The New England Journal of Statistics in Data Science Volume 1, 123–125 (2023)
DOI: https://doi.org/10.51387/23-NEJSDS12EDI

Editorial
Modern Bayesian Methods with Applications in Data Science

Dipak K. DEY, Ming-Hui CHEN, Min-ge XIE, HaiYing WANG, and Jing WU

1. BACKGROUND OF THE SPECIAL ISSUE

This special issue is on "Modern Bayesian Methods with Applications in Data Science", originated from the EAC-ISBA 2021 conference, with the theme of celebrating Dr. James O. Berger's 70th birthday.

This issue brings together a collection of thought-provoking discussions and innovative methodologies that shed light on the interplay between frequentist and Bayesian approaches in statistics.

2. THE DISCUSSIONS

James O. Berger's paper [1] sets the stage by providing a comprehensive examination of different types of frequentism and their compatibility with Bayesian reasoning. Through practical examples, he elucidates the strengths and limitations of various frequentist perspectives, offering valuable insights to researchers and practitioners alike.

Van der Vaart [14] provides personal reflections on Berger's classification of frequentists into different types. It offers appraisal for Type I and Type II frequentism, resonating with the author's pragmatism. The discussion highlights the natural acceptance of empirical frequentism across statistical frameworks and the compatibility of procedural frequentism with Bayesian reasoning, grounded in the notions of consistency and compatibility.

Pericchi [7] emphasizes the importance of distinguishing between different types of frequentism and highlights the scientifically compelling "empirical frequentist" approach. It delves into the convergence between frequentist and Bayesian schools, particularly through the lens of objective Bayesian reasoning.

In Rousseau [10], the attention is drawn to Berger's review of error reporting from a frequentist perspective and its connection to Bayesian thinking. The discussion raises intriguing questions about the justification of Neyman-Pearson procedures from an empirical frequentist viewpoint, stimulating thought-provoking discussions around relevant measures of uncertainty and reported errors.

The rejoinder [2] provides valuable insights from prominent researchers who respond to specific comments and observations. Their contributions further enrich the discussions on empirical error, precision in defining frequentism, and empirical frequentist targets in multiple testing scenarios.

3. RESEARCH ARTICLES

In addition to the discussion paper, this issue also features a diverse range of research articles that explore Bayesian and frequentist inferences in various statistical domains.

Porwal and Raftery [8] talk about Bayesian model averaging (BMA), which accounts for model uncertainty in statistical inference tasks. The authors compare eight different model space priors and three adaptive parameter priors in BMA for variable selection in linear regression models. They assess the performance of these prior specifications for various statistical tasks, including parameter estimation, interval estimation, inference, point and interval prediction, through extensive simulation studies based on 14 real datasets. The authors reveal that the beta-binomial model space priors specified in terms of the prior probability of model size performed the best on average for different statistical tasks and datasets.

Vimalajeewa et al. [15] propose a method for wavelet denoising of signals contaminated with Gaussian noise using level dependent shrinkage rules. The method is particularly useful for denoising tasks when the signal-to-noise ratio is low. Through simulations, the proposed method outperformed several standard wavelet shrinkage methods.

Gu et al. [4] focus on scalable marginalization of latent variables in modeling correlated data. The authors introduce innovative approaches, such as Gaussian processes and sparse representation, to overcome the computational complexity associated with large data sets. These techniques have wide-ranging applications in various domains, including molecular dynamics simulation, cellular migration, and agent-based models.

Halder et al. [5] discuss double generalized linear models, which can vary the mean and dispersion across observations, and are applicable to many commonly used distributions. However, there are challenges with model specification when dealing with many covariates and dependent data. To address these challenges, the authors propose a hierarchical model with a spatial random effect, specifically using a Gaussian process specification. They use Bayesian variable selection with a continuous spike-and-slab prior on fixed effects to address the problem of model specification. They showcase the accuracy of their frameworks through

synthetic experiments and apply them to analyze automobile insurance premiums in Connecticut.

