367** 4273–4293. MR2546388

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

[14] DONOHO, D. L. and MONTANARI, A. (2013). High dimensional robust M-estimation: Asymptotic variance via approximate message passing. Preprint. Available at arXiv:1310.7320.

[15] DONOHO, D. L. and MONTANARI, A. (2015). Variance breakdown of Huber (M)-estimators. $n/p \to m \in (1, \infty)$. Preprint. Available at arXiv:1503.02106.

[16] EBERLIN, L. S. et al. (2014). Molecular assessment of surgical-resection margins of gastric cancer by mass-spectrometric imaging. *Proc. Natl. Acad. Sci. USA* **111** 2436–2441.

[17] EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166

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

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

[20] G'SELL, M., HASTIE, T. and TIBSHIRANI, R. (2013). False variable selection rates in regression. Preprint. Available at arXiv:1302.2303.

[21] JI, P. and ZHAO, Z. (2014). Rate optimal multiple testing procedure in high-dimensional regression. Preprint. Available at arXiv:1404.2961.

[22] LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2016). Exact post-selection inference, with application to the Lasso. *Ann. Statist.* **44** 802–837. MR3485948

[23] LEE, W. et al. (2007). A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.* **39** 1235–1244.

[24] MALEKI, A., ANITORI, L., YANG, Z. and BARANIUK, R. (2013). Asymptotic analysis of complex LASSO via complex approximate message passing (CAMP). *IEEE Trans. Inform. Theory* **59** 4290–4308.

[25] MONTANARI, A. and RICHARD, E. (2014). Non-negative principal component analysis: Message passing algorithms and sharp asymptotics. Preprint. Available at arXiv:1406.4775.

[26] MOUSAVI, A., MALEKI, A. and BARANIUK, R. G. (2013). Asymptotic analysis of LASSO's solution path with implications for approximate message passing. Preprint. Available at arXiv:1309.5979.

[27] PAOLO, F. S., FRICKER, H. A. and PADMAN, L. (2015). Volume loss from Antarctic ice shelves is accelerating. *Science* **348** 327–331.

[28] POKAROWSKI, P. and MIELNICZUK, J. (2015). Combined ℓ_1 and greedy ℓ_0 penalized least squares for linear model selection. *J. Mach. Learn. Res.* **16** 961–992. MR3417774

[29] REEVES, G. and GASTPAR, M. C. (2013). Approximate sparsity pattern recovery: Information-theoretic lower bounds. *IEEE Trans. Inform. Theory* **59** 3451–3465.

[30] SANTOSA, F. and SYMES, W. W. (1986). Linear inversion of band-limited reflection seismograms. *SIAM J. Sci. Statist. Comput.* **7** 1307–1330. MR0857796

[31] SU, W., BOGDAN, M. and CANDÈS, E. (2017). Supplement to "False discoveries occur early on the Lasso path." DOI:10.1214/16-AOS1521SUPP.

[32] TIBSHIRANI, R. (1994). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288.

[33] TIBSHIRANI, R. J. and TAYLOR, J. (2011). The solution path of the generalized lasso. *Ann. Statist.* **39** 1335–1371. MR2850205

[34] TIBSHIRANI, R. J., TAYLOR, J., LOCKHART, R. and TIBSHIRANI, R. (2014). Exact post-selection inference for sequential regression procedures. Preprint. Available at arXiv:1401.3889.

[35] WAINWRIGHT, M. J. (2009). Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inform. Theory* **55** 5728–5741. MR2597190

[36] WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55** 2183–2202. MR2729873

[37] YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. A* **68** 49–67.

[38] YUAN, Y. et al. (2014). Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nat. Biotechnol.* **32** 644–652.

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

[40] ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

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

The Annals of Statistics
2017, Vol. 45, No. 5, 2151–2189
DOI: 10.1214/16-AOS1522

PHASE TRANSITIONS FOR HIGH DIMENSIONAL CLUSTERING AND RELATED PROBLEMS

BY JIASHUN JIN, ZHENG TRACY KE AND WANJIE WANG

Carnegie Mellon University, University of Chicago and University of Pennsylvania

Consider a two-class clustering problem where we observe $X_i = \ell_i \mu + Z_i$, $Z_i \stackrel{i.i.d.}{\sim} N(0, I_p)$, $1 \le i \le n$. The feature vector $\mu \in R^p$ is unknown but is presumably sparse. The class labels $\ell_i \in \{-1, 1\}$ are also unknown and the main interest is to estimate them.

We are interested in the statistical limits. In the two-dimensional phase space calibrating the rarity and strengths of useful features, we find the precise demarcation for the *Region of Impossibility* and *Region of Possibility*. In the former, useful features are too rare/weak for successful clustering. In the latter, useful features are strong enough to allow successful clustering. The results are extended to the case of colored noise using Le Cam's idea on comparison of experiments.

We also extend the study on statistical limits for clustering to that for signal recovery and that for global testing. We compare the statistical limits for three problems and expose some interesting insight.

