Blockwise Missingness meets AI: A Tractable Solution for Semiparametric Inference

Qi Xu[1], Lorenzo Testa[1,2], Jing Lei[1], and Kathryn Roeder[1,3]

[1]Department of Statistics & Data Science, Carnegie Mellon University
[2]L'EMbeDS, Sant'Anna School of Advanced Studies
[3]Department of Computational Biology, Carnegie Mellon University

{qixu,ltesta,jinglei,roeder}@andrew.cmu.edu

September 30, 2025

Abstract

We consider parameter estimation and inference when data feature blockwise, non-monotone missingness. Our approach, rooted in semiparametric theory and inspired by prediction-powered inference, leverages off-the-shelf AI (predictive or generative) models to handle missing completely at random mechanisms, by finding an approximation of the optimal estimating equation through a novel and tractable Restricted Anova hierarchY (RAY) approximation. The resulting **I**nference for **B**lockwise **M**issingness(RAY), or IBM(RAY) estimator incorporates pre-trained AI models and carefully controls asymptotic variance by tuning model-specific hyperparameters. We then extend IBM(RAY) to a general class of estimators. We find the most efficient estimator in this class, which we call IBM(Adaptive), by solving a constrained quadratic programming problem. All IBM estimators are unbiased, and, crucially, asymptotically achieving guaranteed efficiency gains over a naive complete-case estimator, *regardless* of the predictive accuracy of the AI models used. We demonstrate the finite-sample performance and numerical stability of our method through simulation studies and an application to surface protein abundance estimation.

Keywords: Data integration; Non-monotone missingness; Prediction-powered inference; Semiparametric theory; Z-estimation theory.

1 Introduction

Modern machine learning has been transformed by the advent of large-scale predictive and generative models trained on massive datasets and released as pre-trained models. These high-performance models not only offer unprecedented predictive and generative power in natural language processing (Achiam et al., 2023; Team et al., 2023; Liu et al., 2024) and computer vision (Esser et al., 2024; Ramesh et al., 2022), but also empower scientific discovery

through the deployment of sophisticated statistical tools. A recent line of research has focused on incorporating pre-trained AI models to augment statistical efficiency in semi-supervised settings (Wang et al., 2020; Angelopoulos et al., 2023a,b; Miao et al., 2023; Motwani and Witten, 2023; Zrnic and Candès, 2024; Ji et al., 2025; Xu et al., 2025). In this work, we consider a more general blockwise missingness setting, where multiple datasets are accessible, but each of them only contains a subset of the total features. Integrating multiple datasets, resulting in a blockwise missingness setting, can significantly expand the effective sample size. Therefore, it is of great interest to estimate parameters and provide valid inferential tools under this setting, and investigate how to maximize efficiency gains using pre-trained AI models.

Target parameter. Formally, given M sets of features or modalities, we define the full data as $Z = (X_1, \cdots, X_M) \sim \mathscr{P} = \mathscr{P}_{\theta,\eta}, Z \in \mathscr{Z}$, where θ is a finite-dimensional parameter of interest, and η denotes nuisance parameters, possibly infinite-dimensional. For instance, we may be interested in the mean of one feature, or in regression coefficients, while other parameters, like the marginal distribution or the conditional distribution of modalities, may be regarded as nuisances. The target parameter $\theta^\star = \theta(\mathscr{P}) \in \Omega \subset \mathbb{R}^d$ associated with the estimating function $\psi(\cdot,\cdot) : \mathscr{Z} \times \Omega \to \mathbb{R}^d$ is defined as the solution to the estimating equation

$$\mathbb{E}_{\mathscr{P}}[\psi(Z, \theta^\star)] = \mathbf{0}_d \,,$$

where $\mathbb{E}_{\mathscr{P}}[\psi^T(Z,\theta)\psi(Z,\theta)] < \infty$ and $\mathbb{E}_{\mathscr{P}}[\psi(Z^F,\theta)\psi^T(Z^F,\theta)]$ is positive definite for all $\theta \in \Omega$.

Blockwise missing patterns. Given M modalities, we can observe up to $2^M - 1$ missingness patterns. The overall dataset can be split into disjoint datasets based on the observed patterns, that is $\mathscr{D} = \cup_{r\in Q}\mathscr{D}_r = \cup_{r\in Q}(X_{i,r})_{i=1}^{n_r}$, where $r \subseteq [M] = \{X_1, \ldots, X_M\}$ indicates the observed set of features, and Q is the set of all the observed patterns in $\mathscr{D}$ with cardinality $|Q|$. We also use $R \in Q$ to denote a random variable that indicates the observed pattern. In Figure 1 we provide an illustrative example, which includes three modalities X_1, X_2, Y and four observed patterns $Q = \{\{X_1, X_2, Y\}, \{X_1, X_2\}, \{X_1, Y\}, \{X_1\}\}$. In addition, we use $N = \sum_{r\in Q} n_r$ and n_r/N to denote the total sample size and the proportion of observed pattern $R = r$. At the population level, the proportion of each observed pattern is denoted by π_r. Throughout the paper, we make the following assumptions:

Assumption 1.1 (Missing mechanism and positivity assumption)**.** In this work, we assume:

(a) Missing completely at random (MCAR): the distribution of R is independent of $\mathscr{P}$.

(b) Strict positivity: there exists a positive constant c such that $\pi_{[M]} > c > 0$.

Assumption (a) implies that all marginal and conditional distributions among $\mathscr{D}$ are homogeneous. Assumption (b) requires that the proportion of complete data be strictly positive, which is critical in the development of our methods for the target parameters.

1.1 Related work

Prediction-powered inference. Prediction-powered inference (PPI, Angelopoulos et al. 2023a,b; Zrnic and Candès 2024) is a versatile framework for parameter estimation and

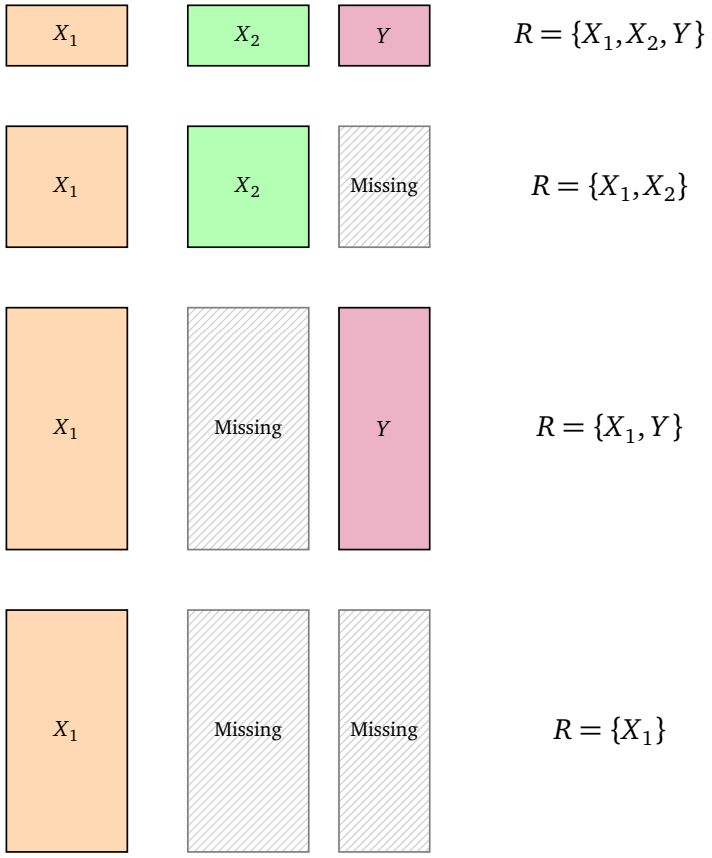

Figure 1: An example of multi-modal blockwise missing data.

inference in a semi-supervised setting. In its vanilla version, `PPI` employs AI models to predict outcomes in an unsupervised dataset, and then corrects the bias of AI models by adding a correction term computed over a supervised dataset. Statistical efficiency is improved when the AI models are accurate and the size of the unsupervised dataset is much larger than that of the supervised dataset. Note that the vanilla `PPI` does not guarantee more efficient estimation compared to the naive estimator. Follow-up works have addressed this issue by introducing tuning parameters (Angelopoulos et al., 2023b) or by exploiting efficient influence functions (Miao et al., 2023; Ji et al., 2025). We combine both these developments in our proposal. Our estimator is in fact motivated by semiparametric theory, but it further controls error-prone AI models by exploiting optimized tuning parameters to minimize asymptotic variance and safeguard efficiency. During the final stage of preparation of our work, Zhao and Candès (2025) and Chen et al. (2025) have proposed imputation-powered inference methods for blockwise missing data under MCAR and MAR assumptions, respectively. Their methods recalibrate each missing pattern using complete data – a strategy termed *pattern stratification* by Chen et al. (2025). We will show that this approach is less efficient than ours and may suffer from numerical instability in limited labeled dataset setting. A more detailed discussion can be found in Section 3.1.

Regression under blockwise missing data. Some recent papers have studied (generalized) linear regression coefficients estimation and variable selection considering blockwise missing data (Kundu et al., 2019; Yu et al., 2020; Xue and Qu, 2021; Xue et al., 2021; Jin and

Rothenhäusler, 2023; Song et al., 2024; Li et al., 2024; Diakité et al., 2025). Based on some specific properties of normal equations or moment conditions of GLM, most of these methods are restricted to regression coefficient estimation and inference. In contrast, our work offers a flexible framework for estimating a much broader class of parameters, which is usually referred to as Z-estimation (Van der Vaart, 2000). However, some of the ideas are related to our work. For example, the weighting strategy for each observed structure in Li et al. (2024) is equivalent to its counterpart in Zhao and Candès (2025); Chen et al. (2025). In addition, our `IBM`(RAY) and `IBM`(Adaptive) estimator resembles the multiple block imputation approach (Xue and Qu, 2021) in that multiple imputations or projections can be applied to the same missingness pattern to improve statistical efficiency. More details can be found in Section 3.2.

Semiparametric theory and non-monotone missing pattern. Classically, estimation and inference with missing data have been extensively studied in the semiparametric theory literature (Robins et al., 1994; Chen, 2004; Tsiatis, 2006). Notably, blockwise missingness is indeed another framing of non-monotone missingness pattern (Robins and Gill, 1997; Robins, 1997). Although the optimal influence function under non-monotone missing pattern is well studied theoretically (Robins et al., 1994; Tsiatis, 2006), its empirical implementation is often rather involved. In recent years, there have been several attempts to seek semiparametric efficient estimators under non-monotone missingness patterns, usually by introducing additional structural assumptions. For instance, Yang (2022) considers a sequential MAR (SMAR) assumption and propose a semiparametric efficient estimator if that assumption holds; Huang et al. (2025) claim that their estimator attains the semiparametric efficiency bound if each dataset either contains all modalities or only one modality, under MCAR assumption. Our approach, in contrast, exploits RAY approximation and assumes MCAR to effectively approximate the theoretically optimal estimating function for general non-monotone missing patterns, thus enlarging the span of applications of these class of methods.

1.2 Contributions

Our contributions can be summarized as follows. First, we show that the main challenge in obtaining the most efficient estimator under non-monotone missingness comes from the composition of conditional expectations. To overcome this, we propose the Restricted Anova HierarchY (RAY) approximation of the optimal estimating function. We show that RAY approximation yields a valid estimating function for the target parameter and leads to a computationally tractable estimator. In this way, our approach strikes a crucial balance between statistical efficiency and computational feasibility. To the best of our knowledge, the RAY approximation is the first general method to tackle this problem in a tractable way, representing a novel contribution to the semiparametric literature.

Second, we introduce a suite of **I**nference for **B**lockwise **M**issing (`IBM`) estimators that target parameters of interest when data present block missingness and leverage pre-trained AI models. A first estimator in this class is `IBM`(Pattern-stratification), or `IBM`(PS) estimator, that is obtained by adopting the weighting strategy proposed in Li et al. (2024), Chen et al. (2025), and Zhao and Candès (2025). A second estimator, the `IBM`(RAY) estimator, is derived from the proposed RAY approximation. Recognizing that neither estimator is uniformly superior,

we unify them into a broader class of estimators encompassing these two estimators as special cases. This class of estimators is parameterized by a vector of tuning parameters. Optimal tuning can be achieved by minimizing the asymptotic variance, leading to the most efficient estimator in this class, which we call the `IBM`(Adaptive) estimator.

1.3 Organization and notation

This paper is organized as follows. In Section 2, we first introduce a simple example, where we aim to estimate the outcome mean in a setting with two modalities. There, we introduce related methods and motivate our proposal. In Section 3, we then present our main methodology, which includes the RAY approximation, and the `IBM`(RAY) and `IBM`(Adaptive) estimators. Here, we also establish their asymptotic behavior. Section 4 provides simulation results that validate the theoretical properties and the finite-sample performance of our methods. In Section 5, we apply our method to a single-cell multi-omics dataset to estimate surface protein abundance levels. Finally, we conclude our paper by providing further potential research directions in Section 6. All technical details and proofs are provided in Appendix A.

Throughout the manuscript, we use the following notation. $[M]$ is used to denote the set $\{X_1, \cdots, X_M\}$, and the cardinality of a set Q is denoted by $|Q|$. Expectation and variance of random variables are denoted as $\mathbb{E}[\cdot]$, $\mathbb{V}[\cdot]$, respectively. The indicator function $\mathbb{I}(\cdot)$ is frequently used, and it returns 1 if the condition within the brackets is satisfied, and 0 otherwise.

