Internship subject: Causal inference meets prediction-powered inference.
Advisors: Charlotte Baey (charlotte.baey@univ-lille.fr, Estelle Kuhn (estelle.kuhn@inrae.fr), Zacharie Naulet (zacharie.naulet@inrae.fr).
Host laboratory: Université Paris-Saclay, INRAE, MaIAGE, 78350, Jouy-en-Josas, France

1 Context

From personalized medicine to algorithmic fairness, causal inference is revolutionizing how we extract meaningful insights from data. This internship will explore modern challenges at the intersection of causality and machine learning where high-dimensional covariates, missing outcomes, and complex dependencies demand innovative solutions.

Causal inference aims to determine whether a treatment T (e.g., a policy, drug, or intervention) has a causal effect on an outcome Y, beyond mere statistical association. Unlike traditional regression, which identifies correlations, causal inference addresses *counterfactuals*: what would have happened if the treatment had not been applied? The *Average Treatment Effect* (ATE) quantifies this causal impact by comparing the mean outcome between treated and control units, accounting for confounding factors W that influences both the treatment and the outcome.

While the Average Treatment Effect (ATE) framework offers a rigorous definition of causal effects, a frequent challenge arises when the outcome of interest Y, is observed for only a limited subset of individuals. Such scenarios, where measuring Y is costly or impractical, are common. A standard approach involves using a surrogate variable X correlated with Y to improve ATE estimation. This concept, introduced by [4], has been further developed in subsequent literature (see [3] and references therein). However, surrogate-based methods often lack theoretical guarantees. For instance, if X is independent of the confounders W conditional on Y (denoted $X \perp W \mid Y$), estimation becomes straightforward. Yet, the dependency structure among (X, T, W, Y) is frequently unclear, and estimates may be severely biased if this conditional independence does not hold (a common occurrence). Additionally, these methods often require developing tailored estimators for specific problems.

Recently, prediction-powered inference (PPI) [1] has emerged as a promising framework that integrates predictive modeling and statistical inference. PPI leverages a small labeled dataset (with observed Y) and a large unlabeled dataset (without Y) to improve estimation while rigorously quantifying uncertainty. The core idea is to train a predictor $g(X)$ for Y using the labeled data and then use this predictor to augment the unlabeled dataset. Although conceptually simple, this approach demands careful implementation. Recently [2] proposed to use PPI to address missing outcomes in ATE estimation. The method is appealing because it provides a systematic approach to ATE estimation when outcomes are missing. The authors suggest using the labeled data to learn a mapping ϕ such that $\phi(X) \perp W \mid Y$, and then augmenting the incomplete dataset $\{(T_i, X_i, W_i)\}$ with $\{\phi(X_i)\}$. Any standard ATE estimation method can then be applied to the augmented dataset. This approach is straightforward to implement and ensures that the ATE estimator remains unbiased, provided $\phi(X)$ is unbiased for Y. This offers a form of weak theoretical guarantee. However, summarizing covariates X through $\phi(X)$ may severely hamper the predictive power of Y. It is also well established in statistics that unbiased estimators are not always optimal.

2 Objectives and research work

This internship is aiming to develop new methodological tools to attain the fundamental limits in estimating causal effects in presence of missing data, and to establish these limits. Statistical decision theory will be used to design predictors of Y based on X that optimally balance the bias-variance trade-off in ATE estimation. The focus will be on high-dimensional settings where X is a high-dimensional vector. Specifically, we will investigate the best achievable performance for ATE stimators in the missing outcome model and assess whether PPI can optimally estimate the ATE.

This internhip is motivated by a dataset of 100 *Brassica napus* (oilseed rape) genotypes, derived from two parental lines. Phenotypic traits (leaf area, leaf count, and carbon/nitrogen content in aerial and root tissues) were recorded under nitrogen-limited conditions for all genotypes, with limited data under non-limiting nitrogen. A key objective is quantifying the causal effects of nitrogen availability on these phenotypes. The candidate may apply developed methods to this problem, with flexibility based on their interests.

3 Desired profile

Candidates should have a BAC+5-level education (Master's degree or engineering school). The ideal candidate should have a strong background in mathematical statistics, with a keen interest in both theoretical developments and real-world applications. Curiosity about applied fields such as genomics, machine learning, or biology is highly valued.

4 Practical details

The internship will take place at the INRAE center in Jouy-en-Josas within the MaIAGE unit. The duration of the internship will be five to six months, between February and September 2026. The monthly stipend is approximately 650 euros (legal rate). The internship will be supervised by Charlotte Baey (Université de Lille), Estelle Kuhn (Université Paris-Saclay, INRAE, MaIAGE), and Zacharie Naulet (Université Paris-Saclay, INRAE, MaIAGE).

This internship can lead to a PhD subject. The candidate must apply for her/his own doctoral school funding.

References

[1] A. N. Angelopoulos, S. Bates, C. Fannjiang, M. I. Jordan, and T. Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023.

[2] R. Cadei, I. Demirel, P. De Bartolomeis, L. Lindorfer, S. Cremer, C. Schmid, and F. Locatello. Causal lifting of neural representations: Zero-shot generalization for causal inferences. *arXiv preprint arXiv:2502.06343*, 2025.

[3] N. Kallus and X. Mao. On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(2):480–509, 2025.

[4] R. L. Prentice. Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in medicine*, 8(4):431–440, 1989.