We propose classical PCA and Important Features PCA (IF-PCA) for clustering. For a threshold $t > 0$, IF-PCA clusters by applying classical PCA to all columns of X with an L^2-norm larger than t. We also propose two aggregation methods. For any parameter in the Region of Possibility, some of these methods yield successful clustering.

We discover a phase transition for IF-PCA. For any threshold $t > 0$, let $\xi^{(t)}$ be the first left singular vector of the post-selection data matrix. The phase space partitions into two different regions. In one region, there is a t such that $\cos(\xi^{(t)}, \ell) \to 1$ and IF-PCA yields successful clustering. In the other, $\cos(\xi^{(t)}, \ell) \le c_0 < 1$ for all $t > 0$.

Our results require delicate analysis, especially on post-selection random matrix theory and on lower bound arguments.

REFERENCES

[1] ADDARIO-BERRY, L., BROUTIN, N., DEVROYE, L. and LUGOSI, G. (2010). On combinatorial testing problems. *Ann. Statist.* **38** 3063–3092. MR2722464

[2] ALDOUS, D. J. (1985). Exchangeability and related topics. In *École D'été de Probabilités de Saint-Flour, XIII—1983*. *Lecture Notes in Math.* **1117** 1–198. Springer, Berlin. MR0883646

MSC2010 subject classifications. Primary 62H30, 62H25; secondary 62G05, 62G10.

Key words and phrases. Clustering, comparison of experiments, feature selection, hypothesis testing, L^1-distance, lower bound, low-rank matrix recovery, phase transition.

[3] AMINI, A. and WAINWRIGHT, M. J. (2008). High-dimensional analysis of semidefinite relaxations for sparse principal components. In *IEEE International Symposium on Information Theory* 2454–2458. IEEE, New York.

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

[5] ARIAS-CASTRO, E. and VERZELEN, N. (2014). Detection and feature selection in sparse mixture models. arXiv:1405.1478.

[6] ARTHUR, D. and VASSILVITSKII, S. (2007). `k-means++`: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms* 1027–1035. ACM, New York. MR2485254

[7] AZIZYAN, M., SINGH, A. and WASSERMAN, L. (2013). Minimax theory for high-dimensional Gaussian mixtures with sparse mean separation. In *NIPS* 2139–2147.

[8] BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **57** 289–300. MR1325392

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

[10] CAI, T., MA, Z. and WU, Y. (2013). Optimal estimation and rank detection for sparse spiked covariance matrices. *Probab. Theory Related Fields* 1–35.

[11] CANDÈS, E. J. and RECHT, B. (2009). Exact matrix completion via convex optimization. *Found. Comput. Math.* **9** 717–772. MR2565240

[12] CHAN, Y. and HALL, P. (2010). Using evidence of mixed populations to select variables for clustering very high-dimensional data. *J. Amer. Statist. Assoc.* **105**.

[13] D'ASPREMONT, A., EL GHAOUI, L., JORDAN, M. I. and LANCKRIET, G. R. G. (2007). A direct formulation for sparse PCA using semidefinite programming. *SIAM Rev.* **49** 434–448. MR2353806

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

[15] DETTLING, M. (2004). BagBoosting for tumor classification with gene expression data. *Bioinformatics* **20** 3583–3593.

[16] DONOHO, D. (2015). 50 years of data science. Manuscript.

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

[18] DONOHO, D. and JIN, J. (2008). Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proc. Natl. Acad. Sci. USA* **105** 14790–14795.

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

[20] DONOHO, D. L. and JOHNSTONE, I. M. (1998). Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26** 879–921. MR1635414

[21] DONOHO, D. L., MALEKI, A., RAHMAN, I. U., SHAHRAM, M. and STODDEN, V. (2009). Reproducible research in computational harmonic analysis. *Comput. Sci. Eng.* **11** 8–18.

[22] HALL, P. and JIN, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist.* **38** 1686–1732. MR2662357

[23] HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*, 2nd ed. Springer, Berlin.

[24] INGSTER, Y. I., POUET, C. and TSYBAKOV, A. B. (2009). Classification of sparse high-dimensional vectors. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **367** 4427–4448. MR2546395

[25] JIN, J. and KE, Z. T. (2016). Rare and weak effects in large-scale inference: Methods and phase diagrams. *Statist. Sinica* **26** 1–34. MR3468343

[26] JIN, J., KE, Z. T. and WANG, W. (2014). Optimal spectral clustering by higher criticism thresholding. Manuscript.

[27] JIN, J., KE, Z. T. and WANG, W. (2017). Supplementary material for "Phase transitions for high dimensional clustering and related problems." DOI:10.1214/16-AOS1522SUPP.

[28] JIN, J. and WANG, W. (2016). Influential features PCA for high dimensional clustering. *Ann. Statist.* **44** 2323–2359. MR3576543

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

[30] KE, Z. T., JIN, J. and FAN, J. (2014). Covariate assisted screening and estimation. *Ann. Statist.* **42** 2202–2242.

