Assumption-lean and data-adaptive post-prediction inference

Jiacheng Miao[*]   JIACHENG.MIAO@WISC.EDU
Department of Biostatistics and Medical Informatics
University of Wisconsin-Madison
Madison, WI 53726, USA

Xinran Miao[*]   XINRAN.MIAO@WISC.EDU
Department of Statistics
University of Wisconsin-Madison
Madison, WI 53706, USA

Yixuan Wu   WU638@WISC.EDU
University of Wisconsin-Madison
Madison, WI 53726, USA

Jiwei Zhao   JIWEI.ZHAO@WISC.EDU
Department of Biostatistics and Medical Informatics, Department of Statistics
University of Wisconsin-Madison
Madison, WI 53726, USA

Qiongshi Lu   QLU@BIOSTAT.WISC.EDU
Department of Biostatistics and Medical Informatics
University of Wisconsin-Madison
Madison, WI 53726, USA

Editor: Edo Airoldi

Abstract

A primary challenge facing modern scientific research is the limited availability of gold-standard data, which can be costly, labor-intensive, or invasive to obtain. With the rapid development of machine learning (ML), scientists can now employ ML algorithms to predict gold-standard outcomes using variables that are easier to obtain. However, these predicted outcomes are often used directly in subsequent statistical analyses, ignoring imprecision and heterogeneity introduced by the prediction procedure. This will likely result in false positive findings and invalid scientific conclusions. In this work, we introduce PoSt-Prediction Adaptive inference (PSPA) that allows valid and powerful inference based on ML-predicted data. Its "assumption-lean" property guarantees reliable statistical inference without assumptions on the ML prediction. Its "data-adaptive" feature guarantees an efficiency gain over existing methods, regardless of the accuracy of ML prediction. We demonstrate the statistical superiority and broad applicability of our method through simulations and real-data applications.

Keywords: Inference with predicted data; Missing data; Semiparametric efficiency

---

1. [*]These authors contribute equally to this work.

1. Introduction

Gene expression data obtained from various tissues and cell types can provide crucial insights into the coordinated biological mechanisms that underlie disease etiology and characterize homeostasis (Lonsdale et al., 2013). However, some important tissues are often difficult to collect, which leads to a limited size of gene expression samples in those tissues. For example, the Genotype-Tissue Expression (GTEx) project is a comprehensive study focusing on gene expression regulation in many human tissues (GTEx Consortium et al., 2015). The percentage of individuals with missing gene expression ranges from 5% to 90% (median 47%) across all tissues in GTEx (GTEx Consortium, 2020). This limits the scientific understanding of transcription regulation across tissue contexts.

The difficulty in obtaining gold-standard data certainly extends beyond human gene expression applications (Wang et al., 2023). While gold-standard data with high reliability are essential to the validity of scientific discoveries, obtaining them is often costly and labor-intensive. Fortunately, the advent and rapid development of machine learning (ML) have enabled the prediction of outcomes using more accessible variables (LeCun et al., 2015; He et al., 2016), showing great potential in reducing the need to acquire gold-standard data.

Despite these benefits, replacing gold-standard data with ML predictions poses new challenges, particularly in maintaining the validity of downstream statistical analyses. This issue is exemplified by the statistical analysis using imputed gene expression in the GTEx project. To address the issue of insufficient sample sizes for rare tissues, several approaches have been proposed to impute gene expression in these tissues using expression data from tissues that are easier to acquire (Wang et al., 2016; Viñas et al., 2023; Basu et al., 2021). However, imputed gene expression is often treated as if it were observed, and used directly in subsequent statistical analyses for scientific discovery (Viñas et al., 2023), such as exploring sex differences and the genetic basis of gene regulation. As demonstrated in Wang et al. (2020); Angelopoulos et al. (2023a), using ML predictions directly in statistical inference will most likely produce false positive findings, leading to invalid scientific conclusions.

The challenge of making valid inferences from ML predictions extends beyond the specific example discussed, affecting numerous scientific disciplines where ML is applied (Bullock et al., 2020). Recent work has introduced a method called prediction-powered inference (`PPI`), which leverages a small set of labeled data with gold-standard outcomes and a large amount of unlabeled data with ML predictions to perform valid statistical inference (Angelopoulos et al., 2023a). Although validity is guaranteed, `PPI` may be less efficient than inference relying solely on labeled data. For instance, when applied to GTEx data, `PPI` identifies fewer sex-biased gene expression compared to the classical approach based on labeled data alone, highlighting a loss of statistical efficiency (Figure 1). This limitation reduces the broader applicability of `PPI` and similar methods.

To address this problem, we introduce a valid, powerful, and widely applicable approach named `PoSt-Prediction Adaptive inference` (PSPA). PSPA is designed to integrate ML prediction with the observed gold-standard data to ensure valid and efficient statistical inference. We highlight two key features of our `PSPA` method:

- <u>Assumption-Lean</u>: The `PSPA` estimator is consistent and asymptotically normal, ensuring valid inference with no assumptions on the ML model or prediction accuracy. This means that `PSPA` can be used with arbitrarily misspecified ML predictions.

- <u>Data-Adaptive</u>: The `PSPA` estimator is adaptive to the accuracy of the ML prediction. It utilizes more information from a better ML prediction to reduce variance and to avoid variance inflation. Below, we demonstrate that `PSPA` achieves element-wise variance reduction compared to existing methods, regardless of the ML prediction quality.

Our contributions are threefold: (i) We present a simple and data-adaptive method for statistical inference problems with predicted data defined through estimating equations, without any assumptions on the ML prediction or parameterization of the true data-generating process. (ii) We establish the consistent and asymptotic normality of our proposed estimator. We further demonstrate the optimality of our estimator in terms of asymptotic variance over a class of estimators that includes existing methods in post-prediction inference. (iii) In contrast to current methods, which only focus on incorporating ML-predicted labels in statistical inference, our approach extends to utilizing both ML-predicted covariates and labels, enhancing the versatility and applicability of post-prediction inference.

The rest of this paper is organized as follows. In Section 2, we formulate our problem and illustrate our method with mean estimation to build an intuition. Next, in Section 3, we introduce a general protocol for applying `PSPA` to estimands defined through estimating equations. Further, we establish the asymptotic properties of the `PSPA` estimator and describe the estimation procedure. We build the connection with the semiparametric efficiency theory in Section 4 and describe extensions in Section 5. Through extensive simulations (Section 6) and real data analysis (Section 7), we validate our theoretical claims and demonstrate the practical utility and effectiveness of the `PSPA` estimator. We then compare our method with a recent method `PPI++` (Angelopoulos et al., 2023b) in Section 8. This paper concludes with a discussion in Section 9.
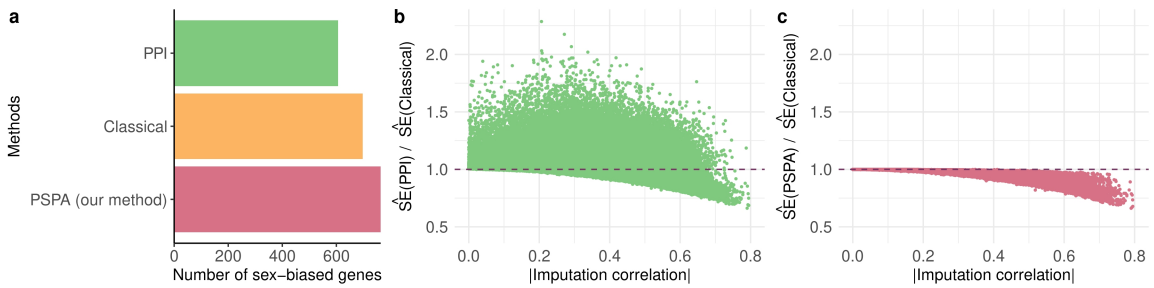

Figure 1: **Comparison of `PPI`, `PSPA`, and classical method in identifying genes with sex-biased expression using GTEx data.** (a) the number of sex-biased genes identified by each of the four approaches. (b) x-axis: absolute value of imputation correlation. y-axis: relative ratio of estimated standard error between `PPI` and classical method. (c) same as b but for our method `PSPA`. The dashed lines represent $y = 1$ in (b)-(c).

2. Problem Formulation

Let Y be a scalar outcome and $\mathbf{X}$ be d-dimensional covariates. The scientific interest is to estimate a q-dimensional parameter $\boldsymbol{\theta}$ defined through an estimating equation,

$$\mathbb{E}\{\boldsymbol{\psi}(Y, \mathbf{X}; \boldsymbol{\theta})\} = \mathbf{0}, \tag{1}$$

where $\boldsymbol{\psi}(\cdot, \cdot; \boldsymbol{\theta})$ is a user-defined function. Such a definition of $\boldsymbol{\theta}$ is very general, including outcome mean, outcome quantile, least squares coefficients, or any other specific quantities of interest involving both $\mathbf{x}$ and y, including the unique minimizer of a loss function and the maximizer of a criterion function (Van der Vaart, 2000).

We observe a random sample where only a small subset is labeled with outcome Y. In addition, we observe the auxiliary variable $\mathbf{Z}$ that is predictive of the outcome Y. Note that $\mathbf{Z}$ could be a subset of $\mathbf{X}$ and vice versa; see Remark 1. Denote this sample as $\mathcal{L} \cup \mathcal{U}$ where $\mathcal{L} = \{(y_i, \mathbf{x}_i, \mathbf{z}_i), i = 1, \cdots, n\}$ and $\mathcal{U} = \{(\mathbf{x}_i, \mathbf{z}_i), i = n+1, \cdots, N+n\}$. Assume that $\mathcal{L}$ and $\mathcal{U}$ are independent and share the same marginal distributions of $(\mathbf{X}, \mathbf{Z})$. Let $n/N \to \rho$ as $n \to \infty$ and $N \to \infty$, and let $\rho = \infty$ if there is no unlabeled data.

We consider the availability of an external and independent prediction algorithm $\widehat{f}(\cdot)$ on $\mathbf{z}$ that produces predictions $\widehat{f}$ for the outcome variable Y. We assume that the operating characteristics of $\widehat{f}$ are unknown to the user, and the data used to fit $\widehat{f}$ are unavailable. $\widehat{f}$ is considered a black box function and can be incorrect in predicting y.

With access to this black-box $\widehat{f}$, our primary goal is to improve efficiency while making inference on the parameter $\boldsymbol{\theta}$. We refer to our problem as post-prediction inference, where inference occurs after predictions are made independently (Wang et al., 2020; Angelopoulos et al., 2023a). Specifically, we make no assumptions on $\widehat{f}$. The resulting protocol can benefit researchers where training a prediction algorithm is unrealistic, and offers guidelines for scientists lacking computational resources or predictive modeling expertise.

Example 1 (Sex-differentiated gene expression) *We present the scientific problem that motivates our setup: identifying sex-biased gene expression in the brain cortex tissue using data from the GTEx project. For this analysis, we have labeled data denoted by $\mathcal{L} = \{(y_i, \widehat{f}_i, \mathbf{x}_i, \mathbf{z}_i), i = 1, \cdots, 205\}$, where y_i represents the gene expression in brain cortex tissue, $\mathbf{x}_i$ includes a binary indicator for biological sex in $\{0, 1\}$, along with other technical factors such as surrogate variables for batch effects, age, RNA integrity number, and total ischemic time. $\mathbf{z}_i$ is the whole-transcriptome gene expression profile in whole blood tissue that is easier to access, and $\widehat{f}$ is the imputed gene expression in the brain cortex tissue using $\mathbf{z}_i$. In addition, our unlabeled data, $\mathcal{U} = \{(\widehat{f}_i, \mathbf{x}_i, \mathbf{z}_i), i = 206, \cdots, 670\}$, contains 465 samples whose gene expression was measured in whole blood but not in brain cortex. Our interest lies in the linear regression coefficient $\boldsymbol{\theta}$ as a solution to $\mathbb{E}[\mathbf{x}(y - \mathbf{x}^\top \boldsymbol{\theta})] = 0$, focusing particularly on the coefficients corresponding to biological sex.*

Remark 1 (The consideration of X and Z) *We split the covariate into two (possibly overlapping) parts, $\mathbf{Z}$ and $\mathbf{X}$, where the former is predictive of the outcome Y and the latter appears in the estimating equation (1). They could be the same, but a wide range of applications suggests they are different. When $\mathbf{Z}$ is not a subset of $\mathbf{X}$, the information it carries about the outcome Y may be the most informative for the estimation problem (1).*

Hence, a better prediction algorithm $\widehat{f}$ does not necessarily correspond to a better estimate of $\boldsymbol{\theta}$. As justified above, this work does not attempt to improve the prediction algorithm, but to empower the inference on the parameter of interest, given any arbitrary black-box $\widehat{f}$.

In the following, Section 2.1 presents the key idea of the proposed method using mean estimation as an example. Section 2.2 reviews related work. In Section 3, we introduce the proposed framework for the estimation problem defined in equation (1).

2.1 Example: estimating the outcome mean

We illustrate the intuition of our approach using mean estimation. Consider $\psi(y, \mathbf{x}; \boldsymbol{\theta}) = y - \theta$. The resulting estimand is $\theta = \mathbb{E}(Y)$, the outcome mean. Then the classical approach calculates an estimator from the sample average of observed outcomes,

$$\widehat{\theta}_{\mathrm{C}} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

This estimator is unbiased, but can have high variance due to the small sample size of labeled data. To increase efficiency with the ML-predicted outcomes in unlabeled data, we propose to introduce another unbiased estimator of zero with a weighting scalar ω,

$$\widehat{\theta}_{\mathrm{PSPA}}(\omega) = \frac{1}{n} \sum_{i=1}^{n} y_i + \omega \left[\frac{1}{N} \sum_{i=n+1}^{n+N} \widehat{f} - \frac{1}{n} \sum_{i=1}^{n} \widehat{f} \right]$$

We refer to this estimator as the `PSPA` estimator. Both the classical and `PPI` estimator (Angelopoulos et al., 2023a) can be viewed as special cases of the `PSPA`. The classical estimator is the `PSPA` estimator with $\omega = 0$. The `PPI` estimator is the `PSPA` estimator with $\omega = 1$:

$$\widehat{\theta}_{\mathrm{PPI}} = \frac{1}{n} \sum_{i=1}^{n} y_i + \left[\frac{1}{N} \sum_{i=n+1}^{n+N} \widehat{f} - \frac{1}{n} \sum_{i=1}^{n} \widehat{f} \right].$$

As detailed in Section 4, another candidate estimator is guided by the efficient influence function (EIF). We refer to this estimator as EIF*-based where * represents the fact that the nuisance function is plugged in as an estimate,

$$\widehat{\theta}_{\mathrm{EIF}^*} = \frac{1}{n} \sum_{i=1}^{n} y_i + \frac{N}{N+n} \left[\frac{1}{N} \sum_{i=n+1}^{n+N} \widehat{f} - \frac{1}{n} \sum_{i=1}^{n} \widehat{f} \right].$$

This is a special case of `PSPA` estimator with $\omega = \frac{N}{N+n}$, achieving the semiparametric efficiency bound when $\widehat{f}$ approximates its true counterpart at a sufficiently fast rate.

Since these estimators are summations of consistent estimators of the population mean and zero, they are consistent for the population mean. Additionally, Wald-type confidence intervals can be constructed using a consistent estimator for the asymptotic variance.

$$\underbrace{\frac{\mathrm{Var}(Y)}{n} + \underbrace{\left(\omega^2 \frac{\mathrm{Var}[\widehat{f}]}{n} - 2\omega \frac{\mathrm{Cov}[Y, \widehat{f}]}{n} + \omega^2 \frac{\mathrm{Var}[\widehat{f}]}{N} \right)}_{\text{Additional terms}}}_{\text{Variance of \texttt{PSPA} estimator}}, \tag{2}$$

To guarantee that the asymptotic variance (2) is no greater than the classical approach, regardless of the quality of $\widehat{f}$, the additional terms need to be no larger than zero. A key insight is that these additional terms are a quadratic function of ω:

$$q(\omega) = \omega^2 \left\{ \frac{\mathrm{Var}[\widehat{f}]}{n} + \frac{\mathrm{Var}[\widehat{f}]}{N} \right\} - 2\omega \frac{\mathrm{Cov}[Y, \widehat{f}]}{n},$$

which achieves its minimum at $\omega^{\mathrm{opt}} = \mathrm{Cov}[Y, \widehat{f}]/\{\mathrm{Var}[\widehat{f}] + n\,\mathrm{Var}[\widehat{f}]/N\}$ with minima

$$q(\omega^{\mathrm{opt}}) = \frac{-\mathrm{Cov}[Y, \widehat{f}]}{n\,\mathrm{Var}[\widehat{f}] + n^2\,\mathrm{Var}[\widehat{f}]/N} \le 0,$$

where the equality holds if and only if $\mathrm{Cov}[Y, \widehat{f}] = 0$. In this case, our PSPA estimator is at least as efficient as the classical estimator. In contrast, both PPI and EIF*-based estimators can be less efficient than the classical estimator, when the ML predictions are not accurate. As we shall see in Section 3, similar patterns appear in estimating the general parameter defined in (1). In addition, the additional terms are below zero if and only if $\mathrm{Cov}[Y, \widehat{f}] > \frac{\omega(N+n)\,\mathrm{Var}[\widehat{f}]}{2N}$. This indicates that in order for PPI to achieve efficiency improvement (where $\omega = 1$), the prediction accuracy of the ML algorithm cannot be low.

