

THE ANNALS *of* STATISTICS

AN OFFICIAL JOURNAL OF THE
INSTITUTE OF MATHEMATICAL STATISTICS

Functional data analysis by matrix completion MARIE-HÉLÈNE DESCARY AND VICTOR M. PANARETOS	1
Bayesian fractional posteriors ANIRBAN BHATTACHARYA, DEBDEEP PATI AND YUN YANG	39
Distribution theory for hierarchical processes FEDERICO CAMERLENGHI, ANTONIO LIJOI, PETER ORBANZ AND IGOR PRÜNSTER	67
Adaptive estimation of the sparsity in the Gaussian vector model ALEXANDRA CARPENTIER AND NICOLAS VERZELEN	93
Approximate optimal designs for multivariate polynomial regression YOHANN DE CASTRO, FABRICE GAMBOA, DIDIER HENRION, ROXANA HESS AND JEAN-BERNARD LASSERRE	127
Efficient estimation of integrated volatility functionals via multiscale jackknife JIA LI, YUNXIAO LIU AND DACHENG XIU	156
Nonasymptotic rates for manifold, tangent space and curvature estimation EDDIE AAMARI AND CLÉMENT LEVRARD	177
Nonparametric testing for multiple survival functions with noninferiority margins HSIN-WEN CHANG AND IAN W. MCKEAGUE	205
Oracle inequalities and adaptive estimation in the convolution structure density model O. V. LEPSKI AND T. WILLER	233
Efficient multivariate entropy estimation via k -nearest neighbour distances THOMAS B. BERRETT, RICHARD J. SAMWORTH AND MING YUAN	288
Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models XUAN CAO, KSHITIJ KHARE AND MALAY GHOSH	319
Locally adaptive confidence bands TIM PATSCHKOWSKI AND ANGELIKA ROHDE	349
Asymptotic distribution-free change-point detection for multivariate and non-Euclidean data LYNNA CHU AND HAO CHEN	382
Statistics on the Stiefel manifold: Theory and applications RUDRASIS CHAKRABORTY AND BABA C. VEMURI	415
Goodness-of-fit tests for the functional linear model based on randomly projected empirical processes JUAN A. CUESTA-ALBERTOS, EDUARDO GARCÍA-PORTUGUÉS, MANUEL FEBRERO-BANDE AND WENCESLAO GONZÁLEZ-MANTEIGA	439
Convolved subsampling estimation with applications to block bootstrap JOHANNES TEWES, DIMITRIS N. POLITIS AND DANIEL J. NORDMAN	468
Feature elimination in kernel machines in moderately high dimensions SAYAN DASGUPTA, YAIR GOLDBERG AND MICHAEL R. KOSOROK	497
High-dimensional covariance matrices in elliptical distributions with application to spherical test JIANG HU, WEIMING LI, ZHI LIU AND WANG ZHOU	527
A critical threshold for design effects in network sampling KARL ROHE	556
Permutation p -value approximation via generalized Stolarsky invariance HERA Y. HE, KINJAL BASU, QINGYUAN ZHAO AND ART B. OWEN	583
Canonical correlation coefficients of high-dimensional Gaussian vectors: Finite rank case ZHIGANG BAO, JIANG HU, GUANGMING PAN AND WANG ZHOU	612
Uniform projection designs FASHENG SUN, YAPING WANG AND HONGQUAN XU	641

THE ANNALS OF STATISTICS

Vol. 47, No. 1, pp. 1–661 February 2019

INSTITUTE OF MATHEMATICAL STATISTICS

(Organized September 12, 1935)

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

IMS OFFICERS

President: Xiao-Li Meng, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138-2901, USA

President-Elect: Susan Murphy, Department of Statistics, Harvard University, Cambridge, Massachusetts 02138-2901, USA

Past President: Alison Etheridge, Department of Statistics, University of Oxford, Oxford, OX1 3LB, United Kingdom

Executive Secretary: Edsel Peña, Department of Statistics, University of South Carolina, Columbia, South Carolina 29208-001, USA

Treasurer: Zhengjun Zhang, Department of Statistics, University of Wisconsin, Madison, Wisconsin 53706-1510, USA

Program Secretary: Ming Yuan, Department of Statistics, Columbia University, New York, NY 10027-5927, USA

IMS EDITORS

The Annals of Statistics. *Editors:* Richard J. Samworth, Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, Cambridge, CB3 0WB, UK. Ming Yuan, Department of Statistics, Columbia University, New York, NY 10027, USA

The Annals of Applied Statistics. *Editor-in-Chief:* Tilmann Gneiting, Heidelberg Institute for Theoretical Studies, HITS gGmbH, Schloss-Wolfsbrunnengasse 35, 69118 Heidelberg, Germany

The Annals of Probability. *Editor:* Amir Dembo, Department of Statistics and Department of Mathematics, Stanford University, Stanford, California 94305, USA

The Annals of Applied Probability. *Editors:* Francois Delarue, Laboratoire J. A. Dieudonné, Université de Nice Sophia-Antipolis, France-06108 Nice Cedex 2. Peter Friz, Institut für Mathematik, Technische Universität Berlin, 10623 Berlin, Germany and Weierstrass-Institut für Angewandte Analysis und Stochastik, 10117 Berlin, Germany

Statistical Science. *Editor:* Cun-Hui Zhang, Department of Statistics, Rutgers University, Piscataway, NJ 08854, USA

The IMS Bulletin. *Editor:* Vlada Limic, UMR 7501 de l'Université de Strasbourg et du CNRS, 7 rue René Descartes, 67084 Strasbourg Cedex, France

The Annals of Statistics [ISSN 0090-5364 (print); ISSN 2168-8966 (online)], Volume 47, Number 1, February 2019. Published bimonthly by the Institute of Mathematical Statistics, 3163 Somerset Drive, Cleveland, Ohio 44122, USA. Periodicals postage paid at Cleveland, Ohio, and at additional mailing offices.

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

FUNCTIONAL DATA ANALYSIS BY MATRIX COMPLETION¹

BY MARIE-HÉLÈNE DESCARY AND VICTOR M. PANARETOS

Ecole Polytechnique Fédérale de Lausanne

Functional data analyses typically proceed by smoothing, followed by functional PCA. This paradigm implicitly assumes that rough variation is due to nuisance noise. Nevertheless, relevant functional features such as time-localised or short scale fluctuations may indeed be rough relative to the global scale, but still smooth at shorter scales. These may be confounded with the global smooth components of variation by the smoothing and PCA, potentially distorting the parsimony and interpretability of the analysis. The goal of this paper is to investigate how both smooth and rough variations can be recovered on the basis of discretely observed functional data. Assuming that a functional datum arises as the sum of two uncorrelated components, one smooth and one rough, we develop identifiability conditions for the recovery of the two corresponding covariance operators. The key insight is that they should possess complementary forms of parsimony: one smooth and finite rank (large scale), and the other banded and potentially infinite rank (small scale). Our conditions elucidate the precise interplay between rank, bandwidth and grid resolution. Under these conditions, we show that the recovery problem is equivalent to rank-constrained matrix completion, and exploit this to construct estimators of the two covariances, without assuming knowledge of the true bandwidth or rank; we study their asymptotic behaviour, and then use them to recover the smooth and rough components of each functional datum by best linear prediction. As a result, we effectively produce separate functional PCAs for smooth and rough variation.

REFERENCES

- [1] BAUSCHKE, H. H. and BORWEIN, J. M. (1996). On projection algorithms for solving convex feasibility problems. *SIAM Rev.* **38** 367–426. [MR1409591](#)
- [2] BOSQ, D. (2000). *Linear Processes in Function Spaces: Theory and Applications*. *Lecture Notes in Statistics* **149**. Springer, New York. [MR1783138](#)
- [3] BOSQ, D. (2014). Computing the best linear predictor in a Hilbert space. Applications to general ARMAH processes. *J. Multivariate Anal.* **124** 436–450. [MR3147336](#)
- [4] CHEN, Y. and WAINWRIGHT, M. J. (2015). Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. Available at [arXiv:1509.03025](#).
- [5] COLEMAN, T. F. and LI, Y. (1994). On the convergence of interior-reflective Newton methods for nonlinear minimization subject to bounds. *Math. Program.* **67** 189–224. [MR1305566](#)
- [6] COLEMAN, T. F. and LI, Y. (1996). An interior trust region approach for nonlinear minimization subject to bounds. *SIAM J. Optim.* **6** 418–445. [MR1387333](#)

MSC2010 subject classifications. Primary 62M; secondary 15A99, 62M15, 60G17.

Key words and phrases. Analyticity, banding, covariance operator, functional PCA, low rank, resolution, scale, smoothing.

- [7] DAUXOIS, J., POUSSE, A. and ROMAIN, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *J. Multivariate Anal.* **12** 136–154. [MR0650934](#)
- [8] DESCARY, M.-H. and PANARETOS, V. M. (2018). Supplement to “Functional data analysis by matrix completion.” DOI:[10.1214/17-AOS1590SUPP](#).
- [9] HALL, P., MÜLLER, H.-G. and WANG, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.* **34** 1493–1517. [MR2278365](#)
- [10] HORVÁTH, L. and KOKOSZKA, P. (2012). *Inference for Functional Data with Applications*. Springer, New York. [MR2920735](#)
- [11] HSING, T. and EUBANK, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley, Chichester. [MR3379106](#)
- [12] KIRÁLY, F. and TOMIOKA, R. (2012). A combinatorial algebraic approach for the identifiability of low-rank matrix completion. In *Proceedings of the 29th International Conference on Machine Learning*.
- [13] KRANTZ, S. G. and PARKS, H. R. (2002). *A Primer of Real Analytic Functions*, 2nd ed. Birkhäuser, Boston, MA. [MR1916029](#)
- [14] LI, Y. and HSING, T. (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Ann. Statist.* **38** 3321–3351. [MR2766854](#)
- [15] MATLAB (2012). Version 8.0.0.783 (R2012b). The MathWorks Inc., Natick, MA.
- [16] OPSOMER, J., WANG, Y. and YANG, Y. (2001). Nonparametric regression with correlated errors. *Statist. Sci.* **16** 134–153. [MR1861070](#)
- [17] PANARETOS, V. M. and TAVAKOLI, S. (2013). Cramér–Karhunen–Loève representation and harmonic principal component analysis of functional time series. *Stochastic Process. Appl.* **123** 2779–2807. [MR3054545](#)
- [18] RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. Springer, New York. [MR2168993](#)
- [19] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York. [MR1385671](#)
- [20] WANG, J. L., CHIOU, J. M. and MÜLLER, H.-G. (2015). Review of functional data analysis. Available at [arXiv:1507.05135](#).
- [21] YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100** 577–590. [MR2160561](#)
- [22] YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005). Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33** 2873–2903. [MR2253106](#)

BAYESIAN FRACTIONAL POSTERIOR

BY ANIRBAN BHATTACHARYA^{*,1}, DEBDEEP PATI^{*,2} AND YUN YANG[†]

Texas A&M University^{} and University of Illinois at Urbana–Champaign[†]*

We consider the *fractional posterior* distribution that is obtained by updating a prior distribution via Bayes theorem with a *fractional likelihood* function, a usual likelihood function raised to a fractional power. First, we analyze the contraction property of the fractional posterior in a general misspecified framework. Our contraction results only require a prior mass condition on certain Kullback–Leibler (KL) neighborhood of the true parameter (or the KL divergence minimizer in the misspecified case), and obviate constructions of test functions and sieves commonly used in the literature for analyzing the contraction property of a regular posterior. We show through a counterexample that some condition controlling the complexity of the parameter space is necessary for the regular posterior to contract, rendering additional flexibility on the choice of the prior for the fractional posterior. Second, we derive a novel Bayesian oracle inequality based on a PAC-Bayes inequality in misspecified models. Our derivation reveals several advantages of averaging based Bayesian procedures over optimization based frequentist procedures. As an application of the Bayesian oracle inequality, we derive a sharp oracle inequality in multivariate convex regression problems. We also illustrate the theory in Gaussian process regression and density estimation problems.

REFERENCES

- [1] ALQUIER, P., RIDGWAY, J. and CHOPIN, N. (2016). On the properties of variational approximations of Gibbs posteriors. *J. Mach. Learn. Res.* **17** Paper No. 239, 41. [MR3595173](#)
- [2] BALÁZS, G., GYÖRGY, A. and SZEPESVÁRI, C. (2015). Near-optimal max-affine estimators for convex regression. In *AISTATS*.
- [3] BARRON, A., SCHERVISH, M. J. and WASSERMAN, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* **27** 536–561. [MR1714718](#)
- [4] BARTLETT, P. L., BOUSQUET, O. and MENDELSON, S. (2005). Local Rademacher complexities. *Ann. Statist.* **33** 1497–1537. [MR2166554](#)
- [5] BARTLETT, P. L. and MENDELSON, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.* **3** 463–482. [MR1984026](#)
- [6] BARTLETT, P. L., MENDELSON, S. and PHILIPS, P. (2004). Local complexities for empirical risk minimization. In *Learning Theory. Lecture Notes in Computer Science* **3120** 270–284. Springer, Berlin. [MR2177915](#)
- [7] BELLEC, P. C. and TSYBAKOV, A. B. (2015). Sharp oracle bounds for monotone and convex regression through aggregation. *J. Mach. Learn. Res.* **16** 1879–1892. [MR3417801](#)

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

Key words and phrases. Posterior contraction, Rényi divergence, misspecified models, PAC-Bayes, oracle inequality, convex regression.

- [8] BHATTACHARYA, A., PATI, D. and YANG, Y. (2019). Supplement to “Bayesian fractional posteriors.” DOI:[10.1214/18-AOS1712SUPP](https://doi.org/10.1214/18-AOS1712SUPP).
- [9] BIRGÉ, L. (1984). Sur un théorème de minimax et son application aux tests. *Probab. Math. Statist.* **3** 259–282. [MR0764150](#)
- [10] BISSIRI, P. G., HOLMES, C. C. and WALKER, S. G. (2016). A general framework for updating belief distributions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 1103–1130. [MR3557191](#)
- [11] CATONI, O. (2007). *Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **56**. IMS, Beachwood, OH. [MR2483528](#)
- [12] CATONI, O. and PICARD, J. (2004). *Statistical Learning Theory and Stochastic Optimization: Ecole D’Eté de Probabilités de Saint-Flour, XXXI-2001*. Springer, Berlin.
- [13] CHATTERJEE, S. (2014). A new perspective on least squares under convex constraint. *Ann. Statist.* **42** 2340–2381. [MR3269982](#)
- [14] CHERNOZHUKOV, V. and HONG, H. (2003). An MCMC approach to classical estimation. *J. Econometrics* **115** 293–346. [MR1984779](#)
- [15] DALALYAN, A. and TSYBAKOV, A. B. (2008). Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Mach. Learn.* **72** 39–61.
- [16] DALALYAN, A. S. and SALMON, J. (2012). Sharp oracle inequalities for aggregation of affine estimators. *Ann. Statist.* **40** 2327–2355. [MR3059085](#)
- [17] DE BLASI, P. and WALKER, S. G. (2013). Bayesian asymptotics with misspecified models. *Statist. Sinica* **23** 169–187. [MR3076163](#)
- [18] FRIEL, N. and PETTITT, A. N. (2008). Marginal likelihood estimation via power posteriors. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 589–607. [MR2420416](#)
- [19] GELMAN, A. and MENG, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statist. Sci.* **13** 163–185. [MR1647507](#)
- [20] GERMAIN, P., LACASSE, A., LAVIOLETTE, F. and MARCHAND, M. (2009). PAC-Bayesian learning of linear classifiers. In *Proceedings of the 26th Annual International Conference on Machine Learning* 353–360. ACM, New York.
- [21] GEYER, C. J. and THOMPSON, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Amer. Statist. Assoc.* **90** 909–920.
- [22] GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. [MR1790007](#)
- [23] 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](#)
- [24] GHOSAL, S. and VAN DER VAART, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.* **35** 697–723. [MR2336864](#)
- [25] GRÜNWALD, P. (2012). The safe Bayesian: Learning the learning rate via the mixability gap. In *Algorithmic Learning Theory. Lecture Notes in Computer Science* **7568** 169–183. Springer, Heidelberg. [MR3042889](#)
- [26] GRÜNWALD, P. and VAN OMMEN, T. (2017). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Anal.* **12** 1069–1103. [MR3724979](#)
- [27] GUEDJ, B. and ALQUIER, P. (2013). PAC-Bayesian estimation and prediction in sparse additive models. *Electron. J. Stat.* **7** 264–291. [MR3020421](#)
- [28] GUNTUBOYINA, A. and SEN, B. (2013). Covering numbers for convex functions. *IEEE Trans. Inform. Theory* **59** 1957–1965. [MR3043776](#)
- [29] GUNTUBOYINA, A. and SEN, B. (2015). Global risk bounds and adaptation in univariate convex regression. *Probab. Theory Related Fields* **163** 379–411. [MR3405621](#)
- [30] HANNAH, L. A. and DUNSON, D. B. (2013). Multivariate convex regression with adaptive partitioning. *J. Mach. Learn. Res.* **14** 3261–3294. [MR3144462](#)

- [31] HOLMES, C. C. and WALKER, S. G. (2017). Assigning a value to a power likelihood in a general Bayesian model. *Biometrika* **104** 497–503. [MR3698270](#)
- [32] JIANG, W. and TANNER, M. A. (2008). Gibbs posterior for variable selection in high-dimensional classification and data mining. *Ann. Statist.* **36** 2207–2231. [MR2458185](#)
- [33] KLEIJN, B. J. K. and VAN DER VAART, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist.* **34** 837–877. [MR2283395](#)
- [34] KOLTCHINSKII, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.* **34** 2593–2656. [MR2329442](#)
- [35] KOLTCHINSKII, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems. Lecture Notes in Math.* **2033**. Springer, Heidelberg. [MR2829871](#)
- [36] KOLTCHINSKII, V. and PANCHENKO, D. (2000). Rademacher processes and bounding the risk of function learning. In *High Dimensional Probability, II (Seattle, WA, 1999). Progress in Probability* **47** 443–457. Birkhäuser, Boston, MA. [MR1857339](#)
- [37] KRUIJER, W., ROUSSEAU, J. and VAN DER VAART, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electron. J. Stat.* **4** 1225–1257. [MR2735885](#)
- [38] LE CAM, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, New York. [MR0856411](#)
- [39] LECAM, L. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** 38–53. [MR0334381](#)
- [40] LEUNG, G. and BARRON, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory* **52** 3396–3410. [MR2242356](#)
- [41] LUGOSI, G. and WEGKAMP, M. (2004). Complexity regularization via localized random penalties. *Ann. Statist.* **32** 1679–1697. [MR2089138](#)
- [42] MARTIN, R., MESS, R. and WALKER, S. G. (2017). Empirical Bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli* **23** 1822–1847. [MR3624879](#)
- [43] MARTIN, R. and WALKER, S. G. (2014). Asymptotically minimax empirical Bayes estimation of a sparse normal mean vector. *Electron. J. Stat.* **8** 2188–2206. [MR3273623](#)
- [44] MARTIN, R. and WALKER, S. G. (2016). Optimal Bayesian posterior concentration rates with empirical priors. Preprint. Available at [arXiv:1604.05734](#).
- [45] MCALLESTER, D. A. (1998). Some PAC-Bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory (Madison, WI, 1998)* 230–234. ACM, New York. [MR1811587](#)
- [46] MILLER, J. W. and DUNSON, D. B. (2015). Robust Bayesian inference via coarsening. Preprint. Available at [arXiv:1506.06101](#).
- [47] MILLER, J. W. and HARRISON, M. T. (2018). Mixture models with a prior on the number of components. *J. Amer. Statist. Assoc.* **113** 340–356. [MR3803469](#)
- [48] O’HAGAN, A. (1995). Fractional Bayes factors for model comparison. *J. Roy. Statist. Soc. Ser. B* **57** 99–138. [MR1325379](#)
- [49] RAKHLIN, A., SRIDHARAN, K. and TSYBAKOV, A. B. (2017). Empirical entropy, minimax regret and minimax risk. *Bernoulli* **23** 789–824. [MR3606751](#)
- [50] RAMAMOORTHY, R. V., SRIRAM, K. and MARTIN, R. (2015). On posterior concentration in misspecified models. *Bayesian Anal.* **10** 759–789. [MR3432239](#)
- [51] ROCKAFELLAR, R. T. (1997). *Convex Analysis*. Princeton Univ. Press, Princeton, NJ. [MR1451876](#)
- [52] SHAWE-TAYLOR, J. and WILLIAMSON, R. C. (1997). A PAC analysis of a Bayesian estimator. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory* 2–9. ACM, New York.
- [53] SHEN, W., TOKDAR, S. T. and GHOSAL, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika* **100** 623–640. [MR3094441](#)
- [54] SHEN, X. and WASSERMAN, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist.* **29** 687–714. [MR1865337](#)

- [55] STEPHENS, M. (2016). False discovery rates: A new deal. *Biostatistics* **18** 275–294.
- [56] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.* **37** 2655–2675. [MR2541442](#)
- [57] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York. [MR1385671](#)
- [58] VAN ERVEN, T. and HARREMOËS, P. (2014). Rényi divergence and Kullback–Leibler divergence. *IEEE Trans. Inform. Theory* **60** 3797–3820. [MR3225930](#)
- [59] VAN DE GEER, S. (2000). *Empirical Processes in M-Estimation*. Cambridge Univ. Press, Cambridge.
- [60] VAPNIK, V. N. and CHERVONENKIS, A. J. (1974). Theory of pattern recognition.
- [61] WALKER, S. and HJORT, N. L. (2001). On Bayesian consistency. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 811–821. [MR1872068](#)
- [62] ZHANG, T. (2006). From ϵ -entropy to KL-entropy: Analysis of minimum information complexity density estimation. *Ann. Statist.* **34** 2180–2210. [MR2291497](#)

DISTRIBUTION THEORY FOR HIERARCHICAL PROCESSES

BY FEDERICO CAMERLENGHI^{*,1}, ANTONIO LIJOI^{†,1,2}, PETER ORBANZ^{‡,3}
 AND IGOR PRÜNSTER^{†,2}

University of Milano–Bicocca^{*}, *Bocconi University*[†] and *Columbia University*[‡]

Hierarchies of discrete probability measures are remarkably popular as nonparametric priors in applications, arguably due to two key properties: (i) they naturally represent multiple heterogeneous populations; (ii) they produce ties across populations, resulting in a shrinkage property often described as “sharing of information.” In this paper, we establish a distribution theory for hierarchical random measures that are generated via normalization, thus encompassing both the hierarchical Dirichlet and hierarchical Pitman–Yor processes. These results provide a probabilistic characterization of the induced (partially exchangeable) partition structure, including the distribution and the asymptotics of the number of partition sets, and a complete posterior characterization. They are obtained by representing hierarchical processes in terms of completely random measures, and by applying a novel technique for deriving the associated distributions. Moreover, they also serve as building blocks for new simulation algorithms, and we derive marginal and conditional algorithms for Bayesian inference.

