Contents

BJPS is an IMS supported journal

ISSN 0103-0752

Brazilian Statistical Association

Brazilian Journal of Probability and Statistics

Volume 39 • Number 1 • March 2025

Brazilian Journal of Probability and Statistics

Volume 39 • Number 1 • March 2025

www.imstat.org/bjps

ISSN 0103-0752 (Print) ISSN 2317-6199 (Online), Volume 39, Number 1, March 2025. Published quarterly by the Brazilian Statistical Association.

POSTMASTER:

Send address changes to Brazilian Journal of Probability and Statistics, Institute of Mathematical Statistics, Dues and Subscriptions Office, PO Box 729, Middletown, Maryland 21769, USA.

Brazilian Statistical Association members should send address changes to Rua do Matão, 1010 sala 250A, 05508-090 São Paulo/SP Brazil (address of the BSA office).

Printed in the United States of America

Partial financial support:
CNPq and CAPES (Brazil).

Brazilian Journal of Probability and Statistics
2025, Vol. 39, No. 1, 1–18
https://doi.org/10.1214/25-BJPS621

Finite-sample bounds to the normal limit under group sequential sampling

Julian Aronowitz[1,a] **and Jay Bartroff**[2,b]

[1]*Google Inc., New York City, U.S.A.,* [a]*julian.aronowitz@gmail.com*
[2]*Department of Statistics & Data Sciences, University of Texas at Austin, Austin, U.S.A.,* [b]*bartroff@austin.utexas.edu*

Abstract. In group sequential analysis, data is collected and analyzed in batches until pre-defined stopping criteria are met. Inference in the parametric setup typically relies on the limiting asymptotic multivariate normality of the repeatedly computed maximum likelihood estimators (MLEs), a result first rigorously proved by Jennison and Turnbull (1997) under general regularity conditions. In this work, using Stein's method we provide optimal order, non-asymptotic bounds on the distance for smooth test functions between the joint group sequential MLEs and the appropriate normal distribution under the same conditions. Our results assume independent observations but allow heterogeneous (i.e., non-identically distributed) data. We examine how the resulting bounds simplify when the data comes from an exponential family. Finally, we present a general result relating multivariate Kolmogorov distance to smooth function distance which, in addition to extending our results to the former metric, may be of independent interest.

References

Anastasiou, A. (2017). Bounds for the normal approximation of the maximum likelihood estimator from m-dependent random variables. *Statistics & Probability Letters* **129**, 171–181. MR3688530 https://doi.org/10.1016/j.spl.2017.04.022

Anastasiou, A. (2018). Assessing the multivariate normal approximation of the maximum likelihood estimator from high-dimensional, heterogeneous data. *Electronic Journal of Statistics* **12**, 3794–3828. MR3881763 https://doi.org/10.1214/18-EJS1492

Anastasiou, A. and Gaunt, R. E. (2020). Multivariate normal approximation of the maximum likelihood estimator via the delta method. *Brazilian Journal of Probability and Statistics* **34**, 136–149. MR4058975 https://doi.org/10.1214/18-BJPS411

Anastasiou, A. and Ley, C. (2017). Bounds for the asymptotic normality of the maximum likelihood estimator using the delta method. *Latin American Journal of Probability and Statistics* **14**, 153–171. MR3622464 https://doi.org/10.30757/alea.v14-09

Anastasiou, A. and Reinert, G. (2017). Bounds for the normal approximation of the maximum likelihood estimator. *Bernoulli* **23**, 191–218. MR3556771 https://doi.org/10.3150/15-BEJ741

Aronowitz, J. and Bartroff, J. (2025). Supplement to "Finite-sample bounds to the normal limit under group sequential sampling." https://doi.org/10.1214/25-BJPS621SUPP

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory* **39**, 930–945. MR1237720 https://doi.org/10.1109/18.256500

Bartroff, J., Lai, T. L. and Shih, M. (2013). *Sequential Experimentation in Clinical Trials: Design and Analysis*. New York: Springer. MR2987767 https://doi.org/10.1007/978-1-4614-6114-2

Bonis, T. (2020). Stein's method for normal approximation in Wasserstein distances with application to the multivariate central limit theorem. *Probability Theory and Related Fields* **178**, 827–860. MR4168389 https://doi.org/10.1007/s00440-020-00989-4

Chen, L. H., Goldstein, L. and Shao, Q.-M. (2010). *Normal Approximation by Stein's Method*. Berlin Heidelberg: Springer. MR2732624 https://doi.org/10.1007/978-3-642-15007-4

Coumans, F. A., Ligthart, S. T., Uhr, J. W. and Terstappen, L. W. (2012). Challenges in the enumeration and phenotyping of CTC. *Clinical Cancer Research* **18**, 5711–5718.

Key words and phrases. Central limit theorem, group sequential, maximum likelihood estimation, Stein's method.

Cucker, F. and Zhou, D. X. (2007). *Learning Theory: An Approximation Theory Viewpoint, Vol. 24*. Cambridge: Cambridge University Press. MR2354721 https://doi.org/10.1017/CBO9780511618796

de A. Cysneiros, F. J., dos Santos, S. J. P. and Cordeiro, G. M. (2001). Skewness and kurtosis for maximum likelihood estimator in one-parameter exponential family models. *Brazilian Journal of Probability and Statistics* **15**, 85–105. MR1907862

Fang, X., Shao, Q.-M. and Xu, L. (2019). Multivariate approximations in Wasserstein distance by Stein's method and Bismut's formula. *Probability Theory and Related Fields* **174**, 945–979. MR3980309 https://doi.org/10.1007/s00440-018-0874-5

Gaunt, R. E. (2016). Rates of convergence in normal approximation under moment conditions via new bounds on solutions of the Stein equation. *Journal of Theoretical Probability* **26**, 231–247. MR3463084 https://doi.org/10.1007/s10959-014-0562-z

Gaunt, R. E. and Li, S. (2023). Bounding Kolmogorov distances through Wasserstein and related integral probability metrics. *Journal of Mathematical Analysis and Applications* **522**, 126985. MR4533915 https://doi.org/10.1016/j.jmaa.2022.126985

Goldstein, L. and Rinott, Y. (1996). Multivariate normal approximations by Stein's method and size bias couplings. *Journal of Applied Probability* **33**, 1–17. MR1371949 https://doi.org/10.1017/s0021900200103675

Hoadley, B. (1971). Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case. *The Annals of Mathematical Statistics* **42**, 1977–1991. MR0297051 https://doi.org/10.1214/aoms/1177693066

Jennison, C. and Turnbull, B. W. (1997a). Group sequential analysis incorporating covariate information. *Journal of the American Statistical Association* **92**, 1330–1341. MR1615245 https://doi.org/10.2307/2965403

Jennison, C. and Turnbull, B. W. (1997b). Distribution theory of group sequential t, χ^2 and F-tests for general linear models. *Sequential Analysis* **16**, 295–317. MR1491639 https://doi.org/10.1080/07474949708836390

Jennison, C. and Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton, Florida: Chapman & Hall. MR1710781

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. London: Chapman and Hill. MR3223057 https://doi.org/10.1007/978-1-4899-3242-6

O'Brien, P. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35**, 549–556.

Peers, H. W. and Iqbal, M. (1985). Asymptotic expansions for confidence limits in the presence of nuisance parameters, with applications. *Journal of the Royal Statistical Society, Series B, Methodological* **47**, 547–554. MR0844486

Pinelis, I. (2017). Optimal-order uniform and nonuniform bounds on the rate of convergence to normality for maximum likelihood estimators. *Electronic Journal of Statistics* **11**, 1160–1179. MR3634332 https://doi.org/10.1214/17-EJS1264

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191–199.

Reinert, G. and Röllin, A. (2009). Multivariate normal approximation with Stein's method of exchangable pairs under a general linearity condition. *Annals of Probability* **36**, 2150–2173. MR2573554 https://doi.org/10.1214/09-AOP467

Smola, A. J. and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing* **14**, 199–222. MR2086398 https://doi.org/10.1023/B:STCO.0000035301.49549.88

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, 583–602. Berkeley, Calif: University of California Press. MR0402873

Ulyanov, V. (1979). On more precise convergence rate estimates in the central limit theorem. *Theory of Probability and Its Applications* **23**, 660–663.

