The New England Journal of Statistics in Data Science Volume 1, 297–298 (2023)
DOI: https://doi.org/10.51387/23-NEJSDS13EDI

Editorial
Design and Analysis of Experiments for Data Science

HaiYing WANG, Xinwei DENG, Devon LIN, Ming-Hui CHEN, Min-ge XIE, and Jing WU

1. INTRODUCTION

We are pleased to introduce this special issue dedicated to "Design and Analysis of Experiments for Data Science". The statistical methodology for design and analysis of experiments has played a pivotal role in various scientific fields for over a century. It guides researchers to make valid inference from experiments with optimized resource allocation. This enduring discipline remains a vibrant and dynamic field of modern research. This collection of papers present cutting-edge research and innovative methodologies related to experimental designs. They provide valuable insights to address challenges of modern experimental designs and demonstrate the pivotal role of experimental designs in various domains, from clinical trials to online experimentation and beyond, in the increasingly data-driven world.

2. RESEARCH ARTICLES

Chen et al. [2] showcase how Particle Swarm Optimization (PSO) can efficiently find optimal designs for longitudinal studies with diverse correlation structures and different models. Longitudinal studies present a unique set of challenges when it comes to experimental designs. The application of PSO in this context is a game-changer. The potential applications are far-reaching, from the Michaelis-Menten model to growth curve studies and HIV dynamic modeling. The optimization provided by PSO opens up new avenues for the scientific community.

Clinical trialists often grapple with the need to balance statistical power, ethical considerations, and control over the type I error rate. Zhu et al. [11] propose an adaptive seamless design (ASD) in conjunction with response adaptive randomization (RAR) to address these challenges. This innovative approach demonstrates how it is possible to achieve efficient and ethical objectives while maintaining control over the type I error rate, a crucial aspect in clinical trials.

The task of discovering significant factors in experiments with a large number of variables and limited observations is a common problem in data science. Qi and Chien [7] introduce a new approach to construct supersaturated designs with low coherence, enhancing their usability for variable selection, particularly in methods like the Lasso. The real-world examples provided illustrate the practical value of this approach.

Hyperparameter tuning is a key factor in the success of deep learning models. However, the computational expense involved can be prohibitive. Shi et al. [9] delve into the exploration of hyperparameters and the collection of informative data for deep learning techniques. The findings demonstrate the superiority of the technique of strong orthogonal array in efficiently collecting data to improve the test accuracy with deep learning, a critical measure of learning performance.

Bui et al. [1] propose a general additive network effect (GANE) model. Innovative experiments are essential for businesses, and social network companies are at the forefront of testing new ideas and product changes. The unique challenges these experiments pose require specialized approaches. The introduction of the GANE model provides a comprehensive framework for understanding treatment effects and network influence. The proposed power-degree specification showcases how specialized experiments can yield precise results, even in the face of model misspecification.

Crossover and interference models have diverse applications, but their complex nature has limited their practical utility. Hao et al. [3] present an algorithm designed to efficiently generate crossover designs under various conditions. The inclusion of a user-friendly interface and an R package broadens its accessibility and usability in a wide range of applications.

Systems with both quantitative and qualitative responses require specialized experimental designs. Kang et al. [4] introduce a Bayesian D-optimal design method that caters to such systems, providing an efficient approach to constructing both local and global D-optimal designs. The inclusion of prior distributions and a point-exchange search algorithm allows for meaningful interpretations and practical application in various contexts.

Personalized decision-making in controlled experiments is of growing interest, particularly in clinical trials and user behavior studies. Li et al. [5] tackle the challenge of optimizing treatment allocation while considering observational covariates. The proposed method seeks to maximize the variance of personalized treatment effects, enhancing precision

in decision-making processes. Numerical studies validate the method's quality and applicability.

Pronzato and Rendas [6] address integrated squared error (ISE) estimation for predicting unknown functions. They calculate ISE estimators as weighted averages of predictor residuals at selected points. Their study shows that minimizing mean squared error of ISE estimators is equivalent to minimizing maximum mean discrepancy (MMD). Sequential Bayesian quadrature is utilized to create nested validation designs that minimize MMD at each step. The optimal ISE estimate is expressed as the integral of a linear reconstruction of interpolator residual squares over the function's domain. The validation designs maintain a space-filling property. Numerical experiments validate the method's strong performance and robustness.

