RESEARCH STATEMENT

Yuki Ohnishi (yohnishi@purdue.edu)

Causal inference is a critical consideration across a broad range of domains in science, technology, engineering, and medicine, as it is the field of study that seeks to determine how manipulating one variable causes a change in an outcome of interest. The inference of causality presents significant challenges, requiring specific assumptions to identify causal relationships accurately. Yet, these assumptions are often compromised in practice, leading to unreliable conclusions about treatment effects. My research focuses on creating statistically valid causal inference methods designed for situations where typical assumptions fail. I aim to devise innovative solutions to overcome the shortcomings of existing techniques, tackling the intricate real-world problems faced in public policy, clinical trials, and social and marketing sciences.

Causal Inference with Violated Assumptions

Three significant sources of complications in causal inference that are increasingly of interest are interference among individuals, nonadherence/noncompliance of individuals to their assigned treatments, and unintended missing outcomes of individuals. Interference exists if the outcome of an individual depends not only on its assigned treatment, but also on the assigned treatments for other units. It arises when limited controls are placed on the interactions of individuals with one another during the course of an experiment. Treatment nonadherence frequently occurs in human subject experiments, as it can be unethical to force an individual to take their assigned treatment. Clinical trials, in particular, typically have subjects that do not adhere to their assigned treatments due to adverse side effects or intercurrent events. Finally, missing values commonly occur in clinical studies. For example, some patients may drop out of the study due to the side effects of the treatment. Failing to account for interference, nonadherence, and missing outcomes will generally yield unstable and biased inferences on treatment effects, even in randomized experiments, which are the gold standard for drawing causal inferences.

In my first project, we introduced a novel Bayesian method that tackles all three challenges simultaneously (Ohnishi and Sabbaghi, 2022). In contrast to existing methods that invoke strong structural assumptions to identify causal effects, our Bayesian approach uses flexible distributional models that can accommodate these complexities, ensuring that causal effects are identifiable. We demonstrated the efficacy of our method through the analyses of real-world data from India's National Health Insurance Program, successfully uncovering more definitive evidence of spillover and overall effects of the intervention that were not recognized in the previous analyses.

My subsequent research (Ohnishi et al., 2023) further addresses one of the limitations of Ohnishi and Sabbaghi (2022): a set of assumptions about interference structures that may be too restrictive in some practical settings. We introduced a concept of the "degree of interference" (DoI), a latent variable capturing the interference structure. This concept allows for handling arbitrary, unknown interference structures to facilitate inference on causal estimands. Additionally, we developed a novel Bayesian semiparametric method, flexible for handling arbitrary interference effects, with an efficient MCMC sampler to draw inferences under our framework. It also segregates units into distinct clusters according to the interference effects they experience, facilitating interpretable group-based analysis. We illustrate the DoI concept and properties of our Bayesian methodology via extensive simulation studies and a real-life case study from additive manufacturing for which interference is a critical concern in achieving geometric accuracy control. Ultimately, our framework enables us to infer causal effects without strong structural assumptions on interference.

While randomized experiments offer a solid foundation for valid causal analysis, people are also interested in conducting causal inference using observational data due to the cost and difficulty of randomized experiments and the wide availability of observational data. Nonetheless, using observational data to infer causality requires us to rely on additional assumptions. A central assumption is that of *ignorability*, which posits that the treatment is randomly assigned based on the variables (covariates) included in the dataset. While crucial, this assumption is often debatable, especially when treatments are assigned sequentially to optimize future outcomes. For instance, in digital marketing, assessing the effectiveness of targeted emails requires consideration of sequential selection bias—where earlier interactions impact subsequent promotions—and noncompliance, such as when recipients fail to open an email. Marketers typically adjust subsequent promotions based on responses to earlier ones and speculate on how customers might have reacted to alternative past promotions. This speculative behavior introduces latent confounders, which must be carefully addressed to prevent biased conclusions. This type of sequential treatment can also be seen as temporal interference, where past assigned treatments affect future outcomes. However, these latent sequential confounders are not adequately addressed by existing methods.

We investigate these issues by studying sequences of promotional emails sent by a US retailer (Ohnishi et al., 2023). We developed a novel Bayesian approach for causal inference from longitudinal observational data that accommodates noncompliance and latent sequential confounding. Our methodology allows us to 1) estimate the average treatment effect of specific email strategies, 2) evaluate the relative effectiveness of these strategies considering varying compliance behaviors, and 3) infer optimal strategies for distinct customer segments. Our research uncovered, among other findings, that certain promotional emails are particularly successful in maintaining engagement among customers who are frequently targeted. Additionally, we observed that individuals who regularly open their emails tend to be less responsive to promotional content.

Future Directions: Causal Inference under Spatiotemporal Interference

While a growing body of research has focused on causal inference in the presence of spatial interference or temporal interference—where units influence each other geographically or temporally—a less explored dimension is spatiotemporal interference, where there is interference in both time and location. For example, if you vaccinate a group of individuals against a disease, this can reduce disease transmission not only within the treated group but potentially in neighboring regions as well due to reduced spread. Similarly, a policy intervention in an economic system may influence not only the present economic conditions but also future conditions because of dynamic interactions. This type of spatiotemporal interference is prevalent in various applied settings, yet there is a noticeable gap in research on how to conduct causal inference in these contexts.