Ulyanov, V. (1986). Normal approximation for sums of non-identically distributed random variables in Hilbert spaces. *Acta Scientiarum Mathematicarum* **50**, 411–419. MR0882054

Ulyanov, V. (1987). Asymptotic expansions for distributions of sums of independent random variables in H. *Theory of Probability and Its Applications* **31**, 25–39. MR0836950

van der Vaart, A. W. (2007). *Asymptotic Statistics*. New York, NY: Cambrige University Press. MR1652247 https://doi.org/10.1017/CBO9780511802256

Verbeek, A. (1992). The compactification of generalized linear models. *Statistica Neerlandica* **46**, 107–142. MR1178475 https://doi.org/10.1111/j.1467-9574.1992.tb01332.x

Williams, A. L., Fitzgerald, J. E., Ivich, F., Sontag, E. D. and Niedre, M. (2020). Short-term circulating tumor cell dynamics in mouse xenograft models and implications for liquid biopsy. *Frontiers in Oncology* **10**, 601085.

Zhu, X., Suo, Y., Ding, N., He, H. and Wei, X. (2017). Circulating tumor cells occur nonuniformly monitored by in vivo flow cytometry. In *International Conference on Photonics and Imaging in Biology and Medicine*, W3A–104.

Zhu, X., Suo, Y., Fu, Y., Zhang, F., Ding, N., Pang, K., Xie, C., Weng, X., Tian, M., He, H., et al (2021). In vivo flow cytometry reveals a circadian rhythm of circulating tumor cells. *Light: Science & Applications* **10**, 110.

Brazilian Journal of Probability and Statistics
2025, Vol. 39, No. 1, 19–38
https://doi.org/10.1214/25-BJPS622

Multidimensional graded response models with hierarchical structure and Q-matrix

Marcelo da Silva[1,a], Jorge Bazán[2,b], Ren Liu[3,c], Edna Possan[4,d] and Silvana Vincenzi[5,e]

[1] *"Luiz de Queiroz" College of Agriculture, University of São Paulo, Piracicaba, Brazil,* [a]*silva.marcelo@usp.br*
[2] *Institute of Mathematics and Computer Sciences, University of São Paulo, São Carlos, Brazil,* [b]*jlbazan@icmc.usp.br*
[3] *University of California, Merced, California, USA,* [c]*rliu45@ucmerced.edu*
[4] *Federal University of Latin American Integration, Foz do Iguaçu, Brazil,* [d]*edna.possan@unila.edu.br*
[5] *Federal University of Santa Catarina, Florianópolis, Brazil,* [e]*silvanaligia@gmail.com*

Abstract. Multidimensional item response theory (MIRT) models estimate multiple latent traits from individual item response data and have been applied to problems in several knowledge areas. This article explores some statistical aspects about the compensatory multidimensional graded response model with two hierarchical structures and Q-matrix. Specifically, the efficiency of the estimation method using No-U-Turn Sampler (NUTS) algorithm and the performance of the three model comparison criteria regarding the adequacy of the models with the same dataset were verified. An application was conducted with recent data from responses to a questionnaire on sustainability perception applied to residents in Brazil verified the model that best fit the data, thus justifying the choice of the appropriate model for the research. The results were used for the selection of a model for the application. The explored models provided detailed information about individuals, which may be useful for the design of public policies aimed at the development of specific educational actions for the economic, environmental and social fields. Codes are available for the proposed methodology to be explored in other applications.

References

Andrade, D. F., Tavares, H. R. and Valle, R. C. (2000). *Teoria da Resposta ao Item: conceitos e aplicações*. Sao Paulo: ABE.

Béguin, A. and Glas, C. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika* **66**, 541–561. MR1961913 https://doi.org/10.1007/BF02296195

Bock, R. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika* **46**, 443–459. MR0668311 https://doi.org/10.1007/BF02293801

Burnham, K. and Anderson, D. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Berlin: Springer. MR1919620

Cai, L. (2010). Metropolis–Hastings Robbins–Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics* **35**, 307–335.

Chen, Y., Liu, J., Xu, G. and Ying, Z. (2015). Statistical analysis of q-matrix based diagnostic classification models. *Journal of the American Statistical Association* **110**, 850–866. MR3367269 https://doi.org/10.1080/01621459.2014.934827

Curtis, S. (2010). BUGS code for item response theory. *Journal of Statistical Software, Code Snippets* **36**, 1–34.

da Silva, M., Bazán, J. and Huggins-Manley, A. (2019). Sensitivity analysis and choosing between alternative polytomous IRT models using Bayesian model comparison criteria. *Communications in Statistics Simulation and Computation* **48**, 601–620. MR3933866 https://doi.org/10.1080/03610918.2017.1390126

da Silva, M., Bazán, J., Liu, R., Possan, E. and Vincenzi, S. (2025). Supplement to "Multidimensional graded response models with hierarchical structure and Q-matrix." https://doi.org/10.1214/25-BJPS622SUPP

Key words and phrases. Bayesian estimation, graded response model, hierarchical multidimensional item response theory, Q-matrix, sustainability perception.

da Silva, M., Liu, R., Huggins-Manley, A. and Bazán, J. (2019). Incorporating the Q-matrix into multidimensional item response theory models. *Educational and Psychological Measurement* **79**, 665–687.
da Silva, M., Liu, R., Huggins-Manley, A. and Bazán, J. (2023). Bayesian estimation of multidimensional polytomous item response theory models with Q-matrices using Stan. *Communications in Statistics Simulation and Computation* **52**, 5178–5194. MR4666701 https://doi.org/10.1080/03610918.2021.1977951
da Silva, M., Oliveira, E., Davier, A. and Bazán, J. (2018). Estimating the DINA model parameters using the no-U-turn sampler. *Biometrical Journal* **60**, 352–368. MR3773539 https://doi.org/10.1002/bimj.201600225
de la Torre, J. and Chiu, C.-Y. (2016). General method of empirical Q-matrix validation. *Psychometrika* **81**, 253–273. MR3505366 https://doi.org/10.1007/s11336-015-9467-8
de la Torre, J., Qiu, X.-L. and Santos, K. C. (2021). An empirical q-matrix validation method for the polytomous g-dina model. *Psychometrika*, 1–32. MR4434012 https://doi.org/10.1007/s11336-021-09821-x
de la Torre, J. and Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher order IRT model approach. *Applied Psychological Measurement* **33**, 620–639. MR2667525 https://doi.org/10.1177/0146621608320523
Dunlap, R. (2016). A brief history of sociological research on environmental concern. In *Green European: Environmental Behaviour and Attitudes in Europe in a Historical and Cross-Cultural Comparative Perspective*.
Fernández-Sánchez, G. and Rodríguez-López, F. (2010). A methodology to identify sustainability indicators in construction project management—application to infrastructure projects in Spain. *Ecological Indicators* **10**, 1193–1201.
Fox, J. (2010). *Bayesian Item Response Modeling*. New York: Springer. MR2657265 https://doi.org/10.1007/978-1-4419-0742-4
Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. and Rubin, D. (2014). *Bayesian Data Analysis*. London: Chapman and Hall. MR3235677
Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457–472.
Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., Kupfer, D. J., Frank, E., Grochocinski, V. J. and Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement* **31**, 4–19. MR2297621 https://doi.org/10.1177/0146621606289485
Girolami, M. and Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society, Series B, Statistical Methodology* **73**, 123–214. MR2814492 https://doi.org/10.1111/j.1467-9868.2010.00765.x
Grant, R., Furr, D., Carpenter, B. and Gelman, A. (2017). Fitting Bayesian item response models in Stata and Stan. *Stata Journal* **17**, 343–357.
Hoffman, M. and Gelman, A. (2014). The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* **15**, 1593–1623. MR3214779
Huang, H.-Y. and Wang, W.-C. (2014). Multilevel higher-order item response theory models. *Educational and Psychological Measurement* **74**, 495–515.
Huang, H.-Y., Wang, W.-C., Chen, P.-H. and Su, C.-M. (2013). Higher-order item response models for hierarchical latent traits. *Applied Psychological Measurement* **37**, 619–637.
Liu, R., Huggins-Manley, A. and Bradshaw, L. (2017). The impact of Q-matrix designs on diagnostic classification accuracy in the presence of attribute hierarchies. *Educational and Psychological Measurement* **77**, 220–240.
Lunn, D., Thomas, A., Best, N. and Spiegelhalter, D. (2000). WinBUGS—a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing* **10**, 325–337.
Luo, Y. and Jiao, H. (2018). Using the Stan program for Bayesian item response theory. *Educational and Psychological Measurement* **78**, 384–408.
Madison, M. and Bradshaw, L. (2015). The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educational and Psychological Measurement* **75**, 491–511.
Mahpour, A. (2018). Prioritizing barriers to adopt circular economy in construction and demolition waste management. *Resources, Conservation and Recycling* **134**, 216–227.
Martelli, I., Matteucci, M. and Mignani, S. (2016). Bayesian estimation of a multidimensional additive graded response model for correlated traits. *Communications in Statistics Simulation and Computation* **45**, 1636–1654. MR3490695 https://doi.org/10.1080/03610918.2014.932804
Mayer, A. (2008). Strengths and weaknesses of common sustainability indices for multidimensional systems. *Environment International* **34**, 277–291.
Menegon, L., Vincenzi, S., Andrade, D., Barbetta, P., Vink, P. and Merino, E. (2019). An aircraft seat discomfort scale using item response theory. *Applied Ergonomics* **77**, 1–8.
Molenaar, W. (1995). Estimation of item parameters. In *Rasch Models: Foundationa, Recent Developments and Applications*, 39–51. Workshop on Rasch Models—Foundations, Recent Developments, and Applications; Conference date: 25-02-1993 Through 27-02-1993.