2 Warm-up example: outcome mean estimation with two sets of covariates

We now introduce a simple setting that we exploit to revisit some existing methods and to motivate our proposal. Consider the problem of outcome mean estimation when data present block missingness as shown in Figure 1. Here, the prototypical full-data observation is $Z^F = (X_1, X_2, Y) \sim \mathscr{P}$, where $X_1 \in \mathbb{R}^{p_1}, X_2 \in \mathbb{R}^{p_2}$ and $Y \in \mathbb{R}$[1]. We are interested in the estimation of $\theta^\star = \mathbb{E}_{\mathscr{P}}[Y]$, and the observed-data sample is $\mathscr{D} = \cup_{r \in Q} \mathscr{D}_r$. For notational simplicity, we denote $\mathscr{D}_L = \mathscr{D}_{\{X_1, X_2, Y\}} \cup \mathscr{D}_{\{X_1, Y\}}$ as *labeled* dataset and $\mathscr{D}_U = \mathscr{D}_{\{X_1, X_2\}} \cup \mathscr{D}_{\{X_1\}}$ as *unlabeled* dataset. The naive estimator of $\theta^\star$ is the sample mean of the outcome in the labeled datasets, that is,

$$\hat{\theta}^{naive} = \frac{1}{N} \sum_{i=1}^{N} \frac{\mathbb{I}(R_i = \{X_1, X_2, Y\} \text{ or } \{X_1, Y\})}{\hat{\pi}_{\{X_1, X_2, Y\}} + \hat{\pi}_{\{X_1, Y\}}} Y_i,$$

whose variance is $\mathbb{V}(Y)/|\mathscr{D}_L|$. With `PPI` (Angelopoulos et al., 2023a), the authors employ a pre-trained AI model $f(X_1)$ to predict the outcome Y, and improve estimation efficiency using

[1] From now on, we use the superscript "F" to denote full-data as opposed to observed-data quantities.

the following estimator:

$$\begin{aligned}\hat{\theta}^{\texttt{PPI}} = &\frac{1}{N}\sum_{i=1}^{N}\frac{\mathbb{I}(R_i=\{X_1,X_2,Y\}\text{ or }\{X_1,Y\})}{\hat{\pi}_{\{X_1,X_2,Y\}}+\hat{\pi}_{\{X_1,Y\}}}Y_i \\ &+\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\mathbb{I}(R_i=\{X_1,X_2\}\text{ or }\{X_1\})}{\hat{\pi}_{\{X_1,X_2\}}+\hat{\pi}_{\{X_1\}}}-\frac{\mathbb{I}(R_i=\{X_1,X_2,Y\}\text{ or }\{X_1,Y\})}{\hat{\pi}_{\{X_1,X_2,Y\}}+\hat{\pi}_{\{X_1,Y\}}}\right)f(X_{i,1}).\end{aligned} \tag{1}$$

The expectation of the augmentation term is zero, so that $\hat{\theta}^{\texttt{PPI}}$ is an unbiased estimator of $\theta^\star$. If $|\mathscr{D}_U| \gg |\mathscr{D}_L|$, and $f(\cdot)$ is an accurate prediction model of Y, the variance of $\hat{\theta}^{\texttt{PPI}} = \mathbb{V}(f)/|\mathscr{D}_U| + \mathbb{V}(Y-f)/|\mathscr{D}_L|$ is dominated by $\mathbb{V}(Y-f)/|\mathscr{D}_L|$ which can be smaller than $\mathbb{V}(Y)/|\mathscr{D}_L|$, the variance of the naive estimator, thus achieving efficiency gains by leveraging $f(\cdot)$. At the same time, (1) can be less efficient than the naive estimator when the AI model $f(\cdot)$ is error-prone. Therefore, follow-up works (Angelopoulos et al., 2023b; Miao et al., 2023; Gronsbell et al., 2024; Ji et al., 2025) have introduced an additional tuning parameter, which depends on $f(\cdot)$, to safeguard estimation efficiency. For example, the PPI++ (Angelopoulos et al., 2023b) estimator is defined as

$$\begin{aligned}\hat{\theta}^{\texttt{PPI++}} = &\frac{1}{N}\sum_{i=1}^{N}\frac{\mathbb{I}(R_i=\{X_1,X_2,Y\}\text{ or }\{X_1,Y\})}{\hat{\pi}_{\{X_1,X_2,Y\}}+\hat{\pi}_{\{X_1,Y\}}}Y_i \\ &+\frac{\alpha_f}{N}\sum_{i=1}^{N}\left(\frac{\mathbb{I}(R_i=\{X_1,X_2\}\text{ or }\{X_1\})}{\hat{\pi}_{\{X_1,X_2\}}+\hat{\pi}_{\{X_1\}}}-\frac{\mathbb{I}(R_i=\{X_1,X_2,Y\}\text{ or }\{X_1,Y\})}{\hat{\pi}_{\{X_1,X_2,Y\}}+\hat{\pi}_{\{X_1,Y\}}}\right)f(X_{i,1}),\end{aligned} \tag{2}$$

where the optimal α_f can be chosen by minimizing asymptotic variance. As a result, $\hat{\theta}^{\texttt{PPI++}}$ is never less efficient than $\hat{\theta}^{\texttt{PPI}}$ and $\hat{\theta}^{naive}$.

While these methods succeed in improving statistical efficiency with a pre-trained model in a semi-supervised setting, it is intriguing to understand whether, and eventually how, additional covariates X_2, accompanied by additional AI models, can further improve efficiency. While in the final stage of preparation of our manuscript, Zhao and Candès (2025); Chen et al. (2025) have considered the same problem under MCAR and MAR assumptions, respectively. Instantiating their method to outcome mean estimation results in the following estimator:

$$\begin{aligned}\hat{\theta}^{\texttt{IBM}}_{\texttt{PS}} = &\frac{1}{N}\sum_{i=1}^{N}\frac{\mathbb{I}(R_i=\{X_1,X_2,Y\}\text{ or }\{X_1,Y\})}{\hat{\pi}_{\{X_1,X_2,Y\}}+\hat{\pi}_{\{X_1,Y\}}}Y_i \\ &+\frac{\alpha_f}{N}\sum_{i=1}^{N}\left(\frac{\mathbb{I}(R_i=\{X_1\})}{\hat{\pi}_{\{X_1\}}}-\frac{\mathbb{I}(R_i=\{X_1,X_2,Y\})}{\hat{\pi}_{\{X_1,X_2,Y\}}}\right)f(X_{i,1}) \\ &+\frac{\alpha_g}{N}\sum_{i=1}^{N}\left(\frac{\mathbb{I}(R_i=\{X_1,X_2\})}{\hat{\pi}_{\{X_1,X_2\}}}-\frac{\mathbb{I}(R=\{X_1,X_2,Y\})}{\hat{\pi}_{\{X_1,X_2,Y\}}}\right)g(X_{i,1},X_{i,2}),\end{aligned} \tag{3}$$

where $g(X_1,X_2)$ is another AI model that uses both X_1 and X_2 as inputs. Again, α_f and α_g are tuning parameters that safeguard estimation efficiency. Hereafter, we denote this estimator as IBM(Pattern-stratification), or briefly IBM(PS), where the terminology "Pattern-stratification" is borrowed from Chen et al. (2025), and refers to the weighting strategy for $f(\cdot)$ and $g(\cdot)$. The general form of this estimator will be presented in Section 3.1.

Although PPI++ and IBM(PS) are guaranteed to be more efficient than the naive estimator, it is unknown whether they can achieve the lowest asymptotic variance given pre-trained $f(\cdot)$ and $g(\cdot,\cdot)$. In the following Section, we find that the semiparametric efficiency bound is not generally attained even if $f(X_1)=\mathbb{E}[Y\mid X_1]$ and $g(X_1,X_2)=\mathbb{E}[Y\mid X_1,X_2]$ are correctly specified. Therefore, we introduce an estimator motivated by a novel RAY approximation of the optimal estimating function, named as IBM(RAY) estimator. Later, we also present a unified class of estimators, including IBM(RAY) and IBM(PS) estimators as special cases, and propose an adaptive estimator achieving the lowest asymptotic variance within this class.

3 Inference for blockwise missingness

3.1 RAY approximation and IBM(RAY) estimator

Motivation from semiparametric theory. The estimation of parameters under non-monotone missingness has been extensively studied in semiparametric theory (Robins et al., 1994; Tsiatis, 2006). Formally, the optimal estimating function achieving semiparametric efficiency is characterized by the following Theorem.

Theorem 3.1. *For a fixed full-data estimating function $\psi^F(Z,\theta^\star)\in\Lambda^{F\perp}$ for estimating $\theta^\star=\theta(\mathscr{P})$, where $\Lambda^{F\perp}$ is the orthogonal complement of nuisance tangent space of $\mathscr{P}_{\theta,\eta}$, the optimal observed-data estimating function is given by*

$$\psi(Z,\theta^\star)=\mathscr{L}\{\mathscr{M}^{-1}[\psi^F(Z,\theta^\star)]\}, \tag{4}$$

where $\mathscr{L}(\cdot)=\sum_{r\in Q}\mathbb{I}(R=r)\mathbb{E}[\cdot\mid X_r]$ and $\mathscr{M}(\cdot)=\sum_{r\in Q}\pi_r\mathbb{E}[\cdot\mid X_r]$. In addition, $\mathscr{M}^{-1}(\cdot)$ exists and is uniquely defined as $\mathscr{M}^{-1}=\sum_{k=0}^{\infty}(\mathscr{I}-\mathscr{M})^k$, where $\mathscr{I}$ is the identity operator.

Theorem 3.1 can be proved using Theorems 10.1, 10.6 and Lemma 10.5 in Tsiatis (2006). Although the optimal estimating function above is contained in a neat formula, its practical implementation is often infeasible. In particular, $\mathscr{M}^{-1}[\psi^F(Z,\theta^\star)]$ does not often have a closed-form expression for general distribution $\mathscr{P}$ and $\psi^F(Z,\theta)$. Instead, it can be approximated by letting $\psi_0(Z,\theta^\star)=\psi^F(Z,\theta^\star)$ and computing $\psi_{k+1}(Z,\theta^\star)=(\mathscr{I}-\mathscr{M})(\psi_k(Z,\theta^\star))+\psi^F(Z,\theta^\star)$ iteratively. This approximation requires composition and interaction of projection operators, which can be rather sophisticated without special structural assumptions.

RAY approximation. Motivated by (4), we further investigate the properties of $\mathscr{M}$, aiming to find an approximation of $\mathscr{L}(\mathscr{M}^{-1})$ that balances computational feasibility and statistical efficiency. For ease of notation, we define $A_r(\cdot)=\mathbb{E}[\cdot\mid X_r]$ as the conditional expectation operator given X_r for any $f\in\sigma(\mathscr{X})$, where $\sigma(\mathscr{X})$ denotes the sigma-algebra generated by $\mathscr{X}$. We propose the following *Restricted Anova hierarchY* (RAY) decomposition.

Lemma 3.2 (RAY decomposition)**.** *For any $f\in\sigma(\mathscr{X})$, f can be decomposed as:*

$$f=\sum_{s\subseteq[M]}P_s(f),\quad P_s(f)=\sum_{r\subseteq s,r\in Q}(-1)^{|s|-|r|}A_r(f),\quad P_\emptyset=A_\emptyset=0. \tag{5}$$

In the above decomposition, $P_s(f)$ can be interpreted as the projection of f onto X_s adjusting for the lower-order effects. The proof can be found in Appendix A.1. Our RAY decomposition is

conceptually similar to the functional ANOVA decomposition (Hooker, 2004; Owen, 2013) and the Hoeffding-Sobol decomposition (Hoeffding, 1948), in that all follow from the inclusion-exclusion principle. However, there are two relevant distinctions: first, $P_s(f)$ is defined as the sum of the signed projections that are *restricted* in Q, and not over all subsets of $[M]$; second, unlike the classical Hoeffding-Sobol decomposition, the RAY decomposition allows for dependence among variables, which breaks orthogonality among the $P_s(f)$ terms. Despite these differences, the RAY decomposition still enjoys nice properties that can be exploited to approximate $\mathscr{M}(\cdot)$. In particular, for any $s \subseteq [M]$, the following holds:

$$\begin{aligned}
\mathscr{M}[P_s(f)] &= \sum_{r\in Q} \pi_r A_r[P_s(f)] \\
&= \sum_{r\in Q, s\subseteq r} \pi_r A_r[P_s(f)] + \sum_{r\in Q, s\not\subseteq r} \pi_r A_r[P_s(f)] \\
&= \left(\sum_{r\in Q, s\subseteq r} \pi_r\right) P_s(f) + \sum_{r\in Q, s\not\subseteq r} \pi_r A_r[P_s(f)] \\
&\triangleq \lambda_s P_s(f) + \mathrm{Rem}_s(f).
\end{aligned} \tag{6}$$

The third equality holds because $P_s(f)$ is measurable with respect to the sigma-algebra generated by the set r since $s \subseteq r$. The second term in the third equality involves the composition of conditional expectations that cannot be simplified for general distribution $\mathscr{P}$, and we denote it by Rem_s hereafter. Again, for a general distribution $\mathscr{P}$, orthogonality between $P_s(f)$ and Rem_s does not hold.

