Journal of Machine Learning Research 24 (2023) 1-24 Submitted 6/22; Revised 8/23; Published 9/23

Augmented Transfer Regression Learning with Semi-non-parametric Nuisance Models

Molei Liu ML4890@CUMC.COLUMBIA.EDU
Department of Biostatistics
Columbia Mailman School of Public Health
New York, NY 10032, USA

Yi Zhang YIZHANG.USTC@GMAIL.COM
Department of Statistics
Harvard University
Cambridge, MA 02138, USA

Katherine P Liao KLIAO@BWH.HARVARD.EDU
Department of Medicine Rheumatology, Immunology
Brigham and Women's Hospital
Boston, MA 02115, USA

Tianxi Cai TCAI.HSPH@GMAIL.COM
Department of Biostatistics
Harvard Chan School of Public Health
Boston, MA 02115, USA

Editor: Jin Tian

Abstract

We develop an augmented transfer regression learning (ATReL) approach that introduces an imputation model to augment the importance weighting equation to achieve double robustness for covariate shift correction. More significantly, we propose a novel semi-non-parametric (SNP) construction framework for the two nuisance models. Compared with existing doubly robust approaches relying on fully parametric or fully non-parametric (machine learning) nuisance models, our proposal is more flexible and balanced to address model misspecification and the curse of dimensionality, achieving a better trade-off in terms of model complexity. The SNP construction presents a new technical challenge in controlling the first-order bias caused by the nuisance estimators. To overcome this, we propose a two-step calibrated estimating approach to construct the nuisance models that ensures the effective reduction of potential bias. Under this SNP framework, our ATReL estimator is $n^{1/2}$-consistent when (i) at least one nuisance model is correctly specified and (ii) the non-parametric components are rate-doubly robust. Simulation studies demonstrate that our method is more robust and efficient than existing methods under various configurations. We also examine the utility of our method through a real transfer learning example of the phenotyping algorithm for rheumatoid arthritis across different time windows. Finally, we propose ways to enhance the intrinsic efficiency of our estimator and to incorporate modern machine-learning methods in the proposed SNP framework.

Keywords: Covariate shift, model misspecification, double robustness, double machine learning, semi-non-parametric model, bias calibration.

1. Introduction

1.1 Background

The shift in the predictor distribution, often referred to as *covariate shift*, is one of the key contributors to poor transportability and generalizability of a supervised learning model from one data set to another. An example that arises often in modern biomedical research is the between health system transportability of prediction algorithms trained from electronic health records (EHR) data (Weng et al., 2020). Frequently encountered heterogeneity between hospital systems includes the underlying patient population and how the EHR system encodes the data. For example, the prevalence of rheumatoid arthritis (RA) among patients with at least one billing code of RA differs greatly among hospitals (Carroll et al., 2012). On the other hand, the conditional distribution of the disease outcome given all important EHR features may remain stable and similar for different cohorts. Nevertheless, a shift in the distribution of these features can still have a large impact on the performance of a prediction algorithm trained in one source cohort on another target cohort (Rasmy et al., 2018). Thus, correcting for the covariate shift is crucial to successful knowledge transfer across multiple heterogeneous studying cohorts.

The robustness of covariate shift correction is an important topic and has been widely studied in the recent literature on statistical learning. A branch of work including Wen et al. (2014); Chen et al. (2016); Reddi et al. (2015); Liu and Ziebart (2017) focused on the covariate shift correction methods that are robust to the extreme importance weight incurred by the high dimensionality. The main concern of their work is the robustness of a learning model's prediction performance on the target data to a small amount of high magnitude importance weight. However, there is a paucity of literature on improving the validity and efficiency of statistical inference under covariate shift, with respect to the robustness of the misspecification or poor estimation of the importance weight model. In this paper, we propose an augmented transfer regression learning (ATReL) procedure in the context of covariate shift by specifying flexible machine learning models for the importance weight model and the outcome model. We establish the validity and efficiency of the proposed method under possible misspecification in one of the specified models. We next state the problem of interest and then highlight the contributions of this paper.

1.2 Problem Statement

The source data $\mathcal{S}$, indexed by $S = 1$, consist of n labeled samples with observed response Y and covariates $\boldsymbol{X} = (X_1, \ldots, X_p)$ while the target data $\mathcal{T}$, indexed by $S = 0$, consist of N unlabeled samples with only observed on $\boldsymbol{X}$. We write the full observed data as $\{(S_iY_i, \boldsymbol{X}_i, S_i) : i = 1, 2, \ldots, n + N\}$, where without loss of generality we let the first n observations be from the source population with $S_i = I(1 \leq i \leq n)$ and remaining from the target population. We assume that $(Y, \boldsymbol{X}) \mid S = s \sim p_s(\boldsymbol{x})q(y \mid \boldsymbol{x})$, where $p_s(\boldsymbol{x})$ denotes the probability density measure of $\boldsymbol{X} \mid S = s$ and $q(y \mid \boldsymbol{x})$ is the conditional density of Y given $\boldsymbol{X}$, which is the same across the two populations. The conditional distribution of $Y \mid \boldsymbol{X}$ shared between the two populations, could be complex and difficult to specify correctly. It is often of interest to infer $\mathbb{E}_0(Y \mid \boldsymbol{A}) = \mathbb{E}(Y \mid \boldsymbol{A}, S = 0)$, the model of $Y \sim \boldsymbol{A}$ on $\mathcal{S}$, where $\boldsymbol{A} \in \mathbb{R}^d$ is a sub-vector of $\boldsymbol{X} = (\boldsymbol{A}^\intercal, \boldsymbol{W}^\intercal)^\intercal$ and $\boldsymbol{W}$ consists of the adjustment covariates.

We assume $Y \mid \boldsymbol{X}$ to be the same between $\mathcal{S}$ and $\mathcal{T}$ but the target $Y \mid \boldsymbol{A}$ can be different. This key assumption, as well as why $Y \mid \boldsymbol{A}$ (but not $Y \sim \boldsymbol{X}$) is of our primary interest, is explained and connected with several real-world application fields in Remark 1.

We consider a working model $g(\boldsymbol{A}^{\intercal}\boldsymbol{\beta}) \to \mathbb{E}_0(Y \mid \boldsymbol{A})$ and define the regression parameter $\boldsymbol{\beta}_0$ as the solution to the estimating equation in the target population $S = 0$:

$$\mathbb{E}\left[\boldsymbol{A}\{Y - g(\boldsymbol{A}^{\intercal}\boldsymbol{\beta})\} \mid S = 0\right] \equiv \mathbb{E}_0[\boldsymbol{A}\{Y - g(\boldsymbol{A}^{\intercal}\boldsymbol{\beta})\}] = \mathbf{0}, \tag{1}$$

where $\mathbb{E}_s$ is the expectation operator on the population $S = s$ and $g(\cdot)$ is a link function, e.g. $g(\theta) = \theta$ represents linear regression and $g(\theta) = 1/(1 + e^{-\theta})$ for logistics regression. Although $g(\boldsymbol{A}^{\intercal}\boldsymbol{\beta})$ may often be misspecified, i.e., $\mathbb{E}_0(Y \mid \boldsymbol{A}) \neq g(\boldsymbol{A}^{\intercal}\boldsymbol{\beta})$, the target $\boldsymbol{\beta}$ could still provide reasonable importance measures and risk prediction (Eguchi and Copas, 2002).

Directly solving an empirical estimating equation for (1) using the source data to estimate $\boldsymbol{\beta}_0$ may result in inconsistency due to the covariate shift. It is important to note that even when $\mathbb{E}_0(Y \mid \boldsymbol{A}) = g(\boldsymbol{A}^{\intercal}\boldsymbol{\beta}_0)$ holds, $\mathbb{E}_1\{\boldsymbol{A}(Y - g(\boldsymbol{A}^{\intercal}\boldsymbol{\beta}_0)\}$ may not be zero in the presence of covariate shift. To correct for the covariate shift bias, it is natural to incorporate importance sampling weighting and estimate $\boldsymbol{\beta}_0$ as $\widehat{\boldsymbol{\beta}}_{\mathsf{IW}}$, the solution to the weighted estimating equation

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{\omega}(\boldsymbol{X}_i)\boldsymbol{A}_i\{Y_i - g(\boldsymbol{A}_i^{\intercal}\boldsymbol{\beta})\} = 0, \tag{2}$$

where $\widehat{\omega}(\boldsymbol{X})$ is an estimate for the density ratio $\mathbb{w}(\boldsymbol{X}) = p_0(\boldsymbol{X})/p_1(\boldsymbol{X})$. However, the validity of $\widehat{\boldsymbol{\beta}}_{\mathsf{IW}}$ heavily relies on the consistency of $\widehat{\omega}(\boldsymbol{X})$ for $\mathbb{w}(\boldsymbol{X})$ and can perform poorly when the density ratio model is misspecified or not well estimated.

Remark 1 *There are a number of real-world applications in which one is primarily interested in $Y \sim \boldsymbol{A}$ rather than the model of Y against $\boldsymbol{X} = (\boldsymbol{A}^{\intercal}, \boldsymbol{W}^{\intercal})^{\intercal}$. For example, in EHR and Biobank based genetic studies (Zhou et al., 2022, e.g.), we are interested in the association between certain genetic variants $\boldsymbol{A}$ and disease Y. Adjustment covariates $\boldsymbol{W}$ are taken as EHR proxies (e.g., counts of diagnostic codes) for Y, which are usually strong surrogates of Y. Hence $Y \mid \boldsymbol{A}$ may differ between two distinguished (gender, ethnicity, etc) cohorts $\mathcal{S}$ and $\mathcal{T}$, while $Y \mid \boldsymbol{A}, \boldsymbol{W}$ is transferable from $\mathcal{S}$ to $\mathcal{T}$.*

In clinical studies, $\boldsymbol{A}$ is the treatment or key risk factors at the baseline, Y is a long-term outcome of our interest, and $\boldsymbol{W}$ is some early-point surrogates or mediators (VanderWeele, 2013) such as the post-treatment tumor response rate that can affect the long-term Y and vary between $\mathcal{S}$ and $\mathcal{T}$. In this case, the model of $Y \sim \boldsymbol{A}$ is more useful for clinical decision-making and, thus, of the primary interest. Meanwhile, it is still more reliable to assume $Y \mid \boldsymbol{A}, \boldsymbol{W}$ to be shared by $\mathcal{S}$ and $\mathcal{T}$ but not $Y \mid \boldsymbol{A}$, due to the distributional shift of $\boldsymbol{W}$. In EHR phenotyping studies, $\boldsymbol{A}$ represents widely available codified features and $\boldsymbol{W}$ may include features extracted from narrative notes via natural language processing (NLP), which can be available for research studies but too costly to include when implementing risk models for broad patient populations. This again makes $Y \sim \boldsymbol{A}$ chosen as the target risk model.

1.3 Literature review and our contribution

We propose an augmented transfer regression learning (ATReL) method for doubly robust estimation of a potentially misspecified regression model. Our method leverages an outcome model $m(\boldsymbol{X})$ imputing the missing Y for the target data to augment the importance-weighted estimating equation (2). More importantly, to construct the nuisance density ratio and outcome models, we propose a novel semi-non-parametric (SNP) approach that is more flexible and balanced between model and rate robustness than existing fully parametric and nonparametric approaches. Meanwhile, our SNP framework presents a uniquely challenging problem of (the first-order) bias inflation, as described in equation (6), and overcome it through a novel calibration approach introduced in Section 2.2. In the remaining part of this section, we shall review relevant literature and highlight the novelty and contribution of our proposed SNP approach.

Doubly robust estimators have been extensively studied for missing data and causal inference problems (Bang and Robins, 2005; Qin et al., 2008; Cao et al., 2009; van der Laan and Gruber, 2010; Tan, 2010; Vermeulen and Vansteelandt, 2015). Estimation of average treatment effect on the treated (ATT) can be viewed as an analog to our covariate shift problem. To improve the DR estimation for ATT, Graham et al. (2016) proposed an auxiliary-to-study tilting method and studied its efficiency. Zhao and Percival (2017) proposed an entropy balancing approach that achieves double robustness without augmentation and Shu and Tan (2018) proposed a DR estimator attaining local and intrinsic efficiency. Besides, existing works like Rotnitzky et al. (2012) and Han (2016) are similar to ours in the sense that their parameters of interest are multidimensional regression coefficients. Properties including intrinsic efficiency and multiple robustness have been studied in their work.

However, all these above-reviewed methods used low dimensional parametric nuisance models (e.g., generalized linear models), which are prone to bias due to model misspecification in practice. For example, in biomedical studies, the risks of many age-related diseases are characterized by an underlying biological age and thus change non-linearly with the observed age (Kim et al., 2021). As another example, in EHR phenotyping, total healthcare utilization, as a normalization factor, is typically adjusted through non-linear models (Yu et al., 2017; Liao et al., 2019).