Neal, R. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*, 116–162. MR2858447

Pacheco, J., Andrade, D. and Bornia, A. (2015). Benchmarking by Item Response Theory (BIRTH): A benchmarking method using IRT to build competitiveness scales for Brazilian technology higher education. *Benchmarking* **22**, 945–962.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. DSC 2003 Working Papers. 1–8.

Rao, C. and Sinharay, S. (2007). *Psychometrics*. Amsterdam: Elsevier.

Reckase, M. (2009). *Multidimensional Item Response Theory*. Berlin: Springer.

Rijmen, F., Jeon, M., von Davier, M. and Rabe-Hesketh, S. (2014). A third-order item response theory model for modeling the effects of domains and subdomains in large-scale educational assessment surveys. *Journal of Educational and Behavioral Statistics* **39**, 235–256. MR3192978

Sahu, S. K. (2002). Bayesian estimation and model choice in item response models. *Journal of Statistical Computation and Simulation* **72**, 217–232. MR1909259 https://doi.org/10.1080/00949650212387

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monographs* **17**.

Santos, T., Julian, C., Andrade, D., Villar, B., Piccinelli, R., González-Gross, M., Gottrand, F., Androutsos, O., Kersting, M., Michels, N., Huybrechts, I., Widhalm, K., Molnár, D., Marcos, A., Castillo-Garzón, M. and Moreno, L. (2019). Measuring nutritional knowledge using item response theory and its validity in European adolescents. *Public Health Nutrition* **22**, 419–430.

Sheng, Y. (2010). Bayesian estimation of MIRT models with general and specific latent traits in MATLAB. *Journal of Statistical Software* **34**, 1–27.

Sheng, Y. and Wikle, C. (2008). Bayesian multidimensional IRT models with a hierarchical structure. *Educational and Psychological Measurement* **68**, 413–430. MR2432233 https://doi.org/10.1177/0013164407308512

Spiegelhalter, D., Best, N., Carlin, B. and Linden, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B* **64**, 583–639. MR1979380 https://doi.org/10.1111/1467-9868.00353

Stan Development Team (2018). Stan Modeling Language Users Guide and Reference Manual, 2.17.3. https://mc-stan.org.

Szopik-Depczynska, K., Cheba, K., Bąk, I., Stajniak, M., Simboli, A. and Ioppolo, G. (2018). The study of relationship in a hierarchical structure of EU sustainable development indicators. *Ecological Indicators* **90**, 120–131.

Tatsuoka, K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement* **20**, 345–354.

Tavares, H., Andrade, D. and Pereira, C. (2004). Detection of determinant genes and diagnostic via item response theory. *Genetics and Molecular Biology* **27**, 679–685.

Tezza, R., Bornia, A. and Andrade, D. (2011). Measuring web usability using item response theory: Principles, features and opportunities. *Interacting with Computers* **23**, 167–175.

Vehtari, A., Gelman, A. and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* **27**, 1413–1432. MR3647105 https://doi.org/10.1007/s11222-016-9696-4

Venturelli, J., Gonçalves, F. and Andrade, D. (2020). Multidimensional Bayesian IRT model for hierarchical latent structures. arXiv preprint. Available at arXiv:2006.09966.

Vincenzi, S., Possan, E., Andrade, D., Pituco, M., Oliveira Santos, T. and Jasse, E. (2018). Assessment of environmental sustainability perception through item response theory: A case study in Brazil. *Journal of Cleaner Production* **170**, 1369–1386.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* **11**, 3571–3594. MR2756194

Brazilian Journal of Probability and Statistics
2025, Vol. 39, No. 1, 39–58
https://doi.org/10.1214/25-BJPS623

The first-order seasonal integer-valued autoregression process with zero-inflated Poisson innovations; application to integer-valued seasonal data analysis with overdispersion

Kang-Hyok Yu[a] and **Song-Guk Kim**

Faculty of Mathematics, Kim Il Sung University, Pyongyang, Democratic People's Republic of Korea,
[a]*kh.yu@ryongnamsan.edu.kp*

Abstract. The first order Poisson integer-valued autoregressive (INAR$(1)_s$) process with a seasonal structure has been used to model the seasonal integer-valued time series. This model, however, may not be suitable for count data with both seasonal structure and overdispersion. This paper introduces the first order seasonal integer-valued autoregressive process with zero-inflated Poisson innovations that can be used to model the seasonal integer-valued time-series with overdispersion. We discuss several probabilistic and statistical properties including stationarity of the process as well as a conditional least squares estimator and a maximum likelihood estimator of the process. Also, throughout our simulations, the applicability of these estimators to finite samples is evaluated and compared. Finally, we illustrate the usefulness of the proposed model by comparing our analysis results with ones by other competitive models considered in previous works on three real data sets.

References

Al-Osh, M. A. and Alzaid, A. A. (1987). First order integer- valued autoregressive (INAR(1)) process. *Journal of Time Series Analysis* **8**, 261–275. MR0903755 https://doi.org/10.1111/j.1467-9892.1987.tb00438.x

Alzaid, A. A. and Al-Osh, M. A. (1990). An integer-valued pth order autoregressive structure (INAR(p)) process. *Journal of Applied Probability* **27**, 314–324. MR1052303 https://doi.org/10.2307/3214650

Bentarzi, M. and Aries, N. (2020a). On some periodic INARMA(p, q) models. *Communications in Statistics Simulation and Computation* **51**, 5773–5793. MR4496155 https://doi.org/10.1080/03610918.2020.1780443

Bentarzi, M. and Aries, N. (2020b). QMLE of periodic integer-valued time series models. *Communications in Statistics Simulation and Computation* **51**, 4973–4999. MR4491664 https://doi.org/10.1080/03610918.2020.1752380

Bourguignon, M., Vasconcellos, K. L. P., Reisen, V. A. and Ispány, M. (2016). A Poisson INAR(1) process with a seasonal structure. *Journal of Statistical Computation and Simulation* **86**, 373–387. MR3406739 https://doi.org/10.1080/00949655.2015.1015127

Consul, P. C. and Famoye, F. (2006). *Lagrangian Probability Distributions*. Boston: Birkhauser. MR2209108

Cui, Y. and Lund, R. B. (2009). A new look at time series of counts. *Biometrika* **96**, 781–792. MR2564490 https://doi.org/10.1093/biomet/asp057

Davis, R. A., Fokianos, K., Holan, S. H., Joe, H., Livsey, J., Lund, R., Pipiras, V. and Ravishanker, N. (2021). Count time series: A methodological review. *Journal of the American Statistical Association* **116**, 1533–1547. MR4309291 https://doi.org/10.1080/01621459.2021.1904957

Du, J. G. and Li, Y. (1991). The integer-valued autoregressive (INAR(p)) model. *Journal of Time Series Analysis* **12**, 129–142. MR1108796 https://doi.org/10.1111/j.1467-9892.1991.tb00073.x

Filho, P. R. P., Reisen, V. A., Bondon, P., Ispaány, M. and Melo, M. M. (2021). A periodic and seasonal statistical model for non-negative integer-valued time series with an application to dispensed medications in respiratory diseases. *Applied Mathematical Modelling* **96**, 545–558. MR4241220 https://doi.org/10.1016/j.apm.2021.03.025

Franke, J. and Subba Rao, T. (1995). Multivariate first-order integer-valued autoregressions. Technical Report, Maths. Dep., UMIST.

Key words and phrases. Seasonal structure, overdispersion, zero-inflated Poisson distribution, maximum likelihood estimator, binomial thinning operator.

