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. 2—April 2017

The Annals of Statistics
2017, Vol. 45, No. 2, 461–499
DOI: 10.1214/15-AOS1433

MIMICKING COUNTERFACTUAL OUTCOMES TO ESTIMATE CAUSAL EFFECTS

BY JUDITH J. LOK[1]

Harvard School of Public Health

In observational studies, treatment may be adapted to covariates at several times without a fixed protocol, in continuous time. Treatment influences covariates, which influence treatment, which influences covariates and so on. Then even time-dependent Cox-models cannot be used to estimate the net treatment effect. Structural nested models have been applied in this setting. Structural nested models are based on counterfactuals: the outcome a person would have had had treatment been withheld after a certain time. Previous work on continuous-time structural nested models assumes that counterfactuals depend deterministically on observed data, while conjecturing that this assumption can be relaxed. This article proves that one can mimic counterfactuals by constructing random variables, solutions to a differential equation, that have the same distribution as the counterfactuals, even given past observed data. These "mimicking" variables can be used to estimate the parameters of structural nested models without assuming the treatment effect to be deterministic.

REFERENCES

[1] BAUER, H. (1972). *Probability Theory and Elements of Measure Theory*. Holt, Rinehart & Winston, New York.
[2] BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York.
[3] COX, D. R. and OAKES, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London. MR0751780
[4] DUISTERMAAT, J. J. and ECKHAUS, W. (1995). *Analyse van Gewone Differentiaalvergelijkingen*. *Epsilon Uitgaven*. Epsilon, Utrecht.
[5] GILL, R. D. and ROBINS, J. M. (2001). Causal inference for complex longitudinal data: The continuous case. *Ann. Statist.* **29** 1785–1811. MR1891746
[6] HERNÁN, M. A., BRUMBACK, B. and ROBINS, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* **11** 561–570.
[7] HOLLAND, P. W. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81** 945–970. MR0867618
[8] KEIDING, N. (1999). Event history analysis and inference from observational epidemiology. *Stat. Med.* **18** 2353–2363.
[9] KEIDING, N., FILIBERTI, M., ESBJERG, S., ROBINS, J. M. and JACOBSEN, N. (1999). The graft versus leukemia effect after bone marrow transplantation: A case study using structural nested failure time models. *Biometrics* **55** 23–28.

MSC2010 subject classifications. Primary 62P10; secondary 62M99, 62N02.
Key words and phrases. Causality in continuous time, dynamic treatments, longitudinal data, observational studies, panel data, rank preservation, stochastic differential equations, structural nested models.

[10] LOK, J. J (2017). Supplement to “Mimicking counterfactual outcomes to estimate causal effects.” DOI:10.1214/15-AOS1433SUPP.

[11] LOK, J. J. (2001). Statistical modelling of causal effects in time, Ph.D. thesis, Department of Mathematics, Free Univ. Amsterdam. http://www.math.vu.nl/research/theses/pdf/lok.pdf.

[12] LOK, J. J. (2007). Structural nested models and standard software: A mathematical foundation through partial likelihood. *Scand. J. Stat.* **34** 186–206. MR2325250

[13] LOK, J. J. (2008). Statistical modelling of causal effects in continuous time. *Ann. Statist.* **36** 1464–1507.

[14] LOK, J. J., GILL, R. D., VAN DER VAART, A. W. and ROBINS, J. M. (2004). Estimating the causal effect of a time-varying treatment on time-to-event using structural nested failure time models. *Stat. Neerl.* **58** 271–295.

[15] MARK, S. D. and ROBINS, J. M. (1993). Estimating the causal effect of smoking cessation in the presence of confounding factors using a rank preserving structural failure time model. *Stat. Med.* **12** 1605–1628.

[16] POLLARD, D. (2001). *A User's Guide to Measure Theoretic Probability*. Cambridge Univ. Press, Cambridge.

[17] ROBINS, J. M. (1987). A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Disease* **40** 139S–161S.

[18] ROBINS, J. M. (1989). The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In *Health Service Research Methodology: A Focus on AIDS* (L. Sechrest, H. Freeman and A. Bailey, eds.) 113–159. NCHSR, U.S. Public Health Service, Washington, D.C.

[19] ROBINS, J. M. (1992). Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika* **79** 321–334.

[20] ROBINS, J. M. (1995). Causal inference from complex longitudinal data. In *Design and Analysis of Follow-up Studies*. The Netherlands Institute for Health Sciences, Rotterdam. Lecture notes.

[21] ROBINS, J. M. (1997). Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality* (*Los Angeles*, *CA*, 1994). *Lect. Notes Stat.* **120** 69–117. Springer, New York. MR1601279

[22] ROBINS, J. M. (1998). Structural nested failure time models. In *Survival Analysis* (P. Armitage and T. Colton, eds.). *Encyclopedia of Biostatistics* **6** 4372–4389. Wiley, Chichester, UK. Section Eds: P. K. Andersen and N. Keiding.

[23] ROBINS, J. M. (1999). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology: The Environment and Clinical Trials* (M. E. Halloran and D. Berry, eds.) **116** 95–134. Springer, New York.

[24] ROBINS, J. M., BLEVINS, D., RITTER, G. and WULFSOHN, M. (1992). G-estimation of the effect of prophylaxis therapy for pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology* **3** 319–336.

[25] ROBINS, J. M. and GREENLAND, S. (1994). Adjusting for differential rates of PCP prophylaxis in high- versus low-dose AZT treatment arms in an AIDS randomized trial. *J. Amer. Statist. Assoc.* **89** 737–749.

[26] ROBINS, J. M., HERNÁN, M. A. and BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11** 550–560.

[27] TILLING, K., STERNE, J. A. and SZKLO, M. (2002). Estimating the effect of cardiovascular risk factors on all-cause mortality and incidence of coronary heart disease using G-estimation: The atherosclerosis risk in communities study. *Am. J. Epidemiol.* **155** 710–718.

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

[29] WILLIAMS, D. (1991). *Probability with Martingales*. Cambridge Univ. Press, Cambridge.
[30] WITTEMAN, J. C. M., D'AGOSTINO, R. B., STIJNEN, T., KANNEL, W. B., COBB, J. C., DE RIDDER, M. A. J., HOFMAN, A. and ROBINS, J. M. (1998). G-estimation of causal effects: Isolated systolic hypertension and cardiovascular death in the Framingham heart study. *American Journal of Epidemiology* **148** 390–401.

The Annals of Statistics
2017, Vol. 45, No. 2, 500–528
DOI: 10.1214/16-AOS1457

LIKELIHOOD-BASED MODEL SELECTION FOR STOCHASTIC BLOCK MODELS[1]

BY Y. X. RACHEL WANG AND PETER J. BICKEL

Stanford University and University of California, Berkeley

The stochastic block model (SBM) provides a popular framework for modeling community structures in networks. However, more attention has been devoted to problems concerning estimating the latent node labels and the model parameters than the issue of choosing the number of blocks. We consider an approach based on the log likelihood ratio statistic and analyze its asymptotic properties under model misspecification. We show the limiting distribution of the statistic in the case of underfitting is normal and obtain its convergence rate in the case of overfitting. These conclusions remain valid when the average degree grows at a polylog rate. The results enable us to derive the correct order of the penalty term for model complexity and arrive at a likelihood-based model selection criterion that is asymptotically consistent. Our analysis can also be extended to a degree-corrected block model (DCSBM). In practice, the likelihood function can be estimated using more computationally efficient variational methods or consistent label estimation algorithms, allowing the criterion to be applied to large networks.

REFERENCES

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

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

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

BALL, B., KARRER, B. and NEWMAN, M. E. J. (2011). An efficient and principled method for detecting communities in networks. *Phys. Rev. E* (3) **84** 036103.

BICKEL, P. and CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106** 21068–73.

BICKEL, P. J. and SARKAR, P. (2016). Hypothesis testing for automated community detection in networks. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 253–273. MR3453655

BICKEL, P., CHOI, D., CHANG, X. and ZHANG, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Ann. Statist.* **41** 1922–1943. MR3127853

CELISSE, A., DAUDIN, J.-J. and PIERRE, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electron. J. Stat.* **6** 1847–1899. MR2988467

MSC2010 subject classifications. 62F05.

Key words and phrases. Stochastic block models, model misspecification, network communities, likelihood ratio statistic.

CHEN, K. and LEI, J. (2014). Network cross-validation for determining the number of communities in network data. Preprint. Available at arXiv:1411.1715.

CHOI, D. S., WOLFE, P. J. and AIROLDI, E. M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika* **99** 273–284. MR2931253

DAUDIN, J.-J., PICARD, F. and ROBIN, S. (2008). A mixture model for random graphs. *Stat. Comput.* **18** 173–183. MR2390817

DECELLE, A., KRZAKALA, F., MOORE, C. and ZDEBOROVÁ, L. (2011). Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E* (3) **84** 066106.

FISHKIND, D. E., SUSSMAN, D. L., TANG, M., VOGELSTEIN, J. T. and PRIEBE, C. E. (2013). Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. *SIAM J. Matrix Anal. Appl.* **34** 23–39. MR3032990

HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Social Networks* **5** 109–137. MR0718088

JOSEPH, A. and YU, B. (2013). Impact of regularization on spectral clustering. Preprint. Available at arXiv:1312.1733.

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

LATOUCHE, P., BIRMELÉ, E. and AMBROISE, C. (2012). Variational Bayesian inference and complexity control for stochastic block models. *Stat. Model.* **12** 93–115. MR2953099

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

LEI, J. (2014). A goodness-of-fit test for stochastic block models. Preprint. Available at arXiv:1412.4857.

LESKOVEC, J. and MCAULEY, J. J. (2012). Learning to discover social circles in ego networks. In *Adv. Neural Inf. Process. Syst.* 539–547. Lake Tahoe.

NEWMAN, M. E. J. (2006a). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **103** 8577–8582.

NEWMAN, M. E. J. (2006b). Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* (3) **74** 036104, 19. MR2282139

PEIXOTO, T. P. (2013). Parsimonious module inference in large networks. *Phys. Rev. Lett.* **110** 148701.

ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915. MR2893856

SALDANA, D. F., YU, Y. and FENG, Y. (2014). How many communities are there? Preprint. Available at arXiv:1412.1684.

WANG, Y. X. R. and BICKEL, P. J. (2016). Supplement to "Likelihood-based model selection for stochastic block models." DOI:10.1214/16-AOS1457SUPP.

WOLFE, P. J. and OLHEDE, S. C. (2013). Nonparametric graphon estimation. Preprint. Available at arXiv:1309.5936.

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

ZHAO, Y., LEVINA, E. and ZHU, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Statist.* **40** 2266–2292. MR3059083

The Annals of Statistics
2017, Vol. 45, No. 2, 529–556
DOI: 10.1214/16-AOS1458

MULTIPLE TESTING OF LOCAL MAXIMA FOR DETECTION OF PEAKS IN RANDOM FIELDS[1]

BY DAN CHENG AND ARMIN SCHWARTZMAN

University of California, San Diego

A topological multiple testing scheme is presented for detecting peaks in images under stationary ergodic Gaussian noise, where tests are performed at local maxima of the smoothed observed signals. The procedure generalizes the one-dimensional scheme of Schwartzman, Gavrilov and Adler [*Ann. Statist.* **39** (2011) 3290–3319] to Euclidean domains of arbitrary dimension. Two methods are developed according to two different ways of computing p-values: (i) using the exact distribution of the height of local maxima, available explicitly when the noise field is isotropic [*Extremes* **18** (2015) 213–240; Expected number and height distribution of critical points of smooth isotropic Gaussian random fields (2015) Preprint]; (ii) using an approximation to the overshoot distribution of local maxima above a pre-threshold, applicable when the exact distribution is unknown, such as when the stationary noise field is nonisotropic [*Extremes* **18** (2015) 213–240]. The algorithms, combined with the Benjamini–Hochberg procedure for thresholding p-values, provide asymptotic strong control of the False Discovery Rate (FDR) and power consistency, with specific rates, as the search space and signal strength get large. The optimal smoothing bandwidth and optimal pre-threshold are obtained to achieve maximum power. Simulations show that FDR levels are maintained in nonasymptotic conditions. The methods are illustrated in the analysis of functional magnetic resonance images of the brain.