Remark 2 (Data-Adaptive Feature) *We provide an intuition for the "data-adaptive" feature of* PSPA. *When the ML algorithm is purely random,* $\mathrm{Cov}[Y, \widehat{f}] = 0$ *and hence the optimal weight* $\omega^{\mathrm{opt}} = 0$. *Therefore,* PSPA *degenerates to the classical estimator that uses only labeled data. This aligns with the intuition that such predictions should not contribute to inference. On the other hand, if the ML prediction is perfect,* $\mathrm{Cov}[Y, \widehat{f}] = \mathrm{Var}[\widehat{f}] = \mathrm{Var}[Y]$ *and hence the optimal weight is* $\omega^{\mathrm{opt}} = N/(N + n)$. *In this case,* PSPA *degenerates to the EIF*-based estimator that achieves the semiparametric efficiency bound. This is also intuitive since perfect ML prediction can be treated as gold-standard data and the estimator should utilize all the (measured and predicted) information and thus is efficient.*

2.2 Related work

Our setting is closely related to a recent method called prediction-powered inference (PPI) (Angelopoulos et al., 2023a), which ensures valid inference with gold-standard labels and arbitrary ML predictions. It corresponds to a special case of our PSPA estimator where $\omega = 1$ (or a vector of all ones). Both theoretical and numerical comparisons suggest that the PPI estimator is less efficient than the proposed PSPA estimator. Our approach builds upon and generalizes the method proposed by PPI++ (Angelopoulos et al., 2023b), called PPI++, which originally enhanced the efficiency of PPI. However, unlike PSPA, PPI++ cannot guarantee element-wise variance reduction. Moreover, PPI++ cannot be applied to statistical inference with ML-predicted covariates. On the other hand, PSPA is capable of handling this type of application, while guaranteeing improved efficiency, element-wise. We present a comprehensive comparison of PSPA and PPI++ in Section 8.

Our proposal of boosting a consistent estimator with a consistent estimator of zero has been a long-standing idea. One famous estimator is the augmented inverse probability weighting estimator in the literature of missing data (Robins et al., 1994) and causal

inference (Robins, 2000), in which a consistent inverse probability weighting estimator is augmented by a weighted residual term with mean zero, resulting in an efficiency gain (Hahn, 1998). In fact, it reaches the semiparametric efficiency bound (Bickel et al., 1993; Hahn, 1998) when nuisance functions are correctly specified or approximated sufficiently fast. Although the pursuit of efficiency under correct specifications differs from our main focus, we establish the efficiency bound of our estimation problem (1) in Section 4 and illustrate the connection between the PSPA estimator and the efficient influence function.

The idea of using unlabeled data fits into the broader context of improving efficiency with auxiliary data, which has been widely applied in survey estimation (Breidt and Opsomer, 2017), missing data (Robins et al., 1994), measurement error models (Chen et al., 2005), causal inference with surrogate outcomes (Kallus and Mao, 2020), and semi-supervised learning (Wang and Shen, 2007; Chakrabortty and Cai, 2018). The literature on semi-supervised learning has adopted a similar data structure to our problem in which labeled data are accompanied by unlabeled data (possibly with surrogate outcomes). Recently, Chakrabortty and Cai (2018) has considered efficiency improvement in linear models under possible model misspecifications. After that, estimation methods have been proposed for mean estimation (Zhang et al., 2019), best linear predictor estimation (Azriel et al., 2022), and general M-estimation problems (Song et al., 2023). Our work is distinct from this line of research in that we consider post-prediction inference with any "black-box" ML predictions for efficiency improvement.

3. PSPA for assumption-lean and data-adaptive post-prediction inference

In this section, we introduce our method PSPA for assumption-lean and data-adaptive post-prediction inference. For the estimand defined by (1) and any ML prediction, it guarantees the validity of inference results and element-wise variance reduction compared with the classical method that uses labeled data alone. We present our estimator and algorithm in Section 3.1 and establish the theoretical guarantees for our method in Section 3.2. Examples of applying our method for regression tasks are provided in Section 3.3.

3.1 PSPA estimator

Bearing in mind that $\widehat{f}$ could be incorrect, one typical approach to estimating $\boldsymbol{\theta}$ is to ignore the unlabeled data and use the labeled data only. This classical estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{C}}$ solves for

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}(y_i, \mathbf{x}_i; \boldsymbol{\theta}) = \mathbf{0}.$$

It is always consistent, but may have low efficiency because it ignores the unlabeled data.

In this work, we aim to provide an estimator that is consistent and more efficient than the classical estimator, regardless of the quality of the $\widehat{f}$. Similar to the mean estimation example, we add an augmentation term, which has a mean of zero and is indexed by a vector $\boldsymbol{\omega} = [\omega_1, \cdots, \omega_q]^{\top}$, to the estimation equation for the classical approach:

$$\mathrm{diag}(\boldsymbol{\omega})\left\{-\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}(\widehat{f}, \mathbf{x}_i; \boldsymbol{\theta}) + \frac{1}{N}\sum_{i=n+1}^{n+N}\boldsymbol{\psi}(\widehat{f}, \mathbf{x}_i; \boldsymbol{\theta})\right\},$$

where $\mathrm{diag}(\boldsymbol{\omega})$ is a diagonal matrix with diagonal elements $\omega_1, \cdots, \omega_q$. We refer to the vector $\boldsymbol{\omega}$ as a weighting vector, as it controls the contribution of labeled outcomes and the arbitrary prediction $\widehat{f}$ in a data-adaptive way. Together, we propose our PSPA estimator $\widehat{\boldsymbol{\theta}}_{\mathtt{PSPA}}(\boldsymbol{\omega})$ that solves the equation,

$$\Psi^{\boldsymbol{\omega}}_{\mathtt{PSPA}}(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\psi}(y_i, \mathbf{x}_i; \boldsymbol{\theta}) + \mathrm{diag}(\boldsymbol{\omega}) \left\{ -\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\psi}(\widehat{f}, \mathbf{x}_i; \boldsymbol{\theta}) + \frac{1}{N} \sum_{i=n+1}^{n+N} \boldsymbol{\psi}(\widehat{f}, \mathbf{x}_i; \boldsymbol{\theta}) \right\} = \mathbf{0},$$

(3)

for a given weighting vector $\boldsymbol{\omega}$. This equation corresponds to a class of estimators $\widehat{\boldsymbol{\theta}}_{\mathtt{PSPA}}(\boldsymbol{\omega})$ for different fixed values of $\boldsymbol{\omega}$. We next establish the consistency and asymptotic normality of these estimators with the following regularity conditions. Condition (C1) is a regularity assumption on the parameter of interest. Conditions (C2)-(C4) guarantee consistency. Condition (C5) is needed for asymptotic normality. All conditions are reasonable and standard (Van der Vaart, 2000).

(C1) The parameter space $\boldsymbol{\Theta}$ is a compact subset of $\mathbb{R}^q$, containing the true parameter $\boldsymbol{\theta}_0$ as a unique solution to (1).

(C2) The function in (3) is continuous and the equation has a unique solution.

(C3) $\sup_{\boldsymbol{\theta} \in \Theta} \mathbb{E}[\|\boldsymbol{\psi}(Y, \mathbf{X}; \boldsymbol{\theta})\|^2] < \infty$ and $\sup_{\boldsymbol{\theta} \in \Theta} \mathbb{E}[\|\mathrm{diag}(\boldsymbol{\omega}) \cdot \boldsymbol{\psi}(\widehat{f}(\mathbf{Z}), \mathbf{X}; \boldsymbol{\theta})\|^2] < \infty$ where $\| \cdot \|$ is the L_2-norm that takes over the true data-generating process.

(C4) Labeled and unlabeled data are i.i.d drawn from the population of interest.

(C5) There exists a function $\boldsymbol{\varphi}$ such that $\mathbb{E}[\sup_{\boldsymbol{\theta} \in \Theta} \boldsymbol{\varphi}^2] < \infty$ and $\mathbb{E}\|\boldsymbol{\psi}(Y, \mathbf{X}; \boldsymbol{\theta}') - \boldsymbol{\psi}(Y, \mathbf{X}; \boldsymbol{\theta}'')\|_2^2 \leq \boldsymbol{\varphi}\|\boldsymbol{\theta}' - \boldsymbol{\theta}''\|^2$ for any $\boldsymbol{\theta}'$ and $\boldsymbol{\theta}''$ in a neighborhood of $\boldsymbol{\theta}$. Furthermore, $\mathbb{E}\{\boldsymbol{\psi}(y, \mathbf{x}; \boldsymbol{\theta})\}$ is differentiable at $\boldsymbol{\theta}$ with a nonsingular derivative matrix.

Theorem 3 *Under Conditions (C1)-(C4), assuming $\frac{n}{N} \to \rho$ as $n \to \infty$ and $N \to \infty$, then the proposed estimator $\widehat{\boldsymbol{\theta}}_{\mathtt{PSPA}}(\boldsymbol{\omega})$ converges to $\boldsymbol{\theta}$ in probability. Assuming additionally Condition (C5), we have*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\mathtt{PSPA}}(\boldsymbol{\omega}) - \boldsymbol{\theta}) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\omega})),$$

where $\boldsymbol{\Sigma}(\boldsymbol{\omega}) = \mathbf{A}^{-1}\mathbf{V}(\boldsymbol{\omega})\mathbf{A}^{-1}$, $\mathbf{A} = \mathbb{E}[\partial \boldsymbol{\psi}(Y, \mathbf{X}; \boldsymbol{\theta})/\partial \boldsymbol{\theta}]$, $\mathbf{V}(\boldsymbol{\omega}) = \mathbf{M}_1 + \mathrm{diag}(\boldsymbol{\omega})(\mathbf{M}_2 + \rho\mathbf{M}_3)\mathrm{diag}(\boldsymbol{\omega}) - 2\mathrm{diag}(\boldsymbol{\omega})\mathbf{M}_4$, $\mathbf{M}_1 = \mathrm{Var}[\boldsymbol{\psi}(Y, \mathbf{X}; \boldsymbol{\theta})]$, $\mathbf{M}_2 = \mathrm{Var}[\boldsymbol{\psi}(\widehat{f}(\mathbf{Z}), \mathbf{X}; \boldsymbol{\theta})]$, $\mathbf{M}_3 = \mathrm{Var}[\boldsymbol{\psi}(\widehat{f}(\mathbf{Z}), \mathbf{X}; \boldsymbol{\theta}_0)]$, $\mathbf{M}_4 = \mathrm{Cov}[\boldsymbol{\psi}(Y, \mathbf{X}; \boldsymbol{\theta}), \boldsymbol{\psi}(\widehat{f}(\mathbf{Z}), \mathbf{X}; \boldsymbol{\theta})]$.

The proof is contained in Appendix B.1. Theorem 3 suggests that the proposed estimator $\widehat{\boldsymbol{\theta}}_{\mathtt{PSPA}}(\boldsymbol{\omega})$ is asymptotically normal with a fixed weighting vector $\boldsymbol{\omega}$. In practice, this weighting vector $\boldsymbol{\omega}$ should be estimated from data with the pursuit of efficiency. We provide an estimation procedure aimed at element-wise variance reduction. This is made possible by a key observation that the j-th diagonal element of the asymptotic covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\omega})$ is a quadratic function of ω_j and does not depend on other components of $\boldsymbol{\omega}$:

$$\boldsymbol{\Sigma}_{jj}(\boldsymbol{\omega}) = \omega_j^2 [\mathbf{A}^{-1}(\mathbf{M}_2 + \rho\mathbf{M}_3)\mathbf{A}^{-1}]_{jj} - 2\omega_j [\mathbf{A}^{-1}\mathbf{M}_4\mathbf{A}^{-1}]_{jj} + [\mathbf{A}^{-1}\mathbf{M}_1\mathbf{A}^{-1}]_{jj}$$

where $[\mathbf{M}]_{jj}$ represents the j-th diagonal element of matrix $\mathbf{M}$. We define the optimal weighting vector $\boldsymbol{\omega}^{\mathrm{opt}} = [\omega_1^{\mathrm{opt}}, \ldots, \omega_q^{\mathrm{opt}}]^\top$ such that each coordinate of the weighting vector minimizes the asymptotic variance for the corresponding coordinate:

$$\omega_j^{\mathrm{opt}} = \arg\min_{\omega_j} \mathbf{\Sigma}_{jj}(\omega_j) = \frac{[\mathbf{A}^{-1}\mathbf{M}_4\mathbf{A}^{-1}]_{jj}}{[\mathbf{A}^{-1}(\mathbf{M}_2 + \rho\mathbf{M}_3)\mathbf{A}^{-1}]_{jj}} \text{ for all } j \in 1, \ldots, q,$$

The resulting asymptotic variance for j-th coordinate is

$$\mathbf{\Sigma}_{jj}(\boldsymbol{\omega}^{\mathrm{opt}}) = [\mathbf{A}^{-1}\mathbf{M}_1\mathbf{A}^{-1}]_{jj} - \frac{[\mathbf{A}^{-1}\mathbf{M}_4\mathbf{A}^{-1}]_{jj}}{[\mathbf{A}^{-1}(\mathbf{M}_2 + \rho\mathbf{M}_3)\mathbf{A}^{-1}]_{jj}}.$$

Such construction of $\boldsymbol{\omega}^{\mathrm{opt}}$ guarantees the optimality of the element-wise asymptotic variance across the class of estimators with any weighting vector $\boldsymbol{\omega}$. Next, we present our algorithm for PSPA estimation and inference.

Algorithm 1 PSPA estimation with ML-predicted labels

Input: Data $\mathcal{L} \cup \mathcal{U}$, pre-trained ML model $\widehat{f}$, error rate $\alpha \in (0,1)$.

1: Obtain the classical estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{C}}$ by solving $\frac{1}{n}\sum_{i=1}^n \boldsymbol{\psi}(y_i, \mathbf{x}_i; \boldsymbol{\theta}) = \mathbf{0}$.

2: Obtain the optimal weighting vector $\widehat{\boldsymbol{\omega}}^{\mathrm{opt}} = [\widehat{\omega}_1^{\mathrm{opt}}, \ldots, \widehat{\omega}_q^{\mathrm{opt}}]^\top$ by

$$\widehat{\omega}_j^{\mathrm{opt}} = \min\left(\frac{[\widehat{\mathbf{A}}_{\mathrm{C}}^{-1}\widehat{\mathbf{M}}_{4,\mathrm{C}}\widehat{\mathbf{A}}_{\mathrm{C}}^{-1}]_{jj}}{[\widehat{\mathbf{A}}_{\mathrm{C}}^{-1}(\widehat{\mathbf{M}}_{2,\mathrm{C}} + \rho\widehat{\mathbf{M}}_{3,\mathrm{C}})\widehat{\mathbf{A}}_{\mathrm{C}}^{-1}]_{jj}}, 1\right) \text{ for all } j \in 1, \ldots, q,$$

where $\widehat{\mathbf{A}}_{\mathrm{C}}$, $\widehat{\mathbf{M}}_{1,\mathrm{C}}$, $\widehat{\mathbf{M}}_{2,\mathrm{C}}$, $\widehat{\mathbf{M}}_{3,\mathrm{C}}$, $\widehat{\mathbf{M}}_{4,\mathrm{C}}$ are sample analogs of $\mathbf{A}$, $\mathbf{M}_1$, $\mathbf{M}_2$, $\mathbf{M}_3$, $\mathbf{M}_4$ with $\widehat{\boldsymbol{\theta}}_{\mathrm{C}}$ plugged in.