Freeland, R. K. and McCabe, B. P. M. (2004). Analysis of low count time series data by Poisson autoregression. *Journal of Time Series Analysis* **25**, 701–722. MR2089191 https://doi.org/10.1111/j.1467-9892.2004.01885.x

Ispány, M., Bondon, P., Reisen, V. A. and Filho, P. R. P. (2024). Existence of a periodic and seasonal INAR process. *Journal of Time Series Analysis*. Available at http://onlinelibrary.wiley.com/doi/10.1111/jtsa.12746. MR4822257 https://doi.org/10.1111/jtsa.12746

Jazi, M. A., Jones, G. and Lai, C. D. (2012). First-order integer valued processes with zero inflated Poisson innovations. *Journal of Time Series Analysis* **33**, 954–963. MR2991911 https://doi.org/10.1111/j.1467-9892.2012.00809.x

Latour, A. (1997). The multivariate GINAR(p) process. *Advances in Applied Probability* **29**, 228–248. MR1432938 https://doi.org/10.2307/1427868

Mckenzie, E. (1985). Some simple models for discrete variate time series. *JAWRA Journal of the American Water Resources Bulletin* **21**, 645–650. MR1973555 https://doi.org/10.1016/S0169-7161(03)21018-X

Monteiro, M., Scotto, M. G. and Pereira, I. (2010). Integer-valued autoregressive processes with periodic structure. *Journal of Statistical Planning and Inference* **140**, 1529–1541. MR2592231 https://doi.org/10.1016/j.jspi.2009.12.015

Moriña, D., Puig, P., Ríos, J., Vilella, A. and Trilla, A. (2011). A statistical model for hospital admissions caused by seasonal diseases. *Statistics in Medicine* **20**, 3125–3136. MR2845682 https://doi.org/10.1002/sim.4336

Nastić, A. S., Laketa, P. N. and Ristić, M. M. (2016). Random environment integer-valued autoregressive process. *Journal of Time Series Analysis* **37**, 267–287. MR3511585 https://doi.org/10.1111/jtsa.12161

Qi, X., Li, Q. and Zhu, F. (2019). Modeling time series of count with excess zeros and ones based on INAR(1) model with zero-and-one inflated Poisson innovations. *Journal of Computational and Applied Mathematics* **346**, 572–590. MR3864182 https://doi.org/10.1016/j.cam.2018.07.043

Quddus, M. A. (2008). Time series count data models: An empirical application to traffic accidents. *Accident Analysis and Prevention* **40**, 1732–1741.

Quoreshi, A. M. M. S. (2006). Bivariate time series modelling of financial count data. *Communications in Statistics - Theory and Methods* **35**, 1343–1358. MR2328481 https://doi.org/10.1080/03610920600692649

Ridout, M., Demetrio, C. G. B. and Hinde, J. (1998). Models for count data with many zeros. In *International Biometric Conference*. Cape Town.

Sadoun, M. and Bentarzi, M. (2020). Efficient estimation in periodic INAR(1) model: Parametric case. *Communications in Statistics Simulation and Computation* **49**, 2014–2034. MR4143429 https://doi.org/10.1080/03610918.2018.1510524

Santos, C., Pereira, I. and Scotto, M. G. (2021). On the theory of periodic multivariate INAR processes. *Statistical Papers* **62**, 1291–1348. MR4262195 https://doi.org/10.1007/s00362-019-01136-5

Steutel, F. W. and Van Harn, K. (1979). Discrete analogues of self-decomposability and stability. *Annals of Probability* **7**, 893–899. MR0542141

Tian, S., Wang, D. and Cui, S. (2020). A seasonal geometric INAR process based on negative binomial thinning operator. *Statistical Papers* **61**, 2561–2581. MR4167342 https://doi.org/10.1007/s00362-018-1060-7

Weiß, C. H. (2008). The combined INAR(p) models for time series of counts. *Statistics & Probability Letters* **78**, 1817–1822. MR2528554 https://doi.org/10.1016/j.spl.2008.01.036

Yu, G. H. and Kim, S. G. (2020). Parameter change test for periodic integer-valued autoregressive process. *Communications in Statistics - Theory and Methods* **49**, 2898–2912. MR4095597 https://doi.org/10.1080/03610926.2019.1584309

Zheng, H., Basawa, I. V. and Datta, A. (2007). First-order random coefficient integer-valued autoregressive processes. *Journal of Statistical Planning and Inference* **173**, 212–229. MR2292852 https://doi.org/10.1016/j.jspi.2005.12.003

Brazilian Journal of Probability and Statistics
2025, Vol. 39, No. 1, 59–78
https://doi.org/10.1214/25-BJPS624

Doubly robust estimation with graphical structure among predictors for integrating probability and non-probability samples

Zhan Liu[1], **Qing Zhou**[2] **and Yingli Pan**[1,a]

[1]*Hubei Key Laboratory of Applied Mathematics, School of Mathematics and Statistics, Hubei University, Wuhan 430062, China,* [a]*ylpan@hubu.edu.cn*

[2]*Department of Mathematical Foundations and Applications, Wuhan Huaxia Institute of Technology, Wuhan 430223, China*

Abstract. Non-probability samples such as web survey data are widely used in different fields because of their low cost, short period and high efficiency. However, it is difficult to infer the population due to the absence of inclusion probabilities of non-probability samples. The propensity score approach, superpopulation model approach and doubly robust approach have been used for inference with non-probability samples. However, when there is a relatively rich set of covariates or predictors, variable selection and correlations between predictors become important in modeling. In this paper, doubly robust estimation methods are proposed via incorporating a penalty and double penalty graphical structures among predictors for non-probability samples with high-dimensional data. Specifically, not only are graphical structures among predictors considered to describe the relationships between covariates in superpopulation modeling, but the inverse of the estimated inclusion probabilities of non-probability samples are also added to the objective function for estimating superpopulation parameters in the proposed methods. The asymptotic properties of the proposed estimators and variance estimation are discussed. Simulation studies are further conducted to verify the performance of the proposed methods. Lastly, the proposed methods are adopted to analyze a Rattus norvegicus eye eQTL experiment data set.

References

Baker, R., Brick, J. M., Bates, N. A., et al (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology* **1**, 90–143.

Chen, J. K. T., Valliant, R. and Elliott, M. R. (2018). Model-assisted calibration of non-probability sample survey data using adaptive LASSO. *Survey Methodology* **44**, 117–145.

Chen, J. K. T., Valliant, R. and Elliott, M. R. (2019). Calibrating non-probability surveys to estimated control totals using LASSO, with an application to political polling. *Journal of the Royal Statistical Society Series C Applied Statistics* **68**, 657–681. MR3937468 https://doi.org/10.1111/rssc.12327

Chen, Y., Li, P. and Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association* **115**, 2011–2021. MR4189773 https://doi.org/10.1080/01621459.2019.1677241

Dobriban, E. and Sheng, Y. (2020). WONDER: Weighted one-shot distributed ridge regression in high dimensions. *Journal of Machine Learning Research* **21**, 2483–2534. MR4095345

Elliott, M. R. and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science* **32**, 249–264. MR3648958 https://doi.org/10.1214/16-STS598

Friedman, J., Hastie, T. and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441.

Genz, A. and Bretz, F. (2009). *Computation of Multivariate Normal and t Probabilities*. *Lecture Notes in Statistics*. Heidelberg: Springer. MR2840595 https://doi.org/10.1007/978-3-642-01689-9

Genz, A., Bretz, F., Miwa, T., et al. (2017). mvtnorm: Multivariate normal and t distribution. R package version 4.2.1.

Key words and phrases. Doubly robust estimation, graphical structure, non-probability sample, high-dimensional data, superpopulation model, propensity score.

Kim, J. K., Park, S., Chen, Y., et al (2021). Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society Series A Statistics in Society* **184**, 941–963. MR4305566 https://doi.org/10.1111/rssa.12696

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA, USA: MIT Press. MR2816736

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics* **22**, 329–349.

Lee, S. and Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research* **37**, 319–343. MR2649463 https://doi.org/10.1177/0049124108329643

Liu, H., Roeder, K. and Wasserman, L. (2010). Stability approach to regularization selection (StARS) for high dimensional graphical models. *Advances in Neural Information Processing Systems* **23**, 1432–1440.

Liu, Z. and Pan, Y. (2021). Research on inference of candidate database web surveys based on superpopulation pseudo design and the combined sample. *Chinese Journal of Applied Probability and Statistics* **35**, 221–232. MR4112111

Liu, Z., Tu, C. and Pan, Y. (2021). Model-assisted SCAD calibration for non-probability samples. *Brazilian Journal of Probability and Statistics* **35**, 772–787. MR4350959 https://doi.org/10.1214/21-bjps506

Liu, Z., Zheng, J. and Pan, Y. (2023). Doubly robust estimation for non-probability samples with modified intertwined probabilistic factors decoupling. *Statistical Analysis and Data Mining* **16**, 224–236. MR4600557 https://doi.org/10.1002/sam.11614

Liu, Z., Zheng, J., Pan, Y., et al (2023). Estimation for volunteer web survey samples using a model-averaging approach. *Journal of Applied Statistics* **50**, 3251–3271. MR4665311 https://doi.org/10.1080/02664763.2022.2107187

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* **34**, 1436–1462. MR2278363 https://doi.org/10.1214/009053606000000281

Miller, C. J., Miller, M. C., biocViews Microarray, O. C., et al. (2013). Package 'simpleaffy'.