REFERENCES

- [1] BLEI, D. M., NG, A. Y. and JORDAN, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3** 993–1022.
- [2] CAMERLENGHI, F., LIJOI, A., ORBANZ, P. and PRÜNSTER, I. (2019). Distribution theory for hierarchical processes: Supplementary material. DOI:[10.1214/17-AOS1678SUPP](https://doi.org/10.1214/17-AOS1678SUPP).
- [3] CAMERLENGHI, F., LIJOI, A. and PRÜNSTER, I. (2017). Bayesian prediction with multiple-samples information. *J. Multivariate Anal.* **156** 18–28. [MR3624682](#)
- [4] CARON, F., TEH, Y. W. and MURPHY, T. B. (2014). Bayesian nonparametric Plackett–Luce models for the analysis of preferences for college degree programmes. *Ann. Appl. Stat.* **8** 1145–1181.
- [5] DALEY, D. J. and VERE-JONES, D. (2008). *An Introduction to the Theory of Point Processes. Vol. II*, 2nd ed. Springer, New York. [MR2371524](#)
- [6] DE BLASI, P., FAVARO, S., LIJOI, A., MENA, R. H., RUGGIERO, M. and PRÜNSTER, I. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Trans. Pattern Anal. Mach. Intell.* **37** 212–229.
- [7] DUNSON, D. B. (2010). Nonparametric Bayes applications to biostatistics. In *Bayesian Nonparametrics* 223–273. Cambridge Univ. Press, Cambridge.
- [8] FAVARO, S., LIJOI, A., MENA, R. H. and PRÜNSTER, I. (2009). Bayesian nonparametric inference for species variety with a two parameter Poisson–Dirichlet process prior. *J. Roy. Statist. Soc. Ser. B* **71** 993–1008.

MSC2010 subject classifications. 60G57, 62G05, 62F15.

Key words and phrases. Bayesian nonparametrics, distribution theory, hierarchical processes, partition structure, posterior distribution, prediction, random measures, species sampling models.

- [9] FERGUSON, T. S. and KLASS, M. J. (1972). A representation of independent increment processes without Gaussian components. *Ann. Math. Stat.* **43** 1634–1643.
- [10] FOTI, N. J. and WILLIAMSON, S. A. (2015). A survey of non-exchangeable priors for Bayesian nonparametric models. *IEEE Trans. Pattern Anal. Mach. Intell.* **37** 359–371.
- [11] GASTHAUS, J. and TEH, Y. W. (2010). Improvements to the sequence memoizer. *Adv. Neural Inf. Process. Syst.* **23**.
- [12] GNEDIN, A. V. and PITMAN, J. (2005). Exchangeable Gibbs partitions and Stirling triangles. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)* **325** 83–102.
- [13] GRIFFIN, J. E., KOLOSSIATIS, M. and STEEL, M. F. J. (2013). Comparing distributions by using dependent normalized random-measure mixtures. *J. Roy. Statist. Soc. Ser. B* **75** 499–529.
- [14] JAMES, L. F. (2005). Bayesian Poisson process partition calculus with an application to Bayesian Lévy moving averages. *Ann. Statist.* **33** 1771–1799. [MR2166562](#)
- [15] JAMES, L. F. (2006). Poisson calculus for spatial neutral to the right processes. *Ann. Statist.* **34** 416–440. [MR2275248](#)
- [16] JAMES, L. F., LIJOI, A. and PRÜNSTER, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scand. J. Stat.* **36** 76–97.
- [17] KALLENBERG, O. (2005). *Probabilistic Symmetries and Invariance Principles*. Springer, New York.
- [18] KINGMAN, J. F. C. (1967). Completely random measures. *Pacific J. Math.* **21** 59–78.
- [19] KINGMAN, J. F. C. (1982). The coalescent. *Stochastic Process. Appl.* **13** 235–248.
- [20] KINGMAN, J. F. C. (1993). *Poisson Processes*. Oxford Univ. Press, Oxford.
- [21] LIJOI, A., MENA, R. H. and PRÜNSTER, I. (2007). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika* **94** 769–786. [MR2416792](#)
- [22] LIJOI, A., MENA, R. H. and PRÜNSTER, I. (2007). Controlling the reinforcement in Bayesian nonparametric mixture models. *J. Roy. Statist. Soc. Ser. B* **69** 715–740.
- [23] LIJOI, A., NIPOTI, B. and PRÜNSTER, I. (2014). Bayesian inference with dependent normalized completely random measures. *Bernoulli* **20** 1260–1291. [MR3217444](#)
- [24] LIJOI, A., NIPOTI, B. and PRÜNSTER, I. (2014). Dependent mixture models: Clustering and borrowing information. *Comput. Statist. Data Anal.* **71** 417–433. [MR3131980](#)
- [25] LIJOI, A., PRÜNSTER, I. and WALKER, S. G. (2008). Bayesian nonparametric estimators derived from conditional Gibbs structures. *Ann. Appl. Probab.* **18** 1519–1547. [MR2434179](#)
- [26] MACÉACHERN, S. N. (1999). Dependent nonparametric processes. In *ASA Proceedings of the SBSS* 50–55. Amer. Statist. Assoc., Alexandria, VA.
- [27] MACÉACHERN, S. N. (2000). Dependent Dirichlet processes. Technical Report, Dept. Statistics, Ohio State Univ., Columbus, OH.
- [28] NGUYEN, X. (2010). Inference of global clusters from locally distributed data. *Bayesian Anal.* **5** 817–846.
- [29] NGUYEN, X. (2016). Borrowing strength in hierarchical Bayes: Convergence of the Dirichlet base measure. *Bernoulli* **22** 1535–1571.
- [30] PITMAN, J. (1996). Some developments of the Blackwell–MacQueen urn scheme. In *Statistics, Probability and Game Theory* (T. S. Ferguson, L. S. Shapley and J. B. MacQueen, eds.). *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **30** 245–267. IMS, Hayward, CA.
- [31] PITMAN, J. (2006). *Combinatorial Stochastic Processes. Ecole d’Été de Probabilités de Saint-Flour XXXII. Lecture Notes in Mathematics N. 1875*. Springer, Berlin. [MR2245368](#)
- [32] PITMAN, J. and YOR, M. (1997). The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25** 855–900. [MR1434129](#)
- [33] REGAZZINI, E., LIJOI, A. and PRÜNSTER, I. (2003). Distributional results for means of normalized random measures with independent increments. *Ann. Statist.* **31** 560–585. [MR1983542](#)

- [34] TEH, Y. W. and JORDAN, M. I. (2010). Hierarchical Bayesian nonparametric models with applications. In *Bayesian Nonparametrics* 158–207. Cambridge Univ. Press, Cambridge.
- [35] TEH, Y. W., JORDAN, M. I., BEAL, M. J. and BLEI, D. M. (2006). Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.* **101** 1566–1581.
- [36] WALKER, S. and DAMIEN, P. (2000). Representations of Lévy processes without Gaussian components. *Biometrika* **87** 477–483.
- [37] WOOD, F., GASTHAUS, J., ARCHAMBEAU, C., JAMES, L. F. and TEH, Y. W. (2011). The sequence memoizer. *Commun. ACM* **54** 91–98.
- [38] ZHOU, M. and CARIN, L. (2015). Negative binomial process count and mixture modeling. *IEEE Trans. Pattern Anal. Mach. Intell.* **37** 307–320.

ADAPTIVE ESTIMATION OF THE SPARSITY IN THE GAUSSIAN VECTOR MODEL

BY ALEXANDRA CARPENTIER¹ AND NICOLAS VERZELEN

OvGU Magdeburg and INRA

Consider the Gaussian vector model with mean value θ . We study the twin problems of estimating the number $\|\theta\|_0$ of nonzero components of θ and testing whether $\|\theta\|_0$ is smaller than some value. For testing, we establish the minimax separation distances for this model and introduce a minimax adaptive test. Extensions to the case of unknown variance are also discussed. Rewriting the estimation of $\|\theta\|_0$ as a multiple testing problem of all hypotheses $\{\|\theta\|_0 \leq q\}$, we both derive a new way of assessing the optimality of a sparsity estimator and we exhibit such an optimal procedure. This general approach provides a roadmap for estimating the complexity of the signal in various statistical models.

REFERENCES

- [1] ARIAS-CASTRO, E., CANDÈS, E. J. and PLAN, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *Ann. Statist.* **39** 2533–2556. [MR2906877](#)
- [2] BARAUD, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli* **8** 577–606. [MR1935648](#)
- [3] BARAUD, Y., HUET, S. and LAURENT, B. (2005). Testing convex hypotheses on the mean of a Gaussian vector. Application to testing qualitative hypotheses on a regression function. *Ann. Statist.* **33** 214–257. [MR2157802](#)
- [4] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Heidelberg. [MR2807761](#)
- [5] CAI, T. T. and GUO, Z. (2016). Accuracy assessment for high-dimensional linear regression. Preprint. Available at [arXiv:1603.03474](#).
- [6] CAI, T. T. and GUO, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Ann. Statist.* **45** 615–646. [MR3650395](#)
- [7] CAI, T. T. and JIN, J. (2010). Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. *Ann. Statist.* **38** 100–145. [MR2589318](#)
- [8] CAI, T. T., JIN, J. and LOW, M. G. (2007). Estimation and confidence sets for sparse normal mixtures. *Ann. Statist.* **35** 2421–2449. [MR2382653](#)
- [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 62C20, 62G10; secondary 62B10.

Key words and phrases. Sparsity estimation and testing, composite-composite testing problems, minimax separation distance in testing problems.

- [11] CAI, T. T. and LOW, M. G. (2011). Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *Ann. Statist.* **39** 1012–1041. [MR2816346](#)
- [12] CARPENTIER, A. (2015). Testing the regularity of a smooth signal. *Bernoulli* **21** 465–488.
- [13] CARPENTIER, A. and VERZELEN, N. (2019). Supplement to “Adaptive estimation of the sparsity in the Gaussian vector model.” DOI:[10.1214/17-AOS1680SUPP](#).
- [14] CELISSE, A. and ROBIN, S. (2010). A cross-validation based estimation of the proportion of true null hypotheses. *J. Statist. Plann. Inference* **140** 3132–3147.
- [15] COLLIER, O., COMMINGES, L. and TSYBAKOV, A. B. (2017). Minimax estimation of linear and quadratic functionals on sparsity classes. *Ann. Statist.* **45** 923–958. [MR3662444](#)
- [16] COMMINGES, L. and DALALYAN, A. S. (2013). Minimax testing of a composite null hypothesis defined via a quadratic functional in the model of regression. *Electron. J. Stat.* **7** 146–190. [MR3020417](#)
- [17] DONOHO, D. and JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994.
- [18] FROMONT, M., LERASLE, M. and REYNAUD-BOURET, P. (2016). Family-wise separation rates for multiple testing. *Ann. Statist.* **44** 2533–2563. [MR3576553](#)
- [19] GAYRAUD, G. and POUET, C. (2005). Adaptive minimax testing in the discrete regression scheme. *Probab. Theory Related Fields* **133** 531–558.
- [20] GINÉ, E. and NICKL, R. (2015). *Mathematical Foundations of Infinite-Dimensional Statistical Models* **40**. Cambridge Univ. Press, Cambridge.
- [21] HÄRDLE, W., KERKYACHARIAN, G., PICARD, D. and TSYBAKOV, A. (2012). *Wavelets, Approximation, and Statistical Applications* **129**. Springer, New York. [MR1618204](#)
- [22] HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*. Springer, New York.
- [23] HOFFMANN, M. and NICKL, R. (2011). On adaptive inference and confidence bands. *Ann. Statist.* **39** 2383–2409. [MR2906872](#)
- [24] INGSTER, Y., TSYBAKOV, A. and VERZELEN, N. (2010). Detection boundary in sparse regression. *Electron. J. Stat.* **4** 1476–1526.
- [25] INGSTER, YU. and SUSLINA, I. A. (2003). *Nonparametric Goodness-of-Fit Testing Under Gaussian Models. Lecture Notes in Statistics* **169**. Springer, New York. [MR1991446](#)
- [26] JIN, J. (2008). Proportion of non-zero normal means: Universal oracle equivalences and uniformly consistent estimators. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 461–493.
- [27] JIN, J. and TONY CAI, T. (2007). Estimating the null and the proportional of nonnull effects in large-scale multiple comparisons. *J. Amer. Statist. Assoc.* **102** 495–506.
- [28] JUDITSKY, A. and NEMIROVSKI, A. (2002). On nonparametric tests of positivity/monotonicity/convexity. *Ann. Statist.* **30** 498–527. [MR1902897](#)
- [29] KALAI, A. T., MOITRA, A. and VALIANT, G. (2012). Disentangling Gaussians. *Commun. ACM* **55** 113–120.
- [30] KESHAVAN, R. H., MONTANARI, A. and OH, S. (2010). Matrix completion from a few entries. *IEEE Trans. Inform. Theory* **56** 2980–2998. [MR2683452](#)
- [31] LANGAAS, M., LINDQVIST, B. H. and FERKINGSTAD, E. (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 555–572.
- [32] LEPSKI, O., NEMIROVSKI, A. and SPOKOINY, V. (1999). On estimation of the L_r norm of a regression function. *Probab. Theory Related Fields* **113** 221–253.
- [33] LI, J. and SIEGMUND, D. (2015). Higher criticism: p -values and criticism. *Ann. Statist.* **43** 1323–1350. [MR3346705](#)
- [34] MAHER, B. (2008). Personal genomes: The case of the missing heritability. *Nature* **456** 18–21.
- [35] MASSART, P. (2007). *Concentration Inequalities and Model Selection. Lecture Notes in Math.* **1896**. Springer, Berlin. [MR2319879](#)

- [36] MEINSHAUSEN, N. and RICE, J. (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist.* **34** 373–393. [MR2275246](#)
- [37] MOSCOVICH, A., NADLER, B. and SPIEGELMAN, C. (2016). On the exact Berk–Jones statistics and their p -value calculation. *Electron. J. Stat.* **10** 2329–2354. [MR3544289](#)
- [38] NICKL, R. and VAN DE GEER, S. (2013). Confidence sets in sparse regression. *Ann. Statist.* **41** 2852–2876. [MR3161450](#)
- [39] PATRA, R. K. and SEN, B. (2015). Estimation of a two-component mixture model with applications to multiple testing. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 869–893. [MR3534354](#)
- [40] STOREY, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 479–498.
- [41] TORO, R. et al. (2015). Genomic architecture of human neuroanatomical diversity. *Mol. Psychiatry* **20** 1011–1016.
- [42] VERZELEN, N. (2012). Minimax risks for sparse regressions: Ultra-high dimensional phenomena. *Electron. J. Stat.* **6** 38–90.

APPROXIMATE OPTIMAL DESIGNS FOR MULTIVARIATE POLYNOMIAL REGRESSION

BY YOHANN DE CASTRO^{*}, FABRICE GAMBOA[†], DIDIER HENRION^{‡,1},
 ROXANA HESS^{‡,1} AND JEAN-BERNARD LASSERRE^{‡,1}

Université Paris Sud and INRIA^{}, Université Paul Sabatier[†]
 and Université de Toulouse[‡]*

We introduce a new approach aiming at computing approximate optimal designs for multivariate polynomial regressions on compact (semialgebraic) design spaces. We use the moment-sum-of-squares hierarchy of semidefinite programming problems to solve numerically the approximate optimal design problem. The geometry of the design is recovered via semidefinite programming duality theory. This article shows that the hierarchy converges to the approximate optimal design as the order of the hierarchy increases. Furthermore, we provide a dual certificate ensuring finite convergence of the hierarchy and showing that the approximate optimal design can be computed numerically with our method. As a byproduct, we revisit the equivalence theorem of the experimental design theory: it is linked to the Christoffel polynomial and it characterizes finite convergence of the moment-sum-of-square hierarchies.

REFERENCES

- [1] BOX, G. E. P. and HUNTER, J. S. (1957). Multi-factor experimental designs for exploring response surfaces. *Ann. Math. Stat.* **28** 195–241. [MR0085679](#)
- [2] DE CASTRO, Y., GAMBOA, F., HENRION, D., HESS, R. and LASSERRE, J.-B. (2019). Supplement to “Approximate optimal designs for multivariate polynomial regression.” DOI:10.1214/18-AOS1683SUPP.
- [3] DETTE, H. and STUDDEN, W. J. (1997). *The Theory of Canonical Moments with Applications in Statistics, Probability, and Analysis*. Wiley, New York. [MR1468473](#)
- [4] GAFFKE, N., GRASSHOFF, U. and SCHWABE, R. (2014). Algorithms for approximate linear regression design with application to a first order model with heteroscedasticity. *Comput. Statist. Data Anal.* **71** 1113–1123. [MR3132032](#)
- [5] HENRION, D., LASSERRE, J.-B. and LÖFBERG, J. (2009). GloptiPoly 3: Moments, optimization and semidefinite programming. *Optim. Methods Softw.* **24** 761–779. [MR2554910](#)
- [6] KIEFER, J. (1974). General equivalence theory for optimum designs (approximate theory). *Ann. Statist.* **2** 849–879. [MR0356386](#)
- [7] KREIN, M. G. and NUDEL’MAN, A. A. (1977). *The Markov Moment Problem and Extremal Problems*. Amer. Math. Soc., Providence, RI. [MR0458081](#)
- [8] LASSERRE, J.-B. and PAUWELS, É. (2016). Sorting out typicality with the inverse moment matrix SOS polynomial. In *Advances in Neural Information Processing Systems* **29**.

MSC2010 subject classifications. Primary 62K05, 90C25; secondary 41A10, 49M29, 90C90, 15A15.

Key words and phrases. Experimental design, semidefinite programming, Christoffel polynomial, linear model, equivalence theorem.

- [9] LASSERRE, J. B. (2010). *Moments, Positive Polynomials and Their Applications. Imperial College Press Optimization Series 1*. Imperial College Press, London. [MR2589247](#)
- [10] LASSERRE, J. B. (2015). A generalization of Löwner–John’s ellipsoid theorem. *Math. Program.* **152** 559–591. [MR3369493](#)
- [11] LASSERRE, J. B. and NETZER, T. (2007). SOS approximations of nonnegative polynomials via simple high degree perturbations. *Math. Z.* **256** 99–112. [MR2282261](#)
- [12] LEDOUX, M. (2004). Differential operators and spectral distributions of invariant ensembles from the classical orthogonal polynomials. The continuous case. *Electron. J. Probab.* **9** 177–208. [MR2041832](#)
- [13] LEWIS, A. S. (1996). Convex analysis on the Hermitian matrices. *SIAM J. Optim.* **6** 164–177. [MR1377729](#)
- [14] LOFBERG, J. (2004). Yalmip: A toolbox for modeling and optimization in Matlab. In *Computer Aided Control Systems Design, 2004 IEEE International Symposium on* 284–289. IEEE, New York.
- [15] MOLCHANOV, I. and ZUYEV, S. (2004). Optimisation in space of measures and optimal design. *ESAIM Probab. Stat.* **8** 12–24. [MR2085602](#)
- [16] MOSEK, A. (2015). The MOSEK optimization toolbox for matlab manual. Version 7.1 (Revision 28).
- [17] NIE, J. (2014). The $\mathcal{A}$ -truncated K -moment problem. *Found. Comput. Math.* **14** 1243–1276. [MR3273678](#)
- [18] PAPP, D. (2012). Optimal designs for rational function regression. *J. Amer. Statist. Assoc.* **107** 400–411. [MR2949369](#)
- [19] PUKELSHEIM, F. (2006). *Optimal Design of Experiments. Classics in Applied Mathematics 50*. SIAM, Philadelphia, PA. [MR2224698](#)
- [20] SAGNOL, G. (2013). On the semidefinite representation of real functions applied to symmetric matrices. *Linear Algebra Appl.* **439** 2829–2843. [MR3116395](#)
- [21] SAGNOL, G. and HARMAN, R. (2015). Computing exact D -optimal designs by mixed integer second-order cone programming. *Ann. Statist.* **43** 2198–2224. [MR3396983](#)
- [22] SCHEIDERER, C. (2016). Semidefinitely representable convex sets. ArXiv e-prints.
- [23] TORSNEY, B. (2009). W-iterations and ripples therefrom. In *Optimal Design and Related Areas in Optimization and Statistics. Springer Optim. Appl.* **28** 1–12. Springer, New York. [MR2513344](#)
- [24] VANDENBERGHE, L., BOYD, S. and WU, S.-P. (1998). Determinant maximization with linear matrix inequality constraints. *SIAM J. Matrix Anal. Appl.* **19** 499–533. [MR1614078](#)

EFFICIENT ESTIMATION OF INTEGRATED VOLATILITY FUNCTIONALS VIA MULTISCALE JACKKNIFE

BY JIA LI¹, YUNXIAO LIU AND DACHENG XIU²

*Duke University, University of North Carolina at Chapel Hill and
University of Chicago*

We propose semiparametrically efficient estimators for general integrated volatility functionals of multivariate semimartingale processes. A plug-in method that uses nonparametric estimates of spot volatilities is known to induce high-order biases that need to be corrected to obey a central limit theorem. Such bias terms arise from boundary effects, the diffusive and jump movements of stochastic volatility and the sampling error from the nonparametric spot volatility estimation. We propose a novel jackknife method for bias correction. The jackknife estimator is simply formed as a linear combination of a few uncorrected estimators associated with different local window sizes used in the estimation of spot volatility. We show theoretically that our estimator is asymptotically mixed Gaussian, semiparametrically efficient, and more robust to the choice of local windows. To facilitate the practical use, we introduce a simulation-based estimator of the asymptotic variance, so that our inference is derivative-free, and hence is convenient to implement.