To improve robustness to model misspecifications, Chen et al. (2008) studied the semi-parametric efficient generalized method of moments (GMM) estimation in the presence of auxiliary data, with the nuisance model estimated by the nonparametric sieve estimators; also see Hirano et al. (2003) and Cattaneo (2010). Chernozhukov et al. (2018a) extended classic nonparametric constructions to the modern machine learning setting with cross-fitting. Their proposed double machine learning (DML) framework facilitates the use of general machine learning methods in semiparametric estimation. This general framework has been recently explored under various settings like the semiparametric logistic model (Liu et al., 2021), the conditional average treatment effect (CATE) characterized by the best linear predictor (Semenova and Chernozhukov, 2021), nonparametric CATE predictors (Kennedy, 2020), and handling in-consistent machine learning estimators (Dukes et al., 2021). Among them, Semenova and Chernozhukov (2021) is the most relevant one to our work as they also aimed at estimating regression coefficients of a potentially misspecified parametric model. In contrast to the parametric approaches, the fully nonparametric strategy is free of misspecification of the nuisance models. However, it is impacted by the excessive fitting errors of nonparametric models with higher complexity than parametric models, and thus subject to the so-called "rate double robustness" assumption (Smucler et al., 2019). Usually, classic nonparametric regression methods like kernel smoothing could not achieve the desirable convergence rates even under a moderate dimensionality, as shortly discussed in Remark 2.

Remark 2 *Assume that the nuisance models are a-times continuously differentiable in $\boldsymbol{x} \in \mathbb{R}^p$, and estimated using the standard sieve estimation as studied in Chen et al. (2008); also see Hirano et al. (2003) and Cattaneo (2010). Then by Chen et al. (2008) (see their Assumption 4 and Theorem 7), the doubly robust (or the semiparametric efficient) estimator is possible to achieve the $\sqrt{n}$-consistency only when $p < 2(a+1)$. Consequently, when $a = 1$, i.e., the nuisance models are continuously differentiable, the desirable $\sqrt{n}$-consistency of the estimators constructed using the classic nonparametric regression approaches is not guaranteed even if the dimension of $\boldsymbol{X}$ is as small as* 4.

Though the "curse of dimensionality" discussed in Remark 2 could be relieved by modern machine learning methods like neural network, theoretical justification for the performance of these methods are relatively inadequate. Even though their convergence rates can sometimes be justified according to recent literature (Farrell et al., 2020, 2021), just similar to the classic smoothing regression, these approaches still require conditions of simultaneously having (i) moderate or large enough sample sizes, (ii) relatively low dimensionality, and (iii) strong enough smoothness, to ensure satisfactory estimation errors. This drawback has become a main concern about the fully nonparametric and DML approaches, which can be seen from our numerical studies.

Our proposed SNP framework can be viewed as a mitigation of the parametric and nonparametric methods, enabling better trade-off on model complexity. It specifies the two nuisance models as generalized partially linear models combining a parametric parts and a nonparametric function of a subset in $\boldsymbol{X}$. Compared to the fully parametric and nonparametric strategies, SNP cannot strictly weaken the model and rate robustness conditions. For example, when reducing the complexity of the nonparametric part, it will become less susceptible to the curse of dimensionality but will be more prone to model misspecification at the same time. However, the SNP framework is more flexible for one to specify the nuisance models and attain more balanced construction, neither too simple as the parametric methods, nor too complex like the nonparametric ones.

In addition, the flexibility of the SNP construction allows us to take advantages of the prior knowledge (if available) on the nuisance models. Taking studies of age-related disease as an example, age may have a non-linear effect on phenotypes like Alzheimer's disease while other covariates like gene variants tend to show small and linear effects. In this case, the SNP construction with nonparametric modeling on age and parametric modeling on other factors can give us more balanced model and rate robustness.

As is highlighted in Section 2.2, the proposed SNP approach is not a trivial extension of the two existing strategies since excessive first-order bias can be caused by the fitting error of the nonparametric components under model misspecification. This presents a new challenge in achieving a $\sqrt{n}$-consistent doubly robust estimator that is unique to the SNP framework. To overcome this challenge, our approach constructs the moment equations of the nuisance models more elaborately to *calibrate* them and achieve certain orthogonality that is effective in removing the potential first-order bias. We take the SNP models with kernel or sieve estimator as our main example for realizing this calibration approach and also present other possibilities including the general machine learning construction. We show that the proposed estimator is $n^{1/2}$-consistent and asymptotically normal when at least one nuisance model is correctly specified and both nonparametric components attain the commonly used $o_p(n^{-1/4})$ convergence rate.

Recent work has been developed to construct model doubly robust estimators using high dimensional sparse parametric nuisance models (Smucler et al., 2019; Tan, 2020; Ning et al., 2020; Dukes and Vansteelandt, 2020; Ghosh and Tan, 2020; Liu et al., 2021, e.g.). Compared with the low-dimensional parametric setting, their main challenge is to remove the excessive high-dimensional regularization bias when the nuisance models could be misspecified, under which the Neyman orthogonality is not naturally satisfied as in the DML framework. To address this problem, they calibrate the potentially wrong nuisance models by solving moment equations corresponding to the Neyman orthogonality. Technically speaking, we adapt a similar high-level idea of calibration in our SNP framework. Nevertheless, our problem setup is essentially different from this track of work. In addition, compared to them, the parametric part in our framework can be specified by arbitrary estimating equations, which provides more flexibility, as well as the possibility to achieve intrinsic efficiency as discussed in Section 6 and Appendix C.3.

We note that a similar idea of constructing semi-non-parametric nuisance models has been considered by Chakrabortty (2016) and Chakrabortty and Cai (2018) to improve the efficiency of linear regression under a semi-supervised setting *without* any covariate shift between the labeled and unlabeled data. They proposed a refitting procedure to adjust for the bias incurred by the nonparametric components in the imputation model while our method can be viewed as their extension leveraging the importance weights and imputation models to correct for the bias of each other, which is substantially more challenging. Also, we use the semi-non-parametric models in estimating the parametric parts of the nuisance models, to ensure their correctness and validity. Chakrabortty (2016) and Chakrabortty and Cai (2018) did not actually elaborate on this point and only used parametric regression to estimate the parametric part, which does not guarantee the model double robustness property achieved by our method.

1.4 Outline of the paper

Remaining of the paper will be organized as follow. In Section 2, we introduce the general doubly robust estimating equation, our semi-non-parametric framework, and specific procedures to estimate the parametric and nonparametric components of nuisance models. In Section 3, we present the large sample properties of our proposed ATReL estimator, i.e. its double robustness concerning model specification and estimation. In Section 4, we present simulation results evaluating the finite sample performance of our ATReL estimator and its relevant performance compared with existing methods under various settings. In Section 5, we apply our ATReL estimation on transferring a phenotyping algorithm for bipolar disorder across two EHR cohorts. Finally, we propose and comment on some potential strategies for improving and extending our method in Section 6.

2. Method

2.1 General form of the doubly robust estimating equation

Let $m(\boldsymbol{x})$ denote an imputation model used to approximate $\mu(\boldsymbol{x}) = \mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x})$, which is equal to $\mathbb{E}_0(Y \mid \boldsymbol{X} = \boldsymbol{x})$ and $\mathbb{E}_1(Y \mid \boldsymbol{X} = \boldsymbol{x})$ under our covariate shift assumption, and $\widehat{m}(\boldsymbol{x})$ denote the estimate of $m(\boldsymbol{x})$ by fitting the model to the labeled source data. We augment the importance sampling weighted estimating equation (2) with the term

$$\frac{1}{N}\sum_{i=n+1}^{N+n} \boldsymbol{A}_i\{\widehat{m}(\boldsymbol{X}_i) - g(\boldsymbol{A}_i^{\mathsf{T}}\boldsymbol{\beta})\} - \frac{1}{n}\sum_{i=1}^{n}\widehat{\omega}(\boldsymbol{X}_i)\boldsymbol{A}_i\{\widehat{m}(\boldsymbol{X}_i) - g(\boldsymbol{A}_i^{\mathsf{T}}\boldsymbol{\beta})\}, \tag{3}$$

which results in the augmented estimating equation:

$$\widehat{\boldsymbol{U}}_{\mathsf{DR}}(\boldsymbol{\beta}) \equiv \frac{1}{n}\sum_{i=1}^{n}\widehat{\omega}(\boldsymbol{X}_i)\boldsymbol{A}_i\{Y_i - \widehat{m}(\boldsymbol{X}_i)\} + \frac{1}{N}\sum_{i=n+1}^{N+n}\boldsymbol{A}_i\{\widehat{m}(\boldsymbol{X}_i) - g(\boldsymbol{A}_i^{\mathsf{T}}\boldsymbol{\beta})\} = \mathbf{0}. \tag{4}$$

We denote its solution as $\widehat{\boldsymbol{\beta}}_{\mathsf{DR}}$. Construction (4) is in a similar spirit with the DR estimators of the average treatment effect on the treated studied in existing literature (Graham et al., 2016; Shu and Tan, 2018, e.g.). When the density ratio model is correctly specified and consistently estimated, equation (4) converges to $\mathbb{E}_0[\boldsymbol{A}_i(Y_i - g(\boldsymbol{A}_i^{\mathsf{T}}\boldsymbol{\beta})\}] = 0$ and hence $\widehat{\boldsymbol{\beta}}_{\mathsf{DR}}$ is consistent for $\boldsymbol{\beta}_0$. When the imputation model is correct, the first term of $\widehat{\boldsymbol{U}}_{\mathsf{DR}}(\boldsymbol{\beta})$ in (4) converges to $\mathbf{0}$ and the second term converges to $\mathbb{E}_0[\boldsymbol{A}_i\{\mathbb{E}_0(Y_i \mid \boldsymbol{X}_i) - g(\boldsymbol{A}_i^{\mathsf{T}}\boldsymbol{\beta})\}] = \mathbb{E}_0[\boldsymbol{A}_i\{Y_i - g(\boldsymbol{A}_i^{\mathsf{T}}\boldsymbol{\beta})\}]$ and hence $\widehat{\boldsymbol{\beta}}_{\mathsf{DR}}$ is also expected to be consistent for $\boldsymbol{\beta}_0$. Thus, the augmented estimating equation (4) is doubly robust to the specification of the two nuisance models.

2.2 Semi-non-parametric (SNP) nuisance models

Now we introduce an SNP construction for the nuisance models in (4) that captures more complex effects in $\mathbb{w}(\boldsymbol{X})$ and $\mu(\boldsymbol{X})$ from a subset of $\boldsymbol{X}$, denoted by $\boldsymbol{Z} \in \mathbb{R}^{p_{\boldsymbol{z}}}$, along with simpler effects for the remainder of $\boldsymbol{X}$ that can be explained via linear effects on a finite set of pre-specified functional bases for approximating $\mathbb{w}(\boldsymbol{X})$ and $\mu(\boldsymbol{X})$, respectively denoted by $\boldsymbol{\Psi} = \boldsymbol{\psi}(\boldsymbol{X}) \in \mathbb{R}^{p_{\boldsymbol{\psi}}}$ and $\boldsymbol{\Phi} = \boldsymbol{\phi}(\boldsymbol{X}) \in \mathbb{R}^{p_{\boldsymbol{\phi}}}$. For example, in EHR data analysis, $\boldsymbol{Z}$ may represent measures of *healthcare utilization* which may differ greatly across healthcare systems and have complex effects on patient outcomes. Similarly, in genetic studies such as Cai et al. (2022), the adjustment variable *age* usually has a non-linear relationship with many diseases, e.g., chronic or age-related diseases such as cardiovascular disease and type II diabetes. In contrast, the genetic variants tend to have moderate and linear effects. In this case, it is appealing to choose our $\boldsymbol{Z}$ in the SNP models as *age*.

Under this framework, we specify the SNP nuisance models as

$$\omega(\boldsymbol{X}) = \exp\{\boldsymbol{\Psi}^{\mathsf{T}}\boldsymbol{\alpha} + h(\boldsymbol{Z})\} \quad \text{and} \quad m(\boldsymbol{X}) = g_m\{\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\gamma} + r(\boldsymbol{Z})\} \tag{5}$$

for $\mathbb{w}(\boldsymbol{X})$ and $\mu(\boldsymbol{X})$, where $\boldsymbol{\Psi}^{\mathsf{T}}\boldsymbol{\alpha}$ and $\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\gamma}$ represent parametric components, the unknown functions $h(\boldsymbol{z})$ and $r(\boldsymbol{z})$ represent the nonparametric components, and $g_m(\cdot)$ is a pre-specified link function that may be either the same or different from $g(\cdot)$. To avoid non-identifiability between the parametric and nonparametric parts, we rule out the scenario that some components in $\boldsymbol{\Phi}$ or $\boldsymbol{\Psi}$ are fully determined by $\boldsymbol{Z}$. For example, in our real-world case study, we first specify $\boldsymbol{Z}$ as a subset of covariates in $\boldsymbol{X}$. Then for $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}$, we only include $\boldsymbol{X}_{\text{-}z}$ (covariates in $\boldsymbol{X}$ excluding $\boldsymbol{Z}$) and the interaction terms in $\boldsymbol{X}$, without any basis functions of $\boldsymbol{Z}$ in them. In the asymptotic analysis, common non-singularity conditions like our Assumption A1 (iii) can be used to distinguish the parametric and nonparametric spaces and avoid the non-identifying issue, which is necessary for the convergence of the SNP nuisance estimators.