3: Obtain the PSPA estimator with a one-step update:

$$\widehat{\boldsymbol{\theta}}_{\mathtt{PSPA}} = \widehat{\boldsymbol{\theta}}_{\mathrm{C}} - \left[\nabla\Psi_{\mathtt{PSPA}}^{\widehat{\boldsymbol{\omega}}^{\mathrm{opt}}}(\widehat{\boldsymbol{\theta}}_{\mathrm{C}})\right]^{-1}\Psi_{\mathtt{PSPA}}^{\widehat{\boldsymbol{\omega}}^{\mathrm{opt}}}(\widehat{\boldsymbol{\theta}}_{\mathrm{C}}), \text{ where } \nabla\Psi_{\mathtt{PSPA}}^{\widehat{\boldsymbol{\omega}}^{\mathrm{opt}}} = \partial\Psi_{\mathtt{PSPA}}^{\widehat{\boldsymbol{\omega}}^{\mathrm{opt}}}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\mathrm{C}}}$$

4: Obtain the asymptotic variance of $\widehat{\boldsymbol{\theta}}_{\mathtt{PSPA}}$:

$$\widehat{\mathbf{\Sigma}}(\widehat{\boldsymbol{\omega}}^{\mathrm{opt}}) = \widehat{\mathbf{A}}_{\mathtt{PSPA}}^{-1}[\widehat{\mathbf{M}}_{1,\mathtt{PSPA}} + \mathrm{diag}(\widehat{\boldsymbol{\omega}}^{\mathrm{opt}})(\widehat{\mathbf{M}}_{2,\mathtt{PSPA}} + \rho\widehat{\mathbf{M}}_{3,\mathtt{PSPA}})\mathrm{diag}(\boldsymbol{\omega}) - 2\mathrm{diag}(\widehat{\boldsymbol{\omega}}^{\mathrm{opt}})\widehat{\mathbf{M}}_{4,\mathtt{PSPA}}]\widehat{\mathbf{A}}_{\mathtt{PSPA}}^{-1},$$

where $\widehat{\mathbf{A}}_{\mathtt{PSPA}}$, $\widehat{\mathbf{M}}_{1,\mathtt{PSPA}}$, $\widehat{\mathbf{M}}_{2,\mathtt{PSPA}}$, $\widehat{\mathbf{M}}_{3,\mathtt{PSPA}}$, $\widehat{\mathbf{M}}_{4,\mathtt{PSPA}}$ are sample analogs of $\mathbf{A}$, $\mathbf{M}_1$, $\mathbf{M}_2$, $\mathbf{M}_3$, $\mathbf{M}_4$ with $\widehat{\boldsymbol{\theta}}_{\mathtt{PSPA}}$ plugged in.

Output: PSPA estimator $\widehat{\boldsymbol{\theta}}_{\mathtt{PSPA}}$, standard error $\sqrt{\frac{\widehat{\mathbf{\Sigma}}(\widehat{\boldsymbol{\omega}}^{\mathrm{opt}})_{jj}}{n}}$, α-level confidence interval $\mathcal{C}_{\alpha,j}^{\mathtt{PSPA}} = \left(\widehat{\boldsymbol{\theta}}_{\mathtt{PSPA}_j} \pm z_{1-\alpha/2}\sqrt{\frac{\widehat{\mathbf{\Sigma}}(\widehat{\boldsymbol{\omega}}^{\mathrm{opt}})_{jj}}{n}}\right)$, and (two-sided) p-value $2(1 - \Phi(|\frac{\widehat{\boldsymbol{\theta}}_{\mathtt{PSPA}_j}}{\sqrt{\frac{\widehat{\mathbf{\Sigma}}(\widehat{\boldsymbol{\omega}}^{\mathrm{opt}})_{jj}}{n}}}|))$ for the j-th coordinate. Here, Φ is the CDF of the standard normal distribution.

3.2 Theoretical guarantees

Next, we establish the theoretical guarantees for our proposed estimator and algorithm. We modify the theorem 3 to reflect that $\widehat{\boldsymbol{\omega}}$ in the algorithm is estimated from the data.

Corollary 4 *Suppose* $\widehat{\boldsymbol{\omega}} \overset{P}{\to} \boldsymbol{\omega}$ *and conditions for Theorem 3 hold, then* $\widehat{\boldsymbol{\theta}}_{\texttt{PSPA}}(\boldsymbol{\omega}) \overset{P}{\to} \boldsymbol{\theta}$ *and*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\texttt{PSPA}}(\boldsymbol{\omega}) - \boldsymbol{\theta}) \overset{D}{\to} N(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\omega})).$$

Proof of Corollary 4 is contained in Appendix B.2. In our algorithm, we substitute the sample analogs for $\mathbf{A}$, $\mathbf{M}_1$, $\mathbf{M}_2$, $\mathbf{M}_3$, $\mathbf{M}_4$ into ω_j^{opt}. Given that these sample analogs are typically consistent estimators, Slutsky's theorem implies that $\widehat{\boldsymbol{\omega}}^{\text{opt}} \overset{P}{\to} \boldsymbol{\omega}^{\text{opt}}$. Therefore, this additional condition is satisfied by our algorithm. The asymptotic normality in Corollary 4 guarantees the validity for inference for our algorithm by the following Corollary 5.

Corollary 5 *Suppose* $\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\omega}}) \overset{P}{\to} \boldsymbol{\Sigma}(\boldsymbol{\omega})$ *and conditions for Theorem 3 hold. Given confidence level* $1 - \alpha \in (0, 1)$,

$$\lim_{n \to \infty} \mathbb{P}(\theta_j \in \mathcal{C}_{\alpha,j}^{\widehat{\boldsymbol{\omega}}}) \geq 1 - \alpha, \ j = 1 \cdots, q,$$

where θ_j *is the j-th coordinate of the parameter and* $\mathcal{C}_{\alpha,j}^{\texttt{PSPA}} = \left(\widehat{\boldsymbol{\theta}}_{\texttt{PSPA}_j}(\widehat{\boldsymbol{\omega}}) \pm z_{1-\alpha/2} \sqrt{\frac{\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\omega}})_{jj}}{n}} \right).$

Proof of Corollary 5 follows from the asymptotic normality of the $\texttt{PSPA}$ estimator in Corollary 4. Since we plug in the consistent estimator for each element in $\boldsymbol{\Sigma}(\boldsymbol{\omega}^{\text{opt}})$, Slutsky's theorem implies that $\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\omega}}^{\text{opt}}) \overset{P}{\to} \boldsymbol{\Sigma}(\boldsymbol{\omega}^{\text{opt}})$, and therefore the condition in Corollary 5 is also satisfied by our algorithm.

The above two corollaries ensure the validity of the inference for our algorithm. We next demonstrate that our method achieves element-wise variance reduction when compared to all baseline methods.

Proposition 6 *Suppose* $\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\texttt{PSPA}}(\boldsymbol{\omega}) - \boldsymbol{\theta}) \overset{D}{\to} N(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\omega}))$, *given* $\boldsymbol{\omega}^{\text{opt}}$ *defined by equation 3.1,* $\boldsymbol{\Sigma}(\boldsymbol{\omega}^{\text{opt}})_{jj} \leq \boldsymbol{\Sigma}(\boldsymbol{\omega})_{jj}$ *for all* $j \in \{1, \cdots, q\}$.

Proof of Proposition 6 is contained in Appendix B.3. By setting the weighting vector $\boldsymbol{\omega}$ to $\boldsymbol{\omega}^{\text{C}} = [0, \ldots, 0]^{\top}$ or $\boldsymbol{\omega}^{\texttt{PPI}} = [1, \ldots, 1]^{\top}$, our method reduces to classical method and $\texttt{PPI}$, respectively. Therefore, $\texttt{PSPA}$ guarantees an element-wise reduction in the asymptotic variance compared to both the classical method and $\texttt{PPI}$, for arbitrary ML models.

3.3 Examples: linear and logistic regression

Here, we present two examples to illustrate the intuitions behind the $\texttt{PSPA}$ estimation procedure. We begin with the estimation of coefficients in a linear regression model, where the resulting estimator is expressed in a closed form.

Example 2 (Linear Regression) *We consider the ordinary least squares problem, where* $\boldsymbol{\psi}(y, \mathbf{x}; \boldsymbol{\theta}) = \mathbf{x}(y - \mathbf{x}^{\top}\boldsymbol{\theta})$. *Therefore, the estimand is* $\boldsymbol{\theta}_0 = \mathbb{E}[\mathbf{X}\mathbf{X}^{\top}]^{-1}\mathbb{E}[\mathbf{X}Y]$. *From equation (3), the* $\texttt{PSPA}$ *estimator is the solution to the equation*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i(y_i - \mathbf{x}_i^{\top}\boldsymbol{\theta}) + \frac{1}{N}\sum_{i=n+1}^{n+N} \text{diag}(\boldsymbol{\omega})\mathbf{x}_i\{\widehat{f} - \mathbf{x}_i^{\top}\boldsymbol{\theta}\} - \frac{1}{n}\sum_{i=1}^{n} \text{diag}(\boldsymbol{\omega})\mathbf{x}_i\{\widehat{f} - \mathbf{x}_i^{\top}\boldsymbol{\theta}\} = \mathbf{0},$$

yielding a closed-form solution for `PSPA` *linear regression*

$$\widehat{\boldsymbol{\theta}}_{\text{PSPA}} = \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{\top} + \text{diag}(\widehat{\boldsymbol{\omega}}^{\text{opt}}) \left[\frac{1}{N} \sum_{i=n+1}^{n+N} \mathbf{x}_i \mathbf{x}_i^{\top} - \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{\top} \right] \right\}^{-1}$$

$$\left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i y_i + \text{diag}(\widehat{\boldsymbol{\omega}}^{\text{opt}}) \left[\frac{1}{N} \sum_{i=n+1}^{n+N} \mathbf{x}_i \widehat{f}(\mathbf{z}_i) - \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \widehat{f} \right] \right\},$$

where $\widehat{\boldsymbol{\omega}}^{\text{opt}}$ *is a consistent estimator of optimal* $\boldsymbol{\omega}^{\text{opt}}$, *obtainable through Algorithm 1.*

The second example is logistic regression, which lacks a closed-form solution but can be solved using Algorithm 1.

Example 3 (Logistic Regression) *For logistic regression,* $\boldsymbol{\psi}(y, \mathbf{x}; \boldsymbol{\theta}) = -\mathbf{x}y + \mathbf{x}\gamma_{\boldsymbol{\theta}}(\mathbf{x}),$ *where* $\gamma_{\boldsymbol{\theta}}(\mathbf{x}) = 1 / \left\{ 1 + \exp\left(-\mathbf{x}^{\top} \boldsymbol{\theta} \right) \right\}.$ *Using estimation equation (3), the* `PSPA` *estimator* $\widehat{\boldsymbol{\theta}}_{\text{PSPA}}$ *is the solution to the equation*

$$\frac{1}{n} \sum_{i=1}^{n} \left\{ -\mathbf{x}_i y_i + \mathbf{x}_i \gamma_{\boldsymbol{\theta}}(\mathbf{x}_i) \right\} + \text{diag}(\boldsymbol{\omega}) \left\{ \frac{1}{N} \sum_{i=n+1}^{n+N} [-\mathbf{x}_i \widehat{f}(\mathbf{z}_i) + \mathbf{x}_i \gamma_{\boldsymbol{\theta}}(\mathbf{x}_i)] - \frac{1}{n} \sum_{i=1}^{n} [-\mathbf{x}_i \widehat{f} + \mathbf{x}_i \gamma_{\boldsymbol{\theta}}(\mathbf{x}_i)] \right\} = \mathbf{0}$$

4. Relationship with Efficient Influence Function

To connect our method with the literature on semiparametric efficiency (Bickel et al., 1993; Tsiatis, 2006), in this section, we state the efficient influence function (EIF) and the semiparametric efficiency bound for the estimation problem (1) (Proposition 7) and connect it with our proposed `PSPA` estimator. To ease notation, we introduce a random variable R to indicate the labeling; $r_1 = \cdots = r_n = 1$ and $r_{n+1} = \cdots = r_{n+N} = 1$. In this section, suppose $\pi \equiv \text{pr}(R = 1) = n/(n + N)$ is a fixed constant between zero and one. Note that the introduction of R is for notational simplicity, and we make no inference on it.

Proposition 7 (Efficient Influence Function) *The EIF for estimating* $\boldsymbol{\theta}$ *is*

$$\boldsymbol{\phi}(y_i, \mathbf{x}_i, \boldsymbol{\theta}) = \frac{r}{\pi} \mathbf{A}^{-1} [\boldsymbol{\psi}(y, \mathbf{x}; \boldsymbol{\theta}) - \mathbb{E}\{\boldsymbol{\psi}(Y, \mathbf{x}; \boldsymbol{\theta}) \mid \mathbf{x}, \mathbf{z}\}] + \mathbf{A}^{-1} \mathbb{E}\{\boldsymbol{\psi}(Y, \mathbf{x}; \boldsymbol{\theta}) \mid \mathbf{x}, \mathbf{z}\}.$$

The proof is contained in Appendix B.4. Guided by Proposition 7, an EIF-based estimator would be the solution to the empirical equation,

$$\mathbf{0} = \frac{1}{n + N} \sum_{i=1}^{n+N} \boldsymbol{\phi}(y_i, \mathbf{x}_i, \widehat{\boldsymbol{\theta}}_{\text{EIF}}). \tag{4}$$

With the derived EIF, the EIF-based estimator is asymptotically normal and achieves the semiparametric efficiency bound by section 8.4 of Molenberghs et al. (2014).

11

Corollary 8 (Semiparametric Efficiency Bound) *Under Conditions (C1) and (C3), assume* $\mathbb{E}\{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \phi(Y,\mathbf{X},\boldsymbol{\theta})\} < \infty$ *and equation (4) has a unique solution, then as* $n \to \infty$ *and* $N \to \infty$,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\mathrm{EIF}} - \boldsymbol{\theta}_0) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathbb{E}\{\boldsymbol{\phi}(Y,\mathbf{X},\boldsymbol{\theta}_0)\boldsymbol{\phi}(Y,\mathbf{X},\boldsymbol{\theta}_0)^{\top}\pi\}).$$

If the nuisance function $\mathbb{E}\{\boldsymbol{\psi}(Y,\mathbf{x};\boldsymbol{\theta}) \mid \mathbf{x},\mathbf{z}\}$ that appears in the EIF can be correctly specified or well approximated, then an EIF-based estimator will achieve the semiparametric efficiency bound. However, this is nearly impossible in reality with limited knowledge of the data generation process or a small computational budget. Instead, it's typical to construct an estimator with nuisance functions plugged in as their estimates. We refer to such estimators as "EIF*-based", because the influence function of the resulting estimator, denoted by $\widehat{\boldsymbol{\theta}}_{\mathrm{EIF}^*}$, may or may not be the efficient influence function.

In our setting, the nuisance function $\mathbb{E}\{\boldsymbol{\psi}(Y,\mathbf{x};\boldsymbol{\theta}) \mid \mathbf{x},\mathbf{z}\}$ can be estimated with $\mathbb{E}\{\boldsymbol{\psi}(\widehat{f},\mathbf{x};\boldsymbol{\theta}) \mid \mathbf{x},\mathbf{z}\}$. Upon simple calculation, we can write the resulting $\widehat{\boldsymbol{\theta}}_{\mathrm{EIF}^*}$ as the solution to

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{A}^{-1}\boldsymbol{\psi}(y_i,\mathbf{x}_i;\boldsymbol{\theta}) + (1-\pi)\mathbf{A}^{-1}\left\{-\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}(\widehat{f},\mathbf{x}_i;\boldsymbol{\theta}) + \frac{1}{N}\sum_{i=n+1}^{n+N}\boldsymbol{\psi}(\widehat{f},\mathbf{x}_i;\boldsymbol{\theta})\right\} = \mathbf{0}. \quad (5)$$

Similar to the equation (3) that the PSPA estimator solves, the EIF*-based estimator (5) solves an equation with two parts: one term that can produce a consistent estimator with labeled data only, and the other term that utilizes predictions $\widehat{f}$, which always have a mean of zero. If one multiples both sides of (3) with matrix $\mathbf{A}^{-1}$, then (5) corresponds to that equation weight equals to a vector of $1 - \pi$. In estimating the outcome mean (Section 2.1) where $\mathbf{A}$ is essentially the scalar one, the EIF*-based estimator is the PSPA estimator with weight being $1 - \pi = N/(n + N)$. The difference between (3) and (5) clarifies one of the main distinctions between the proposed PSPA estimator and an EIF-based estimator. The former seeks variance reduction in a data-adaptive way, while the latter attains minimum variance at correct specifications. In real-world situations, the proposed PSPA estimator can be much more efficient in that it utilizes information in a data-adaptive fashion. Numeric comparisons are provided in Section 6.