[31] LEE, A. B., LUCA, D. and ROEDER, K. (2010). A spectral graph approach to discovering genetic ancestry. *Ann. Appl. Stat.* **4** 179–202.

[32] LEI, J. and VU, V. Q. (2015). Sparsistency and agnostic inference in sparse PCA. *Ann. Statist.* **43** 299–322.

[33] LE CAM, L. and YANG, G. L. (2000). *Asymptotics in Statistics*: *Some Basic Concepts*, 2nd ed. Springer, New York. MR1784901

[34] MA, Z. and WU, Y. (2015). Computational barriers in minimax submatrix detection. *Ann. Statist.* **43** 1089–1116.

[35] PAN, W. and SHEN, X. (2007). Penalized model-based clustering with application to variable selection. *J. Mach. Learn. Res.* **8** 1145–1164.

[36] RAFTERY, A. E. and DEAN, N. (2006). Variable selection for model-based clustering. *J. Amer. Statist. Assoc.* **101** 168–178. MR2268036

[37] ROGERS, C. A. (1963). Covering a sphere with spheres. *Mathematika* **10** 157–164.

[38] SHORACK, G. and WELLNER, J. (1986). *Empirical Processes with Applications to Statistics*. John Wiley & Sons, New York.

[39] SPIEGELHALTER, D. J. (2014). Statistics. The future lies in uncertainty. *Science* **345** 264–265.

[40] SUN, W., WANG, J., FANG, Y. et al. (2012). Regularized k-means clustering of high-dimensional data and its asymptotic consistency. *Electron. J. Stat.* **6** 148–167.

[41] VAN DER VAART, A. (2000). *Asymptotic Statistics* **3**. Cambridge Univ. Press, Cambridge.

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

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

[44] WANG, Z., LU, H. and LIU, H. (2014). Nonconvex statistical optimization: Minimax-optimal sparse PCA in polynomial time. arXiv:1408.5352.

[45] WEYL, H. (1912). Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung). *Math. Ann.* **71** 441–479.

[46] WITTEN, D. M. and TIBSHIRANI, R. (2012). A framework for feature selection in clustering. *J. Amer. Statist. Assoc.* **105** 713–726.

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

The Annals of Statistics
2017, Vol. 45, No. 5, 2190–2217
DOI: 10.1214/16-AOS1523

BAYESIAN DETECTION OF IMAGE BOUNDARIES[1]

BY MENG LI AND SUBHASHIS GHOSAL

Duke University and North Carolina State University

Detecting boundary of an image based on noisy observations is a fundamental problem of image processing and image segmentation. For a d-dimensional image ($d = 2, 3, \ldots$), the boundary can often be described by a closed smooth $(d - 1)$-dimensional manifold. In this paper, we propose a nonparametric Bayesian approach based on priors indexed by $\mathbb{S}^{d-1}$, the unit sphere in $\mathbb{R}^d$. We derive optimal posterior contraction rates for Gaussian processes or finite random series priors using basis functions such as trigonometric polynomials for 2-dimensional images and spherical harmonics for 3-dimensional images. For 2-dimensional images, we show a rescaled squared exponential Gaussian process on $\mathbb{S}^1$ achieves four goals of guaranteed geometric restriction, (nearly) minimax optimal rate adapting to the smoothness level, convenience for joint inference and computational efficiency. We conduct an extensive study of its reproducing kernel Hilbert space, which may be of interest by its own and can also be used in other contexts. Several new estimates on modified Bessel functions of the first kind are given. Simulations confirm excellent performance and robustness of the proposed method.

REFERENCES

[1] ARBEL, J., GAYRAUD, G. and ROUSSEAU, J. (2013). Bayesian optimal adaptive estimation using a sieve prior. *Scand. J. Stat.* **40** 549–570. MR3091697

[2] BANERJEE, S. and GELFAND, A. E. (2006). Bayesian wombling: Curvilinear gradient assessment under spatial process models. *J. Amer. Statist. Assoc.* **101** 1487–1501. MR2279474

[3] BASU, M. (2002). Gaussian-based edge-detection methods-a survey. *IEEE Trans. Syst. Man Cybern., Part C Appl. Rev.* **32** 252–260.

[4] BHARDWAJ, S. and MITTAL, A. (2012). A survey on various edge detector techniques. *Proc. Technol.* **4** 220–226.

[5] CARLSTEIN, E. and KRISHNAMOORTHY, C. (1992). Boundary estimation. *J. Amer. Statist. Assoc.* **87** 430–438. MR1173808

[6] CASTILLO, I. (2012). A semiparametric Bernstein–von Mises theorem for Gaussian process priors. *Probab. Theory Related Fields* **152** 53–99. MR2875753

[7] CASTILLO, I., KERKYACHARIAN, G. and PICARD, D. (2014). Thomas Bayes' walk on manifolds. *Probab. Theory Related Fields* **158** 665–710. MR3176362

[8] CHEN, J. and GUPTA, A. K. (2012). *Parametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance*, 2nd ed. Birkhäuser/Springer, New York. MR3025631

[9] DAI, F. and XU, Y. (2013). *Approximation Theory and Harmonic Analysis on Spheres and Balls*. Springer, New York. MR3060033

MSC2010 subject classifications. Primary 62G20, 62H35; secondary 62F15, 60G15.