REFERENCES

[1] ADLER, R. J., SUBAG, E. and TAYLOR, J. E. (2012). Rotation and scale space random fields and the Gaussian kinematic formula. *Ann. Statist.* **40** 2910–2942. MR3097964

[2] ADLER, R. J. and TAYLOR, J. E. (2007). *Random Fields and Geometry*. Springer, New York. MR2319516

[3] ADLER, R. J., TAYLOR, J. E. and WORSLEY, K. J. (2010). Applications of random fields and geometry: Foundations and case studies. Preprint. Available at http://webee.technion.ac.il/people/adler/publications.html.

[4] AZAÏS, J.-M. and WSCHEBOR, M. (2009). *Level Sets and Extrema of Random Processes and Fields*. Wiley, Hoboken, NJ. MR2478201

[5] BENJAMINI, Y. and HELLER, R. (2007). False discovery rates for spatial signals. *J. Amer. Statist. Assoc.* **102** 1272–1281. MR2412549

[6] 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

MSC2010 subject classifications. Primary 62H35; secondary 62H15.

Key words and phrases. False discovery rate, Gaussian random field, kernel smoothing, image analysis, overshoot distribution, selective inference, topological inference.

[7] BRUTTI, P., GENOVESE, C. R., MILLER, C. J., NICHOL, R. C. and WASSERMAN, L. (2005). Spike hunting in galaxy spectra. Technical report, Libera Università Internazionale degli Studi Sociali Guido Carli di Roma. Available at http://www.stat.cmu.edu/tr/tr828/tr828.html.

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

[9] CHENG, D. and SCHWARTZMAN, A. (2016). Supplement to "Multiple testing of local maxima for detection of peaks in random fields." DOI:10.1214/16-AOS1458SUPP.

[10] CHENG, D. and SCHWARTZMAN, A. (2015). Distribution of the height of local maxima of Gaussian random fields. *Extremes* **18** 213–240. MR3351815

[11] CHENG, D. and SCHWARTZMAN, A. (2015). Expected number and height distribution of critical points of smooth isotropic Gaussian random fields. Preprint. Available at arXiv:1511.06835.

[12] CHENG, D. and SCHWARTZMAN, A. (2015). On the explicit height distribution and expected number of local maxima of isotropic Gaussian random fields. Preprint. Available at arXiv:1503.01328.

[13] CHUMBLEY, J. R. and FRISTON, K. J. (2009). False discovery rate revisited: FDR and topological inference using Gaussian random fields. *Neuroimage* **44** 62–70.

[14] CHUMBLEY, J. R., WORSLEY, K., FLANDIN, G. and FRISTON, K. J. (2010). Topological FDR for neuroimaging. *Neuroimage* **49** 3057–3064.

[15] CRAMÉR, H. and LEADBETTER, M. R. (1967). *Stationary and Related Stochastic Processes. Sample Function Properties and Their Applications*. Wiley, New York. MR0217860

[16] EGNER, A., GEISLER, C., VON MIDDENDORFF, C., BOCK, H., WENZEL, D., MEDDA, R., ANDRESEN, M., STIEL, A. C., JAKOBS, S., EGGELING, C., SCHÖNLE, A. and HELL, S. W. (2007). Fluorescence nanoscopy in whole cells by asynchronous localization of photoswitching emitters. *Biophys. J.* **93** 3285–3290.

[17] FITHIAN, W., SUN, D. and TAYLOR, J. (2015). Optimal inference after model selection. Preprint. Available at arXiv:1410.2597.

[18] FYODOROV, Y. V. (2004). Complexity of random energy landscapes, glass transition, and absolute value of the spectral determinant of random matrices. *Phys. Rev. Lett.* **92** 240601, 4. MR2115095

[19] GEISLER, C., SCHÖNLE, A., VON MIDDENDORFF, C., BOCK, H., EGGELING, C., EGNER, A. and HELL, S. W. (2007). Resolution of $\lambda/10$ in fluorescence microscopy using fast single molecule photo-switching. *Applied Physics A* **88** 223–226.

[20] GENOVESE, C. R., LAZAR, N. A. and NICHOLS, T. E. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage* **15** 870–878.

[21] HELLER, R., STANLEY, D., YEKUTIELI, D., RUBIN, N. and BENJAMINI, Y. (2006). Cluster-based analysis of FMRI data. *Neuroimage* **33** 599–608.

[22] LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2015). Exact post-selection inference, with application to the lasso. Preprint. Available at arXiv:1311.6238.

[23] MORAN, J. M., JOLLY, E. and MITCHELL, J. P. (2012). Social-cognitive deficits in normal aging. *J. Neurosci.* **32** 5553–5561.

[24] NICHOLS, T. and HAYASAKA, S. (2003). Controlling the familywise error rate in functional neuroimaging: A comparative review. *Stat. Methods Med. Res.* **12** 419–446. MR2005445

[25] PERONE PACIFICO, M., GENOVESE, C., VERDINELLI, I. and WASSERMAN, L. (2004). False discovery control for random fields. *J. Amer. Statist. Assoc.* **99** 1002–1014. MR2109490

[26] PERONE PACIFICO, M., GENOVESE, C., VERDINELLI, I. and WASSERMAN, L. (2007). Scan clustering: A false discovery approach. *J. Multivariate Anal.* **98** 1441–1469. MR2364129

[27] PITERBARG, V. I. (1996). *Asymptotic Methods in the Theory of Gaussian Processes and Fields. Translations of Mathematical Monographs* **148**. Amer. Math. Soc., Providence, RI. MR1361884

[28] POLINE, J.-B., WORSLEY, K. J., EVANS, A. C. and FRISTON, K. J. (1997). Combining spatial extent and peak intensity to test for activations in functional imaging. *Neuroimage* **5** 83–96.

[29] PRATT, W. K. (1991). *Digital Image Processing*. Wiley, New York.

[30] SCHWARTZMAN, A., GAVRILOV, Y. and ADLER, R. J. (2011). Multiple testing of local maxima for detection of peaks in 1D. *Ann*. *Statist*. **39** 3290–3319. MR3012409

[31] SIEGMUND, D. O., ZHANG, N. R. and YAKIR, B. (2011). False discovery rate for scanning statistics. *Biometrika* **98** 979–985. MR2860337

[32] SIMON, M. (1995). *Digital Communication Techniques*: *Signal Design and Detection*. Prentice Hall, Englewood Cliffs, NJ.

[33] SMITH, S. M. and NICHOLS, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage* **44** 83–98.

[34] SUBAG, E. (2015). The complexity of spherical p-spin models—a second moment approach. Preprint. Available at arXiv:1504.02251.

[35] SUN, W., REICH, B. J., CAI, T. T., GUINDANI, M. and SCHWARTZMAN, A. (2015). False discovery control in large-scale spatial multiple testing. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **77** 59–83. MR3299399

[36] TAYLOR, J. E., LOFTUS, J. R. and TIBSHIRANI, R. J. (2016). Inference in adaptive regression via the Kac–Rice formula. *Ann*. *Statist*. **44** 743–770. MR3476616

[37] TAYLOR, J. E. and WORSLEY, K. J. (2007). Detecting sparse signals in random fields, with an application to brain mapping. *J*. *Amer*. *Statist*. *Assoc*. **102** 913–928. MR2354405

[38] WORSLEY, K. J., MARRETT, S., NEELIN, P. and EVANS, A. C. (1996). Searching scale space for activation in PET images. *Hum*. *Brain Mapp*. **4** 74–90.

[39] WORSLEY, K. J., MARRETT, S., NEELIN, P., VANDAL, A. C., FRISTON, K. J. and EVANS, A. C. (1996). A unified statistical approach for determining significant signals in images of cerebral activation. *Hum*. *Brain Mapp*. **4** 58–73.

[40] WORSLEY, K. J., TAYLOR, J. E., TOMAIUOLO, F. and LERCH, J. (2004). Unified univariate and multivariate random field theory. *Neuroimage* **23** S189–195.

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

[42] ZHANG, H., NICHOLS, T. E. and JOHNSON, T. D. (2009). Cluster mass inference via random field theory. *Neuroimage* **44** 51–61.

The Annals of Statistics
2017, Vol. 45, No. 2, 557–590
DOI: 10.1214/16-AOS1459

A RATE OPTIMAL PROCEDURE FOR RECOVERING SPARSE DIFFERENCES BETWEEN HIGH-DIMENSIONAL MEANS UNDER DEPENDENCE

BY JUN LI AND PING-SHOU ZHONG

Kent State University and Michigan State University

The paper considers the problem of recovering the sparse different components between two high-dimensional means of column-wise dependent random vectors. We show that dependence can be utilized to lower the identification boundary for signal recovery. Moreover, an optimal convergence rate for the marginal false nondiscovery rate (mFNR) is established under dependence. The convergence rate is faster than the optimal rate without dependence. To recover the sparse signal bearing dimensions, we propose a Dependence-Assisted Thresholding and Excising (DATE) procedure, which is shown to be rate optimal for the mFNR with the marginal false discovery rate (mFDR) controlled at a pre-specified level. Extensions of the DATE to recover the differences in contrasts among multiple population means and differences between two covariance matrices are also provided. Simulation studies and case study are given to demonstrate the performance of the proposed signal identification procedure.

REFERENCES

ALLEN, G. I. and TIBSHIRANI, R. (2012). Inference with transposable data: Modelling the effects of row and column correlations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 721–743. MR2965957

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

BENJAMINI, Y. and YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29** 1165–1188. MR1869245

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

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

CAI, T., LIU, W. and XIA, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *J. Amer. Statist. Assoc.* **108** 265–277. MR3174618

CAI, T. T., LIU, W. and XIA, Y. (2014). Two-sample test of high dimensional means under dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 349–372. MR3164870

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

EFRON, B. (2007). Correlation and large-scale simultaneous significance testing. *J. Amer. Statist. Assoc.* **102** 93–103. MR2293302

MSC2010 subject classifications. Primary 62H15; secondary 62G20.

Key words and phrases. False discovery rate, high dimensional data, multiple testing, sparse signals, thresholding.

FAN, J., HAN, X. and GU, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *J. Amer. Statist. Assoc.* **107** 1019–1035. MR3010887

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

GENOVESE, C. and WASSERMAN, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 499–517. MR1924303

GENOVESE, C. R., JIN, J., WASSERMAN, L. and YAO, Z. (2012). A comparison of the lasso and marginal regression. *J. Mach. Learn. Res.* **13** 2107–2143. MR2956354

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

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

JI, P. and ZHAO, Z. (2014). Rate optimal multiple testing procedure in high-dimensional regression. Technical report.

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

JIN, J., ZHANG, C.-H. and ZHANG, Q. (2014). Optimality of graphlet screening in high dimensional variable selection. *J. Mach. Learn. Res.* **15** 2723–2772. MR3270749

KE, Z. T., JIN, J. and FAN, J. (2014). Covariate assisted screening and estimation. *Ann. Statist.* **42** 2202–2242. MR3269978

LI, J. and CHEN, S. X. (2012). Two sample tests for high-dimensional covariance matrices. *Ann. Statist.* **40** 908–940. MR2985938

LI, J. and ZHONG, P. (2016). Supplement to "A rate optimal procedure for recovering sparse differences between high-dimensional means under dependence." DOI:10.1214/16-AOS1459SUPP.