The component P_s can be viewed as an almost-eigen-operator[2] of $\mathscr{M}$ with eigenvalue λ_s in (6). The eigenvalue λ_s is the proportion of datasets containing the modalities in s. According to Assumption 1.1 on positivity, $\lambda_s \geq \pi_{[M]} > c > 0$ for all subsets $s \subseteq [M]$. The remainder terms Rem_S's contain compositions of projections – in the example in Section 2, the remainder contains $A_{X_1,X_2}[A_{X_1,Y}(f)]$, which represents a major challenge in solving the optimal estimating equation (4). Furthermore, since $\mathscr{M}$ is a linear operator, its inverse $\mathscr{M}^{-1}$ is also linear, leading to the following property:

$$\begin{aligned}
\mathscr{M}^{-1}[P_s(f)] &= \frac{1}{\lambda_s} P_s(f) - \frac{1}{\lambda_s} \mathscr{M}^{-1}[\mathrm{Rem}_s(f)], \\
\mathscr{M}^{-1}(f) &= \sum_{s\subseteq[M]} \frac{1}{\lambda_s} P_s(f) - \sum_{s\subseteq[M]} \frac{1}{\lambda_s} \mathscr{M}^{-1}[\mathrm{Rem}_s(f)].
\end{aligned} \tag{7}$$

Based on the above derivation, we can readily separate $\mathscr{M}^{-1}$ into two terms: a tractable term including projections onto subsets of observed patterns, and an intractable term involving composition of projections.

Remark 3.3. Another property of the components $P_s(f)$'s is that $\mathscr{M}(f) = \sum_{s\subseteq[M]} \lambda_s P_s(f)$, or equivalently, $\sum_{s\subseteq[M]} \mathrm{Rem}_s(f) = 0$, which implies that the remainder terms can be ignored as a whole to represent $\mathscr{M}$ with P_s. However, since we are interested in $\mathscr{M}^{-1}$, it is still crucial to inspect the impact of each remainder term. Here, we consider a special example

[2]Similar concepts appear in several areas of mathematics, for example, almost-eigenvector in random regular graph (BACKHAUSZ and SZEGEDY, 2019).

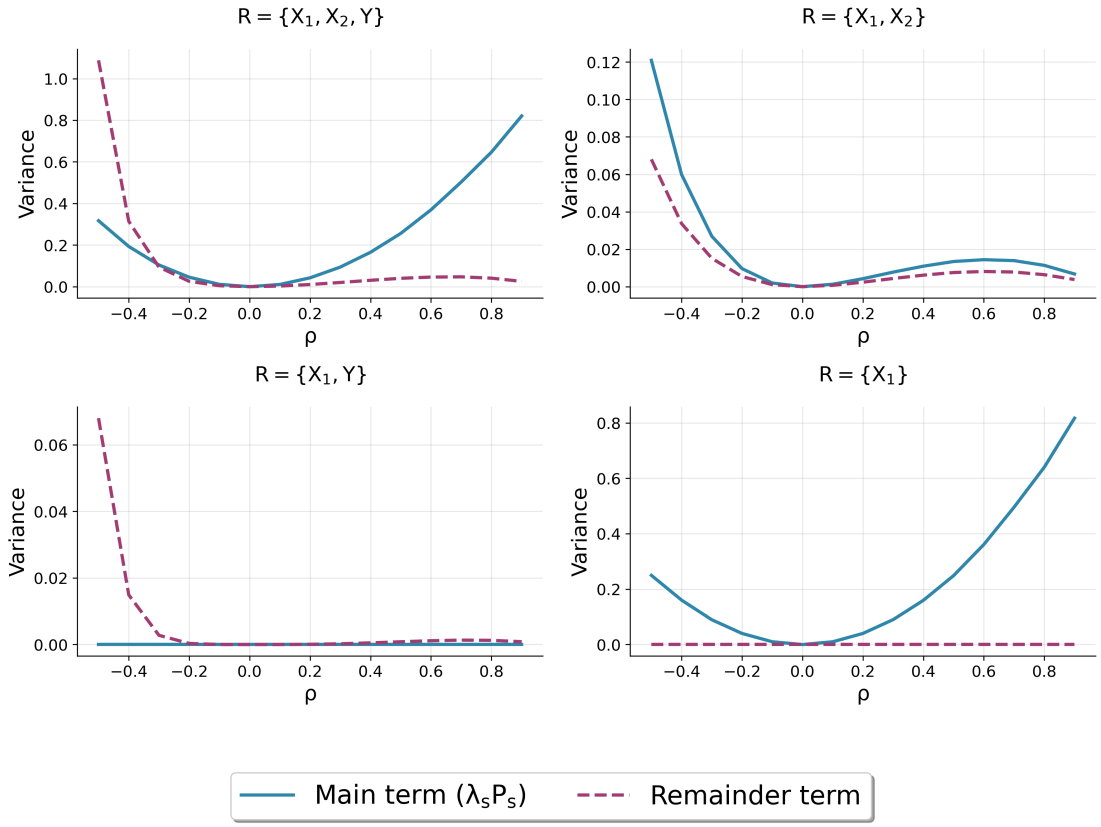

Figure 2: Comparison of the estimated variance between main term ($\lambda_s P_s$) and remainder terms in the example of Section 2 when variables (X_1, X_2, Y) come from a multivariate Gaussian distribution with various degrees of correlation ρ. Details on the simulation are described in Remark 3.3.

related to the observed patterns in Section 2, and let (X_1, X_2, Y) follow a three-dimensional Gaussian distribution with zero mean and exchangeable covariance matrix Σ where diagonal entries are 1 and off-diagonal entries ρ's can range in $(-0.5, 0.9)$. We let $f = Y$, so that the expectations of $\lambda_s P_s(f)$ and $\mathrm{Rem}_s(f)$ are all zeros. Then we calculate the variance of $\lambda_s P_s(f)$ and $\mathrm{Rem}_s(f)$ from samples generated from the above distribution and visualize them in Figure 2. It appears clear that the remainder terms are negligible compared to the main terms $\lambda_s P_s$ *if* the correlations among the variables are positive, which in turn suggests that the approximation of $\mathscr{M}$ by the $\lambda_s P_s$'s is justifiable. However, the remainder terms can dominate when correlations are very negative, indicating that ignoring remainder terms would cause a large efficiency loss compared to the semiparametric efficiency lower bound.

According to (7), we can write the optimal estimating function (4) given $\psi^F(Z,\theta)$ as

$$\begin{aligned}\mathscr{L}\{\mathscr{M}^{-1}[\psi^F(Z,\theta^\star)]\} &= \sum_{r\in Q}\mathbb{I}(R=r)A_r\{\mathscr{M}^{-1}[\psi^F(Z,\theta^\star)]\}\\ &= \sum_{r\in Q}\mathbb{I}(R=r)A_r\left\{\sum_{s\subseteq r}\frac{1}{\lambda_s}P_s[\psi^F(Z,\theta^\star)]\right\}\\ &+ \sum_{r\in Q}\mathbb{I}(R=r)A_r\left\{\sum_{s\nsubseteq r, s\subseteq[M]}\frac{1}{\lambda_s}P_s[\psi^F(Z,\theta^\star)]\right\}\\ &+ \sum_{r\in Q}\mathbb{I}(R=r)A_r\left\{\sum_{s\subseteq[M]}\frac{1}{\lambda_s}\mathscr{M}^{-1}\{\mathrm{Rem}_s[\psi^F(Z,\theta^\star)]\}\right\},\end{aligned} \tag{8}$$

where the first term can be further simplified as in (6), while the second and the third terms include composition of projection terms which are computationally intractable.

Remark 3.4. The second and third terms in (8) can be equal to zero under certain conditions. For the example in Section 2, we can write out the remainder terms in (6) as follows:

- $\mathrm{Rem}_{\{X_1,X_2,Y\}}(f) = -\left(\pi_{\{X_1,X_2\}}+\pi_{\{X_1,Y\}}\right)\left\{A_{\{X_1,X_2\}}[A_{\{X_1,Y\}}(f)]+A_{\{X_1,Y\}}[A_{\{X_1,X_2\}}(f)]-2A_{\{X_1\}}(f)\right\}$
- $\mathrm{Rem}_{\{X_1,X_2\}}(f) = \pi_{\{X_1,Y\}}\left\{A_{\{X_1,X_2\}}[A_{\{X_1,Y\}}(f)]-A_{\{X_1\}}(f)\right\}$
- $\mathrm{Rem}_{\{X_1,Y\}}(f) = \pi_{\{X_1,X_2\}}\left\{A_{\{X_1,Y\}}[A_{\{X_1,X_2\}}(f)]-A_{\{X_1\}}(f)\right\}$

with all the other remainder terms being equal to zero. If X_2 and Y are conditionally independent given X_1, that is, $X_2 \perp\!\!\!\perp Y \mid X_1$, we have $A_{\{X_1\}}(f) = A_{\{X_1,X_2\}}[A_{\{X_1,Y\}}(f)] = A_{\{X_1,Y\}}[A_{\{X_1,X_2\}}(f)]$, which implies that all remainder terms are zero (Figure 2 also validates this claim when conditional independence holds, that is when $\rho = 0$). Similarly, under the same conditional independence assumption, one can verify that the second term in (8) is also zero. In this case, the semiparametric efficiency bound is empirically attainable with projections onto observed patterns. In other words, conditional independence implies that X_2 does not provide additional predictive power for Y, so that there is no efficiency gain by incorporating X_2 in the prediction model. This degenerates the problem into the typical semi-supervised setting with features X_1 and Y, so that the semiparametric efficiency bound can be attained.

For ease of computation, we ignore the second and the third terms in (8), which involve composition of projections, and we consider the following estimating function, that we refer to as *RAY approximation* afterward:

$$\begin{aligned}\psi(Z,\theta^\star) &= \sum_{r\in Q}\mathbb{I}(R=r)\sum_{s\subseteq r}\frac{1}{\lambda_s}P_s[\psi^F(Z,\theta^\star)]\\ &= \frac{\mathbb{I}(R=[M])}{\pi_{[M]}}\psi^F(Z,\theta^\star)+\sum_{r\in Q, r\neq[M]}\omega_r A_r[\psi^F(Z,\theta^\star)],\end{aligned} \tag{9}$$

where the second equality highlights the role played by each projection according to the different observed patterns. The random variable ω_r, which we call *RAY random variable*, is

defined as:

$$\omega_r = \sum_{s \supseteq r} (-1)^{|s|-|r|} \frac{\mathbb{I}(R \supseteq s)}{\lambda_s}. \tag{10}$$

In Appendix A.2, we show that $\mathbb{E}[\omega_r] = 0$ for $r \in Q$, $r \neq [M]$. This leads us to the following result.

Proposition 3.5. *The function given in* (9) *is a valid estimating function in the sense that*

$$\mathbb{E}[\psi(Z, \theta^\star)] = 0$$

is a valid estimating equation for the target $\theta^\star$. The expectation is taken over the observed-data distribution induced by $\mathscr{P}$ and the distribution of R following Assumption 1.1.

Accordingly, we are able to provide an estimator of $\theta^\star$ with all data in $\mathscr{D}$ by solving the following estimating equation:

$$\frac{1}{N}\sum_{i=1}^{N} \frac{\mathbb{I}(R_i = [M])}{\pi_{[M]}} \psi^F(Z_i, \theta) + \frac{1}{N}\sum_{i=1}^{N} \sum_{r \in q, r \neq [M]} \omega_r \mathbb{E}[\psi^F(Z_i, \theta) \mid X_r] = 0. \tag{11}$$

The semiparametric estimator obtained by solving (11) is a *multiply robust* estimator in the sense that it is consistent and asymptotic normal even if any conditional expectation model $\mathbb{E}[\psi^F(Z, \theta) \mid X_r]$ is misspecified, as long as π_r can be consistently estimated, for which sample proportions suffice under Assumption 1.1. The formal statement of this property is omitted for brevity. In practical applications, if no pre-trained AI model is available, we can follow standard semiparametric procedures that fit flexible machine learning algorithms to estimate nuisance parameters $\mathbb{E}[\psi^F(Z, \theta) \mid X_r]$ using cross-fitting in order to avoid overfitting bias (Chernozhukov et al., 2018).

To summarize, we have analyzed the properties of $\mathscr{M}$ through RAY decomposition, which revealed that the semiparametric efficiency bound can be attained only in very special cases (see Remark 3.4). In addition, we have proposed a semiparametric estimator based on the RAY approximation of the optimal estimating function (4). We are now ready to plug pre-trained AI models into (11) as to boost statistical efficiency.

Blockwise missing prediction-powered inference. We now show how to incorporate pre-trained AI models to improve estimation based on (11). In essence, pre-trained AI models can be used to proxy nuisance parameters $\mathbb{E}[\psi^F(Z_i, \theta^\star) \mid X_r]$. For instance, in the warm-up example in Section 2, if we choose $\psi^F(Z, \theta^\star) = Y - \theta^\star$ as full-data estimating function for the target parameter $\theta^\star = \mathbb{E}[Y]$, then $f(X_1) - \theta^\star$ replaces $\mathbb{E}[Y - \theta^\star \mid X_1]$ and $g(X_1, X_2) - \theta^\star$ replaces $\mathbb{E}[Y - \theta^\star \mid X_1, X_2]$. Similarly, if our target is linear regression coefficients, the full-data estimating function is $\psi^F(Z) = X(Y - X^T\theta^\star)$, so that

$$\mathbb{E}[\psi^F(Z, \theta^\star) \mid X_1] = \begin{pmatrix} X_1\mathbb{E}[Y \mid X_1] - X_1X_1^T\theta_1^\star - X_1\mathbb{E}[X_2^T \mid X_1]\theta_2^\star \\ \mathbb{E}[X_2Y \mid X_1] - \mathbb{E}[X_2 \mid X_1]X_1^T\theta_1^\star - \mathbb{E}[X_2X_2^T \mid X_1]\theta_2^\star. \end{pmatrix}. \tag{12}$$

In this case, we need multiple AI models for the various conditional expectations. If we have access to conditional generative models $\mu(W; T = t)$ that generate samples of W from the

conditional distribution $W \mid T = t$ for any two random variables W, T, then we are able to estimate the moments in (12) by Monte Carlo method (Metropolis and Ulam, 1949). We refer to this approach as "expectation" mode. However, in practice, some required AI models are likely inaccessible. Therefore, we can follow Chen et al. (2025); Zhao and Candès (2025) and approximate $\mathbb{E}[\psi^F(Z, \theta^\star) \mid X_r]$ using $\psi(\mathbb{E}[Z \mid X_r], \theta^\star)$. With this approach, which we term "imputation" mode, only one AI model $\mathbb{E}[Z \mid X_r]$ is necessary. In classical semiparametric estimation when nuisance parameters are estimated from collected data, the "expectation" mode leads, in theory, to more efficient estimators. In contrast, when we deploy potentially biased pre-trained AI models, it is unclear which among "expectation" and "imputation" mode will dominate. In practice, we suggest that practitioners adopt "imputation" mode if only essential AI models are accessible and "expectation" mode if all required AI models are available. In any case, regardless of the choice between "expectation" and "imputation" modes and the quality of pre-trained AI models, our estimation based on (11) is still consistent and asymptotically normal due to the multiple-robustness property detailed above. In the following, to streamline our results for both modes, we use $\mathscr{F}_r$ to denote either $\mathbb{E}[\psi^F(Z, \theta^\star) \mid X_r]$ or $\psi^F(\mathbb{E}[Z \mid X_r], \theta^\star)$.