Maity and Basu [6] also focus on variable selection, an important topic in data analytics applications. The authors propose a Bayesian approach to selecting the model with the highest posterior probability. The authors use simulated annealing to perform this optimization over the model space and show its feasibility in high-dimensional problems through various simulation studies. They provide theoretical justifications and discuss applications to high-dimensional datasets. The proposed method is implemented in an R package called `sahpm` and is available in R CRAN.

Shen et al. [11] compare tail probabilities between the Bayesian and frequentist methods. The authors investigate why the Bayesian estimator for tail probability is consistently higher than the frequentist estimator and establish sufficient conditions for this phenomenon, using both Jensen's inequality and Taylor series approximations. These analyses point to the convexity of the distribution function. The authors bring up the example of a rainfall in Venezuela that caused over 30,000 deaths, which was not captured by simple frequentist extreme value techniques but was predicted by Bayesian inference using parameter uncertainty and full available data.

Prothero et al. [9] discuss the under-examined aspect of centering in data analysis, specifically in functional data analysis (FDA). The authors suggest that centering along a dimension other than the default can identify a useful mode of variation not previously explored in FDA. They propose a unified framework and new terminology for centering operations, as well as a series of diagnostics for determining the best choice of centering for a given dataset. The authors clearly demonstrate the intuition behind and consequences of each centering choice through informative graphics and explore the application of their diagnostics in several FDA settings. The article also addresses ambiguities in matrix orientation and nomenclature.

Shen et al. [12] consider the envelope model, a dimension reduction method for multivariate linear regression that has gained attention for its modeling flexibility and improved estimation and prediction efficiency. The authors incorporate the partial response envelope model and the simultaneous envelope model into a Bayesian framework and propose a novel Bayesian simultaneous partial envelope model that addresses some of the limitations of both approaches. The method has the flexibility of incorporating prior information and enables coherent quantification of all modeling uncertainty through the posterior distribution of model parameters. A block Metropolis-within-Gibbs algorithm for Markov chain Monte Carlo sampling from the posterior is developed.

Thornton et al. [13] study approximate confidence distribution computing (ACDC), which is a likelihood-free inference method within a frequentist framework that provides frequentist validation for computational inference in problems with unknown or intractable likelihoods. The main theoretical contribution of this work is the identification of a matching condition necessary for frequentist validity of inference from this method, connecting Bayesian and frequentist inferential paradigms. The authors present a data-driven approach to drive ACDC in both Bayesian or frequentist contexts, using a data-dependent proposal function that is general and adaptable to many settings. The ACDC development does not require to use a sufficient statistic, sidestepping a constraint for making a valid inference in an Approximate Bayesian Computing (ABC) method. The paper also includes numerical studies to empirically validate the theoretical results and suggests instances where ACDC outperforms the ABC methods.

Dey et al. [3] introduce the use of graphical Gaussian processes for modeling multivariate spatial data, which is an area that has seen significant growth and usage in spatial data science. While much of the literature has focused on a single or few spatially dependent outcomes, recent attention has been given to modeling and inference for a large number of outcomes. The focus of the article is on scalable graphical models that exploit the notion of conditional independence among a large number of spatial processes to enable fully model-based Bayesian analysis.

4. REMARK

We hope that this special issue will inspire researchers to further explore the fascinating bridges between frequentism and Bayesianism, and simulate further developments of novel methodologies to advance statistics and data science.

ACKNOWLEDGEMENT

We extend our gratitude to all the authors, reviewers, and contributors who have made this issue possible. Their dedication and expertise have ensured the quality and relevance of the papers presented.

REFERENCES

[1] Berger, J. (2022). Four Types of Frequentism and Their Interplay with Bayesianism. *The New England Journal of Statistics in Data Science* 1–12. https://doi.org/10.51387/22-NEJSDS4.

[2] Berger, J. (2022). Rejoinder of "Four Types of Frequentism and Their Interplay with Bayesianism". *The New England Journal of Statistics in Data Science* 1–2. https://doi.org/10.51387/22-NEJSDS4REJ.