One primary research direction is developing a statistical test to discern spatiotemporal interference effects from longitudinal data. While there is existing literature on detecting spatial interference, the multifaceted nature of spatiotemporal interference necessitates a more advanced methodology. A secondary direction involves defining causal estimands that remain robust under spatiotemporal interference. Subsequently, I aim to develop both design-based and model-based inferential strategies to deduce spatiotemporal spillover effects. Potential avenues for the model-based approach might encompass extending our Degree of Interference (DoI) framework to account for temporal interference, implying a time-varying DoI that addresses both temporal and spatial interference.

Causal Inference on Privatized Data

In the era of digital expansion, the secure handling of sensitive data poses an intricate challenge that significantly influences research, policy-making, and technological innovation. As the collection of sensitive data becomes more widespread across academic, governmental, and corporate sectors, the threat to individual privacy grows increasingly acute. Despite these risks, such data is essential for informed and evidence-based decision-making processes. Addressing the complex balance between making data accessible and safeguarding private information requires the development of sophisticated methods for analysis and reporting, which must include stringent privacy protections. Currently, the gold standard for maintaining this balance is Differential Privacy (DP). DP offers mathematically provable privacy protection against arbitrary privacy breaches while enabling the secure sharing of summary statistics and synthetic data, ensuring privacy without hindering data utility. This probabilistic guarantee is often achieved by adding random noise to the data. One DP model is *local* differential privacy (LDP). In this model, people do not directly provide their data to the data curator; instead, users apply the DP mechanism to their data locally before sending it to the curator. LDP is a preferable model if the data curators are not trusted by users. The LDP framework is receiving growing attention as major organizations, such as Google and Apple, integrate it to enhance their privacy protections.

On the other hand, randomized experiments, where individuals are randomly allocated to treatment or control groups with a specified probability, set the benchmark for conducting valid causal inference. Such experiments are widespread across various fields in science and business, including clinical trials and A/B testing in marketing strategies. Nonetheless, inferring causal relationships from privatized data presents challenges, even within the framework of randomized experiments. While the added random noise helps safeguard individuals' privacy, the noise introduced to protect privacy can obscure actual data patterns, potentially leading to biased conclusions. This issue becomes even more pronounced in the LDP method, where each data point is individually privatized before it is aggregated. Therefore, when trying to understand cause-and-effect relationships using this protected data, researchers must exercise extra caution to ensure their interpretations remain accurate and unbiased.

In our study (Ohnishi and Awan, 2023), we presented how to perform statistically valid causal inferences on locally privatized data in a variety of different privacy scenarios. A fundamental problem we tackled was quantifying and correcting the bias of naive frequentist approaches and quantifying the error of our estimators. We also constructed novel LDP estimators, which we prove are minimax optimal. To faciliate uncertainty quantification of the privatized estimators, we develop both confidence intervals as well as a Bayesian methodology along with an efficient sampler designed for locally privatized data.

Future Directions: Causal Inference on Privatized Data in Observational Studies

Our methodological work for the privacy setting has primarily concentrated on randomized experiments, but the interest in causal inference from observational data is also significant. For example, technological companies often accumulate extensive user data in non-experimental settings and are keen to derive causal insights from this data reservoir. Consequently, a natural extension of our research is to adapt our methodology for use in observational studies.

Two commonly used estimators in observational studies are the inverse probability weighting (IPW) estimator and the doubly robust (DR) estimator. Specifically, the DR estimator combines outcome modeling and treatment assignment modeling. Notably, it is well-known that the DR estimator possesses several desirable statistical properties. One such property is its greater efficiency compared to the IPW estimator in standard, non-private settings. However, in private settings, this is not always the case. This difference arises because the DR estimator requires privacy costs for inferring both outcome and treatment assignment models, potentially compromising efficiency. In contrast, the IPW estimator only requires the treatment assignment model. The goal of this project is to thoroughly understand the trade-off between privacy costs and efficiency. This presents not only practical relevance but also unique theoretical challenges in quantifying efficiency in terms of privacy budgets.

References

Ohnishi, Y. and J. Awan (2023). Locally private causal inference for randomized experiments. *Submitted to Journal of Machine Learning Research*.

Ohnishi, Y., B. Karmakar, and W. Kar (2023). Inferring causal effect of a digital communication strategy under a latent sequential ignorability assumption and treatment noncompliance. *Submitted to Journal of the American Statistical Association*.

Ohnishi, Y., B. Karmakar, and A. Sabbaghi (2023). Degree of interference: A general framework for causal inference under interference. *Submitted to Biometrika*.

Ohnishi, Y. and A. Sabbaghi (2022). A Bayesian Analysis of Two-Stage Randomized Experiments in the Presence of Interference, Treatment Nonadherence, and Missing Outcomes. *Bayesian Analysis*, 1 – 30.