Correspondingly, we denote the estimators used in (4) as $\widehat{\omega}(\boldsymbol{X}) = \exp\{\boldsymbol{\Psi}^\intercal\widehat{\boldsymbol{\alpha}} + \widehat{h}(\boldsymbol{Z})\}$ and $\widehat{m}(\boldsymbol{X}) = g_m\{\boldsymbol{\Phi}^\intercal\widehat{\boldsymbol{\gamma}} + \widehat{r}(\boldsymbol{Z})\}$. Here and in the sequel, we let $\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}}$ denote the ATReL estimator derived from (4) with this specific construction of $\widehat{m}(\cdot)$ and $\widehat{\omega}(\cdot)$. Unlike $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\gamma}}$, estimation errors of $\widehat{h}(\cdot)$ and $\widehat{r}(\cdot)$ are larger in rate than the desirable parametric rate $n^{-1/2}$ since they are estimated using non-parametric approaches like kernel smoothing. In addition, removing the large non-parametric estimation biases from the biases of the resulting $\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}}$ is particularly challenging due to the bias and variance trade-off in non-parametric regression. To motivate our strategy for mitigating such biases, we consider the estimation of $\boldsymbol{c}^\intercal\boldsymbol{\beta}_0$, an arbitrary linear functional of $\boldsymbol{\beta}_0$ where $\|\boldsymbol{c}\|_2 = 1$, and study the first order (over-fitting) bias incurred by $\widehat{h}(\cdot)$ and $\widehat{r}(\cdot)$ in $\boldsymbol{c}^\intercal\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}}$. The essential bias terms of $n^{1/2}(\boldsymbol{c}^\intercal\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}} - \boldsymbol{c}^\intercal\boldsymbol{\beta}_0)$ arising from the non-parametric components can be asymptotically expressed as

$$\begin{aligned}
\Delta_1 =& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\bar{\omega}(\boldsymbol{X}_i)\boldsymbol{\kappa}_{i,\boldsymbol{\beta}_0}\{Y_i - \bar{m}(\boldsymbol{X}_i)\}\{\widehat{h}(\boldsymbol{Z}_i) - \bar{h}(\boldsymbol{Z}_i)\};\\
\Delta_2 =& \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\bar{\omega}(\boldsymbol{X}_i)\boldsymbol{\kappa}_{i,\boldsymbol{\beta}_0}\breve{g}_m\{\bar{m}(\boldsymbol{X}_i)\}\{\widehat{r}(\boldsymbol{Z}_i) - \bar{r}(\boldsymbol{Z}_i)\}\\
&-\frac{\sqrt{n}}{N}\sum_{i=n+1}^{N+n}\boldsymbol{\kappa}_{i,\boldsymbol{\beta}_0}\breve{g}_m\{\bar{m}(\boldsymbol{X}_i)\}\{\widehat{r}(\boldsymbol{Z}_i) - \bar{r}(\boldsymbol{Z}_i)\},
\end{aligned}\tag{6}$$

where $\boldsymbol{\kappa}_{i,\boldsymbol{\beta}} = \boldsymbol{c}^\intercal\boldsymbol{J}_{\boldsymbol{\beta}}^{-1}\boldsymbol{A}_i$ $\breve{g}_m(a) = \dot{g}_m\{g_m^{-1}(a)\}$, $\dot{g}_m(x) = dg_m(x)/dx > 0$, $\boldsymbol{J}_{\boldsymbol{\beta}} = \mathbb{E}_0\{\dot{g}(\boldsymbol{A}^\intercal\boldsymbol{\beta})\boldsymbol{A}\boldsymbol{A}^\intercal\}$ is the limit of $\widehat{\boldsymbol{J}}_{\boldsymbol{\beta}} = N^{-1}\sum_{i=n+1}^{n+N}\dot{g}(\boldsymbol{A}_i^\intercal\boldsymbol{\beta})\boldsymbol{A}_i\boldsymbol{A}_i^\intercal$, $\bar{\omega}(\boldsymbol{X}) = \exp\{\boldsymbol{\Psi}^\intercal\bar{\boldsymbol{\alpha}} + \bar{h}(\boldsymbol{Z})\}$, $\bar{m}(\boldsymbol{X}) = g_m\{\boldsymbol{\Phi}^\intercal\bar{\boldsymbol{\gamma}} + \bar{r}(\boldsymbol{Z})\}$, $\bar{h}(\boldsymbol{Z})$, $\bar{r}(\boldsymbol{Z})$, $\bar{\boldsymbol{\alpha}}$, $\bar{\boldsymbol{\gamma}}$, and $\bar{\boldsymbol{\beta}}$ are the respective limits of $\widehat{h}(\boldsymbol{Z})$, $\widehat{r}(\boldsymbol{Z})$, $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}}$. These limiting values are not necessarily true model parameter values due to potential model misspecification.

When $m(\boldsymbol{X})$ and $\omega(\boldsymbol{X})$ are specified fully nonparametrically as those in Rothe and Firpo (2015) and Chernozhukov et al. (2018a), a standard cross-fitting strategy can removing terms like Δ_1 and Δ_2 by leveraging $\bar{m}(\boldsymbol{X}) = \mu(\boldsymbol{X})$ and $\bar{\omega}(\boldsymbol{X}) = \mathbb{w}(\boldsymbol{X})$ and utilizing the orthogonality between the "residual" of S or Y on the covariates $\boldsymbol{X}$ and the functional space of $\boldsymbol{X}$. However, simply adopting cross-fitting is not sufficient for the current setting because such orthogonality does not hold due to the potential misspecification of $m(\cdot)$ and $\omega(\cdot)$ in (5). To overcome this challenge, we impose moment condition constraints on the nonparametric components $\bar{r}(\boldsymbol{Z})$ and $\bar{h}(\boldsymbol{Z})$ in that: for any measurable function $f(\cdot)$ of the covariates $\boldsymbol{Z}$,

$$\mathbb{E}_1\left[\mathbb{w}(\boldsymbol{X})\boldsymbol{\kappa}_{\boldsymbol{\beta}_0}\left(Y - g_m\left\{\boldsymbol{\Phi}^\intercal\bar{\boldsymbol{\gamma}} + \bar{r}(\boldsymbol{Z})\right\}\right)f(\boldsymbol{Z})\right] = 0;\tag{7}$$

$$\mathbb{E}_1\left[\exp\{\boldsymbol{\Psi}^\intercal\bar{\boldsymbol{\alpha}} + \bar{h}(\boldsymbol{Z})\}\boldsymbol{\kappa}_{\boldsymbol{\beta}_0}\breve{g}_m\{\mu(\boldsymbol{X})\}f(\boldsymbol{Z})\right] = \mathbb{E}_0\left[\boldsymbol{\kappa}_{\boldsymbol{\beta}_0}\breve{g}_m\{\mu(\boldsymbol{X})\}f(\boldsymbol{Z})\right].\tag{8}$$

Remark 3 *When the density ratio model is correct, moment condition (8) is naturally satisfied, and solving (8) for $\bar{h}(\cdot)$ leads to the true $h_0(\cdot)$. Constructing $\bar{r}(\cdot)$ under the moment condition (7) will enable us to remove excess bias arising from the empirical error in estimating $\bar{h}(\cdot)$. On the other hand, when the imputation model $m(\boldsymbol{X})$ is correct, condition (7) holds and solving (7) for $\bar{r}(\cdot)$ leads to $r_0(\cdot)$. And similarly, constructing $\bar{h}(\cdot)$ under (8) will enable us to remove bias from the error in estimating $\bar{r}(\cdot)$. See our theoretical analyses given in Section 3 and Appendix A for more details on these points.*

Note that when the corresponding nuisance models are wrong, $\bar{r}(\cdot)$ and $\bar{h}(\cdot)$ could only be interpreted as some bias-calibration functions introduced due to technical reasons. This interpretability issue also appeared in recent methods calibrating high-dimensional sparse nuisance models to construct doubly robust estimation (Tan, 2020; Smucler et al., 2019, e.g.). One could show that when the true distribution functions of the data are smooth on $\boldsymbol{Z}$, the above-defined $\bar{r}(\cdot)$ and $\bar{h}(\cdot)$ will be also smooth even under wrongly specified models. This enables us to derive good estimators of them. See Remark A1 for a detailed discussion on this point.

2.3 Estimation Procedure for $\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}}$

We next detail estimation procedures for $\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}}$ under the constraints of the moment conditions (7) and (8). Here we mainly focus on classic local regression approaches for low dimensional and smooth nonparametric components $r(\cdot)$ and $h(\cdot)$. In Appendix C.2, we propose a more general construction procedure that can learn $r(\cdot)$ and $h(\cdot)$ using arbitrary modern machine learning algorithms (e.g. random forest and neural network). Similar to Chernozhukov et al. (2018a), we adopt cross-fitting on the source sample to eliminate the dependence between the estimators and the samples on which they are evaluated and remove the first order bias Δ_1 and Δ_2 through concentration. Specifically, we randomly split the source samples into K equal sized disjoint sets, indexed by $\mathcal{I}_1, \ldots, \mathcal{I}_K$, with $\{1, ..., n\} = \cup_{k=1}^{K}\mathcal{I}_k$ and denote $\mathcal{I}_{\text{-}k} = \{1, .., n\} \setminus \mathcal{I}_k$.

Equations (7) and (8) involve not only $r(\cdot)$ and $h(\cdot)$ but also other unknown parameters that needed to be estimated. To this end, we propose a two-step construction procedure that first obtains preliminary estimators for $\omega(\boldsymbol{X})$ and $m(\boldsymbol{X})$ via standard semiparametric regression as $\widetilde{\omega}^{[\text{-}k]}(\boldsymbol{X}) = \exp\{\boldsymbol{\Psi}^{\intercal}\widetilde{\boldsymbol{\alpha}}^{[\text{-}k]} + \widetilde{h}^{[\text{-}k]}(\boldsymbol{Z})\}$ and $\widetilde{m}^{[\text{-}k]}(\boldsymbol{X}) = g_m\{\boldsymbol{\Phi}^{\intercal}\widetilde{\boldsymbol{\gamma}}^{[\text{-}k]} + \widetilde{r}^{[\text{-}k]}(\boldsymbol{Z})\}$ on $\mathcal{I}_{\text{-}k} \cup \{n+1, \ldots, n+N\}$, where the nonparametric components can be estimated with either sieve (Beder, 1987) or profile kernel (Lin and Carroll, 2006). Here, we take the sieve as an example. Let $\boldsymbol{b}(\boldsymbol{Z})$ be some basis of $\boldsymbol{Z}$ with growing dimension, e.g. the Hermite polynomials, and the partitioning-based series. Denote by $\boldsymbol{\Psi}^{\boldsymbol{b}} = (\boldsymbol{\Psi}^{\intercal}, \boldsymbol{b}(\boldsymbol{Z})^{\intercal})^{\intercal}$ and $\boldsymbol{\Phi}^{\boldsymbol{b}} = (\boldsymbol{\Phi}^{\intercal}, \boldsymbol{b}(\boldsymbol{Z})^{\intercal})^{\intercal}$. We solve

$$\frac{K}{n(K-1)}\sum_{i\in\mathcal{I}_{\text{-}k}} \boldsymbol{\Psi}_i^{\boldsymbol{b}} \exp(\boldsymbol{\theta}_w^{\intercal}\boldsymbol{\Psi}_i^{\boldsymbol{b}}) + \lambda_1(0, \boldsymbol{\theta}_{w,\text{-}1}^{\intercal})^{\intercal} = \frac{1}{N}\sum_{i=n+1}^{n+N} \boldsymbol{\Psi}_i^{\boldsymbol{b}}; \quad \text{with } \boldsymbol{\theta}_w = (\boldsymbol{\alpha}^{\intercal}, \boldsymbol{\eta}^{\intercal})^{\intercal} \tag{9}$$

$$\frac{K}{n(K-1)}\sum_{i\in\mathcal{I}_{\text{-}k}} \boldsymbol{\Phi}_i^{\boldsymbol{b}} \left\{Y_i - g_m(\boldsymbol{\theta}_m^{\intercal}\boldsymbol{\Phi}_i^{\boldsymbol{b}})\right\} + \lambda_2(0, \boldsymbol{\theta}_{m,\text{-}1}^{\intercal})^{\intercal} = \mathbf{0}, \qquad \text{with } \boldsymbol{\theta}_m = (\boldsymbol{\gamma}^{\intercal}, \boldsymbol{\xi}^{\intercal})^{\intercal} \tag{10}$$

to obtain the estimators $\widetilde{\boldsymbol{\theta}}_w^{[-k]} = (\widetilde{\boldsymbol{\alpha}}^{[-k]^\intercal}, \widetilde{\boldsymbol{\eta}}^{[-k]^\intercal})^\intercal$, $\widetilde{\boldsymbol{\theta}}_m^{[-k]} = (\widetilde{\boldsymbol{\gamma}}^{[-k]^\intercal}, \widetilde{\boldsymbol{\xi}}^{[-k]^\intercal})^\intercal$ for $\boldsymbol{\theta}_w$ and $\boldsymbol{\theta}_m$, and $\widetilde{h}^{[-k]}(\boldsymbol{Z}) = \boldsymbol{b}^\intercal(\boldsymbol{Z})\widetilde{\boldsymbol{\eta}}^{[-k]}$, $\widetilde{r}^{[-k]}(\boldsymbol{Z}) = \boldsymbol{b}^\intercal(\boldsymbol{Z})\widetilde{\boldsymbol{\xi}}^{[-k]}$. Here we include ridge penalties to improve the training stability, with the two tuning parameters $\lambda_1, \lambda_2 = o_p(n^{-1/2})$. Suppose that $\widetilde{\omega}^{[-k]}(\boldsymbol{X})$ and $\widetilde{m}^{[-k]}(\boldsymbol{X})$ approach some limiting models denoted as $\omega^*(\boldsymbol{X}) = \exp\{\boldsymbol{\Psi}^\intercal\boldsymbol{\alpha}^* + h^*(\boldsymbol{Z})\}$ and $m^*(\boldsymbol{X}) = g_m\{\boldsymbol{\Phi}^\intercal\boldsymbol{\gamma}^* + r^*(\boldsymbol{Z})\}$. Certainly, we have that $\omega^*(\boldsymbol{X}) = \mathbb{w}(\boldsymbol{X})$ when the density ratio model is correctly specified, and $m^*(\boldsymbol{X}) = \mu(\boldsymbol{X})$ when imputation model is correct. Then we solve the estimating equation for $\boldsymbol{\beta}$:

$$\frac{K}{n(K-1)}\sum_{i\in\mathcal{I}_{-k}}\widetilde{\omega}^{[-k]}(\boldsymbol{X}_i)\boldsymbol{A}_i\{Y_i - \widetilde{m}^{[-k]}(\boldsymbol{X}_i)\} + \frac{1}{N}\sum_{i=n+1}^{N+n}\boldsymbol{A}_i\{\widetilde{m}^{[-k]}(\boldsymbol{X}_i) - g(\boldsymbol{A}_i^\intercal\boldsymbol{\beta})\} = \boldsymbol{0},$$

Denote its solution as $\widetilde{\boldsymbol{\beta}}^{[-k]}$, a preliminary estimator consistent for $\boldsymbol{\beta}_0$ when at least one nuisance model is correct but typically not achieving the desirable parametric rate as our final goal.