[3] Dey, D., Datta, A. and Banerjee, S. (2023). Modeling Multivariate Spatial Dependencies Using Graphical Models. *The New England Journal of Statistics in Data Science* 1–13. https://doi.org/10.51387/23-NEJSDS47.

[4] Gu, M., Liu, X., Fang, X. and Tang, S. (2022). Scalable Marginalization of Correlated Latent Variables with Applications to Learning Particle Interaction Kernels. *The New England Journal of Statistics in Data Science* 1–15. https://doi.org/10.51387/22-NEJSDS13.

[5] Halder, A., Mohammed, S. and Dey, D. K. (2023). Bayesian Variable Selection in Double Generalized Linear Tweedie Spatial Process Models. *The New England Journal of Statistics in Data Science* 1–13. https://doi.org/10.51387/23-NEJSDS37.

[6] Maity, A. K. and Basu, S. (2023). Highest Posterior Model Computation and Variable Selection via Simulated Annealing. *The New England Journal of Statistics in Data Science* 1–8. https://doi.org/10.51387/23-NEJSDS40.

[7] Pericchi, L. (2023). Invited Discussion of J.O. Berger: Four Types of Frequentism and Their Interplay with Bayesianism. *The New England Journal of Statistics in Data Science* 1–3. https://doi.org/10.51387/23-NEJSDS4B.

[8] Porwal, A. and Raftery, A. E. (2022). Effect of Model Space Priors on Statistical Inference with Model Uncertainty. *The New England Journal of Statistics in Data Science* 1–10. https://doi.org/10.51387/22-NEJSDS14.

[9] Prothero, J., Hannig, J. and Marron, J. S. (2023). New Perspectives on Centering. *The New England Journal of Statistics in Data Science* 1–21. https://doi.org/10.51387/23-NEJSDS31.

[10] Rousseau, J. (2023). Discussion of: Four Types of Frequentism and Their Interplay with Bayesianism, by J. Berger. *The New England Journal of Statistics in Data Science* 1–2. https://doi.org/10.51387/23-NEJSDS4C.

[11] Shen, N., González-Arévalo, B. and Pericchi, L. R. (2023). Comparison Between Bayesian and Frequentist Tail Probability Estimates. *The New England Journal of Statistics in Data Science* 1–8. https://doi.org/10.51387/23-NEJSDS39.

[12] Shen, Y., Park, Y., Chakraborty, S. and Zhang, C. (2023). Bayesian Simultaneous Partial Envelope Model with Application to an Imaging Genetics Analysis. *The New England Journal of Statistics in Data Science* 1–33. https://doi.org/10.51387/23-NEJSDS23.

[13] Thornton, S., Li, W. and Xie, M. (2023). Approximate Confidence Distribution Computing. *The New England Journal of Statistics in Data Science* 1–13. https://doi.org/10.51387/23-NEJSDS38.

[14] van der Vaart, A. (2022). Frequentism. *The New England Journal of Statistics in Data Science* 1–4. https://doi.org/10.51387/22-NEJSDS4A.

[15] Vimalajeewa, D., DasGupta, A., Ruggeri, F. and Vidakovic, B. (2023). Gamma-Minimax Wavelet Shrinkage for Signals with Low SNR. *The New England Journal of Statistics in Data Science* 1–13. https://doi.org/10.51387/23-NEJSDS43.

Dipak K. Dey. Department of Statistics, University of Connecticut, USA.
E-mail address: dipak.dey@uconn.edu

Ming-Hui Chen. Department of Statistics, University of Connecticut, USA.
E-mail address: ming-hui.chen@uconn.edu

Min-ge Xie. Department of Statistics, Rutgers University, USA.
E-mail address: mxie@stat.rutgers.edu

HaiYing Wang. Department of Statistics, University of Connecticut, USA.
E-mail address: haiying.wang@uconn.edu

Jing Wu. Department of Computer Science and Statistics, University of Rhode Island, USA.
E-mail address: jing_wu@uri.edu