MEINSHAUSEN, N. and BÜHLMANN, P. (2005). Lower bounds for the number of false null hypotheses for multiple testing of associations under general dependence structures. *Biometrika* **92** 893–907. MR2234193

QIU, X., KLEBANOV, L. and YAKOVLEV, A. (2005). Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes. *Stat. Appl. Genet. Mol. Biol.* **4** Art. 34. MR2183944

RICHARDSON, A., WANG, Z., NICOLO, A., LU, X., BROWN, M., MIRON, A., LIAO, X., IGLEHART, J., LIVINGSTON, D. and GANESAN, S. (2006). X chromosomal abnormalities in basal-like human breast cancer. *Cancer Cell* **9** 121–132.

SCHOTT, J. R. (2007). A test for the equality of covariance matrices when the dimension is large relative to the sample sizes. *Comput. Statist. Data Anal.* **51** 6535–6542. MR2408613

SCHWEDER, T. and SPJØTVOLL, E. (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika* **69** 493–502.

SRIVASTAVA, M. S. and YANAGIHARA, H. (2010). Testing the equality of several covariance matrices with fewer observations than the dimension. *J. Multivariate Anal.* **101** 1319–1329. MR2609494

STOREY, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 479–498. MR1924302

SUN, W. and CAI, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.* **102** 901–912. MR2411657

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

XIE, J., CAI, T. T. and LI, H. (2011). Sample size and power analysis for sparse signal recovery in genome-wide association studies. *Biometrika* **98** 273–290. MR2806428

ZHAO, T., LIU, H., ROEDER, K., LAFFERTY, J. and WASSERMAN, L. (2012). The `huge` package for high-dimensional undirected graph estimation in R. *J. Mach. Learn. Res.* **13** 1059–1062. MR2930633

The Annals of Statistics
2017, Vol. 45, No. 2, 591–614
DOI: 10.1214/16-AOS1460

ONLINE ESTIMATION OF THE GEOMETRIC MEDIAN IN HILBERT SPACES: NONASYMPTOTIC CONFIDENCE BALLS

BY HERVÉ CARDOT, PEGGY CÉNAC AND ANTOINE GODICHON-BAGGIONI

Université de Bourgogne Franche Comté

Estimation procedures based on recursive algorithms are interesting and powerful techniques that are able to deal rapidly with very large samples of high dimensional data. The collected data may be contaminated by noise so that robust location indicators, such as the geometric median, may be preferred to the mean. In this context, an estimator of the geometric median based on a fast and efficient averaged nonlinear stochastic gradient algorithm has been developed by [*Bernoulli* **19** (2013) 18–43]. This work aims at studying more precisely the nonasymptotic behavior of this nonlinear algorithm by giving nonasymptotic confidence balls in general separable Hilbert spaces. This new result is based on the derivation of improved L^2 rates of convergence as well as an exponential inequality for the nearly martingale terms of the recursive nonlinear Robbins–Monro algorithm.

REFERENCES

[1] ARNAUDON, M., DOMBRY, C., PHAN, A. and YANG, L. (2012). Stochastic algorithms for computing means of probability measures. *Stochastic Process. Appl.* **122** 1437–1455. MR2914758

[2] BACH, F. (2014). Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *J. Mach. Learn. Res.* **15** 595–627. MR3190851

[3] BACH, F. and MOULINES, E. (2013). Non-strongly-convex smooth stochastic approximation with convergence rate $o(1/n)$. In *Advances in Neural Information Processing Systems* 773–781. Curran Associates, Inc., Red Hook, NY.

[4] BALI, J. L., BOENTE, G., TYLER, D. E. and WANG, J.-L. (2011). Robust functional principal components: A projection-pursuit approach. *Ann. Statist.* **39** 2852–2882. MR3012394

[5] BALSUBRAMANI, A., DASGUPTA, S. and FREUND, Y. (2013). The fast convergence of incremental PCA. In *Advances in Neural Information Processing Systems* 3174–3182. Curran Associates, Inc., Red Hook, NY.

[6] BECK, A. and SABACH, S. (2015). Weiszfeld's method: Old and new results. *J. Optim. Theory Appl.* **164** 1–40. MR3296283

[7] CARDOT, H., CÉNAC, P. and GODICHON-BAGGIONI, A. (2016). Supplement to "Online estimation of the geometric median in Hilbert spaces: Nonasymptotic confidence balls." DOI:10.1214/16-AOS1460SUPP.

[8] CARDOT, H., CÉNAC, P. and ZITT, P.-A. (2013). Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli* **19** 18–43. MR3019484

MSC2010 subject classifications. Primary 62G05; secondary 62L20.

Key words and phrases. Functional data analysis, martingales in Hilbert spaces, recursive estimation, robust statistics, spatial median, stochastic gradient algorithms.

[9] CHAKRABORTY, A. and CHAUDHURI, P. (2014). The spatial distribution in infinite dimensional spaces and related quantiles and depths. *Ann. Statist.* **42** 1203–1231. MR3224286

[10] CHAUDHURI, P. (1992). Multivariate location estimation using extension of R-estimates through U-statistics type approach. *Ann. Statist.* **20** 897–916. MR1165598

[11] FLETCHER, P. T., VENKATASUBRAMANIAN, S. and JOSHI, S. (2009). The geometric median on Riemannian manifolds with application to robust atlas estimation. *NeuroImage* **45** S143–S152.

[12] GERVINI, D. (2008). Robust functional estimation using the median and spherical principal components. *Biometrika* **95** 587–600. MR2443177

[13] HALDANE, J. B. S. (1948). Note on the median of a multivariate distribution. *Biometrika* **35** 414–417.

[14] KEMPERMAN, J. H. B. (1987). The median of a finite measure on a Banach space. In *Statistical Data Analysis Based on the* L_1*-Norm and Related Methods* (*Neuchâtel*, 1987) 217–230. North-Holland, Amsterdam. MR0949228

[15] KRAUS, D. and PANARETOS, V. M. (2012). Dispersion operators and resistant second-order functional data analysis. *Biometrika* **99** 813–832.

[16] KUHN, H. W. (1973). A note on Fermat's problem. *Math. Program.* **4** 98–107. MR0316102

[17] LOCANTORE, N., MARRON, J. S., SIMPSON, D. G., TRIPOLI, N., ZHANG, J. T. and COHEN, K. L. (1999). Robust principal component analysis for functional data. *TEST* **8** 1–73. MR1707596

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

[19] MINSKER, S., SRIVASTAVA, S., LIN, L. and DUNSON, D. (2014). Scalable and robust Bayesian inference via the median posterior. In *Proceedings of the* 31*st International Conference on Machine Learning* (*ICML*-14) 1656–1664.

[20] MÖTTÖNEN, J., NORDHAUSEN, K. and OJA, H. (2010). Asymptotic theory of the spatial median. In *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis*: *A Festschrift in Honor of Professor Jana Jurečková*. *Inst. Math. Stat. Collect.* **7** 182–193. IMS, Beachwood, OH. MR2808378

[21] PELLETIER, M. (2000). Asymptotic almost sure efficiency of averaged stochastic algorithms. *SIAM J. Control Optim.* **39** 49–72 (electronic). MR1780908

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

[23] POLYAK, B. T. and JUDITSKY, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.* **30** 838–855. MR1167814

[24] RUDELSON, M. and VERSHYNIN, R. (2010). Non-asymptotic theory of random matrices: Extreme singular values. In *Proceedings of the International Congress of Mathematicians. Volume III* 1576–1602. Hindustan Book Agency, New Delhi. MR2827856

[25] SMALL, C. G. (1990). A survey of multidimensional medians. *Int. Stat. Rev.* **58** 263–277.

[26] TARRÈS, P. and YAO, Y. (2014). Online learning as stochastic approximation of regularization paths: Optimality and almost-sure convergence. *IEEE Trans. Inform. Theory* **60** 5716–5735. MR3252416

[27] VARDI, Y. and ZHANG, C.-H. (2000). The multivariate L_1-median and associated data depth. *Proc. Natl. Acad. Sci. USA* **97** 1423–1426 (electronic). MR1740461

[28] WEISZFELD, E. (1937). Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tohoku Math. J.* (2) **43** 355–386.

[29] WOODROOFE, M. (1972). Normal approximation and large deviations for the Robbins–Monro process. *Z. Wahrsch. Verw. Gebiete* **21** 329–338. MR0314098

The Annals of Statistics
2017, Vol. 45, No. 2, 615–646
DOI: 10.1214/16-AOS1461

CONFIDENCE INTERVALS FOR HIGH-DIMENSIONAL LINEAR REGRESSION: MINIMAX RATES AND ADAPTIVITY[1]

BY T. TONY CAI AND ZIJIAN GUO

University of Pennsylvania

Confidence sets play a fundamental role in statistical inference. In this paper, we consider confidence intervals for high-dimensional linear regression with random design. We first establish the convergence rates of the minimax expected length for confidence intervals in the oracle setting where the sparsity parameter is given. The focus is then on the problem of adaptation to sparsity for the construction of confidence intervals. Ideally, an adaptive confidence interval should have its length automatically adjusted to the sparsity of the unknown regression vector, while maintaining a pre-specified coverage probability. It is shown that such a goal is in general not attainable, except when the sparsity parameter is restricted to a small region over which the confidence intervals have the optimal length of the usual parametric rate. It is further demonstrated that the lack of adaptivity is not due to the conservativeness of the minimax framework, but is fundamentally caused by the difficulty of learning the bias accurately.

REFERENCES

[1] BARAUD, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli* **8** 577–606. MR1935648

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

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

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

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

[6] CAI, T. T. and GUO, Z. (2015). Supplement to "Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity." DOI:10.1214/16-AOS1461SUPP.

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

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

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

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

MSC2010 subject classifications. Primary 62G15; secondary 62C20, 62H35.
Key words and phrases. Adaptivity, confidence interval, coverage probability, expected length, high-dimensional linear regression, minimaxity, sparsity.

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

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

[13] COLLIER, O., COMMINGES, L. and TSYBAKOV, A. B. (2015). Minimax estimation of linear and quadratic functionals on sparsity classes. Preprint. Available at arXiv:1502.00665.

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

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

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

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

[18] JAVANMARD, A. and MONTANARI, A. (2015). De-biasing the Lasso: Optimal sample size for Gaussian designs. Preprint. Available at arXiv:1508.02757.

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

[20] RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *J. Mach. Learn. Res.* **11** 2241–2259. MR2719855

[21] REN, Z., SUN, T., ZHANG, C.-H. and ZHOU, H. H. (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *Ann. Statist.* **43** 991–1026. MR3346695

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

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

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

[25] VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285

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

[27] 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

The Annals of Statistics
2017, Vol. 45, No. 2, 647–674
DOI: 10.1214/16-AOS1462

ESTIMATING THE EFFECT OF JOINT INTERVENTIONS FROM OBSERVATIONAL DATA IN SPARSE HIGH-DIMENSIONAL SETTINGS

BY PREETAM NANDY*,[1], MARLOES H. MAATHUIS*,[1] AND THOMAS S. RICHARDSON†,[2]

ETH Zürich and University of Washington†*

We consider the estimation of joint causal effects from observational data. In particular, we propose new methods to estimate the effect of multiple simultaneous interventions (e.g., multiple gene knockouts), under the assumption that the observational data come from an unknown linear structural equation model with independent errors. We derive asymptotic variances of our estimators when the underlying causal structure is partly known, as well as high-dimensional consistency when the causal structure is fully unknown and the joint distribution is multivariate Gaussian. We also propose a generalization of our methodology to the class of nonparanormal distributions. We evaluate the estimators in simulation studies and also illustrate them on data from the DREAM4 challenge.