Key words and phrases. Boundary detection, Gaussian process on sphere, image, posterior contraction rate, random series, squared exponential periodic kernel, Bayesian adaptation.

[10] DONOHO, D. L. (1999). Wedgelets: Nearly minimax estimation of edges. *Ann. Statist.* **27** 859–897. MR1724034

[11] DUDLEY, R. M. (1974). Metric entropy of some classes of sets with differentiable boundaries. *J. Approx. Theory* **10** 227–236. MR0358168

[12] ENNIS, D. (2005). Spherical Harmonics. http://www.mathworks.com/matlabcentral/fileexchange/8638-spherical-harmonics. MATLAB Central File Exchange.

[13] FITZPATRICK, M. C., PREISSER, E. L., PORTER, A., ELKINTON, J., WALLER, L. A., CARLIN, B. P. and ELLISON, A. M. (2010). Ecological boundary detection using Bayesian areal wombling. *Ecology* **91** 3448–3455.

[14] GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6** 721–741.

[15] GHOSAL, S. and VAN DER VAART, A. (2007). Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.* **35** 192–223. MR2332274

[16] GU, K., PATI, D. and DUNSON, D. B. (2014). Bayesian multiscale modeling of closed curves in point clouds. *J. Amer. Statist. Assoc.* **109** 1481–1494. MR3293605

[17] HALL, P., PENG, L. and RAU, C. (2001). Local likelihood tracking of fault lines and boundaries. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 569–582. MR1858403

[18] HASTIE, T. and STUETZLE, W. (1989). Principal curves. *J. Amer. Statist. Assoc.* **84** 502–516. MR1010339

[19] JACKSON, D. (1994). *The Theory of Approximation. American Mathematical Society Colloquium Publications* **11**. Amer. Math. Soc., Providence, RI. Reprint of the 1930 original. MR1451140

[20] KILLICK, R. and ECKLEY, I. A. (2011). Changepoint: An R package for changepoint analysis. R Package Version 0.6, http://CRAN.R-project.Org/package=changepoint.

[21] KILLICK, R., FEARNHEAD, P. and ECKLEY, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.* **107** 1590–1598. MR3036418

[22] KOROSTELËV, A. P. and TSYBAKOV, A. B. (1993). *Minimax Theory of Image Reconstruction. Lecture Notes in Statistics* **82**. Springer, New York. MR1226450

[23] KUELBS, J. and LI, W. V. (1993). Metric entropy and the small ball problem for Gaussian measures. *J. Funct. Anal.* **116** 133–157. MR1237989

[24] LI, M. and GHOSAL, S. (2017). Supplement to "Bayesian detection of image boundaries." DOI:10.1214/16-AOS1523SUPP.

[25] LI, W. V. and LINDE, W. (1999). Approximation, metric entropy and small ball estimates for Gaussian measures. *Ann. Probab.* **27** 1556–1578. MR1733160

[26] LU, H. and CARLIN, B. P. (2005). Bayesian areal wombling for geographical boundary analysis. *Geogr. Anal.* **37** 265–285.

[27] MACKAY, D. J. (1998). Introduction to Gaussian processes. In *NATO ASI Series F Computer and Systems Sciences* **168** 133–166.

[28] MAMMEN, E. and TSYBAKOV, A. B. (1995). Asymptotical minimax recovery of sets with smooth boundaries. *Ann. Statist.* **23** 502–524. MR1332579

[29] MÜLLER, H.-G. and SONG, K. S. (1994). Maximin estimation of multidimensional boundaries. *J. Multivariate Anal.* **50** 265–281. MR1293046

[30] NEAL, R. M. (2003). Slice sampling. *Ann. Statist.* **31** 705–767. With discussions and a rejoinder by the author. MR1994729

[31] POLZEHL, J. and SPOKOINY, V. (2003). Image denoising: Pointwise adaptive approach. *Ann. Statist.* **31** 30–57. MR1962499

[32] QIU, P. (2005). *Image Processing and Jump Regression Analysis*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ. MR2111430

[33] QIU, P. (2007). Jump surface estimation, edge detection, and image restoration. *J. Amer. Statist. Assoc.* **102** 745–756. MR2370864

[34] QIU, P. and SUN, J. (2007). Local smoothing image segmentation for spotted microarray images. *J. Amer. Statist. Assoc.* **102** 1129–1144. MR2412538

[35] QIU, P. and SUN, J. (2009). Using conventional edge detectors and postsmoothing for segmentation of spotted microarray images. *J. Comput. Graph. Statist.* **18** 147–164. MR2649642

[36] QIU, P. and YANDELL, B. (1997). Jump detection in regression surfaces. *J. Comput. Graph. Statist.* **6** 332–354. MR1466871

