Efficient Prediction-Powered Inference under Covariate Shift

PHS 7065 Fall 2024 - Final Report

2024-12-15

Contents

1 Introduction

A substantial proportion of missing data is a common challenge in Electronic Health Records (EHR) data analysis, which has the potential to undermine the validity of research findings (Sterne et al., 2009). Machine learning algorithms can be employed to predict missing values based on observed data. However, questions remain about the validity of conclusions drawn from such predicted data. To address this concern, Angelopoulos et al. (2023) proposed Prediction-Powered Inference (PPI), a framework designed to enable provably valid statistical inference when predictions are used as data. The PPI method leverages a gold-standard dataset, consisting of features paired with observed outcomes, to quantify and correct errors made by the machine-learning algorithm on the unlabeled dataset. This allows the constructed confidence intervals to achieve the best between two extremes: using only labeled data and relying solely on predicted unlabeled data: (1) the intervals are valid, as they contain the true value of the estimand of interest, and (2) they are more efficient, with narrower widths achieved by incorporating information from the larger sample size of unlabeled data.

Despite its advantages, the PPI framework has notable limitations. When the provided predictions are inaccurate, the constructed intervals can perform worse than the "classical intervals" derived solely from labeled data. To improve the statistical efficiency of PPI, Angelopoulos et al. (2024) developed Efficient Prediction-Powered Inference (PPI++), which incorporates a weighting parameter, λ, to minimize the asymptotic variance of the prediction-powered estimator. Additionally, Zrnic and Candès (2024) proposed Cross-Prediction-Powered Inference (Cross-PPI), an extension of PPI that ensures validity by splitting the labeled data to train the predictive model.

Our study aims to evaluate the performance of the PPI, PPI++, Cross-PPI, and the combination of PPI++ and Cross-PPI (Cross-PPI++) methods through a comprehensive simulation study. Specifically, we will examine their statistical properties, including the validity and efficiency of the constructed confidence intervals, under various simulated scenarios. Initially, we will evaluate these methods in settings where the labeled and unlabeled data are drawn from the same distribution. Furthermore, we will expand our focus to include the common challenge in EHR data analysis known as covariate shift, where the distribution of labeled data differs from that of unlabeled data. To address this, we will incorporate the density ratio into the PPI methods.

In the simulation study, we will focus on the simplest estimand, the mean outcome, and evaluate the performance of these methods across various settings, including different ratios of labeled to unlabeled data, varying levels of predictive model accuracy, diverse feature distributions, and degrees of covariate shift. The simulation will be replicated 100 times, and performance metrics will include the coverage probability of confidence intervals, the mean width of intervals, and the root mean squared error of the point estimates. By systematically exploring these scenarios, we aim to identify the conditions under which each method performs optimally or encounters limitations, offering practical guidance for their application in real-world EHR datasets.

2 Methods

2.1 PPI

Below are the the fundamental notations used in the PPI framework:

- Let $(X, Y) \in (\mathcal{X} \times \mathcal{Y})^n$ denote the labeled or the gold-standard dataset, where $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$.
- Similarly, let $(\tilde{X}, \tilde{Y}) \in (\mathcal{X} \times \mathcal{Y})^N$ denote the unlabeled dataset, where the outcomes $\tilde{Y}$ are not observed.

- In this section, we assume that (X, Y) and $(\tilde{X}, \tilde{Y})$ are independently and identically distributed (i.i.d.) samples from a common distribution, $p(x, y)$.
- We have a prediction rule, $f : X \to Y$ that is independent of the observed data. Thus, $f(X_i)$ denote the predictions for the labeled data and $f(\tilde{X}_i)$ denote the predictions for the unlabeled data.

The goal of PPI is to construct a confidence interval $\mathcal{C}^{PP}$ for the estimand θ. We will focus on the mean outcome as the estimand in our study. Specifically, we are interested in the mean outcome in the unlabeled data, $E[\tilde{Y}]$. The key conceptual innovation of PPI lies in the introduction of measure of fit m_θ and rectifier Δ_θ:

- Measure of fit m_θ is computed on the predictions for the unlabeled data $(\tilde{X}, f(\tilde{X}))$ and quantifies how close the estimate of θ it to its true value.
- Rectifier Δ_θ is defined as the difference of the measure of fit m_θ computed using the labeled data and its predictions $(X, Y, f(X))$. If the predictions are perfect, the rectifier is equal to zero.

The PPI method computes a confidence interval as $\mathcal{C}_\alpha^{PP} = \{\theta \text{ such that} |m_\theta + \Delta_\theta| \leq t_\theta(\alpha)\}$, where $t_\theta(\alpha)$ is a constant depending on the error level α.