REFERENCES

[1] ALIFERIS, C. F., STATNIKOV, A., TSAMARDINOS, I., MANI, S. and KOUTSOUKOS, X. D. (2010). Local causal and Markov blanket induction for causal discovery and feature selection for classification. Part I: Algorithms and empirical evaluation. *J. Mach. Learn. Res.* **11** 171–234. MR2591625

[2] BALKE, A. A. and PEARL, J. (1994). Probabilistic evaluation of counterfactual queries. In *AAAI* 1994. Seattle, WA.

[3] CASTELO, R. and ROVERATO, A. (2006). A robust procedure for Gaussian graphical model search from microarray data with p larger than n. *J. Mach. Learn. Res.* **7** 2621–2650. MR2274453

[4] CLAASSEN, T., MOOIJ, J. and HESKES, T. (2013). Learning sparse causal models is not NP-hard. In *Proc. UAI* 2013, 172–181. AUAI Press, Corvallis, OR.

[5] COCHRAN, W. G. (1938). The omission or addition of an independent variable in multiple linear regression. *J. R. Statist. Soc. Suppl.* **5** 171–176.

[6] COLOMBO, D. and MAATHUIS, M. H. (2014). Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.* **15** 3741–3782. MR3291411

[7] COLOMBO, D., MAATHUIS, M. H., KALISCH, M. and RICHARDSON, T. S. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann. Statist.* **40** 294–321. MR3014308

[8] COX, D. R. and WERMUTH, N. (2003). A general condition for avoiding effect reversal after marginalization. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **65** 937–941. MR2017879

MSC2010 subject classifications. 62M99, 62H12, 62P10.

Key words and phrases. Causal inference, directed acyclic graph (DAG), linear structural equation model (linear SEM), multiple simultaneous interventions, joint causal effects, nonparanormal distribution, high-dimensional data.

[9] COX, D. R. and WERMUTH, N. (2004). Causality: A statistical view. *Int. Stat. Rev.* **72** 285–305.

[10] DOR, D. and TARSI, M. (1992). A simple algorithm to construct a consistent extension of a partially oriented graph. Technical Report No. R-185, Cognitive Systems Laboratory, Univ. California, Los Angeles.

[11] DRTON, M., FOX, C. and KÄUFL, A. (2012). Comments on: Sequences of regressions and their independencies [MR2935353]. *TEST* **21** 255–261. MR2935355

[12] HARRIS, N. and DRTON, M. (2013). PC algorithm for nonparanormal graphical models. *J. Mach. Learn. Res.* **14** 3365–3383. MR3144465

[13] JASNOS, L. and KORONA, R. (2007). Epistatic buffering of fitness loss in yeast double deletion strains. *Nat. Genet.* **39** 550–554.

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

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

[16] LIU, H., HAN, F., YUAN, M., LAFFERTY, J. and WASSERMAN, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.* **40** 2293–2326. MR3059084

[17] LIU, H., LAFFERTY, J. and WASSERMAN, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.* **10** 2295–2328. MR2563983

[18] MAATHUIS, M. H. and COLOMBO, D. (2015). A generalized back-door criterion. *Ann. Statist.* **43** 1060–1088. MR3346697

[19] MAATHUIS, M. H., COLOMBO, D., KALISCH, M. and BÜHLMANN, P. (2010). Predicting causal effects in large-scale systems from observational data. *Nat. Methods* **7** 247–248.

[20] MAATHUIS, M. H., KALISCH, M. and BÜHLMANN, P. (2009). Estimating high-dimensional intervention effects from observational data. *Ann. Statist.* **37** 3133–3164. MR2549555

[21] MARBACH, D., PRILL, R. J., SCHAFFTER, T., MATTIUSSI, C., FLOREANO, D. and STOLOVITZKY, G. (2010). Revealing strengths and weaknesses of methods for gene network inference. *PNAS* **107** 6286–6291.

[22] MEEK, C. (1995). Causal inference and causal explanation with background knowledge. In *Proc. UAI* 1995, 403–410. Morgan Kaufmann, San Francisco, CA.

[23] MOOIJ, J. M. and HESKES, T. (2013). Cyclic causal discovery from continuous equilibrium data. In *Proc. UAI* 2013, 431–439. AUAI Press, Corvallis, OR.

[24] NANDY, P., MAATHUIS, M. H. and RICHARDSON, T. S. (2016). Supplement to "Estimating the effect of joint interventions from observational data in sparse high-dimensional settings." DOI:10.1214/16-AOS1462SUPP.

[25] PEARL, J. (2009). Causal inference in statistics: An overview. *Stat. Surv.* **3** 96–146. MR2545291

[26] PERKOVIC, E., TEXTOR, J., KALISCH, M. and MAATHUIS, M. H. (2015). A complete adjustment criterion. In *Proc. UAI* 2015, 682–691. AUAI Press, Corvallis, OR.

[27] POURAHMADI, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika* **86** 677–690. MR1723786

[28] RAMSEY, J. (2006). A PC-style Markov blanket search for high dimensional datasets. Technical Report No. 177, Dept. Philosophy, Carnegie Mellon Univ., Pittsburgh, PA.

[29] ROBINS, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math. Modelling* **7** 1393–1512. MR0877758

[30] ROBINS, J. M., HERNAN, M. A. and BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11** 550–560.

[31] SCHAFFTER, T., MARBACH, D. and FLOREANO, D. (2011). GeneNetWeaver: In silico benchmark generation and performance profiling of network inference methods. *Bioinform*. **27** 2263–2270.

[32] SHPITSER, I., VANDERWEELE, T. J. and ROBINS, J. M. (2010). On the validity of covariate adjustment for estimating causal effects. In *Proc*. *UAI* 2010, 527–536. AUAI Press, Corvallis, OR.

[33] SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation*, *Prediction*, *and Search*, 2nd ed. MIT Press, Cambridge, MA. MR1815675

[34] STEKHOVEN, D. J., MORAES, I., SVEINBJÖRNSSON, G., HENNING, L., MAATHUIS, M. H. and BÜHLMANN, P. (2012). Causal stability ranking. *Bioinform*. **28** 2819–2823.

[35] TERAMOTO, R., SAITO, C. and FUNAHASHI, S. (2014). Estimating causal effects with a nonparanormal method for the design of efficient intervention experiments. *BMC Bioinformatics* **15** 1–14.

[36] TIAN, J. and PEARL, J. (2001). Causal discovery from changes. In *Proc*. *UAI* 2001, 512–521. Morgan Kaufmann, San Francisco, CA.

[37] VAN DER ZANDER, B., LISKIEWICZ, M. and TEXTOR, J. (2014). Constructing separators and adjustment sets in ancestral graphs. In *Proc*. *UAI* 2014, 907–916. AUAI Press, Corvallis, OR.

[38] VELENICH, A. and GORE, J. (2013). The strength of genetic interactions scales weakly with mutational effects. *Genome Biol*. **14** R76.

[39] WRIGHT, S. (1921). Correlation and causation. *J*. *Agric*. *Res*. **20** 557–585.

The Annals of Statistics
2017, Vol. 45, No. 2, 675–707
DOI: 10.1214/16-AOS1464

IDENTIFIABILITY OF RESTRICTED LATENT CLASS MODELS WITH BINARY RESPONSES

BY GONGJUN XU

University of Minnesota

Statistical latent class models are widely used in social and psychological researches, yet it is often difficult to establish the identifiability of the model parameters. In this paper, we consider the identifiability issue of a family of restricted latent class models, where the restriction structures are needed to reflect pre-specified assumptions on the related assessment. We establish the identifiability results in the strict sense and specify which types of restriction structure would give the identifiability of the model parameters. The results not only guarantee the validity of many of the popularly used models, but also provide a guideline for the related experimental design, where in the current applications the design is usually experience based and identifiability is not guaranteed. Theoretically, we develop a new technique to establish the identifiability result, which may be extended to other restricted latent class models.

REFERENCES

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

CHEN, Y., LIU, J., XU, G. and YING, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *J. Amer. Statist. Assoc.* **110** 850–866. MR3367269

CHIU, C.-Y., DOUGLAS, J. A. and LI, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika* **74** 633–665. MR2565331

DECARLO, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, class sizes, and the Q-matrix. *Appl. Psychol. Meas.* **35** 8–26.

DE LA TORRE, J. (2011). The generalized DINA model framework. *Psychometrika* **76** 179–199. MR2788881

DE LA TORRE, J. and DOUGLAS, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika* **69** 333–353. MR2272454

DIBELLO, L. V., STOUT, W. F. and ROUSSOS, L. A. (1995). Unified cognitive psychometric diagnostic assessment likelihood-based classification techniques. In *Cognitively Diagnostic Assessment* (P. D. Nichols, S. F. Chipman and R. L. Brennan, eds.) 361–390. Erlbaum Associates, Hillsdale, NJ.

ELMORE, R., HALL, P. and NEEMAN, A. (2005). An application of classical invariant theory to identifiability in nonparametric mixtures. *Ann. Inst. Fourier* (*Grenoble*) **55** 1–28. MR2141286

GABRIELSEN, A. (1978). Consistency and identifiability. *J. Econometrics* **8** 261–263. MR0547145

GOODMAN, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61** 215–231. MR0370936

MSC2010 subject classifications. 62E10.

Key words and phrases. Identifiability, restricted latent class models, Q-matrix, cognitive diagnosis models, multivariate Bernoulli mixture, Kruskal's tensor decomposition.

GYLLENBERG, M., KOSKI, T., REILINK, E. and VERLAAN, M. (1994). Nonuniqueness in probabilistic numerical identification of bacteria. *J. Appl. Probab.* **31** 542–548. MR1274807

HAGENAARS, J. A. (1993). *Loglinear Models with Latent Variables* **94**. Sage, Thousand Oaks, CA.

HARTZ, S. M. (2002). A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality. Ph.D. thesis, Univ. Illinois, Urbana-Champaign.

HENSON, R. A., TEMPLIN, J. L. and WILLSE, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika* **74** 191–210. MR2507377

JUNKER, B. W. and SIJTSMA, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Meas.* **25** 258–272. MR1842982

KOOPMANS, T. C., ed. (1950). *Statistical Inference in Dynamic Economic Models*. *Cowles Commission Monograph No*. 10. Wiley, New York. MR0038640

KOOPMANS, T. C. and REIERSØL, O. (1950). The identification of structural characteristics. *Ann. Math. Stat.* **21** 165–181. MR0039967

KRUSKAL, J. B. (1976). More factors than subjects, tests and treatments: An indeterminacy theorem for canonical decomposition and individual differences scaling. *Psychometrika* **41** 281–293. MR0488592

KRUSKAL, J. B. (1977). Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra Appl.* **18** 95–138. MR0444690

LEIGHTON, J. P., GIERL, M. J. and HUNKA, S. M. (2004). The attribute hierarchy model for cognitive assessment: A variation on Tatsuoka's rule-space approach. *J. Educ. Meas.* **41** 205–237.

MARIS, E. (1999). Estimating multiple classification latent class models. *Psychometrika* **64** 187–212. MR1700708

MARIS, G. and BECHGER, T. M. (2009). Equivalent diagnostic classification models. *Measurement* **7** 41–46.

MCHUGH, R. B. (1956). Efficient estimation and local identification in latent class analysis. *Psychometrika* **21** 331–347. MR0082427

ROTHENBERG, T. J. (1971). Identification in parametric models. *Econometrica* **39** 577–591. MR0436944

RUPP, A. A., TEMPLIN, J. L. and HENSON, R. A. (2010). *Diagnostic Measurement*: *Theory*, *Methods*, *and Applications*. Guilford Press, New York.

TATSUOKA, C. (2009). Diagnostic models as partially ordered sets. *Measurement* **7** 49–53.