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

[38] RUDEMO, M. and STRYHN, H. (1994). Approximating the distribution of maximum likelihood contour estimators in two-region images. *Scand. J. Stat.* **21** 41–55. MR1267042

[39] SHEN, W. and GHOSAL, S. (2015). Adaptive Bayesian procedures using random series priors. *Scand. J. Stat.* **42** 1194–1213. MR3426318

[40] TERRAS, A. (2013). *Harmonic Analysis on Symmetric Spaces—Euclidean Space, the Sphere, and the Poincaré Upper Half-Plane*, 2nd ed. Springer, New York. MR3100414

[41] VAN DER VAART, A. and VAN ZANTEN, H. (2007). Bayesian inference with rescaled Gaussian process priors. *Electron. J. Stat.* **1** 433–448. MR2357712

[42] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.* **36** 1435–1463. MR2418663

[43] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.* **37** 2655–2675. MR2541442

[44] WALLER, L. A. and GOTWAY, C. A. (2004). *Applied Spatial Statistics for Public Health Data*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ. MR2075123

[45] WATSON, G. N. (1995). *A Treatise on the Theory of Bessel Functions*. Cambridge Univ. Press, Cambridge. Reprint of the second (1944) edition. MR1349110

The Annals of Statistics
2017, Vol. 45, No. 5, 2218–2247
DOI: 10.1214/16-AOS1525

SPECTRUM ESTIMATION FROM SAMPLES

BY WEIHAO KONG AND GREGORY VALIANT[1]

Stanford University

We consider the problem of approximating the set of eigenvalues of the covariance matrix of a multivariate distribution (equivalently, the problem of approximating the "population spectrum"), given access to samples drawn from the distribution. We consider this recovery problem in the regime where the sample size is comparable to, or even sublinear in the dimensionality of the distribution. First, we propose a theoretically optimal and computationally efficient algorithm for recovering the moments of the eigenvalues of the population covariance matrix. We then leverage this accurate moment recovery, via a Wasserstein distance argument, to accurately reconstruct the vector of eigenvalues. Together, this yields an eigenvalue reconstruction algorithm that is asymptotically consistent as the dimensionality of the distribution and sample size tend toward infinity, even in the sublinear sample regime where the ratio of the sample size to the dimensionality tends to zero. In addition to our theoretical results, we show that our approach performs well in practice for a broad range of distributions and sample sizes.

REFERENCES

[1] ALON, N., YUSTER, R. and ZWICK, U. (1997). Finding and counting given length cycles. *Algorithmica* **17** 209–223. MR1425734
[2] ANDERSON, T. W. (1963). Asymptotic theory for principal component analysis. *Ann. Math. Stat.* **34** 122–148. MR0145620
[3] BAI, Z., CHEN, J. and YAO, J. (2010). On estimation of the population spectral distribution from a high-dimensional sample covariance matrix. *Aust. N. Z. J. Stat.* **52** 423–437. MR2791528
[4] BAI, Z. and SILVERSTEIN, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed. Springer, New York. MR2567175
[5] BICKEL, P. J. and LEVINA, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969
[6] BURDA, Z., GÖRLICH, A., JAROSZ, A. and JURKIEWICZ, J. (2004). Signal and noise in correlation matrix. *Phys. A* **343** 295–310. MR2094415
[7] BURDA, Z., JURKIEWICZ, J. and WACŁAW, B. (2005). Spectral moments of correlated Wishart matrices. *Phys. Rev. E* (3) **71** 026111. MR2139960
[8] CAROTHERS, N. (2000). A short course on approximation theory.
[9] DE VILLIERS, J. (2012). *Mathematics of Approximation*. *Mathematics Textbooks for Science and Engineering* **1**. Atlantis Press, Paris. MR2962227
[10] DEY, D. K. and SRINIVASAN, C. (1985). Estimation of a covariance matrix under Stein's loss. *Ann. Statist.* **13** 1581–1591. MR0811511

MSC2010 subject classifications. 62H12, 62H10.
Key words and phrases. Spectrum estimation, eigenvalues of covariance matrices, sublinear sample size, method of moments, random matrix theory, high-dimensional inference.

[11] DONOHO, D. L., GAVISH, M. and JOHNSTONE, I. M. (2013). Optimal shrinkage of eigenvalues in the spiked covariance model. Preprint. Available at arXiv:1311.0851.

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

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

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

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

[16] HARDT, M., LIGETT, K. and MCSHERRY, F. (2012). A simple and practical algorithm for differentially private data release. In *Advances in Neural Information Processing Systems* 2339–2347.

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

[18] KANE, D. M., NELSON, J. and WOODRUFF, D. P. (2010). On the exact space complexity of sketching and streaming small norms. In *Proceedings of the Twenty-First Annual ACM–SIAM Symposium on Discrete Algorithms* 1161–1178. SIAM, Philadelphia, PA. MR2809734

[19] KANTOROVIČ, L. V. and RUBINŠTEĬN, G. Š. (1957). On a functional space and certain extremum problems. *Dokl. Akad. Nauk SSSR (N.S.)* **115** 1058–1061. MR0094707

[20] KONG, W. and VALIANT, G. (2017). Supplement to "Spectrum estimation from samples." DOI:10.1214/16-AOS1525SUPP.