Pan, Y., Cai, W. and Liu, Z. (2022). Inference for non-probability samples under high-dimensional covariate-adjusted superpopulation model. *Statistical Methods and Applications* **31**, 955–979. MR4480330 https://doi.org/10.1007/s10260-021-00619-w

Rafei, A. (2021). Robust and efficient Bayesian inference for large-scale non-probability samples. The University of Michigan. MR4352285

Rafei, A., Flannagan, C. A. C. and West, B. T. (2022). Robust Bayesian inference for big data: Combining sensor-based records with traditional survey data. *Annals of Applied Statistics* **16**, 1038–1070. MR4438823 https://doi.org/10.1214/21-aoas1531

Scheetz, T. E., Kim, K. Y. A., Swiderski, R. E., et al (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 14429–14434.

Stephenson, M. (2018). Doubly sparse regularized regression incorporating graphical structure among predictors. The University of Guelph, Guelph, Ontario, Canada. MR4035798 https://doi.org/10.1002/cjs.11521

Stephenson, M., Ali, R. A. and Darlington, G. A. (2019). Doubly sparse regression incorporating graphical structure among predictors. *Canadian Journal of Statistics* **47**, 729–747. MR4035798 https://doi.org/10.1002/cjs.11521

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B, Methodological* **58**, 267–288. MR1379242

Valliant, R. (2019). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology* **8**, 231–263.

Valliant, R. and Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research* **40**, 105–137. MR2758301 https://doi.org/10.1177/0049124110392533

Valliant, R., Dorfman, A. H. and Royall, R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley. MR1784794

Var Der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. New York: Springer. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2

Wang, L., Valliant, R. and Li, Y. (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer—based epidemiologic cohorts. *Statistics in Medicine* **40**, 5237–5250. MR4330549 https://doi.org/10.1002/sim.9122

Yang, S., Kim, J. and Hwang, Y. (2021). Integration of survey data and big observational data for finite population inference using mass imputation. *Survey Methodology* **47**, 29–58.

Yang, S., Kim, J. K. and Song, R. (2020). Doubly robust inference when combining probability and non-probability samples with high dimensional data. *Journal of the Royal Statistical Society, Series B, Statistical Methodology* **82**, 445–465. MR4084171

Yu, G. and Liu, Y. (2016). Sparse regression incorporating graphical structure among predictors. *Journal of the American Statistical Association* **111**, 707–720. MR3538699 https://doi.org/10.1080/01621459.2015.1034319
Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B, Methodological* **68**, 49–67. MR2212574 https://doi.org/10.1111/j.1467-9868.2005.00532.x
Zhao, T., Li, X., Liu, H., et al. (2015). huge: High-dimensional undirected graph estimation. R package version 4.2.1.
Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429. MR2279469 https://doi.org/10.1198/016214506000000735

Brazilian Journal of Probability and Statistics
2025, Vol. 39, No. 1, 79–108
https://doi.org/10.1214/25-BJPS625

Nonlinear log-wavelet-variance regression for perturbed 2D long memory Gaussian random fields

Jing Wang[a] **and Xiaojiang Yu**[b]

Department of Statistics, Tiangong University, Tianjin, P.R. China, [a]*jingwang2019@qq.com*, [b]*yuxia2016@qq.com*

Abstract. In this paper, we apply the wavelet method to the estimation of the memory parameter of a two-dimensional (intrinsically) stationary Gaussian long memory random field, which is perturbed by an independent, additive and stationary Gaussian short memory noise field. We propose the nonlinear log-wavelet-variance regression (NLWVR) estimator, and determine its consistency with convergence rate as well as asymptotic normality, basically when the memory parameter of the signal field is larger than -1. The finite sample performance of the NLWVR estimator of perturbed stationary long memory random fields is examined by Monte Carlo simulation. In the simulation, we also include the comparison between the NLWVR estimator and the classic wavelet-variance-based linear regression estimator.

References

Arteche, J. (2004). Gaussian semiparametric estimation in long memory in stochastic volatility and signal plus noise models. *Journal of Econometrics* **119**, 131–154. MR2041895 https://doi.org/10.1016/S0304-4076(03)00158-1

Arteche, J. (2006). Semiparametric estimation in perturbed long memory series. *Computational Statistics & Data Analysis* **51**, 2118–2141. MR2307490 https://doi.org/10.1016/j.csda.2006.07.023

Baran, S. and Pap, G. (2009). On the least squares estimator in a nearly unstable sequence of stationary spatial AR models. *Journal of Multivariate Analysis* **100**, 686–698. MR2478191 https://doi.org/10.1016/j.jmva.2008.07.003

Beran, J., Feng, Y., Ghosh, S. and Kulik, R. (2013). *Long Memory Processes-Probabilistic Properties and Statistical Methods*. Berlin: Springer. MR3075595 https://doi.org/10.1007/978-3-642-35512-7

Chilès, J. and Delfiner, P. (2012). *Geostatistics: Modeling Spatial Uncertainty*, 2nd ed. Hoboken: John Wiley & Sons. MR2850475 https://doi.org/10.1002/9781118136188

Daubechies, I. (1992). *Ten Lectures on Wavelets*. Philadelphia: SIAM. MR1162107 https://doi.org/10.1137/1.9781611970104

Deo, R. and Hurvich, C. (2001). On the log periodogram regression estimator of the memory parameter in long memory stochastic volatility models. *Econometric Theory* **17**, 686–710. MR1862372 https://doi.org/10.1017/S0266466601174025

Dutta, S. and Mondal, D. (2016a). REML estimation with intrinsic Matérn dependence in the spatial linear mixed model. *Electronic Journal of Statistics* **10**, 2856–2893. MR3553914 https://doi.org/10.1214/16-EJS1125

Dutta, S. and Mondal, D. (2016b). Variogram calculations for random fields on regular lattices using quadrature methods. *EnvironMetrics* **27**, 380–395. MR3556660 https://doi.org/10.1002/env.2390

Faÿ, G., Moulines, E., Roueff, F. and Taqqu, M. (2008). Estimators of long-memory: Fourier versus wavelets. *Journal of Econometrics* **151**, 159–177. MR2559823 https://doi.org/10.1016/j.jeconom.2009.03.005

Fletcher, R. (1987). *Practical Methods of Optimization*. Chichester: John Wiley & Sons. MR0955799

Frederiksen, P., Nielsen, F. and Nielsen, M. (2012). Local polynomial Whittle estimation of perturbed fractional processes. *Journal of Econometrics* **167**, 426–447. MR2892086 https://doi.org/10.1016/j.jeconom.2011.09.026

Frías, M., Alonso, F., Ruiz-Medina, M. and Angulo, J. (2008). Semiparametric estimation of spatial long-range dependence. *Journal of Statistical Planning and Inference* **138**, 1479–1495. MR2389146 https://doi.org/10.1016/j.jspi.2007.07.005

Key words and phrases. Long memory random field, noise perturbation, intrinsically stationary, semiparametric estimation, discrete wavelet transform.