TATSUOKA, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *J. Educ. Meas.* **20** 345–354.

TATSUOKA, K. K. (2009). *Cognitive Assessment*: *An Introduction to the Rule Space Method*. Routledge, New York.

TEICHER, H. (1967). Identifiability of mixtures of product measures. *Ann. Math. Stat.* **38** 1300–1302. MR0216635

TEMPLIN, J. L. and HENSON, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychol. Methods* **11** 287–305.

VON DAVIER, M. (2008). A general diagnostic model applied to language testing data. *Br. J. Math. Stat. Psychol.* **61** 287–307. MR2649038

XU, G. (2013). Statistical inference for diagnostic classification models. Ph.D. thesis, Columbia Univ., New York, NY.

XU, G. and ZHANG, S. (2015). Identifiability of diagnostic classification models. *Psychometrika*. To appear. DOI:10.1007/s11336-015-9471-z.

The Annals of Statistics
2017, Vol. 45, No. 2, 708–743
DOI: 10.1214/16-AOS1465

A BERNSTEIN-TYPE INEQUALITY FOR SOME MIXING PROCESSES AND DYNAMICAL SYSTEMS WITH AN APPLICATION TO LEARNING

BY HANYUAN HANG AND INGO STEINWART

University of Stuttgart

We establish a Bernstein-type inequality for a class of stochastic processes that includes the classical geometrically ϕ-mixing processes, Rio's generalization of these processes and many time-discrete dynamical systems. Modulo a logarithmic factor and some constants, our Bernstein-type inequality coincides with the classical Bernstein inequality for i.i.d. data. We further use this new Bernstein-type inequality to derive an oracle inequality for generic regularized empirical risk minimization algorithms and data generated by such processes. Applying this oracle inequality to support vector machines using the Gaussian kernels for binary classification, we obtain essentially the same rate as for i.i.d. processes, and for least squares and quantile regression; it turns out that the resulting learning rates match, up to some arbitrarily small extra term in the exponent, the optimal rates for i.i.d. processes.

REFERENCES

[1] ADAMCZAK, R. (2008). A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.* **13** 1000–1034. MR2424985

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

[3] ALQUIER, P., LI, X. and WINTENBERGER, O. (2013). Prediction of time series by statistical learning: general losses and fast rates. *Dependence Modeling* **1** 65–93.

[4] AMBROSIO, L., FUSCO, N. and PALLARA, D. (2000). *Functions of Bounded Variation and Free Discontinuity Problems*. Oxford Univ. Press, New York. MR1857292

[5] ARAÚJO, V., GALATOLO, S. and PACIFICO, M. J. (2014). Decay of correlations for maps with uniformly contracting fibers and logarithm law for singular hyperbolic attractors. *Math. Z.* **276** 1001–1048. MR3175169

[6] BALADI, V. (2000). *Positive Transfer Operators and Decay of Correlations*. *Advanced Series in Nonlinear Dynamics* **16**. World Scientific, River Edge, NJ. MR1793194

[7] BALADI, V. (2001). Decay of correlations. In *Smooth Ergodic Theory and Its Applications* (*Seattle*, *WA*, 1999). *Proc. Sympos. Pure Math.* **69** 297–325. Amer. Math. Soc., Providence, RI. MR1858537

[8] BARTLETT, P. L., JORDAN, M. I. and MCAULIFFE, J. D. (2006). Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.* **101** 138–156. MR2268032

[9] BELOMESTNY, D. (2011). Spectral estimation of the Lévy density in partially observed affine models. *Stochastic Process. Appl.* **121** 1217–1244. MR2794974

MSC2010 subject classifications. Primary 60E15; secondary 60G10, 37D20, 60F10, 68T05, 62G08, 62M10.

Key words and phrases. Bernstein-type inequalities, mixing processes, dynamical systems, nonparametric classification and regression, support vector machines (SVMs).

[10] BENEDICKS, M. and YOUNG, L.-S. (2000). Markov extensions and decay of correlations for certain Hénon maps. *Astérisque* **261** 13–56. MR1755436
[11] BLANCHARD, G., LUGOSI, G. and VAYATIS, N. (2004). On the rate of convergence of regularized boosting classifiers. *J. Mach. Learn. Res.* **4** 861–894. MR2076000
[12] BOSQ, D. (1993). Bernstein-type large deviations inequalities for partial sums of strong mixing processes. *Statistics* **24** 59–70. MR1238263
[13] BOWEN, R. (1975). *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms*. *Lecture Notes in Mathematics* **470**. Springer, Berlin. MR0442989
[14] BRADLEY, R. C. (2007). *Introduction to Strong Mixing Conditions*. *Vol*. 1. Kendrick Press, Heber City, UT. MR2325294
[15] CHAZOTTES, J.-R., COLLET, P. and SCHMITT, B. (2005). Statistical consequences of the Devroye inequality for processes. Applications to a class of non-uniformly hyperbolic dynamical systems. *Nonlinearity* **18** 2341–2364. MR2165706
[16] CHAZOTTES, J.-R., COLLET, P. and SCHMITT, B. (2005). Devroye inequality for a class of non-uniformly hyperbolic dynamical systems. *Nonlinearity* **18** 2323–2340. MR2166315
[17] CHAZOTTES, J.-R. and GOUËZEL, S. (2012). Optimal concentration inequalities for dynamical systems. *Comm. Math. Phys.* **316** 843–889. MR2993935
[18] CHERNOV, N. (1999). Decay of correlations and dispersing billiards. *J. Stat. Phys.* **94** 513–556. MR1675363
[19] COLLET, P., MARTINEZ, S. and SCHMITT, B. (2002). Exponential inequalities for dynamical measures of expanding maps of the interval. *Probab. Theory Related Fields* **123** 301–322. MR1918536
[20] DAVYDOV, Y. A. (1968). Convergence of distributions generated by stationary stochastic processes. *Theory Probab. Appl.* **13** 691–696.
[21] 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
[22] DEDECKER, J. and PRIEUR, C. (2005). New dependence coefficients. Examples and applications to statistics. *Probab. Theory Related Fields* **132** 203–236. MR2199291
[23] DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (1996). *A Probabilistic Theory of Pattern Recognition*. *Applications of Mathematics* (*New York*) **31**. Springer, New York. MR1383093
[24] DEVROYE, L. and LUGOSI, G. (2001). *Combinatorial Methods in Density Estimation*. Springer, New York. MR1843146
[25] EBERTS, M. and STEINWART, I. (2013). Optimal regression rates for SVMs using Gaussian kernels. *Electron. J. Stat.* **7** 1–42. MR3020412
[26] GYÖRFI, L., KOHLER, M., KRZYŻAK, A. and WALK, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York. MR1920390
[27] HANG, H. and STEINWART, I. (2014). Fast learning from α-mixing observations. *J. Multivariate Anal.* **127** 184–199. MR3188886
[28] HANG, H. and STEINWART, I. (2016). Supplement to "A Bernstein-type inequality for some mixing processes and dynamical systems with an application to learning." DOI:10.1214/16-AOS1465SUPP.
[29] HOFBAUER, F. and KELLER, G. (1982). Ergodic properties of invariant measures for piecewise monotonic transformations. *Math. Z.* **180** 119–140. MR0656227
[30] IBRAGIMOV, I. A. (1962). Some limit theorems for stationary processes. *Theory Probab. Appl.* **7** 349–382.
[31] JAGER, L., MAES, J. and NINET, A. (2015). Exponential decay of correlations for a real-valued dynamical system embedded in $\mathbb{R}_2$. *C. R. Math. Acad. Sci. Paris* **353** 1041–1045. MR3419857
[32] KELLER, G. and NOWICKI, T. (1992). Spectral theory, zeta functions and the distribution of periodic points for Collet–Eckmann maps. *Comm. Math. Phys.* **149** 31–69. MR1182410

[33] LASOTA, A. and MACKEY, M. C. (1985). *Probabilistic Properties of Deterministic Systems*. Cambridge Univ. Press, Cambridge. MR0832868

[34] LIVERANI, C. (1995). Decay of correlations. *Ann. of Math.* (2) **142** 239–301. MR1343323

[35] LUZZATTO, S. and MELBOURNE, I. (2013). Statistical properties and decay of correlations for interval maps with critical points and singularities. *Comm. Math. Phys.* **320** 21–35. MR3046988

[36] MAMMEN, E. and TSYBAKOV, A. B. (1999). Smooth discrimination analysis. *Ann. Statist.* **27** 1808–1829. MR1765618

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

[38] MAUME-DESCHAMPS, V. (2006). Exponential inequalities and functional estimations for weak dependent data; applications to dynamical systems. *Stoch. Dyn.* **6** 535–560. MR2285515

[39] MCGOFF, K., MUKHERJEE, S., NOBEL, A. and PILLAI, N. (2015). Consistency of maximum likelihood estimation for some dynamical systems. *Ann. Statist.* **43** 1–29. MR3285598

[40] MCGOFF, K., MUKHERJEE, S. and PILLAI, N. S. (2012). Statistical inference for dynamical systems: A review. Preprint. Available at arXiv:1204.6265.

[41] MERLEVÈDE, F., PELIGRAD, M. and RIO, E. (2009). Bernstein inequality and moderate deviations under strong mixing conditions. In *High Dimensional Probability V*: *The Luminy Volume*. *Inst. Math. Stat. Collect.* **5** 273–292. IMS, Beachwood, OH. MR2797953

[42] MODHA, D. S. and MASRY, E. (1996). Minimum complexity regression estimation with weakly dependent observations. *IEEE Trans. Inform. Theory* **42** 2133–2145. MR1447519

[43] RIO, E. (1996). Sur le théorème de Berry–Esseen pour les suites faiblement dépendantes. *Probab. Theory Related Fields* **104** 255–282. MR1373378

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

[45] RUELLE, D. (1976). A measure associated with axiom-A attractors. *Amer. J. Math.* **98** 619–654. MR0415683

[46] RUNST, T. and SICKEL, W. (1996). *Sobolev Spaces of Fractional Order, Nemytskij Operators, and Nonlinear Partial Differential Equations*. *De Gruyter Series in Nonlinear Analysis and Applications* **3**. de Gruyter, Berlin. MR1419319

[47] RYCHLIK, M. (1983). Bounded variation and invariant measures. *Studia Math.* **76** 69–80. MR0728198

[48] SAMSON, P.-M. (2000). Concentration of measure inequalities for Markov chains and Φ-mixing processes. *Ann. Probab.* **28** 416–461. MR1756011

[49] SHUB, M. (1987). *Global Stability of Dynamical Systems*. Springer, New York. MR0869255

[50] SINAĬ, J. G. (1972). Gibbs measures in ergodic theory. *Russ. Math. Surveys* **27** 21–69.