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

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

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

[24] LEDOIT, O. and WOLF, M. (2013). Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions. Available at SSRN 2198287.

[25] LI, W., CHEN, J., QIN, Y., BAI, Z. and YAO, J. (2013). Estimation of the population spectral distribution from a large dimensional sample covariance matrix. *J. Statist. Plann. Inference* **143** 1887–1897.

[26] LI, W. and YAO, J. (2014). A local moment estimator of the spectrum of a large dimensional covariance matrix. *Statist. Sinica* **24** 919–936. MR3235405

[27] LI, Y., NGUYEN, H. L. and WOODRUFF, D. P. (2014). On sketching matrix norms and the top singular vector. In *Proceedings of the Twenty-Fifth Annual ACM–SIAM Symposium on Discrete Algorithms* 1562–1581. ACM, New York. MR3376474

[28] MAHONEY, M. W. (2011). Randomized algorithms for matrices and data. *Found. Trends Mach. Learn.* **3** 123–224.

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

[30] MESTRE, X. (2008). Improved estimation of eigenvalues and eigenvectors of covariance matrices using their sample estimates. *IEEE Trans. Inform. Theory* **54** 5113–5129.

[31] RAO, N. R., MINGO, J. A., SPEICHER, R. and EDELMAN, A. (2008). Statistical eigen-inference from large Wishart matrices. *Ann. Statist.* **36** 2850–2885. MR2485015

[32] SCHÄFER, J. and STRIMMER, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.* **4** Art. 32, 28. MR2183942

[33] SILVERSTEIN, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *J. Multivariate Anal.* **55** 331–339.

[34] STEIN, C. (1975). Estimation of a covariance matrix. *Rietz Lecture, 39th Annual IMS Meeting, Atlanta, GA*.

[35] STEIN, C. (1977). Lectures on the theory of estimation of many parameters. In *Studies in the Statistical Theory of Estimation I. Proceedings of Scientific Seminars of the Steklov Institute*, *Leningrad Division* **74** 4–65.

[36] TAKEMURA, A. (1984). An orthogonally invariant minimax estimator of the covariance matrix of a multivariate normal population. *Tsukuba J. Math.* **8** 367–376. MR0767967

[37] VALIANT, G. and VALIANT, P. (2011). Estimating the unseen: An $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing* 685–694. ACM, New York.

[38] YIN, Y. Q. and KRISHNAIAH, P. R. (1983). A limit theorem for the eigenvalues of product of two random matrices. *J. Multivariate Anal.* **13** 489–507.

The Annals of Statistics
2017, Vol. 45, No. 5, 2248–2273
DOI: 10.1214/16-AOS1526

ON THE CONTRACTION PROPERTIES OF SOME HIGH-DIMENSIONAL QUASI-POSTERIOR DISTRIBUTIONS

BY YVES A. ATCHADÉ[1]

University of Michigan

We study the contraction properties of a quasi-posterior distribution $\check{\Pi}_{n,d}$ obtained by combining a quasi-likelihood function and a sparsity inducing prior distribution on $\mathbb{R}^d$, as both n (the sample size), and d (the dimension of the parameter) increase. We derive some general results that highlight a set of sufficient conditions under which $\check{\Pi}_{n,d}$ puts increasingly high probability on sparse subsets of $\mathbb{R}^d$, and contracts toward the true value of the parameter. We apply these results to the analysis of logistic regression models, and binary graphical models, in high-dimensional settings. For the logistic regression model, we shows that for well-behaved design matrices, the posterior distribution contracts at the rate $O(\sqrt{s_\star \log(d)/n})$, where $s_\star$ is the number of nonzero components of the parameter. For the binary graphical model, under some regularity conditions, we show that a quasi-posterior analog of the neighborhood selection of [*Ann. Statist.* **34** (2006) 1436–1462] contracts in the Frobenius norm at the rate $O(\sqrt{(p+S)\log(p)/n})$, where p is the number of nodes, and S the number of edges of the true graph.

REFERENCES

[1] ALQUIER, P. and LOUNICI, K. (2011). PAC-Bayesian bounds for sparse regression estimation with exponential weights. *Electron. J. Stat.* **5** 127–145. MR2786484

[2] ARIAS-CASTRO, E. and LOUNICI, K. (2014). Estimation and variable selection with exponential weights. *Electron. J. Stat.* **8** 328–354. MR3195119

[3] ATCHADÉ, Y. F. (2014). Estimation of high-dimensional partially-observed discrete Markov random fields. *Electron. J. Stat.* **8** 2242–2263.

[4] ATCHADÉ, Y. F. (2015). A Moreau-Yosida approximation scheme for high-dimensional posterior and quasi-posterior distributions. Available at arXiv:1505.07072.