5. Extensions

Next, we extend PSPA to address more general scenarios. Specifically, Section 5.1 outlines an inference procedure for fully labeled outcomes with partially labeled covariate $\mathbf{x}$, and Section 5.2 discusses cases where both y and $\mathbf{x}$ are partially labeled.

5.1 ML-predicted covariates

We adjusted the previously described setup by modifying it to use ML to predict the covariate $\mathbf{X}$ instead of the outcome Y. Our data can be divided into two parts: $\mathcal{L} \cup \mathcal{U}$. The $\mathcal{L}$ part includes data points $(y_i, \mathbf{x}_i, \mathbf{z}_i)$ for $i = 1, \cdots, n$ and the $\mathcal{U}$ part includes data points $(y_i, \mathbf{z}_i)$ for $i = n+1, \cdots, N+n$. In addition, we use an external prediction algorithm with the notation $\widehat{q}(\cdot)$. This algorithm is applied to $\mathbf{z}$ to generate the predicted value $\widehat{q}(\mathbf{z})$ of the covariate $\mathbf{X}$. However, it is important to note that $\widehat{q}$ may produce inaccurate or biased predictions of $\mathbf{X}$. The estimand is also defined by the estimating equation (1). Similar to

our proposal to handle the ML-predicted outcome, we propose our PSPA estimator $\widehat{\boldsymbol{\theta}}_{\text{PSPA}'}(\boldsymbol{\omega})$ with ML-predicted covariates that solves the equation

$$\Psi^{\boldsymbol{\omega}}_{\text{PSPA}*}(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}(y_i, \mathbf{x}_i; \boldsymbol{\theta}) + \text{diag}(\boldsymbol{\omega})\left\{-\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}(y_i, \widehat{q}; \boldsymbol{\theta}) + \frac{1}{N}\sum_{i=n+1}^{n+N}\boldsymbol{\psi}(y_i, \widehat{q}; \boldsymbol{\theta})\right\} = \mathbf{0},$$

Denote $\mathbf{A} = \mathbb{E}[\partial\boldsymbol{\psi}(Y, \mathbf{X}; \boldsymbol{\theta})/\partial\boldsymbol{\theta}]$, $\mathbf{M}_1 = \text{Var}[\boldsymbol{\psi}(Y, \mathbf{X}; \boldsymbol{\theta})]$, $\mathbf{M}'_2 = \text{Var}[\boldsymbol{\psi}(Y, \widehat{q}; \boldsymbol{\theta}_0)]$, $\mathbf{M}'_3 = \text{Var}[\boldsymbol{\psi}(Y, \widehat{q}; \boldsymbol{\theta}_0)]$, $\mathbf{M}'_4 = \text{Cov}[\boldsymbol{\psi}(Y, \mathbf{X}; \boldsymbol{\theta}_0), \boldsymbol{\psi}(Y, \widehat{q}; \boldsymbol{\theta}_0)]$. Upon calculations similar to Theorem 3, the j-th diagonal element of the asymptotic variance of $\widehat{\boldsymbol{\theta}}_{\text{PSPA}'}(\boldsymbol{\omega})$ is

$$\boldsymbol{\Sigma}'_{jj}(\boldsymbol{\omega}) = \omega_j^2[\mathbf{A}^{-1}(\mathbf{M}'_2 + \rho\mathbf{M}'_3)\mathbf{A}^{-1}]_{jj} - 2\omega_j[\mathbf{A}^{-1}\mathbf{M}'_4\mathbf{A}^{-1}]_{jj} + [\mathbf{A}^{-1}\mathbf{M}'_1\mathbf{A}^{-1}]_{jj},$$

which achieve its minimum with $\omega_j^{\text{opt}} = \frac{[\mathbf{A}^{-1}\mathbf{M}'_4\mathbf{A}^{-1}]_{jj}}{[\mathbf{A}^{-1}(\mathbf{M}'_2+\rho\mathbf{M}'_3)\mathbf{A}^{-1}]_{jj}}$. We modify our Algorithm 1 to the following algorithm to incorporate ML-predicted covariates.

Algorithm 2 PSPA estimation with ML-predicted covariates

Input: Data $\mathcal{L} \cup \mathcal{U}$, pre-trained ML model $\widehat{f}$, error rate $\alpha \in (0, 1)$.

1: Obtain the classical estimator $\widehat{\boldsymbol{\theta}}_{\text{C}}$ by solving $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}(y_i, \mathbf{x}_i; \boldsymbol{\theta}) = \mathbf{0}$.

2: Obtain the optimal weighting vector $\widehat{\boldsymbol{\omega}}^{\text{opt}} = [\widehat{\omega}_1^{\text{opt}}, \ldots, \widehat{\omega}_q^{\text{opt}}]^{\top}$ by

$$\widehat{\omega}_j^{\text{opt}} = \begin{cases} \min\left(\frac{[\widehat{\mathbf{A}}_{\text{C}}^{-1}\widehat{\mathbf{M}}'_{4,\text{C}}\widehat{\mathbf{A}}_{\text{C}}^{-1}]_{jj}}{[\widehat{\mathbf{A}}_{\text{C}}^{-1}(\widehat{\mathbf{M}}'_{2,\text{C}}+\rho\widehat{\mathbf{M}}'_{3,\text{C}})\widehat{\mathbf{A}}_{\text{C}}^{-1}]_{jj}}, \frac{\lambda_{\min,+}[\frac{1}{n}\nabla\boldsymbol{\psi}(y_i,\mathbf{x}_i;\boldsymbol{\theta})]}{\lambda_{\max}[\frac{1}{n}\nabla\boldsymbol{\psi}(y_i,\widehat{q};\boldsymbol{\theta})]}\right) & \text{if } \frac{[\widehat{\mathbf{A}}_{\text{C}}^{-1}\widehat{\mathbf{M}}'_{4,\text{C}}\widehat{\mathbf{A}}^{-1}]_{jj}}{[\widehat{\mathbf{A}}_{\text{C}}^{-1}(\widehat{\mathbf{M}}'_{2,\text{C}}+\rho\widehat{\mathbf{M}}'_{3,\text{C}})\widehat{\mathbf{A}}_{\text{C}}^{-1}]_{jj}} > 0 \\ \max\left(\frac{[\widehat{\mathbf{A}}_{\text{C}}^{-1}\widehat{\mathbf{M}}'_{4,\text{C}}\widehat{\mathbf{A}}_{\text{C}}^{-1}]_{jj}}{[\widehat{\mathbf{A}}_{\text{C}}^{-1}(\widehat{\mathbf{M}}'_{2,\text{C}}+\rho\widehat{\mathbf{M}}'_{3,\text{C}})\widehat{\mathbf{A}}_{\text{C}}^{-1}]_{jj}}, 0\right) & \text{if } \frac{[\widehat{\mathbf{A}}_{\text{C}}^{-1}\widehat{\mathbf{M}}'_{4,\text{C}}\widehat{\mathbf{A}}^{-1}]_{jj}}{[\widehat{\mathbf{A}}_{\text{C}}^{-1}(\widehat{\mathbf{M}}'_{2,\text{C}}+\rho\widehat{\mathbf{M}}'_{3,\text{C}})\widehat{\mathbf{A}}_{\text{C}}^{-1}]_{jj}} \leq 0 \end{cases}$$

for all $j \in 1, \ldots, q$. Here, $\widehat{\mathbf{A}}_{\text{C}}, \widehat{\mathbf{M}}'_{1,\text{C}}, \widehat{\mathbf{M}}'_{2,\text{C}}, \widehat{\mathbf{M}}'_{3,\text{C}}, \widehat{\mathbf{M}}'_{4,\text{C}}$ are sample analogs of $\mathbf{A}, \mathbf{M}'_1, \mathbf{M}'_2, \mathbf{M}'_3, \mathbf{M}'_4$ for $\Psi^{\boldsymbol{\omega}}_{\text{PSPA}*}(\boldsymbol{\theta})$ with $\widehat{\boldsymbol{\theta}}_{\text{C}}$ plugged in. $\lambda_{\min,+}[\mathbf{M}]$ and $\lambda_{\max}[\mathbf{M}]$ represent the smallest non-negative and largest eigenvalue of matrix $\mathbf{M}$, respectively.

3: Obtain the PSPA* estimator with a one-step update:

$$\widehat{\boldsymbol{\theta}}_{\text{PSPA}'} = \widehat{\boldsymbol{\theta}}_{\text{C}} - \left[\nabla\Psi^{\widehat{\boldsymbol{\omega}}^{\text{opt}}}_{\text{PSPA}'}(\widehat{\boldsymbol{\theta}}_{\text{C}})\right]^{-1}\Psi^{\widehat{\boldsymbol{\omega}}^{\text{opt}}}_{\text{PSPA}'}(\widehat{\boldsymbol{\theta}}_{\text{C}}), \text{ where } \nabla\Psi^{\widehat{\boldsymbol{\omega}}^{\text{opt}}}_{\text{PSPA}'} = \partial\Psi^{\widehat{\boldsymbol{\omega}}^{\text{opt}}}_{\text{PSPA}'}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_{\text{C}}}.$$

4: Obtain the asymptotic variance of $\widehat{\boldsymbol{\theta}}_{\text{PSPA}'}$:

$$\widehat{\boldsymbol{\Sigma}}'(\widehat{\boldsymbol{\omega}}^{\text{opt}}) = \widehat{\mathbf{A}}_{\text{PSPA}'}^{-1}[\widehat{\mathbf{M}}'_{1,\text{PSPA}} + \text{diag}(\widehat{\boldsymbol{\omega}}^{\text{opt}})(\widehat{\mathbf{M}}'_{2,\text{PSPA}} + \rho\widehat{\mathbf{M}}'_{3,\text{PSPA}})\text{diag}(\boldsymbol{\omega}) - 2\text{diag}(\widehat{\boldsymbol{\omega}}^{\text{opt}})\widehat{\mathbf{M}}'_{4,\text{PSPA}}]\widehat{\mathbf{A}}_{\text{PSPA}'}^{-1},$$

where $\widehat{\mathbf{A}}'_{\text{PSPA}}, \widehat{\mathbf{M}}'_{1,\text{PSPA}}, \widehat{\mathbf{M}}'_{2,\text{PSPA}}, \widehat{\mathbf{M}}'_{3,\text{PSPA}}, \widehat{\mathbf{M}}'_{4,\text{PSPA}}$ are sample analogs of $\mathbf{A}, \mathbf{M}'_1, \mathbf{M}'_2, \mathbf{M}'_3, \mathbf{M}'_4$ for $\Psi^{\boldsymbol{\omega}}_{\text{PSPA}'}(\boldsymbol{\theta})$ with $\widehat{\boldsymbol{\theta}}_{\text{PSPA}}$ plugged in.

Output: PSPA$'$ estimator $\widehat{\boldsymbol{\theta}}_{\text{PSPA}'}$, standard error $\sqrt{\frac{\widehat{\boldsymbol{\Sigma}}'(\widehat{\boldsymbol{\omega}}^{\text{opt}})_{jj}}{n}}$, α-level confidence interval $\mathcal{C}^{\text{PSPA}'}_{\alpha,j} = \left(\widehat{\boldsymbol{\theta}}_{\text{PSPA}'_j} \pm z_{1-\alpha/2}\sqrt{\frac{\widehat{\boldsymbol{\Sigma}}'(\widehat{\boldsymbol{\omega}}^{\text{opt}})_{jj}}{n}}\right)$, and (two-sided) p-value $2(1 - \Phi(|\frac{\widehat{\boldsymbol{\theta}}_{\text{PSPA}'_j}}{\sqrt{\frac{\widehat{\boldsymbol{\Sigma}}'(\widehat{\omega}^{\text{opt}})_{jj}}{n}}}|))$ for the j-th coordinate. Here, Φ is the CDF of the standard normal distribution.

Corollary 9 states the asymptotic distribution that enables interval estimation, with its proof following from the proof of Theorem 3, Corollary 4 and 5, and thus omitted.

Corollary 9 *Under Conditions (C1), (C2) on equation 5.1, and (C3)-(C4), assuming $\frac{n}{N} \to \rho$ as $n \to \infty$ and $N \to \infty$, and $\widehat{\boldsymbol{\omega}} \xrightarrow{P} \boldsymbol{\omega}$, then the proposed estimator $\widehat{\boldsymbol{\theta}}_{\mathtt{PSPA}}(\widehat{\boldsymbol{\omega}})$ converges to $\boldsymbol{\theta}_0$ in probability. Assuming additionally Condition (C5), we have that as $n \to \infty$ and $N \to \infty$,*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}'_{\mathtt{PSPA}}(\widehat{\boldsymbol{\omega}}) - \boldsymbol{\theta}_0) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}'(\boldsymbol{\omega})).$$

where $\boldsymbol{\Sigma}'(\boldsymbol{\omega}) = \mathbf{A}^{-1}\mathbf{V}'(\boldsymbol{\omega})\mathbf{A}^{-1}$, $\mathbf{A} = \mathbb{E}[\partial\boldsymbol{\psi}(Y, \mathbf{X}; \boldsymbol{\theta})/\partial\boldsymbol{\theta}]$, $\mathbf{V}'(\boldsymbol{\omega}) = \mathbf{M}_1 + \mathrm{diag}(\boldsymbol{\omega})(\mathbf{M}'_2 + \rho\mathbf{M}'_3)\mathrm{diag}(\boldsymbol{\omega}) - 2\mathrm{diag}(\boldsymbol{\omega})\mathbf{M}'_4$, $\mathbf{M}_1 = \mathrm{Var}[\boldsymbol{\psi}(Y, \mathbf{X}; \boldsymbol{\theta})]$, $\mathbf{M}'_2 = \mathrm{Var}[\boldsymbol{\psi}(Y, \widehat{q}; \boldsymbol{\theta}_0)]$, $\mathbf{M}'_3 = \mathrm{Var}[\boldsymbol{\psi}(Y, \widehat{q}; \boldsymbol{\theta}_0)]$, $\mathbf{M}'_4 = \mathrm{Cov}[\boldsymbol{\psi}(Y, \mathbf{X}; \boldsymbol{\theta}_0), \boldsymbol{\psi}(Y, \widehat{q}; \boldsymbol{\theta}_0)]$. With $\widehat{\boldsymbol{\Sigma}}'(\widehat{\boldsymbol{\omega}}) \xrightarrow{P} \boldsymbol{\Sigma}'(\boldsymbol{\omega})$,

$$\lim_{n \to \infty} \mathbb{P}(\theta_j \in \mathcal{C}'^{\widehat{\boldsymbol{\omega}}}_{\alpha, j}) \geq 1 - \alpha, \ j = 1, \cdots, q,$$

where $\mathcal{C}'^{\widehat{\boldsymbol{\omega}}}_{\alpha, j} = (\widehat{\boldsymbol{\theta}}'_{\mathtt{PSPA}_j}(\widehat{\boldsymbol{\omega}}) \pm z_{1-\alpha/2}\sqrt{\frac{\widehat{\boldsymbol{\Sigma}}'(\widehat{\boldsymbol{\omega}})_{jj}}{n}})$, $1 - \alpha \in (0, 1)$ is the confidence level, and θ_j is the j-th coordinate of the parameter.