To account for the potential bias in the pre-trained AI models, we introduce the tuning parameter α_r for each $\mathscr{F}_r$ and define:

$$\psi_\alpha(Z; \theta^\star) = \frac{\mathbb{I}(R = [M])}{\pi_{[M]}} \psi^F(Z, \theta^\star) + \sum_{r \in Q, r \neq [M]} \alpha_r \omega_r \mathscr{F}_r(X_r), \tag{13}$$

where $\alpha = \{\alpha_r : r \in Q, r \neq [M]\}$. We can show that even (13) is a valid estimating function.

Proposition 3.6. *For any α, the function given in (13) is a valid estimating function in the sense that*

$$\mathbb{E}[\psi_\alpha(Z, \theta^\star)] = 0$$

is a valid estimating equation for the target $\theta^\star$. The expectation is taken over the observed-data distribution induced by $\mathscr{P}$ and the distribution of R following Assumption 1.1.

Now we can formally define the Inference for Blockwise Missingness (RAY) or IBM(RAY) estimator, which we denote $\hat{\theta}^{\texttt{IBM}}_{\texttt{RAY}}$, as the solution of estimating equation:

$$\frac{1}{N} \sum_{i=1}^{N} \frac{\mathbb{I}(R_i = [M])}{\pi_{[M]}} \psi^F(Z_i, \theta) + \frac{1}{N} \sum_{i=1}^{N} \sum_{r \in Q, r \neq [M]} \alpha_r \omega_r \mathscr{F}_r(X_{i,r}) = 0\,. \tag{14}$$

Remark 3.7. The IBM(PS) estimator proposed by Chen et al. (2025); Zhao and Candès (2025) and instantiated before in (3) can be derived as the solution to (14) with $\omega_r = \mathbb{I}(R = r)/\pi_r - \mathbb{I}(R = [M])/\pi_{[M]}$. Therefore, consistency and asymptotic normality of both the IBM(PS) and the IBM(RAY) estimators can be established under the same framework.

In the following, we analyze the asymptotic behaviors of the IBM(RAY) and IBM(PS) estimators given pre-trained AI models and α.

Theorem 3.8. *Define* $A = \mathbb{E}[\partial \psi^F / \partial \theta^T]$, $\gamma_r = \mathbb{E}[\omega_r \mathbb{I}(R = [M])/\pi_{[M]}]$, *and* $\eta_{r,r'} = \mathbb{E}[\omega_r \omega_{r'}]$. *Then, under Assumption 1.1 and suitable regularity conditions detailed in Appendix A.3, we have for any* α

(a) (Consistency) $\hat{\theta}^{\texttt{IBM}}_{\texttt{RAY/PS}} \xrightarrow{p} \theta^\star$;

(b) (Asymptotic normality) $\sqrt{N}(\hat{\theta}^{\texttt{IBM}}_{\texttt{RAY/PS}} - \theta^\star) \rightsquigarrow \mathcal{N}(\mathbf{0}_d, \Sigma_\alpha)$, *where*

$$\Sigma_\alpha = A^{-1}\frac{\mathbb{V}(\psi^F)}{\pi_{[M]}}A^{-1} + A^{-1}\Bigg\{\sum_{r\in Q, r\neq[M]} \alpha_r\gamma_r[Cov(\psi^F, \mathscr{F}_r) + Cov(\mathscr{F}_r, \psi^F)] + \sum_{r,r'\in Q, r,r'\neq[M]} \alpha_r\alpha_{r'}\eta_{r,r'}Cov(\mathscr{F}_r, \mathscr{F}_{r'})\Bigg\}A^{-1}. \tag{15}$$

The first term in (15) corresponds to the asymptotic variance of the naive estimator that estimates $\theta^\star$ using only complete data. The second term is due to the contribution of the pre-trained AI models under each observed pattern r. In principle, if this second term is large, Σ_α can be worse than the variance of the naive estimator. Therefore, we propose an optimization scheme for the tuning parameters α_r's, which minimizes a linear criterion ℓ of the asymptotic variance matrix. For example, we can minimize the trace of asymptotic variance matrix $\ell(\Sigma_\alpha) = \mathrm{Tr}(\Sigma_\alpha)$, or, if we are interested in estimating a linear contrast $c^T\theta^\star$, we can minimize $\ell(\Sigma_\alpha) = c^T\Sigma_\alpha c$. Since $\ell(\Sigma_\alpha)$ is a quadratic form of α, we have the following result.

Corollary 3.9. *The optimal* $\alpha^\star$ *that minimizes a linear scalarization of the covariance matrix* $\ell(\Sigma_\alpha)$ *is given by*

$$\alpha^\star = -G^{-1}L\,, \tag{16}$$

where $G = (\eta_{r,r'}\ell(A^{-1}Cov(\mathscr{F}_r, \mathscr{F}_{r'})A^{-1}))_{r,r'\in Q, r,r'\neq[M]}$, *and* $L = (\gamma_r\ell(A^{-1}Cov(\psi^F, \mathscr{F}_r)A^{-1}))_{r\in Q, r\neq[M]}$. *Under mild conditions on* $\mathscr{F}_r$*'s detailed in Appendix A.4, the matrix* G *is positive definite, and the solution* $\alpha^\star$ *detailed above is unique. The resulting* $\ell(\Sigma_{\alpha^\star})$ *is*

$$\ell(\Sigma_{\alpha^\star}) = \frac{\ell(A^{-1}\mathbb{V}(\psi^F)A^{-1})}{\pi_{[M]}} - L^TG^{-1}L\,,$$

where $L^TG^{-1}L$ *indicates the efficiency gain over the naive estimator.*

Based on Theorem 3.8 and Corollary 3.9, we propose Algorithm 1 to estimate $\theta^\star$. Given that the procedure to estimate $\alpha^\star$ in (16) is data-dependent, Algorithm 1 uses cross-fitting to avoid overfitting. In particular, we split the data so that an estimate for $\alpha^\star$ and an estimate for $\theta^\star$ are derived from different independent samples.

Algorithm 1 Inference for Blockwise Missing (IBM) algorithm

Step 1: Randomly split dataset $\mathscr{D}$ into two folds: $\mathscr{D}^{(1)}$ and $\mathscr{D}^{(2)}$.
Step 2: On $\mathscr{D}^{(k)}$, compute the covariance matrix $\hat{G}^{(k)}$, $\hat{L}^{(k)}$, and obtain $\hat{\alpha}^{(k)}$ by (16) for $k = 1, 2$.
Step 3: Solve estimating equation (14) on $\mathscr{D}^{(k')}$ with plugged-in $\hat{\alpha}^{(k)}$, obtain $\hat{\theta}^{(k')}$ and $\hat{\Sigma}^{(k')}$, for $k', k \in \{1, 2\}$ and $k \neq k'$.
Step 4: Compute the final estimate: $\hat{\theta}^{\texttt{IBM}} = (\hat{\theta}^{(1)} + \hat{\theta}^{(2)})/2$ with variance $(\hat{\Sigma}^{(1)} + \hat{\Sigma}^{(2)})/4$.

Remark 3.10. In some applications, the sample size of the complete dataset may be far less than some other datasets, i.e., $\pi_{[M]} \ll \pi_r$ for some r (Testa et al., 2025b; Zhang et al., 2023). This may raise concerns about numerical stability. The matrix G is the Hadamard product between $(\eta_{r,r'})_{r,r' \in Q, r,r' \neq [M]}$ and $(\ell(A^{-1}\mathrm{Cov}(\mathscr{F}_r, \mathscr{F}_{r'})A^{-1}))_{r,r' \in Q, r,r' \neq [M]\}}$. According to the Schur product theorem (Horn and Johnson, 2012), G is positive definite if these two matrices are positive definite, and the eigenvalues of G are functions of the eigenvalues of these two matrices. If $\omega_r = \mathbb{I}(R = r)/\pi_r - \mathbb{I}(R = [M])/\pi_{[M]}$ as in the IBM(PS) estimator (Chen et al., 2025; Zhao and Candès, 2025), then $\eta_{r,r'} = \mathbb{E}[\omega_r \omega_{r'}] = \mathbb{I}(r = r')/\pi_r + 1/\pi_{[M]}$. In this case, $1/\pi_{[M]}$ dominates $\eta_{r,r'}$, so that the smallest eigenvalue of $(\eta_{r,r'})_{r,r' \in \{2,\cdots,Q\}}$ approaches zero. As a result, G is close to be rank deficit, leading IBM(PS) to be numerically unstable and to perform potentially worse than the naive estimator empirically. In stark contrast, G is well-behaved if ω_r is taken as RAY random variable, because $\pi_{[M]}$ appears in the denominator of each term of $\eta_{r,r'}$. Later in the paper, we will explore and confirm this insight through simulations, e.g. in Figure 4.

3.2 IBM(Adaptive) estimator

The previous analysis on the two IBM estimators raises the practical question of how to choose between them in real-world applications. In fact, the performance of these two estimators depends on several factors, including both the quality of pre-trained AI models and the proportion of each observed patterns. Unfortunately, there is no conclusive answer as to which estimator is more efficient – see the numerical experiment in Figure 4 for a simulated example, and see Appendix A.6 for an analytical comparison. Therefore, instead of detailing the conditions under which one estimator can be more efficient than the other, we extend our analysis by considering an entire class of estimators that encompasses these two estimators as special cases. Then, we pursue the optimal estimator in this class that achieves the lowest asymptotic variance.

Compared with the IBM(PS) estimator, the IBM(RAY) estimator applies pre-trained AI models to all datasets where required sets of features are observed. For example, in the warm-up example, the IBM(RAY) estimator applies $f(X_1)$ to the entire dataset $\mathscr{D}$ instead of just $\mathscr{D}_{\{X_1,X_2,Y\}}$ and $\mathscr{D}_{\{X_1\}}$ as in IBM(PS). A similar idea has been considered in Xue and Qu (2021), where multiple block imputations are implemented to improve estimation efficiency. Motivated by this idea, we consider the following estimating function:

$$\psi_\alpha(Z, \theta^\star) = \frac{\mathbb{I}(R = [M])}{\pi_{[M]}} \psi^F(Z, \theta^\star) + \sum_{r \in Q, r \neq [M]} \left[\sum_{s \supseteq r} \frac{\mathbb{I}(R = s)}{\pi_s} \alpha_{r,s} \right] \mathscr{F}_r(X_r), \tag{17}$$

where $\alpha = \{\alpha_{r,s} \in \mathbb{R}, r \in Q, r \neq [M], s \supseteq r\}$ are tuning parameters satisfying

$$\sum_{s \supseteq r} \alpha_{r,s} = 0, \quad \forall r \in Q, r \neq [M]. \tag{18}$$

With this estimating function, each AI model $\mathscr{F}_r$ is applied to $\mathscr{D}_s$ if $s \supseteq r$ in order to empower the estimation efficiency with all available datasets. Under the constraints given by (18), it is easy to show that the estimating function (17) is a valid estimating function as in Proposition 3.6. In addition, we can show that the estimating functions of the IBM(PS) and IBM(RAY) estimators

are special cases of (17). In Appendix A.5, we characterize these two estimators in the framework of (17), confirming that the constraints in (18) are satisfied.

Similarly to Theorem 3.8 and Corollary 3.9, we can analyze the asymptotic behavior of the estimator obtained by solving the estimating equation associated with (17) and optimize a linear scalarization of the asymptotic variance matrix with respect to α under the constraints in (18), which is a quadratic programming problem and can be easily solved. We refer this estimator to as IBM(Adaptive) hereafter to emphasize its adaptation to the quality of the pre-trained AI models and to the proportion of each observed pattern, in order to achieve the optimal efficiency among the class of estimators defined in (17) and (18).