One might improve the convergence rate of the remainder bias of $\widetilde{\boldsymbol{\alpha}}^{[-k]}$ and $\widetilde{\boldsymbol{\gamma}}^{[-k]}$ by further using cross-fitting on the nonparametric components in estimating equations (9) and (10); see Newey and Robins (2018). While $\widetilde{\boldsymbol{\alpha}}^{[-k]}$ and $\widetilde{\boldsymbol{\gamma}}^{[-k]}$ can be shown to be $n^{1/2}$-consistent and asymptotically normal under certain smoothness and regularity conditions (Shen, 1997; Chen, 2007; Belloni et al., 2015; Cattaneo et al., 2020), and thus satisfy our requirement (see Assumption 3 and Proposition 1). Therefore, one could simply set $\widehat{\boldsymbol{\alpha}}^{[-k]} = \widetilde{\boldsymbol{\alpha}}^{[-k]}$ and $\widehat{\boldsymbol{\gamma}}^{[-k]} = \widetilde{\boldsymbol{\gamma}}^{[-k]}$ as the estimator of the parametric components in the final nuisance models. Consequently, their limiting (true) values are also identical: $\bar{\boldsymbol{\alpha}} = \boldsymbol{\alpha}^*$ and $\bar{\boldsymbol{\gamma}} = \boldsymbol{\gamma}^*$. In the following part of this section, we choose this construction.

Remark 4 *Equations (9) and (10) are not the only choices for specifying $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$. In our framework, $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ could be estimated through any estimating equations ensuring their $n^{1/2}$-consistency for some limiting parameters equal to the true ones when the corresponding nuisance models are correct. This flexibility is particularly useful when the intrinsic efficiency (Tan, 2010; Rotnitzky et al., 2012) of our estimator is further desirable, i.e. $\boldsymbol{c}^\intercal\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}}$ is the most efficient among all the doubly robust estimators when $\omega(\cdot)$ is correct and $m(\cdot)$ has some wrong specification. Interestingly, we find that one could elaborate an estimating procedure for $\boldsymbol{\gamma}$ to realize this property and shall leave relevant details in Appendix C.3.*

Then as the second step, we construct the calibrated estimating equations for the nonparametric nuisance components based on $\widehat{\boldsymbol{\alpha}}^{[-k]}$, $\widehat{\boldsymbol{\gamma}}^{[-k]}$ and the preliminary estimators. Let $K(\cdot)$ represent some kernel function satisfying $\int_{\mathbb{R}^{p_{\boldsymbol{z}}}} K(\boldsymbol{z})d\boldsymbol{z} = 1$ and define that $K_h(\boldsymbol{z}) = K(\boldsymbol{z}/h)$. Localizing the terms in (7) and (8) with $K_h(\cdot)$, we solve for $r(\boldsymbol{z})$ and $h(\boldsymbol{z})$ respectively from

$$\begin{aligned}
&\frac{1}{|\mathcal{I}_{-k}|}\sum_{i\in\mathcal{I}_{-k}}K_h(\boldsymbol{Z}_i - \boldsymbol{z})\widehat{\boldsymbol{\kappa}}_{i,\widetilde{\boldsymbol{\beta}}^{[-k]}}\widetilde{\omega}^{[-k]}(\boldsymbol{X}_i)\left[Y_i - g_m\left\{\boldsymbol{\Phi}_i^\intercal\widehat{\boldsymbol{\gamma}}^{[-k]} + r(\boldsymbol{z})\right\}\right] = \boldsymbol{0};\\
&\frac{1}{|\mathcal{I}_{-k}|}\sum_{i\in\mathcal{I}_{-k}}K_h(\boldsymbol{Z}_i - \boldsymbol{z})\widehat{\boldsymbol{\kappa}}_{i,\widetilde{\boldsymbol{\beta}}^{[-k]}}\breve{g}_m\{\widetilde{m}^{[-k]}(\boldsymbol{X}_i)\}\exp\left\{\boldsymbol{\Psi}_i^\intercal\widehat{\boldsymbol{\alpha}}^{[-k]} + h(\boldsymbol{z})\right\} \qquad (11)\\
&= \frac{1}{N}\sum_{i=n+1}^{n+N}K_h(\boldsymbol{Z}_i - \boldsymbol{z})\widehat{\boldsymbol{\kappa}}_{i,\widetilde{\boldsymbol{\beta}}^{[-k]}}\breve{g}_m\{\widetilde{m}^{[-k]}(\boldsymbol{X}_i)\}.
\end{aligned}$$

where $\widehat{\boldsymbol{\kappa}}_{i,\boldsymbol{\beta}} = \boldsymbol{c}^{\mathsf{T}} \widehat{\boldsymbol{J}}_{\boldsymbol{\beta}}^{-1} \boldsymbol{A}_i$. Equations in (11) calibrate the nonparametric components to ensure the orthogonality between their score functions and the functional space of $\boldsymbol{Z}$, which is necessary for removing the bias terms introduced in (6). In contrast, the parametric component could include different sets of covariates from $\boldsymbol{Z}$, and there is no need to calibrate them. This substantially distinguishes our framework from existing methods (Smucler et al., 2019; Tan, 2020, e.g.) utilizing a similar calibration idea to handle high dimensional sparse nuisance models.

Remark 5 *If the weights $\widehat{\boldsymbol{\kappa}}_{i,\widehat{\boldsymbol{\beta}}^{[\text{-}k]}} = \boldsymbol{c}^{\mathsf{T}} \widehat{\boldsymbol{J}}_{\widehat{\boldsymbol{\beta}}^{[\text{-}k]}}^{-1} \boldsymbol{A}_i$ have the same sign for a majority of the subjects $i \in \mathcal{I}_{\text{-}k} \cup \{n+1, \ldots, n+N\}$, both equations in (11) have an unique solution for each $\boldsymbol{z}$, denoted as $\widehat{r}^{[\text{-}k]}(\boldsymbol{Z})$ and $\widehat{h}^{[\text{-}k]}(\boldsymbol{Z})$. In practice, it is more likely that $\widehat{\boldsymbol{\kappa}}_{i,\widehat{\boldsymbol{\beta}}^{[\text{-}k]}}$ can be positive for some subjects and negative for others, in which case (11) can be irregular and ill-posed, leading to inefficient estimation. One simple strategy to overcome this is to expand the nuisance imputation models to allow h and r to differ among those with $\widehat{\boldsymbol{\kappa}}_{i,\widehat{\boldsymbol{\beta}}^{[\text{-}k]}} \geq 0$ versus those with $\widehat{\boldsymbol{\kappa}}_{i,\widehat{\boldsymbol{\beta}}^{[\text{-}k]}}$. Specifically, we may solve for*

$$\begin{aligned}
&\frac{1}{|\mathcal{I}_{\text{-}k}|} \sum_{i \in \mathcal{I}_{\text{-}k}} \begin{bmatrix} \widehat{I}_{+,i}^{[\text{-}k]} \\ \widehat{I}_{-,i}^{[\text{-}k]} \end{bmatrix} K_h(\boldsymbol{Z}_i - \boldsymbol{z}) \widehat{\boldsymbol{\kappa}}_{i,\widehat{\boldsymbol{\beta}}^{[\text{-}k]}} \widetilde{\omega}^{[\text{-}k]}(\boldsymbol{X}_i) \left[Y_i - g_m \left\{ \boldsymbol{\Phi}_i^{\mathsf{T}} \widehat{\boldsymbol{\gamma}}^{[\text{-}k]} + \widehat{I}_{+,i}^{[\text{-}k]} r_+(\boldsymbol{z}) + \widehat{I}_{-,i}^{[\text{-}k]} r_-(\boldsymbol{z}) \right\} \right] = \mathbf{0}; \\
&\frac{1}{|\mathcal{I}_{\text{-}k}|} \sum_{i \in \mathcal{I}_{\text{-}k}} \begin{bmatrix} \widehat{I}_{+,i}^{[\text{-}k]} \\ \widehat{I}_{-,i}^{[\text{-}k]} \end{bmatrix} K_h(\boldsymbol{Z}_i - \boldsymbol{z}) \widehat{\boldsymbol{\kappa}}_{i,\widehat{\boldsymbol{\beta}}^{[\text{-}k]}} \breve{g}_m \{ \widetilde{m}^{[\text{-}k]}(\boldsymbol{X}_i) \} \exp \left\{ \boldsymbol{\Psi}_i^{\mathsf{T}} \widehat{\boldsymbol{\alpha}}^{[\text{-}k]} + \widehat{I}_{+,i}^{[\text{-}k]} h_+(\boldsymbol{z}) + \widehat{I}_{-,i}^{[\text{-}k]} h_-(\boldsymbol{z}) \right\} \\
&= \frac{1}{N} \sum_{i=n+1}^{n+N} \begin{bmatrix} \widehat{I}_{+,i}^{[\text{-}k]} \\ \widehat{I}_{-,i}^{[\text{-}k]} \end{bmatrix} K_h(\boldsymbol{Z}_i - \boldsymbol{z}) \widehat{\boldsymbol{\kappa}}_{i,\widehat{\boldsymbol{\beta}}^{[\text{-}k]}} \breve{g}_m \{ \widetilde{m}^{[\text{-}k]}(\boldsymbol{X}_i) \},
\end{aligned} \tag{12}$$

where $\widehat{I}_{+,i}^{[\text{-}k]} = I(\widehat{\boldsymbol{\kappa}}_{i,\widehat{\boldsymbol{\beta}}^{[\text{-}k]}} \geq 0)$ and $\widehat{I}_{-,i}^{[\text{-}k]} = I(\widehat{\boldsymbol{\kappa}}_{i,\widehat{\boldsymbol{\beta}}^{[\text{-}k]}} < 0)$. Then we take

$$\widehat{m}^{[\text{-}k]}(\boldsymbol{X}_i) = g_m\{ \boldsymbol{\Phi}_i^{\mathsf{T}} \widehat{\boldsymbol{\gamma}}^{[\text{-}k]} + \widehat{I}_{+,i}^{[\text{-}k]} r_+(\boldsymbol{Z}_i) + \widehat{I}_{-,i}^{[\text{-}k]} r_-(\boldsymbol{Z}_i) \};$$

$$\widehat{\omega}^{[\text{-}k]}(\boldsymbol{X}_i) = \exp \left\{ \boldsymbol{\Psi}_i^{\mathsf{T}} \widehat{\boldsymbol{\alpha}}^{[\text{-}k]} + \widehat{I}_{+,i}^{[\text{-}k]} h_+(\boldsymbol{Z}_i) + \widehat{I}_{-,i}^{[\text{-}k]} h_-(\boldsymbol{Z}_i) \right\}.$$

With this modification, our construction still effectively removes Δ_1 and Δ_2 as one could trivially analyze the two disjoint sets of samples separately, and combine their convergence rates at last.