Online experimentation frequently encounters incomplete metric data, making imputation an essential step for analysis. Shen et al. [8] introduce a clustering-based imputation method, considering both experiment-specific features and user activities, to improve the analysis of online experiments. This research lays the foundation for more efficient imputation of large-scale data in online experimentation, benefiting both simulations and real-world experiments.

The explosion of big data requires efficient subdata selection methods to perform inferences while managing computational costs. Singh and Stufken [10] present the information-based optimal subdata selection method for selecting subdata with strong statistical properties in linear regression models. The idea of combining lasso and subdata selection extends the capabilities of subdata selection, offering an efficient solution for handling large datasets with many variables.

3. REMARK

We hope that this special issue, featuring eleven diverse and innovative research papers, will inspire further exploration and innovation in the field of experimental designs for data science. The methodologies and insights presented in these papers have the potential to transform how experiments are conducted and analyzed in our data-driven world. We invite you to delve into these papers to discover the latest advancements and ideas on experimental designs in data science.

ACKNOWLEDGEMENTS

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] Bui, T., Steiner, S. and Stevens, N. (2023). General Additive Network Effect Models. *The New England Journal of Statistics in Data Science* 1–19. https://doi.org/10.51387/23-NEJSDS29.

[2] Chen, P.-Y., Chen, R.-B. and Wong, W. K. (2023). Particle Swarm Optimization for Finding Efficient Longitudinal Exact Designs for Nonlinear Models. *The New England Journal of Statistics in Data Science* 1–15. https://doi.org/10.51387/23-NEJSDS45.

[3] Hao, S., Yang, M. and Zheng, W. (2023). Algorithm-Based Optimal and Efficient Exact Experimental Designs for Crossover and Interference Models. *The New England Journal of Statistics in Data Science* 1–10. https://doi.org/10.51387/23-NEJSDS41.

[4] Kang, L., Deng, X. and Jin, R. (2023). Bayesian D-Optimal Design of Experiments with Quantitative and Qualitative Responses. *The New England Journal of Statistics in Data Science* 1–15. https://doi.org/10.51387/23-NEJSDS30.

[5] Li, Y., Zhang, Q., Khademi, A. and Yang, B. (2023). Optimal Design of Controlled Experiments for Personalized Decision Making in the Presence of Observational Covariates. *The New England Journal of Statistics in Data Science* 1–8. https://doi.org/10.51387/23-NEJSDS22.

[6] Pronzato, L. and Rendas, M.-J. (2023). Validation of Machine Learning Prediction Models. *The New England Journal of Statistics in Data Science* 1–21. https://doi.org/10.51387/23-NEJSDS50.

[7] Qi, Y. and Chien, P. (2023). Construction of Supersaturated Designs with Small Coherence for Variable Selection. *The New England Journal of Statistics in Data Science* 1–11. https://doi.org/10.51387/23-NEJSDS34.

[8] Shen, S., Mao, H., Zhang, Z., Chen, Z., Nie, K. and Deng, X. (2023). Clustering-Based Imputation for Dropout Buyers in Large-Scale Online Experimentation. *The New England Journal of Statistics in Data Science* 1–11. https://doi.org/10.51387/23-NEJSDS33.

[9] Shi, C., Chiu, A. K. and Xu, H. (2023). Evaluating Designs for Hyperparameter Tuning in Deep Neural Networks. *The New England Journal of Statistics in Data Science* 1–8. https://doi.org/10.51387/23-NEJSDS26.

[10] Singh, R. and Stufken, J. (2023). Subdata Selection With a Large Number of Variables. *The New England Journal of Statistics in Data Science* 1–13. https://doi.org/10.51387/23-NEJSDS36.

[11] Zhu, H., Yu, J., Lai, D. and Wang, L. (2023). Seamless Clinical Trials with Doubly Adaptive Biased Coin Designs. *The New England Journal of Statistics in Data Science* 1–9. https://doi.org/10.51387/23-NEJSDS25.

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

Xinwei Deng. Department of Statistics, Virginia Tech, USA. E-mail address: xdeng@vt.edu

Devon Lin. Department of Mathematics and Statistics, Queen's University, Canada. E-mail address: devon.lin@queensu.ca

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

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