[5] ATCHADÉ, Y. F. (2015). A scalable quasi-Bayesian framework for Gaussian graphical models. Available at arXiv:1512.07934.

[6] ATCHADÉ, Y. F. (2017). Supplement to "On the contraction properties of some high-dimensional quasi-posterior distributions." DOI:10.1214/16-AOS1526SUPP.

[7] 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.

[8] BANERJEE, S. and GHOSAL, S. (2015). Bayesian structure learning in graphical models. *J. Multivariate Anal.* **136** 147–162. MR3321485

[9] BARBER, R. F. and DRTON, M. (2015). High-dimensional Ising model selection with Bayesian information criteria. *Electron. J. Stat.* **9** 567–607.

MSC2010 subject classifications. 62F15, 62Jxx.

Key words and phrases. Quasi-Bayesian inference, high-dimensional inference, Bayesian asymptotics, logistic regression models, discrete graphical models.

[10] Baricz, A. (2008). Mills' ratio: Monotonicity patterns and functional inequalities. *J. Math. Anal. Appl.* **340** 1362–1370.

[11] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **36** 192–236. MR0373208

[12] Castillo, I., Schmidt-Hieber, J. and van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.* **43** 1986–2018.

[13] Castillo, I. and van der Vaart, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Ann. Statist.* **40** 2069–2101.

[14] Catoni, O. (2004). *Statistical Learning Theory and Stochastic Optimization*. *Lecture Notes in Math.* **1851**. Springer, Berlin. MR2163920

[15] Chernozhukov, V. and Hong, H. (2003). An MCMC approach to classical estimation. *J. Econometrics* **115** 293–346.

[16] Dalalyan, A. S. and Tsybakov, A. B. (2007). Aggregation by exponential weighting and sharp oracle inequalities. In *Learning Theory*. *Lecture Notes in Computer Science* **4539** 97–111. Springer, Berlin. MR2397581

[17] Florens, J.-P. and Simoni, A. (2012). Nonparametric estimation of an instrumental regression: A quasi-Bayesian approach based on regularized posterior. *J. Econometrics* **170** 458–475.

[18] Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics*. *Springer Series in Statistics*. Springer, New York. MR1992245

[19] Höfling, H. and Tibshirani, R. (2009). Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. *J. Mach. Learn. Res.* **10** 883–906. MR2505138

[20] Kato, K. (2013). Quasi-Bayesian analysis of nonparametric instrumental variables models. *Ann. Statist.* **41** 2359–2390. MR3127869

[21] Kleijn, B. J. K. and van der Vaart, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist.* **34** 837–877. MR2283395

[22] Li, C. and Jiang, W. (2014). Model selection for likelihood-free Bayesian methods based on moment conditions: Theory and numerical examples. Available at arXiv:1405.6693v1.

[23] Li, Y.-H., Scarlett, J., Ravikumar, P. and Cevher, V. (2014). Sparsistency of ℓ_1-regularized M-estimators. Preprint. Available at arXiv:1410.7605v1.

[24] Liao, Y. and Jiang, W. (2011). Posterior consistency of nonparametric conditional moment restricted models. *Ann. Statist.* **39** 3003–3031. MR3012399

[25] Marin, J.-M., Pudlo, P., Robert, C. P. and Ryder, R. J. (2012). Approximate Bayesian computational methods. *Stat. Comput.* **22** 1167–1180. MR2992292

[26] McAllester, D. A. (1999). Some pac-Bayesian theorems. *Mach. Learn.* **37** 355–363.

[27] Meinshausen, N. and Buhlmann, P. (2006). High-dimensional graphs with the lasso. *Ann. Statist.* **34** 1436–1462.

[28] Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83** 1023–1036. MR0997578

[29] 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.

[30] Ravikumar, P., Wainwright, M. J. and Lafferty, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.* **38** 1287–1319.

[31] Schreck, A., Fort, G., Le Corff, S. and Moulines, E. (2013). A shrinkage-thresholding Metropolis adjusted Langevin algorithm for Bayesian variable selection. Available at arXiv:1312.5658.

[32] Sun, T. and Zhang, C.-H. (2013). Sparse matrix inversion with scaled lasso. *J. Mach. Learn. Res.* **14** 3385–3418. MR3144466

[33] Yang, W. and He, X. (2012). Bayesian empirical likelihood for quantile regression. *Ann. Statist.* **40** 1102–1131.

[34] ZHANG, T. (2006). From ε-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *Ann*. *Statist*. **34** 2180–2210. MR2291497

The Annals of Statistics
2017, Vol. 45, No. 5, 2274–2298
DOI: 10.1214/16-AOS1528

NONASYMPTOTIC ANALYSIS OF SEMIPARAMETRIC REGRESSION MODELS WITH HIGH-DIMENSIONAL PARAMETRIC COEFFICIENTS[1]

BY YING ZHU

Michigan State University

We consider a two-step projection based Lasso procedure for estimating a partially linear regression model where the number of coefficients in the linear component can exceed the sample size and these coefficients belong to the l_q-"balls" for $q \in [0, 1]$. Our theoretical results regarding the properties of the estimators are nonasymptotic. In particular, we establish a new nonasymptotic "oracle" result: Although the error of the nonparametric projection *per se* (with respect to the prediction norm) has the scaling t_n in the first step, it only contributes a scaling t_n^2 in the l_2-error of the second-step estimator for the linear coefficients. This new "oracle" result holds for a large family of nonparametric least squares procedures and regularized nonparametric least squares procedures for the first-step estimation and the driver behind it lies in the projection strategy. We specialize our analysis to the estimation of a semiparametric sample selection model and provide a simple method with theoretical guarantees for choosing the regularization parameter in practice.

