ESTIMATING TREATMENT EFFECTS FROM COSTLY SAMPLES OF POPULATION-SCALE MODELS

Abdulrahman A. Ahmed[a]

[a]Department of Industrial Engineering, University of Pittsburgh, PA, USA
aba173@pitt.edu

ABSTRACT

Large-scale simulations require computationally exhaustive resources, especially when the number of treatment conditions is large. Therefore, it requires careful study to propose methods that can estimate treatment effects accurately with relatively few samples. In this study, I will show the accomplished work during my PhD and the possible extensions. The contribution can be broken down into three parts: 1) Sample allocation method and how to choose the best method that estimates accurately with as few samples as possible and has the best bias-variance tradeoff, 2) Model selection given a specific sample size to estimate the treatment effect and what are the main components that control the performance, and 3) What could be a suitable method in inferring the epidemic dynamics in spatiotemporal problem formation.

1 INTRODUCTION

Agent-based simulation provides a useful way to build insights into epidemic dynamics. However, as the simulation becomes very large, getting samples from that simulation becomes computationally exhaustive. Therefore, we need to search for estimation methods that can accurately estimate treatment effects with as few simulation run samples as possible. In this study, we use simulation samples from software named FRED. FRED is an agent-based simulation software that the Public Health Dynamics Laboratory (PHDL) has developed at the University of Pittsburgh School of Public Health (based on U.S. Census data) to simulate spatial and temporal behaviors of epidemic [1]. To understand the Opioid Use Disorder (OUD) epidemic's dynamics better, PHDL develops an OUD model in FRED calibrated with real U.S. data.

2 CONTRIBUTIONS ACCOMPLISHED

The work can be divided into three parts: Methods for sample allocation to estimate treatment effects accurately, model selection to estimate treatment effects given a sample size, and inferring epidemic dynamics from spatiotemporal modeling. First, given a computational exhaustive simulation, what would be the best way to estimate treatment effects with as few samples as possible? The direct method could be brute force (i.e., getting samples until accuracy is achieved). However, this will not scale as the number of treatment effects becomes large, so we need to think of other methods where regression could be one of them. We experiment with different settings for the methods and provide a comparison between them. The results show that regression-based methods performed better than direct estimation methods like the greedy method, which requires 23350 runs to get the same estimate for the ten treatment effects estimated by model-based methods, whereas the first model-based method required 9950 runs and the second required 5650, which is almost the half and quarter respectively of the runs required by model-free method [2]. Second, how to select a method based on a given sample size? In this work, we tested different methods over different sample

Proc. of the 2024 Annual Simulation Conference (ANNSIM'24), May 20-23, 2024, American University, DC, USA
P.J. Giabbanelli, I. David, C. Ruiz-Martin, B. Oakes and R. Cárdenas, eds.

sizes to see which would perform better based on the sample size. The result shows that simpler methods would work better in small sample sizes, and as sample sizes become larger, more complex methods will perform better (until a very large sample size where direct estimation of the average of samples will perform better than other methods). In addition, we provide a theoretical analysis where we show which elements control the MSE equation and how this reflects on a given sample size [3]. Third, can Gaussian process regression (GPR) learn epidemic dynamics from spatiotemporal data? In this study, we explore the potential of GPR and learn the dynamics of two different epidemics, given both cases started from the same settings and the same place [4].

3 REMAINING WORK

The ultimate aim of this research is to provide a suitable method to estimate large-scale simulation treatment effects (especially when the number of treatment effects becomes large). In addition, in the previous section we showed that we develop methods to learn across treatment effects. However, learning across different Counties could also improve the methods for solving the problem of sample allocation to simulate the U.S. population. This is important when interventions in a region can affect the neighboring regions. This problem formation can be explored using spatial Gaussian process which showed potential in [4], as the geospatial factor incorporated with the treatment effect element and the problem becomes more complex. For the model selection problem, we examined the equations for two methods; we plan to extend our study to other methods to see how our results generalize and whether the same variables have the same controlling effect on the MSE equation.

4 SUMMARY

Estimating treatment effects in computationally exhaustive simulations is a challenging task. This requires the development of methods that can get an accurate estimate of treatment effect with as few samples as possible, where we show that model-based methods exceed the performance of model-free methods. In addition, we explored the problem from another perspective, i.e., if we have already the sample size, what would be the best method to select for estimation? We plan to extend this work by exploring new methods for sample allocation that can have a better bias-variance tradeoff, extending the theoretical analysis by examining other methods in the analysis to get more generalized insights.

ACKNOWLEDGMENTS

I would like to thank my supervisor, Dr. M. Amin Rahimian, for the support, encouragement, and patient guidance he has provided since starting my time as a PhD student until now.

This research was funded by contract 75D30121C12574 from the Centers for Disease Control and Prevention. The findings and conclusions in this work are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

This research was partly supported by the University of Pittsburgh Center for Research Computing, RRID:SCR_022735, through the resources provided. Specifically, this work used the HTC and VIZ clusters, which are supported by NIH award number S10OD028483.

REFERENCES

[1] H. Guclu, S. Kumar, D. Galloway, M. Krauland, R. Sood, A. Bocour, T. B. Hershey, E. van Nostrand, and M. Potter, "An agent-based model for addressing the impact of a disaster on access to primary care services," *Disaster Medicine and Public Health Preparedness*, vol. 10, no. 3, pp. 386–393, 2016.

[2] A. A. Ahmed, M. A. Rahimian, and M. S. Roberts, "Estimating treatment effects using costly simulation samples from a population-scale model of opioid use disorder," in *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2023, pp. 1–4.

[3] ——, "Optimized model selection for estimating treatment effects from costly simulations of the us opioid epidemic," under review.

[4] ——, "Inferring epidemic dynamics using gaussian process emulation of agent-based simulations," in *Proceedings of Winter Simulation Conference (WSC)*. IEEE, 2023.