After obtaining $\widehat{r}^{[\text{-}k]}(\cdot)$ and $\widehat{h}^{[\text{-}k]}(\cdot)$ for each $k \in \{1, 2, \ldots, K\}$, we take $\widehat{\omega}^{[\text{-}k]}(\boldsymbol{X}_i) = \exp\{ \boldsymbol{\Psi}_i^{\mathsf{T}} \widehat{\boldsymbol{\alpha}}^{[\text{-}k]} + \widehat{h}^{[\text{-}k]}(\boldsymbol{Z}_i) \}$, $\widehat{m}^{[\text{-}k]}(\boldsymbol{X}_i) = g_m\{ \boldsymbol{\Phi}_i^{\mathsf{T}} \widehat{\boldsymbol{\gamma}}^{[\text{-}k]} + \widehat{r}^{[\text{-}k]}(\boldsymbol{Z}_i) \}$, $\widehat{m}(\boldsymbol{X}_i) = K^{-1} \sum_{k=1}^{K} \widehat{m}^{[\text{-}k]}(\boldsymbol{X}_i)$, and plug them into the cross-fitted version of the estimating equation (4) written as:

$$\frac{1}{n} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \widehat{\omega}^{[\text{-}k]}(\boldsymbol{X}_i) \boldsymbol{A}_i \left\{ Y_i - \widehat{m}^{[\text{-}k]}(\boldsymbol{X}_i) \right\} + \frac{1}{N} \sum_{i=n+1}^{N+n} \boldsymbol{A}_i \{ \widehat{m}(\boldsymbol{X}_i) - g(\boldsymbol{A}_i^{\mathsf{T}} \boldsymbol{\beta}) \} = \mathbf{0}. \tag{13}$$

Let the solution of (13) be $\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}}$ and we take $\boldsymbol{c}^{\mathsf{T}} \widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}}$ as the estimation for $\boldsymbol{c}^{\mathsf{T}} \boldsymbol{\beta}_0$. For uncertainty quantification and interval estimation of $\boldsymbol{c}^{\mathsf{T}} \boldsymbol{\beta}_0$, we use a standard multiplier bootstrap approach. Since the first-order impact of the nonparametric component estimators has been removed through calibration, we only need to refit $\widehat{\boldsymbol{\alpha}}^{[\text{-}k]} = \widetilde{\boldsymbol{\alpha}}^{[\text{-}k]}$ and $\widehat{\boldsymbol{\gamma}}^{[\text{-}k]} = \widetilde{\boldsymbol{\gamma}}^{[\text{-}k]}$ with the bootstrap samples and plug them into the bootstrap version of the equation (13) to solve for the resampled $\boldsymbol{c}^{\intercal}\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}}$. Implementation details are presented in Appendix D. Alternatively, one can estimate the standard error of $\boldsymbol{c}^{\intercal}\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}}$ by directly estimating its asymptotic variance (with the method of moments), whose form is given by equation (A7) in Appendix A. Based on the asymptotic normality of $n^{1/2}(\boldsymbol{c}^{\intercal}\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}} - \boldsymbol{c}^{\intercal}\boldsymbol{\beta}_0)$ given by Theorem 1 in the next section, both the bootstrap and the straightforward moment estimation can provide consistent uncertainty quantification and valid confidence intervals (under Assumptions 1–3 to be introduced in the next section).

3. Theoretical analysis

Assume that $\rho = n/N = O(1)$, $K = O(1)$. For any vector $\boldsymbol{a}$, let $\|\boldsymbol{a}\|_2$ represent its ℓ_2-norm. Let $\mathcal{Z}$ and $\mathcal{X}$ represent the domains of $\boldsymbol{Z}$ and $\boldsymbol{X}$ respectively. Assume that dimensionality of $\boldsymbol{A}$, $p_{\boldsymbol{\phi}}$ and $p_{\boldsymbol{\psi}}$ are fixed. We then introduce three sets of assumptions as follows.

Assumption 1 (Regularity conditions) *There exists a constant $C_L > 0$ such that $|\dot{g}(a) - \dot{g}(b)| \leq C_L|a-b|$ and $|\dot{g}_m(a) - \dot{g}_m(b)| \leq C_L|a-b|$ for any $a, b \in \mathbb{R}$. $\boldsymbol{\beta}_0$ belongs to a compact space. $\boldsymbol{A}_i$ belong to a compact set and has a continuous differential density on both populations $\mathcal{S}$ and $\mathcal{T}$. There exists a constant $C_U > 0$ such that $\mathbb{E}_j|Y|^2 + \mathbb{E}_1\bar{\omega}^4(\boldsymbol{X}) + \mathbb{E}_j\breve{g}_m^4\{\bar{m}(\boldsymbol{X})\} + \mathbb{E}_j\|\boldsymbol{\Phi}\|_2^4 + \mathbb{E}_j\|\boldsymbol{\Psi}\|_2^8 < C_U$, for $j \in \{0,1\}$. The information matrix $\boldsymbol{J}_{\boldsymbol{\beta}_0}$ has its all eigenvalues bounded away from 0 and ∞.*

Assumption 2 (Specification of the nuisance models) *At least one of the following two conditions holds (i) $\mathrm{w}(\boldsymbol{X}) = \exp\{\boldsymbol{\Psi}^{\intercal}\boldsymbol{\alpha}_0 + h_0(\boldsymbol{Z})\}$ for some $\boldsymbol{\alpha}_0$ and $h_0(\cdot)$; or (ii) $\mu(\boldsymbol{X}) = g_m\{\boldsymbol{\Phi}^{\intercal}\boldsymbol{\gamma}_0 + r_0(\boldsymbol{Z})\}$ for some $\boldsymbol{\gamma}_0$ and $r_0(\cdot)$.*

Assumption 3 (Estimation error of the nuisance models) *The nuisance estimators satisfy that (i) $n^{1/2}(\widehat{\boldsymbol{\alpha}}^{[\text{-}k]} - \bar{\boldsymbol{\alpha}})$ and $n^{1/2}(\widehat{\boldsymbol{\gamma}}^{[\text{-}k]} - \bar{\boldsymbol{\gamma}})$ is asymptotically normal with mean $\mathbf{0}$ and finite variance; (ii) for every $k \in \{1, 2, \ldots, K\}$ and $j \in \{0,1\}$:*

$$\mathbb{E}_1\{\widehat{h}^{[\text{-}k]}(\boldsymbol{Z}) - \bar{h}(\boldsymbol{Z})\}^2 + \mathbb{E}_j\{\widehat{r}^{[\text{-}k]}(\boldsymbol{Z}) - \bar{r}(\boldsymbol{Z})\}^2 = o_p(n^{-1/2});$$
$$\sup_{\boldsymbol{z}\in\mathcal{Z}} |\widehat{h}^{[\text{-}k]}(\boldsymbol{z}) - \bar{h}(\boldsymbol{z})| + |\widehat{r}^{[\text{-}k]}(\boldsymbol{z}) - \bar{r}(\boldsymbol{z})| = o_p(1).$$

Remark 6 *Assumption 1 is reasonable and commonly used for the asymptotic analysis of M-estimation such as logistic regression (Van der Vaart, 2000). The assumption on the compactness of the domain of $\boldsymbol{A}_i$ could be relaxed to accommodate unbounded covariates with regular tail behaviors. Assumption 2 assumes that at least one nuisance model is correctly specified, and the nonparametric component in the possibly wrong model satisfies the moment constraints (7) or (8). Similar to the classic double robustness condition for the parametric nuisance models (Bang and Robins, 2005; Qin et al., 2008), the parametric part from the wrong model in our method could be arbitrarily specified.*

Assumption 3(ii) assumes that both the nonparametric components have their mean squared errors (MSE) below $o_p(n^{-1/2})$, known as the rate doubly robust assumption (Smucler et al., 2019). With a similar spirit to Chernozhukov et al. (2018a), our Assumption 3 is imposed directly on the calibrated estimators $\widehat{h}^{[\text{-}k]}(\cdot)$ and $\widehat{r}^{[\text{-}k]}(\cdot)$ regardless of their specific estimation procedures, to preserve the generality. Justification of Assumption 3 for the nuisance estimators obtained through smooth regression introduced in Section 2.3 is not standard because the estimating equations in (11) involve the nuisance preliminary estimators impacting the calibrated estimator through their empirical errors. We present this as Proposition 1 and include its proof in Appendix B, obtained leveraging results established in early literature about sieve and kernel methods (Carroll et al., 1998; Chen, 2007). We note that sharper results for the nuisance estimators might be derived using recent literature like Belloni et al. (2015) and Cattaneo et al. (2020); see more discussion in Appendix B.

Proposition 1 *Under Assumption 1 and Assumptions A1–A3 presented in Appendix B about regularity, smoothness, and specification of the sieve and kernel functions, Assumption 3 holds for our mainly proposed nuisance estimators in Section 2.3.*

In Proposition 1, we use under-smoothing on sieve to achieve the asymptotic normality and unbiasedness of the parametric parts $\widehat{\boldsymbol{\alpha}}^{[\text{-}k]}$ and $\widehat{\boldsymbol{\gamma}}^{[\text{-}k]}$; see Assumption A3(i). We note that recent robust bias correction methods for kernel and sieve (Calonico et al., 2018; Qu et al., 2022, e.g.) can be used as an alternative strategy to under-smoothing that could minimize the coverage error and improve robustness to tuning parameter choice when performing inference. Since the variation of $\widehat{\boldsymbol{\alpha}}^{[\text{-}k]}$ and $\widehat{\boldsymbol{\gamma}}^{[\text{-}k]}$ affects the asymptotic behaviour of $\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}}$ (see Remark 7), this could further enhance the validity of our interval estimation of $\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}}$.

Different from the sieve and kernel approaches introduced in Section 2.3, when there is high dimensional $\boldsymbol{Z}$ and the nonparametric components are estimated using modern machine learning approaches like lasso and random forest, our debiased method introduced in Appendix C is used to construct the parametric nuisance components. We demonstrate in Appendix C that such debiased estimation will satisfy Assumptions 3(i) when the machine learning estimators for the nonparametric components have good quality.

Compared to the fully parametric and nonparametric methods, our (model and rate) doubly robust assumptions, i.e., Assumptions 2 and 3 are neither strictly stronger nor strictly weaker. When more covariates are included in the nonparametric parts, Assumption 2 will become weaker. As a price to pay, Assumption 3 will be harder to satisfy due to the increasing dimensionality of $\boldsymbol{Z}$. Thus, the SNP models need to be carefully designed to make Assumptions 2 and 3 reasonable at the same time. This has a similar spirit with the well-known bias-variance trade-off in statistical learning theory. The fully parametric and nonparametric strategies are actually the two extreme choices in terms of the model complexity. The main advantage of the SNP framework is that one can specify the nuisance models more flexibly to balance the model correctness and rate conditions, achieving a better trade-off on model complexity.

Now we present the main theoretical results about the consistency and asymptotic validity of our estimator $\boldsymbol{c}^{\intercal}\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}}$ in Theorem 1 with its proof found in Appendix A.

Theorem 1 *Under Assumptions 1 to 3, it holds that* $\|\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}}-\boldsymbol{\beta}_0\|_2=o_p(1)$ *and*

$$\sqrt{n}(\boldsymbol{c}^{\intercal}\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}}-\boldsymbol{c}^{\intercal}\boldsymbol{\beta}_0)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{c}^{\intercal}F_i^{\mathcal{S}}+\frac{\sqrt{n}}{N}\sum_{n+1}^{n+N}\boldsymbol{c}^{\intercal}F_i^{\mathcal{T}}+\sqrt{n}\boldsymbol{\zeta}_{\alpha}^{\intercal}(\widehat{\boldsymbol{\alpha}}-\bar{\boldsymbol{\alpha}})+\sqrt{n}\boldsymbol{\zeta}_{\gamma}^{\intercal}(\widehat{\boldsymbol{\gamma}}-\bar{\boldsymbol{\gamma}})+o_p(1),$$

where $F_i^{\mathcal{S}} = \bar{\omega}(\boldsymbol{X}_i)\boldsymbol{A}_i\left\{Y_i - \bar{m}(\boldsymbol{X}_i)\right\}$, $F_i^{\mathcal{T}} = \boldsymbol{A}_i\{\bar{m}(\boldsymbol{X}_i) - g(\boldsymbol{A}_i^{\intercal}\boldsymbol{\beta})\}$,

$$\boldsymbol{\zeta}_\alpha = \mathbb{E}_1\bar{\omega}(\boldsymbol{X})\boldsymbol{\kappa}_{\beta_0}\left[Y - g_m\{\boldsymbol{\Phi}^{\intercal}\bar{\boldsymbol{\gamma}} + \bar{r}(\boldsymbol{Z})\}\right]\boldsymbol{\Psi},$$
$$\boldsymbol{\zeta}_\gamma = \mathbb{E}_1\bar{\omega}(\boldsymbol{X})\boldsymbol{\kappa}_{\beta_0}\breve{g}_m\{\bar{m}(\boldsymbol{X})\}\boldsymbol{\Phi} - \mathbb{E}_0\boldsymbol{\kappa}_{\beta_0}\breve{g}_m\{\bar{m}(\boldsymbol{X})\}\boldsymbol{\Phi},$$

$\widehat{\boldsymbol{\alpha}} = K^{-1}\sum_{k=1}^{K}\widehat{\boldsymbol{\alpha}}^{[\text{-}k]}$, *and* $\widehat{\boldsymbol{\gamma}} = K^{-1}\sum_{k=1}^{K}\widehat{\boldsymbol{\gamma}}^{[\text{-}k]}$. *Consequently,* $n^{1/2}(\boldsymbol{c}^{\intercal}\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}} - \boldsymbol{c}^{\intercal}\boldsymbol{\beta}_0)$ *weakly converges to Gaussian distribution with mean* $\mathbf{0}$ *and the variance given by equation (A7) in Appendix A.*

Remark 7 *When Assumption 2(i) holds, i.e. the density ratio is correctly specified, we have that* $\boldsymbol{\zeta}_\gamma = \mathbf{0}$ *so* $\widehat{\boldsymbol{\gamma}}^{[\text{-}k]} - \bar{\boldsymbol{\gamma}}$ *has no impact on the asymptotic expansion* $\boldsymbol{c}^{\intercal}\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}}$. *Similarly, when the imputation model is correct,* $\boldsymbol{\zeta}_\alpha = \mathbf{0}$ *and* $\widehat{\boldsymbol{\alpha}}^{[\text{-}k]} - \bar{\boldsymbol{\alpha}}$ *has no impact on* $\boldsymbol{c}^{\intercal}\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}}$. *Thus, when both nuisance models are correctly specified, estimating equations for the nuisance estimators will not affect the asymptotic efficiency of our estimator.*