When the estimand of interest is the mean outcome in unlabeled data, the algorithm for constructing the confidence interval $\mathcal{C}_\alpha^{PP}$ at error level $\alpha \in (0, 1)$ is as follows:

1. Point estimate of θ by PPI: $\hat{\theta}^{PP} \leftarrow \tilde{\theta}^f - \hat{\Delta} := \frac{1}{N}\sum_{i=1}^{N} f(\tilde{X}_i) - \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)$
2. Empirical variance of estimate based on unlabeled data: $\hat{\sigma}_f^2 \leftarrow \frac{1}{N}\sum_{i=1}^{N}(f(\tilde{X}_i) - \tilde{\theta}^f)^2$
3. Empirical variance of the rectifier: $\hat{\sigma}_\Delta^2 \leftarrow \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i - \hat{\Delta})^2$
4. Normal approximation: $t_\alpha \leftarrow z_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}_\Delta^2}{n} + \frac{\hat{\sigma}_f^2}{N}}$
5. Construct prediction-powered confidence set: $\mathcal{C}_\alpha^{PP} = \left(\hat{\theta}^{PP} \pm t_\alpha\right)$

2.2 PPI++

$$\hat{\theta}^{PP} = \frac{1}{N}\sum_{i=1}^{N} f(\tilde{X}_i) + \frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n}\sum_{i=1}^{n} f(X_i) = \frac{1}{n}\sum_{i=1}^{n} Y_i + \left(\frac{1}{N}\sum_{i=1}^{N} f(\tilde{X}_i) - \frac{1}{n}\sum_{i=1}^{n} f(X_i)\right)$$

Adding the parameter λ to minimize the asymptotic variance of the prediction-powered estimator, we have:

$$\hat{\theta} \text{ by Efficient PPI under covariate shift} = \frac{1}{n}\sum_{i=1}^{n} w(X_i)Y_i + \lambda\left[\frac{1}{N}\sum_{i=1}^{N} f(\tilde{X}_i) - \frac{1}{n}\sum_{i=1}^{n} w(X_i)f(X_i)\right]$$

2.3 Cross-PPI

2.4 Cross-PPI++

3 Extensions under Covariate Shift

Although the foundational PPI framework assumes that the labeled and unlabeled data are drawn from the same distribution, this assumption does not always hold in practice. One common violation is

covariate shift, where the distribution of input features, $p(X)$, differs between the labeled and unlabeled dataset, while the conditional distribution of the outcome given the feature $p(Y|X)$ remains the same (Shimodaira, 2000). To address this mismatch, the density ratio $w(x) = \tilde{p}(x)/p(x)$ is introduced, where $\tilde{p}(x)$ is the distribution of the features in the unlabeled data and $p(x)$ is the distribution of the features in the gold-standard data (Sugiyama et al., 2007). Incorporating the density ratio allows for adjusting model predictions to account for the covariate shift, ensuring valid and unbiased inference in scenarios where the covariate shift occurs.

Denote $w(x) = p^*(x)/\tilde{p}(x)$ the density ratio.

$$\hat{\theta} \text{ by standard PPI under covariate shift} = \frac{1}{n}\sum_{i=1}^{n} w(X_i)Y_i + \frac{1}{N}\sum_{i=1}^{N} f(\tilde{X}_i) - \frac{1}{n}\sum_{i=1}^{n} w(X_i)f(X_i)$$

Adding the parameter λ to minimize the asymptotic variance of the prediction-powered estimator, we have:

$$\hat{\theta} \text{ by Efficient PPI under covariate shift} = \frac{1}{n}\sum_{i=1}^{n} w(X_i)Y_i + \lambda\left[\frac{1}{N}\sum_{i=1}^{N} f(\tilde{X}_i) - \frac{1}{n}\sum_{i=1}^{n} w(X_i)f(X_i)\right]$$

We can find the parameter λ at:

$$\begin{aligned}
\hat{\lambda} &= argmin_{\lambda} E\left\{\frac{1}{n}\sum_{i=1}^{n} w(X_i)Y_i - \theta^* + \lambda\left[\frac{1}{N}\sum_{i=1}^{N} f(\tilde{X}_i) - \frac{1}{n}\sum_{i=1}^{n} w(X_i)f(X_i)\right]\right\}^2 \\
&= \frac{Cov\left(\frac{1}{n}\sum_{i=1}^{n} w(X_i)Y_i, \frac{1}{N}\sum_{i=1}^{N} f(\tilde{X}_i) - \frac{1}{n}\sum_{i=1}^{n} w(X_i)f(X_i)\right)}{Var\left(\frac{1}{N}\sum_{i=1}^{N} f(\tilde{X}_i) - \frac{1}{n}\sum_{i=1}^{n} w(X_i)f(X_i)\right)} \\
&= \frac{\frac{1}{n}Cov\left(w(X)Y, w(X)f(X)\right)}{\frac{1}{N}Var\left(f(\tilde{X})\right) + \frac{1}{n}Var\left(w(X)f(X)\right)}
\end{aligned}$$