Remark 10 *Condition (C2) requires equation 5.1 to have a unique solution. To ensure this, we impose a constraint on $\boldsymbol{\omega}$. Equation 5.1 has a unique solution if its derivative, $\frac{1}{n}\sum_{i=1}^{n}[\nabla\boldsymbol{\psi}(y_i, \mathbf{x}_i; \boldsymbol{\theta}) - \mathrm{diag}(\boldsymbol{\omega})\nabla\boldsymbol{\psi}(y_i, \widehat{q}; \boldsymbol{\theta})] + \frac{1}{N}\sum_{i=n+1}^{n+N}\nabla\boldsymbol{\psi}(y_i, \widehat{q}; \boldsymbol{\theta})$, is positive semi-definite. The second term, $\frac{1}{N}\sum_{i=n+1}^{n+N}\nabla\boldsymbol{\psi}(y_i, \widehat{q}; \boldsymbol{\theta})$, is already positive semi-definite, so we only need to ensure the first term is positive semi-definite. According to Lemma 13 in the appendix, a sufficient condition for this is: $0 \leq \omega_j \leq \dfrac{\lambda_{\min,+}[\frac{1}{n}\sum_{i=1}^{n}\nabla\boldsymbol{\psi}(y_i, \mathbf{x}_i; \boldsymbol{\theta})]}{\lambda_{\max}[\frac{1}{n}\sum_{i=1}^{n}\nabla\boldsymbol{\psi}(y_i, \widehat{q}; \boldsymbol{\theta})]}$ for all $j = 1, \cdots, q$, where $\lambda_{\min,+}[\mathbf{M}]$ and $\lambda_{\max}[\mathbf{M}]$ represent the smallest positive and largest eigenvalue of matrix $\mathbf{M}$, respectively. By imposing this constraint on the weighting vector in the algorithm, we ensure that condition (C2) holds. In practice, one can relax this constraint by substituting the prefixed $\boldsymbol{\omega}$ into the derivative and confirming its positive semi-definiteness. If necessary, one can iteratively reduce the value of each positive ω_j (or increase negative ω_j) until the derivative achieves positive semi-definiteness.*

Next, we state the theoretical guarantees of PSPA for element-wise variance reduction compared with the classical approach with weighting vector $\boldsymbol{\omega}^{\mathrm{C}} = [0, \ldots, 0]^\top$.

Corollary 11 *Suppose $\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\mathtt{PSPA}'}(\boldsymbol{\omega}^{\mathrm{C}}) - \boldsymbol{\theta}) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Sigma}(\mathbf{0}))$. Denote $\widehat{\boldsymbol{\omega}}^{\mathrm{opt}} = [\widehat{\omega}_1^{\mathrm{opt}}, \ldots, \widehat{\omega}_q^{\mathrm{opt}}]^\top$, where*

$$\widehat{\omega}_j^{\mathrm{opt}} = \begin{cases} \min\left(\dfrac{[\widehat{\mathbf{A}}^{-1}\widehat{\mathbf{M}}'_{4,C}\widehat{\mathbf{A}}^{-1}]_{jj}}{[\widehat{\mathbf{A}}^{-1}(\widehat{\mathbf{M}}'_{2,C}+\rho\widehat{\mathbf{M}}'_{3,C})\widehat{\mathbf{A}}^{-1}]_{jj}}, \dfrac{\lambda_{\min,+}[\frac{1}{n}\nabla\boldsymbol{\psi}(y_i,\mathbf{x}_i;\boldsymbol{\theta})]}{\lambda_{\max}[\frac{1}{n}\nabla\boldsymbol{\psi}(y_i,\widehat{q};\boldsymbol{\theta})]}\right) & \text{if } \dfrac{[\widehat{\mathbf{A}}^{-1}\widehat{\mathbf{M}}'_{4,C}\widehat{\mathbf{A}}^{-1}]_{jj}}{[\widehat{\mathbf{A}}^{-1}(\widehat{\mathbf{M}}'_{2,C}+\rho\widehat{\mathbf{M}}'_{3,C})\widehat{\mathbf{A}}^{-1}]_{jj}} > 0 \\[3mm] \max\left(\dfrac{[\widehat{\mathbf{A}}^{-1}\widehat{\mathbf{M}}'_{4,C}\widehat{\mathbf{A}}^{-1}]_{jj}}{[\widehat{\mathbf{A}}^{-1}(\widehat{\mathbf{M}}'_{2,C}+\rho\widehat{\mathbf{M}}'_{3,C})\widehat{\mathbf{A}}^{-1}]_{jj}}, 0\right) & \text{if } \dfrac{[\widehat{\mathbf{A}}^{-1}\widehat{\mathbf{M}}'_{4,C}\widehat{\mathbf{A}}^{-1}]_{jj}}{[\widehat{\mathbf{A}}^{-1}(\widehat{\mathbf{M}}'_{2,C}+\rho\widehat{\mathbf{M}}'_{3,C})\widehat{\mathbf{A}}^{-1}]_{jj}} \leq 0 \end{cases}$$

for all $j \in 1, \ldots, q$. Suppose $\widehat{\boldsymbol{\omega}}^{\mathrm{opt}} \xrightarrow{P} \boldsymbol{\omega}^{\mathrm{opt}}$, then $\boldsymbol{\Sigma}(\boldsymbol{\omega}^{\mathrm{opt}})_{jj} \leq \boldsymbol{\Sigma}(\boldsymbol{\omega}^{\mathrm{C}})_{jj}$.

The proof is contained in Appendix B.6.

14

5.2 ML-predicted outcome and covariates

Next, we explore a situation where the outcome and covariates are not directly observed in the unlabeled data. Instead, we employ machine learning techniques to predict both. Additionally, we possess much smaller data with both outcome and covariates measured. This full data can be categorized into two segments: $\mathcal{L} \cup \mathcal{U}$. The first segment, $\mathcal{L}$, includes the data points $(y_i, \mathbf{x}_i, \mathbf{z}_i)$ for $i = 1, \cdots, n$, while the second segment, $\mathcal{U}$, comprises the instances of $\mathbf{z}_i$ for $i = n+1, \cdots, N+n$, and $N >> n$. The variable $\mathbf{z}$ is to predict both the outcome and the covariates. The ML-predicted outcome and covariates are represented by $\widehat{f}$ and $\widehat{q}(\mathbf{z})$, respectively, where $\widehat{f}$ and $\widehat{q}$ are independently obtained. The `PSPA` estimator for this situation is to solve the equation,

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\psi}(y_i, \mathbf{x}_i; \boldsymbol{\theta}) + \mathrm{diag}(\boldsymbol{\omega})\left\{ -\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}(\widehat{f}, \widehat{q}; \boldsymbol{\theta}) + \frac{1}{N}\sum_{i=n+1}^{n+N}\boldsymbol{\psi}(\widehat{f}, \widehat{q}; \boldsymbol{\theta}) \right\} = \mathbf{0},$$

Replacing the estimating equation in Algorithm 2 with Equation 5.2 gives an algorithm for estimation and statistical inference for this task. The consistency and asymptotic normality of the resulting estimator can be established by the proof of Theorem 3.

6. Simulations

We conduct simulations to evaluate the performance of `PSPA` using linear and logistic regression across two settings: one with ML-predicted labels and another with ML-predicted covariates. In the ML-predicted labels setting, we considered classical, `PPI`, and EIF*-based methods. The EIF*-based method employs a weighting vector defined as $\boldsymbol{\omega}^{\mathrm{EIF}^*} = \left[\frac{N}{N+n}, \ldots, \frac{N}{N+n}\right]^{\top}$, where N and n represent the sample sizes of unlabeled and labeled data, respectively. For the ML-predicted covariates setting, we used classical and imputation-based methods as baselines. We present the implementation details in Section 6.1 and the simulation results in Section 6.2.

6.1 Implementation details

In all simulations, the ground truth coefficients are obtained by a Monte Carlo approximation with 5×10^4 samples. The labeled data is with 500 samples, and the unlabeled data is with 500, 1500, 2500, 5000, or 10000 samples in different settings. A pre-trained random forest with 100 trees to grow is obtained from a hold-out dataset with 1000 samples. All simulations are repeated 1000 times.

6.1.1 ML-PREDICTED LABELS

We simulated the labels Y_i and covariates $X_{1i}, \ldots, X_{50i}$ for the linear regression setting by $X_{1i}, \ldots, X_{50i} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1), Z_i \sim \mathcal{N}(0,1), \theta_1, \ldots, \theta_{10} = \frac{0.1}{\sqrt{10}}; , \theta_{11}, \ldots, \theta_{50} = 0, Y_i = \sum_{k=1}^{50} X_{ki}\theta_k + rZ_i + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, \tau_\epsilon)$ where τ_ϵ such that $\mathrm{Var}(Y_i) = 1$, where r is set to 0.8 for settings with different sample sizes of unlabeled data and 0, 0.2, 0.4, 0.6, or 0.8 for settings with different levels of imputation accuracy. We used the $X_{1i}, \ldots, X_{50i}$, and Z_i as inputs to train the random forest model, which aimed to predict Y_i in unlabeled data. For logistic regression,

15

we adopted the same data-generating process except that we generated the label $\tilde{Y}_i$ by $\tilde{Y}_i = \mathbb{1}(Y_i > \text{median}(Y_i))$. Our parameter of interest is the regression coefficient for X_{1i}.

6.1.2 ML-PREDICTED COVARIATES

We simulated the labels Y_i and covariates $X_{1i}, \ldots, X_{10i}$ for linear regression setting by $X_{1i}, \ldots, X_{10i} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0,1), Z_i \sim \mathcal{N}(0,1), \theta_1 = 0.1, \theta_2, \ldots, \theta_{10} = 0, Y_i = \sum_{k=1}^{10} X_{ki}\theta_k + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, \tau_\epsilon)$ where τ_ϵ is set to be the value such that $\text{Var}(Y_i) = 1, Z_i = 0.1Y_i + rX_{1i} + \delta_i$ where $\delta_i \sim \mathcal{N}(0, \tau_\delta)$ where τ_δ is set to be the value such that $\text{Var}(Z_i) = 1$ where r is set to 0.8 for settings with different samples size of unlabeled data and 0, 0.2, 0.4, 0.6, or 0.8 for settings with different levels of imputation accuracy. We employed the variables Z_i as inputs to train the random forest model, which aimed to predict X_{1i} in unlabeled data. For logistic regression, we used the same data-generating process except that we generated the label $\tilde{Y}_i$ by $\tilde{Y}_i = \mathbb{1}(Y_i > \text{median}(Y_i))$. We are interested in the regression coefficient for X_{1i}.
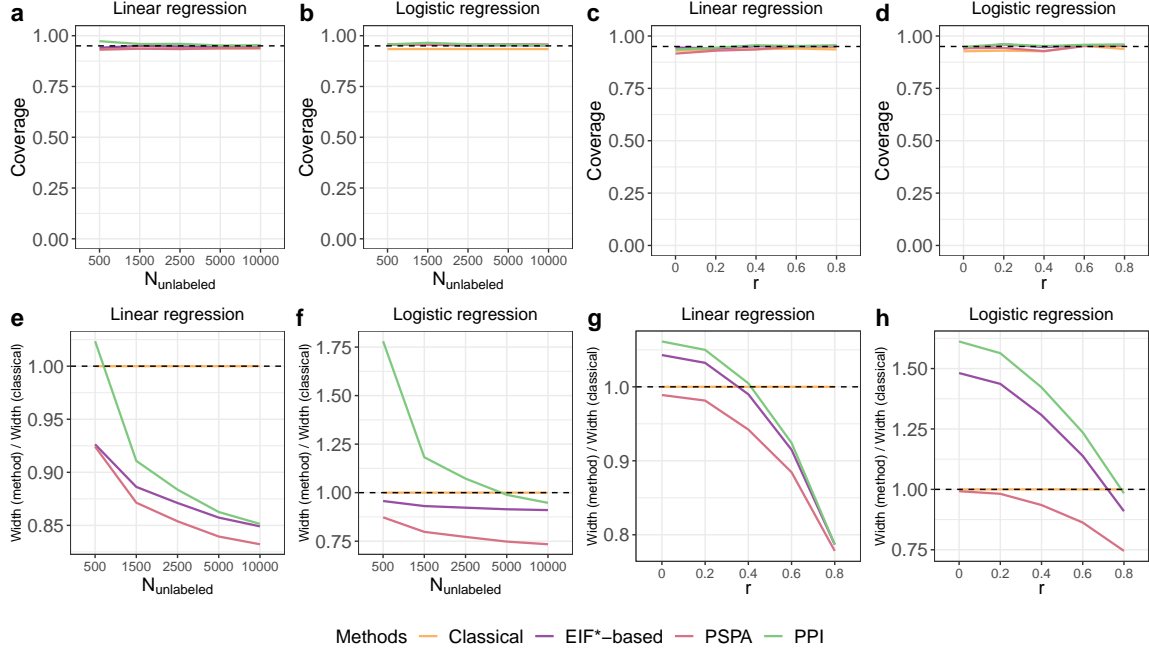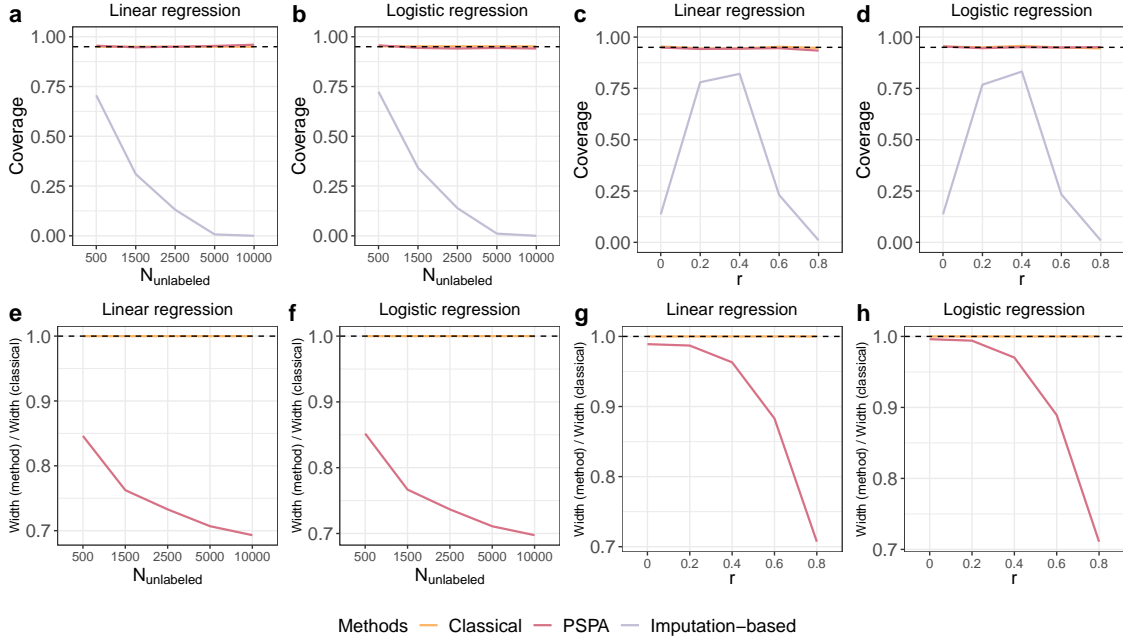

Figure 2: **Coverage of the confidence interval and relative ratio of its width compared to the classical method for linear and logistic regression.** ML is used to predict the labels. Panels (a)-(d) show the coverage of the confidence interval. Panels (e)-(h) show the relative ratio of the width of the confidence interval in comparison with the classical method. Panels (a), (b), (e), and (f) correspond to settings with varying sample sizes of unlabeled data. Panels (c), (d), (g), and (h) correspond to settings with different levels of imputation accuracy. The dashed lines represent $y = 0.95$ in (a)-(d) and $y = 1$ in (e)-(h).

6.2 Results

6.2.1 ML-PREDICTED LABELS

Figure 2 shows the simulation results for the setting of ML-predicted labels. All evaluated methods achieved confidence interval coverage rates close to 95%, indicating the inference validity across different unlabeled sample sizes and different levels of ML prediction accuracy. PPI and EIF*-based methods may be less efficient than classical methods when the ML has poor prediction accuracy or limited unlabeled data. PSPA has narrower confidence intervals compared to the classical method and other baseline methods, especially as the unlabeled sample size and ML prediction performance increase.

6.2.2 ML-PREDICTED COVARIATES

Figure 3 shows the simulation results for the setting with ML-predicted covariates. The imputation-based method fails to achieve the correct confidence interval coverage. Both classical and PSPA methods lead to coverage rates around 95%, which suggests valid inference results. PSPA has narrower confidence intervals compared to the classical method, regardless of the unlabeled sample size and the accuracy of the ML model.

Figure 3: **Coverage of the confidence interval and relative ratio of its width compared to the classical method for linear and logistic regression.** ML is used to predict the covariates. Panels (a)-(d) show the coverage of the confidence interval. Panels (e)-(h) show the relative ratio of the width of the confidence interval in comparison with the classical method. Panels (a), (b), (e), and (f) correspond to settings with varying sample sizes of unlabeled data. Panels (c), (d), (g), and (h) correspond to settings with different levels of imputation accuracy. The dashed lines represent $y = 0.95$ in (a)-(d) and $y = 1$ in (e)-(h).