4. Simulation studies

We conduct simulation studies to investigate the performance of the ATReL method and compare it with existing doubly robust approaches. We consider four different data-generating mechanisms concerning the specification of the nuisance models. Throughout, we let $n = 500$ and $N = 1000$. In this and the next sections, we take both the link functions as $g(a) = g_m(a) = e^a/(1 + e^a)$. To generate the data, we first generate $\boldsymbol{V} = (V_1, V_2, ..., V_7)^{\intercal}$ from $\mathcal{N}(\mathbf{0}, \Sigma_V)$ where $\Sigma_V = (\sigma_{ij})_{7\times 7}$, $\sigma_{ij} = 1$ when $i = j$, $\sigma_{ij} = 0.3$ when (i, j) or $(j, i) \in \{(1, 2), (1, 3), (3, 4), (3, 5)\}$, $\sigma_{ij} = 0.15$ when (i, j) or $(j, i) \in \{(1, 6), (1, 7), (5, 6), (5, 7)\}$, and $\sigma_{ij} = 0$ otherwise. Then we obtain each $\widetilde{X}_j$ by truncating V_j with $(-1.5, 1.5)$ and standardizing it, and take

$$\boldsymbol{W} = \left\{1, \exp(0.5\widetilde{X}_1), \frac{\widetilde{X}_2}{1 + \exp(\widetilde{X}_3)}, \left(\frac{\widetilde{X}_1\widetilde{X}_3}{5} + 0.6\right)^3, \widetilde{X}_4, ..., \widetilde{X}_7\right\}^{\intercal}$$

as a nonlinear transformation of $\widetilde{\boldsymbol{X}} = (1, \widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_7)^{\intercal}$. Based on this, we consider four configurations for the underlying data-generating mechanisms introduced below as the configurations indexed by (i)–(iv). First, we set $Z = \widetilde{X}_1$ and generate the source indication S given $\widetilde{\boldsymbol{X}}$ by $\mathrm{P}(S = 1 \mid \widetilde{\boldsymbol{X}}) = g_m\{\mathbf{a}_w^{\intercal}\boldsymbol{W} + \mathbf{a}_x^{\intercal}\widetilde{\boldsymbol{X}} + h_x(Z)\}$ where

(i) $\mathbf{a}_w = (-1, 0, -0.4, -0.4, -0.15, -0.15, 0, 0)^{\intercal}$, $\mathbf{a}_x = \mathbf{0}$, and $h_x(Z) = 0.6Z^2 \cdot I(|Z| < 1.5) + \{0.6(|Z| - 1.5) + 1.35\} \cdot I(|Z| \geq 1.5)$.

(ii) The same as Configurations (i).

(iii) $\mathbf{a}_w = \mathbf{0}$, $\mathbf{a}_x = (0, -0.2, -0.4, -0.4, -0.2, -0.2, 0, 0)^{\intercal}$, and $h_x(Z) = 0.5|Z|^3 \cdot I(|Z| < 1.5) + \{0.5 \cdot 1.5^3 + (|Z| - 1.5)\} \cdot I(|Z| \geq 1.5)$.

(iv) $\mathbf{a}_w = \mathbf{0}$, $\mathbf{a}_x = (0, -0.4, -0.4, -0.4, -0.15, -0.15, 0, 0)^{\intercal}$, and $h_x(Z) = 0$.

In Configurations (i) and (ii), set the observed covariates as $\boldsymbol{X} = (1, X_1, X_2, \ldots, X_7)^{\intercal}$ where

$$X_2 = 0.8\widetilde{X}_2 - 0.2sin(\frac{3}{4}\pi\widetilde{X}_1) \cdot I(S = 0); \quad X_3 = 0.8\widetilde{X}_3 - 0.2sin(\frac{3}{4}\pi\widetilde{X}_1) \cdot I(S = 0),$$

and $X_j = \widetilde{X}_j$ for all $j \neq 2, 3$. While in Configurations (iii) and (iv), we simply set $\boldsymbol{X} = \widetilde{\boldsymbol{X}}$. Then we generate Y by $\mathrm{P}(Y = 1 \mid \boldsymbol{X}) = g_m\{\mathbf{b}_w^{\intercal}\boldsymbol{W} + \mathbf{b}_x^{\intercal}\boldsymbol{X} + r_x(Z)\}$, where

(i) $\mathbf{b}_w = \mathbf{0}$, $\mathbf{b}_x = (0, 0.5, 0.5, 0.5, 0.3, 0.3, 0.15, 0.15)^{\intercal}$, $r_x(Z) = -0.4 \cdot sin(\frac{3}{4}\pi Z)$.

(ii) $\mathbf{b}_w = \mathbf{0}$, $\mathbf{b}_x = (0, 0.5, 0.5, 0.5, 0.3, 0.3, 0.15, 0.15)^{\intercal}$, $r_x(Z) = 0$.

(iii) $\mathbf{b}_w = (-0.5, 0.5, 0.8, 0.3, -0.3, -0.2, 0.15, 0.15)^{\intercal}$, $\mathbf{b}_x = \mathbf{0}$, $r_x(Z) = -0.6 \cdot sin(\frac{3}{4}\pi Z)$.

(iv) $\mathbf{b}_w = (-0.8, 0.5, 0.5, 0.5, 0.3, 0.3, 0.15, 0.15)^{\intercal}$, $\mathbf{b}_x = \mathbf{0}$, $r_x(Z) = -0.4 \cdot sin(\frac{3}{4}\pi Z)$.

In all four configurations, we set $\boldsymbol{A} = (1, X_1, ..., X_3)^{\intercal}$. For each generated data set, we fit the following nuisance models to estimate $\boldsymbol{\beta}_0$:

(a) Parametric nuisance models (Parametric): the importance weight model is chosen as the logistic model of S against $\boldsymbol{\Psi} = \boldsymbol{X}$ and the imputation model is specified as the logistic model of Y against $\boldsymbol{\Phi} = \boldsymbol{X}$.

(b) SNP nuisance models (ATReL): $\mathrm{P}(S = 1 \mid \boldsymbol{X}) = g_m\{\boldsymbol{\Psi}^{\intercal}\boldsymbol{\alpha} + h(Z)\}$ and $\mathrm{P}(Y = 1 \mid \boldsymbol{X}) = g_m\{\boldsymbol{\Phi}^{\intercal}\boldsymbol{\gamma} + r(Z)\}$, where $\boldsymbol{\Psi} = \boldsymbol{X}_{\text{-}1}$, $\boldsymbol{\Phi} = \boldsymbol{X}_{\text{-}1}$, and $Z = X_1$, where $\boldsymbol{X}_{\text{-}j}$ represents the vector of components in $\boldsymbol{X}$ excluding X_j.

(c) Double machine learning with flexible basis expansions ($\mathrm{DML}_{\mathsf{BE}}$): the nuisance models regress Y or S on features combining together $\boldsymbol{X}$, natural splines of each X_j with order 4 and all the interaction terms of these natural splines. Due to the high dimensionality of the bases, we use a combination of ℓ_1 and ℓ_2 penalties for regularization.

(d) Double machine learning with kernel machine ($\mathrm{DML}_{\mathsf{KM}}$): both models are estimated using the support vector machine with the radial basis function kernel.

Our data generation and model specification have a similar spirit as Kang and Schafer (2007) and Tan (2020). In Configurations (i) and (ii), our SNP imputation model correctly characterizes $Y \mid \boldsymbol{X}$ while our importance weight model is misspecified. Parametric approach (a) has its imputation model correctly specified under Configuration (ii) but misses the nonlinear function $r(Z)$ under (i). Also note that under (ii), the nonparametric component included in the imputation model of our method is redundant for the logistic linear model of $\mathrm{P}(Y = 1 \mid \boldsymbol{X})$. Similar logic applies to Configurations (iii) and (iv) with the status of the imputation model and importance weight model interchanged. More implementing details of (a)–(d) are presented in Appendix D.

Performance of the four approaches is evaluated through root mean square error, bias, and coverage probability of the 95% confidence interval in terms of estimating and inferring

$\beta_0, \beta_1, \beta_2, \beta_3$, as summarized in Tables A2–A5 of Appendix D for configurations (i)–(iv) respectively. The mean square error and absolute bias averaged over the target parameters, and the maximum deviance of the coverage probability from the nominal level 0.95 among all parameters are summarized in Table 1.

Table 1: Average root mean square error (RMSE), average absolute bias (|Bias|), and maximum deviance of coverage probability (CP) of the constructed CI from its nominal level 0.95 over all parameters of the doubly robust estimators with different modeling strategies for the nuisance models: Parametric, ATReL, $\mathrm{DML}_{\mathsf{BE}}$ and $\mathrm{DML}_{\mathsf{KM}}$ under Configurations (i)–(iv), as introduced in Section 4.

Configurations		Parametric	ATReL	$\mathrm{DML}_{\mathsf{BE}}$	$\mathrm{DML}_{\mathsf{KM}}$
(i)	Average RMSE	0.141	**0.123**	0.179	0.153
	Average |Bias|	0.065	**0.030**	0.108	0.058
	Deviance of CP	0.04	0.02	0.11	0.10
(ii)	Average RMSE	**0.117**	**0.123**	0.186	0.148
	Average |Bias|	**0.005**	0.016	0.114	0.061
	Deviance of CP	0.04	0.02	0.13	0.05
(iii)	Average RMSE	0.207	**0.134**	0.142	0.144
	Average |Bias|	0.092	**0.019**	0.036	0.062
	Deviance of CP	0.13	0.02	0.02	0.09
(vi)	Average RMSE	**0.131**	**0.122**	0.145	0.128
	Average |Bias|	**0.005**	**0.009**	0.058	0.044
	Deviance of CP	0.01	0.02	0.22	0.09

Under all configurations, ATReL achieves better performance, especially at least 48% smaller average bias, than the two DML approaches. Also, ATReL performs well in interval estimation with coverage probabilities on all parameters under all configurations falling in ±0.02 of the nominal level. In comparison, the Parametric method fails obviously on interval estimation of β_1 under (iii) because in the importance weighting model, the nonparametric component is placed on the corresponding predictor. The two DML approaches fail apparently on interval estimation of certain parameters, for example, Additive fails on interval estimation of β_0 under Configuration (i), (ii), and (iv), and kernel machine fails on β_1 under Configuration (i), (iii) and (iv). These demonstrate that our method achieves better balance on the model complexity than the fully nonparametric/machine learning constructions, leading to consistently better performance on point and interval estimation.

Our method has a significantly smaller root mean square error than Parametric under (i) (relative efficiency being 0.89) and (iii) (relative efficiency being 0.65), with nonlinear effects in the nuisance models captured by our method and missed by the parametric approach. Under these two configurations, our method also has a smaller average absolute bias than Parametric (55% under (i) and 79% under (iii)). While for (ii) and (iv) with the nonparametric components in our construction being redundant, the performance of our method is close to the parametric approach. Thus, our nonparametric components modeling helps to reduce bias and improve estimation efficiency in the presence of nonlinear effects while they basically do not hurt the efficiency when being redundant.

One should note that the above-discussed advantages of our method rely on Assumption 2, the correctness of the SNP nuisance models. When both models are severely misspecified, one cannot expect ATReL to be valid and efficient. We demonstrate this limitation in Appendix D, through an additional simulation study in which the nuisance models' non-linearity creates large covariate shift bias and is not adequately captured by our SNP construction. See Appendix D for more details about data generation and Table A1 for the results. In this scenario, ATReL produces a large estimation bias, e.g., around 1.3 on β_2 that occupies a large proportion of its RMSE. The Parametric method fails in a similar way. In comparison, $\mathrm{DML}_{\mathsf{KM}}$ with fully nonparametric nuisance estimators has a small bias of 0.05 on β_2. Thus, in practice, one needs to be aware of the danger of severe model misspecification when specifying the nuisance models in ATReL.

5. Transfer EHR phenotyping of rheumatoid arthritis across different time windows

The growing availability of EHR data opens more opportunities for translational biomedical research (Kohane et al., 2012). However, a major obstacle to realizing the full translational potential of EHR is the lack of precise definitions of disease phenotypes needed for clinical studies. With a small number of gold standard labels for phenotypes, machine learning phenotyping algorithms based on both codified EHR features and clinical note mentions extracted using natural language processing (NLP) have been derived to improve the phenotype definition Liao et al. (2019). For example, several phenotyping algorithms for rheumatoid arthritis (RA), a common autoimmune disease, have been developed and validated at multiple institutions in recent years (Liao et al., 2010; Carroll et al., 2012; Yu et al., 2017). Once the phenotyping algorithms become available, they are used to classify disease status for downstream tasks such as genomic association studies using EHR-linked biobank data (Kohane, 2011).

Once a phenotyping algorithm is developed, it is often used repeatedly to classify disease status for patients in an EHR database which is often updated over time. For example, the RA algorithm developed by Liao et al. (2010) at Mass General Brigham (MGB) was trained in 2009 and validated again in 2020 Huang et al. (2020). Significant changes have occurred between 2009 and 2020: the EHR system at MGB was switched to EPIC and the International Classification of Diseases (ICD) system was changed from version 9 to version 10 around 2015 - 2016. Although the algorithm trained in Liao et al. (2010) appears to have stable performance for the 2020 data Huang et al. (2020), it had poorer performance in 2016. Thus, we investigated to what extent transfer learning can be used to automatically update the phenotyping algorithm over time, particularly in 2016 when the algorithm performed less well. To this end, we considered training an RA EHR phenotyping algorithm to classify RA status for patients with EHR data from 2016 at MGB using training data from 2009.