REFERENCES

- [1] AÏT-SAHALIA, Y., FAN, J. and XIU, D. (2010). High-frequency covariance estimates with noisy and asynchronous financial data. *J. Amer. Statist. Assoc.* **105** 1504–1517. [MR2796567](#)
- [2] AÏT-SAHALIA, Y. and XIU, D. (2015). Principal component analysis of high frequency data. Technical report.
- [3] ANDERSEN, T. G., BOLLERSLEV, T., DIEBOLD, F. X. and LABYS, P. (2003). Modeling and forecasting realized volatility. *Econometrica* **71** 579–625. [MR1958138](#)
- [4] BARNDORFF-NIELSEN, O. E. and SHEPHARD, N. (2004). Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics. *Econometrica* **72** 885–925. [MR2051439](#)
- [5] BARNDORFF-NIELSEN, O. E., HANSEN, P. R., LUNDE, A. and SHEPHARD, N. (2008). Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica* **76** 1481–1536. [MR2468558](#)
- [6] BARNDORFF-NIELSEN, O. E., HANSEN, P. R., LUNDE, A. and SHEPHARD, N. (2011). Multivariate realised kernels: Consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. *J. Econometrics* **162** 149–169. [MR2795610](#)
- [7] BIBINGER, M., HAUTSCH, N., MALEC, P. and REISS, M. (2014). Estimating the quadratic covariation matrix from noisy observations: Local method of moments and efficiency. *Ann. Statist.* **42** 1312–1346. [MR3226158](#)

MSC2010 subject classifications. 60F05, 60G44, 62F12.

Key words and phrases. High-frequency data, jackknife, semimartingale, spot volatility.

- [8] CARR, P., GEMAN, H., MADAN, D. B. and YOR, M. (2002). The fine structure of asset returns: An empirical investigation. *J. Bus.* **75** 305–332.
- [9] CLÉMENT, E., DELATTRE, S. and GLOTER, A. (2013). An infinite dimensional convolution theorem with applications to the efficient estimation of the integrated volatility. *Stochastic Process. Appl.* **123** 2500–2521. [MR3054534](#)
- [10] DUFFIE, D., PAN, J. and SINGLETON, K. (2000). Transform analysis and asset pricing for affine jump-diffusions. *Econometrica* **68** 1343–1376. [MR1793362](#)
- [11] EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. CBMS–NSF Regional Conference Series in Applied Mathematics **38**. SIAM, Philadelphia, PA. [MR0659849](#)
- [12] HAYASHI, T. and YOSHIDA, N. (2005). On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli* **11** 359–379. [MR2132731](#)
- [13] JACOD, J. and PROTTER, P. (2012). *Discretization of Processes*. *Stochastic Modelling and Applied Probability* **67**. Springer, Heidelberg. [MR2859096](#)
- [14] JACOD, J. and ROSENBAUM, M. (2013). Quarticity and other functionals of volatility: Efficient estimation. *Ann. Statist.* **41** 1462–1484. [MR3113818](#)
- [15] JACOD, J. and ROSENBAUM, M. (2015). Estimation of volatility functionals: The case of a $\sqrt{n}$ window. In *Large Deviations and Asymptotic Methods in Finance* (P. Friz, J. Gatheral, A. Gulisashvili, A. Jacquier and J. Teichmann, eds.). *Springer Proc. Math. Stat.* **110** 559–590. Springer, Cham. [MR3375197](#)
- [16] JACOD, J., LI, Y., MYKLAND, P. A., PODOLSKIJ, M. and VETTER, M. (2009). Microstructure noise in the continuous case: The pre-averaging approach. *Stochastic Process. Appl.* **119** 2249–2276. [MR2531091](#)
- [17] KALNINA, I. and XIU, D. (2017). Nonparametric estimation of the leverage effect: A trade-off between robustness and efficiency. *J. Amer. Statist. Assoc.* **112** 384–396. [MR3646579](#)
- [18] KRISTENSEN, D. (2010). Nonparametric filtering of the realized spot volatility: A kernel-based approach. *Econometric Theory* **26** 60–93. [MR2587103](#)
- [19] LI, J., LIU, Y. and XIU, D. (2019). Supplement to “Efficient estimation of integrated volatility functionals via multiscale jackknife.” DOI:[10.1214/18-AOS1684SUPP](#).
- [20] LI, J., TODOROV, V. and TAUCHEN, G. (2016). Inference theory for volatility functional dependencies. *J. Econometrics* **193** 17–34. [MR3500174](#)
- [21] LI, J., TODOROV, V. and TAUCHEN, G. (2017). Adaptive estimation of continuous-time regression models using high-frequency data. *J. Econometrics* **200** 36–47. [MR3683660](#)
- [22] LI, J. and XIU, D. (2016). Generalized method of integrated moments for high-frequency data. *Econometrica* **84** 1613–1633. [MR3537165](#)
- [23] MYKLAND, P. A. and ZHANG, L. (2017). Assessment of uncertainty in high frequency data: The observed asymptotic variance. *Econometrica* **85** 197–231. [MR3611770](#)
- [24] REISS, M. (2011). Asymptotic equivalence for inference on the volatility from noisy observations. *Ann. Statist.* **39** 772–802. [MR2816338](#)
- [25] RENAULT, E., SARISOY, C. and WERKER, B. J. M. (2017). Efficient estimation of integrated volatility and related processes. *Econometric Theory* **33** 439–478. [MR3600049](#)
- [26] SHEPHARD, N. and XIU, D. (2017). Econometric analysis of multivariate realised QML: Estimation of the covariation of equity prices under asynchronous trading. *J. Econometrics* **201** 19–42. [MR3704675](#)
- [27] ZHANG, L. (2011). Estimating covariation: Epps effect, microstructure noise. *J. Econometrics* **160** 33–47. [MR2745865](#)
- [28] ZHANG, L., MYKLAND, P. A. and AÏT-SAHALIA, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *J. Amer. Statist. Assoc.* **100** 1394–1411. [MR2236450](#)

NONASYMPTOTIC RATES FOR MANIFOLD, TANGENT SPACE AND CURVATURE ESTIMATION

BY EDDIE AAMARI^{1,2,3} AND CLÉMENT LEVRARD^{1,2}

University of California, San Diego and Université Paris-Diderot

Given a noisy sample from a submanifold $M \subset \mathbb{R}^D$, we derive optimal rates for the estimation of tangent spaces $T_X M$, the second fundamental form Π_X^M and the submanifold M . After motivating their study, we introduce a quantitative class of C^k -submanifolds in analogy with Hölder classes. The proposed estimators are based on local polynomials and allow to deal simultaneously with the three problems at stake. Minimax lower bounds are derived using a conditional version of Assouad’s lemma when the base point X is random.

REFERENCES

- [1] AAMARI, E. and LEVRARD, C. (2015). Stability and minimax optimality of tangential delaunay complexes for manifold reconstruction. ArXiv E-prints, [arxiv:1512.02857](https://arxiv.org/abs/1512.02857).
- [2] AAMARI, E. and LEVRARD, C. (2019). Supplement to “Nonasymptotic rates for manifold, tangent space and curvature estimation.” DOI:[10.1214/18-AOS1685SUPP](https://doi.org/10.1214/18-AOS1685SUPP).
- [3] ARIAS-CASTRO, E., LERMAN, G. and ZHANG, T. (2017). Spectral clustering based on local PCA. *J. Mach. Learn. Res.* **18** Paper No. 9, 57. [MR3634876](https://arxiv.org/abs/1605.01333)
- [4] ARIAS-CASTRO, E., PATEIRO-LÓPEZ, B. and RODRÍGUEZ-CASAL, A. (2016). Minimax estimation of the volume of a set with smooth boundary. ArXiv E-prints, [arxiv:1605.01333](https://arxiv.org/abs/1605.01333).
- [5] BOISSONNAT, J.-D. and GHOSH, A. (2014). Manifold reconstruction using tangential Delaunay complexes. *Discrete Comput. Geom.* **51** 221–267. [MR3148657](https://arxiv.org/abs/1605.01333)
- [6] CAZALS, F. and POUGET, M. (2005). Estimating differential quantities using polynomial fitting of osculating jets. *Comput. Aided Geom. Design* **22** 121–146. [MR2116098](https://arxiv.org/abs/1605.01333)
- [7] CHENG, S.-W. and CHIU, M.-K. (2016). Tangent estimation from point samples. *Discrete Comput. Geom.* **56** 505–557. [MR3544007](https://arxiv.org/abs/1605.01333)
- [8] CHENG, S.-W., DEY, T. K. and RAMOS, E. A. (2005). Manifold reconstruction from point samples. In *Proceedings of the Sixteenth Annual ACM–SIAM Symposium on Discrete Algorithms* 1018–1027. ACM, New York. [MR2298361](https://arxiv.org/abs/1605.01333)
- [9] DO CARMO, M. P. (1992). *Riemannian Geometry*. Birkhäuser, Boston, MA. [MR1138207](https://arxiv.org/abs/1605.01333)
- [10] FASY, B. T., LECCI, F., RINALDO, A., WASSERMAN, L., BALAKRISHNAN, S. and SINGH, A. (2014). Confidence sets for persistence diagrams. *Ann. Statist.* **42** 2301–2339. [MR3269981](https://arxiv.org/abs/1605.01333)
- [11] FEDERER, H. (1959). Curvature measures. *Trans. Amer. Math. Soc.* **93** 418–491. [MR0110078](https://arxiv.org/abs/1605.01333)
- [12] FEFFERMAN, C., IVANOV, S., KURYLEV, Y., LASSAS, M. and NARAYANAN, H. (2015). Reconstruction and interpolation of manifolds I: The geometric Whitney problem. ArXiv E-prints, [arxiv:1508.00674](https://arxiv.org/abs/1508.00674).

MSC2010 subject classifications. 62G05, 62C20.

Key words and phrases. Geometric inference, minimax, manifold learning.

- [13] GASHLER, M. S. and MARTINEZ, T. (2011). Tangent space guided intelligent neighbor finding. In *Proceedings of the IEEE International Joint Conference on Neural Networks IJCNN'11* 2617–2624. IEEE Press.
- [14] GENOVESE, C. R., PERONE-PACIFICO, M., VERDINELLI, I. and WASSERMAN, L. (2012). Manifold estimation and singular deconvolution under Hausdorff loss. *Ann. Statist.* **40** 941–963. [MR2985939](#)
- [15] GENOVESE, C. R., PERONE-PACIFICO, M., VERDINELLI, I. and WASSERMAN, L. (2012). Minimax manifold estimation. *J. Mach. Learn. Res.* **13** 1263–1291. [MR2930639](#)
- [16] GOLUB, G. H. and VAN LOAN, C. F. (1996). *Matrix Computations*, 3rd ed. Johns Hopkins Univ. Press, Baltimore, MD. [MR1417720](#)
- [17] GUMHOLD, S., WANG, X. and MACLEOD, R. (2001). Feature extraction from point clouds. In *10th International Meshing Roundtable* 293–305. Sandia National Laboratories, Albuquerque, NM.
- [18] HARTMAN, P. (1951). On geodesic coordinates. *Amer. J. Math.* **73** 949–954. [MR0046087](#)
- [19] KIM, A. K. H. and ZHOU, H. H. (2015). Tight minimax rates for manifold estimation under Hausdorff loss. *Electron. J. Stat.* **9** 1562–1582. [MR3376117](#)
- [20] MAGGIONI, M., MINSKER, S. and STRAWN, N. (2016). Multiscale dictionary learning: Non-asymptotic bounds and robustness. *J. Mach. Learn. Res.* **17** Paper No. 2, 51. [MR3482922](#)
- [21] MERIGOT, Q., OVSIANKOV, M. and GUIBAS, L. J. (2011). Voronoi-based curvature and feature estimation from point clouds. *IEEE Trans. Vis. Comput. Graph.* **17** 743–756. DOI:[10.1109/TVCG.2010.261](#).
- [22] NIYOGI, P., SMALE, S. and WEINBERGER, S. (2008). Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.* **39** 419–441. [MR2383768](#)
- [23] ROWEIS, S. T. and SAUL, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* **290** 2323–2326.
- [24] RUSINKIEWICZ, S. (2004). Estimating curvatures and their derivatives on triangle meshes. In *Proceedings of the 3D Data Processing, Visualization, and Transmission, 2nd International Symposium. 3DPVT '04* 486–493. IEEE Computer Society, Washington, DC.
- [25] SINGER, A. and WU, H.-T. (2012). Vector diffusion maps and the connection Laplacian. *Comm. Pure Appl. Math.* **65** 1067–1144. [MR2928092](#)
- [26] TENENBAUM, J. B., DE SILVA, V. and LANGFORD, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* **290** 2319–2323.
- [27] TYAGI, H., VURAL, E. and FROSSARD, P. (2013). Tangent space estimation for smooth embeddings of Riemannian manifolds. *Inf. Inference* **2** 69–114. [MR3311444](#)
- [28] USEVICH, K. and MARKOVSKY, I. (2014). Optimization on a Grassmann manifold with application to system identification. *Automatica J. IFAC* **50** 1656–1662. [MR3214912](#)
- [29] WASSERMAN, L. (2016). Topological data analysis. ArXiv E-prints, [arxiv:1609.08227](#).
- [30] YU, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam* 423–435. Springer, New York. [MR1462963](#)

NONPARAMETRIC TESTING FOR MULTIPLE SURVIVAL FUNCTIONS WITH NONINFERIORITY MARGINS

BY HSIN-WEN CHANG¹ AND IAN W. McKEAGUE²

Academia Sinica and Columbia University

New nonparametric tests for the ordering of multiple survival functions are developed with the possibility of right censorship taken into account. The motivation comes from noninferiority trials with multiple treatments. The proposed tests are based on nonparametric likelihood ratio statistics, which are known to provide more powerful tests than Wald-type procedures, but in this setting have only been studied for pairs of survival functions or in the absence of censoring. We introduce a novel type of pool adjacent violator algorithm that leads to a complete solution of the problem. The limit distributions can be expressed as weighted sums of squares involving projections of certain Gaussian processes onto the given ordered alternative. A simulation study shows that the new procedures have superior power to a competing combined-pairwise Cox model approach. We illustrate the proposed methods using data from a three-arm noninferiority trial.

REFERENCES

- ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. and KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York. [MR1198884](#)
- BERGER, R. L. and HSU, J. C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statist. Sci.* **11** 283–302. [MR1445984](#)
- BORGAN, Ø. and LIESTØL, K. (1990). A note on confidence intervals and bands for the survival function based on transformations. *Scand. J. Stat.* **17** 35–41. [MR1062844](#)
- BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge Univ. Press, Cambridge. [MR2061575](#)
- CHANG, H., EL BARMÍ, H. and McKEAGUE, I. W. (2016). Tests for stochastic ordering under biased sampling. *J. Nonparametr. Stat.* **28** 659–682. [MR3555451](#)
- CHANG, H. and McKEAGUE, I. W. (2016). Empirical likelihood based tests for stochastic ordering under right censorship. *Electron. J. Stat.* **10** 2511–2536. [MR3545467](#)
- CHANG, H. and McKEAGUE, I. W. (2019). Supplement to “Nonparametric testing for multiple survival functions with noninferiority margins.” DOI:[10.1214/18-AOS1686SUPP](https://doi.org/10.1214/18-AOS1686SUPP).
- CHI, Y. (2002). Ordered tests for right-censored survival data. *Aust. N. Z. J. Stat.* **44** 367–380. [MR1919199](#)
- DICKMAN, P. W., SLOGGETT, A., HILLS, M. and HAKULINEN, T. (2004). Regression models for relative survival. *Stat. Med.* **23** 51–64.
- DIGNAM, J. J., HUANG, L., RIES, L., REICHMAN, M., MARIOTTO, A. and FEUER, E. (2009). Estimating breast cancer-specific and other-cause mortality in clinical trial and population-based cancer registry cohorts. *Cancer* **115** 5272–5283.

MSC2010 subject classifications. Primary 62G10; secondary 62G20, 62F30.

Key words and phrases. Empirical likelihood, equivalence test, order-restricted inference.

- DYKSTRA, R. L. and FELTZ, C. J. (1989). Nonparametric maximum likelihood estimation of survival functions with a general stochastic ordering and its dual. *Biometrika* **76** 331–341. [MR1016024](#)
- EL BARMÍ, H. and MCKEAGUE, I. W. (2013). Empirical likelihood-based tests for stochastic ordering. *Bernoulli* **19** 295–307. [MR3019496](#)
- EL BARMÍ, H. and MUKERJEE, H. (2005). Inferences under a stochastic ordering constraint: The k -sample case. *J. Amer. Statist. Assoc.* **100** 252–261. [MR2156835](#)
- FINNER, H. and STRASSBURGER, K. (2002). The partitioning principle: A powerful tool in multiple decision theory. *Ann. Statist.* **30** 1194–1213. [MR1926174](#)
- FREITAG, G., LANGE, S. and MUNK, A. (2006). Non-parametric assessment of non-inferiority with censored data. *Stat. Med.* **25** 1201–1217. [MR2225587](#)
- GEHAN, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* **52** 203–223. [MR0207130](#)
- GUYOT, P., ADES, A. E., OUWENS, M. J. N. M. and WELTON, N. J. (2012). Enhanced secondary analysis of survival data: Reconstructing the data from published Kaplan–Meier survival curves. *BMC Med. Res. Methodol.* **12** 1–13.
- HAUSCHKE, D. and PIGEOT, I. (2005). Establishing efficacy of a new experimental treatment in the ‘gold standard’ design. *Biom. J.* **47** 782–786. [MR2221912](#)
- HIDA, E. and TANGO, T. (2011). On the three-arm non-inferiority trial including a placebo with a prespecified margin. *Stat. Med.* **30** 224–231. [MR2758874](#)
- KITAMURA, Y., SANTOS, A. and SHAIKH, A. M. (2012). On the asymptotic optimality of empirical likelihood for testing moment restrictions. *Econometrica* **80** 413–423. [MR2920761](#)
- KOMBRINK, K., MUNK, A. and FRIEDE, T. (2013). Design and semiparametric analysis of non-inferiority trials with active and placebo control for censored time-to-event data. *Stat. Med.* **32** 3055–3066. [MR3073783](#)
- LI, G. (1995). On nonparametric likelihood ratio estimation of survival probabilities for censored data. *Statist. Probab. Lett.* **25** 95–104. [MR1365025](#)
- LIU, P. Y. and TSAI, W. Y. (1999). A modified logrank test for censored survival data under order restrictions. *Statist. Probab. Lett.* **41** 57–63. [MR1659237](#)
- MCDANIEL, L. S., YU, M. and CHAPPELL, R. (2016). Sample size under the additive hazards model. *Clin. Trials* **13** 188–198.
- MCKEAGUE, I. W. and ZHAO, Y. (2002). Simultaneous confidence bands for ratios of survival functions via empirical likelihood. *Statist. Probab. Lett.* **60** 405–415. [MR1947180](#)
- MIELKE, M., MUNK, A. and SCHACHT, A. (2008). The assessment of non-inferiority in a gold standard design with censored, exponentially distributed endpoints. *Stat. Med.* **27** 5093–5110. [MR2516746](#)
- MÜTZE, T., KONIETSCHKE, F., MUNK, A. and FRIEDE, T. (2017). A studentized permutation test for three-arm trials in the ‘gold standard’ design. *Stat. Med.* **36** 883–898. [MR3606651](#)
- OWEN, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75** 237–249. [MR0946049](#)
- OWEN, A. B. (2001). *Empirical Likelihood*. Chapman & Hall/CRC, Boca Raton, FL.
- PARZEN, M. I., WEI, L. J. and YING, Z. (1997). Simultaneous confidence intervals for the difference of two survival functions. *Scand. J. Stat.* **24** 309–314. [MR1481417](#)
- ROMANO, J. P. (2005). Optimal testing of equivalence hypotheses. *Ann. Statist.* **33** 1036–1047. [MR2195627](#)
- ROTHMANN, M. D., WIENS, B. L. and CHAN, I. S. F. (2011). *Design and Analysis of Non-inferiority Trials*. CRC Press, Boca Raton, FL. [MR2760027](#)
- THOMAS, D. R. and GRUNKEMEIER, G. L. (1975). Confidence interval estimation of survival probabilities for censored data. *J. Amer. Statist. Assoc.* **70** 865–871. [MR0405766](#)
- USMANI, R. A. (1994). Inversion of a tridiagonal Jacobi matrix. *Linear Algebra Appl.* **212/213** 413–414. [MR1306990](#)

- VAN DER VAART, A. W. (2000). *Asymptotic Statistics*. Cambridge Univ. Press, Cambridge. [MR1652247](#)
- WELLEK, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority*, 2nd ed. CRC Press, Boca Raton, FL. [MR2676002](#)
- WHITE, W. B., WEBER, M. A., SICA, D., BAKRIS, G. L., PEREZ, A., CAO, C. and KUPFER, S. (2011). Effects of the angiotensin receptor blocker azilsartan medoxomil versus olmesartan and valsartan on ambulatory and clinic blood pressure in patients with stages 1 and 2 hypertension. *Hypertension* **57** 413–420.

ORACLE INEQUALITIES AND ADAPTIVE ESTIMATION IN THE CONVOLUTION STRUCTURE DENSITY MODEL¹

BY O. V. LEPSKI AND T. WILLER

Aix-Marseille Université

We study the problem of nonparametric estimation under $\mathbb{L}_p$ -loss, $p \in [1, \infty)$, in the framework of the convolution structure density model on $\mathbb{R}^d$. This observation scheme is a generalization of two classical statistical models, namely density estimation under direct and indirect observations. The original pointwise selection rule from a family of “kernel-type” estimators is proposed. For the selected estimator, we prove an $\mathbb{L}_p$ -norm oracle inequality and several of its consequences. Next, the problem of adaptive minimax estimation under $\mathbb{L}_p$ -loss over the scale of anisotropic Nikol’skii classes is addressed. We fully characterize the behavior of the minimax risk for different relationships between regularity parameters and norm indexes in the definitions of the functional class and of the risk. We prove that the proposed selection rule leads to the construction of an optimally or nearly optimally (up to logarithmic factors) adaptive estimator.

REFERENCES

- AKAKPO, N. (2012). Adaptation to anisotropy and inhomogeneity via dyadic piecewise polynomial selection. *Math. Methods Statist.* **21** 1–28. [MR2901269](#)
- BIRGÉ, L. (2014). Model selection for density estimation with $\mathbb{L}_2$ -loss. *Probab. Theory Related Fields* **158** 533–574. [MR3176358](#)
- BUTUCEA, C. and TSYBAKOV, A. B. (2008). Sharp optimality in density deconvolution with dominating bias. I, II. *Theory Probab. Appl.* **52** 111–128, 237–249. [MR2354572](#)
- COMTE, F. and LACOUR, C. (2013). Anisotropic adaptive kernel deconvolution. *Ann. Inst. Henri Poincaré Probab. Stat.* **49** 569–609. [MR3088382](#)
- COMTE, F., ROZENHOLC, Y. and TAUPIN, M.-L. (2006). Penalized contrast estimator for adaptive density deconvolution. *Canad. J. Statist.* **34** 431–452. [MR2328553](#)
- DE GUZMÁN, M. (1975). *Differentiation of Integrals in $\mathbb{R}^n$* . *Lecture Notes in Mathematics* **481**. Springer, Berlin. [MR0457661](#)
- DEVROYE, L. and LUGOSI, G. (1997). Nonasymptotic universal smoothing factors, kernel complexity and Yatracos classes. *Ann. Statist.* **25** 2626–2637. [MR1604428](#)
- DONOHU, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.* **24** 508–539. [MR1394974](#)
- DUVAL, C. (2017). A note on a fixed-point method for deconvolution. *Statistics* **51** 347–362. [MR3609324](#)
- EFROIMOVICH, S. YU. (1986). Nonparametric estimation of a density of unknown smoothness. *Theory Probab. Appl.* **30** 557–568. [MR0805304](#)

MSC2010 subject classifications. 62G05, 62G20.