Hurvich, C. M., Moulines, E. and Soulier, P. (2005). Estimating long memory in volatility. *Econometrica* **73**, 1283–1328. MR2149248 https://doi.org/10.1111/j.1468-0262.2005.00616.x
Hurvich, C. M. and Ray, R. K. (2003). The local Whittle estimator of long-memory stochastic volatility. *Journal of Financial Econometrics* **1**, 445–470.
Kohli, P. (2016). Fractional bivariate exponential estimator for long-range dependent random field. *Spatial Statistics* **15**, 22–38. MR3457666 https://doi.org/10.1016/j.spasta.2015.11.001
Künsch, H. (1987). Intrinsic autoregressions and related models on the two-dimensional lattice. *Biometrika* **74**, 517–524. MR0909356 https://doi.org/10.1093/biomet/74.3.517
Lee, T. and Berman, M. (1997). Nonparametric estimation and simulation of two-dimensional Gaussian image texture. *Graphical Models and Image Processing* **59**, 434–445.
Liu, J., Hwang, W. and Chen, M. (2000). Estimation of 2-D noisy fractional Brownian motion and its applications using wavelets. *IEEE Transactions on Image Processing* **9**, 1407–1419. MR1808278 https://doi.org/10.1109/83.855435
Matheron, G. (1973). The intrinsic random functions and their applications. *Advances in Applied Probability* **5**, 439–468. MR0356209 https://doi.org/10.2307/1425829
Mondal, D. and Percival, D. (2012). Wavelet variance analysis for random fields on a regular lattice. *IEEE Transactions on Image Processing* **21**, 537–549. MR2920639 https://doi.org/10.1109/TIP.2011.2164412
Moulines, E., Roueff, F. and Taqqu, M. (2007a). On the spectral density of the wavelet coefficients of long memory time series with application to the log-regression estimation of the memory parameter. *Journal of Time Series Analysis* **28**, 155–187. MR2345656 https://doi.org/10.1111/j.1467-9892.2006.00502.x
Moulines, E., Roueff, F. and Taqqu, M. (2007b). Central limit theorem for the log-regression wavelet estimation of the memory parameter in the Gaussian semi-parametric context. *Fractals* **15**, 301–313. MR2396718 https://doi.org/10.1142/S0218348X07003721
Nicolis, O., Ramírez-Cobo, P. and Vidakovic, B. (2011). 2D wavelet-based spectra with applications. *Computational Statistics & Data Analysis* **55**, 738–751. MR2736593 https://doi.org/10.1016/j.csda.2010.06.020
Rawlings, J., Pantula, S. and Dickey, D. (1998). *Applied Regression Analysis: A Research Tool*, 2nd ed. New York: Springer. MR1631919 https://doi.org/10.1007/b98890
Reed, I., Lee, P. and Truong, T. (1995). Spectral representation of fractional Brownian motion in n dimensions and its properties. *IEEE Transactions on Information Theory* **41**, 1439–1451. MR1366329 https://doi.org/10.1109/18.412687
Richard, F. (2016). Tests of isotropy for rough textures of trended images. *Statistica Sinica* **26**, 1279–1304. MR3559953
Stein, M. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. New York: Springer. MR1697409 https://doi.org/10.1007/978-1-4612-1494-6
Sun, Y. and Phillips, P. (2003). Nonlinear log-periodogram regression for perturbed fractional processes. *Journal of Econometrics* **115**, 355–389. MR1984781 https://doi.org/10.1016/S0304-4076(03)00115-5
Wang, J. and Yu, X. (2025). Supplement to "Nonlinear Log-Wavelet-Variance Regression for Perturbed 2D Long Memory Gaussian Random Fields." https://doi.org/10.1214/25-BJPS625SUPP
Wang, L. (2009). Memory parameter estimation for long range dependent random fields. *Statistics & Probability Letters* **79**, 2297–2306. MR2591985 https://doi.org/10.1016/j.spl.2009.07.030
Wang, L. and Wang, J. (2014). Wavelet estimation of the memory parameter for long range dependent random fields. *Statistical Papers* **55**, 1145–1158. MR3266386 https://doi.org/10.1007/s00362-013-0558-2
Yajima, Y. and Matsuda, Y. (2020). Log-periodogram regression of two-dimensional intrinsically stationary random fields. *Japanese Journal of Statistics and Data Science* **3**, 333–347. MR4125486 https://doi.org/10.1007/s42081-020-00078-9
Yajima, Y. and Matsuda, Y. (2023). Gaussian semiparametric estimation of two-dimensional intrinsically stationary random fields. DSSR Discussion Papers, Graduate. https://ideas.repec.org/s/toh/dssraa.html.
Zhu, Z. and Stein, M. (2002). Parameter estimation for fractional Brownian surface. *Statistica Sinica* **12**, 863–883. MR1929968

Brazilian Journal of Probability and Statistics
2025, Vol. 39, No. 1, 109–133
https://doi.org/10.1214/25-BJPS626

A performance comparison of KDE fixed-bandwidth selectors in density estimation

David de Melo Souza, Luan Gomes M. de Carvalho and Rafael Antunes Nóbrega[a]

[1] *Faculdade de Engenharia, Universidade Federal de Juiz de Fora, Juiz de Fora, Brazil,* [a] *rafael.nobrega@ufjf.br*

Abstract. Kernel Density Estimation stands as the foundation for most contemporary non-parametric probability density function estimation methods. The essence of this technique centers around the bandwidth parameter, and a variety of bandwidth selectors have been created to precisely determine the optimal value for refinement. Despite the existence of numerous articles on this topic, selecting an appropriate method for a specific database remains a challenging issue, and it continues to be an active area of research. This work explores classic and prominent bandwidth selectors with the intention of evaluating their performance, specifically in terms of the error area between the model and its estimation, across different scenarios. Fourteen methods are evaluated in total, comprising six Plug-in and eight Cross-validation based methods. This research provides fresh perspectives on the subject by conducting a comparative study encompassing a diverse set of distributions varying in complexity and a wide range of sample sizes. In addition, it includes a comparison of KDE methods that have not been previously assessed together. This paper provides an in-depth exploration of the distinctions among the examined methods, providing means for a conscious choice of which bandwidth selector to use based on some prior knowledge of the experiment in hand and its data characteristics. Among the tested approaches, some methods show remarkable robustness and accuracy, particularly in handling complex distributions. Key challenges, such as overestimating or underestimating bandwidth, were identified and addressed, offering improved solutions for diverse scenarios.

References

Berlinet, A. and Devroye, L. (1994). A comparison of kernel density estimates. *Publications de L'Institut de Statistique de L'Université de Paris* **38**, 3–59. MR1743393

Bhatia, V. and Mulgrew, B. (2007). Non-parametric likelihood based channel estimator for Gaussian mixture noise. *Signal Processing* **87**, 2569–2586. https://doi.org/10.1016/j.sigpro.2007.04.006

Borrajo, M. I., González-Manteiga, W. and Martínez-Miranda, M. D. (2017). Bandwidth selection for kernel density estimation with length-biased data. *Journal of Nonparametric Statistics* **29**, 636–668. MR3670313 https://doi.org/10.1080/10485252.2017.1339309

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71**, 353–360. MR0767163 https://doi.org/10.1093/biomet/71.2.353

Cao, R., Cuevas, A. and Manteiga, W. G. (1994). A comparative study of several smoothing methods in density estimation. *Computational Statistics & Data Analysis* **17**, 153–176. https://doi.org/10.1016/0167-9473(92)00066-z

Chiu, S.-T. (1996). A comparative review of bandwidth selection for kernel density estimation. *Statistica Sinica* **6**, 129–145. MR1379053

Devroye, D., Beirlant, J., Cao, R., Fraiman, R., Hall, P., Jones, M., Lugosi, G., Mammen, E., Marron, J., Sánchez-Sellero, C., et al (1997). Universal smoothing factor selection in density estimation: Theory and practice. *Test* **6**, 223–320. MR1616896 https://doi.org/10.1007/bf02564701

Devroye, L. (1989). The double kernel method in density estimation. *Annales de L'Institut Henri Poincaré Section B Probabilités et Statistiques* **25**, 533–580. MR1045250

Key words and phrases. Density estimation, bandwidth selectors.

Du, R., Wu, Q., He, X. and Yang, J. (2013). MIL-SKDE: Multiple-instance learning with supervised kernel density estimation. *Signal Processing* **93**, 1471–1484. https://doi.org/10.1016/j.sigpro.2012.07.024

Duong, T. (2007). ks: Kernel Density Estimation and Kernel Discriminant Analysis for Multivariate Data inR. *Journal of Statistical Software* **21**. MR0666869 https://doi.org/10.18637/jss.v021.i07

Erdogmus, D., Agrawal, R. and Principe, J. C. (2005). A mutual information extension to the matched filter. *Signal Processing* **85**, 927–935. https://doi.org/10.1016/j.sigpro.2004.11.018

Feluch, W. and Koronacki, J. (1992). A note on modified cross-validation in density estimation. *Computational Statistics & Data Analysis* **13**, 143–151. MR1160643 https://doi.org/10.1016/0167-9473(92)90002-w

Friedman, J. H. (1997). On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery* 55–77. https://doi.org/10.1023/A:1009778005914

Ghosh, A. K., Chaudhuri, P. and Sengupta, D. (2006). Classification using kernel density estimates. *Technometrics* **48**, 120–132. MR2236534 https://doi.org/10.1198/004017005000000391

Guidoum, A. C. (2024). Kernel estimator and bandwidth selection for density and its derivatives. Available at https://cran.r-project.org/web/packages/kedd/vignettes/kedd.pdf.

Habbema, J. (1974). A stepwise discriminant analysis program using density estimetion. In *Compstat*, 101–110. Heidelberg: Physica-Verlag.