There are a total of 200 labeled patients with true RA status, Y, manually annotated via chart review. There are a total of $p = 9$ demographic or EHR features, $\boldsymbol{X}$, available for training RA algorithm, including the total healthcare utilization (X_1), NLP count of RA

(X_2), NLP mention of tumor necrosis factor (TNF) inhibitor (X_3), NLP mention of bone erosion (X_4), age (X_5), gender (X_6), ICD count of RA (X_7), presence of TNF inhibitor prescription (X_8), and tested negative for rheumatoid factor (X_9), where we use $x \to \log(x+1)$ transformation for all count variables. Since NLP mentions of clinical terms are less sensitive to changes to the EHR coding system, we aim to develop an NLP feature-only model for predicting Y using $\boldsymbol{A} = (X_1, X_2, X_3, X_4)^\intercal$, for the EHR cohort of 2016 using labeled data from 2009 via transfer learning. Due to the co-linearity among $\boldsymbol{A}$, we convert X_2 into its orthogonal complement to X_1. For simplicity, we still denote the transformed covariates as $(X_1, X_2, X_3, X_4)^\intercal$.

We implemented the doubly robust transfer learning approaches introduced in Section 4, including Parametric, ATReL, $\mathrm{DML}_{\mathsf{BE}}$, and $\mathrm{DML}_{\mathsf{KM}}$. The specific construction of the nuisance models in the four approaches is presented in Appendix E. We also include the logistic model for $Y \sim \boldsymbol{A}$ simply fitted on the source data without adjusting for covariate shift, named Source. For ATReL, we choose Z as the NLP count of RA for non-parametric modeling as it is believed to be the most predictive feature (see Table A6 in Appendix). If the pattern of effects was similar across all predictors in $\boldsymbol{X}$, this choice could help us to capture as much non-linear effect as possible under the SNP framework.

To evaluate the performance of the transfer learning, we additionally performed chart-reviewing on 150 subjects from the target population in 2016, denoted as $\mathcal{L}_{16}$. We fit a logistic regression $Y \sim \boldsymbol{A}$ using these labeled observations in $\mathcal{L}_{16}$ and denote the estimate for $\boldsymbol{\beta}$ as $\widehat{\boldsymbol{\beta}}_{\mathsf{Valid}}$ to serve as a gold standard benchmark. Fitted intercepts and coefficients of all methods are presented in Table A6 of Appendix E. To evaluate the estimation performance of a derived estimator $\widehat{\boldsymbol{\beta}}$ according to our practical needs, we calculate the following metrics:

AUC. Area under the receiver operating characteristic (ROC) curve evaluated with the labels. For the Target estimator $\widehat{\boldsymbol{\beta}}_{\mathsf{Valid}}$, we use repeated sample-splitting for evaluation.

RMSPE. Relative mean square prediction error to $\widehat{\boldsymbol{\beta}}_{\mathsf{Valid}}$ evaluated on the target data:

$$\frac{\widehat{\mathbb{E}}_0\{g(\boldsymbol{A}^\intercal\widehat{\boldsymbol{\beta}}_{\mathsf{Valid}}) - g(\boldsymbol{A}^\intercal\widehat{\boldsymbol{\beta}})\}^2}{\widehat{\mathbb{E}}_0\{g(\boldsymbol{A}^\intercal\widehat{\boldsymbol{\beta}}_{\mathsf{Valid}})\}^2}.$$

CC with $\widehat{\boldsymbol{\beta}}_{\mathsf{Valid}}$. Classifier's correlation with that of $\widehat{\boldsymbol{\beta}}_{\mathsf{Valid}}$:

$$\widehat{\mathrm{Corr}}_0\left\{I\left(g(\boldsymbol{A}^\intercal\widehat{\boldsymbol{\beta}}_{\mathsf{Valid}}) \geq \widehat{\mathbb{E}}_0[g(\boldsymbol{A}^\intercal\widehat{\boldsymbol{\beta}}_{\mathsf{Valid}})]\right), I\left(g(\boldsymbol{A}^\intercal\widehat{\boldsymbol{\beta}}) \geq \widehat{\mathbb{E}}_0[g(\boldsymbol{A}^\intercal\widehat{\boldsymbol{\beta}})]\right)\right\},$$

FCR v.s. $\widehat{\boldsymbol{\beta}}_{\mathsf{Valid}}$. False classification rate of $\widehat{\boldsymbol{\beta}}$'s classifier against that of $\widehat{\boldsymbol{\beta}}_{\mathsf{Valid}}$:

$$\widehat{\mathbb{P}}_0\left\{I\left(g(\boldsymbol{A}^\intercal\widehat{\boldsymbol{\beta}}_{\mathsf{Valid}}) \geq \widehat{\mathbb{E}}_0[g(\boldsymbol{A}^\intercal\widehat{\boldsymbol{\beta}}_{\mathsf{Valid}})]\right) \neq I\left(g(\boldsymbol{A}^\intercal\widehat{\boldsymbol{\beta}}) \geq \widehat{\mathbb{E}}_0[g(\boldsymbol{A}^\intercal\widehat{\boldsymbol{\beta}})]\right)\right\}.$$

Here $\widehat{\mathbb{E}}_0$, $\widehat{\mathbb{P}}_0$, and $\widehat{\mathrm{Corr}}_0(\cdot,\cdot)$ represent the empirical expectation, probability measure, and Pearson correlation on the target population. Evaluation results obtained with the target data and the validation labels are presented in Table 2. Our ATReL method attains the smallest estimation error among all the methods under comparison, with its relative efficiency of RMSPE being 0.21 to the naive source estimator, 0.23 to doubly robust estimator with parametric nuisance models, 0.17 to DML with flexible basis expansions, and 0.46 to DML with kernel machine. Also, among Source and all the transfer learning estimators, ATReL produces the largest AUC, as well as the closest classifiers to the gold standard target data estimator, i.e. attaining the largest CC with $\widehat{\boldsymbol{\beta}}_{\mathsf{Valid}}$ and smallest FCR v.s. $\widehat{\boldsymbol{\beta}}_{\mathsf{Valid}}$. Thus, by trading off the parametric and nonparametric modeling strategies in a better way to adjust for the covariate shift, our method achieves better estimation performance than all existing methods.

Table 2: Estimation performance of the source or transfer learning estimators evaluated with the validation labeled data and validation estimator denoted as Target. All included methods are as described in Sections 4 and 5. The evaluation metrics, as introduced in Section 5, include AUC: area under the ROC curve; RMSPE: relative mean square prediction error; CC with $\widehat{\boldsymbol{\beta}}_{\mathsf{Valid}}$: classifier's correlation with that of $\widehat{\boldsymbol{\beta}}_{\mathsf{Valid}}$; FCR v.s. $\widehat{\boldsymbol{\beta}}_{\mathsf{Valid}}$: false classification rate against $\widehat{\boldsymbol{\beta}}_{\mathsf{Valid}}$.

	Source	Parametric	ATReL	DML$_{\mathsf{BE}}$	DML$_{\mathsf{KM}}$	Target
AUC	0.908	0.904	**0.916**	0.907	**0.911**	0.922
RMSPE	0.052	0.048	**0.011**	0.064	0.024	0
Prevalence	0.376	0.336	0.323	0.329	0.330	0.340
CC with $\widehat{\boldsymbol{\beta}}_{\mathsf{Valid}}$	0.89	0.88	**0.97**	0.91	0.93	1
FCR v.s. $\widehat{\boldsymbol{\beta}}_{\mathsf{Valid}}$	0.05	0.06	**0.01**	0.05	0.03	0

6. Discussion

Contribution and limitation. In this paper, we propose ATReL, a transfer regression learning approach using an imputation model to augment the importance weighting equation to achieve double robustness. Interestingly, our target $\boldsymbol{\beta}$ is defined as the solution to some pre-specified estimating equations but not necessarily underlying true models. This fact connects our work to the comprehensive semiparametric inference literature. Moreover, we propose a novel SNP framework to construct the two nuisance models that achieves a better model complexity trade-off than existing doubly robust or DML approaches. To control the excessive first-order bias incurred by the nonparametric component under model misspecification, appearing as a unique challenge of the SNP construction, we develop a novel set of calibrating estimating equations constructed through a two-step procedure. The $n^{1/2}$-consistency of our proposed estimator is guaranteed by a hybrid of the model double robustness of the parametric component and the rate double robustness of the nonparametric component. Simulation studies and a real example also demonstrate that our method is more robust and efficient than the existing methods.

Note that the covariate shift studied in this paper is closely relevant to the standard causal inference problems (Imbens and Rubin, 2015, e.g.). Our indicator of the source data $S \in \{0, 1\}$ can be viewed as an analog of the treatment variable. We shall comment on the connection between the standard causal assumptions and ours, which sheds light on the potential generalization of the proposed SNP framework. First, the unconfoundedness assumption corresponds to our assumption that the distribution of $Y \mid \boldsymbol{X}$ remains the same between the source and target populations. Second, the strict overlap (positivity) assumption, i.e., the importance weights staying away from δ and δ^{-1} for some $\delta > 0$, is actually implied by our Assumption A1 (i, ii). At last, under our data generation mechanism with independent samples from both $\mathcal{S}$ and $\mathcal{T}$, the consistency assumption and the stable unit treatment value assumption (SUTVA) are as given.

Proper choices of $\boldsymbol{Z}$ in the nonparametric part are crucial. In our additional simulation described in Appendix D, the failure of capturing the strong non-linear effects in the nuisance models causes severe bias in our method. Besides leveraging potential prior knowledge mentioned in Section 2.2, the users could try to put (moderately) more variables in $\boldsymbol{Z}$ if model misspecification is really the main concern in their applications. It is desirable to further develop data-driven approaches for model selection among different choices on $\boldsymbol{Z}$. This could be more challenging than the usual model selection since the purpose here is to find SNP nuisance estimators leading to a small bias on the target estimator. In addition, compared to the parametric doubly robust and DML approaches, the implementation of ATReL, including fitting the preliminary SNP nuisance models, and calibrating the estimators, is more complicated and involve more tuning parameters to select. This could raise concerns about the stability of ATReL's performance, as well as its utility in practice.

The covariate shift correction problem is closely related to the estimation of ATT, whose semiparametric efficiency under both correct nuisance models has been well-studied (Hahn, 2004; Shu and Tan, 2018). So one could further study the efficiency of our estimator by extending the previous work on ATT to our case. Also, we currently specify the parametric parts in our nuisance models with some low-dimensional bases. Our method could be extended to address high-dimensional $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ with sparse coefficients. For this purpose, some recent bias-correction methods like Kozbur (2021) and Tan (2020) need to be incorporated to handle the excessive estimation errors of the nonparametric and the parametric parts respectively. Meanwhile, we find more concrete ideas and directions to generalize our current proposal and introduce them shortly as below with more details presented in Appendix C.

Sieve or modern machine learning estimation of the nonparametric parts. We propose some other choices in constructing the nuisance estimators alternative to the kernel smoothing method introduced in Section 2.3. Detailed construction procedures under these choices, including sieve and modern (black-box) machine learning algorithms are presented in Appendix C. First, we note that the sieve can be naturally incorporated to solve the calibrated equations in (11) and achieve the same convergence properties as the kernel estimators. More importantly, we propose a construction procedure using arbitrary modern (nonparametric) machine learning algorithms to learn the nonparametric components in the nuisance models under our framework. This is substantially more challenging than the kernel or sieve constructions since we consider arbitrary black-box machine learning algorithms with no special forms, and thus it becomes more involving to derive nuisance estimators satisfying the moment conditions (7) and (8). To our best knowledge, a similar problem has not been solved in the existing literature.

The $N \gg n$ scenario. In many application fields like EHR phenotyping studied in this paper, the sample size of unlabeled data N can usually be much larger than the size of labeled data n. Analysis of our method under such a $N \gg n$ scenario is of particular interest. It has been established that semi-supervised learning with $N \gg n$ unlabeled samples enables estimating various types of target parameters more efficiently than the supervised method (Kawakita and Kanamori, 2013; Azriel et al., 2016; Gronsbell and Cai, 2018; Chakrabortty and Cai, 2018; Gronsbell et al., 2020, e.g.). However, existing work is restricted to the setting where the unlabeled and labeled data are from the same population. In the presence of covariate shift, it is of interest to further investigate whether having $N \gg n$ (unlabeled) target samples would benefit our estimator. When the importance weight model is correct, similar results as Kawakita and Kanamori (2013) should apply in our case and the asymptotic variance of ATReL could be reduced compared with the estimator obtained under the $N \asymp n$ or $N < n$ scenarios. The study of this problem warrants future work.

Intrinsic efficient estimator. When the importance weight model is correctly specified while the imputation model may be wrong, the asymptotic variance of our estimator is dependent on the parameters $\bar{\gamma}$ and $\bar{r}(\cdot)$. For purely fixed dimensional parametric nuisance models, there exist certain moment equations for the imputation parameters that grant one to get the most efficient doubly robust estimator among those with the same specification of the imputation model. This property is known as intrinsic efficiency (Tan, 2010; Rotnitzky et al., 2012). Under our semi-nonparametric framework, flexibility in specifying the parametric parts of the nuisance models makes the intrinsic efficiency of our proposed estimator worthwhile considering. In Appendix C.3, we introduce a modified construction procedure for $\widehat{m}^{[\text{-}k]}(\cdot)$ that calibrates its nonparametric part and ensures the intrinsic efficiency of the estimator of $\boldsymbol{c}^{\intercal}\boldsymbol{\beta}_0$, or more generally, any given smooth function of $\boldsymbol{\beta}_0$.