Key words and phrases. Deconvolution model, density estimation, oracle inequality, adaptive estimation, kernel estimators, $\mathbb{L}_p$ -risk, anisotropic Nikol’skii class.

- FAN, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.* **19** 1257–1272. [MR1126324](#)
- FAN, J. (1993). Adaptively local one-dimensional subproblems with application to a deconvolution problem. *Ann. Statist.* **21** 600–610. [MR1232507](#)
- FAN, J. and KOO, J.-Y. (2002). Wavelet deconvolution. *IEEE Trans. Inform. Theory* **48** 734–747. [MR1889978](#)
- FOLLAND, G. B. (1999). *Real Analysis: Modern Techniques and Their Applications*, 2nd ed. Wiley, New York. [MR1681462](#)
- GACH, F., NICKL, R. and SPOKOINY, V. (2013). Spatially adaptive density estimation by localised Haar projections. *Ann. Inst. Henri Poincaré Probab. Stat.* **49** 900–914. [MR3112439](#)
- GINÉ, E. and NICKL, R. (2009). An exponential inequality for the distribution function of the kernel density estimator, with applications to adaptive estimation. *Probab. Theory Related Fields* **143** 569–596. [MR2475673](#)
- GOLDENSHLUGER, A. and LEPSKI, O. (2011). Bandwidth selection in kernel density estimation: Oracle inequalities and adaptive minimax optimality. *Ann. Statist.* **39** 1608–1632. [MR2850214](#)
- GOLDENSHLUGER, A. and LEPSKI, O. (2014). On adaptive minimax density estimation on R^d . *Probab. Theory Related Fields* **159** 479–543. [MR3230001](#)
- GOLUBEV, G. K. (1992). Non-parametric estimation of smooth probability densities. *Probl. Inf. Transm.* **1** 52–62.
- GRAFAKOS, L. (2008). *Classical Fourier Analysis*, 2nd ed. *Graduate Texts in Mathematics* **249**. Springer, New York. [MR2445437](#)
- HALL, P. and MEISTER, A. (2007). A ridge-parameter approach to deconvolution. *Ann. Statist.* **35** 1535–1558. [MR2351096](#)
- HASMINSKII, R. and IBRAGIMOV, I. (1990). On density estimation in the view of Kolmogorov's ideas in approximation theory. *Ann. Statist.* **18** 999–1010. [MR1062695](#)
- HESSE, C. H. (1995). Deconvolving a density from partially contaminated observations. *J. Multivariate Anal.* **55** 246–260. [MR1370403](#)
- JUDITSKY, A. and LAMBERT-LACROIX, S. (2004). On minimax density estimation on $\mathbb{R}$. *Bernoulli* **10** 187–220. [MR2046772](#)
- KERKYACHARIAN, G., LEPSKI, O. and PICARD, D. (2001). Nonlinear estimation in anisotropic multi-index denoising. *Probab. Theory Related Fields* **121** 137–170. [MR1863916](#)
- KERKYACHARIAN, G., PHAM NGOC, T. M. and PICARD, D. (2011). Localized spherical deconvolution. *Ann. Statist.* **39** 1042–1068. [MR2816347](#)
- LEPSKI, O. (2013). Multivariate density estimation under sup-norm loss: Oracle approach, adaptation and independence structure. *Ann. Statist.* **41** 1005–1034. [MR3099129](#)
- LEPSKI, O. (2015). Adaptive estimation over anisotropic functional classes via oracle approach. *Ann. Statist.* **43** 1178–1242. [MR3346701](#)
- LEPSKI, O. V. (2018). A new approach to estimator selection. *Bernoulli* **24**. To appear. Available at [arXiv:1603.03934v1](#).
- LEPSKI, O. V. and WILLER, T. (2017). Lower bounds in the convolution structure density model. *Bernoulli* **23** 884–926. [MR3606754](#)
- LOUNICI, K. and NICKL, R. (2011). Global uniform risk bounds for wavelet deconvolution estimators. *Ann. Statist.* **39** 201–231. [MR2797844](#)
- MASRY, E. (1993). Strong consistency and rates for deconvolution of multivariate densities of stationary processes. *Stochastic Process. Appl.* **47** 53–74. [MR1232852](#)
- MEISTER, A. (2009). *Deconvolution Problems in Nonparametric Statistics. Lecture Notes in Statistics* **193**. Springer, Berlin. [MR2768576](#)
- NIKOL'SKII, S. M. (1977). *Priblizhenie Funktsii Mnogikh Peremennykh i Teoremy Vlozheniya* [Approximation of Functions of Several Variables and Imbedding Theorems], 2nd ed., revised and supplemented. Nauka, Moscow (in Russian).

- PENSKY, M. and VIDAKOVIC, B. (1999). Adaptive wavelet estimator for nonparametric density deconvolution. *Ann. Statist.* **27** 2033–2053. [MR1765627](#)
- REBELLES, G. (2016). Structural adaptive deconvolution under $\mathbb{L}_p$ -losses. *Math. Methods Statist.* **25** 26–53. [MR3480609](#)
- REYNAUD-BOURET, P., RIVOIRARD, V. and TULEAU-MALOT, C. (2011). Adaptive density estimation: A curse of support? *J. Statist. Plann. Inference* **141** 115–139. [MR2719482](#)
- RIGOLLET, P. (2006). Adaptive density estimation using the blockwise Stein method. *Bernoulli* **12** 351–370. [MR2218559](#)
- RIGOLLET, PH. and TSYBAKOV, A. B. (2007). Linear and convex aggregation of density estimators. *Math. Methods Statist.* **16** 260–280. [MR2356821](#)
- SAMAROV, A. and TSYBAKOV, A. (2007). Aggregation of density estimators and dimension reduction. In *Advances in Statistical Modeling and Inference. Ser. Biostat.* **3** 233–251. World Sci. Publ., Hackensack, NJ. [MR2416118](#)
- STEFANSKI, L. A. (1990). Rates of convergence of some estimators in a class of deconvolution problems. *Statist. Probab. Lett.* **9** 229–235. [MR1045189](#)
- STEFANSKI, L. and CARROLL, R. J. (1990). Deconvoluting kernel density estimators. *Statistics* **21** 169–184. [MR1054861](#)
- YUAN, M. and CHEN, J. (2002). Deconvolving multidimensional density from partially contaminated observations. *J. Statist. Plann. Inference* **104** 147–160. [MR1900523](#)

EFFICIENT MULTIVARIATE ENTROPY ESTIMATION VIA k -NEAREST NEIGHBOUR DISTANCES

BY THOMAS B. BERRETT^{*,1}, RICHARD J. SAMWORTH^{*,2} AND
 MING YUAN^{†,3}

University of Cambridge^{*} and *University of Wisconsin–Madison*[†]

Many statistical procedures, including goodness-of-fit tests and methods for independent component analysis, rely critically on the estimation of the entropy of a distribution. In this paper, we seek entropy estimators that are efficient and achieve the local asymptotic minimax lower bound with respect to squared error loss. To this end, we study weighted averages of the estimators originally proposed by Kozachenko and Leonenko [*Probl. Inform. Transm.* **23** (1987), 95–101], based on the k -nearest neighbour distances of a sample of n independent and identically distributed random vectors in $\mathbb{R}^d$. A careful choice of weights enables us to obtain an efficient estimator in arbitrary dimensions, given sufficient smoothness, while the original unweighted estimator is typically only efficient when $d \leq 3$. In addition to the new estimator proposed and theoretical understanding provided, our results facilitate the construction of asymptotically valid confidence intervals for the entropy of asymptotically minimal width.

REFERENCES

- BARBOUR, A. D. and CHEN, L. H. Y., eds. (2005). *An Introduction to Stein's Method. Lecture Notes Series. Institute for Mathematical Sciences. National University of Singapore* **4**. Singapore Univ. Press, Singapore. [MR2235447](#)
- BEIRLANT, J., DUDEWICZ, E. J., GYÖRFI, L. and VAN DER MEULEN, E. C. (1997). Nonparametric entropy estimation: An overview. *Int. J. Math. Stat. Sci.* **6** 17–39. [MR1471870](#)
- BERRETT, T. B., SAMWORTH, R. J. and YUAN, M. (2019). Supplement to “Efficient multivariate entropy estimation via k -nearest neighbour distances.” DOI:[10.1214/18-AOS1688SUPP](#).
- BIAU, G. and DEVROYE, L. (2015). *Lectures on the Nearest Neighbor Method*. Springer, Cham. [MR3445317](#)
- CAI, T. T. and LOW, M. G. (2011). Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *Ann. Statist.* **39** 1012–1041. [MR2816346](#)
- CRESSIE, N. (1976). On the logarithms of high-order spacings. *Biometrika* **63** 343–355. [MR0428583](#)
- DELATTRE, S. and FOURNIER, N. (2017). On the Kozachenko–Leonenko entropy estimator. *J. Statist. Plann. Inference* **185** 69–93. [MR3612672](#)
- EL HAJE HUSSEIN, F. and GOLUBEV, YU. (2009). On entropy estimation by m -spacing method. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)* **363** 151–181, 189. [MR2749122](#)
- GAO, W., OH, S. and VISWANATH, P. (2016). Demystifying fixed k -nearest neighbor information estimators. Available at [arXiv:1604.03006](#).

MSC2010 subject classifications. 62G05, 62G20.

Key words and phrases. Efficiency, entropy estimation, Kozachenko–Leonenko estimator, weighted nearest neighbours.

- GORIA, M. N., LEONENKO, N. N., MERGEL, V. V. and NOVI INVERARDI, P. L. (2005). A new class of random vector entropy estimators and its applications in testing statistical hypotheses. *J. Nonparametr. Stat.* **17** 277–297. [MR2129834](#)
- HALL, P. and MORTON, S. C. (1993). On the estimation of entropy. *Ann. Inst. Statist. Math.* **45** 69–88. [MR1220291](#)
- KANTOROVIČ, L. V. and RUBINŠTEIN, G. Š. (1958). On a space of completely additive functions. *Vestnik Leningrad Univ. Math.* **13** 52–59. [MR0102006](#)
- KELLERER, H. G. (1985). Duality theorems and probability metrics. In *Proceedings of the Seventh Conference on Probability Theory (Braşov, 1982)* 211–220. VNU Sci. Press, Utrecht. [MR0867434](#)
- KOZACHENKO, L. F. and LEONENKO, N. N. (1987). Sample estimate of the entropy of a random vector. *Probl. Inf. Transm.* **23** 95–101.
- KWAK, N. and CHOI, C. (2002). Input feature selection by mutual information based on Parzen window. *IEEE Trans. Pattern Anal. Mach. Intell.* **24** 1667–1671.
- LAURENT, B. (1996). Efficient estimation of integral functionals of a density. *Ann. Statist.* **24** 659–681. [MR1394981](#)
- LEARNED-MILLER, E. G. and FISHER, J. W. III (2004). ICA using spacings estimates of entropy. *J. Mach. Learn. Res.* **4** 1271–1295. [MR2103630](#)
- LEPSKI, O., NEMIROVSKI, A. and SPOKOINY, V. (1999). On estimation of the L_r norm of a regression function. *Probab. Theory Related Fields* **113** 221–253. [MR1670867](#)
- MNATSAKANOV, R. M., MISRA, N., LI, SH. and HARNER, E. J. (2008). k_n -nearest neighbor estimators of entropy. *Math. Methods Statist.* **17** 261–277. [MR2448950](#)
- MOON, K. R., SRICHARAN, K., GREENEWALD, K. and HERO, A. O. (2016). Nonparametric ensemble estimation of distributional functionals. <https://arxiv.org/abs/1601.06884v2>.
- PANINSKI, L. (2003). Estimation of entropy and mutual information. *Neural Comput.* **15** 1191–1253.
- PANINSKI, L. and YAJIMA, M. (2008). Undersmoothed kernel entropy estimators. *IEEE Trans. Inform. Theory* **54** 4384–4388. [MR2451978](#)
- SHORACK, G. R. and WELLNER, J. A. (2009). *Empirical Processes with Applications to Statistics. Classics in Applied Mathematics* **59**. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. [MR3396731](#)
- SINGH, S. and PÓCZOS, B. (2016). Analysis of k nearest neighbor distances with application to entropy estimation. *NIPS* **29** 1217–1225.
- SINGH, H., MISRA, N., HNZDO, V., FEDOROWICZ, A. and DEMCHUK, E. (2003). Nearest neighbor estimates of entropy. *Amer. J. Math. Management Sci.* **23** 301–321. [MR2045530](#)
- SRICHARAN, K., WEI, D. and HERO, A. O. III (2013). Ensemble estimators for multivariate entropy estimation. *IEEE Trans. Inform. Theory* **59** 4374–4388. [MR3071335](#)
- TSYBAKOV, A. B. and VAN DER MEULEN, E. C. (1996). Root- n consistent estimators of entropy for densities with unbounded support. *Scand. J. Stat.* **23** 75–83. [MR1380483](#)
- VAN DER VAART, A. W. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. [MR1652247](#)
- VASICEK, O. (1976). A test for normality based on sample entropy. *J. Roy. Statist. Soc. Ser. B* **38** 54–59. [MR0420958](#)

POSTERIOR GRAPH SELECTION AND ESTIMATION CONSISTENCY FOR HIGH-DIMENSIONAL BAYESIAN DAG MODELS

BY XUAN CAO, KSHITIJ KHARE AND MALAY GHOSH

University of Florida

Covariance estimation and selection for high-dimensional multivariate datasets is a fundamental problem in modern statistics. Gaussian directed acyclic graph (DAG) models are a popular class of models used for this purpose. Gaussian DAG models introduce sparsity in the Cholesky factor of the inverse covariance matrix, and the sparsity pattern in turn corresponds to specific conditional independence assumptions on the underlying variables. A variety of priors have been developed in recent years for Bayesian inference in DAG models, yet crucial convergence and sparsity selection properties for these models have not been thoroughly investigated. Most of these priors are adaptations/generalizations of the Wishart distribution in the DAG context. In this paper, we consider a flexible and general class of these “DAG-Wishart” priors with multiple shape parameters. Under mild regularity assumptions, we establish strong graph selection consistency and establish posterior convergence rates for estimation when the number of variables p is allowed to grow at an appropriate subexponential rate with the sample size n .

REFERENCES

- [1] ALTOMARE, D., CONSONNI, G. and LA ROCCA, L. (2013). Objective Bayesian search of Gaussian directed acyclic graphical models for ordered variables with non-local priors. *Biometrics* **69** 478–487. [MR3071066](#)
- [2] ARAGAM, B., AMINI, A. and ZHOU, Q. (2015). Learning directed acyclic graphs with penalized neighbourhood regression. Available at <https://arxiv.org/abs/1511.08963>.
- [3] BANERJEE, S. and GHOSAL, S. (2014). Posterior convergence rates for estimating large precision matrices using graphical models. *Electron. J. Stat.* **8** 2111–2137. [MR3273620](#)
- [4] BANERJEE, S. and GHOSAL, S. (2015). Bayesian structure learning in graphical models. *J. Multivariate Anal.* **136** 147–162. [MR3321485](#)
- [5] BEN-DAVID, E., LI, T., MASSAM, H. and RAJARATNAM, B. (2016). High dimensional Bayesian inference for Gaussian directed acyclic graph models. Technical report. Available at <http://arxiv.org/abs/1109.4371>.
- [6] BICKEL, P. J. and LEVINA, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. [MR2387969](#)
- [7] CAO, X., KHARE, K. and GHOSH, M. (2019). Supplement to “Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models.” DOI:[10.1214/18-AOS1689SUPP](https://doi.org/10.1214/18-AOS1689SUPP).

MSC2010 subject classifications. Primary 62F15; secondary 62G20.

Key words and phrases. Posterior consistency, high-dimensional data, Bayesian DAG models, covariance estimation, graph selection.

- [8] CONSONNI, G., LA ROCCA, L. and PELUSO, S. (2017). Objective Bayes covariate-adjusted sparse graphical model selection. *Scand. J. Stat.* **44** 741–764. [MR3687971](#)
- [9] EL KAROUI, N. (2008). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Ann. Statist.* **36** 2757–2790. [MR2485012](#)
- [10] GEIGER, D. and HECKERMAN, D. (2002). Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *Ann. Statist.* **30** 1412–1440. [MR1936324](#)
- [11] HUANG, J. Z., LIU, N., POURAHMADI, M. and LIU, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93** 85–98. [MR2277742](#)
- [12] JOHNSON, V. E. and ROSSELL, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 143–170. [MR2830762](#)
- [13] JOHNSON, V. E. and ROSSELL, D. (2012). Bayesian model selection in high-dimensional settings. *J. Amer. Statist. Assoc.* **107** 649–660. [MR2980074](#)
- [14] KHARE, K., OH, S., RAHMAN, S. and RAJARATNAM, B. (2017). A convex framework for high-dimensional sparse cholesky based covariance estimation in gaussian dag models. Technical report. Available at <https://arxiv.org/abs/1610.02436>.
- [15] KOLLER, D. and FRIEDMAN, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, MA. [MR2778120](#)
- [16] LETAC, G. and MASSAM, H. (2007). Wishart distributions for decomposable graphs. *Ann. Statist.* **35** 1278–1323. [MR2341706](#)
- [17] NARISSETTY, N. N. and HE, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *Ann. Statist.* **42** 789–817. [MR3210987](#)
- [18] PAULSEN, V. I., POWER, S. C. and SMITH, R. R. (1989). Schur products and matrix completions. *J. Funct. Anal.* **85** 151–178. [MR1005860](#)
- [19] POURAHMADI, M. (2007). Cholesky decompositions and estimation of a covariance matrix: Orthogonality of variance-correlation parameters. *Biometrika* **94** 1006–1013. [MR2376812](#)
- [20] ROTHMAN, A. J., LEVINA, E. and ZHU, J. (2010). A new approach to Cholesky-based covariance regularization in high dimensions. *Biometrika* **97** 539–550. [MR2672482](#)
- [21] RUDELSON, M. and VERSHYNIN, R. (2013). Hanson–Wright inequality and sub-Gaussian concentration. *Electron. Commun. Probab.* **18** no. 82, 9. [MR3125258](#)
- [22] RÜTIMANN, P. and BÜHLMANN, P. (2009). High dimensional sparse covariance estimation via directed acyclic graphs. *Electron. J. Stat.* **3** 1133–1160. [MR2566184](#)
- [23] SHOJAIE, A. and MICHAELIDIS, G. (2010). Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika* **97** 519–538. [MR2672481](#)
- [24] SMITH, M. and KOHN, R. (2002). Parsimonious covariance matrix estimation for longitudinal data. *J. Amer. Statist. Assoc.* **97** 1141–1153. [MR1951266](#)
- [25] VAN DE GEER, S. and BÜHLMANN, P. (2013). ℓ_0 -penalized maximum likelihood for sparse directed acyclic graphs. *Ann. Statist.* **41** 536–567. [MR3099113](#)
- [26] XIANG, R., KHARE, K. and GHOSH, M. (2015). High dimensional posterior convergence rates for decomposable graphical models. *Electron. J. Stat.* **9** 2828–2854. [MR3439186](#)
- [27] YU, G. and BIEN, J. (2017). Learning local dependence in ordered data. *J. Mach. Learn. Res.* **18** Paper No. 42, 60. [MR3655307](#)

LOCALLY ADAPTIVE CONFIDENCE BANDS¹

BY TIM PATSCHKOWSKI AND ANGELIKA ROHDE

Ruhr-Universität Bochum and Albert-Ludwigs-Universität Freiburg

We develop honest and locally adaptive confidence bands for probability densities. They provide substantially improved confidence statements in case of inhomogeneous smoothness, and are easily implemented and visualized. The article contributes conceptual work on locally adaptive inference as a straightforward modification of the global setting imposes severe obstacles for statistical purposes. Among others, we introduce a statistical notion of local Hölder regularity and prove a correspondingly strong version of local adaptivity. We substantially relax the straightforward localization of the self-similarity condition in order not to rule out prototypical densities. The set of densities permanently excluded from the consideration is shown to be pathological in a mathematically rigorous sense. On a technical level, the crucial component for the verification of honesty is the identification of an asymptotically least favorable stationary case by means of Slepian's comparison inequality.

REFERENCES

- BARAUD, Y. (2004). Confidence balls in Gaussian regression. *Ann. Statist.* **32** 528–551. [MR2060168](#)
- BICKEL, P. J. and ROSENBLATT, M. (1973). On some global measures of the deviations of density function estimates. *Ann. Statist.* **1** 1071–1095. [MR0348906](#)
- BULL, A. D. (2012). Honest adaptive confidence bands and self-similar functions. *Electron. J. Stat.* **6** 1490–1516. [MR2988456](#)
- BULL, A. D. and NICKL, R. (2013). Adaptive confidence sets in L^2 . *Probab. Theory Related Fields* **156** 889–919. [MR3078289](#)
- CAI, T. T. and LOW, M. G. (2004). An adaptation theory for nonparametric confidence intervals. *Ann. Statist.* **32** 1805–1840. [MR2102494](#)
- CAI, T. T. and LOW, M. G. (2006). Adaptive confidence balls. *Ann. Statist.* **34** 202–228. [MR2275240](#)
- CARPENTIER, A. (2013). Honest and adaptive confidence sets in L_p . *Electron. J. Stat.* **7** 2875–2923. [MR3148371](#)
- CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2014a). Anti-concentration and honest, adaptive confidence bands. *Ann. Statist.* **42** 1787–1818. [MR3262468](#)
- CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2014b). Gaussian approximation of suprema of empirical processes. *Ann. Statist.* **42** 1564–1597. [MR3262461](#)
- DAVIES, P. L., KOVAC, A. and MEISE, M. (2009). Nonparametric regression, confidence regions and regularization. *Ann. Statist.* **37** 2597–2625. [MR2541440](#)
- DÜMBGEN, L. (1998). New goodness-of-fit tests and their application to nonparametric confidence sets. *Ann. Statist.* **26** 288–314. [MR1611768](#)

MSC2010 subject classifications. 62G07, 62G15.