7. Data Applications

7.1 Sex-differentiated gene expression

We used the PSPA method to assess the effect of biological sex on gene expression across 44 human tissues using GTEx data. We aim to use an ML algorithm called hypergraph factorization (HYFA) to impute the missing gene expression (Viñas et al., 2023) in the uncollected tissue. The imputed gene expression is then used in PSPA to increase the statistical power to identify sex-biased genes. We compared PSPA with classical, PPI, and EIF*-based approaches. In particular, we processed the GTEx-v8 data following the GTEx pipeline (GTEx Consortium et al., 2015). We then used HYFA to predict transcript levels for each uncollected tissue for each individual in the data, resulting in a gene expression dataset of 834 individuals across 44 tissues. HYFA is a parameter-efficient graph representation learning approach for multi-tissue prediction of gene expression. We used cross-validation to predict expression levels in the labeled data to avoid over-fitting. Prediction accuracy is measured by the correlation between measured and imputed gene expression in the labeled data. We then used linear regression to assess the effect of sex on gene expression while controlling for technical factors, including surrogate variables, age, RNA integrity number, and total ischemic time for a sample.
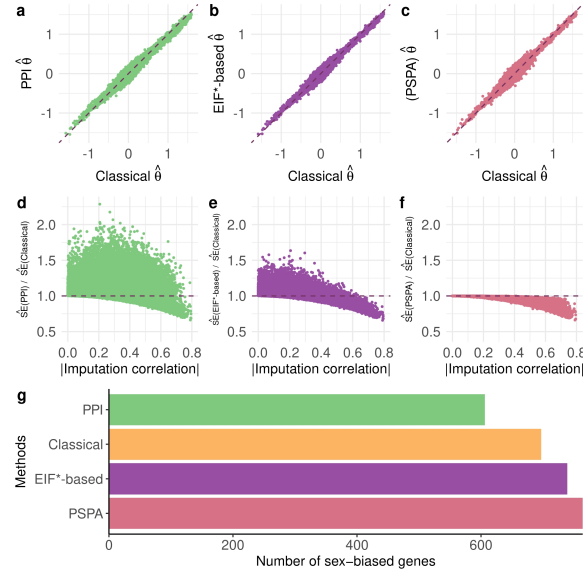

Figure 4: **Comparison of PSPA, classical, PPI, and EIF*-based approaches in identifying sex-biased gene expression using GTEx data.** Each panel illustrates a different aspect of comparison on the y- and x- axes: point estimates between the (a) classical and PPI approaches; (b) classical and EIF*-based approaches; (c) classical and PSPA approaches; estimated standard errors between the (d) classical and PPI approaches; (e) classical and EIF*-based approaches; (f) classical and PSPA approaches; (g) number of sex-biased genes identified by each of the four approaches. The dashed lines represent $y = x$ in (a)-(c) and $y = 1$ in (d)-(f).

Results are shown in Figure 4. Each dot represents the result of the inference for one gene in one tissue in Figure 4 (a)-(f). The point estimates of all methods are close to the classical approach (Figure 4 (a)-(c)). However, both PPI and EIF*-based are less efficient than the classical approach when the prediction accuracy is low (Figure 4 (d)-(e)). In contrast, the PSPA estimator is always at least as efficient as the classical approach, regardless of the prediction accuracy (Figure 4 (f)). Moreover, PSPA identifies more sex-biased genes than other approaches, demonstrating its improved efficiency over alternatives (Figure 4 (g)).

7.2 Risk factors for bone mineral density

Next, we applied PSPA to a multi-dimensional linear regression task. The goal is to identify the associations between dual-energy x-ray absorptiometry (DXA)-derived total bone mineral density (DXA-BMD) and several covariates, including biological sex, age, physical activities (PA), sedentary behavior (SB), smoking status (current smoker or not), and frequency of alcohol intake. DXA-BMD serves as the primary diagnostic marker for osteoporosis and fracture risk in clinical settings. For PA, we assigned individuals three levels (low, medium, and high) scores according to the International Physical Activity Questionnaire guidelines. SB was quantified as an integer value representing the combined hours spent driving, using a computer, and watching television.

Table 1: **Comparison of different methods in identifying risk factors for bone mineral density.** Estimate and $\widehat{\text{SE}}$ represent the estimated linear regression coefficient and its corresponding standard error for each covariate. $\widehat{\text{SE}}$ ratio indicates the ratio of the standard error for a specific method compared with the classical method. PA denotes physical activity, and SB refers to sedentary behavior. The bold fonts represent the method that gives the smallest $\widehat{\text{SE}}$ ratio for each covariate.

		Biological sex	Age	PA	SB	Smoking	Alcohol
Classical	Estimate	-0.616	-0.190	0.019	0.040	0.006	-0.008
	$\widehat{\text{SE}}$	4.20E-03	4.12E-03	4.07E-03	4.18E-03	4.13E-03	4.20E-03
	$\widehat{\text{SE}}$ ratio	1	1	1	1	1	1
PPI	Estimate	-0.604	-0.201	0.009	0.043	-0.026	-0.002
	$\widehat{\text{SE}}$	5.73E-03	6.26E-03	6.35E-03	6.48E-03	6.39E-03	6.47E-03
	$\widehat{\text{SE}}$ ratio	1.363	1.517	1.560	1.553	1.549	1.541
EIF*-based	Estimate	-0.605	-0.200	0.010	0.042	-0.023	-0.003
	$\widehat{\text{SE}}$	5.28E-03	5.74E-03	5.83E-03	5.95E-03	5.87E-03	5.93E-03
	$\widehat{\text{SE}}$ ratio	1.257	1.391	1.433	1.424	1.423	1.412
PSPA	Estimate	-0.614	-0.181	0.010	0.040	-0.008	-5.15E-07
	$\widehat{\text{SE}}$	3.74E-03	3.74E-03	3.74E-03	3.82E-03	3.80E-03	3.81E-03
	$\widehat{\text{SE}}$ ratio	**0.892**	**0.907**	**0.919**	**0.914**	**0.920**	**0.908**

We regressed DXA-BMD on these variables using data from the UK Biobank (UKB). In the UKB, DXA-BMD measurements are available for only 10% of the participants. Therefore, we employed the Softimpute algorithm to impute DXA-BMD values for the remaining 90% individuals in the unlabeled dataset. To prevent overfitting, cross-validation was applied for the imputation in the labeled data. We consider classical, `PPI`, and EIF*-based method as the baseline method.

Table 1 presents the inference results. Both the `PPI` and EIF*-based methods show larger standard errors across all covariates compared to the classical method. In contrast, `PSPA` shows smaller standard errors for all coordinates when compared to the classical method, demonstrating its feature for element-wise variance reduction.

In conclusion, the real data results are consistent with our theoretical and simulation results, demonstrating that our method `PSPA` provides more efficient inference results in real-world post-prediction inference applications.

8. Comparison with `PPI++`

Finally, we present a comparison between `PPI++` (Angelopoulos et al., 2023b) and our method `PSPA`, both theoretically and empirically. First, we note that `PPI++` can be reviewed as a special case of `PSPA` when each element of the weighting vector $\boldsymbol{\omega}$ is constrained to the same values. `PSPA` degenerates to `PPI++` for the one-dimensional estimation task such as mean estimation. Moreover, `PPI++` cannot be applied to the case where the covariates are predicted by ML instead of the labels. Theoretically, since `PPI++` can be reviewed as a special case of `PSPA`, `PPI++` is less efficient than `PSPA` by Proposition 6. Moreover, `PPI++` fails to guarantee element-wise variance reduction since it only incorporates one scalar for variance reduction.

Empirically, we compared `PPI++` with `PSPA` for all experiments detailed in this paper. We present the simulation results in Figure 5 and real data applications in Figure 6 and Table A, respectively. Simulation results show that both `PPI++` and `PSPA` have the correct confidence interval coverage. However, `PSPA` has narrower confidence intervals compared with `PPI++`. In real data applications, `PSPA` identifies more sex-biased gene expression than `PPI++` in the GTEx example and has a smaller estimated standard error compared to `PPI++` for every covariate in the DXA-BMD example. These results are consistent with Proposition 6, showing `PSPA` is more efficient than `PPI++` in post-prediction inference applications.

9. Discussion

We have provided a simple yet powerful method, `PSPA`, to improve the efficiency of statistical inference with arbitrary ML predictions. To enable inference, we establish the consistency and asymptotic normality of the proposed estimator and prove its superiority over baseline methods. Through extensive simulations and real data applications, we demonstrate the superiority of the `PSPA` estimator over alternatives.

This work fits into a variety of scientific problems and unlocks several directions for future research. One can apply the same principle to more complicated data structures and missing mechanisms, for example, in cases where a multi-dimensional outcome or covariate is subject to missingness. One promising direction of future work is to carefully discuss

relationships between $\mathbf{X}$ and $\mathbf{Z}$ and how they may affect the magnitude of the efficiency improvement.

The R codes to implement `PSPA`, benchmark methods, and replicate the simulation and real data analysis, is available at `https://github.com/qlu-lab/pspa`.
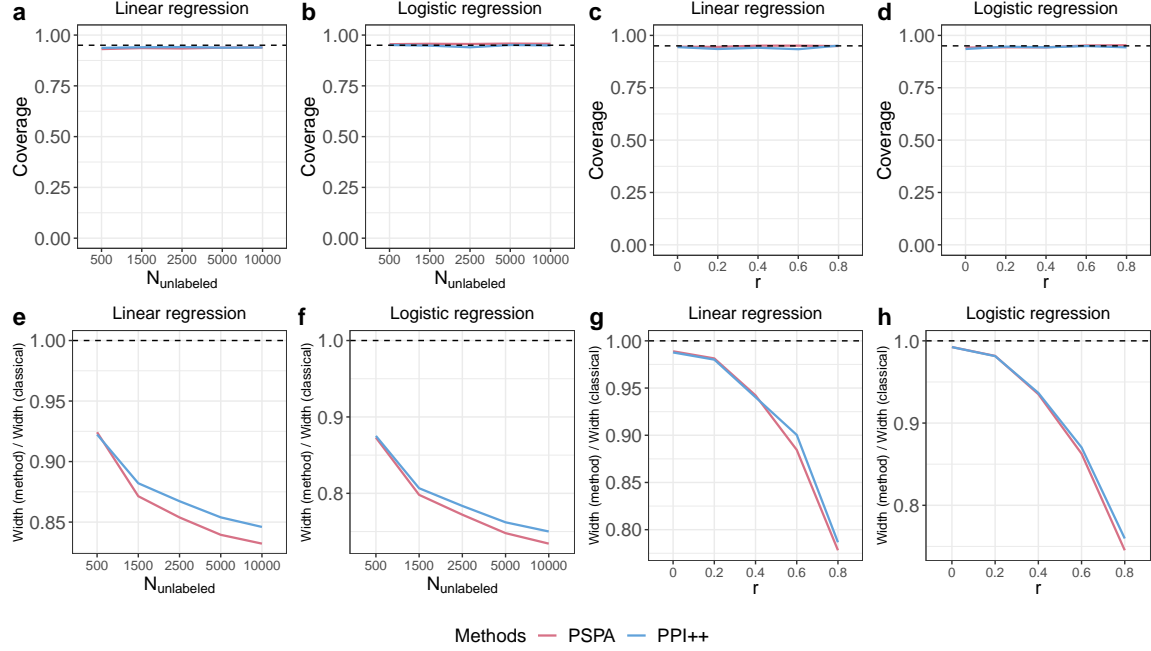
Acknowledgments

Appendix A. Supplementary Figures and Tables

Figure 5: **Comparison of `PPI++` and `PSPA` in simulation.** This figure shows the coverage of the confidence interval and the relative ratio of the width of the confidence interval compared to the classical method for linear and logistic regression. ML is used to predict the labels. Panels (a)-(d) show the coverage of the confidence interval. Panels (e)-(h) show the relative ratio of the width of the confidence interval in comparison with the classical method. Panels (a), (b), (e), and (f) correspond to settings with varying sample sizes of unlabeled data. Panels (c), (d), (g), and (h) correspond to settings with different levels of imputation accuracy. The dashed line represents $y = 0.95$ in (a)-(d) and $y = 1$ in (e)-(h).
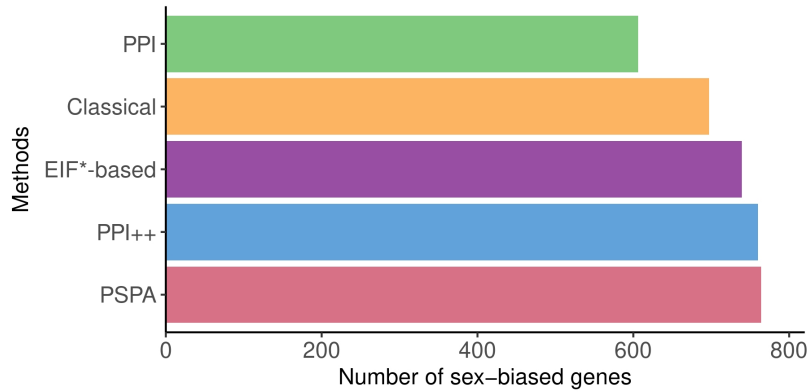
Figure 6: Comparison of `PPI++` and `PSPA` in identifying sex-biased genes in GTEx.

Table 2: **Comparison of `PPI++` and `PSPA` in identifying risk factors for bone mineral density.** Estimate and $\widehat{\text{SE}}$ represent the estimated linear regression coefficient and its corresponding standard error for each covariate, respectively. $\widehat{\text{SE}}$ ratio indicates the ratio of the standard error for a specific method compared with the classical method. PA denotes physical activity, and SB refers to sedentary behavior. The bold font represents the method that gives the smallest $\widehat{\text{SE}}$ ratio for each covariate.

		Biological sex	Age	PA	SB	Smoking	Alcohol
`PPI++`	Estimate	-0.613	-0.193	0.016	0.041	-0.002	-0.006
	$\widehat{\text{SE}}$	3.86E-03	3.79E-03	3.84E-03	3.91E-03	3.87E-03	3.89E-03
	$\widehat{\text{SE}}$ ratio	0.920	0.919	0.943	0.935	0.939	0.926
`PSPA`	Estimate	-0.614	-0.181	0.010	0.040	-0.008	-5.15E-07
	$\widehat{\text{SE}}$	3.74E-03	3.74E-03	3.74E-03	3.82E-03	3.80E-03	3.81E-03
	$\widehat{\text{SE}}$ ratio	**0.892**	**0.907**	**0.919**	**0.914**	**0.920**	**0.908**

Appendix B. Proofs

B.1 Proof of Theorem 3

We first present a lemma that will be used later.

Lemma 12 *Suppose that $X_1, \ldots, X_n$ are i.i.d., and the parameter space Θ is compact and $L(X, \theta)$ is continuous in $\theta \in \Theta$ almost everywhere. Moreover, there exists a function $h(X)$ satisfying $\|L(X, \theta)\| \le h(X)$ for arbitrary $\theta \in \Theta$ and $\mathbb{E}\{h(X)\} < \infty$, then $\mathbb{E}\{L(X, \theta)\}$ is continuous in θ and*

$$\sup_{\theta \in \Theta} \left\| \frac{1}{n} \sum_{i=1}^n L(X_i, \theta) - \mathbb{E}\{L(X, \theta)\} \right\| \xrightarrow{P} 0.$$

Lemma 12 follows from Tauchen (1985, Lemma 1) and hence the proof is omitted.

Condition (C5): There exists a small neighborhood of $\boldsymbol{\theta}$, denoted as B, in which $\boldsymbol{\psi}(y, \mathbf{x}, \boldsymbol{\theta})$ is continuously differentiable with respect to $\boldsymbol{\theta}$ almost everywhere and the partial derivative satisfies $\mathbb{E}\sup_{\boldsymbol{\theta} \in B}\|\partial\boldsymbol{\psi}(Y, \mathbf{X}, \boldsymbol{\theta})/\partial\boldsymbol{\theta}\| < \infty$ and $\mathbb{E}\sup_{\boldsymbol{\theta} \in B}\|\partial\boldsymbol{\psi}(\widehat{f}(\mathbf{Z}), \mathbf{X}, \boldsymbol{\theta})/\partial\boldsymbol{\theta}\| < \infty$.