Theorem 3.11. *Define* $A = \mathbb{E}[\partial \psi^F / \partial \theta^T]$. *Under Assumption 1.1 and suitable regularity conditions as in Theorem 3.8, we have for any* α *satisfying (18)*

(a) (Consistency) $\hat{\theta}_{\texttt{ad}}^{\texttt{IBM}} \xrightarrow{p} \theta^\star$;

(b) (Asymptotic normality) $\sqrt{N}(\hat{\theta}_{ad}^{\texttt{IBM}} - \theta^\star) \rightsquigarrow \mathcal{N}(\mathbf{0}_d, \Sigma_\alpha)$, *where*

$$\Sigma_\alpha = A^{-1}\frac{\mathbb{V}(\psi^F)}{\pi_{[M]}}A^{-1} + \sum_{r\in Q, r\neq[M]} \frac{1}{\pi_{[M]}}\alpha_{r,[M]}\Big[Cov(\psi^F, \mathscr{F}_r) + Cov(\psi^F, \mathscr{F}_r)^T\Big] + \sum_{r\in Q, r\neq[M]} \frac{1}{\pi_r}A^{-1}\mathbb{V}\Big[\sum_{s\subseteq r}\alpha_{s,r}\mathscr{F}_s\Big]A^{-1}.$$

For any given linear scalarization function $\ell(\cdot)$ (e.g., trace or linear contrast), the optimal α can be obtained by minimizing $\ell(\Sigma_\alpha)$ with constraints given in (18). Since this is a constrained quadratic programming, either numerical algorithms or closed-form solutions are available.

4 Simulation study

In this Section, we evaluate the finite sample performance of the three proposed IBM estimators for two targets: outcome mean and linear regression coefficients. We use the same observed structures as the one introduced in Section 2. We also compare their performance against the naive estimator and the PPI++ method (Angelopoulos et al., 2023b) (if applicable) to illustrate the efficiency gains of incorporating additional sets of modalities.

4.1 Outcome mean estimation

We consider the following data generating process. We assume $X_1 \sim \mathcal{N}(\mathbf{1}_{10}, \mathbf{I}_{10})$, $X_2 \sim \mathcal{N}(\mathbf{1}_{20}, \mathbf{I}_{20})$, and $Y = X_1^T\mathbf{1}_{10} + X_2^T\mathbf{1}_{20} + \epsilon$ where $\epsilon \sim \mathcal{N}(0,1)$. The sample size of the unlabeled datasets $\mathscr{D}_{\{X_1,X_2\}}$ and $\mathscr{D}_{\{X_1\}}$ is fixed to $n_{\{X_1,X_2\}} = n_{\{X_1\}} = 20000$, while the sample size of the labeled dataset $n_{\{X_1,X_2,Y\}} = n_{\{X_1,Y\}}$ varies from 100 to 12800. Larger unlabeled datasets resemble the realistic setting in which labeled data are more difficult to collect. To simulate potentially biased pre-trained AI models, we consider two settings as follows. In the "correct predictor" scenario, we generate 40000 independent model training samples of (X_1, X_2, Y) following the above data generating process. In the "misspecified predictor" scenario, we

generate $Y = X_1^T \mathbf{1}_{10} + X_2^T \mathbf{1}_{20} + (X_1^2)^T \mathbf{1}_{10} + \epsilon$, as to simulate the conditional distribution discrepancy between our observed dataset and the model training dataset. In both scenarios, we fit a linear regression model on the training dataset to learn the pre-trained models $f(X_1)$ and $g(X_1, X_2)$.

For the PPI++ estimator, we ignore the set of features X_2 and select the tuning parameter according to Example 6.1 in (Angelopoulos et al., 2023b). For all three IBM estimators, we adopt the cross-fitting Algorithm 1. Notice that for the outcome mean, "expectation" and "imputation" modes coincide. We run experiments 1,000 times for each scenario to evaluate estimation and inference performance. We track three key metrics: root mean squared error (RMSE), empirical coverage, and confidence interval widths. Figure 3 shows these metrics under varying sample sizes and for each of the five estimators under consideration. All three IBM estimators consistently achieve lower RMSE than both PPI++ and the naive estimator *regardless* of the accuracy of the predictive AI models, emphasizing the benefits of incorporating more informative features in the accuracy of the estimate. Moreover, the IBM(Adaptive) estimator always performs better than the others, illustrating the advantage of pursuing the lowest asymptotic variance among the estimation class in (17). The coverage and confidence interval width plots confirm that every estimator achieves nominal coverage level around 95%, with our proposed estimators usually attaining it with shorter confidence intervals.

To further assess the robustness of these estimators against the quality of pre-trained AI models, we let $f_q(X_1) = (1-q)\mathbb{E}[Y \mid X_1] + q\varepsilon$, $g_q(X_1, X_2) = (1-q)\mathbb{E}[Y \mid X_1, X_2] + q\varepsilon, 0 \leq q \leq 1$, which are convex combinations of true predictors with random noise $\varepsilon \sim \mathcal{N}(0, 2)$. In a way, q can be interpreted as proxy of the rate of convergence of the AI models to their corresponding population quantities (see Das et al. 2024; Testa et al. 2025a for a more explicit characterization). Although all estimators considered protect the estimation efficiency against the naive estimator asymptotically, it is still interesting to verify their finite-sample performance, especially to verify their numerical stability in small labeled data setting as mentioned in Remark 3.10. Therefore, we consider $n_{\{X_1, X_2, Y\}} = n_{\{X_1, Y\}} = 100$ to inspect their performance in a limited labeled data setting, and $n_{\{X_1, X_2, Y\}} = n_{\{X_1, Y\}} = 1000$ to confirm their asymptotic behavior. Again, we repeat the experiments 1000 times for each setting.

The estimation accuracy curves with respect to the quality of the pre-trained models (controlled by q) are provided in Figure 4. As explained in Remark 3.10, the performance of IBM(PS) can be even worse than that of the naive estimator in the small labeled data setting until the quality of the AI model is far from random noise. In comparison, all other estimators perform no worse than the naive estimator regardless of the quality of the pre-trained model. In the large labeled data setting, all estimators safeguard the estimation accuracy for any pre-trained model quality. In addition, neither IBM(PS) nor IBM(RAY) estimators is universally superior to the other. Lastly, IBM(Adaptive) always achieves the smallest estimation errors among these estimators.

4.2 Linear regression coefficients

We now move to examine the finite-sample performance of our proposed estimators and the naive estimator in a low-dimensional linear regression problem. Again, we consider the

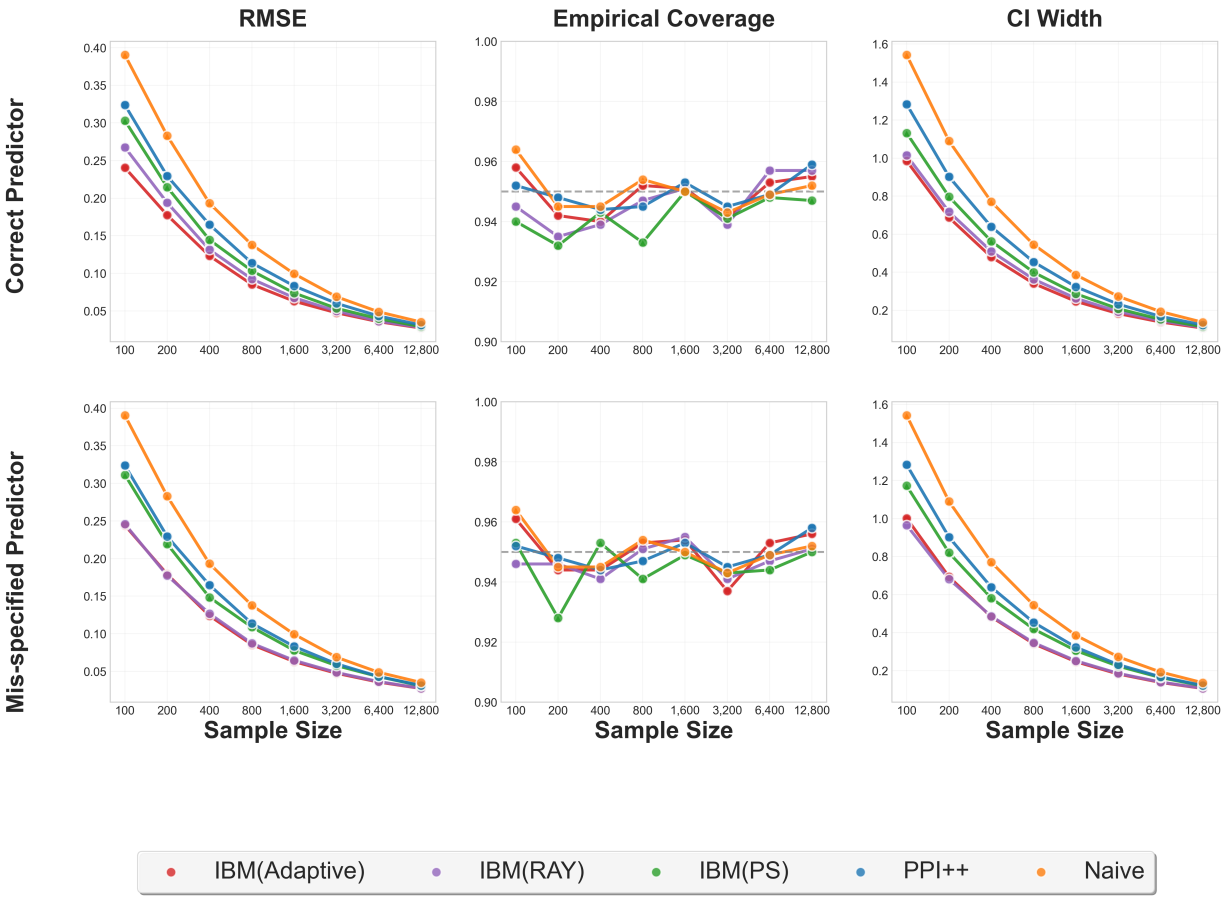

Figure 3: Performance comparison of five estimators for the estimation of the outcome mean target. Top row: "correct predictor" scenario. Bottom row: "misspecified predictor" scenario. RMSE, empirical coverage and confidence interval width are displayed for each scenario.

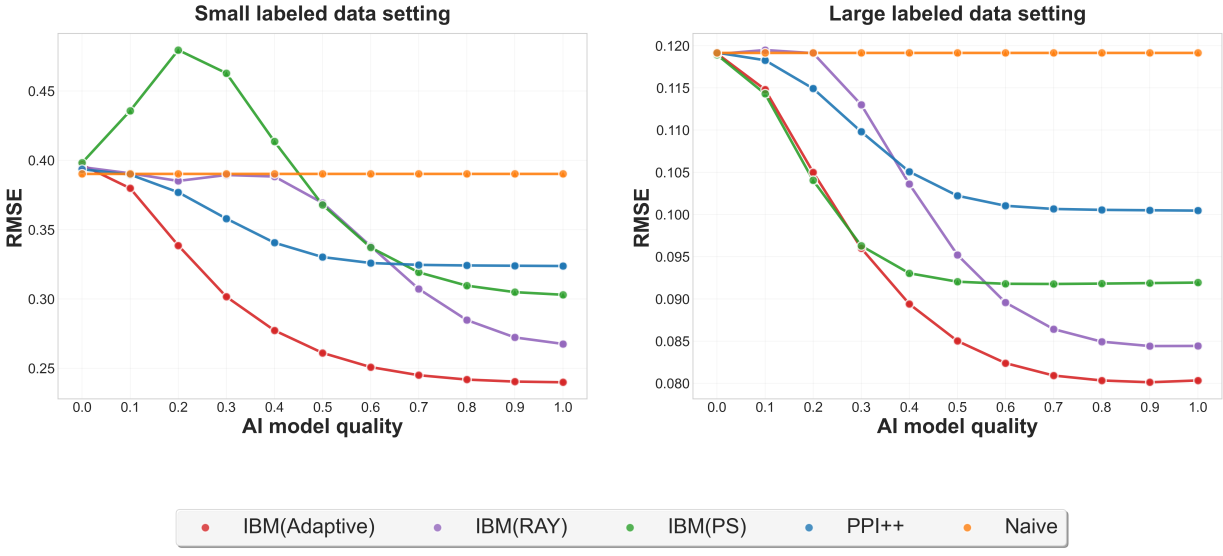

Figure 4: Estimation accuracy v.s. AI model quality of five estimators in small labeled data setting ($n_{\{X_1,X_2,Y\}} = n_{\{X_1,Y\}} = 100$) and large labeled data setting ($n_{\{X_1,X_2,Y\}} = n_{\{X_1,Y\}} = 1000$).

observing patterns in Figure 1, and generate simulated data as follows: $X_1 \in \mathbb{R}^2, X_2 \in \mathbb{R}^2$, $X = (X_1, X_2)$ are sampled from $\mathcal{N}(\mathbf{0}, \Sigma)$, where all diagonal entries of Σ are one and off-diagonal entries are 0.4. The linear model is specified as $Y = X_1^T \mathbf{1} + X_2^T \mathbf{1} + \epsilon$ where ϵ is standard Gaussian noise. Let the sample size of the complete data vary from 400 to 6400, and set the sample size of incomplete datasets at 10000. An independent dataset of 40000 samples is generated to train all the necessary pre-trained models for the imputation mode, where the pre-trained models in this experiment are specified as conditional VAE models (Sohn et al., 2015). All tuning parameters in three `IBM` estimators are designed to minimize the trace of the asymptotic variance. Again, we repeat the experiment 1000 times to evaluate the performance of the estimation and inference. Figure 5 shows the RMSE, the coverage and the trace of the finite-sample variance of four estimators under varying sample sizes. The three `IBM` estimators perform similarly, improving the RMSE and trace by at least 10% compared to the naive estimator, while maintaining the nominal coverage level.