3.1 Cross Efficient PPI under Covariate Shift

Use the trained models to impute predictions and compute the cross-prediction estimator, defined as:

$$\begin{aligned}
\hat{\theta} \text{ by Cross PPI} &= \frac{1}{KN}\sum_{j=1}^{K}\sum_{i=1}^{N} f^{(j)}(\tilde{X}_i) + \frac{1}{n}\sum_{j=1}^{K}\sum_{i\in I_j}(Y_i - f^{(j)}(X_i)) \\
&= \frac{1}{n}\sum_{j=1}^{K}\sum_{i\in I_j} Y_i + \frac{1}{KN}\sum_{j=1}^{K}\sum_{i=1}^{N} f^{(j)}(\tilde{X}_i) - \frac{1}{n}\sum_{j=1}^{K}\sum_{i\in I_j} f^{(j)}(X_i)
\end{aligned}$$

Consider the covariate shift situation and adding density ratio

$$\begin{aligned}
&\hat{\theta} \text{ by Cross PPI under covariate shift} \\
=&\frac{1}{n}\sum_{j=1}^{K}\sum_{i\in I_j} w(X_i)Y_i + \frac{1}{KN}\sum_{j=1}^{K}\sum_{i=1}^{N} f^{(j)}(\tilde{X}_i) - \frac{1}{n}\sum_{j=1}^{K}\sum_{i\in I_j} w(X_i)f^{(j)}(X_i)
\end{aligned}$$

Adding the parameter λ to minimize the asymptotic variance of the prediction-powered estimator, we have:

$$
\begin{aligned}
&\hat{\theta} \text{ by Cross Efficient PPI under covariate shift} \\
=&\frac{1}{n}\sum_{j=1}^{K}\sum_{i\in I_j} w(X_i)Y_i + \lambda\left[\frac{1}{KN}\sum_{j=1}^{K}\sum_{i=1}^{N} f^{(j)}(\tilde{X}_i) - \frac{1}{n}\sum_{j=1}^{K}\sum_{i\in I_j} w(X_i)f^{(j)}(X_i)\right] \\
&\hat{\lambda} \text{ with k-fold cross-validation} \\
=&\frac{Cov\left(\frac{1}{n}\sum_{i=1}^{n} w(X_i)Y_i, \frac{1}{KN}\sum_{j=1}^{K}\sum_{i=1}^{N} f^{(j)}(\tilde{X}_i) - \frac{1}{n}\sum_{j=1}^{K}\sum_{i\in I_j} w(X_i)f^{(j)}(X_i)\right)}{Var\left(\frac{1}{KN}\sum_{j=1}^{K}\sum_{i=1}^{N} f^{(j)}(\tilde{X}_i) - \frac{1}{n}\sum_{j=1}^{K}\sum_{i\in I_j} w(X_i)f^{(j)}(X_i)\right)} \\
=&\frac{\frac{1}{n}Cov\left(w(X)Y, w(X)f^{(j)}(X)\right)}{\frac{1}{K^2N}\sum_{j=1}^{K} Var\left(f^{(j)}(\tilde{X})\right) + \frac{1}{n}Var\left(w(X)f^{(\cdot)}(X)\right)}
\end{aligned}
$$

4 Simulation Study

4.1 Set up

- $N = 10,000$ as the sample size in the unlabeled data
- $n = 100, 200, 500, 1000, 2000, 5000$ as the sample size in the labeled data
- Data generating process: $Y = 1 + 2X_1 + 3X_2 + \epsilon$, where $\epsilon \sim N(0, \sigma_e^2), \sigma_e^2 = 0.01, 1, 4, 25$, ϵ is the noise, which is the difference between the true value and the predicted value $\hat{Y} = 1 + 2X_1 + 3X_2$
- Distribution of labeled X: continuous variable $X_1 \sim N(\mu_0, \sigma_0^2)$, where $\mu_0 = 0$ and $\sigma_0 = 1$, and binary variable $X_2 \sim Bernoulli(p_0)$, where $p_0 = 0.5$. X_1 is independent from X_2.

We consider three scenarios of covariate shift:

1. $X_1 \sim N(\mu_1, \sigma_0^2)$, where $\mu_1 = 0.2, 1, 2$, and $X_2 \sim Bernoulli(p_0)$. In this case, the density ratio is:
$$w(x) = \frac{p^*(x)}{\tilde{p}(x)} = \exp\left\{\frac{1}{2\sigma_0^2}(2x_1 - \mu_1 - \mu_0)(\mu_1 - \mu_0)\right\}$$
2. $X_1 \sim N(\mu_0, \sigma_1^2)$, where $\sigma_1 = 0.1, 0.5, 2$, and $X_2 \sim Bernoulli(p_0)$. In this case, the density ratio is:
$$w(x) = \frac{p^*(x)}{\tilde{p}(x)} = \frac{\sigma_0}{\sigma_1}\exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right)(x_1 - \mu_0)^2\right\}$$
3. $X_1 \sim N(\mu_0, \sigma_0^2)$, and $X_2 \sim Bernoulli(p_1)$, where $p_1 = 0.55, 0.7, 0.85$. In this case, the density ratio is:
$$w(x) = \frac{p^*(x)}{\tilde{p}(x)} = \left(\frac{p_1}{p_0}\right)^{x_2}\left(\frac{1 - p_1}{1 - p_0}\right)^{1 - x_2}$$