Key words and phrases. Local adaptivity, confidence bands in density estimation.

- DÜMBGEN, L. (2003). Optimal confidence bands for shape-restricted curves. *Bernoulli* **9** 423–449. [MR1997491](#)
- GENOVESE, C. R. and WASSERMAN, L. (2005). Confidence sets for nonparametric wavelet regression. *Ann. Statist.* **33** 698–729. [MR2163157](#)
- GENOVESE, C. and WASSERMAN, L. (2008). Adaptive confidence bands. *Ann. Statist.* **36** 875–905. [MR2396818](#)
- GINÉ, E. and NICKL, R. (2010). Confidence bands in density estimation. *Ann. Statist.* **38** 1122–1170. [MR2604707](#)
- HENGARTNER, N. W. and STARK, P. B. (1995). Finite-sample confidence envelopes for shape-restricted densities. *Ann. Statist.* **23** 525–550. [MR1332580](#)
- HOFFMANN, M. and NICKL, R. (2011). On adaptive inference and confidence bands. *Ann. Statist.* **39** 2383–2409. [MR2906872](#)
- JUDITSKY, A. and LAMBERT-LACROIX, S. (2003). Nonparametric confidence set estimation. *Math. Methods Statist.* **12** 410–428. [MR2054156](#)
- KERKYACHARIAN, G., NICKL, R. and PICARD, D. (2012). Concentration inequalities and confidence bands for needlet density estimators on compact homogeneous manifolds. *Probab. Theory Related Fields* **153** 363–404. [MR2925578](#)
- KUEH, A. (2012). Locally adaptive density estimation on the unit sphere using needlets. *Constr. Approx.* **36** 433–458. [MR2996439](#)
- LEPSKIĬ, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Teor. Veroyatn. Primen.* **35** 459–470. [MR1091202](#)
- LOW, M. G. (1997). On nonparametric confidence intervals. *Ann. Statist.* **25** 2547–2554. [MR1604412](#)
- NICKL, R. and SZABÓ, B. (2016). A sharp adaptive confidence ball for self-similar functions. *Stochastic Process. Appl.* **126** 3913–3934. [MR3565485](#)
- PATSKOWSKI, T. and ROHDE, A. (2019). Supplement to “Locally adaptive confidence bands.” DOI:[10.1214/18-AOS1690SUPP](#).
- PICARD, D. and TRIBOULEY, K. (2000). Adaptive confidence interval for pointwise curve estimation. *Ann. Statist.* **28** 298–335. [MR1762913](#)
- ROBINS, J. and VAN DER VAART, A. (2006). Adaptive nonparametric confidence sets. *Ann. Statist.* **34** 229–253. [MR2275241](#)
- SEURET, S. and VÉHEL, J. L. (2002). The local Hölder function of a continuous function. *Appl. Comput. Harmon. Anal.* **13** 263–276. [MR1942744](#)
- TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York. Revised and extended from the 2004 French original, translated by Vladimir Zaiats. [MR2724359](#)

ASYMPTOTIC DISTRIBUTION-FREE CHANGE-POINT DETECTION FOR MULTIVARIATE AND NON-EUCLIDEAN DATA¹

BY LYNNA CHU AND HAO CHEN

University of California, Davis

We consider the testing and estimation of change-points, locations where the distribution abruptly changes, in a sequence of multivariate or non-Euclidean observations. We study a nonparametric framework that utilizes similarity information among observations, which can be applied to various data types as long as an informative similarity measure on the sample space can be defined. The existing approach along this line has low power and/or biased estimates for change-points under some common scenarios. We address these problems by considering new tests based on similarity information. Simulation studies show that the new approaches exhibit substantial improvements in detecting and estimating change-points. In addition, under some mild conditions, the new test statistics are asymptotically distribution-free under the null hypothesis of no change. Analytic p -value approximations to the significance of the new test statistics for the single change-point alternative and changed interval alternative are derived, making the new approaches easy off-the-shelf tools for large datasets. The new approaches are illustrated in an analysis of New York taxi data.

REFERENCES

- CARLSTEIN, E., MÜLLER, H.-G. and SIEGMUND, D., eds. (1994). *Change-Point Problems. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **23**. IMS, Hayward, CA. Papers from the AMS-IMS-SIAM Summer Research Conference held at Mt. Holyoke College, South Hadley, MA, July 11–16, 1992. [MR1477909](#)
- CHEN, H., CHEN, X. and SU, Y. (2017). A weighted edge-count two-sample test for multivariate and object data. *J. Amer. Statist. Assoc.* **112**. To appear. DOI:[10.1080/01621459.2017.1307757](#).
- CHEN, H. and FRIEDMAN, J. H. (2017). A new graph-based two-sample test for multivariate and object data. *J. Amer. Statist. Assoc.* **112** 397–409. [MR3646580](#)
- CHEN, J. and GUPTA, A. K. (2012). *Parametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance*, 2nd ed. Birkhäuser/Springer, New York. [MR3025631](#)
- CHEN, L. H. and SHAO, Q.-M. (1994). *Stein's Method for Normal Approximation*. In *An Introduction to Stein's Method. Lecture Notes Series* **4** 1–59. World Scientific, Singapore.
- CHEN, H. and ZHANG, N. (2015). Graph-based change-point detection. *Ann. Statist.* **43** 139–176. [MR3285603](#)
- CHU, L. and CHEN, H. (2019). Supplement to “Asymptotic distribution-free change-point detection for multivariate and non-Euclidean data.” DOI:[10.1214/18-AOS1691SUPP](#).
- CSÖRGŐ, M. and HORVÁTH, L. (1997). *Limit Theorems in Change-Point Analysis*. Wiley, Chichester. [MR2743035](#)

MSC2010 subject classifications. Primary 62G32; secondary 60K35.

Key words and phrases. Change-point, graph-based tests, nonparametric, scan statistic, tail probability, high-dimensional data, network data, non-Euclidean data.

- CULE, M., SAMWORTH, R. and STEWART, M. (2010). Maximum likelihood estimation of a multi-dimensional log-concave density. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 545–607. [MR2758237](#)
- DESOBRY, F., DAVY, M. and DONCARLI, C. (2005). An online kernel change detection algorithm. *IEEE Trans. Signal Process.* **53** 2961–2974. [MR2169647](#)
- FRIEDMAN, J. H. and RAFSKY, L. C. (1979). Multivariate generalizations of the Wald–Wolfowitz and Smirnov two-sample tests. *Ann. Statist.* **7** 697–717. [MR0532236](#)
- HEARD, N. A., WESTON, D. J., PLATANIOTI, K. and HAND, D. J. (2010). Bayesian anomaly detection methods for social networks. *Ann. Appl. Stat.* **4** 645–662. [MR2758643](#)
- JIRAK, M. (2015). Uniform change point tests in high dimension. *Ann. Statist.* **43** 2451–2483. [MR3405600](#)
- KOSSINETIS, G. and WATTS, D. J. (2006). Empirical analysis of an evolving social network. *Science* **311** 88–90. [MR2192483](#)
- LUNG-YUT-FONG, A., LÉVY-LEDUC, C. and CAPPÉ, O. (2015). Homogeneity and change-point detection tests for multivariate data using rank statistics. *J. SFS* **156** 133–162. [MR3436651](#)
- MATTESON, D. S. and JAMES, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *J. Amer. Statist. Assoc.* **109** 334–345. [MR3180567](#)
- PARK, Y., WANG, H., NÖBAUER, T., VAZIRI, A. and PRIEBE, C. E. (2015). Anomaly detection on whole-brain functional imaging of neuronal activity using graph scan statistics. In *ACM Conference on Knowledge Discovery and Data Mining (KDD), Workshop on Outlier Definition, Detection, and Description (ODDx3)*.
- SIEGMUND, D. and YAKIR, B. (2007). *The Statistics of Gene Mapping. Statistics for Biology and Health*. Springer, New York. [MR2301277](#)
- WANG, H., TANG, M., PARK, Y. and PRIEBE, C. E. (2014). Locality statistics for anomaly detection in time series of graphs. *IEEE Trans. Signal Process.* **62** 703–717. [MR3160307](#)
- XIE, Y. and SIEGMUND, D. (2013). Sequential multi-sensor change-point detection. *Ann. Statist.* **41** 670–692. [MR3023983](#)
- ZHANG, N. R., SIEGMUND, D. O., JI, H. and LI, J. Z. (2010). Detecting simultaneous changepoints in multiple sequences. *Biometrika* **97** 631–645. [MR2672488](#)

STATISTICS ON THE STIEFEL MANIFOLD: THEORY AND APPLICATIONS¹

BY RUDRASIS CHAKRABORTY AND BABA C. VEMURI

University of Florida

A Stiefel manifold of the compact type is often encountered in many fields of engineering including, signal and image processing, machine learning, numerical optimization and others. The Stiefel manifold is a Riemannian homogeneous space but not a symmetric space. In previous work, researchers have defined probability distributions on symmetric spaces and performed statistical analysis of data residing in these spaces. In this paper, we present original work involving definition of Gaussian distributions on a homogeneous space and show that the maximum-likelihood estimate of the location parameter of a Gaussian distribution on the homogeneous space yields the Fréchet mean (FM) of the samples drawn from this distribution. Further, we present an algorithm to sample from the Gaussian distribution on the Stiefel manifold and recursively compute the FM of these samples. We also prove the weak consistency of this recursive FM estimator. Several synthetic and real data experiments are then presented, demonstrating the superior computational performance of this estimator over the gradient descent based nonrecursive counter part as well as the stochastic gradient descent based method prevalent in literature.

REFERENCES

- ABSIL, P.-A., MAHONY, R. and SEPULCHRE, R. (2004). Riemannian geometry of Grassmann manifolds with a view on algorithmic computation. *Acta Appl. Math.* **80** 199–220. [MR2035507](#)
- AFSARI, B. (2011). Riemannian L^p center of mass: Existence, uniqueness, and convexity. *Proc. Amer. Math. Soc.* **139** 655–673. [MR2736346](#)
- ANDO, T., LI, C.-K. and MATHIAS, R. (2004). Geometric means. *Linear Algebra Appl.* **385** 305–334. [MR2063358](#)
- ARNAUDON, M., BARBARESCO, F. and YANG, L. (2013). Riemannian medians and means with applications to radar signal processing. *IEEE J. Sel. Top. Signal Process.* **7** 595–604.
- BHATIA, R. (1997). *Matrix Analysis. Graduate Texts in Mathematics* **169**. Springer, New York. [MR1477662](#)
- BHATTACHARYA, A. and BHATTACHARYA, R. (2008). Statistics on Riemannian manifolds: Asymptotic distribution and curvature. *Proc. Amer. Math. Soc.* **136** 2959–2967. [MR2399064](#)
- BONNABEL, S. (2013). Stochastic gradient descent on Riemannian manifolds. *IEEE Trans. Automat. Control* **58** 2217–2229. [MR3101606](#)
- CETINGUL, H. E. and VIDAL, R. (2009). Intrinsic mean shift for clustering on Stiefel and Grassmann manifolds. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on 1896–1902*. IEEE, New York.

MSC2010 subject classifications. Primary 62F12; secondary 58A99.

Key words and phrases. Homogeneous space, Stiefel manifold, Fréchet mean, Gaussian distribution.

- CHAKRABORTY, R., BANERJEE, M. and VEMURI, B. (2017). Statistics on the space of trajectories for longitudinal data analysis. In *IEEE International Symposium on Biomedical Imaging*.
- CHAKRABORTY, R. and VEMURI, B. C. (2015). Recursive Frechet mean computation on the Grassmannian and its applications to computer vision. In *The IEEE International Conference on Computer Vision (ICCV)*.
- CHARFI, M., CHEBBI, Z., MOAKHER, M. and VEMURI, B. C. (2013). Bhattacharyya median of symmetric positive-definite matrices and application to the denoising of diffusion-tensor fields. In *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on* 1227–1230. IEEE, New York.
- CHEEGER, J. and EBIN, D. G. (1975). *Comparison Theorems in Riemannian Geometry*. North-Holland Mathematical Library **9**. North-Holland, Amsterdam. [MR0458335](#)
- CHENG, G. and VEMURI, B. C. (2013). A novel dynamic system in the space of SPD matrices with applications to appearance tracking. *SIAM J. Imaging Sci.* **6** 592–615. [MR3035470](#)
- CHIKUSE, Y. (1991). Asymptotic expansions for distributions of the large sample matrix resultant and related statistics on the Stiefel manifold. *J. Multivariate Anal.* **39** 270–283. [MR1147122](#)
- CRAMÉR, H. (1946). *Mathematical Methods of Statistics*. Princeton Mathematical Series **9**. Princeton Univ. Press, Princeton, NJ. [MR0016588](#)
- DALAL, N. and TRIGGS, B. (2005). Histograms of oriented gradients for human detection. In *CVPR* **1** 886–893.
- DORETTO, G., CHIUSO, A., WU, Y. N. and SOATTO, S. (2003). Dynamic textures. *Int. J. Comput. Vis.* **51** 91–109.
- DOWNES, T., LIEBMAN, J. and MACKAY, W. (1971). Statistical methods for vectorcardiogram orientations. In *Vectorcardiography 2: Proc. XIth International Symp. on Vectorcardiography*.
- FLETCHER, P. T. (2013). Geodesic regression and the theory of least squares on Riemannian manifolds. *Int. J. Comput. Vis.* **105** 171–185. [MR3104017](#)
- FLETCHER, P. T. and JOSHI, S. (2007). Riemannian geometry for the statistical analysis of diffusion tensor data. *Signal Process.* **87** 250–262.
- FLETCHER, P. T., LU, C., PIZER, S. M. and JOSHI, S. (2004). Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Trans. Med. Imag.* **23** 995–1005.
- FRAIKIN, C., HÜPER, K. and DOOREN, P. V. (2007). Optimization over the Stiefel manifold. *PAMM* **7** 1062205–1062206.
- FRÉCHET, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. H. Poincaré* **10** 215–310. [MR0027464](#)
- GOODALL, C. R. and MARDIA, K. V. (1999). Projective shape analysis. *J. Comput. Graph. Statist.* **8** 143–168. [MR1706381](#)
- GRENANDER, U. (2008). *Probabilities on Algebraic Structures*. Dover Publications, Mineola, NY.
- GROISSER, D. (2004). Newton’s method, zeroes of vector fields, and the Riemannian center of mass. *Adv. in Appl. Math.* **33** 95–135. [MR2064359](#)
- HARTLEY, R., TRUMPF, J., DAI, Y. and LI, H. (2013). Rotation averaging. *Int. J. Comput. Vis.* **103** 267–305. [MR3067537](#)
- HAUBERG, S., FERAGEN, A. and BLACK, M. J. (2014). Grassmann averages for scalable robust PCA. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 3810–3817.
- HELGASON, S. (1978). *Differential Geometry, Lie Groups, and Symmetric Spaces*. Pure and Applied Mathematics **80**. Academic Press, New York. [MR0514561](#)
- HENDRIKS, H. and LANDSMAN, Z. (1998). Mean location and sample mean location on manifolds: Asymptotics, tests, confidence regions. *J. Multivariate Anal.* **67** 227–243. [MR1659156](#)
- HO, J., CHENG, G., SALEHIAN, H. and VEMURI, B. (2013). Recursive Karcher expectation estimators and geometric law of large numbers. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics* 325–332.

- KANEKO, T., FIORI, S. and TANAKA, T. (2013). Empirical arithmetic averaging over the compact Stiefel manifold. *IEEE Trans. Signal Process.* **61** 883–894. [MR3018336](#)
- KARCHER, H. (1977). Riemannian center of mass and mollifier smoothing. *Comm. Pure Appl. Math.* **30** 509–541. [MR0442975](#)
- KENDALL, W. S. (1990). Probability, convexity, and harmonic maps with small image. I. Uniqueness and fine existence. *Proc. Lond. Math. Soc.* (3) **61** 371–406. [MR1063050](#)
- LUI, Y. M. (2012). Advances in matrix manifolds for computer vision. *Image Vis. Comput.* **30** 380–388.
- MARDIA, K. V. and JUPP, P. E. (2000). *Directional Statistics*. Wiley, Chichester. Revised reprint of *Statistics of directional data* by Mardia [[MR0336854](#) (49 #1627)]. [MR1828667](#)
- MOAKHER, M. (2005). A differential geometric approach to the geometric mean of symmetric positive-definite matrices. *SIAM J. Matrix Anal. Appl.* **26** 735–747. [MR2137480](#)
- MOAKHER, M. (2006). On the averaging of symmetric positive-definite tensors. *J. Elasticity* **82** 273–296. [MR2231065](#)
- PATRANGENARU, V. and MARDIA, K. V. (2003). Affine shape analysis and image analysis. In *22nd Leeds Annual Statistics Research Workshop*.
- PENNEC, X. (2006). Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. *J. Math. Imaging Vision* **25** 127–154. [MR2254442](#)
- PENNEC, X., FILLARD, P. and AYACHE, N. (2006). A Riemannian framework for tensor computing. *Int. J. Comput. Vis.* **66** 41–66.
- PHAM, D.-S. and VENKATESH, S. (2008). Robust learning of discriminative projection for multiclass classification on the Stiefel manifold. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* 1–7. IEEE, New York.
- RAO, C. R. (1987). Differential metrics in probability spaces. *Differential Geometry in Statistical Inference* **10** 217–240.
- RAO, C. R. (1992). Information and the accuracy attainable in the estimation of statistical parameters. In *Breakthroughs in Statistics* 235–247. Springer, New York.
- SAID, S., HAJRI, H., BOMBRUN, L. and VEMURI, B. C. (2016). Gaussian distributions on Riemannian symmetric spaces: Statistical learning with structured covariance matrices. ArXiv Preprint [ArXiv:1607.06929](#).
- SAID, S., BOMBRUN, L., BERTHOUMIEU, Y. and MANTON, J. H. (2017). Riemannian Gaussian distributions on the space of symmetric positive definite matrices. *IEEE Trans. Inform. Theory* **63** 2153–2170. [MR3626862](#)
- SALEHIAN, H., CHAKRABORTY, R., OFORI, E., VAILLANCOURT, D. and VEMURI, B. C. (2015). An efficient recursive estimator of the Fréchet mean on a hypersphere with applications to medical image analysis. In *Mathematical Foundations of Computational Anatomy*.
- SCHULDT, C., LAPTEV, I. and CAPUTO, B. (2004). Recognizing human actions: A local SVM approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on* **3** 32–36.
- SRIVASTAVA, A., JERMYN, I. and JOSHI, S. (2007). Riemannian analysis of probability density functions with applications in vision. In *CVPR* 1–8.
- STURM, K.-T. (2003). Probability measures on metric spaces of nonpositive curvature. In *Heat Kernels and Analysis on Manifolds, Graphs, and Metric Spaces (Paris, 2002)*. *Contemp. Math.* **338** 357–390. Amer. Math. Soc., Providence, RI. [MR2039961](#)
- TUCH, D. S., REESE, T. G., WIEGELL, M. R. and WEDEEN, V. J. (2003). Diffusion MRI of complex neural architecture. *Neuron* **40** 885–895.
- TURAGA, P., VEERARAGHAVAN, A. and CHELLAPPA, R. (2008). Statistical analysis on Stiefel and Grassmann manifolds with applications in computer vision. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* 1–8. IEEE, New York.
- WONG, Y. (1968). Sectional curvatures of Grassmann manifolds. *Proc. Natl. Acad. Sci. USA* **60** 75–79. [MR0229173](#)

- ZHANG, M. and FLETCHER, P. T. (2013). Probabilistic principal geodesic analysis. In *Advances in Neural Information Processing Systems* 1178–1186.
- ZILLER, W. (2007). Examples of Riemannian manifolds with non-negative sectional curvature. In *Surveys in Differential Geometry. Vol. XI. Surv. Differ. Geom.* **11** 63–102. Int. Press, Somerville, MA. [MR2408264](#)
- ZIMMERMANN, R. (2017). A matrix-algebraic algorithm for the Riemannian logarithm on the Stiefel manifold under the canonical metric. *SIAM J. Matrix Anal. Appl.* **38** 322–342. [MR3639577](#)

GOODNESS-OF-FIT TESTS FOR THE FUNCTIONAL LINEAR MODEL BASED ON RANDOMLY PROJECTED EMPIRICAL PROCESSES

BY JUAN A. CUESTA-ALBERTOS^{*,1} EDUARDO GARCÍA-PORTUGUÉS^{†,2,3},
 MANUEL FEBRERO-BANDE^{‡,2} AND WENCESLAO GONZÁLEZ-MANTEIGA^{‡,2}

University of Cantabria^{}, Carlos III University of Madrid[†] and
 University of Santiago de Compostela[‡]*

We consider marked empirical processes indexed by a randomly projected functional covariate to construct goodness-of-fit tests for the functional linear model with scalar response. The test statistics are built from continuous functionals over the projected process, resulting in computationally efficient tests that exhibit root- n convergence rates and circumvent the curse of dimensionality. The weak convergence of the empirical process is obtained conditionally on a random direction, whilst the almost surely equivalence between the testing for significance expressed on the original and on the projected functional covariate is proved. The computation of the test in practice involves calibration by wild bootstrap resampling and the combination of several p -values, arising from different projections, by means of the false discovery rate method. The finite sample properties of the tests are illustrated in a simulation study for a variety of linear models, underlying processes, and alternatives. The software provided implements the tests and allows the replication of simulations and data applications.

REFERENCES

- ANEIROS-PÉREZ, G. and VIEU, P. (2006). Semi-functional partial linear regression. *Statist. Probab. Lett.* **76** 1102–1110. [MR2269280](#)
- 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 ROSENBLATT, M. (1973). On some global measures of the deviations of density function estimates. *Ann. Statist.* **1** 1071–1095. [MR0348906](#)
- BIERENS, H. J. (1982). Consistent model specification tests. *J. Econometrics* **20** 105–134. [MR0685673](#)
- BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York. [MR0233396](#)
- BÜCHER, A., DETTE, H. and WIECZOREK, G. (2011). Testing model assumptions in functional regression models. *J. Multivariate Anal.* **102** 1472–1488. [MR2819963](#)
- CARDOT, H., MAS, A. and SARDA, P. (2007). CLT in functional linear regression models. *Probab. Theory Related Fields* **138** 325–361. [MR2299711](#)
- CARDOT, H., FERRATY, F., MAS, A. and SARDA, P. (2003). Testing hypotheses in the functional linear model. *Scand. J. Stat.* **30** 241–255. [MR1965105](#)

MSC2010 subject classifications. Primary 62G10, 62J05; secondary 62G09.