[51] STEINWART, I. (2009). Two oracle inequalities for regularized boosting classifiers. *Stat. Interface* **2** 271–284. MR2540086

[52] STEINWART, I. and ANGHEL, M. (2009). Consistency of support vector machines for forecasting the evolution of an unknown ergodic dynamical system from observations with unknown noise. *Ann. Statist.* **37** 841–875. MR2502653

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

[54] STEINWART, I. and CHRISTMANN, A. (2011). Estimating conditional quantiles with the help of the pinball loss. *Bernoulli* **17** 211–225. MR2797989

[55] STEINWART, I., HUSH, D. and SCOVEL, C. (2006). An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. *IEEE Trans. Inform. Theory* **52** 4635–4643. MR2300845

[56] STEINWART, I., HUSH, D. and SCOVEL, C. (2009). Optimal rates for regularized least squares regression. In *Proceedings of the* 22*nd Annual Conference on Learning Theory* (S. Dasgupta and A. Klivans, eds.) 79–93. Available at http://www.cs.mcgill.ca/~colt2009/papers/038.pdf.
[57] TAKEUCHI, I., LE, Q. V., SEARS, T. D. and SMOLA, A. J. (2006). Nonparametric quantile estimation. *J. Mach. Learn. Res.* **7** 1231–1264. MR2274404
[58] TRIEBEL, H. (2010). *Theory of Function Spaces*. Birkhäuser/Springer, Basel. MR3024598
[59] TSYBAKOV, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** 135–166. MR2051002
[60] VIANA, M. (1997). *Stochastic Dynamics of Deterministic Systems* **21**. IMPA, Brazil.
[61] WINTENBERGER, O. (2010). Deviation inequalities for sums of weakly dependent time series. *Electron. Commun. Probab.* **15** 489–503. MR2733373
[62] YOUNG, L.-S. (1998). Statistical properties of dynamical systems with some hyperbolicity. *Ann. of Math.* (2) **147** 585–650. MR1637655
[63] ZHANG, J. (2004). Sieve estimates via neural network for strong mixing processes. *Stat. Inference Stoch. Process.* **7** 115–135. MR2061181

The Annals of Statistics
2017, Vol. 45, No. 2, 744–770
DOI: 10.1214/16-AOS1466

CONSISTENCY OF LIKELIHOOD ESTIMATION FOR GIBBS POINT PROCESSES[1]

BY DAVID DEREUDRE* AND FRÉDÉRIC LAVANCIER†,‡

University of Lille 1, University of Nantes† and Inria‡*

Strong consistency of the maximum likelihood estimator (MLE) for parametric Gibbs point process models is established. The setting is very general. It includes pairwise pair potentials, finite and infinite multibody interactions and geometrical interactions, where the range can be finite or infinite. The Gibbs interaction may depend linearly or nonlinearly on the parameters, a particular case being hardcore parameters and interaction range parameters. As important examples, we deduce the consistency of the MLE for all parameters of the Strauss model, the hardcore Strauss model, the Lennard–Jones model and the area-interaction model.

REFERENCES

[1] BADDELEY, A. J. and VAN LIESHOUT, M. N. M. (1995). Area-interaction point processes. *Ann. Inst. Statist. Math.* **47** 601–619. MR1370279

[2] CHIU, S. N., STOYAN, D., KENDALL, W. S. and MECKE, J. (2013). *Stochastic Geometry and Its Applications*, 3rd ed. Wiley, Chichester. MR3236788

[3] DEREUDRE, D. (2009). The existence of Quermass-interaction processes for nonlocally stable interaction and nonbounded convex grains. *Adv. in Appl. Probab.* **41** 664–681. MR2571312

[4] DEREUDRE, D. (2016). Variational principle for Gibbs point processes with finite range interaction. *Electron. Commun. Probab.* **21** Paper No. 10, 11. MR3485379

[5] DEREUDRE, D. and LAVANCIER, F. (2016). Supplement to "Consistency of likelihood estimation for Gibbs point processes." DOI:10.1214/16-AOS1466SUPP.

[6] DIGGLE, P. J., FIKSEL, T., GRABARNIK, P., OGATA, Y., STOYAN, D. and TANEMURA, M. (1994). On parameter estimation for pairwise interaction point processes. In *International Statistical Review/Revue Internationale de Statistique* 99–117.

[7] DIGGLE, P. J. and GRATTON, R. J. (1984). Monte Carlo methods of inference for implicit statistical models. *J. Roy. Statist. Soc. Ser. B* **46** 193–227. MR0781880

[8] GEORGII, H.-O. (1988). *Gibbs Measures and Phase Transitions*. *de Gruyter Studies in Mathematics* **9**. de Gruyter, Berlin. MR0956646

[9] GEORGII, H.-O. (1994). Large deviations and the equivalence of ensembles for Gibbsian particle systems with superstable interaction. *Probab. Theory Related Fields* **99** 171–195. MR1278881

[10] GEORGII, H.-O. (1995). The equivalence of ensembles for classical systems of particles. *J. Stat. Phys.* **80** 1341–1378. MR1349785

[11] GEYER, C. J. and MØLLER, J. (1994). Simulation procedures and likelihood inference for spatial point processes. *Scand. J. Statist.* **21** 359–373. MR1310082

MSC2010 subject classifications. 62F12, 62M30, 60G55, 60D05.

Key words and phrases. Parametric estimation, variational principle, Strauss model, Lennard–Jones model, area-interaction model.

[12] GUYON, X. (1995). *Random Fields on a Network*: *Modeling*, *Statistics*, *and Applications*. Springer, New York. MR1344683

[13] ILLIAN, J., PENTTINEN, A., STOYAN, H. and STOYAN, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. Wiley, Chichester. MR2384630

[14] JENSEN, J. L. (1993). Asymptotic normality of estimates in spatial point processes. *Scand*. *J*. *Statist*. **20** 97–109. MR1229287

[15] KENDALL, W. S., VAN LIESHOUT, M. N. M. and BADDELEY, A. J. (1999). Quermass-interaction processes: Conditions for stability. *Adv. in Appl*. *Probab.* **31** 315–342. MR1724554

[16] KÜNSCH, H. (1981). Thermodynamics and statistical analysis of Gaussian random fields. *Z. Wahrsch*. *Verw. Gebiete* **58** 407–421. MR0639149

[17] MASE, S. (1992). Uniform LAN condition of planar Gibbsian point processes and optimality of maximum likelihood estimators of soft-core potential functions. *Probab. Theory Related Fields* **92** 51–67. MR1156450

[18] MASE, S. (2002). Asymptotic properties of MLE's of Gibbs models on $\mathbb{R}^d$. Unpublished manuscript.

[19] MATEU, J. and MONTES, F. (2001). Likelihood inference for Gibbs processes in the analysis of spatial point patterns. *Int*. *Stat*. *Rev*. **69** 81–104.

[20] MØLLER, J. and WAAGEPETERSEN, R. (2003). *Statistical Inference and Simulation for Spatial Point Processes*. Chapman & Hall, Boca Raton.

[21] OGATA, Y. and TANEMURA, M. (1984). Likelihood analysis of spatial point patterns. *J*. *Roy. Statist*. *Soc*. *Ser. B* **46** 496–518. MR0790635

[22] PICKARD, D. K. (1987). Inference for discrete Markov fields: The simplest nontrivial case. *J*. *Amer. Statist*. *Assoc*. **82** 90–96. MR0883337

[23] PRESTON, C. (1976). *Random Fields*. Springer, Berlin. MR0448630

[24] RUELLE, D. (1969). *Statistical Mechanics*: *Rigorous Results*. Benjamin, New York. MR0289084

[25] RUELLE, D. (1970). Superstable interactions in classical statistical mechanics. *Comm*. *Math*. *Phys*. **18** 127–159. MR0266565

[26] VAN LIESHOUT, M. N. M. (2000). *Markov Point Processes and Their Applications*. World Scientific, Singapore.

[27] VARADHAN, S. R. S. (1988). Large deviations and applications. In *École D'Été de Probabilités de Saint-Flour XV–XVII*, 1985–87. *Lecture Notes in Math*. **1362** 1–49. Springer, Berlin. MR0983371

[28] WIDOM, B. and ROWLINSON, J. S. (1970). New model for the study of liquid-vapor phase transitions. *J. Chem*. *Phys*. **52** 1670–1684.

The Annals of Statistics
2017, Vol. 45, No. 2, 771–799
DOI: 10.1214/16-AOS1467

TESTS FOR HIGH-DIMENSIONAL DATA BASED ON MEANS, SPATIAL SIGNS AND SPATIAL RANKS

BY ANIRVAN CHAKRABORTY AND PROBAL CHAUDHURI

Indian Statistical Institute

Tests based on mean vectors and spatial signs and ranks for a zero mean in one-sample problems and for the equality of means in two-sample problems have been studied in the recent literature for high-dimensional data with the dimension larger than the sample size. For the above testing problems, we show that under suitable sequences of alternatives, the powers of the mean-based tests and the tests based on spatial signs and ranks tend to be same as the data dimension tends to infinity for any sample size when the coordinate variables satisfy appropriate mixing conditions. Further, their limiting powers do not depend on the heaviness of the tails of the distributions. This is in striking contrast to the asymptotic results obtained in the classical multivariate setting. On the other hand, we show that in the presence of stronger dependence among the coordinate variables, the spatial-sign- and rank-based tests for high-dimensional data can be asymptotically more powerful than the mean-based tests if, in addition to the data dimension, the sample size also tends to infinity. The sizes of some mean-based tests for high-dimensional data studied in the recent literature are observed to be significantly different from their nominal levels. This is due to the inadequacy of the asymptotic approximations used for the distributions of those test statistics. However, our asymptotic approximations for the tests based on spatial signs and ranks are observed to work well when the tests are applied on a variety of simulated and real datasets.

REFERENCES

BAI, Z. and SARANADASA, H. (1996). Effect of high dimension: By an example of a two sample problem. *Statist. Sinica* **6** 311–329. MR1399305

BRADLEY, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surv.* **2** 107–144. MR2178042

CAI, T. T., LIU, W. and XIA, Y. (2014). Two-sample test of high dimensional means under dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 349–372. MR3164870

CAPPÉ, O., MOULINES, E. and RYDÉN, T. (2005). *Inference in Hidden Markov Models*. Springer, New York. MR2159833

CHAKRABORTY, A. and CHAUDHURI, P. (2016). Supplement to "Tests for high-dimensional data based on means, spatial signs and spatial ranks." DOI:10.1214/16-AOS1467SUPP.

CHEN, S. X. and QIN, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.* **38** 808–835. MR2604697

MSC2010 subject classifications. Primary 62H15, 62G10; secondary 60G10, 62E20.

Key words and phrases. ARMA processes, heavy tailed distributions, permutation tests, ρ-mixing, randomly scaled ρ-mixing, spherical distributions, stationary sequences.

CHOI, K. and MARDEN, J. (1997). An approach to multivariate rank tests in multivariate analysis of variance. *J. Amer. Statist. Assoc.* **92** 1581–1590. MR1615267

FAN, J. and LIN, S.-K. (1998). Test of significance when data are curves. *J. Amer. Statist. Assoc.* **93** 1007–1021. MR1649196

FENG, L., ZOU, C., WANG, Z. and ZHU, L. (2015). Two-sample Behrens–Fisher problem for high-dimensional data. *Statist. Sinica* **25** 1297–1312. MR3409068

GREGORY, K. B., CARROLL, R. J., BALADANDAYUTHAPANI, V. and LAHIRI, S. N. (2015). A two-sample test for equality of means in high dimension. *J. Amer. Statist. Assoc.* **110** 837–849. MR3367268

HETTMANSPERGER, T. P. and MCKEAN, J. W. (2011). *Robust Nonparametric Statistical Methods*, 2nd ed. *Monographs on Statistics and Applied Probability* **119**. CRC Press, Boca Raton, FL. MR2779026

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

KALLENBERG, O. (2005). *Probabilistic Symmetries and Invariance Principles*. Springer, New York. MR2161313

KATAYAMA, S. and KANO, Y. (2014). A new test on high-dimensional mean vector without any assumption on population covariance matrix. *Comm. Statist. Theory Methods* **43** 5290–5304. MR3282878

KOLMOGOROV, A. N. and ROZANOV, J. A. (1960). On a strong mixing condition for stationary Gaussian processes. *Teor. Verojatnost. i Primenen.* **5** 222–227. MR0133175

LIN, Z. and LU, C. (1996). *Limit Theory for Mixing Dependent Random Variables*. *Mathematics and Its Applications* **378**. Kluwer, Dordrecht. MR1486580

MARDEN, J. I. (1999). Multivariate rank tests. In *Multivariate Analysis, Design of Experiments, and Survey Sampling*. *Statist. Textbooks Monogr.* **159** 401–432. Dekker, New York. MR1719102