For simplicity of notation, we re-write the estimating equation (3) as

$$\Psi(\widehat{\boldsymbol{\theta}}_{\text{PSPA}}, \boldsymbol{\omega}) = \Psi_n^Y + \text{diag}(\boldsymbol{\omega})(-\Psi_n^{\widehat{f}} + \Psi_N^{\widehat{f}}) = \mathbf{0}, \text{ where}$$

$$\Psi_n^Y = \frac{1}{n}\sum_{i=1}^n \psi(y_i, \mathbf{x}_i; \widehat{\boldsymbol{\theta}}_{\text{PSPA}}), \Psi_n^{\widehat{f}} = \frac{1}{n}\sum_{i=1}^n \psi(\widehat{f}, \mathbf{x}_i; \widehat{\boldsymbol{\theta}}_{\text{PSPA}}), \text{ and } \Psi_N^{\widehat{f}} = \frac{1}{N}\sum_{i=n+1}^{n+N} \psi(\widehat{f}, \mathbf{x}_i; \widehat{\boldsymbol{\theta}}_{\text{PSPA}}).$$

We first establish consistency following Van der Vaart (2000, Theorem 5.9) by checking its two conditions. The deterministic condition is implied by Condition (C1) and (C2). We are

left to verify the uniform convergence condition $\|\Psi(\widehat{\boldsymbol{\theta}}_{\text{PSPA}}) - \mathbb{E}(\Psi_n^Y)\| = o_p(1)$. By Condition (C3) and Lemma 12, we have

$$\|\text{diag}(\boldsymbol{\omega})\Psi_n^{\widehat{f}} - \mathbb{E}(\text{diag}(\boldsymbol{\omega})\Psi_n^{\widehat{f}})\| = o_p(1),$$
$$\|\text{diag}(\boldsymbol{\omega})\Psi_N^{\widehat{f}} - \mathbb{E}(\text{diag}(\boldsymbol{\omega})\Psi_N^{\widehat{f}})\| = o_p(1), \text{ and}$$
$$\|\Psi_n^Y - \mathbb{E}(\Psi_n^Y)\| = o_p(1).$$

Hence, by triangular inequality,

$$
\begin{aligned}
&\|\Psi(\widehat{\boldsymbol{\theta}}_{\text{PSPA}}) - \mathbb{E}(\Psi_n^Y)\| \\
\leq &\|\Psi(\widehat{\boldsymbol{\theta}}_{\text{PSPA}}) - \Psi_n^Y\| + \|\Psi_n^Y - \mathbb{E}(\Psi_n^Y)\| \\
= &\| - \text{diag}(\boldsymbol{\omega})\Psi_n^{\widehat{f}} + \text{diag}(\boldsymbol{\omega})\Psi_N^{\widehat{f}}\| + \|\Psi_n^Y - \mathbb{E}(\Psi_n^Y)\| \\
\leq &\| - \boldsymbol{\omega} \cdot \Psi_n^{\widehat{f}} + \mathbb{E}(\text{diag}(\boldsymbol{\omega})\Psi_n^{\widehat{f}})\| + \|\text{diag}(\boldsymbol{\omega})\Psi_N^{\widehat{f}} - \mathbb{E}(\text{diag}(\boldsymbol{\omega})\Psi_N^{\widehat{f}})\| \\
&+ \|\mathbb{E}(\text{diag}(\boldsymbol{\omega})\Psi_N^{\widehat{f}} - \text{diag}(\boldsymbol{\omega})\Psi_n^{\widehat{f}})\| + \|\Psi_n^Y - \mathbb{E}(\Psi_n^Y)\| \\
= &o_p(1),
\end{aligned}
$$

which completes the proof of consistency, i.e., $\widehat{\boldsymbol{\theta}}_{\text{PSPA}} \xrightarrow{P} \boldsymbol{\theta}$. Next, we establish the asymptotic normality. Expanding $\Psi(\widehat{\boldsymbol{\theta}}_{\text{PSPA}}, \boldsymbol{\omega})$ at the true value $\boldsymbol{\theta}$,

$$\mathbf{0} = \Psi(\boldsymbol{\theta}, \boldsymbol{\omega}) + (\widehat{\boldsymbol{\theta}}_{\text{PSPA}} - \boldsymbol{\theta})\partial\Psi(\bar{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}$$

for some $\bar{\boldsymbol{\theta}}$ between $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}_{\text{PSPA}}$. Multiplying both sides with $\sqrt{n}$ and with a direct calculation,

$$\left\{\partial\Psi(\bar{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}\right\}\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\text{PSPA}} - \boldsymbol{\theta}) = \sqrt{n}\Psi(\boldsymbol{\theta}, \boldsymbol{\omega}).$$

In order to show the asymptotic normality, by Slutsky's theorem, it's sufficient to verify the following:

$$\|\partial\Psi(\bar{\boldsymbol{\theta}}, \boldsymbol{\omega})/\partial\boldsymbol{\theta} - \mathbb{E}\{\partial\boldsymbol{\psi}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\}\| = o_p(1),$$
$$\sqrt{n}\Psi(\boldsymbol{\theta}, \boldsymbol{\omega}) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathbf{V}(\boldsymbol{\omega})).$$

To show (B.1), we note that within the neighborhood B in Condition (C5),

$$
\begin{aligned}
&\left\|\frac{\partial\Psi(\boldsymbol{\theta}, \boldsymbol{\omega})}{\partial\boldsymbol{\theta}} - \frac{\partial\Psi_n^Y(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right\| \\
= &\left\|\text{diag}(\boldsymbol{\omega})\left(-\partial\Psi_n^{\widehat{f}}/\partial\boldsymbol{\theta} + \mathbb{E}\left[\partial\Psi_n^{\widehat{f}}/\partial\boldsymbol{\theta}\right]\right) + \text{diag}(\boldsymbol{\omega})\left(\partial\Psi_N^{\widehat{f}}/\partial\boldsymbol{\theta} - \mathbb{E}\left[\partial\Psi_N^{\widehat{f}}/\partial\boldsymbol{\theta}\right]\right)\right\| \\
\leq &\left\|\text{diag}(\boldsymbol{\omega})\left(-\partial\Psi_n^{\widehat{f}}/\partial\boldsymbol{\theta} + \mathbb{E}\left[\partial\Psi_n^{\widehat{f}}/\partial\boldsymbol{\theta}\right]\right)\right\| + \left\|\text{diag}(\boldsymbol{\omega})\left(\partial\Psi_N^{\widehat{f}}/\partial\boldsymbol{\theta} - \mathbb{E}\left[\partial\Psi_N^{\widehat{f}}/\partial\boldsymbol{\theta}\right]\right)\right\|.
\end{aligned}
$$

Then together with Condition (C4) and Lemma 12, we have

$$\sup_{\widetilde{\boldsymbol{\theta}} \in B}\left\|\partial\Psi(\widetilde{\boldsymbol{\theta}})/\partial\widetilde{\boldsymbol{\theta}} - \partial\Psi_n^Y(\widetilde{\boldsymbol{\theta}})/\partial\widetilde{\boldsymbol{\theta}}\right\| = o_p(1).$$

24

By Conditions (C1, (C4) and Lemma 12, we have

$$\sup_{\widetilde{\boldsymbol{\theta}} \in B} \left\| \partial \Psi_n^Y(\widetilde{\boldsymbol{\theta}})/\partial\widetilde{\boldsymbol{\theta}} - \mathbb{E}\left[\partial \Psi_n^Y(\widetilde{\boldsymbol{\theta}})/\partial\widetilde{\boldsymbol{\theta}}\right] \right\| = o_p(1).$$

Equations (B.1) and (B.1) give $\sup_{\widetilde{\boldsymbol{\theta}} \in B} \left\| \partial\Psi(\widetilde{\boldsymbol{\theta}})/\partial\widetilde{\boldsymbol{\theta}} - \mathbb{E}\left[\partial\Psi_n^Y(\widetilde{\boldsymbol{\theta}})/\partial\widetilde{\boldsymbol{\theta}}\right] \right\| = o_p(1)$. Then the consistency of $\widehat{\boldsymbol{\theta}}_{\texttt{PSPA}}$ and the continuous mapping theorem imply that

$$\begin{aligned}
&\left\| \partial\Psi(\bar{\boldsymbol{\theta}}, \boldsymbol{\omega})/\partial\boldsymbol{\theta} - \mathbb{E}\left[\partial\boldsymbol{\psi}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\right] \right\| \\
&\leq \sup_{\widetilde{\boldsymbol{\theta}} \in B} \left\| \partial\Psi(\widetilde{\boldsymbol{\theta}})/\partial\widetilde{\boldsymbol{\theta}} - \mathbb{E}\left[\partial\Psi_n^Y(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\right] \right\| + \left\| \mathbb{E}\left[\partial\boldsymbol{\psi}(\bar{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}\right] - \mathbb{E}\left[\partial\boldsymbol{\psi}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\right] \right\| \\
&\leq o_p(1).
\end{aligned}$$

Therefore, condition (B.1) has been verified. Meanwhile, (B.1) holds by Central Limit Theorem. The proof of asymptotic normality is completed.

B.2 Proof of Corollary 4

We denote the corresponding estimator as $\widehat{\boldsymbol{\theta}}(\widehat{\boldsymbol{\omega}})$. Then the consistency of $\widehat{\boldsymbol{\theta}}(\widehat{\boldsymbol{\omega}})$ follows from the consistency of $\widehat{\boldsymbol{\omega}}$ and the proof in Section B.1. For asymptotic normality of $\widehat{\boldsymbol{\theta}}(\widehat{\boldsymbol{\omega}})$, applying Taylor expansion on $0 = \Psi(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\omega}})$ yields

$$0 = \Psi(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\omega}}) = \Psi(\boldsymbol{\theta}, \boldsymbol{\omega}^{\mathrm{opt}}) + \mathrm{diag}(\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}^{\mathrm{opt}})\left(-\Psi_n^{\widehat{f}} + \Psi_N^{\widehat{f}}\right) + (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})\partial\Psi(\bar{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}$$

for some $\bar{\boldsymbol{\theta}}$ between $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}(\widehat{\boldsymbol{\omega}})$. With the same proof technique in Section B.1, we have

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}(\widehat{\boldsymbol{\omega}}) - \boldsymbol{\theta}) \overset{D}{\to} N(\boldsymbol{0}, \boldsymbol{\Sigma}(\boldsymbol{\omega}^{\mathrm{opt}})),$$

which implies Corollary 4.

B.3 Proof of Proposition 6

The j-th diagonal element of the asymptotic covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\omega})$ of $\widehat{\boldsymbol{\theta}}_{\texttt{PSPA}}(\boldsymbol{\omega})$ given weighting vector $\boldsymbol{\omega}$ is

$$\boldsymbol{\Sigma}_{jj}(\boldsymbol{\omega}) = \omega_j^2[\mathbf{A}^{-1}(\mathbf{M}_2 + \rho\mathbf{M}_3)\mathbf{A}^{-1}]_{jj} - 2\omega_j[\mathbf{A}^{-1}\mathbf{M}_4\mathbf{A}^{-1}]_{jj} + [\mathbf{A}^{-1}\mathbf{M}_1\mathbf{A}^{-1}]_{jj}.$$

It is a quadratic function of ω_j with a unique minimizer

$$\omega_j^{\mathrm{opt}} = \frac{[\mathbf{A}^{-1}\mathbf{M}_4\mathbf{A}^{-1}]_{jj}}{[\mathbf{A}^{-1}(\mathbf{M}_2 + \rho\mathbf{M}_3)\mathbf{A}^{-1}]_{jj}}.$$

Therefore, given any weighting vector $\boldsymbol{\omega}^* = [\omega_1^*, \cdots, \omega_q^*]^\top$, $\boldsymbol{\Sigma}_{jj}(\boldsymbol{\omega}^{\mathrm{opt}}) \leq \boldsymbol{\Sigma}_{jj}(\boldsymbol{\omega}^*)$, which completes the proof.

B.4 Proof of Proposition 7

The joint log-likelihood for a generic observation $(r, \mathbf{x}, \mathbf{z}, y)$ is

$$l = r\mathrm{log}p(y \mid \mathbf{x}, \mathbf{z}) + \mathrm{log}p(\mathbf{x} \mid \mathbf{z}) + \mathrm{log}p(\mathbf{z}) + r\mathrm{log}(\pi) + (1 - r)\mathrm{log}(1 - \pi).$$

Taking the semiparametric approach (Bickel et al., 1993; Tsiatis, 2006), we consider the Hilbert space $\mathcal{H}$ that contains all one-dimensional zero-mean measurable functions of the observed data with finite variance, equipped with the inner product $\langle h_1, h_2 \rangle = \mathbb{E}\{h_1(\cdot)h_2(\cdot)\}$ where $h_1, h_2 \in \mathcal{H}$. To estimate $\boldsymbol{\theta}$, we regard $p(y \mid \mathbf{x}, \mathbf{z})$, $p(\mathbf{x} \mid \mathbf{z})$ and $p(\mathbf{z})$ as nuisance functions. Denote their nuisance tangent spaces by $\mathcal{T}_y$, $\mathcal{T}_{\mathbf{x}}$, and $\mathcal{T}_{\mathbf{z}}$, which are defined as the mean squared closure of the nuisance tangent spaces of parametric submodels spanned by the nuisance score vectors. We have that $\mathcal{T} = \mathcal{T}_y \oplus \mathcal{T}_{\mathbf{x}} \oplus \mathcal{T}_{\mathbf{z}}$ where

$$\mathcal{T}_y = [r\mathbf{a}_1(y, \mathbf{x}, \mathbf{z}) : \mathbb{E}\{\mathbf{a}_1(Y, \mathbf{x}, \mathbf{z}) \mid \mathbf{x}, \mathbf{z}\} = \mathbf{0}],$$
$$\mathcal{T}_{\mathbf{x}} = [\mathbf{a}_2(\mathbf{x}, \mathbf{z}) : \mathbb{E}\{\mathbf{a}_2(X, \mathbf{z}) \mid \mathbf{z}\} = \mathbf{0}], \text{ and}$$
$$\mathcal{T}_{\mathbf{z}} = [\mathbf{a}_3(\mathbf{z}), \mathbb{E}\{\mathbf{a}_3(Z)\} = \mathbf{0}],$$

respectively, and the notation $\oplus$ represents the direct sum of two spaces that are orthogonal to each other.

