DRP Reflection

In the Direct Reading Program this quarter, I explored a novel statistical method known as Post-Prediction Adaptive Inference (PSPA), introduced by Miao et al. (2024) in their work "Assumption-Lean and Data-Adaptive Post-Prediction Inference." This approach is particularly engineered to overcome the validity challenges faced by existing methods such as Prediction-Powered Inference (PPI), particularly in scientific fields where gold-standard data are scarce or expensive to obtain. Traditional reliance on machine learning (ML) predictions in statistical analyses often leads to biased outcomes due to the imprecision of prediction models, posing the need for a more sophisticated method to integrate ML predictions with observed data.

PSPA enhances the reliability and efficiency of statistical inference by being both assumption-lean and data-adaptive. The "assumption-lean" quality means it does not rely on stringent assumptions regarding the accuracy of the ML models used for generating predictions. This is essential for maintaining validity even when the predictive model is somewhat mis-specified. Furthermore, its "data-adaptive" quality ensures that PSPA optimizes the use of available information. It improves inference by leveraging more data from accurate predictions and reducing reliance on less accurate ones. This adaptability leads to greater efficiency in statistical analyses, accommodating the inherent variability and uncertainty of the predicted data.

Working with the graduate students, I implemented the PSPA simulation in R. First, we constructed three sets of variables, X_i, Y_i and Z_i, to simulate the typical data scenarios encountered in real-world applications. Here, X_i represents the set of covariates or independent variables, Y_i is the scalar outcome variable, and Z_i is the auxiliary variables that are predictive of Y_i but more easily obtainable than Y_i itself. To generate realistic predictions, we employed a random forest model trained on the labeled dataset containing X_i, Y_i and Z_i to generate predictions for Y_i in our unlabeled dataset containing X_i and Z_i. After obtaining the imputed Y_i, we can compute the PSPA estimator using a specific estimating equation derived from the paper. We start by calculating a weighted combination of the observed outcomes Y_i and the ML predicted Y_i made from the auxiliary data Z_i. The weights, or adjustment factors, are optimized based on the accuracy of the ML predictions to minimize variance and bias. The core of the PSPA methodology requires solving the estimating equation of a function including Y_i, X_i and θ equal to zero, where θ represents a parameter of interest. This process effectively reconciles the discrepancies between observed and predicted datasets in statistical inference.

The results demonstrated that PSPA outperforms other methods like the PPI estimator in variance reduction and efficiency, particularly as the size of the unlabeled data and the accuracy of the ML predictions increased. The simulations highlight PSPA's capacity to improve the reliability of statistical analyses in research fields where large portions of data are predicted rather than directly measured, and thus can be a valuable tool for researchers dealing with imperfect data sources. Overall, the "Assumption-Lean and Data-Adaptive Post-Prediction Inference" paper introduces a nuanced approach that significantly enhances the

integration of ML predictions into statistical analyses. Participating in this course of DRP has broadened my understanding and curiosity into more potentials of machine learning and statistical inferences. This experience also expanded my tool box for tackling similar challenges in my own research projects.

References

Miao, J., Miao, X., Wu, Y., Zhao, J., & Lu, Q. (2024). Assumption-Lean and Data-Adaptive Post-Prediction Inference (No. arXiv:2311.14220). arXiv. http://arxiv.org/abs/2311.14220