MÖTTÖNEN, J. and OJA, H. (1995). Multivariate spatial sign and rank methods. *J. Nonparametr. Stat.* **5** 201–213. MR1346895

MÖTTÖNEN, J., OJA, H. and TIENARI, J. (1997). On the efficiency of multivariate spatial sign and rank tests. *Ann. Statist.* **25** 542–552. MR1439313

OJA, H. (2010). *Multivariate Nonparametric Methods with R: An Approach Based on Spatial Signs and Ranks*. *Lecture Notes in Statistics* **199**. Springer, New York. MR2598854

PURI, M. L. and SEN, P. K. (1971). *Nonparametric Methods in Multivariate Analysis*. Wiley, New York. MR0298844

SRIVASTAVA, M. S., KATAYAMA, S. and KANO, Y. (2013). A two sample test in high dimensional data. *J. Multivariate Anal.* **114** 349–358. MR2993891

WANG, L., PENG, B. and LI, R. (2015). A high-dimensional nonparametric multivariate test for mean vector. *J. Amer. Statist. Assoc.* **110** 1658–1669. MR3449062

WEI, S., LEE, C., WICHERS, L. and MARRON, J. S. (2016). Direction-projection-permutation for high-dimensional hypothesis tests. *J. Comput. Graph. Statist.* **25** 549–569. MR3499694

The Annals of Statistics
2017, Vol. 45, No. 2, 800–832
DOI: 10.1214/16-AOS1468

INFERENCE ON THE MODE OF WEAK DIRECTIONAL SIGNALS: A LE CAM PERSPECTIVE ON HYPOTHESIS TESTING NEAR SINGULARITIES

BY DAVY PAINDAVEINE[1] AND THOMAS VERDEBOUT

Université libre de Bruxelles

We revisit, in an original and challenging perspective, the problem of testing the null hypothesis that the mode of a directional signal is equal to a given value. Motivated by a real data example where the signal is weak, we consider this problem under asymptotic scenarios for which the signal strength goes to zero at an arbitrary rate η_n. Both under the null and the alternative, we focus on rotationally symmetric distributions. We show that, while they are asymptotically equivalent under fixed signal strength, the classical Wald and Watson tests exhibit very different (null and nonnull) behaviours when the signal becomes arbitrarily weak. To fully characterize how challenging the problem is as a function of η_n, we adopt a Le Cam, convergence-of-statistical-experiments, point of view and show that the resulting limiting experiments crucially depend on η_n. In the light of these results, the Watson test is shown to be *adaptively* rate-consistent and essentially adaptively Le Cam optimal. Throughout, our theoretical findings are illustrated via Monte-Carlo simulations. The practical relevance of our results is also shown on the real data example that motivated the present work.

REFERENCES

BOENTE, G., RODRIGUEZ, D. and GONZÁLEZ MANTEIGA, W. (2014). Goodness-of-fit test for directional data. *Scand. J. Stat.* **41** 259–275. MR3181143

CAI, T. T., JIN, J. and LOW, M. G. (2007). Estimation and confidence sets for sparse normal mixtures. *Ann. Statist.* **35** 2421–2449. MR2382653

CHANG, T. and RIVEST, L.-P. (2001). M-estimation for location and regression parameters in group models: A case study using Stiefel manifolds. *Ann. Statist.* **29** 784–814. MR1865341

CUTTING, C., PAINDAVEINE, D. and VERDEBOUT, T. (2016a). Testing uniformity on high-dimensional spheres against monotone rotationally symmetric alternatives. *Ann. Statist.* To appear.

CUTTING, C., PAINDAVEINE, D. and VERDEBOUT, T. (2016b). Supplement to "Testing uniformity on high-dimensional spheres against monotone rotationally symmetric alternatives". *Ann. Statist.* To appear.

DUFOUR, J.-M. (1997). Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica* **65** 1365–1387. MR1604305

DUFOUR, J.-M. (2006). Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics. *J. Econometrics* **133** 443–477. MR2252905

MSC2010 subject classifications. Primary 62G10, 62G20; secondary 62G35, 62H11.

Key words and phrases. Contiguity, convergence of statistical experiments, directional statistics, robust tests, rotationally symmetric distributions.

Fisher, N. I., Lewis, T. and Embleton, B. J. J. (1987). *Statistical Analysis of Spherical Data*. Cambridge Univ. Press, Cambridge. MR0899958

Forchini, G. (2009). Some properties of tests for parameters that can be arbitrarily close to being unidentified. *J. Statist. Plann. Inference* **139** 3193–3199. MR2535193

Forchini, G. and Hillier, G. (2003). Conditional inference for possibly unidentified structural equations. *Econometric Theory* **19** 707–743. MR2002577

Hayakawa, T. (1990). On tests for the mean direction of the Langevin distribution. *Ann. Inst. Statist. Math.* **42** 359–373. MR1064794

Hayakawa, T. and Puri, M. L. (1985). Asymptotic expansions of the distributions of some test statistics. *Ann. Inst. Statist. Math.* **37** 95–108. MR0790378

Larsen, P. V. and Jupp, P. E. (2003). Parametrization-invariant Wald tests. *Bernoulli* **9** 167–182. MR1963676

Ley, C., Paindaveine, D. and Verdebout, T. (2015). High-dimensional tests for spherical location and spiked covariance. *J. Multivariate Anal.* **139** 79–91. MR3349481

Ley, C., Sabbah, C. and Verdebout, T. (2014). A new concept of quantiles for directional data and the angular Mahalanobis depth. *Electron. J. Stat.* **8** 795–816. MR3217789

Ley, C., Swan, Y., Thiam, B. and Verdebout, T. (2013). Optimal R-estimation of a spherical location. *Statist. Sinica* **23** 305–332. MR3076169

Paindaveine, D. and Verdebout, T. (2015). Optimal rank-based tests for the location parameter of a rotationally symmetric distribution on the hypersphere. In *Mathematical Statistics and Limit Theorems* (M. Hallin, D. Mason, D. Pfeifer and J. Steinebach, eds.). 249–269. Springer, Cham. MR3380740

Pötscher, B. M. (2002). Lower risk bounds and properties of confidence sets for ill-posed estimation problems with applications to spectral density and persistence estimation, unit roots, and estimation of long memory parameters. *Econometrica* **70** 1035–1065. MR1910411

Preuss, P., Vetter, M. and Dette, H. (2013). Testing semiparametric hypotheses in locally stationary processes. *Scand. J. Stat.* **40** 417–437. MR3091690

Toyoda, Y., Suga, K., Murakami, K., Hasegawa, H., Shibata, S., Domingo, V., Escobar, I., Kamata, K., Bradt, H., Clark, G. and La Pointe, M. (1965). Studies of primary cosmic rays in the energy region 1014 eV to 1017 eV (Bolivian air shower joint experiment). *Proc. Int. Conf. Cosmic Rays* (*London*) **2** 708–711.

Tsai, M.-T. and Sen, P. K. (2007). Locally best rotation-invariant rank tests for modal location. *J. Multivariate Anal.* **98** 1160–1179. MR2326245

van der Vaart, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247

Watson, G. S. (1983). *Statistics on Spheres*. *University of Arkansas Lecture Notes in the Mathematical Sciences* **6**. Wiley, New York. MR0709262

The Annals of Statistics
2017, Vol. 45, No. 2, 833–865
DOI: 10.1214/16-AOS1469

ASYMPTOTIC BEHAVIOUR OF THE EMPIRICAL BAYES POSTERIORS ASSOCIATED TO MAXIMUM MARGINAL LIKELIHOOD ESTIMATOR

BY JUDITH ROUSSEAU[1,*,†] AND BOTOND SZABO[2,‡,§]

University Paris Dauphine, CREST-ENSAE,† Budapest University of Technology‡ and Leiden University§*

We consider the asymptotic behaviour of the marginal maximum likelihood empirical Bayes posterior distribution in general setting. First, we characterize the set where the maximum marginal likelihood estimator is located with high probability. Then we provide oracle type of upper and lower bounds for the contraction rates of the empirical Bayes posterior. We also show that the hierarchical Bayes posterior achieves the same contraction rate as the maximum marginal likelihood empirical Bayes posterior. We demonstrate the applicability of our general results for various models and prior distributions by deriving upper and lower bounds for the contraction rates of the corresponding empirical and hierarchical Bayes posterior distributions.

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] BABENKO, A. and BELITSER, E. (2010). Oracle convergence rate of posterior under projection prior and Bayesian model selection. *Math. Methods Statist.* **19** 219–245. MR2742927

[3] BELITSER, E. and ENIKEEVA, F. (2008). Empirical Bayesian test of the smoothness. *Math. Methods Statist.* **17** 1–18. MR2400361

[4] BELITSER, E. and GHOSAL, S. (2003). Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *Ann. Statist.* **31** 536–559. MR1983541

[5] CASTILLO, I. (2008). Lower bounds for posterior rates with Gaussian process priors. *Electron. J. Stat.* **2** 1281–1299. MR2471287

[6] CASTILLO, I. and ROUSSEAU, J. (2015). A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *Ann. Statist.* **43** 2353–2383. MR3405597

[7] COX, D. D. (1993). An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.* **21** 903–923. MR1232525

[8] DONNET, S., RIVOIRARD, V., ROUSSEAU, J. and SCRICCIOLO, C. (2014). Posterior concentration rates for empirical Bayes procedures, with applications to Dirichlet Process mixtures. Available at arXiv:1406.4406v1.

[9] FLORENS, J.-P. and SIMONI, A. (2012). Regularized posteriors in linear ill-posed inverse problems. *Scand. J. Stat.* **39** 214–235. MR2927022

[10] GAO, C. and ZHOU, H. H. (2016). Rate exact Bayesian adaptation with modified block priors. *Ann. Statist.* **44** 318–345. MR3449770

MSC2010 subject classifications. Primary 62G20, 62G05, 60K35; secondary 62G08, 62G07.

Key words and phrases. Posterior contraction rates, adaptation, empirical Bayes, hierarchical Bayes, nonparametric regression, density estimation, Gaussian prior, truncation prior.

[11] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007

[12] 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

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

[14] KNAPIK, B. T., SZABÓ, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2016). Bayes procedures for adaptive inference in inverse problems for the white noise model. *Probab. Theory Related Fields* **164** 771–813. MR3477780

[15] KNAPIK, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2011). Bayesian inverse problems with Gaussian priors. *Ann. Statist.* **39** 2626–2657. MR2906881

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

[17] LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation*, 2nd ed. Springer, New York. MR1639875

[18] LIAN, H. (2014). Adaptive rates of contraction of posterior distributions in Bayesian wavelet regression. *J. Statist. Plann. Inference* **145** 92–101. MR3125351

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

[20] RAY, K. (2013). Bayesian inverse problems with non-conjugate priors. *Electron. J. Stat.* **7** 2516–2549. MR3117105

[21] RIVOIRARD, V. and ROUSSEAU, J. (2012). Bernstein–von Mises theorem for linear functionals of the density. *Ann. Statist.* **40** 1489–1523. MR3015033

[22] RIVOIRARD, V. and ROUSSEAU, J. (2012). Posterior concentration rates for infinite dimensional exponential families. *Bayesian Anal.* **7** 311–333. MR2934953

[23] ROUSSEAU, J. and SZABO, B. (2016). Supplement to "Asymptotic behaviour of the empirical Bayes posteriors associated to maximum marginal likelihood estimator." DOI:10.1214/16-AOS1469SUPP.

[24] SERRA, P. and KRIVOBOKOVA, T. (2014). Adaptive empirical Bayesian smoothing splines. Available at arXiv:1411.6860.