We introduce parametric submodels $p_{\boldsymbol{\eta}_y}(y \mid \mathbf{x}, \mathbf{z})$, $p_{\boldsymbol{\eta}_{\mathbf{x}}}(\mathbf{x} \mid \mathbf{z})$, and $p_{\boldsymbol{\eta}_{\mathbf{z}}}(\mathbf{z})$, where $\boldsymbol{\eta} = [\boldsymbol{\eta}_y^\top, \boldsymbol{\eta}_{\mathbf{x}}^\top, \boldsymbol{\eta}_{\mathbf{z}}^\top, \boldsymbol{\eta}_r^\top]^\top$ is a vector of nuisance parameters. Let

$$\mathbf{S}_{\boldsymbol{\eta}_y}(y, \mathbf{x}, \mathbf{z}) = \frac{\partial \mathrm{log}p_{\boldsymbol{\eta}_y}(y, \mathbf{x}, \mathbf{z})}{\partial \boldsymbol{\eta}_y}, \ \mathbf{S}_{\boldsymbol{\eta}_{\mathbf{x}}}(\mathbf{x}, \mathbf{z}) = \frac{\partial \mathrm{log}p_{\boldsymbol{\eta}_{\mathbf{x}}}(\mathbf{x}, \mathbf{z})}{\partial \boldsymbol{\eta}_{\mathbf{x}}}, \ \text{and} \ \mathbf{S}_{\boldsymbol{\eta}_{\mathbf{z}}}(\mathbf{z}) = \frac{\partial \mathrm{log}p_{\boldsymbol{\eta}_{\mathbf{z}}}(\mathbf{z})}{\partial \boldsymbol{\eta}_{\mathbf{z}}},$$

then these scores functions satisfy $r\mathbf{S}_{\boldsymbol{\eta}_y}(y, \mathbf{x}, \mathbf{z}) \in \mathcal{T}_y$, $\mathbf{S}_{\boldsymbol{\eta}_{\mathbf{x}}}(\mathbf{x}, \mathbf{z}) \in \mathcal{T}_{\mathbf{x}}$, and $\mathbf{S}_{\boldsymbol{\eta}_{\mathbf{z}}}(\mathbf{z}) \in \mathcal{T}_{\mathbf{z}}$. Recall the definition of parameter of interest, i.e.,

$$\iiint \boldsymbol{\psi}(y, \mathbf{x}; \boldsymbol{\theta})p(y \mid \mathbf{x}, \mathbf{z})p(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z}\mathrm{d}\mathbf{x}\mathrm{d}y = \mathbf{0}.$$

Taking derivative of $\boldsymbol{\theta}$ with respect to nuisance parameters,

$$\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}_y^\top} = \mathbf{A}^{-1}\frac{\partial \mathbb{E}\{\boldsymbol{\psi}(Y, \mathbf{X}; \boldsymbol{\theta})\}}{\partial \boldsymbol{\eta}_y^\top} = \mathbf{A}^{-1}\mathbb{E}[\boldsymbol{\psi}(Y, \mathbf{X}; \boldsymbol{\theta})\mathbf{S}_{\boldsymbol{\eta}_y}^\top(Y, \mathbf{X}, \mathbf{Z})],$$
$$\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}_{\mathbf{x}}^\top} = \mathbf{A}^{-1}\frac{\partial \mathbb{E}\{\boldsymbol{\psi}(Y, \mathbf{X}; \boldsymbol{\theta})\}}{\partial \boldsymbol{\eta}_{\mathbf{x}}^\top} = \mathbf{A}^{-1}\mathbb{E}\{\boldsymbol{\psi}(Y, \mathbf{X}; \boldsymbol{\theta})\mathbf{S}_{\boldsymbol{\eta}_{\mathbf{x}}}^\top(\mathbf{X}, \mathbf{Z})\}, \text{ and}$$
$$\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}_{\mathbf{z}}^\top} = \mathbf{A}^{-1}\frac{\partial \mathbb{E}\{\boldsymbol{\psi}(Y, \mathbf{X}; \boldsymbol{\theta})\}}{\partial \boldsymbol{\eta}_{\mathbf{z}}^\top} = \mathbf{A}^{-1}\mathbb{E}\{\boldsymbol{\psi}(Y, \mathbf{X}; \boldsymbol{\theta})\mathbf{S}_{\boldsymbol{\eta}_{\mathbf{z}}}^\top(\mathbf{Z})\}.$$

Let $\boldsymbol{\phi}(y, \mathbf{x}; \boldsymbol{\theta}) = r\mathbf{a}_1(y, \mathbf{x}, \mathbf{z}) + \mathbf{a}_2(\mathbf{x}, \mathbf{z}) + \mathbf{a}_3(\mathbf{z})$, where

$$\mathbf{a}_1(y, \mathbf{x}, \mathbf{z}) = \frac{1}{\pi(\mathbf{z})}\mathbf{A}^{-1}\left[\boldsymbol{\psi}(y, \mathbf{x}; \boldsymbol{\theta}) - \mathbb{E}\{\boldsymbol{\psi}(Y, \mathbf{x}; \boldsymbol{\theta}) \mid \mathbf{x}, \mathbf{z}\}\right],$$
$$\mathbf{a}_2(\mathbf{x}, \mathbf{z}) = \mathbf{A}^{-1}\left[\mathbb{E}\{\boldsymbol{\psi}(Y, \mathbf{x}; \boldsymbol{\theta}) \mid \mathbf{x}, \mathbf{z}\} - \mathbb{E}\{\boldsymbol{\psi}(Y, \mathbf{x}; \boldsymbol{\theta}) \mid \mathbf{x}, \mathbf{z}\}\right], \text{ and}$$
$$\mathbf{a}_3(\mathbf{z}) = \mathbf{A}^{-1}\mathbb{E}\{\boldsymbol{\psi}(Y, \mathbf{x}; \boldsymbol{\theta}) \mid \mathbf{x}, \mathbf{z}\}.$$

Then it can be easily verified that $r\mathbf{a}_1(y, \mathbf{x}, \mathbf{z}) \in \mathcal{T}_y$, $\mathbf{a}_2(\mathbf{x}, \mathbf{z}) \in \mathcal{T}_{\mathbf{x}}$, $\mathbf{a}_3(\mathbf{z}) \in \mathcal{T}_{\mathbf{x}}$, and

$$\mathbb{E}\{\phi(Y, \mathbf{X}; \boldsymbol{\theta}) R \mathbf{S}_{\boldsymbol{\eta}_y}^\top (Y, \mathbf{X}, \mathbf{Z})\} = \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}_y^\top}, \ \mathbb{E}\left\{\phi(Y, \mathbf{X}; \boldsymbol{\theta}) \mathbf{S}_{\boldsymbol{\eta}_\mathbf{x}}^\top (\mathbf{X}, \mathbf{Z}))\right\} = \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}_\mathbf{x}^\top}, \text{ and } \mathbb{E}\{\phi(Y, \mathbf{X}; \boldsymbol{\theta}) \mathbf{S}_{\boldsymbol{\eta}_\mathbf{z}}^\top (\mathbf{Z})\} = \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}_\mathbf{z}^\top}.$$

Therefore, $\phi(y, \mathbf{x}; \boldsymbol{\theta})$ is the efficient influence function. The proof of Proposition 7 is completed.

B.5 Lemma for `PSPA` when applied to ML-predicted covariates

Lemma 13 *Letting $\mathbf{A}$ and $\mathbf{B}$ be two $q \times q$ gram matrices, and $\boldsymbol{\omega}$ is a q-dimensional vector. If $0 \leq \omega_j \leq \dfrac{\lambda_{\min,+}(\mathbf{A})}{\lambda_{\max}(\mathbf{B})}$ for all $j \in \{1, \cdots, p\}$ then $\mathbf{A} - diag(\boldsymbol{\omega})\mathbf{B}$ is positive semi-definite.*

To prove the lemma, we will verify that $\mathbf{A} - \mathrm{diag}(\boldsymbol{\omega})\mathbf{B}$ is positive semi-definite by showing that for any vector $\mathbf{x}$, the quadratic form $\mathbf{x}^T(\mathbf{A} - \mathrm{diag}(\boldsymbol{\omega})\mathbf{B})\mathbf{x} \geq 0$.

Let $\mathbf{x}$ be any vector in $\mathbb{R}^p$. Since

$$\mathbf{x}^\top \mathbf{A} x \geq \lambda_{\min,+}(\mathbf{A}) \|\mathbf{x}\|^2$$
$$\mathbf{x}^\top \mathrm{diag}(\boldsymbol{\omega}) \mathbf{B} \mathbf{x} \leq \lambda_{\max}(\mathrm{diag}(\boldsymbol{\omega})\mathbf{B}) \|x\|^2 \leq \max_j \omega_j \lambda_{\max} \mathbf{B} \|\mathbf{x}\|^2$$

Therefore

$$\mathbf{x}^\top \mathbf{A} \mathbf{x} - \mathbf{x}^\top \mathrm{diag}(\boldsymbol{\omega}) \mathbf{B} \geq (\lambda_{\min,+}(\mathbf{A}) - \max_j \omega_j \lambda_{\max} \mathbf{B}) \|\mathbf{x}\|^2$$

To ensure $\mathbf{x}^\top \mathbf{A} \mathbf{x} - \mathbf{x}^\top \mathrm{diag}(\boldsymbol{\omega}) \mathbf{B} \geq 0$, we require $\lambda_{\min,+}(\mathbf{A}) - \max_j \omega_j \lambda_{\max}(\mathbf{B}) \geq 0$, that is

$$\max_j \omega_j \leq \frac{\lambda_{\min,+}(\mathbf{A})}{\lambda_{\max}(\mathbf{B})}$$

B.6 Proof of Corollary 11

The j-th diagonal element of the asymptotic covariance matrix $\boldsymbol{\Sigma}'(\boldsymbol{\omega})$ of $\widehat{\boldsymbol{\theta}}_{\mathtt{PSPA}'}$ given weighting vector $\boldsymbol{\omega}$ is

$$\boldsymbol{\Sigma}'_{jj}(\boldsymbol{\omega}) = \omega_j^2 [\mathbf{A}^{-1}(\mathbf{M}_2' + \rho \mathbf{M}_3')\mathbf{A}^{-1}]_{jj} - 2\omega_j [\mathbf{A}^{-1}\mathbf{M}_4'\mathbf{A}^{-1}]_{jj} + [\mathbf{A}^{-1}\mathbf{M}_1'\mathbf{A}^{-1}]_{jj},$$

It is a quadratic function of ω_j with a unique minimizer

$$\omega_j^{\mathrm{opt},*} = \frac{[\mathbf{A}^{-1}\mathbf{M}_4'\mathbf{A}^{-1}]_{jj}}{[\mathbf{A}^{-1}(\mathbf{M}_2' + \rho \mathbf{M}_3')\mathbf{A}^{-1}]_{jj}}.$$

If $\omega_j^{\mathrm{opt},*} > 0$, $\boldsymbol{\Sigma}'_{jj}(\boldsymbol{\omega})$ is a decreasing function with $\omega_j \in [0, \omega_j^{\mathrm{opt},*}]$. If $\omega_j^{\mathrm{opt},*} \leq 0$, $\boldsymbol{\Sigma}'_{jj}(\boldsymbol{\omega})$ $\boldsymbol{\Sigma}'_{jj}(\boldsymbol{\omega})$ is an increasing function with $\omega_j \in [\omega_j^{\mathrm{opt},*}, 0]$. We also have $\frac{\lambda_{\min,+}[\frac{1}{n}\nabla \boldsymbol{\psi}(y_i, \mathbf{x}_i; \boldsymbol{\theta})]}{\lambda_{\max}[\frac{1}{n}\nabla \boldsymbol{\psi}(y_i, \widehat{q}; \boldsymbol{\theta})]} > 0$

Therefore, given $\hat{\boldsymbol{\omega}}^{\mathrm{opt}} = [\widehat{\omega}_1^{\mathrm{opt}}, \ldots, \widehat{\omega}_q^{\mathrm{opt}}]^\top$, where

$$\widehat{\omega}_j^{\mathrm{opt}} = \begin{cases} \min\big(\frac{[\widehat{\mathbf{A}}^{-1}\widehat{\mathbf{M}}_{4,\mathrm{C}}'\widehat{\mathbf{A}}^{-1}]_{jj}}{[\widehat{\mathbf{A}}^{-1}(\widehat{\mathbf{M}}_{2,\mathrm{C}}'+\rho\widehat{\mathbf{M}}_{3,\mathrm{C}}')\widehat{\mathbf{A}}^{-1}]_{jj}}, \frac{\lambda_{\min,+}[\frac{1}{n}\nabla\boldsymbol{\psi}(y_i,\mathbf{x}_i;\boldsymbol{\theta})]}{\lambda_{\max}[\frac{1}{n}\nabla\boldsymbol{\psi}(y_i,\widehat{q};\boldsymbol{\theta})]}\big) & \text{if } \frac{[\widehat{\mathbf{A}}^{-1}\widehat{\mathbf{M}}_{4,\mathrm{C}}'\widehat{\mathbf{A}}^{-1}]_{jj}}{[\widehat{\mathbf{A}}^{-1}(\widehat{\mathbf{M}}_{2,\mathrm{C}}'+\rho\widehat{\mathbf{M}}_{3,\mathrm{C}}')\widehat{\mathbf{A}}^{-1}]_{jj}} > 0 \\ \max\big(\frac{[\widehat{\mathbf{A}}^{-1}\widehat{\mathbf{M}}_{4,\mathrm{C}}'\widehat{\mathbf{A}}^{-1}]_{jj}}{[\widehat{\mathbf{A}}^{-1}(\widehat{\mathbf{M}}_{2,\mathrm{C}}'+\rho\widehat{\mathbf{M}}_{3,\mathrm{C}}')\widehat{\mathbf{A}}^{-1}]_{jj}}, 0\big) & \text{if } \frac{[\widehat{\mathbf{A}}^{-1}\widehat{\mathbf{M}}_{4,\mathrm{C}}'\widehat{\mathbf{A}}^{-1}]_{jj}}{[\widehat{\mathbf{A}}^{-1}(\widehat{\mathbf{M}}_{2,\mathrm{C}}'+\rho\widehat{\mathbf{M}}_{3,\mathrm{C}}')\widehat{\mathbf{A}}^{-1}]_{jj}} \leq 0 \end{cases}$$

for all $j \in 1, \ldots, q$, and the weighting vector $\boldsymbol{\omega}^{\mathrm{C}} = [0, \cdots, 0]^\top$ that correspond to the classical estimator, $\boldsymbol{\Sigma}_{jj}(\boldsymbol{\omega}^{\mathrm{opt}}) \leq \boldsymbol{\Sigma}_{jj}(\boldsymbol{\omega}^{\mathrm{C}})$, which completes the proof.

Acknowledgments

References

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023a. doi: 10.1126/science.adi6000. URL https://www.science.org/doi/abs/10.1126/science.adi6000.

Anastasios N Angelopoulos, John C Duchi, and Tijana Zrnic. Ppi++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023b.

David Azriel, Lawrence D Brown, Michael Sklar, Richard Berk, Andreas Buja, and Linda Zhao. Semi-supervised linear regression. *Journal of the American Statistical Association*, 117(540):2238–2251, 2022.

Mahashweta Basu, Kun Wang, Eytan Ruppin, and Sridhar Hannenhalli. Predicting tissue-specific gene expression from whole blood transcriptome. *Science Advances*, 7(14): eabd6991, 2021.

Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Ya'acov Ritov, J Klaassen, Jon A Wellner, and YA'Acov Ritov. *Efficient and adaptive estimation for semiparametric models*, volume 4. Springer, 1993.

F Jay Breidt and Jean D Opsomer. Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 2017.

Eric L Bullock, Curtis E Woodcock, Carlos Souza Jr, and Pontus Olofsson. Satellite-based estimates reveal widespread forest degradation in the amazon. *Global Change Biology*, 26 (5):2956–2969, 2020.

Abhishek Chakrabortty and Tianxi Cai. Efficient and adaptive linear regression in semi-supervised settings. *The Annals of Statistics*, 2018.

Xiaohong Chen, Han Hong, and Elie Tamer. Measurement error models with auxiliary data. *The Review of Economic Studies*, 72(2):343–366, 2005.

GTEx Consortium. The gtex consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020.

GTEx Consortium, Kristin G Ardlie, David S Deluca, Ayellet V Segrè, Timothy J Sullivan, Taylor R Young, Ellen T Gelfand, Casandra A Trowbridge, Julian B Maller, Taru Tukiainen, et al. The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Nathan Kallus and Xiaojie Mao. On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *arXiv preprint arXiv:2003.12408*, 2020.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553): 436–444, 2015.

John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580–585, 2013.

Geert Molenberghs, Garrett Fitzmaurice, Michael G Kenward, Anastasios Tsiatis, and Geert Verbeke. *Handbook of missing data methodology*. CRC Press, 2014.

James M Robins. Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association*, volume 1999, pages 6–10. Indianapolis, IN, 2000.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

Shanshan Song, Yuanyuan Lin, and Yong Zhou. A general m-estimation theory in semi-supervised framework. *Journal of the American Statistical Association*, pages 1–11, 2023.

George Tauchen. Diagnostic testing and evaluation of maximum likelihood models. *Journal of Econometrics*, 30(1):415–443, 1985. ISSN 0304-4076. doi: https://doi. org/10.1016/0304-4076(85)90149-6. URL https://www.sciencedirect.com/science/article/pii/0304407685901496.

Anastasios A Tsiatis. *Semiparametric theory and missing data*. Springer, 2006.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Ramon Viñas, Chaitanya K Joshi, Dobrik Georgiev, Phillip Lin, Bianca Dumitrascu, Eric R Gamazon, and Pietro Liò. Hypergraph factorization for multi-tissue gene expression imputation. *Nature machine intelligence*, 5(7):739–753, 2023.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

Jiebiao Wang, Eric R Gamazon, Brandon L Pierce, Barbara E Stranger, Hae Kyung Im, Robert D Gibbons, Nancy J Cox, Dan L Nicolae, and Lin S Chen. Imputing gene expression in uncollected tissues within and beyond gtex. *The American Journal of Human Genetics*, 98(4):697–708, 2016.

Junhui Wang and Xiaotong Shen. Large margin semi-supervised learning. *Journal of Machine Learning Research*, 8(8), 2007.

Siruo Wang, Tyler H McCormick, and Jeffrey T Leek. Methods for correcting inference based on outcomes predicted by machine learning. *Proceedings of the National Academy of Sciences*, 117(48):30266–30275, 2020.

Anru Zhang, Lawrence D Brown, and T Tony Cai. Semi-supervised inference: General theory and estimation of means. *The Annals of Statistics*, 2019.