Hall, P. and Marron, J. S. (1987a). Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. *Probability Theory and Related Fields* **74**, 567–581. MR0876256 https://doi.org/10.1007/BF00363516

Hall, P. and Marron, J. S. (1987b). Estimation of integrated squared density derivatives. *Statistics & Probability Letters* **6**, 109–115. MR0907270 https://doi.org/10.1016/0167-7152(87)90083-6

Hall, P. and Marron, J. S. (1991). Local minima in cross-validation functions. *Journal of the Royal Statistical Society, Series B, Methodological* **53**, 245–252. MR1094284 https://doi.org/10.1111/j.2517-6161.1991.tb01822.x

Hang, H., Steinwart, I., Feng, Y. and Suykens, J. A. (2018). Kernel density estimation for dynamical systems. *Journal of Machine Learning Research* **19**, 1260–1308. MR3862442

Hansen, B. E. (2005). Exact mean integrated squared error of higher order kernel estimators. *Econometric Theory* **21**, 1031–1057. MR2200984 https://doi.org/10.1017/S0266466605050528

Hart, J. D. and Yi, S. (1998). One-sided cross-validation. *Journal of the American Statistical Association* **93**, 620–631. MR1631337 https://doi.org/10.2307/2670113

Heidenreich, N.-B., Schindler, A. and Sperlich, S. (2013). Bandwidth selection for kernel density estimation: A review of fully automatic selectors. *AStA Advances in Statistical Analysis* **97**, 403–433. MR3105590 https://doi.org/10.1007/s10182-013-0216-y

Jones, M. and Kappenman, R. (1992). On a class of kernel density estimate bandwidth selectors. *Scandinavian Journal of Statistics*, 337–349. MR1211788

Jones, M. C., Marron, J. S. and Sheather, S. J. (1996a). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association* **91**, 401–407. MR1394097 https://doi.org/10.2307/2291420

Jones, M. C., Marron, J. S. and Sheather, S. J. (1996b). Progress in data-based bandwidth selection for kernel density estimation. *Computational Statistics* **11**, 337–381. MR1415761

Ledl, T. (2016). Kernel density estimation: Theory and application in discriminant analysis. *Austrian Journal of Statistics* **33**, 267. https://doi.org/10.17713/ajs.v33i3.441

Lee, T. C. M. (2001). A stabilized bandwidth selection method for kernel smoothing of the periodogram. *Signal Processing* **81**, 419–430. https://doi.org/10.1016/s0165-1684(00)00218-8

Li, J. and Nehorai, A. (2018). Gaussian mixture learning via adaptive hierarchical clustering. *Signal Processing* **150**, 116–121. https://doi.org/10.1016/j.sigpro.2018.04.013

Liu, B., Yang, Y., Webb, G. I. and Boughton, J. (2009). A comparative study of bandwidth choice in kernel density estimation for naive Bayesian classification. In *Advances in Knowledge Discovery and Data Mining* 302–313. Berlin: Springer. https://doi.org/10.1007/978-3-642-01307-2_29

Loader, C. R., et al (1999). Bandwidth selection: Classical or plug-in? *The Annals of Statistics* **27**, 415–438. MR1714723 https://doi.org/10.1214/aos/1018031201

Marron, J. S. and Chung, S. S. (2001). Presentation of smoothers: The family approach. *Computational Statistics* **16**, 195–207. MR1854198 https://doi.org/10.1007/s001800100059

Matioli, L. C., Santos, S. R., Kleina, M. and Leite, E. A. (2017). A new algorithm for clustering based on kernel density estimation. *Journal of Applied Statistics* **45**, 347–366. MR3764731 https://doi.org/10.1080/02664763.2016.1277191

Mulgrew, B. and Chen, S. (2001). Adaptive minimum-BER decision feedback equalisers for binary signalling. *Signal Processing* **81**, 1479–1489. https://doi.org/10.1016/s0165-1684(01)00035-4

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics* **9**, 65–78. MR0668683

Savchuk, O. Y. (2020). One-sided cross-validation for nonsmooth density functions. *Computational Statistics* **35**, 1253–1272. MR4133116 https://doi.org/10.1007/s00180-019-00938-3
Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: Wiley. MR1191168 https://doi.org/10.1002/9780470316849
Scott, D. W. and Terrell, G. R. (1987). Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association* **82**, 1131–1146. MR0922178 https://doi.org/10.1080/01621459.1987.10478550
Sheather, S. J. (2004). Density estimation. *Statistical Science*, 588–597. MR2185580 https://doi.org/10.1214/088342304000000297
Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B, Methodological* **53**, 683–690. MR1125725
Shwartz, S., Zibulevsky, M. and Schechner, Y. Y. (2005). Fast kernel entropy estimation and optimization. *Signal Processing* **85**, 1045–1058. https://doi.org/10.1016/j.sigpro.2004.11.022
Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall. MR0848134 https://doi.org/10.1007/978-1-4899-3324-9
Souza, D. M. (2020). Estatística não-paramétrica: Estimação, classificação e uma nova abordagem de seleção automática para largura de banda. PhD thesis, Universidade Federal de Juiz de Fora. Available at. https://repositorio.ufjf.br/jspui/handle/ufjf/11934.
Souza, D. M., de Carvalho, L. G. M. and Nóbrega, R. A. (2025). Supplement to “A performance comparison of KDE fixed-bandwidth selectors in density estimation.” https://doi.org/10.1214/25-BJPS626SUPP
Stute, W. (1992). Modified cross-validation in density estimation. *Journal of Statistical Planning and Inference* **30**, 293–305. MR1165648 https://doi.org/10.1016/0378-3758(92)90157-N

Brazilian Journal of Probability and Statistics
2025, Vol. 39, No. 1, 134–149
https://doi.org/10.1214/25-BJPS627

Optimal variable acceptance sampling based on decision tree method: A Bayesian approach under Type-II censoring

A. M. Mathai[1,a], **M. Kumar**[1,b] **and J. T. Thomas**[2,c]

[1]*Department of Mathematics, National Institute of Technology, Calicut, Pin Code - 673601, India,* [a]*ashlyn_p190073ma@nitc.ac.in,* [b]*mahesh@nitc.ac.in*

[2]*School of Engineering, University of Wollongong in Dubai, Dubai Knowledge Park, PO Box 20183, UAE,* [c]*juliat@uow.edu.au*

Abstract. Acceptance sampling plans are essential for quality assurance in reliability testing. This study introduces a novel approach to variable acceptance sampling plans under Type-II censoring, assuming an exponential distribution for failure times with an inverse gamma prior. The proposed model applies a decision tree algorithm, simplifying the process compared to traditional methods that rely on complex nonlinear or stochastic optimization. By using Bayesian inference, the model incorporates prior knowledge and updates decisions with new data, minimizing the number of failures required to terminate the test, while ensuring product reliability and reducing testing costs. The decision tree method with backward induction is used to evaluate costs from the leaf nodes to the root, ensuring optimal decision-making with minimum cost. A real-life case study illustrates the practical applications of this method, and a comparative analysis with an existing model demonstrates that it is not only simpler but also has a significantly higher power of test, making it suitable for real-world applications. A sensitivity analysis study is also conducted using simulated data to assess how changes in cost parameters and prior parameters influence the optimal decision cost.

References

AlSultan, R. and Al-Omari, A. (2023). Zeghdoudi distribution in acceptance sampling plans based on truncated life tests with real data application. *Decision Making: Applications in Management and Engineering* **6**, 432–448. https://doi.org/10.31181/dmame05012023a

Aminzadeh, M. S. (2003). Bayesian economic variable acceptance-sampling plan using inverse Gaussian model and step-loss function. *Communications in Statistics - Theory and Methods* **32**, 961–982. MR1982762 https://doi.org/10.1081/STA-120019956

Azam, M., Aslam, M. and Niaki, S. T. A. (2020). A repetitive sampling plan using decision trees method. *Journal of Statistics & Management Systems* **23**, 789–807. https://doi.org/10.1080/09720510.2019.1694738

Balakrishnan, N. and Cramer, E. (2014). *The Art of Progressive Censoring*. *Statistics for Industry and Technology*. New York: Springer. MR3309531 https://doi.org/10.1007/978-0-8176-4807-7

Balakrishnan, N. and Kundu, D. (2013). Hybrid censoring: Models, inferential results and applications. *Computational Statistics & Data Analysis* **57**, 166–209. MR2981081 https://doi.org/10.1016/j.csda.2012.03.025

Chakrabarty, J. B., Roy, S. and Chowdhury, S. (2023). On the economic design of optimal sampling plan under accelerated life test setting. *International Journal of Quality and Reliability Management* **40**, 1683–1705. https://doi.org/10.1108/IJQRM-07-2021-0235

Chen, L.-S., Liang, T. and Yang, M.-C. (2022). Designing Bayesian sampling plans for simple step-stress of accelerated life test on censored data. *Journal of Statistical Computation and Simulation* **92**, 395–415. MR4356781 https://doi.org/10.1080/00949655.2021.1961771

Chen, L.-S., Yang, M.-C. and Liang, T. (2016). Curtailed Bayesian sampling plans for exponential distributions based on Type-II censored samples. *Journal of Statistical Computation and Simulation* **87**, 1160–1178. MR3606844 https://doi.org/10.1080/00949655.2016.1254215

Epstein, B. (1954). Truncated life tests in the exponential case. *The Annals of Mathematical Statistics* **25**, 555–564. MR0064372 https://doi.org/10.1214/aoms/1177728723

Key words and phrases. Acceptance sampling plan, Bayseian estimation, Type-II censoring, exponential distribution, decision tree method.