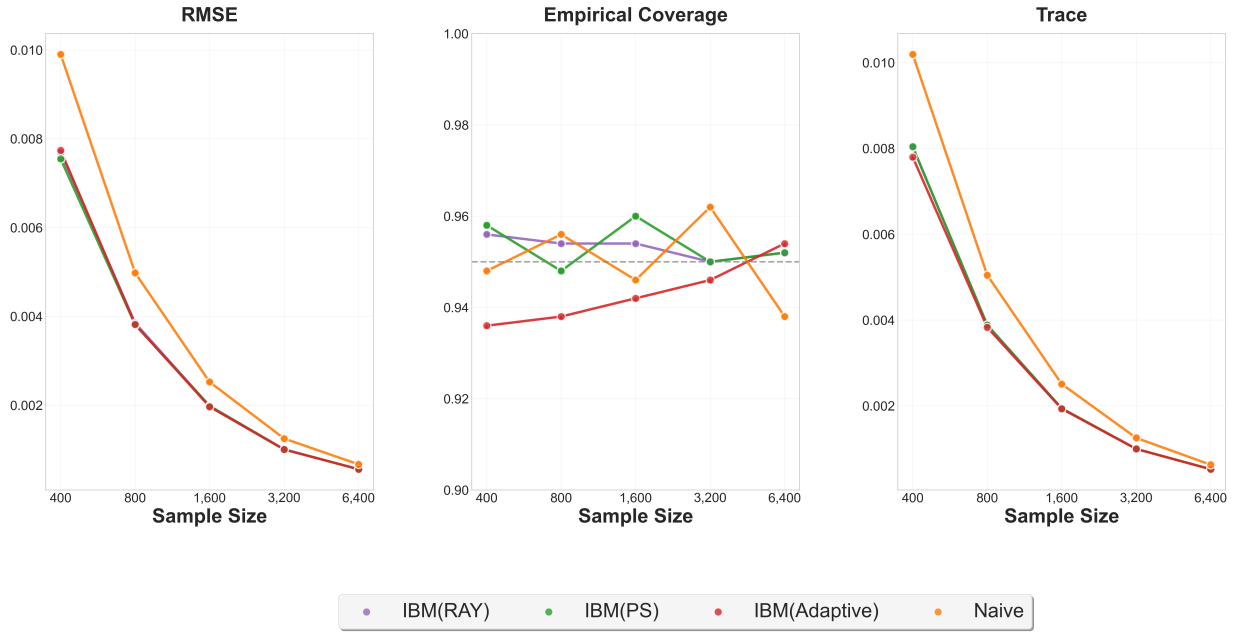

Figure 5: Performance comparison of four estimators of low-dimensional linear regression coefficients.

5 Application to surface protein abundance data

In this Section, we showcase the versatility of our `IBM` estimators by applying them to the surface protein abundance estimation problem. Recent advances in single-cell multi-omic technology have promised a detailed and integrated view of cell function and regulation. For example, Mimitou et al. (2021) have introduced the DOGMA-seq technique that can simultaneously measure chromatin accessibility (ATAC), gene expression (RNA) and cell-surface protein levels (ADT). Measuring these three omics captures the process by which DNA is transcribed into RNA, and then translated into protein, which is known as *central dogma* in molecular biology. Earlier techniques, such as CITE-seq, measure RNA and ADT, while ASAP-seq measures ATAC and ADT. The quantities measured for these techniques are

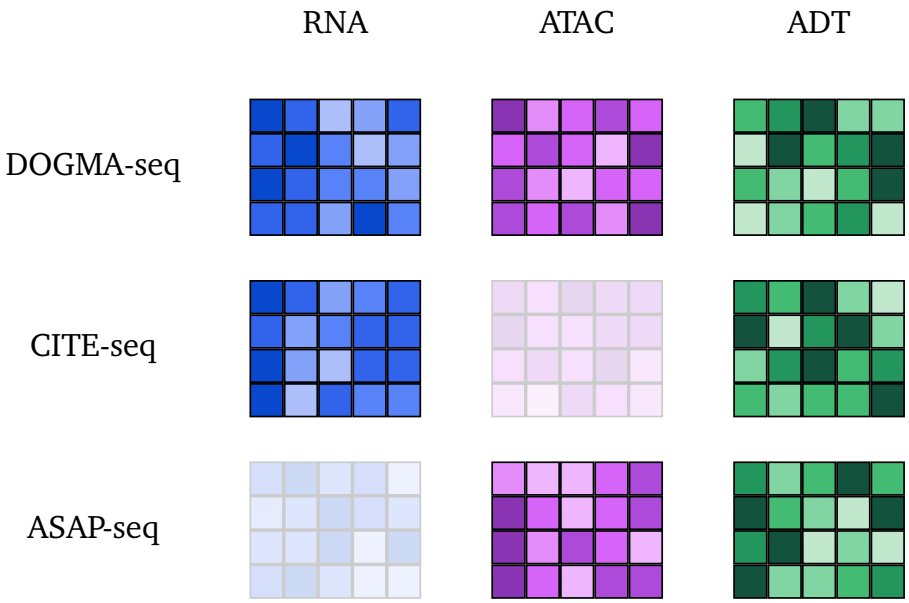

Figure 6: Observed omics for DOGMA-seq, CITE-seq and ASAP-seq datasets. DOGMA-seq observes RNA, ATAC and ADT modalities; CITE-seq only observes RNA and ADT modalities; ASAP-seq only observes ATAC and ADT modalities.

shown in Figure 6. Estimating mean protein abundance levels from a collection of cells is of great interest in understanding cell functions and states, which are crucial for downstream tasks such as disease diagnosis and drug development. However, when all relevant data are accumulated, RNA and/or ATAC are often measured when ADT is not. Because proteins are the final products of transcription and translation, RNA and ATAC measurements should be able to impute protein abundance. Using all available data, we expect to improve the accuracy of the estimation of protein abundance. In this experiment, we show that surface protein abundance estimation can be boosted by our proposed estimators.

In particular, we have access to three datasets using the aforementioned techniques applied to PBMCs (peripheral blood mononuclear cells); CD4 cells are our target in this study. After quality control and pre-processing including batch correction, we select 880 genes for the RNA-seq, 3288 peaks for the ATAC-seq, and 208 proteins for the ADT. To pretrain AI models to predict ADT from RNA or ATAC, we used CITE-seq and ASAP-seq datasets, respectively. To pretrain an AI model to predict ADT from both RNA and ATAC we utilize 20% of the cells randomly selected from DOGMA-seq. All of these AI models are specified as feedforward neural nets with ReLU activation functions. We simulate blockwise missingness using the remaining 80% cells from DOGMA-seq. We manually create the missing patterns by randomly splitting these cells into four datasets: $\mathscr{D}_1 = \{\text{RNA, ATAC, ADT}\}$, $\mathscr{D}_2 = \{\text{RNA, ATAC}\}$, $\mathscr{D}_3 = \{\text{RNA}\}$ and $\mathscr{D}_4 = \{\text{ATAC}\}$, with a proportion of 10%, 30%, 30% and 30%. Then, our proposed `IBM` estimators and the naive estimator (estimating protein abundance only by $\mathscr{D}_1$) can be applied to these datasets to estimate the mean abundance levels for each protein. Note that `PPI`-type estimators are not naturally applicable in this problem, so they are not evaluated in this study. Since protein abundance levels of all cells in DOGMA-seq are actually observed, we consider the sample mean abundance levels across all cells as *true* values in the following

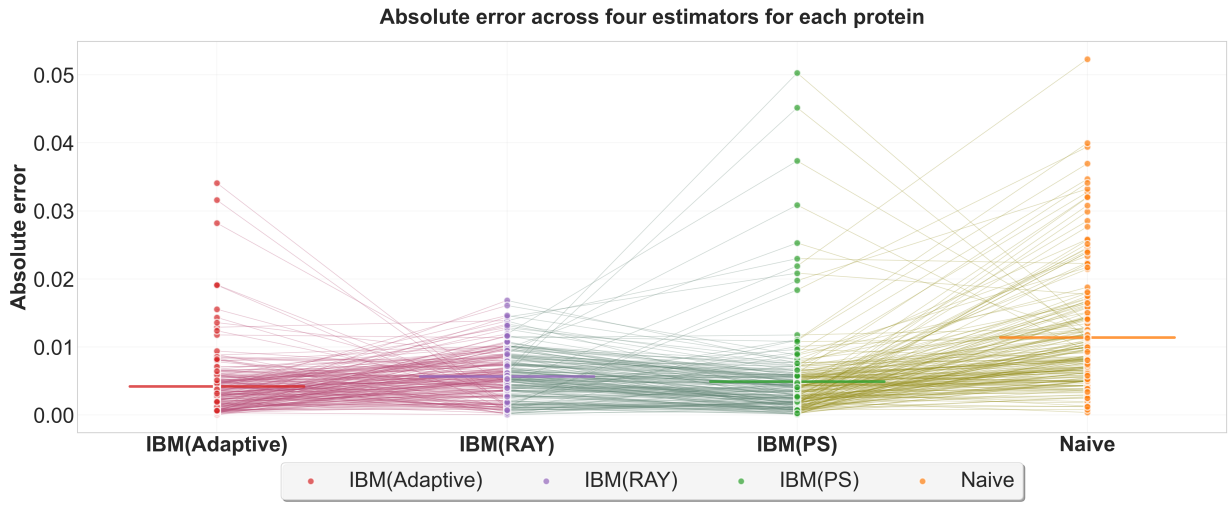

Figure 7: Estimation accuracy of each surface protein abundance level by four estimators. The horizontal bars are average absolute error over all proteins for each estimator.

evaluation.

First, we evaluate the accuracy of the estimation of each protein abundance level by absolute error compared to the *true* values, shown in Figure 7. Specifically, each dot represents the absolute error of a specific protein from the method indicated by its color, and the horizontal bars correspond to the average absolute error across all proteins of the corresponding estimators. It is evident that the three `IBM` estimators achieve a lower average estimation error compared to the naive estimator. In this Figure, line segments connect the same proteins between adjacent estimators. By doing so, we can measure the overall tendency of the estimators to provide similar estimates. `IBM`(Adaptive) and `IBM`(PS) exhibit a lower estimation error for most proteins with few exceptions, while the performance of `IBM`(RAY) across all proteins is less variable but with a slightly higher average estimation error.

Next, we assess the estimation efficiency by comparing the confidence interval width of the four estimators. For ease of illustration, we select proteins that are among the top 5% abundance levels and show four confidence intervals for these proteins in Figure 8. For a better comparison, we standardize the confidence intervals by subtracting the *true* protein mean. In addition, relative efficiencies for all these proteins were measured by

$$\frac{\text{CI width of the naive estimator}}{\text{CI width of each estimator}}$$

are plotted. All three `IBM` estimators achieve an efficiency higher than that of the naive estimator for all of these proteins, as expected. Among `IBM` estimators, `IBM`(Adaptive) is universally the most efficient, supporting its optimality among the class of estimators defined in (17).

6 Discussion

In this work, we have proposed a suite of `IBM` estimators which incorporate pre-trained AI (predictive or generative) models to improve statistical efficiency for data presenting

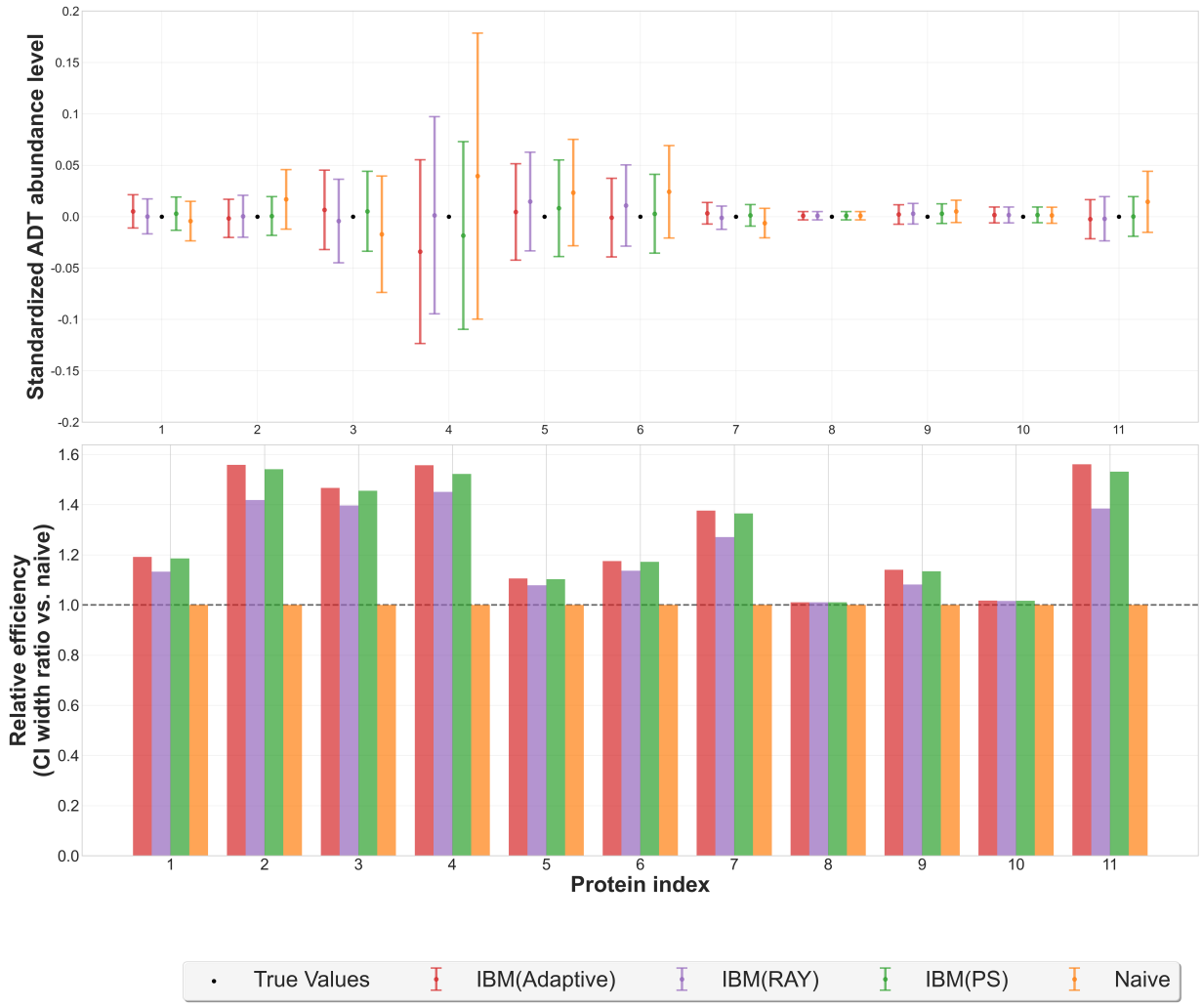

Figure 8: Confidence intervals and relative efficiency of three `IBMestimators` compared to the naive estimator, illustrated on top 5% abundance level proteins.

blockwise missingness. First, we revisit semiparametric theory for non-monotone missingness, and propose a novel and tractable RAY approximation to the optimal estimating function, which – we believe – is a significant contribution to the literature. Second, we propose a suite of `IBM` estimators that include `IBM`(RAY), `IBM`(PS) and `IBM`(Adaptive). In particular, `IBM`(Adaptive) achieves the lowest asymptotic variance among the class containing `IBM`(PS) and `IBM`(RAY) as special cases.