Key words and phrases. Empirical process, functional data, functional linear model, functional principal components, goodness-of-fit, random projections.

- CHIOU, J.-M. and MÜLLER, H.-G. (2007). Diagnostics for functional regression via residual processes. *Comput. Statist. Data Anal.* **51** 4849–4863. [MR2364544](#)
- CUESTA-ALBERTOS, J. A. and FEBRERO-BANDE, M. (2010). A simple multiway ANOVA for functional data. *TEST* **19** 537–557. [MR2746001](#)
- CUESTA-ALBERTOS, J. A., FRAIMAN, R. and RANSFORD, T. (2007). A sharp form of the Cramér–Wold theorem. *J. Theoret. Probab.* **20** 201–209. [MR2324526](#)
- CUESTA-ALBERTOS, J. A., DEL BARRIO, E., FRAIMAN, R. and MATRÁN, C. (2007). The random projection method in goodness of fit for functional data. *Comput. Statist. Data Anal.* **51** 4814–4831. [MR2364542](#)
- CUESTA-ALBERTOS, J. A., GARCÍA-PORTUGUÉS, E., FEBRERO-BANDE, M. and GONZÁLEZ-MANTEIGA, W. (2019). Supplement to “Goodness-of-fit tests for the functional linear model based on randomly projected empirical processes.” DOI:[10.1214/18-AOS1693SUPP](#).
- D’AGOSTINO, R. B. and STEPHENS, M. A., eds. (1986) Goodness-of-Fit Techniques. *Statistics: Textbooks and Monographs* **68**. Dekker, Inc., New York. [MR0874534](#)
- DELSOL, L., FERRATY, F. and VIEU, P. (2011a). Structural test in regression on functional variables. *J. Multivariate Anal.* **102** 422–447. [MR2755007](#)
- DELSOL, L., FERRATY, F. and VIEU, P. (2011b). Structural tests in regression on functional variable. In *Recent Advances in Functional Data Analysis and Related Topics* 77–83. Physica-Verlag/Springer, Heidelberg. [MR2815564](#)
- DURBIN, J. (1973). Weak convergence of the sample distribution function when parameters are estimated. *Ann. Statist.* **1** 279–290. [MR0359131](#)
- ESCANCIANO, J. C. (2006). A consistent diagnostic test for regression models using projections. *Econometric Theory* **22** 1030–1051. [MR2328527](#)
- FEBRERO-BANDE, M. and OVIEDO DE LA FUENTE, M. (2017). *fda.usc: Functional Data Analysis and Utilities for Statistical Computing (fda.usc)*. R package version 1.3.1. Available at <http://cran.r-project.org/web/packages/fda.usc/>.
- FERRATY, F. and VIEU, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer, New York. [MR2229687](#)
- FRAIMAN, R. and MUNIZ, G. (2001). Trimmed means for functional data. *TEST* **10** 419–440. [MR1881149](#)
- GARCÍA-PORTUGUÉS, E., GONZÁLEZ-MANTEIGA, W. and FEBRERO-BANDE, M. (2014). A goodness-of-fit test for the functional linear model with scalar response. *J. Comput. Graph. Statist.* **23** 761–778. [MR3224655](#)
- GONZÁLEZ-MANTEIGA, W. and CRUJEIRAS, R. M. (2013). An updated review of goodness-of-fit tests for regression models. *TEST* **22** 361–411. [MR3093195](#)
- HÄRDLE, W. and MAMMEN, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.* **21** 1926–1947. [MR1245774](#)
- HILGERT, N., MAS, A. and VERZELEN, N. (2013). Minimax adaptive tests for the functional linear model. *Ann. Statist.* **41** 838–869. [MR3099123](#)
- HOFFMANN-JØRGENSEN, J. and PISIER, G. (1976). The law of large numbers and the central limit theorem in Banach spaces. *Ann. Probab.* **4** 587–599. [MR0423451](#)
- HORVÁTH, L. and KOKOSZKA, P. (2012). *Inference for Functional Data with Applications*. Springer, New York. [MR2920735](#)
- HORVÁTH, L. and REEDER, R. (2013). A test of significance in functional quadratic regression. *Bernoulli* **19** 2120–2151. [MR3129046](#)
- KOKOSZKA, P., MASLOVA, I., SOJKA, J. and ZHU, L. (2008). Testing for lack of dependence in the functional linear model. *Canad. J. Statist.* **36** 207–222. [MR2431682](#)
- LAVERGNE, P. and PATILEA, V. (2008). Breaking the curse of dimensionality in nonparametric testing. *J. Econometrics* **143** 103–122. [MR2384435](#)
- MCQUARRIE, A. D. (1999). A small-sample correction for the Schwarz SIC model selection criterion. *Statist. Probab. Lett.* **44** 79–86. [MR1706323](#)

- NADARAJA, È. A. (1964). On a regression estimate. *Teor. Veroyatn. Primen.* **9** 157–159. [MR0166874](#)
- PATILEA, V., SÁNCHEZ-SELLERO, C. and SAUMARD, M. (2012). Projection-based nonparametric testing for functional covariate effect. Preprint. Available at [arXiv:1205.5578](#).
- PATILEA, V., SÁNCHEZ-SELLERO, C. and SAUMARD, M. (2016). Testing the predictor effect on a functional response. *J. Amer. Statist. Assoc.* **111** 1684–1695. [MR3601727](#)
- RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. Springer, New York. [MR2168993](#)
- STUTE, W. (1997). Nonparametric model checks for regression. *Ann. Statist.* **25** 613–641. [MR1439316](#)
- STUTE, W., GONZÁLEZ MANTEIGA, W. and PRESEDO QUINDIMIL, M. (1998). Bootstrap approximations in model checks for regression. *J. Amer. Statist. Assoc.* **93** 141–149. [MR1614600](#)
- WATSON, G. S. (1964). Smooth regression analysis. *Sankhya, Ser. A* **26** 359–372. [MR0185765](#)
- ZHENG, J. X. (1996). A consistent test of functional form via nonparametric estimation techniques. *J. Econometrics* **75** 263–289. [MR1413644](#)

CONVOLVED SUBSAMPLING ESTIMATION WITH APPLICATIONS TO BLOCK BOOTSTRAP

BY JOHANNES TEWES¹, DIMITRIS N. POLITIS AND DANIEL J. NORDMAN²

*Ruhr-Universität Bochum, University of California, San Diego and
 Iowa State University*

The block bootstrap approximates sampling distributions from dependent data by resampling data blocks. A fundamental problem is establishing its consistency for the distribution of a sample mean, as a prototypical statistic. We use a structural relationship with subsampling to characterize the bootstrap in a new and general manner. While subsampling and block bootstrap differ, the block bootstrap distribution of a sample mean equals that of a k -fold self-convolution of a subsampling distribution. Motivated by this, we provide simple necessary and sufficient conditions for a convolved subsampling estimator to produce a normal limit that matches the target of bootstrap estimation. These conditions may be linked to consistency properties of an original subsampling distribution, which are often obtainable under minimal assumptions. Through several examples, the results are shown to validate the block bootstrap for means under significantly weakened assumptions in many existing (and some new) dependence settings, which also addresses a standing conjecture of Politis, Romano and Wolf [*Subsampling* (1999) Springer]. Beyond sample means, convolved subsampling may not match the block bootstrap, but instead provides an alternative resampling estimator that may be of interest. Under minimal dependence conditions, results also broadly establish convolved subsampling for general statistics having normal limits.

REFERENCES

- [1] ATHREYA, K. B. and LAHIRI, S. N. (2006). *Measure Theory and Probability Theory*. Springer Texts in Statistics. Springer, New York. [MR2247694](#)
- [2] BAI, S. and TAQQU, M. S. (2017). On the validity of resampling methods under long memory. *Ann. Statist.* **45** 2365–2399. [MR3737895](#)
- [3] BERAN, J. (1994). *Statistics for Long-Memory Processes*. Monographs on Statistics and Applied Probability **61**. Chapman & Hall, New York. [MR1304490](#)
- [4] BETKEN, A. and WENDLER, M. (2018). Subsampling for general statistics under long range dependence. *Statist. Sinica*. **28** 1199–1224. [MR3821001](#)
- [5] BICKEL, P. J. and FREEDMAN, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** 1196–1217. [MR0630103](#)
- [6] CAMBANIS, S., HOUDRÉ, C., HURD, H. and LEŚKOW, J. (1994). Laws of large numbers for periodically and almost periodically correlated processes. *Stochastic Process. Appl.* **53** 37–54. [MR1290706](#)
- [7] CORDUNEANU, C. (1989). *Almost Periodic Functions*. Chelsea, New York.

MSC2010 subject classifications. Primary 62G09; secondary 62G20, 62J05, 62M10.

Key words and phrases. Convolution, mixing, moving blocks, nonstationary.

- [8] DAVYDOV, Y. A. (1970). The invariance principle for stationary processes. *Theory Probab. Appl.* **15** 487–498.
- [9] DEHLING, H. and WENDLER, M. (2010). Central limit theorem and the bootstrap for U -statistics of strongly mixing data. *J. Multivariate Anal.* **101** 126–137. [MR2557623](#)
- [10] DOBRUSHIN, R. L. and MAJOR, P. (1979). Non-central limit theorems for nonlinear functionals of Gaussian fields. *Z. Wahrsch. Verw. Gebiete* **50** 27–52. [MR0550122](#)
- [11] EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26. [MR0515681](#)
- [12] FITZENBERGER, B. (1998). The moving blocks bootstrap and robust inference for linear least squares and quantile regressions. *J. Econometrics* **82** 235–287. [MR1613422](#)
- [13] HOSKING, J. R. M. (1981). Fractional differencing. *Biometrika* **68** 165–176. [MR0614953](#)
- [14] HURD, H. L. (1991). Correlation theory of almost periodically correlated processes. *J. Multivariate Anal.* **37** 24–45. [MR1097303](#)
- [15] TEWES, J., POLITIS, D. N. and NORDMAN, D. J. (2019). Supplement to “Convolved subsampling estimation with applications to block bootstrap.” DOI:[10.1214/18-AOS1695SUPP](#).
- [16] KIM, Y. M. and NORDMAN, D. J. (2011). Properties of a block bootstrap under long-range dependence. *Sankhya A* **73** 79–109. [MR2887088](#)
- [17] KÜNSCH, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Ann. Statist.* **17** 1217–1241. [MR1015147](#)
- [18] LAHIRI, S. N. (1993). On the moving block bootstrap under long range dependence. *Statist. Probab. Lett.* **18** 405–413. [MR1247453](#)
- [19] LAHIRI, S. N. (2003). *Resampling Methods for Dependent Data*. Springer Series in Statistics. Springer, New York. [MR2001447](#)
- [20] LAHIRI, S. N. (2003). Central limit theorems for weighted sums of a spatial process under a class of stochastic and fixed designs. *Sankhya* **65** 356–388. [MR2028905](#)
- [21] LENART, Ł. (2011). Asymptotic distributions and subsampling in spectral analysis for almost periodically correlated time series. *Bernoulli* **17** 290–319. [MR2797993](#)
- [22] LENART, Ł. (2016). Generalized resampling scheme with application to spectral density matrix in almost periodically correlated class of times series. *J. Time Series Anal.* **37** 369–404. [MR3512963](#)
- [23] LEUCHT, A. (2012). Degenerate U - and V -statistics under weak dependence: Asymptotic theory and bootstrap consistency. *Bernoulli* **18** 552–585. [MR2922461](#)
- [24] LIU, R. Y. (1988). Bootstrap procedures under some non-i.i.d. models. *Ann. Statist.* **16** 1696–1708. [MR0964947](#)
- [25] LIU, R. Y. and SINGH, K. (1992). Moving blocks jackknife and bootstrap capture weak dependence. In *Exploring the Limits of Bootstrap (East Lansing, MI, 1990)*. Wiley Ser. Probab. Math. Statist. Probab. Math. Statist. 225–248. Wiley, New York. [MR1197787](#)
- [26] MANDELBROT, B. B. and VAN NESS, J. W. (1968). Fractional Brownian motions, fractional noises and applications. *SIAM Rev.* **10** 422–437. [MR0242239](#)
- [27] NORDMAN, D. J. and LAHIRI, S. N. (2005). Validity of the sampling window method for long-range dependent linear processes. *Econometric Theory* **21** 1087–1111. [MR2200986](#)
- [28] POLITIS, D. N. and ROMANO, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.* **22** 2031–2050. [MR1329181](#)
- [29] POLITIS, D. N., ROMANO, J. P. and WOLF, M. (1999). *Subsampling*. Springer Series in Statistics. Springer, New York. [MR1707286](#)
- [30] RADULOVIĆ, D. (1996). The bootstrap of the mean for strong mixing sequences under minimal conditions. *Statist. Probab. Lett.* **28** 65–72. [MR1394420](#)
- [31] RADULOVIĆ, D. (2012). Necessary and sufficient conditions for the moving blocks bootstrap central limit theorem of the mean. *J. Nonparametr. Stat.* **24** 343–357. [MR2921140](#)
- [32] SHARIPOV, O. SH., TEWES, J. and WENDLER, M. (2016). Bootstrap for U -statistics: A new approach. *J. Nonparametr. Stat.* **28** 576–594. [MR3514512](#)

- [33] SHARIPOV, O. SH. and WENDLER, M. (2012). Bootstrap for the sample mean and for U -statistics of mixing and near-epoch dependent processes. *J. Nonparametr. Stat.* **24** 317–342. [MR2921139](#)
- [34] SYNOWIECKI, R. (2007). Consistency and application of moving block bootstrap for non-stationary time series with periodic and almost periodic structure. *Bernoulli* **13** 1151–1178. [MR2364230](#)
- [35] TAQQU, M. S. (1974/75). Weak convergence to fractional Brownian motion and to the Rosenblatt process. *Z. Wahrsch. Verw. Gebiete* **31** 287–302. [MR0400329](#)
- [36] ZHANG, T., HO, H.-C., WENDLER, M. and WU, W. B. (2013). Block sampling under strong dependence. *Stochastic Process. Appl.* **123** 2323–2339. [MR3038507](#)

FEATURE ELIMINATION IN KERNEL MACHINES IN MODERATELY HIGH DIMENSIONS¹

BY SAYAN DASGUPTA^{*}, YAIR GOLDBERG[†] AND MICHAEL R. KOSOROK^{*}

University of North Carolina at Chapel Hill^{} and University of Haifa[†]*

We develop an approach for feature elimination in statistical learning with kernel machines, based on recursive elimination of features. We present theoretical properties of this method and show that it is uniformly consistent in finding the correct feature space under certain generalized assumptions. We present a few case studies to show that the assumptions are met in most practical situations and present simulation results to demonstrate performance of the proposed approach.

REFERENCES

- AKSU, Y. (2014). Fast SVM-based feature elimination utilizing data radius, hard-margin, soft-margin. Preprint. Available at [arXiv:1210.4460v4](https://arxiv.org/abs/1210.4460v4).
- AKSU, Y., MILLER, D. J., KESIDIS, G. and YANG, Q. X. (2010). Margin-maximizing feature elimination methods for linear and nonlinear kernel-based discriminant functions. *IEEE Trans. Neural Netw.* **21** 701–717.
- ALLEN, G. I. (2013). Automatic feature selection via weighted kernels and regularization. *J. Comput. Graph. Statist.* **22** 284–299. [MR3173715](https://doi.org/10.1198/106186012E118)
- BRADLEY, P. S. and MANGASARIAN, O. L. (1998). Feature selection via concave minimization and support vector machines. In *Machine Learning Proceedings of the Fifteenth International Conference (ICML 1998)* 82–90. Morgan Kaufmann, San Francisco, CA.
- CHAPELLE, O., HAFNER, P. and VAPNIK, V. N. (1999). Support vector machines for histogram-based image classification. *IEEE Trans. Neural Netw.* **10** 1055–1064.
- DASGUPTA, S., GOLDBERG, Y. and KOSOROK, M. R. (2019). Supplement to “Feature elimination in kernel machines in moderately high dimensions.” DOI:[10.1214/18-AOS1696SUPP](https://doi.org/10.1214/18-AOS1696SUPP).
- 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](https://doi.org/10.1111/j.1467-9868.2007.00564.x)
- GOLDBERG, Y. and KOSOROK, M. R. (2017). Support vector regression for right censored data. *Electron. J. Stat.* **11** 532–569. [MR3619316](https://doi.org/10.1214/15-EJS1036)
- GUYON, I., WESTON, J., BARNHILL, S. and VAPNIK, V. (2002). Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46** 389–422.
- HIDALGO-MUÑOZ, A. R., LÓPEZ, M. M., SANTOS, I. M., PEREIRA, A. T., VÁZQUEZ-MARRUFO, M., GALVAO-CARMONA, A. and TOMÉ, A. M. (2013). Application of SVM-RFE on EEG signals for detecting the most relevant scalp regions linked to affective valence processing. *Expert Syst. Appl.* **40** 2102–2108.
- HU, X., SCHWARZ, J. K., LEWIS, J. S., HUETTNER, P. C., RADER, J. S., DEASY, J. O., GRIGSBY, P. W. and WANG, X. (2010). A microRNA expression signature for cervical cancer prognosis. *Cancer Res.* **70** 1441–1448.

MSC2010 subject classifications. 62G20.

Key words and phrases. Kernel machines, support vector machines, variable selection, recursive feature elimination.

- JOLLIFFE, I. T. (2002). *Principal Component Analysis*, 2nd ed. Springer, New York. [MR2036084](#)
- LESLIE, C. S., ESKIN, E., COHEN, A., WESTON, J. and NOBLE, W. S. (2004). Mismatch string kernels for discriminative protein classification. *Bioinformatics* **20** 467–476.
- LI, R., ZHONG, W. and ZHU, L. (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc.* **107** 1129–1139. [MR3010900](#)
- LIU, Y., ZHANG, H. H., PARK, C. and AHN, J. (2007). Support vector machines with adaptive L_q penalty. *Comput. Statist. Data Anal.* **51** 6380–6394. [MR2408601](#)
- LOUW, N. and STEEL, S. J. (2006). Variable selection in kernel Fisher discriminant analysis by means of recursive feature elimination. *Comput. Statist. Data Anal.* **51** 2043–2055. [MR2307560](#)
- MICCHELLI, C. A., XU, Y. and ZHANG, H. (2006). Universal kernels. *J. Mach. Learn. Res.* **7** 2651–2667. [MR2274454](#)
- MUNDRA, P. and RAJAPAKSE, J. C. (2010). SVM-RFE with MRMR filter for gene selection. *IEEE Trans. Nanobiosci.* **9** 31–37.
- RAKOTOMAMONJY, A. (2003). Variable selection using SVM-based criteria. *J. Mach. Learn. Res.* **3** 1357–1370. [MR2020764](#)
- SCHIELE, B. and CROWLEY, J. L. (1996). Object recognition using multidimensional receptive field histograms. In *Computer Vision ECCV'96* 610–619. Springer, Berlin.
- STEINWART, I. and CHRISTMANN, A. (2008). *Support Vector Machines*. Springer, New York. [MR2450103](#)
- STEINWART, I. and SCOVEL, C. (2007). Fast rates for support vector machines using Gaussian kernels. *Ann. Statist.* **35** 575–607. [MR2336860](#)
- SWAIN, M. J. and BALLARD, D. H. (1992). Indexing via color histograms. In *Active Perception and Robot Vision* 261–273. Springer, Berlin.
- TANG, Y., ZHANG, Y. Q. and HUANG, Z. (2007). Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **4** 365–381.
- TSYBAKOV, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32** 135–166. [MR2051002](#)
- WANG, L., ZHU, J. and ZOU, H. (2006). The doubly regularized support vector machine. *Statist. Sinica* **16** 589–615. [MR2267251](#)
- WESTON, J., ELISSEEFF, A., SCHÖLKOPF, B. and TIPPING, M. (2003). Use of the zero-norm with linear models and kernel methods. *J. Mach. Learn. Res.* **3** 1439–1461. [MR1983013](#)
- ZHANG, H. H., AHN, J., LIN, X. and PARK, C. (2006a). Gene selection using support vector machines with non-convex penalty. *Bioinformatics* **22** 88–95.
- ZHANG, X., LU, X., SHI, Q., XU, X., HON-CHIU, E. L., HARRIS, L. N., IGLEHART, J. D., MIRON, A., LIU, J. S. and WONG, W. H. (2006b). Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinform.* **7** Art. ID 197.
- ZHU, J., ROSSET, S., HASTIE, T. and TIBSHIRANI, R. (2003). 1-norm support vector machines. In *Neural Information Processing Systems* 16 49–56. MIT Press, Cambridge, MA.

HIGH-DIMENSIONAL COVARIANCE MATRICES IN ELLIPTICAL DISTRIBUTIONS WITH APPLICATION TO SPHERICAL TEST

BY JIANG HU^{*,†,1}, WEIMING LI^{‡,2}, ZHI LIU^{§,3} AND WANG ZHOU^{¶,4}

Northeast Normal University^{*}, *Key Laboratory for Applied Statistics of MOE*[†],
Shanghai University of Finance and Economics[‡], *University of Macau*[§] and
National University of Singapore[¶]

This paper discusses fluctuations of linear spectral statistics of high-dimensional sample covariance matrices when the underlying population follows an elliptical distribution. Such population often possesses high order correlations among their coordinates, which have great impact on the asymptotic behaviors of linear spectral statistics. Taking such kind of dependency into consideration, we establish a new central limit theorem for the linear spectral statistics in this paper for a class of elliptical populations. This general theoretical result has wide applications and, as an example, it is then applied to test the sphericity of elliptical populations.

REFERENCES

- ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley, Hoboken, NJ. [MR1990662](#)
- BAI, Z. D. and SILVERSTEIN, J. W. (1998). No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *Ann. Probab.* **26** 316–345. [MR1617051](#)
- BAI, Z. D. and SILVERSTEIN, J. W. (2004). CLT for linear spectral statistics of large-dimensional sample covariance matrices. *Ann. Probab.* **32** 553–605. [MR2040792](#)
- BAI, Z. D. and SILVERSTEIN, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*, 2nd ed. Springer, New York. [MR2567175](#)
- BAI, Z. D. and ZHOU, W. (2008). Large sample covariance matrices without independence structures in columns. *Statist. Sinica* **18** 425–442. [MR2411613](#)
- BAI, Z. D., JIANG, D. D., YAO, J.-F. and ZHENG, S. R. (2009). Corrections to LRT on large-dimensional covariance matrix by RMT. *Ann. Statist.* **37** 3822–3840. [MR2572444](#)
- BAI, Z. D., HU, J., PAN, G. M. and ZHOU, W. (2015). Convergence of the empirical spectral distribution function of Beta matrices. *Bernoulli* **21** 1538–1574. [MR3352053](#)
- BERTHET, Q. and RIGOLLET, P. (2013). Optimal detection of sparse principal components in high dimension. *Ann. Statist.* **41** 1780–1815. [MR3127849](#)
- BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York. [MR0233396](#)
- BILLINGSLEY, P. (1995). *Probability and Measure*, 3rd ed. Wiley, New York. [MR1324786](#)
- DONOHU, D. and JIN, J. (2015). Higher criticism for large-scale inference, especially for rare and weak effects. *Statist. Sci.* **30** 1–25. [MR3317751](#)

MSC2010 subject classifications. Primary 62H15; secondary 62H10.