REFERENCES

AHN, H. and POWELL, J. L. (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *J. Econometrics* **58** 3–29. MR1230978

BARTLETT, P. L. and MENDELSON, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.* **3** 463–482. MR1984026

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

BOBKOV, S. G. and LEDOUX, M. (2000). From Brunn–Minkowski to Brascamp–Lieb and to logarithmic Sobolev inequalities. *Geom. Funct. Anal.* **10** 1028–1052.

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

BUNEA, F. (2004). Consistent covariate selection and post model selection inference in semiparametric regression. *Ann. Statist.* **32** 898–927.

BUNEA, F. and WEGKAMP, M. H. (2004). Two-stage model selection procedures in partially linear regression. *Canad. J. Statist.* **32** 105–118. MR2064395

CHEN, H. (1988). Convergence rates for parametric components in a partly linear model. *Ann. Statist.* **16** 136–146.

DONALD, S. G. and NEWEY, W. K. (1994). Series estimation of semilinear models. *J. Multivariate Anal.* **50** 30–40.

MSC2010 subject classifications. Primary 62J02; secondary 62N01, 62N02, 62G08, 62J12.

Key words and phrases. High-dimensional statistics, Lasso, nonasymptotic analysis, partially linear models, sample selection.

ENGLE, R., GRANGER, C., RICE, J. and WEISS, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Amer. Statist. Assoc.* **81** 310–320.

FAN, J. and LI, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *J. Amer. Statist. Assoc.* **99** 710–723. MR2090905

GRONAU, R. (1973). The effects of children on the housewife's value of time. *J. Polit. Econ.* **81** S168–S199.

HÄRDLE, W., LIANG, H. and GAO, J. T. (2000). *Partially Linear Models*. Springer, Heidelberg.

HECKMAN, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Ann. Econ. Soc. Meas.* **5** 475–492.

KAKADE, S., KALAI, A. T., KANADE, V. and SHAMIR, O. (2011). Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems* **23** 927–935.

KOLTCHINSKII, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.* **34** 2593–2656. MR2329442

LEDOUX, M. (1995/97). On Talagrand's deviation inequalities for product measures. *ESAIM Probab. Stat.* **1** 63–87. MR1399224

LIANG, H. and LI, R. (2009). Variable selection for partially linear models with measurement errors. *J. Amer. Statist. Assoc.* **104** 234–248. MR2504375

LIANG, H., LIU, X., LI, R. and TSAI, C.-L. (2010). Estimation and testing for partially linear single-index models. *Ann. Statist.* **38** 3811–3836.

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

MEINSHAUSEN, N. and YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37** 246–270. MR2488351

MENDELSON, S. (2002). Geometric parameters of kernel machines. In *Proceedings of COLT* 29–43.

MINSKER, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli* **21** 2308–2335. MR3378468

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

NEWEY, W. K. (2009). Two-step series estimation of sample selection models. *Econom. J.* **12** S217–S229. MR2543488

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

ROBINSON, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica* **56** 931–954. MR0951762

ROSENBAUM, M. and TSYBAKOV, A. B. (2013). Improved matrix uncertainty selector. In *From Probability to Statistics and Back: High-Dimensional Models and Processes—A Festschrift in Honor of Jon* (A. Wellner, M. Banerjee et al., eds.). *IMS Collections* **9** 276–290. IMS, Beachwood, OH.

RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003). *Semiparametric Regression*. *Cambridge Series in Statistical and Probabilistic Mathematics* **12**. Cambridge Univ. Press, Cambridge. MR1998720

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

VAN DE GEER, S. (2000). *Empirical Processes in M-Estimation*. Cambridge Univ. Press, Cambridge.

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

WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55** 2183–2202. MR2729873

WAINWRIGHT, J. M. (2015). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Univ. California, Berkeley. In preparation.

WOOLDRIDGE, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge.

YATCHEW, A. (2003). *Semiparametric Regression for the Applied Econometrician*. Cambridge Univ. Press, Cambridge.

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

YU, Y. and RUPPERT, D. (2002). Penalized spline estimation for partially linear single-index models. *J. Amer. Statist. Assoc.* **97** 1042–1054. MR1951258

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

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

ZHU, Y. (2017). Supplement to "Nonasymptotic analysis of semiparametric regression models with high-dimensional parametric coefficients." DOI:10.1214/16-AOS1528SUPP.

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