Through an extensive simulation study, we support the theoretical derivations on the asymptotic behavior of our estimators showing their reliability even in finite-sample scenarios, where we observe consistent improvements in terms of estimation accuracy, valid coverage, and shrunk confidence intervals. These numerical experiments indicate that our proposed method is both theoretically sound and practically feasible. Further, we apply our method to single-cell multi-omic datasets, and show superior estimation accuracy and efficiency for surface protein abundance estimation.

Beyond the aforementioned contributions, there are several future research directions worth investigating. First, the MCAR assumption in this work is quite stringent for many real-data applications. Despite some attempts to extend the theory to MAR scenarios in the simpler semi-supervised setting (Testa et al., 2025b; Ying et al., 2025), moving beyond MCAR for blockwise missingness data seems quite challenging. The estimation of propensity score under non-monotone MAR is challenging: for example, the method proposed by Sun and Tchetgen Tchetgen (2018) could have convergence issue with finite samples. Despite some work (Chen et al., 2025) that has considered this problem under MAR, it is still unclear how far the proposed solution is from the most efficient estimator with a tractable computation. Further, it would also be interesting to study whether the RAY approximation can be applied to more general missing mechanisms such as MAR. Finally, for some scientific applications, traditional missing mechanisms including MCAR, MAR and even missing not at random (MNAR) may not be meaningful for blockwise missing data – see Mitra et al. (2023) for a more comprehensive discussion on the challenges experienced when handling data exhibiting blockwise missingness.

Acknowledgment

This project was supported by National Institute of Mental Health (NIMH) grant R01MH123184 and NSF grant DMS-2310764 and NSF grant DMS-2515687.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Angelopoulos, A. N., Duchi, J. C., and Zrnic, T. (2023b). Ppi++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*.

BACKHAUSZ, Á. and SZEGEDY, B. (2019). On the almost eigenvectors of random regular graphs. *The Annals of Probability*, 47(3):1677–1725.

Chen, H. Y. (2004). Nonparametric and semiparametric models for missing covariates in parametric regression. *Journal of the American Statistical Association*, 99(468):1176–1189.

Chen, X., McCormick, T., Mukherjee, B., and Wu, Z. (2025). A unified framework for inference with general missingness patterns and machine learning imputation. *arXiv preprint arXiv:2508.15162*.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

Das, M., Kennedy, E. H., and Jewell, N. P. (2024). Doubly robust capture-recapture methods for estimating population size. *Journal of the American Statistical Association*, 119(546):1309–1321.

Diakité, A. O., Moreau, C., Bezgin, G., Bhagwat, N., Rosa-Neto, P., Poline, J.-B., Girard, S., Barry, A., Initiative, A. D. N., et al. (2025). Adapdiscom: An adaptive sparse regression method for high-dimensional multimodal data with block-wise missingness and measurement errors. *arXiv preprint arXiv:2508.00120*.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Muller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Lacey, K., Goodwin, A., Marek, Y., and Rombach, R. (2024). Scaling rectified flow transformers for high-resolution image synthesis. *ArXiv*, abs/2403.03206.

Gronsbell, J., Gao, J., Shi, Y., McCaw, Z. R., and Cheng, D. (2024). Another look at inference after prediction. *arXiv preprint arXiv:2411.19908*.

Hoeffding, W. (1948). A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics*, 19(3):293 – 325.

Hooker, G. (2004). Discovering additive structure in black box functions. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 575–580.

Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.

Huang, J., Wang, H., Lei, Y., and Chen, Y. (2025). Efficient semiparametric inference for distributed data with blockwise missingness. *arXiv preprint arXiv:2508.16902*.

Ji, W., Lei, L., and Zrnic, T. (2025). Predictions as surrogates: Revisiting surrogate outcomes in the age of ai. *arXiv preprint arXiv:2501.09731*.

Jin, Y. and Rothenhäusler, D. (2023). Modular regression: improving linear models by incorporating auxiliary data. *Journal of Machine Learning Research*, 24(351):1–52.

Kundu, P., Tang, R., and Chatterjee, N. (2019). Generalized meta-analysis for multiple regression models across studies with disparate covariate information. *Biometrika*, 106(3):567–585.

Li, Y., Yang, X., Wei, Y., and Liu, M. (2024). Adaptive and efficient learning with blockwise missing and semi-supervised data. *arXiv preprint arXiv:2405.18722*.

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al. (2024). Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Metropolis, N. and Ulam, S. (1949). The monte carlo method. *Journal of the American statistical association*, 44(247):335–341.

Miao, J., Miao, X., Wu, Y., Zhao, J., and Lu, Q. (2023). Assumption-lean and data-adaptive post-prediction inference. *arXiv preprint arXiv:2311.14220*.

Mimitou, E. P., Lareau, C. A., Chen, K. Y., Zorzetto-Fernandes, A. L., Hao, Y., Takeshima, Y., Luo, W., Huang, T.-S., Yeung, B. Z., Papalexi, E., et al. (2021). Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nature biotechnology*, 39(10):1246–1258.

Mitra, R., McGough, S. F., Chakraborti, T., Holmes, C., Copping, R., Hagenbuch, N., Biedermann, S., Noonan, J., Lehmann, B., Shenvi, A., et al. (2023). Learning from data with structured missingness. *Nature Machine Intelligence*, 5(1):13–23.

Motwani, K. and Witten, D. (2023). Revisiting inference after prediction. *J. Mach. Learn. Res.*, 24:394:1–394:18.

Owen, A. B. (2013). Monte carlo theory, methods and examples.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125.

Robins, J. M. (1997). Non-response models for the analysis of non-monotone non-ignorable missing data. *Statistics in medicine*, 16(1):21–37.

Robins, J. M. and Gill, R. D. (1997). Non-response models for the analysis of non-monotone ignorable missing data. *Statistics in medicine*, 16(1):39–56.

Robins, J. M., Rotnitzky, A., and and, L. P. Z. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

Sohn, K., Lee, H., and Yan, X. (2015). Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28.

Song, S., Lin, Y., and Zhou, Y. (2024). Semi-supervised inference for block-wise missing data without imputation. *Journal of Machine Learning Research*, 25(99):1–36.

Sun, B. and Tchetgen Tchetgen, E. J. (2018). On inverse probability weighting for nonmonotone missing at random data. *Journal of the American Statistical Association*, 113(521):369–379.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., et al. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Testa, L., Boschi, T., Chiaromonte, F., Kennedy, E. H., and Reimherr, M. (2025a). Doubly-robust functional average treatment effect estimation. *arXiv preprint arXiv:2501.06024*.

Testa, L., Xu, Q., Lei, J., and Roeder, K. (2025b). Semiparametric semi-supervised learning for general targets under distribution shift and decaying overlap. *arXiv preprint arXiv:2505.06452*.

Tsiatis, A. A. (2006). *Semiparametric theory and missing data*, volume 4. Springer.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

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

Xu, Z., Witten, D., and Shojaie, A. (2025). A unified framework for semiparametrically efficient semi-supervised learning. *arXiv preprint arXiv:2502.17741*.

Xue, F., Ma, R., and Li, H. (2021). Statistical inference for high-dimensional linear regression with blockwise missing data. *arXiv preprint arXiv:2106.03344*.

Xue, F. and Qu, A. (2021). Integrating multisource block-wise missing data in model selection. *Journal of the American Statistical Association*, 116(536):1914–1927.

Yang, S. (2022). A double robust approach for non-monotone missingness in multi-stage data. *arXiv preprint arXiv:2201.01010*.

Ying, C., Jin, J., Guo, Y., Li, X., Liang, M., and Zhao, J. (2025). Towards the efficient inference by incorporating automated computational phenotypes under covariate shift. *arXiv preprint arXiv:2505.22632*.

Yu, G., Li, Q., Shen, D., and Liu, Y. (2020). Optimal sparse linear prediction for block-missing multi-modality data without imputation. *Journal of the American Statistical Association*, 115(531):1406–1419.

Zhang, Y., Chakrabortty, A., and Bradic, J. (2023). Double robust semi-supervised inference for the mean: selection bias under mar labeling with decaying overlap. *Information and Inference: A Journal of the IMA*, 12(3):2066–2159.

Zhao, S. and Candès, E. (2025). Imputation-powered inference. *arXiv preprint arXiv:2509.13778*.

Zrnic, T. and Candès, E. J. (2024). Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15):e2322083121.

Appendix

A Technical details

A.1 Proof of Lemma 3.2

Proof. Based on the definition of RAY decomposition, we have

$$\begin{aligned}\sum_{s\subseteq[M]} P_s(f) &= \sum_{s\subseteq[M]}\sum_{r\subseteq s, r\in Q}(-1)^{|s|-|r|}A_r(f)\\ &= \sum_{r\in Q}\Big(\sum_{s\supseteq r, s\subseteq[M]}(-1)^{|s|-|r|}\Big)A_r(f)\end{aligned}$$

Now we analyze the coefficient term for each $A_r(f)$. If $r \neq [M]$, the coefficient term is actually the summation over all subsets of $[M]$ that contains r. Alternatively, we can write $s = r \cup s'$, $s' \subseteq [M]$ and $s' \cap r = \emptyset$. By this change of sets trick, the coefficient term is

$$\begin{aligned}\sum_{s\supseteq r}(-1)^{|s|-|r|} &= \sum_{s'\subseteq[M]\setminus r}(-1)^{|s'|}\\ &= \sum_{k=0}^{n}\binom{n}{k}(-1)^k \qquad (n = |[M]\setminus r|)\\ &= \sum_{k=0}^{n}\binom{n}{k}1^{n-k}(-1)^k\\ &= (1-1)^n = 0.\end{aligned}$$

For $r = [M]$, it is evident the coefficient is 1 and $A_{[M]}(f) = f$ since $f \in \sigma(\mathscr{Z})$. □

A.2 Proof of Proposition 3.5 and Proposition 3.6

Proof. We prove that the estimating function $\psi(Z,\theta)$ and $\psi_\alpha(Z,\theta)$ is valid by showing $\mathbb{E}[\psi(Z,\theta^\star)] = \mathbb{E}[\psi_\alpha(Z,\theta^\star)] = \mathbf{0}_d$. The estimating function is defined as:

$$\begin{aligned}\psi(Z,\theta) &= \frac{\mathbb{I}(R=[M])}{\pi_{[M]}}\psi^F(Z,\theta) + \sum_{r\in Q, r\neq[M]\}}\omega_r A_r[\psi^F(Z,\theta)]\\ \psi_\alpha(Z,\theta) &= \frac{\mathbb{I}(R=[M])}{\pi_{[M]}}\psi^F(Z,\theta) + \sum_{r\in Q, r\neq[M]\}}\alpha_r\omega_r\mathscr{F}_r(X_r).\end{aligned}$$

For the first terms of both estimating function, we have

$$\mathbb{E}[\frac{\mathbb{I}(R=[M])}{\pi_{[M]}}\psi^F(Z,\theta^\star)] = \mathbb{E}[\frac{\mathbb{I}(R=[M])}{\pi_{[M]}}]\mathbb{E}[\psi^F(Z,\theta^\star)] = 0.$$

For the second terms, it is suffice to show that $\mathbb{E}[\omega_r] = 0$ for $r \in Q, r \neq [M]$. From the definition of ω_r in (10), its expectation is:

$$\mathbb{E}[\omega_r] = \mathbb{E}\left[\sum_{s\supseteq r}(-1)^{|s|-|r|}\frac{\sum_{t\supseteq s}\mathbb{I}(R=t)}{\lambda_s}\right] = \sum_{s\supseteq r}\frac{(-1)^{|s|-|r|}}{\lambda_s}\sum_{t\supseteq s}\pi_t$$

Since $\lambda_s = \sum_{t \supseteq s} \pi_t$ by definition, the expression simplifies to $\mathbb{E}[\omega_r] = \sum_{s \supseteq r} (-1)^{|s|-|r|} = 0$ by the conclusion in Appendix A.1. The proof is complete. □

A.3 Proof of Theorem 3.8

Proof. The proof follows from classical Z-estimator theory. The estimator $\hat{\theta}^{\texttt{IBM}}_{\texttt{RAY/PS}}$ is the solution to the estimating equation $U_N(\theta) = \frac{1}{N}\sum_{i=1}^{N} \psi_\alpha(Z_i;\theta) = \mathbf{0}_d$. We assume the standard regularity conditions: the parameter $\theta^\star$ is well-identified as the unique root of $\mathbb{E}[\psi_\alpha(Z;\theta)]$; ψ_α is continuously differentiable with respect to θ with a bounded derivative; the expected Jacobian matrix is invertible; and the variance of ψ_α is finite.