Key words and phrases. Covariance matrix, high-dimensional data, elliptical distribution, sphericity test.

- EL KAROUI, N. (2009). Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond. *Ann. Appl. Probab.* **19** 2362–2405. [MR2588248](#)
- FANG, K. T. and ZHANG, Y. T. (1990). *Generalized Multivariate Analysis*. Springer, Berlin. [MR1079542](#)
- GAO, J. T., HAN, X., PAN, G. M. and YANG, Y. R. (2017). High dimensional correlation matrices: The central limit theorem and its applications. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 677–693. [MR3641402](#)
- GUPTA, A. K., VARGA, T. and BODNAR, T. (2013). *Elliptically Contoured Models in Statistics and Portfolio Theory*, 2nd ed. Springer, New York. [MR3112145](#)
- HU, J., LI, W. M., LIU, Z. and ZHOU, W. (2019). Supplement to “High-dimensional covariance matrices in elliptical distributions with application to spherical test.” DOI:[10.1214/18-AOS1699SUPP](#).
- JOHN, S. (1972). The distribution of a statistic used for testing sphericity of normal distributions. *Biometrika* **59** 169–173. [MR0312619](#)
- JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. [MR1863961](#)
- JONSSON, D. (1982). Some limit theorems for the eigenvalues of a sample covariance matrix. *J. Multivariate Anal.* **12** 1–38. [MR0650926](#)
- KELKER, D. (1970). Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhyā Ser. A* **32** 419–438. [MR0287628](#)
- KOTZ, S. (1975). Multivariate distributions at a cross-road. In *Statistical Distributions in Scientific Work* **1** 247–270.
- LEDOIT, O. and WOLF, M. (2002). Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. *Ann. Statist.* **30** 1081–1102. [MR1926169](#)
- LI, W. M. and YAO, J. F. (2018). On structure testing for component covariance matrices of a high-dimensional mixture. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 293–318.
- ONATSKI, A., MOREIRA, M. J. and HALLIN, M. (2013). Asymptotic power of sphericity tests for high-dimensional data. *Ann. Statist.* **41** 1204–1231. [MR3113808](#)
- ONATSKI, A., MOREIRA, M. J. and HALLIN, M. (2014). Signal detection in high dimension: The multitasked case. *Ann. Statist.* **42** 225–254. [MR3189485](#)
- PAINDAVEINE, D. and VERDEBOUT, T. (2016). On high-dimensional sign tests. *Bernoulli* **22** 1745–1769. [MR3474832](#)
- PAN, G. M. and ZHOU, W. (2008). Central limit theorem for signal-to-interference ratio of reduced rank linear receiver. *Ann. Appl. Probab.* **18** 1232–1270. [MR2418244](#)
- RAMA, C. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quant. Finance* **1** 223–236.
- 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](#)
- SILVERSTEIN, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. *J. Multivariate Anal.* **55** 331–339. [MR1370408](#)
- SILVERSTEIN, J. W. and BAI, Z. D. (1995). On the empirical distribution of eigenvalues of a class of large-dimensional random matrices. *J. Multivariate Anal.* **54** 175–192. [MR1345534](#)
- TIAN, X. T., LU, Y. T. and LI, W. M. (2015). A robust test for sphericity of high-dimensional covariance matrices. *J. Multivariate Anal.* **141** 217–227. [MR3390068](#)
- WANG, Q. W. and YAO, J. F. (2013). On the sphericity test with large-dimensional observations. *Electron. J. Stat.* **7** 2164–2192. [MR3104916](#)
- YANG, Y. R. and PAN, G. M. (2015). Independence test for high dimensional data based on regularized canonical correlation coefficients. *Ann. Statist.* **43** 467–500. [MR3316187](#)

- YIN, Y. Q., BAI, Z. D. and KRISHNAIAH, P. R. (1988). On the limit of the largest eigenvalue of the large-dimensional sample covariance matrix. *Probab. Theory Related Fields* **78** 509–521. [MR0950344](#)
- ZOU, C. L., PENG, L. H., FENG, L. and WANG, Z. J. (2014). Multivariate sign-based high-dimensional tests for sphericity. *Biometrika* **101** 229–236. [MR3180668](#)

A CRITICAL THRESHOLD FOR DESIGN EFFECTS IN NETWORK SAMPLING

BY KARL ROHE¹

University of Wisconsin–Madison

Web crawling, snowball sampling, and respondent-driven sampling (RDS) are three types of network sampling techniques used to contact individuals in hard-to-reach populations. This paper studies these procedures as a Markov process on the social network that is indexed by a tree. Each node in this tree corresponds to an observation and each edge in the tree corresponds to a referral. Indexing with a tree (instead of a chain) allows for the sampled units to refer multiple future units into the sample.

In survey sampling, the design effect characterizes the additional variance induced by a novel sampling strategy. If the design effect is some value DE , then constructing an estimator from the novel design makes the variance of the estimator DE times greater than it would be under a simple random sample with the same sample size n . Under certain assumptions on the referral tree, the design effect of network sampling has a critical threshold that is a function of the referral rate m and the clustering structure in the social network, represented by the second eigenvalue of the Markov transition matrix, λ_2 . If $m < 1/\lambda_2^2$, then the design effect is finite (i.e., the standard estimator is $\sqrt{n}$ -consistent). However, if $m > 1/\lambda_2^2$, then the design effect grows with n (i.e., the standard estimator is no longer $\sqrt{n}$ -consistent). Past this critical threshold, the standard error of the estimator converges at the slower rate of $n^{\log_m \lambda_2}$. The Markov model allows for nodes to be resampled; computational results show that the findings hold in without-replacement sampling. To estimate confidence intervals that adapt to the correct level of uncertainty, a novel resampling procedure is proposed. Computational experiments compare this procedure to previous techniques.

REFERENCES

- ABDUL-QUADER, A. S., HECKATHORN, D. D., MCKNIGHT, C., BRAMSON, H., NEMETH, C., SABIN, K., GALLAGHER, K. and DES JARLAIS, D. C. (2006). Effectiveness of respondent-driven sampling for recruiting drug users in New York City: Findings from a pilot study. *J. Urban Health* **83** 459–476.
- ARAYASIRIKUL, S., CAI, X. and WILSON, E. C. (2015). A qualitative examination of respondent-driven sampling (RDS) peer referral challenges among young transwomen in the San Francisco bay area. *JMIR Public Health Surveill.* **1** e9.
- ATHREYA, K. B. and NEY, P. E. (1972). *Branching Processes. Die Grundlehren der Mathematischen Wissenschaften* **196**. Springer, New York. [MR0373040](#)

MSC2010 subject classifications. Primary 62D99; secondary 60J20.

Key words and phrases. Stochastic blockmodel, social network, link-tracing, Galton–Watson.

- BARAFF, A. J., MCCORMICK, T. H. and RAFTERY, A. E. (2016). Estimating uncertainty in respondent-driven sampling using a tree bootstrap method. *Proc. Natl. Acad. Sci. USA* 201617258.
- BENJAMINI, I. and PERES, Y. (1994). Markov chains indexed by trees. *Ann. Probab.* **22** 219–243. [MR1258875](#)
- CHUNG, F. R. K. (1997). *Spectral Graph Theory*. CBMS Regional Conference Series in Mathematics **92**. Published for the Conference Board of the Mathematical Sciences, Washington, DC; by the Amer. Math. Soc., Providence, RI. [MR1421568](#)
- GILE, K. J. (2011). Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. *J. Amer. Statist. Assoc.* **106** 135–146. [MR2816708](#)
- GILE, K. J. and HANDCOCK, M. S. (2010). Respondent-driven sampling: An assessment of current methodology. *Sociol. Method.* **40** 285–327.
- GILE, K. J., JOHNSTON, L. G. and SALGANIK, M. J. (2015). Diagnostics for respondent-driven sampling. *J. Roy. Statist. Soc. Ser. A* **178** 241–269. [MR3291770](#)
- GOEL, S. and SALGANIK, M. J. (2009). Respondent-driven sampling as Markov chain Monte Carlo. *Stat. Med.* **28** 2202–2229. [MR2751515](#)
- GOEL, S. and SALGANIK, M. J. (2010). Assessing respondent-driven sampling. *Proc. Natl. Acad. Sci. USA* **107** 6743–6747.
- HANDCOCK, M. S., FELLOWS, I. E. and GILE, K. J. (2016). RDS: Respondent-driven sampling. Los Angeles, CA, R package version 0.7-5. <http://CRAN.R-project.org/package=RDS>.
- HECKATHORN, D. D. (1997). Respondent-driven sampling: A new approach to the study of hidden populations. *Soc. Probl.* **44** 174–199.
- HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. [MR0718088](#)
- JOHNSTON, L. G., CHEN, Y.-H., SILVA-SANTISTEBAN, A. and RAYMOND, H. F. (2013). An empirical examination of respondent driven sampling design effects among HIV risk groups from studies conducted around the world. *AIDS Behav.* **17** 2202–2210.
- KHABBAZIAN, M., HANLON, B., RUSSEK, Z. and ROHE, K. (2017). Novel sampling design for respondent-driven sampling. *Electron. J. Stat.* **11** 4769–4812.
- LEVIN, D. A., PERES, Y. and WILMER, E. L. (2009). *Markov Chains and Mixing Times*. Amer. Math. Soc., Providence, RI. [MR2466937](#)
- LI, X. and ROHE, K. (2017). Central limit theorems for network driven sampling. *Electron. J. Stat.* **11** 4871–4895. [MR3733297](#)
- LU, X., BENGTSOON, L., BRITTON, T., CAMITZ, M., KIM, B. J., THORSON, A. and LILJEROS, F. (2012). The sensitivity of respondent-driven sampling. *J. Roy. Statist. Soc. Ser. A* **175** 191–216. [MR2873802](#)
- MCCREESH, N., FROST, S., SEELEY, J., KATONGOLE, J., TARSH, M. N., NDUNGUSE, R., JICHI, F., LUNEL, N. L., MAHER, D., JOHNSTON, L. G. et al. (2012). Evaluation of respondent-driven sampling. *Epidemiology* **23** 138.
- ROCH, S. and ROHE, K. (2017). Generalized least squares can overcome the critical threshold in respondent-driven sampling. ArXiv Preprint [ArXiv:1708.04999](#).
- ROHE, K. (2019). Supplement to “A critical threshold for design effects in network sampling.” DOI:[10.1214/18-AOS1700SUPP](#).
- ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915. [MR2893856](#)
- SALGANIK, M. J. (2006). Variance estimation, design effects, and sample size calculations for respondent-driven sampling. *J. Urban Health* **83** 98–112.
- SALGANIK, M. J. and HECKATHORN, D. D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociol. Method.* **34** 193–240.

- SZWARCWALD, C. L., DE SOUZA JÚNIOR, P. R. B., DAMACENA, G. N., JUNIOR, A. B. and KENDALL, C. (2011). Analysis of data collected by RDS among sex workers in 10 Brazilian cities, 2009: Estimation of the prevalence of HIV, variance, and design effect. *JAIDS J. Acquir. Immune Defic. Syndr.* **57** S129–S135.
- VERDERY, A. M., MOUW, T., BAULDRY, S. and MUCHA, P. J. (2015). Network structure and biased variance estimation in respondent driven sampling. *PLoS ONE* **10** e0145296.
- VOLZ, E. and HECKATHORN, D. D. (2008). Probability based estimation theory for respondent driven sampling. *J. Off. Stat.* **24** 79.
- VON LUXBURG, U. (2007). A tutorial on spectral clustering. *Stat. Comput.* **17** 395–416. [MR2409803](#)
- WHITE, R. G., HAKIM, A. J., SALGANIK, M. J., SPILLER, M. W., JOHNSTON, L. G., KERR, L., KENDALL, C., DRAKE, A., WILSON, D., ORROTH, K. et al. (2015). Strengthening the reporting of observational studies in epidemiology for respondent-driven sampling studies: STROBE-RDS statement. *J. Clin. Epidemiol.* **68** 1463–1471.
- WORLD HEALTH ORGANIZATION and UNAIDS (2013). *Introduction To HIV/AIDS And Sexually Transmitted Infection Surveillance Module 4: Introduction to Respondent-drive Sampling*. World Health Organization & UNAIDS. http://applications.emro.who.int/dsaf/EMRPUB_2013_EN_1539.pdf.

PERMUTATION p -VALUE APPROXIMATION VIA GENERALIZED STOLARSKY INVARIANCE¹

BY HERA Y. HE^{*}, KINJAL BASU[†], QINGYUAN ZHAO[‡] AND ART B. OWEN^{*}

Stanford University^{}, LinkedIn[†] and University of Pennsylvania[‡]*

It is common for genomic data analysis to use p -values from a large number of permutation tests. The multiplicity of tests may require very tiny p -values in order to reject any null hypotheses and the common practice of using randomly sampled permutations then becomes very expensive. We propose an inexpensive approximation to p -values for two sample linear test statistics, derived from Stolarsky's invariance principle. The method creates a geometrically derived reference set of approximate p -values for each hypothesis. The average of that set is used as a point estimate $\hat{p}$ and our generalization of the invariance principle allows us to compute the variance of the p -values in that set. We find that in cases where the point estimate is small, the variance is a modest multiple of the square of that point estimate, yielding a relative error property similar to that of saddlepoint approximations. On a Parkinson's disease data set, the new approximation is faster and more accurate than the saddlepoint approximation. We also obtain a simple probabilistic explanation of Stolarsky's invariance principle.

REFERENCES

- ACKERMANN, M. and STRIMMER, K. (2009). A general modular framework for gene set enrichment analysis. *BMC Bioinform.* **10** 1–20.
- BARNARD, G. A. (1963). Discussion of the spectral analysis of point processes (by M. S. Bartlett). *J. Roy. Statist. Soc. Ser. B* **25** 294. [MR0171334](#)
- BILYK, D., DAI, F. and MATZKE, R. (2018). Stolarsky principle and energy optimization on the sphere. *Constr. Approx.* **48** 31–60.
- BLANCHET, J. and GLYNN, P. (2008). Efficient rare-event simulation for the maximum of heavy-tailed random walks. *Ann. Appl. Probab.* **18** 1351–1378. [MR2434174](#)
- BRAUCHART, J. S. and DICK, J. (2013). A simple proof of Stolarsky's invariance principle. *Proc. Amer. Math. Soc.* **141** 2085–2096. [MR3034434](#)
- FADISTA, J., MANNING, A. K., FLOREZ, J. C. and GROOP, L. (2016). The (in) famous GWAS p -value threshold revisited and updated for low-frequency variants. *Eur. J. Hum. Genet.* **24** 1202–1205.
- HE, H. Y. (2016). Efficient permutation-based p -value estimation for gene set tests. Ph.D. thesis, Stanford Univ., Stanford, CA.
- HE, H. Y., BASU, K., ZHAO, Q. and OWEN, A. B. (2019). Supplement to “Permutation p -value approximation via generalized Stolarsky invariance.” DOI:[10.1214/18-AOS1702SUPP](#).
- JIANG, Z. and GENTLEMAN, R. (2005). Reproducible research: A bioinformatics case study. *Bioinformatics* **23** 306–313.

MSC2010 subject classifications. Primary 62G10; secondary 11K38, 62G09.

Key words and phrases. Discrepancy, gene sets, hypothesis testing, quasi-Monte Carlo.

- KNIJNENBURG, T. A., WESSELS, L. F. A., REINDERS, M. J. T. and SHMULEVICH, I. (2009). Fewer permutations, more accurate p -values. *Bioinformatics* **25** i161–i168.
- LARSON, J. L. and OWEN, A. B. (2015). Moment based gene set tests. *BMC Bioinform.* **16** 132.
- LEE, Y. and KIM, W. C. (2014). Concise formulas for the surface area of the intersection of two hyperspherical caps. Technical report, Korea Advanced Institute of Science and Technology.
- LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. Springer, New York. [MR2135927](#)
- MARDIA, K. V. and JUPP, P. E. (2000). *Directional Statistics*. Wiley, Chichester. [Revised reprint of Mardia, K. V. (1972). *Statistics of Directional Data. Probability and Mathematical Statistics* **13**. Academic Press, London. [MR0336854](#).] [MR1828667](#)
- MORAN, L. B., DUKE, D. C., DEPREZ, M., DEXTER, D. T., PEARCE, R. K. B. and GRAEBER, M. B. (2006). Whole genome expression profiling of the medial and lateral substantia nigra in Parkinson's. *Disease Neurogenet.* **7** 1–11.
- NARCOWICH, F. J., SUN, X., WARD, J. D. and WU, Z. (2010). LeVeque type inequalities and discrepancy estimates for minimal energy configurations on spheres. *J. Approx. Theory* **162** 1256–1278. [MR2643729](#)
- NIEDERREITER, H. (1992). *Random Number Generation and Quasi-Monte Carlo Methods. CBMS-NSF Regional Conference Series in Applied Mathematics* **63**. SIAM, Philadelphia, PA. [MR1172997](#)
- R CORE TEAM (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- REID, N. (1988). Saddlepoint methods and statistical inference. *Statist. Sci.* **3** 213–238. [MR0968390](#)
- ROBINSON, J. (1982). Saddlepoint approximations for permutation tests and confidence intervals. *J. Roy. Statist. Soc. Ser. B* **44** 91–101. [MR0655378](#)
- SCHERZER, C. R., EKLUND, A. C., MORSE, L. J., LIAO, Z., LOCASCIO, J. J., FEFER, D., SCHWARZSCHILD, M. A., SCHLOSSMACHER, M. G., HAUSER, M. A., VANCE, J. M., SUDARSKY, L. R., STANDAERT, D. G., GROWDON, J. H., JENSEN, R. V. and GULLANS, S. R. (2007). Molecular markers of early Parkinson's disease based on gene expression in blood. *Proc. Natl. Acad. Sci. USA* **104** 955–960.
- STOLARSKY, K. B. (1973). Sums of distances between points on a sphere. II. *Proc. Amer. Math. Soc.* **41** 575–582. [MR0333995](#)
- TIAN, L., GREENBERG, S. A., KONG, S. W., ALTSCHULER, J., KOHANE, I. S. and PARK, P. J. (2005). Discovering statistically significant pathways in expression profiling studies. *Proc. Natl. Acad. Sci. USA* **102** 13544–13549.
- ZHANG, Y., JAMES, M., MIDDLETON, F. A. and DAVIS, R. L. (2005). Transcriptional analysis of multiple brain regions in Parkinson's disease supports the involvement of specific protein processing, energy metabolism, and signaling pathways, and suggests novel disease mechanisms. *Amer. J. Med. Genet. B Neuropsych. Genet.* **137B** 5–16.
- ZHOU, C., WANG, H. J. and WANG, Y. M. (2009). Efficient moments-based permutation tests. *Adv. Neural Inf. Process. Syst.* **22** 2277–2285. [MR2714169](#)

CANONICAL CORRELATION COEFFICIENTS OF HIGH-DIMENSIONAL GAUSSIAN VECTORS: FINITE RANK CASE

BY ZHIGANG BAO^{1,*}, JIANG HU^{2,†}, GUANGMING PAN^{3,‡} AND
WANG ZHOU^{4,§}

*Hong Kong University of Science and Technology**, *Northeast Normal University†*, *Nanyang Technological University‡* and *National University of Singapore§*

Consider a Gaussian vector $\mathbf{z} = (\mathbf{x}', \mathbf{y}')'$, consisting of two sub-vectors $\mathbf{x}$ and $\mathbf{y}$ with dimensions p and q , respectively. With n independent observations of $\mathbf{z}$, we study the correlation between $\mathbf{x}$ and $\mathbf{y}$, from the perspective of the canonical correlation analysis. We investigate the high-dimensional case: both p and q are proportional to the sample size n . Denote by Σ_{uv} the population cross-covariance matrix of random vectors $\mathbf{u}$ and $\mathbf{v}$, and denote by S_{uv} the sample counterpart. The canonical correlation coefficients between $\mathbf{x}$ and $\mathbf{y}$ are known as the square roots of the nonzero eigenvalues of the canonical correlation matrix $\Sigma_{xx}^{-1} \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma_{yx}$. In this paper, we focus on the case that Σ_{xy} is of finite rank k , that is, there are k nonzero canonical correlation coefficients, whose squares are denoted by $r_1 \geq \dots \geq r_k > 0$. We study the sample counterparts of r_i , $i = 1, \dots, k$, that is, the largest k eigenvalues of the sample canonical correlation matrix $S_{xx}^{-1} S_{xy} S_{yy}^{-1} S_{yx}$, denoted by $\lambda_1 \geq \dots \geq \lambda_k$. We show that there exists a threshold $r_c \in (0, 1)$, such that for each $i \in \{1, \dots, k\}$, when $r_i \leq r_c$, λ_i converges almost surely to the right edge of the limiting spectral distribution of the sample canonical correlation matrix, denoted by d_+ . When $r_i > r_c$, λ_i possesses an almost sure limit in $(d_+, 1]$, from which we can recover r_i 's in turn, thus provide an estimate of the latter in the high-dimensional scenario. We also obtain the limiting distribution of λ_i 's under appropriate normalization. Specifically, λ_i possesses Gaussian type fluctuation if $r_i > r_c$, and follows Tracy–Widom distribution if $r_i < r_c$. Some applications of our results are also discussed.

REFERENCES

- [1] ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ. [MR1990662](#)
- [2] BAI, Z., CHOI, K. P. and FUJIKOSHI, Y. (2018). Consistency of AIC and BIC in estimating the number of significant components in high-dimensional principal component analysis. *Ann. Statist.* **46** 1050–1076. [MR3797996](#)
- [3] BAI, Z., HU, J., PAN, G. and ZHOU, W. (2015). Convergence of the empirical spectral distribution function of Beta matrices. *Bernoulli* **21** 1538–1574. [MR3352053](#)

MSC2010 subject classifications. 62H20, 60B20, 60F99.

Key words and phrases. Canonical correlation analysis, random matrices, MANOVA ensemble, high-dimensional data, finite rank perturbation, largest eigenvalues.