Fallahnezhad, M. S. and Aslam, M. (2013). A new economical design of acceptance sampling models using Bayesian inference. *Accreditation and Quality Assurance* **18**, 187–195. https://doi.org/10.1007/s00769-013-0984-9

Fallahnezhad, M. S. and Babadi, A. Y. (2015). A new acceptance sampling plan using Bayesian approach in the presence of inspection errors. *Transactions of the Institute of Measurement and Control* **37**, 1060–1073. https://doi.org/10.1177/0142331214555395

Fallahnezhad, M. S., Niaki, S. T. A. and Vahdat Zad, M. A. (2012). A new acceptance sampling design using Bayesian modeling and backward induction. *International Journal of Engineering* **25**, 45–54. https://doi.org/10.5829/idosi.ije.2012.25.01c.06

González, C. and Palomo, G. (2003). Bayesian acceptance sampling plans following economic criteria: An application to paper pulp manufacturing. *Journal of Applied Statistics* **30**, 319–333. MR1963204 https://doi.org/10.1080/0266476022000030093

Khoolenjani, N. B. and Shahsanaie, F. (2016). Estimating the parameter of exponential distribution under Type-II censoring from fuzzy data. *Journal of Statistical Theory and Applications* **15**, 181–195. MR3519349 https://doi.org/10.2991/jsta.2016.15.2.8

Kumar, M., Bajeel, P. N. and Ramyamol, P. C. (2020). Optimal design of reliability acceptance sampling plan based on data obtained from partially accelerated life test (PALT). *International Journal of Reliability, Quality, and Safety Engineering* **27**, 2040002. MR3970471 https://doi.org/10.1142/S0218539320400021

Kumar, M. and Ramyamol, P. C. (2016). Design of optimal reliability acceptance sampling plans for exponential distribution. *Economic Quality Control* **31**, 23–36. MR3507939 https://doi.org/10.1515/eqc-2015-0005

Lee, A. H., Wu, C.-W., Wang, T.-C. and Kuo, M.-H. (2024). Construction of acceptance sampling schemes for exponential lifetime products with progressive Type-II right censoring. *Reliability Engineering & System Safety* **243**, 109843. https://doi.org/10.1016/j.ress.2023.109843

Liang, T. and Yang, M.-C. (2011). Optimal Bayesian sampling plans for exponential distributions based on hybrid censored samples. *Journal of Statistical Computation and Simulation* **83**, 922–940. MR3054776 https://doi.org/10.1080/00949655.2011.642378

Lin, C.-T., Huang, Y.-L. and Balakrishnan, N. (2008). Exact Bayesian variable sampling plans for the exponential distribution based on Type-I and Type-II hybrid censored samples. *Communications in Statistics—Simulation and Computation* **37**, 1101–1116. MR2528263 https://doi.org/10.1080/03610910801923869

Lou, Y., Zheng, H. and Yang, J. (2023). Optimal design of reliability acceptance sampling plan based on constant-stress accelerated life testing considering the uncertainty of accelerated factor. *Quality and Reliability Engineering International* **39**, 1318–1333. https://doi.org/10.1002/qre.3295

Mathai, A. M. and Kumar, M. (2024). Optimal variable acceptance sampling plan for exponential distribution using Bayesian estimate under Type-I hybrid censoring. *Hacettepe Journal of Mathematics and Statistics* **53**, 1178–1195. MR4791091 https://doi.org/10.15672/hujms.1355348

Noughabi, H. A. (2015). Testing exponentiality based on the likelihood ratio and power comparison. *Annals of Data Science* **2**, 195–204. https://doi.org/10.1007/s40745-015-0041-0

Peng, B., Yang, K. and Dong, X. (2024). Variable selection for quantile autoregressive model: Bayesian methods versus classical methods. *Journal of Applied Statistics* **51**, 1098–1130. MR4732507 https://doi.org/10.1080/02664763.2023.2178642

Peng, X.-Y. and Yan, Z.-Z. (2013). Bayesian estimation for generalized exponential distribution based on progressive Type-I interval censoring. *Acta Mathematicae Applicatae Sinica English Series* **29**, 391–402. MR3090250 https://doi.org/10.1007/s10255-013-0222-6

Prajapat, K., Koley, A., Mitra, S. and Kundu, D. (2023). An optimal Bayesian sampling plan for two-parameter exponential distribution under Type-I hybrid censoring. *Sankhya Series A* **85**, 512–539. MR4540807 https://doi.org/10.1007/s13171-021-00263-2

Prajapati, D. and Kundu, D. (2023). Bayesian acceptance sampling plan for simple step-stress model. *Communications in Statistics Simulation and Computation* **53**, 6411–6431. MR4837413 https://doi.org/10.1080/03610918.2023.2234664

Ramyamol, P. C. and Kumar, M. (2020). Optimal reliability acceptance sampling plan based on accelerated life test data. *International Journal of Productivity and Quality Management* **30**, 354–370. https://doi.org/10.1504/IJPQM.2020.108377

Thomas, J. T. and Kumar, M. (2023). An optimal Bayesian acceptance sampling plan using decision tree method. *International Journal of Applied Management Science* **15**, 311–325. https://doi.org/10.1504/IJAMS.2023.134457

Tsai, T.-R., Chiang, J.-Y., Liang, T. and Yang, M.-C. (2012). Efficient Bayesian sampling plans for exponential distributions with Type-I censored samples. *Journal of Statistical Computation and Simulation* **84**, 964–981. MR3169373 https://doi.org/10.1080/00949655.2012.736513

Yang, K., Yu, X., Zhang, Q. and Dong, X. (2022). On MCMC sampling in self-exciting integer-valued threshold time series models. *Computational Statistics & Data Analysis* **169**, 107410. MR4359382 https://doi.org/10.1016/j.csda.2021.107410

Yeh, L. (1990). An optimal single variable sampling plan with censoring. *Journal of the Royal Statistical Society Series D The Statistician* **39**, 53–66. https://doi.org/10.2307/2348194

Yeh, L. (1994). Bayesian variable sampling plans for the exponential distribution with Type-I censoring. *The Annals of Statistics* **22**, 696–711. MR1292536 https://doi.org/10.1214/aos/1176325491

Yeh, L. and Choy, S. T. B. (1995). Bayesian variable sampling plans for the exponential distribution with uniformly distributed random censoring. *Journal of Statistical Planning and Inference* **47**, 277–293. MR1364998 https://doi.org/10.1016/0378-3758(94)00136-J

AIMS AND SCOPE

The Brazilian Journal of Probability and Statistics aims to publish high quality research papers in applied probability, applied statistics, computational statistics, mathematical statistics, probability theory and stochastic processes.

More specifically, the following types of contributions will be considered:

(i) Original articles dealing with methodological developments, comparison of competing techniques or their computational aspects;
(ii) Original articles developing theoretical results;
(iii) Articles that contain novel applications of existing methodologies to practical problems. For these papers the focus is in the importance and originality of the applied problem, as well as, applications of the best available methodologies to solve it.
(iv) Survey articles containing a thorough coverage of topics of broad interest to probability and statistics. The journal will occasionally publish book reviews, invited papers and essays on the teaching of statistics.

GENERAL INFORMATION

Journal Homepage: https://www.imstat.org/journals-and-publications/brazilian-journal-of-probability-and-statistics/

Submissions: Manuscripts for *Brazilian Journal of Probability and Statistics* should be submitted online. Authors may access the Electronic Journals Management System (EJMS) at https://www.e-publications.org/ims/submission/.

Preparation of Manuscripts: https://imstat.org/journals-and-publications/brazilian-journal-of-probability-and-statistics/brazilian-journal-of-probability-and-statistics-manuscript-submission/

Correspondence. Mail concerning membership, subscriptions, nonreceipt claims, copyright permissions, advertising or back issues should be sent to the IMS Dues and Subscription Office, PO Box 729, Middletown, Maryland 21769, USA. Mail concerning submissions or editorial content should be sent to the Editor at nancy@ime.unicamp.br. Mail concerning the production of this journal should be sent to: Geri Mattson at bjps@mattsonpublishing.com.

Individual and Organizational Memberships:
https://www.imstat.org/individual-membership/

Individual and General Subscriptions:
https://imstat.org/journals-and-publications/subscriptions-and-orders/

Electronic Access to IMS Journals:
https://www.imstat.org/journals-and-publications/electronic-access/

The *Brazilian Journal of Probability and Statistics* is an IMS supported journal:
https://www.imstat.org/journals-and-publications/ims-supported-journals/