(a) Consistency

Under the stated regularity conditions, consistency follows from the uniform law of large numbers. The sample objective function $U_N(\theta)$ converges uniformly in probability to its population counterpart $\mathbb{E}[\psi_\alpha(Z;\theta)]$ over a compact neighborhood of $\theta^\star$. Since $\theta^\star$ is the unique solution to $\mathbb{E}[\psi_\alpha(Z;\theta)] = \mathbf{0}_d$, the root of the sample equation $\hat{\theta}^{\texttt{IBM}}_{\texttt{RAY/PS}}$ must converge in probability to $\theta^\star$.

(b) Asymptotic Normality

Given that $\hat{\theta}^{\texttt{IBM}}_{\texttt{RAY/PS}} \xrightarrow{p} \theta^\star$, we perform a first-order Taylor expansion of $U_N(\hat{\theta}^{\texttt{IBM}}_{\texttt{RAY/PS}})$ around $\theta^\star$:

$$\mathbf{0}_d = U_N(\hat{\theta}^{\texttt{IBM}}_{\texttt{RAY/PS}}) = U_N(\theta^\star) + \left(\frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial \theta^T}\psi_\alpha(Z_i;\tilde{\theta})\right)(\hat{\theta}^{\texttt{IBM}}_{\texttt{RAY/PS}} - \theta^\star)$$

where $\tilde{\theta}$ is a point on the line segment between $\hat{\theta}^{\texttt{IBM}}_{\texttt{RAY/PS}}$ and $\theta^\star$. Rearranging gives:

$$\sqrt{N}(\hat{\theta}^{\texttt{IBM}}_{\texttt{RAY/PS}} - \theta^\star) = -\left(\frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial \theta^T}\psi_\alpha(Z_i;\tilde{\theta})\right)^{-1} \sqrt{N} U_N(\theta^\star)$$

The Jacobian term converges in probability to its expectation at $\theta^\star$ by the law of large numbers and the consistency of $\hat{\theta}^{\texttt{IBM}}_{\texttt{RAY/PS}}$:

$$\frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial \theta^T}\psi_\alpha(Z_i;\tilde{\theta}) \xrightarrow{p} A \triangleq \mathbb{E}\left[\frac{\partial}{\partial \theta^T}\psi_\alpha(Z;\theta^\star)\right]$$

By the Central Limit Theorem, $\sqrt{N}U_N(\theta^\star) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \psi_\alpha(Z_i;\theta^\star)$ converges in distribution to a multivariate normal random variable $\mathcal{N}(\mathbf{0}_d, \Sigma)$, where $\Sigma = \mathbb{V}(\psi_\alpha(Z;\theta^\star))$. By Slutsky's theorem, the asymptotic distribution is $\sqrt{N}(\hat{\theta}^{\texttt{IBM}}_{\texttt{RAY/PS}} - \theta^\star) \rightsquigarrow \mathcal{N}(\mathbf{0}_d, A^{-1}\Sigma(A^T)^{-1})$.

We now derive the explicit forms of A and Σ. The Jacobian is $A = \mathbb{E}\left[\frac{\mathbb{I}(R=[M])}{\pi_{[M]}}\frac{\partial \psi^F}{\partial \theta^T}\right] + \sum_{r=2}^{|Q|} \alpha_r \mathbb{E}\left[\omega_r \frac{\partial \mathscr{F}_r}{\partial \theta^T}\right]$. As shown in the proof of Proposition 3.5, terms involving $\mathbb{E}[\omega_r]$ are zero. This leaves $A = \mathbb{E}[\frac{\partial \psi^F}{\partial \theta^T}]$.

The covariance matrix $\Sigma = \mathbb{V}(\psi_\alpha(Z;\theta^\star))$ is given by:

$$\Sigma = \frac{\mathbb{V}\left(\psi^F\right)}{\pi_{[M]}} + \sum_{r,r'\in Q, r,r'\neq[M]} \alpha_r \alpha_{r'} \mathrm{Cov}(\omega_r \mathscr{F}_r, \omega_{r'}\mathscr{F}_{r'}) + \sum_{r\in Q, r\neq[M]} \alpha_r \mathrm{Cov}\left(\frac{\mathbb{I}(R=[M])}{\pi_{[M]}}\psi^F, \omega_r \mathscr{F}_r\right) + \sum_{r\in Q, r\neq[M]} \alpha_r \mathrm{Cov}\left(\frac{\mathbb{I}(R=[M])}{\pi_{[M]}}\psi^F, \omega_r \mathscr{F}_r\right)^T$$

Evaluating these variance and covariance terms using MCAR assumption yields the coefficients presented in the theorem. For example, $\mathbb{V}\left(\frac{\mathbb{I}(R=[M])}{\pi_{[M]}}\psi^F\right) = \frac{1}{\pi_{[M]}}\mathbb{V}(\psi^F)$, and the covariance terms simplify based on expectations of products of the ω_r coefficients and covariances of the functions ψ^F and $\mathscr{F}$. This leads to the final asymptotic variance matrix Σ_α presented in Theorem 3.8, where the coefficients γ_r and $\eta_{r,r'}$ arise from evaluating the expectations involving ω_r and the observation indicators. This completes the proof. □

A.4 Proof of Corollary 3.9

Proof. The objective is to find the vector of tuning parameters α that minimizes the scalar criterion $\ell(\Sigma_\alpha)$, where $\ell(\cdot)$ is a linear functional and Σ_α is the asymptotic variance matrix from Theorem 3.8. The objective function $f(\alpha) = \ell(\Sigma_\alpha)$ is a quadratic function of α:

$$f(\alpha) = \ell(\Sigma_\alpha) = \frac{\ell\left(A^{-1}\mathbb{V}(\psi^F)A^{-1}\right)}{\pi_{[M]}} + \sum_{r,r'} \alpha_r \alpha_{r'} \eta_{r,r'} \ell\left(A^{-1}\mathrm{Cov}(\mathscr{F}_r, \mathscr{F}_{r'})A^{-1}\right) + 2\sum_r \alpha_r \gamma_r \ell\left(A^{-1}\mathrm{Cov}(\psi^F, \mathscr{F}_r)A^{-1}\right)$$

Using the linearity of $\ell(\cdot)$, we can write this in a standard matrix quadratic form:

$$f(\alpha) = C_0 + \alpha^T M \alpha + 2L^T \alpha$$

where $C_0 = \frac{\ell(A^{-1}\mathbb{V}(\psi^F)A^{-1})}{\pi_{[M]}}$ is a constant with respect to α, and the matrix G and vector L are defined as:

$$G = \left(\eta_{r,r'}\ell(A^{-1}\mathrm{Cov}(\mathscr{F}_r, \mathscr{F}_{r'})A^{-1})\right)_{r,r'\in Q, r,r'\neq[M]}$$
$$L = \left(\gamma_r \ell(A^{-1}\mathrm{Cov}(\psi^F, \mathscr{F}_r)A^{-1})\right)_{r\in Q, r\neq[M]}$$

To find the minimum of this convex quadratic function, we compute the gradient with respect to α and set it to zero.

$$\nabla_\alpha f(\alpha) = 2M\alpha + 2L = \mathbf{0}$$

This yields the equation $G\alpha = -L$. Under the mild condition that the matrix M is positive definite (which holds if the functions $\mathscr{F}_r$ are not linearly dependent and the matrix $(\eta_{r,r'})$ is positive definite), M is invertible. This condition also ensures that the solution is a unique minimum. The optimal $\alpha^\star$ is therefore:

$$\alpha^\star = -G^{-1}L$$

To find the minimum value of the objective function, we substitute $\alpha^\star$ back into the expression for $f(\alpha)$:

$$\ell(\Sigma_{\alpha^\star}) = \frac{\ell(A^{-1}\mathbb{V}(\psi^F)A^{-1})}{\pi_{[M]}} - L^T G^{-1} L$$

The term $L^T G^{-1} L \geq 0$ represents the efficiency gain by the optimal choice of $\alpha^\star$. □

A.5 Characterization of IBM(PS) and IBM(RAY) in the framework of estimating function (17)

In this part, we show that ω_r's in IBM(PS) and IBM(RAY) are equivalent to certain α in (18), and satisfy the constraint.

For each AI model $\mathscr{F}_r$, the corresponding RAY random variable is

$$\begin{aligned}\omega_r &= \sum_{t\supseteq r}(-1)^{|t|-|r|}\frac{\mathbb{I}(R\supseteq t)}{\lambda_t}\\ &= \sum_{t\supseteq r}(-1)^{|t|-|r|}\frac{\sum_{s\supseteq t}\mathbb{I}(R=s)}{\lambda_t}\\ &= \sum_{s\supseteq r, s\in Q}\frac{\mathbb{I}(R=s)}{\pi_s}\pi_s\Big(\sum_{t:r\subseteq t\subseteq s}\frac{(-1)^{|t|-|r|}}{\lambda_t}\Big),\end{aligned}$$

which indicates that RAY random variables correspond to $\alpha_{r,s}=\pi_s\Big(\sum_{t:r\subseteq t\subseteq s}\frac{(-1)^{|t|-|r|}}{\lambda_t}\Big)$. Next, we show that $\sum_{s\supseteq r}\pi_s\Big(\sum_{t:r\subseteq t\subseteq s}\frac{(-1)^{|t|-|r|}}{\lambda_t}\Big)=0$ for $r\in Q, r\neq[M]$:

$$\begin{aligned}\sum_{s\supseteq r}\pi_s\Big(\sum_{t:r\subseteq t\subseteq s}\frac{(-1)^{|t|-|r|}}{\lambda_t}\Big) &= \sum_{t\supseteq r}\Big(\frac{(-1)^{|t|-|r|}}{\lambda_t}\sum_{s\supseteq t}\pi_s\Big)\\ &= \sum_{t\supseteq r}(-1)^{|t|-|r|}=0\end{aligned}$$

where the last equation holds by the conclusion in Appendix A.1.

For pattern-stratification random variable, $\omega_r=\frac{\mathbb{I}(R=r)}{\pi_r}-\frac{\mathbb{I}(R=[M])}{\pi_{[M]}}$, corresponding to $\alpha_{r,r}=1$ and $\alpha_{r,[M]}=-1$, which satisfies the constraint apparently.

A.6 Comparison of efficiency gain of IBM(RAY) and IBM(PS)

Following Theorem 3.8 and Corollary 3.9, comparing IBM(RAY) and IBM(PS) is essentially considering $L^TG^{-1}L$ under different choices of ω_r. Since the matrix G is the Hadamard product between $(\eta_{r,r'})_{r,r'\in Q,r,r'\neq[M]\}}$ and $(\ell(A^{-1}\mathrm{Cov}(\mathscr{F}_r,\mathscr{F}_{r'})A^{-1}))_{r,r'\in Q,r,r'\neq[M]\}}$, its inverse is intractable to analyze in a closed-form. Therefore, we consider a special scenario that $\mathrm{Cov}(\mathscr{F}_r,\mathscr{F}_{r'})=0$ if $r\neq r'$, such that G is a diagonal matrix. As a result, the efficiency gain of two estimators is of the form

$$\sum_{r\in Q,r\neq[M]}\frac{\gamma_r^2\ell^2(A^{-1}\mathrm{Cov}(\mathscr{F}_r,\psi^F)A^{-1})}{\eta_{rr}\ell(A^{-1}\mathrm{Cov}(\mathscr{F}_r,\mathscr{F}_r)A^{-1})}, \tag{19}$$

where γ_r and η_{rr} vary between IBM(RAY) and IBM(PS) according to their choice of ω_r. Given fixed AI model $\mathscr{F}_r$'s, $\ell^2(A^{-1}\mathrm{Cov}(\mathscr{F}_r,\psi^F)A^{-1})/\ell(A^{-1}\mathrm{Cov}(\mathscr{F}_r,\mathscr{F}_r)A^{-1})$ is a fixed constant, denoted as ℓ_r. For IBM(RAY), the Equation (19) writes out as

$$EG_{RAY}=\sum_{r\in Q,r\neq[M]}\frac{\sum_{t_1\supseteq r}\sum_{t_2\supseteq r}(-1)^{|t_1|+|t_2|-2|r|}/\lambda_{t_1}\lambda_{t_2}}{\sum_{t_1\supseteq r}\sum_{t_2\supseteq r}(-1)^{|t_1|+|t_2|-2|r|}\lambda_{t_1\cap t_2}/\lambda_{t_1}\lambda_{t_2}}\ell_r,$$

while the efficiency gain for IBM(PS) is

$$EG_{PS} = \sum_{r \in Q, r \neq [M]} \frac{\pi_r}{\pi_{[M]}^2 + \pi_{[M]}\pi_r} \ell_r.$$

Therefore, identifying the condition in which $EG_{RAY} > EG_{PS}$ or vice versa is equivalent to analyzing the behavior of a multi-variate rational function over the domain $\sum_{r \in Q} \pi_r = 1$ and $\pi_r > 0, \forall r \in Q$, which is rather intricate.