- [4] BAI, Z. and YAO, J. (2008). Central limit theorems for eigenvalues in a spiked population model. *Ann. Inst. Henri Poincaré Probab. Stat.* **44** 447–474. [MR2451053](#)
- [5] BAIK, J., BEN AROUS, G. and PÉCHÉ, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *Ann. Probab.* **33** 1643–1697. [MR2165575](#)
- [6] BAIK, J. and SILVERSTEIN, J. W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *J. Multivariate Anal.* **97** 1382–1408. [MR2279680](#)
- [7] BAO, Z., HU, J., PAN, G. and ZHOU, W. (2017). Test of independence for high-dimensional random vectors based on freeness in block correlation matrices. *Electron. J. Stat.* **11** 1527–1548. [MR3635921](#)
- [8] BAO, Z. and HU, J. (2018) High-dimensional CCA with general population. (In preparation).
- [9] BAO, Z., HU, J., PAN, G. and ZHOU, W. (2019). Supplement to “Canonical correlation coefficients of high-dimensional Gaussian vectors: finite rank case.” DOI:[10.1214/18-AOS1704SUPP](#).
- [10] BELINSCHI, S. T., BERCOVICI, H., CAPITAINÉ, M. and FÉVRIER, M. (2017). Outliers in the spectrum of large deformed unitarily invariant models. *Ann. Probab.* **45** 3571–3625. [MR3729610](#)
- [11] BENAYCH-GEORGES, F., GUIONNET, A. and MAIDA, M. (2011). Fluctuations of the extreme eigenvalues of finite rank deformations of random matrices. *Electron. J. Probab.* **16** 1621–1662. [MR2835249](#)
- [12] BENAYCH-GEORGES, F. and NADAKUDITI, R. R. (2011). The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Adv. Math.* **227** 494–521. [MR2782201](#)
- [13] BRETHERTON, C. S., SMITH, C. and WALLACE, J. M. (1992). An intercomparison of methods for finding coupled patterns in climate data. *J. Climate* **5** 541–560.
- [14] CAPITAINÉ, M., DONATI-MARTIN, C. and FÉRAL, D. (2009). The largest eigenvalues of finite rank deformation of large Wigner matrices: Convergence and nonuniversality of the fluctuations. *Ann. Probab.* **37** 1–47. [MR2489158](#)
- [15] CAPITAINÉ, M., DONATI-MARTIN, C. and FÉRAL, D. (2012). Central limit theorems for eigenvalues of deformations of Wigner matrices. *Ann. Inst. Henri Poincaré Probab. Stat.* **48** 107–133. [MR2919200](#)
- [16] DAVIDSON, K. R. and SZAREK, S. J. (2001). Local operator theory, random matrices and Banach spaces. In *Handbook of the Geometry of Banach Spaces, Vol. I* 317–366. North-Holland, Amsterdam. [MR1863696](#)
- [17] DUTILLEUL, P., PELLETIER, B. and ALPARGU, G. (2008). Modified F tests for assessing the multiple correlation between one spatial process and several others. *J. Statist. Plann. Inference* **138** 1402–1415. [MR2388019](#)
- [18] EDELMAN, A. and RAO, N. R. (2005). Random matrix theory. *Acta Numer.* **14** 233–297. [MR2168344](#)
- [19] FÉRAL, D. and PÉCHÉ, S. (2007). The largest eigenvalue of rank one deformation of large Wigner matrices. *Comm. Math. Phys.* **272** 185–228. [MR2291807](#)
- [20] FÉRAL, D. and PÉCHÉ, S. (2009). The largest eigenvalues of sample covariance matrices for a spiked population: Diagonal case. *J. Math. Phys.* **50** 073302, 33. [MR2548630](#)
- [21] FUJIKOSHI, Y. (2016). High-Dimensional Asymptotic Distributions of Characteristic Roots in multivariate linear models and canonical correlation analysis. Technical report.
- [22] FUJIKOSHI, Y. and SAKURAI, T. (2016). High-dimensional consistency of rank estimation criteria in multivariate linear model. *J. Multivariate Anal.* **149** 199–212. [MR3507324](#)
- [23] GAO, C., MA, Z., REN, Z. and ZHOU, H. H. (2015). Minimax estimation in sparse canonical correlation analysis. *Ann. Statist.* **43** 2168–2197. [MR3396982](#)
- [24] GAO, C., MA, Z. and ZHOU, H. H. (2017). Sparse CCA: Adaptive estimation and computational barriers. *Ann. Statist.* **45** 2074–2101. [MR3718162](#)

- [25] GITTINS, R. (1985). *Canonical Analysis. Biomathematics* **12**. Springer, Berlin, Heidelberg.
- [26] HAN, X., PAN, G. and YANG, Q. (2018). A unified matrix model including both CCA and F matrices in multivariate analysis: The largest eigenvalue and its applications. *Bernoulli* **24** 3447–3468. [MR3788178](#)
- [27] HAN, X., PAN, G. and ZHANG, B. (2016). The Tracy–Widom law for the largest eigenvalue of F type matrices. *Ann. Statist.* **44** 1564–1592. [MR3519933](#)
- [28] HOTELLING, H. (1936). Relations between two sets of variates. *Biometrika* **28** 321–377.
- [29] HYODO, M., SHUTOH, N., NISHIYAMA, T. and PAVLENKO, T. (2015). Testing block-diagonal covariance structure for high-dimensional data. *Stat. Neerl.* **69** 460–482. [MR3414707](#)
- [30] JIANG, D., BAI, Z. and ZHENG, S. (2013). Testing the independence of sets of large-dimensional variables. *Sci. China Math.* **56** 135–147. [MR3016588](#)
- [31] JIANG, T. and YANG, F. (2013). Central limit theorems for classical likelihood ratio tests for high-dimensional normal distributions. *Ann. Statist.* **41** 2029–2074. [MR3127857](#)
- [32] JOHNSTONE, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295–327. [MR1863961](#)
- [33] JOHNSTONE, I. M. (2008). Multivariate analysis and Jacobi ensembles: Largest eigenvalue, Tracy–Widom limits and rates of convergence. *Ann. Statist.* **36** 2638–2716. [MR2485010](#)
- [34] JOHNSTONE, I. M. (2009). Approximate null distribution of the largest root in multivariate analysis. *Ann. Appl. Stat.* **3** 1616–1633. [MR2752150](#)
- [35] JOHNSTONE, I. M. and ONATSKI, A. (2015). Testing in high-dimensional spiked models. [arXiv:1509.07269](#).
- [36] KARGIN, V. (2015). Subordination for the sum of two random matrices. *Ann. Probab.* **43** 2119–2150. [MR3353823](#)
- [37] KATZ-MOSES, B. (2012). *Small Deviations for the Beta–Jacobi Ensemble*. ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)—University of Colorado at Boulder. [MR3034656](#)
- [38] KNOWLES, A. and YIN, J. (2013). The isotropic semicircle law and deformation of Wigner matrices. *Comm. Pure Appl. Math.* **66** 1663–1750. [MR3103909](#)
- [39] MUIRHEAD, R. J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York. [MR0652932](#)
- [40] ODA, R., YANAGIHARA, H. and FUJIKOSHI, Y. (2016). Asymptotic non-null distributions of test statistics for redundancy in the high-dimensional canonical correlation analysis Technical report.
- [41] PASSEMIER, D. and YAO, J.-F. (2012). On determining the number of spikes in a high-dimensional spiked population model. *Random Matrices Theory Appl.* **1** 1150002, 19. [MR2930380](#)
- [42] PAUL, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica* **17** 1617–1642. [MR2399865](#)
- [43] PÉCHÉ, S. (2006). The largest eigenvalue of small rank perturbations of Hermitian random matrices. *Probab. Theory Related Fields* **134** 127–173. [MR2221787](#)
- [44] RENCHER, A. C. and PUN, F. C. (1980). Inflation of R^2 in best subset regression. *Technometrics* **22** 49.
- [45] WACHTER, K. W. (1980). The limiting empirical measure of multiple discriminant ratios. *Ann. Statist.* **8** 937–957. [MR0585695](#)
- [46] WANG, Q. and YAO, J. (2017). Extreme eigenvalues of large-dimensional spiked Fisher matrices with application. *Ann. Statist.* **45** 415–460. [MR3611497](#)
- [47] YAMADA, Y., HYODO, M. and NISHIYAMA, T. (2017). Testing block-diagonal covariance structure for high-dimensional data under non-normality. *J. Multivariate Anal.* **155** 305–316. [MR3607897](#)
- [48] YANG, Y. and PAN, G. (2012). The convergence of the empirical distribution of canonical correlation coefficients. *Electron. J. Probab.* **17** no. 64, 13. [MR2968671](#)

- [49] YANG, Y. and PAN, G. (2015). Independence test for high dimensional data based on regularized canonical correlation coefficients. *Ann. Statist.* **43** 467–500. [MR3316187](#)
- [50] ZHENG, S., JIANG, D., BAI, Z. and HE, X. (2014). Inference on multiple correlation coefficients with moderately high dimensional data. *Biometrika* **101** 748–754. [MR3254916](#)

UNIFORM PROJECTION DESIGNS

BY FASHENG SUN¹, YAPING WANG² AND HONGQUAN XU³

*Northeast Normal University, East China Normal University and
 University of California, Los Angeles*

Efficient designs are in high demand in practice for both computer and physical experiments. Existing designs (such as maximin distance designs and uniform designs) may have bad low-dimensional projections, which is undesirable when only a few factors are active. We propose a new design criterion, called uniform projection criterion, by focusing on projection uniformity. Uniform projection designs generated under the new criterion scatter points uniformly in all dimensions and have good space-filling properties in terms of distance, uniformity and orthogonality. We show that the new criterion is a function of the pairwise L_1 -distances between the rows, so that the new criterion can be computed at no more cost than a design criterion that ignores projection properties. We develop some theoretical results and show that maximin L_1 -equidistant designs are uniform projection designs. In addition, a class of asymptotically optimal uniform projection designs based on good lattice point sets are constructed. We further illustrate an application of uniform projection designs via a multidrug combination experiment.

REFERENCES

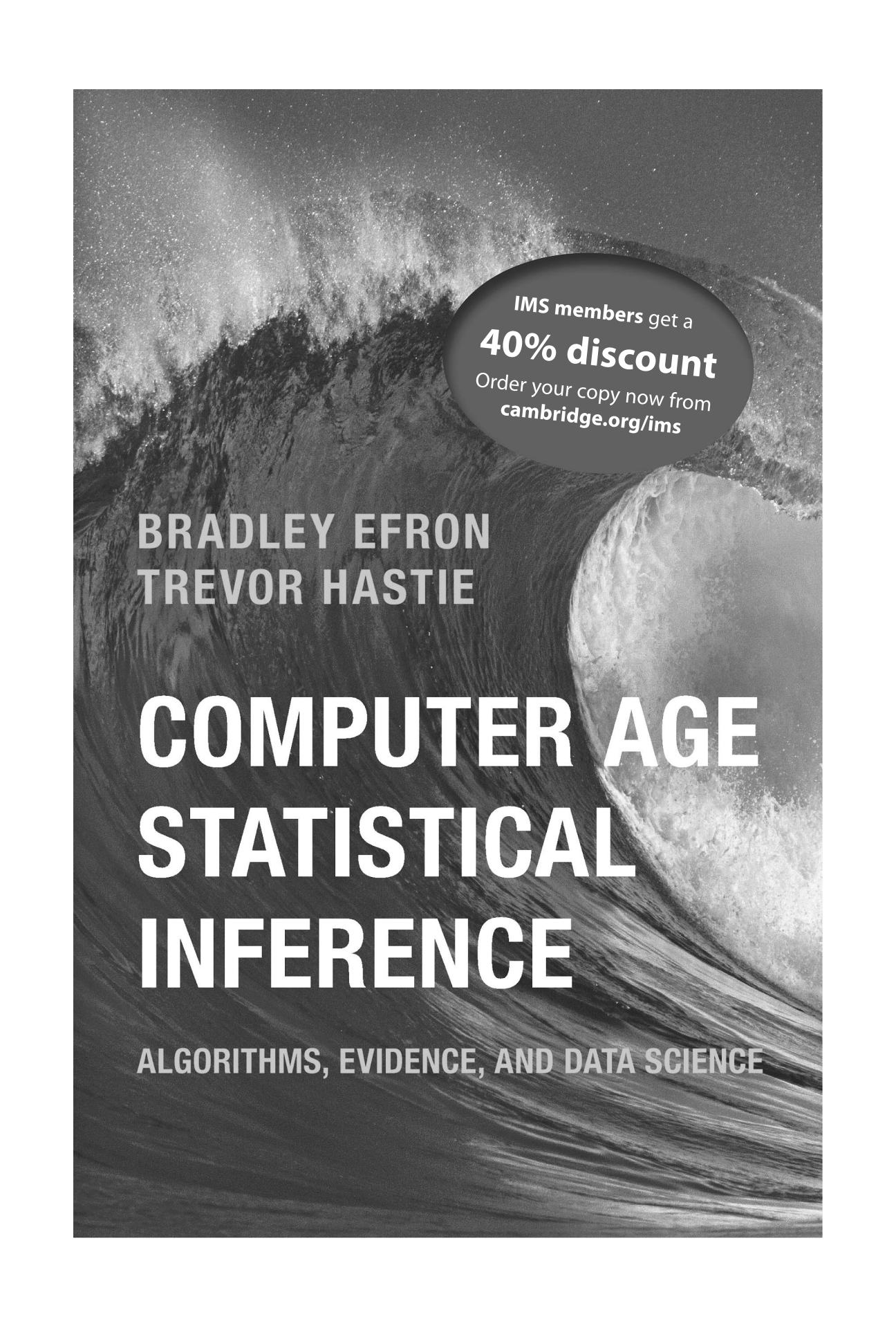
- AL-SHYOUKH, I., YU, F., FENG, J., YAN, K., DUBINETT, S., HO, C. M., SHAMMA, J. S. and SUN, R. (2011). Systematic quantitative characterization of cellular responses induced by multiple signals. *BMC Syst. Biol.* **5** 88.
- BA, S., MYERS, W. R. and BRENNEMAN, W. A. (2015). Optimal sliced Latin hypercube designs. *Technometrics* **57** 479–487. [MR3425485](#)
- FANG, K.-T., LI, R. and SUDJANTO, A. (2006). *Design and Modeling for Computer Experiments*. Chapman & Hall/CRC, Boca Raton, FL. [MR2223960](#)
- FANG, K.-T. and MUKERJEE, R. (2000). A connection between uniformity and aberration in regular fractions of two-level factorials. *Biometrika* **87** 193–198. [MR1766839](#)
- FANG, K.-T., LIN, D. K. J., WINKER, P. and ZHANG, Y. (2000). Uniform design: Theory and application. *Technometrics* **42** 237–248. [MR1801031](#)
- HE, Y., CHENG, C.-S. and TANG, B. (2018). Strong orthogonal arrays of strength two plus. *Ann. Statist.* **46** 457–468. [MR3782373](#)
- HE, Y. and TANG, B. (2013). Strong orthogonal arrays and associated Latin hypercubes for computer experiments. *Biometrika* **100** 254–260. [MR3034340](#)
- HE, Y. and TANG, B. (2014). A characterization of strong orthogonal arrays of strength three. *Ann. Statist.* **42** 1347–1360. [MR3226159](#)
- HICKERNELL, F. J. and LIU, M.-Q. (2002). Uniform designs limit aliasing. *Biometrika* **89** 893–904. [MR1946518](#)

MSC2010 subject classifications. 62K99, 62K15.

Key words and phrases. Computer experiment, discrepancy, Latin hypercube design, maximin distance design, space-filling design, uniform design.

- JOHNSON, M. E., MOORE, L. M. and YLVISAKER, D. (1990). Minimax and maximin distance designs. *J. Statist. Plann. Inference* **26** 131–148. [MR1079258](#)
- JOSEPH, V. R., GUL, E. and BA, S. (2015). Maximum projection designs for computer experiments. *Biometrika* **102** 371–380. [MR3371010](#)
- JOSEPH, V. R. and HUNG, Y. (2008). Orthogonal-maximin Latin hypercube designs. *Statist. Sinica* **18** 171–186. [MR2416907](#)
- KLEIJNEN, J. P. (2017). Design and analysis of simulation experiments: Tutorial. In *Advances in Modeling and Simulation* (A. Tolk, J. Fowler, G. Shao and E. Yucesan, eds.) 135–158. Springer, New York.
- LIANG, Y. Z., FANG, K. T. and XU, Q. S. (2001). Uniform design and its applications in chemistry and chemical engineering. *Chemom. Intell. Lab. Syst.* **58** 43–57.
- LIN, C. D. and TANG, B. (2015). Latin hypercubes and space-filling designs. In *Handbook of Design and Analysis of Experiments* (A. Dean, M. Morris, J. Stufken and D. Bingham, eds.) 593–625. CRC Press, Boca Raton, FL. [MR3699363](#)
- MA, C.-X., FANG, K.-T. and LIN, D. K. J. (2003). A note on uniformity and orthogonality. *J. Statist. Plann. Inference* **113** 323–334. [MR1963050](#)
- MOON, H., DEAN, A. and SANTNER, T. (2011). Algorithms for generating maximin Latin hypercube and orthogonal designs. *J. Stat. Theory Pract.* **5** 81–98. [MR2829824](#)
- MOON, H., DEAN, A. M. and SANTNER, T. J. (2012). Two-stage sensitivity-based group screening in computer experiments. *Technometrics* **54** 376–387. [MR3006387](#)
- NING, S., XU, H., AL-SHYOUKH, I., FENG, J. and SUN, R. (2014). An application of a Hill-based response surface model for a drug combination experiment on lung cancer. *Stat. Med.* **33** 4227–4236. [MR3267406](#)
- OWEN, A. B. (1994). Controlling correlations in Latin hypercube samples. *J. Amer. Statist. Assoc.* **89** 1517–1522.
- SANTNER, T. J., WILLIAMS, B. J. and NOTZ, W. I. (2003). *The Design and Analysis of Computer Experiments*. Springer, New York. [MR2160708](#)
- SUN, F. and TANG, B. (2017). A general rotation method for orthogonal Latin hypercubes. *Biometrika* **104** 465–472. [MR3698266](#)
- SUNDARAM, R. K. (1996). *A First Course in Optimization Theory*. Cambridge Univ. Press, Cambridge. [MR1402910](#)
- TANG, B. (1993). Orthogonal array-based Latin hypercubes. *J. Amer. Statist. Assoc.* **88** 1392–1397. [MR1245375](#)
- TANG, B. (1998). Selecting Latin hypercubes using correlation criteria. *Statist. Sinica* **8** 965–977. [MR1651518](#)
- TANG, Y., XU, H. and LIN, D. K. J. (2012). Uniform fractional factorial designs. *Ann. Statist.* **40** 891–907. [MR2985937](#)
- UNIFORM DESIGN WEBSITE. <http://www.math.hkbu.edu.hk/UniformDesign/>.
- WANG, L., XIAO, Q. and XU, H. (2018). Optimal maximin L_1 -distance Latin hypercube designs based on good lattice point designs. *Ann. Statist.* **46** 3741–3766. [MR3852667](#)
- WANG, Y., YANG, J. and XU, H. (2018). On the connection between maximin distance designs and orthogonal designs. *Biometrika* **105** 471–477. DOI:10.1093/biomet/asy005.
- WILLIAMS, E. J. (1949). Experimental designs balanced for the estimation of residual effects of treatments. *Aust. J. Sci. Res., Ser. A* **2** 149–168. [MR0033508](#)
- WOODS, D. C. and LEWIS, S. M. (2016). Design of experiments for screening. In *Handbook of Uncertainty Quantification* (R. Ghanem, D. Higdon and H. Owhadi, eds.) 1143–1185. Springer, New York.
- XIAO, Q., WANG, L. and XU, H. (2017). Application of Kriging models for a drug combination experiment on lung cancer. Manuscript.
- XIAO, Q. and XU, H. (2017). Construction of maximin distance Latin squares and related Latin hypercube designs. *Biometrika* **104** 455–464. [MR3698265](#)

- XIAO, Q. and XU, H. (2018). Construction of maximin distance designs via level permutation and expansion. *Statist. Sinica.* **28** 1395–1414. [MR3821010](#)
- YE, K. Q. (1998). Orthogonal column Latin hypercubes and their application in computer experiments. *J. Amer. Statist. Assoc.* **93** 1430–1439. [MR1666638](#)
- ZHOU, Y.-D. and XU, H. (2014). Space-filling fractional factorial designs. *J. Amer. Statist. Assoc.* **109** 1134–1144. [MR3265686](#)
- ZHOU, Y. and XU, H. (2015). Space-filling properties of good lattice point sets. *Biometrika* **102** 959–966. [MR3431565](#)



IMS members get a
40% discount

Order your copy now from
cambridge.org/ims

BRADLEY EFRON
TREVOR HASTIE

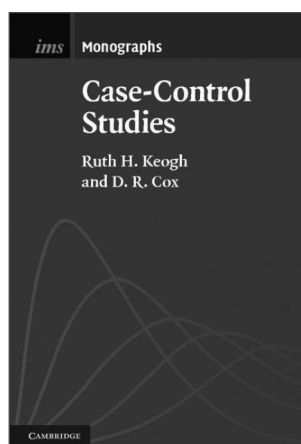
COMPUTER AGE STATISTICAL INFERENCE

ALGORITHMS, EVIDENCE, AND DATA SCIENCE



The Institute of Mathematical Statistics presents

IMS MONOGRAPHS



Case-Control Studies

Ruth H. Keogh
and D. R. Cox

The case-control approach is a powerful method for investigating factors that may explain a particular event. It is extensively used in epidemiology to study disease incidence, one of the best-known examples being Bradford Hill and Doll's investigation of the possible connection between cigarette smoking and lung cancer. More recently, case-control studies have been increasingly used in other fields, including sociology and econometrics.

With a particular focus on statistical analysis, this book is ideal for applied and theoretical statisticians wanting an up-to-date introduction to the field. It covers the fundamentals of case-control study design and analysis as well as more recent developments, including two-stage studies, case-only studies and methods for case-control sampling in time. The latter have important applications in large prospective cohorts which require case-control sampling designs to make efficient use of resources. More theoretical background is provided in an appendix for those new to the field.

**IMS member? Claim
your 40% discount:
www.cambridge.org/ims**

**Hardback price
US\$48.00
(non-member price
\$80.00)**

Cambridge University Press, in conjunction with the Institute of Mathematical Statistics, established the IMS Monographs and IMS Textbooks series of high-quality books. The Series Editors are Xiao-Li Meng, Susan Holmes, Ben Hambly, D. R. Cox and Alan Agresti.