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

[26] SNIEKERS, S. and VAN DER VAART, A. (2015). Adaptive Bayesian credible sets in regression with a Gaussian process prior. *Electron. J. Stat.* **9** 2475–2527. MR3425364

[27] SZABÓ, B., VAN DER VAART, A. and VAN ZANTEN, H. (2015). Honest Bayesian confidence sets for the L^2-norm. *J. Statist. Plann. Inference* **166** 36–51. MR3390132

[28] SZABÓ, B., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *Ann. Statist.* **43** 1391–1428. MR3357861

[29] SZABÓ, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2013). Empirical Bayes scaling of Gaussian priors in the white noise model. *Electron. J. Stat.* **7** 991–1018. MR3044507

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

[31] 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

[32] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2008). Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the Limits of Contemporary Statistics*: *Contributions in Honor of Jayanta K. Ghosh*. *Inst. Math. Stat. Collect.* **3** 200–222. IMS, Beachwood, OH. MR2459226

[33] VERDINELLI, I. and WASSERMAN, L. (1998). Bayesian goodness-of-fit testing using infinite-dimensional exponential families. *Ann. Statist.* **26** 1215–1241. MR1647645

[34] ZHAO, L. H. (2000). Bayesian aspects of some nonparametric problems. *Ann. Statist.* **28** 532–552. MR1790008

The Annals of Statistics
2017, Vol. 45, No. 2, 866–896
DOI: 10.1214/16-AOS1471

STATISTICAL CONSISTENCY AND ASYMPTOTIC NORMALITY FOR HIGH-DIMENSIONAL ROBUST M-ESTIMATORS

BY PO-LING LOH

University of Wisconsin-Madison

We study theoretical properties of regularized robust M-estimators, applicable when data are drawn from a sparse high-dimensional linear model and contaminated by heavy-tailed distributions and/or outliers in the additive errors and covariates. We first establish a form of local statistical consistency for the penalized regression estimators under fairly mild conditions on the error distribution: When the derivative of the loss function is bounded and satisfies a local restricted curvature condition, all stationary points within a constant radius of the true regression vector converge at the minimax rate enjoyed by the Lasso with sub-Gaussian errors. When an appropriate nonconvex regularizer is used in place of an ℓ_1-penalty, we show that such stationary points are in fact unique and equal to the local oracle solution with the correct support; hence, results on asymptotic normality in the low-dimensional case carry over immediately to the high-dimensional setting. This has important implications for the efficiency of regularized nonconvex M-estimators when the errors are heavy-tailed. Our analysis of the local curvature of the loss function also has useful consequences for optimization when the robust regression function and/or regularizer is nonconvex and the objective function possesses stationary points outside the local region. We show that as long as a composite gradient descent algorithm is initialized within a constant radius of the true regression vector, successive iterates will converge at a linear rate to a stationary point within the local region. Furthermore, the global optimum of a convex regularized robust regression function may be used to obtain a suitable initialization. The result is a novel two-step procedure that uses a convex M-estimator to achieve consistency and a nonconvex M-estimator to increase efficiency. We conclude with simulation results that corroborate our theoretical findings.

REFERENCES

AGARWAL, A., NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Fast global convergence of gradient methods for high-dimensional statistical recovery. *Ann. Statist.* **40** 2452–2482. MR3097609

BAI, Z. D. and WU, Y. (1997). General M-estimation. *J. Multivariate Anal.* **63** 119–135. MR1491570

BERTSEKAS, D. P. (1999). *Nonlinear Programming*. Athena Scientific, Belmont, MA.

BICKEL, P. J. (1975). One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc.* **70** 428–434. MR0386168

MSC2010 subject classifications. 62F12.

Key words and phrases. Robust regression, high-dimensional statistics, statistical consistency, support recovery, nonconvex optimization.

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

BOX, G. E. P. (1953). Non-normality and tests on variances. *Biometrika* **40** 318–335. MR0058937

BRADIC, J., FAN, J. and WANG, W. (2011). Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **73** 325–349. MR2815779

CLARKE, F. H. (1983). *Optimization and Nonsmooth Analysis*. Wiley, New York. MR0709590

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

FAN, J., LI, Q. and WANG, Y. (2014). Robust estimation of high-dimensional mean regression. Preprint. Available at arXiv:1410.2150.

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

FREEDMAN, D. A. and DIACONIS, P. (1982). On inconsistent M-estimators. *Ann. Statist.* **10** 454–461. MR0653520

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

HAMPEL, F. R. (1968). Contributions to the theory of robust estimation Ph.D. thesis, Univ. of California, Berkeley.

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

HE, X. and SHAO, Q.-M. (1996). A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs. *Ann. Statist.* **24** 2608–2630. MR1425971

HE, X. and SHAO, Q.-M. (2000). On parameters of increasing dimensions. *J. Multivariate Anal.* **73** 120–135. MR1766124

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

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

HUBER, P. J. (1981). *Robust Statistics*. Wiley, New York. MR0606374

LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation*. Springer, New York. MR1639875

LI, G., PENG, H. and ZHU, L. (2011). Nonconcave penalized M-estimation with a diverging number of parameters. *Statist. Sinica* **21** 391–419. MR2796868

LOH, P. (2016). Supplement to "Statistical consistency and asymptotic normality for high-dimensional robust M-estimators." DOI:10.1214/16-AOS1471SUPP.

LOH, P. and WAINWRIGHT, M. J. (2014). Support recovery without incoherence: A case for nonconvex regularization. Preprint. Available at arXiv:1412.5632.

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

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

MAMMEN, E. (1989). Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *Ann. Statist.* **17** 382–400. MR0981457

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

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

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

NESTEROV, Y. (2007). Gradient methods for minimizing composite objective function. CORE Discussion Papers No. 2007076, Université Catholique de Louvain, Center for Operations Research and Econometrics (CORE).

NOLAN, J. P. (2015). *Stable Distributions—Models for Heavy Tailed Data*. Birkhauser, Boston.

PORTNOY, S. (1985). Asymptotic behavior of M estimators of p regression parameters when p^2/n is large. II. Normal approximation. *Ann. Statist.* **13** 1403–1417. MR0811499

RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *J. Mach. Learn. Res.* **11** 2241–2259. MR2719855

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. MR2662343

ROUSSEEUW, P. J. and LEROY, A. M. (2005). *Robust Regression and Outlier Detection*. Wiley, New York.

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

SHEVLYAKOV, G., MORGENTHALER, S. and SHURYGIN, A. (2008). Redescending M-estimators. *J. Statist. Plann. Inference* **138** 2906–2917. MR2526216

SIMPSON, D. G., RUPPERT, D. and CARROLL, R. J. (1992). On one-step GM estimates and stability of inferences in linear regression. *J. Amer. Statist. Assoc.* **87** 439–450. MR1173809

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

TUKEY, J. W. (1960). A survey of sampling from contaminated distributions. In *Contributions to Probability and Statistics* 448–485. Stanford Univ. Press, Stanford, CA. MR0120720

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

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

WELSH, A. H. and RONCHETTI, E. (2002). A journey in single steps: Robust one-step M-estimation in linear regression. *J. Statist. Plann. Inference* **103** 287–310. MR1896997

YOHAI, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Ann. Statist.* **15** 642–656. MR0888431

YOHAI, V. J. and MARONNA, R. A. (1979). Asymptotic behavior of M-estimators for the linear model. *Ann. Statist.* **7** 258–268. MR0520237

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

The Annals of Statistics
2017, Vol. 45, No. 2, 897–922
DOI: 10.1214/16-AOS1474

INTERACTION PURSUIT IN HIGH-DIMENSIONAL MULTI-RESPONSE REGRESSION VIA DISTANCE CORRELATION[1]

BY YINFEI KONG*, DAOJI LI†, YINGYING FAN‡ AND JINCHI LV‡

California State University, Fullerton, University of Central Florida† and University of Southern California‡*

Feature interactions can contribute to a large proportion of variation in many prediction models. In the era of big data, the coexistence of high dimensionality in both responses and covariates poses unprecedented challenges in identifying important interactions. In this paper, we suggest a two-stage interaction identification method, called the interaction pursuit via distance correlation (IPDC), in the setting of high-dimensional multi-response interaction models that exploits feature screening applied to transformed variables with distance correlation followed by feature selection. Such a procedure is computationally efficient, generally applicable beyond the heredity assumption, and effective even when the number of responses diverges with the sample size. Under mild regularity conditions, we show that this method enjoys nice theoretical properties including the sure screening property, support union recovery and oracle inequalities in prediction and estimation for both interactions and main effects. The advantages of our method are supported by several simulation studies and real data analysis.

REFERENCES

[1] BIEN, J., TAYLOR, J. and TIBSHIRANI, R. (2013). A LASSO for hierarchical interactions. *Ann. Statist.* **41** 1111–1141. MR3113805

[2] CHEN, L. and HUANG, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *J. Amer. Statist. Assoc.* **107** 1533–1545. MR3036414

[3] CHOI, N. H., LI, W. and ZHU, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *J. Amer. Statist. Assoc.* **105** 354–364. MR2656056

[4] CHUN, H. and KELEŞ, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 3–25. MR2751241

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

[6] FAN, Y., KONG, Y., LI, D. and LV, J. (2016). Interaction pursuit with feature screening and selection. Preprint. Available at arXiv:1605.08933.

[7] FAN, Y., KONG, Y., LI, D. and ZHENG, Z. (2015). Innovated interaction screening for high-dimensional nonlinear classification. *Ann. Statist.* **43** 1243–1272. MR3346702

[8] HALL, P. and XUE, J.-H. (2014). On selecting interacting features from high-dimensional data. *Comput. Statist. Data Anal.* **71** 694–708. MR3132000

MSC2010 subject classifications. Primary 62H12, 62J02; secondary 62F07, 62F12.

Key words and phrases. Interaction pursuit, distance correlation, square transformation, multi-response regression, high dimensionality, sparsity.

[9] HAO, N. and ZHANG, H. H. (2014). Interaction screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc.* **109** 1285–1301. MR3265697

[10] HUO, X. and SZÉKELY, G. J. (2016). Fast computing for distance covariance. *Technometrics*. To appear.

[11] JIANG, B. and LIU, J. S. (2014). Variable selection for general index models via sliced inverse regression. *Ann. Statist.* **42** 1751–1786. MR3262467

[12] KONG, Y., LI, D., FAN, Y. and LV, J. (2016). Supplement to "Interaction pursuit in high-dimensional multi-response regression via distance correlation." DOI:10.1214/16-AOS1474SUPP.

[13] LI, J., ZHONG, W., LI, R. and WU, R. (2014). A fast algorithm for detecting gene-gene interactions in genome-wide association studies. *Ann. Appl. Stat.* **8** 2292–2318. MR3292498

[14] LI, R., ZHONG, W. and ZHU, L. (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc.* **107** 1129–1139. MR3010900

[15] LV, J. (2013). Impacts of high dimensionality in finite samples. *Ann. Statist.* **41** 2236–2262. MR3127865

[16] SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York. MR0595165

[17] SPELLMAN, P. T., SHERLOCK, G., ZHANG, M. Q., IYER, V. R., ANDERS, K., EISEN, M. B., BROWN, P. O., BOTSTEIN, D. and FUTCHER, B. (1998). Combined expression trait correlations and expression quantitative trait locus mapping. *Mol. Biol. Cell* **9** 3273–3297.

[18] SZÉKELY, G. J., RIZZO, M. L. and BAKIROV, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35** 2769–2794. MR2382665

[19] YUAN, M., JOSEPH, V. R. and ZOU, H. (2009). Structured variable selection and estimation. *Ann. Appl. Stat.* **3** 1738–1757. MR2752156

[20] ZHONG, W. and ZHU, L. (2015). An iterative approach to distance correlation-based sure independence screening. *J. Stat. Comput. Simul.* **85** 2331–2345. MR3339301