References

David Azriel, Lawrence D Brown, Michael Sklar, Richard Berk, Andreas Buja, and Linda Zhao. Semi-supervised linear regression. *arXiv preprint arXiv:1612.02391*, 2016.

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Jay H Beder. A sieve estimator for the mean of a gaussian process. *The Annals of Statistics*, pages 59–78, 1987.

Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345–366, 2015.

Tianxi Cai, Mengyan Li, and Molei Liu. Semi-supervised triply robust inductive transfer learning. *arXiv preprint arXiv:2209.04977*, 2022.

Sebastian Calonico, Matias D Cattaneo, and Max H Farrell. On the effect of bias estimation on coverage accuracy in nonparametric inference. *Journal of the American Statistical Association*, 113(522):767–779, 2018.

Weihua Cao, Anastasios A Tsiatis, and Marie Davidian. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96(3):723–734, 2009.

Raymond J Carroll, David Ruppert, and Alan H Welsh. Local estimating equations. *Journal of the American Statistical Association*, 93(441):214–227, 1998.

Robert J Carroll, Will K Thompson, Anne E Eyler, Arthur M Mandelin, Tianxi Cai, Raquel M Zink, Jennifer A Pacheco, Chad S Boomershine, Thomas A Lasko, Hua Xu, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *Journal of the American Medical Informatics Association*, 19(e1):e162–e169, 2012.

Matias D Cattaneo. Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics*, 155(2):138–154, 2010.

Matias D Cattaneo, Max H Farrell, and Yingjie Feng. Large sample properties of partitioning-based series estimators. *The Annals of Statistics*, 48(3):1718–1741, 2020.

Abhishek Chakrabortty. *Robust Semi-Parametric Inference in Semi-Supervised Settings.* PhD thesis, 2016.

Abhishek Chakrabortty and Tianxi Cai. Efficient and adaptive linear regression in semi-supervised settings. *The Annals of Statistics*, 46(4):1541–1572, 2018.

Xiangli Chen, Mathew Monfort, Anqi Liu, and Brian D Ziebart. Robust covariate shift regression. In *Artificial Intelligence and Statistics*, pages 1270–1279, 2016.

Xiaohong Chen. Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, 6:5549–5632, 2007.

Xiaohong Chen, Han Hong, and Alessandro Tarozzi. Semiparametric efficiency in GMM models with auxiliary data. *The Annals of Statistics*, 36(2):808–843, 2008.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018a.

Victor Chernozhukov, Whitney K Newey, and James Robins. Double/debiased machine learning using regularized riesz representers. Technical report, cemmap working paper, 2018b.

Oliver Dukes and Stijn Vansteelandt. Inference on treatment effect parameters in potentially misspecified high-dimensional models. *Biometrika*, 2020.

Oliver Dukes, Stijn Vansteelandt, and David Whitney. On doubly robust inference for double machine learning. *arXiv preprint arXiv:2107.06124*, 2021.

Shinto Eguchi and John Copas. A class of logistic-type discriminant functions. *Biometrika*, 89(1):1–22, 2002.

Jianqing Fan, Nancy E Heckman, and Matt P Wand. Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association*, 90(429):141–150, 1995.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep learning for individual heterogeneity: An automatic inference framework. *arXiv preprint arXiv:2010.14694*, 2020.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

Satyajit Ghosh and Zhiqiang Tan. Doubly robust semiparametric inference using regularized calibrated estimation with high-dimensional data. *arXiv preprint arXiv:2009.12033*, 2020.

Bryan S Graham, Cristine Campos de Xavier Pinto, and Daniel Egel. Efficient estimation of data combination models by the method of auxiliary-to-study tilting (ast). *Journal of Business & Economic Statistics*, 34(2):288–301, 2016.

Jessica Gronsbell, Molei Liu, Lu Tian, and Tianxi Cai. Efficient estimation and evaluation of prediction rules in semi-supervised settings under stratified sampling. *arXiv preprint arXiv:2010.09443*, 2020.

Jessica L Gronsbell and Tianxi Cai. Semi-supervised approaches to efficient evaluation of model prediction performance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):579–594, 2018.

Jinyong Hahn. Functional restriction and efficiency in causal inference. *The Review of Economics and Statistics*, 86(1):73–76, 2004.

Peisong Han. Intrinsic efficiency and multiple robustness in longitudinal studies with drop-out. *Biometrika*, 103(3):683–700, 2016.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Sicong Huang, Jie Huang, Tianrun Cai, Kumar P Dahal, Andrew Cagan, Zeling He, Jacklyn Stratton, Isaac Gorelik, Chuan Hong, Tianxi Cai, et al. Impact of icd10 and secular changes on electronic medical record rheumatoid arthritis algorithms. *Rheumatology*, 59 (12):3759–3766, 2020.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press, 2015.

Joseph DY Kang and Joseph L Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539, 2007.

Masanori Kawakita and Takafumi Kanamori. Semi-supervised learning with density-ratio estimation. *Machine learning*, 91(2):189–209, 2013.

Edward H Kennedy. Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 2020.

Seung-Soo Kim, Adam D Hudgins, Brenda Gonzalez, Sofiya Milman, Nir Barzilai, Jan Vijg, Zhidong Tu, and Yousin Suh. A compendium of age-related phewas and gwas traits for human genetic association studies, their networks and genetic correlations. *Frontiers in Genetics*, 12:842, 2021.

Isaac S Kohane. Using electronic health records to drive discovery in disease genomics. *Nature Reviews Genetics*, 12(6):417–428, 2011.

Isaac S Kohane, Susanne E Churchill, and Shawn N Murphy. A translational engine at the national scale: informatics for integrating biology and the bedside. *Journal of the American Medical Informatics Association*, 19(2):181–185, 2012.

Damian Kozbur. Inference in additively separable models with a high-dimensional set of conditioning variables. *Journal of Business & Economic Statistics*, 39(4):984–1000, 2021.

Katherine P Liao, Tianxi Cai, Vivian Gainer, Sergey Goryachev, Qing Zeng-treitler, Soumya Raychaudhuri, Peter Szolovits, Susanne Churchill, Shawn Murphy, Isaac Kohane, et al. Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis care & research*, 62(8):1120–1127, 2010.

Katherine P Liao, Jiehuan Sun, Tianrun A Cai, Nicholas Link, Chuan Hong, Jie Huang, Jennifer E Huffman, Jessica Gronsbell, Yichi Zhang, and Yuk-Lam Ho. High-throughput multimodal automated phenotyping (map) with application to phewas. *Journal of the American Medical Informatics Association*, 26(11):1255–1262, 2019.

Xihong Lin and Raymond J Carroll. Semiparametric estimation in general repeated measures problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):69–88, 2006.

Anqi Liu and Brian D Ziebart. Robust covariate shift prediction with general losses and feature views. *arXiv preprint arXiv:1712.10043*, 2017.

Molei Liu, Yi Zhang, and Doudou Zhou. Double/debiased machine learning for logistic partially linear model. *The Econometrics Journal*, 24(3):559–588, 2021.

Whitney K Newey. Convergence rates and asymptotic normality for series estimators. *Journal of econometrics*, 79(1):147–168, 1997.

Whitney K Newey and James R Robins. Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*, 2018.

Yang Ning, Peng Sida, and Kosuke Imai. Robust estimation of causal effects via a high-dimensional covariate balancing propensity score. *Biometrika*, 107(3):533–554, 2020.

David Pollard. Empirical processes: theory and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages i–86. JSTOR, 1990.

Jing Qin, Jun Shao, and Biao Zhang. Efficient and doubly robust imputation for covariate-dependent missing responses. *Journal of the American Statistical Association*, 103(482): 797–810, 2008.

Zhongjun Qu, Jungmo Yoon, and Pierre Perron. Inference on conditional quantile processes in partially linear models with applications to the impact of unemployment benefits. *Review of Economics and Statistics*, pages 1–45, 2022.

Laila Rasmy, Yonghui Wu, Ningtao Wang, Xin Geng, W Jim Zheng, Fei Wang, Hulin Wu, Hua Xu, and Degui Zhi. A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous ehr data set. *Journal of biomedical informatics*, 84:11–16, 2018.

Sashank Jakkam Reddi, Barnabas Poczos, and Alex Smola. Doubly robust covariate shift correction. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Christoph Rothe and Sergio Firpo. Semiparametric two-step estimation using doubly robust moment conditions, 2015.

Andrea Rotnitzky, Quanhong Lei, Mariela Sued, and James M Robins. Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99(2):439–456, 2012.

Vira Semenova and Victor Chernozhukov. Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2):264–289, 2021.

Xiaotong Shen. On methods of sieves and penalization. *The Annals of Statistics*, pages 2555–2591, 1997.

Heng Shu and Zhiqiang Tan. Improved estimation of average treatment effects on the treated: Local efficiency, double robustness, and beyond. *arXiv preprint arXiv:1808.01408*, 2018.

Ezequiel Smucler, Andrea Rotnitzky, and James M Robins. A unifying approach for doubly-robust ℓ_1-regularized estimation of causal contrasts. *arXiv preprint arXiv:1904.03737*, 2019.

Zhiqiang Tan. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97(3):661–682, 2010.

Zhiqiang Tan. Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *Annals of Statistics*, 48(2):811–837, 2020.

Mark J van der Laan and Susan Gruber. Collaborative double robust targeted maximum likelihood estimation. *The international journal of biostatistics*, 6(1), 2010.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge University Press, 2000.

Tyler J VanderWeele. Surrogate measures and consistent surrogates. *Biometrics*, 69(3): 561–565, 2013.

Karel Vermeulen and Stijn Vansteelandt. Bias-reduced doubly robust estimation. *Journal of the American Statistical Association*, 110(511):1024–1036, 2015.

Junfeng Wen, Chun-Nam Yu, and Russell Greiner. Robust learning under uncertain test distributions: Relating covariate shift to model misspecification. In *ICML*, pages 631–639, 2014.

Chunhua Weng, Nigam H Shah, and George Hripcsak. Deep phenotyping: Embracing complexity and temporality—towards scalability, portability, and interoperability. *Journal of Biomedical Informatics*, 105:103433, 2020.

Sheng Yu, Abhishek Chakrabortty, Katherine P Liao, Tianrun Cai, Ashwin N Ananthakrishnan, Vivian S Gainer, Susanne E Churchill, Peter Szolovits, Shawn N Murphy, Isaac S Kohane, et al. Surrogate-assisted feature extraction for high-throughput phenotyping. *Journal of the American Medical Informatics Association*, 24(e1):e143–e149, 2017.

Qingyuan Zhao and Daniel Percival. Entropy balancing is doubly robust. *Journal of Causal Inference*, 5(1), 2017.

Doudou Zhou, Molei Liu, Mengyan Li, and Tianxi Cai. Doubly robust augmented model accuracy transfer inference with high dimensional features. *arXiv preprint arXiv:2208.05134*, 2022.

Appendix

Appendix A. Proof of Theorem 1

Proof Let $\|\cdot\|_\infty$ represent the maximum norm of a vector or matrix. Without loss of generality, assume $\|\boldsymbol{c}\|_2 = 1$. First, we derive the error rate for the whole $\widehat{\boldsymbol{\beta}}_{\textsf{ATReL}}$ vector, which is above the parametric rate but useful in analyzing the second-order error terms. Inspired by Chen et al. (2016), we expand the left side of (13) as

$$\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\widehat{\omega}^{[\text{-}k]}(\boldsymbol{X}_i)\boldsymbol{A}_i\left\{Y_i-\widehat{m}^{[\text{-}k]}(\boldsymbol{X}_i)\right\}+\frac{1}{N}\sum_{i=n+1}^{N+n}\boldsymbol{A}_i\{\widehat{m}(\boldsymbol{X}_i)-g(\boldsymbol{A}_i^{\intercal}\boldsymbol{\beta})\}\\
=&\frac{1}{n}\sum_{i=1}^{n}\bar{\omega}(\boldsymbol{X}_i)\boldsymbol{A}_i\left\{Y_i-\bar{m}(\boldsymbol{X}_i)\right\}+\frac{1}{N}\sum_{i=n+1}^{N+n}\boldsymbol{A}_i\{\bar{m}(\boldsymbol{X}_i)-g(\boldsymbol{A}_i^{\intercal}\boldsymbol{\beta})\}\\
&+\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\{\widehat{\omega}^{[\text{-}k]}(\boldsymbol{X}_i)-\bar{\omega}(\boldsymbol{X}_i)\}\boldsymbol{A}_i\{\widehat{m}^{[\text{-}k]}(\boldsymbol{X}_i)-\bar{m}(\boldsymbol{X}_i)\}\\
&+\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\bar{\omega}(\boldsymbol{X}_i)\boldsymbol{A}_i\{\widehat{m}^{[\text{-}k]}(\boldsymbol{X}_i)-\bar{m}(\boldsymbol{X}_i)\}-\frac{1}{N}\sum_{i=n+1}^{N+n}\boldsymbol{A}_i\{\widehat{m}(\boldsymbol{X}_i)-\bar{m}(\boldsymbol{X}_i)\}\\
&+\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\{\widehat{\omega}^{[\text{-}k]}(\boldsymbol{X}_i)-\bar{\omega}(\boldsymbol{X}_i)\}\boldsymbol{A}_i\left\{Y_i-\bar{m}(\boldsymbol{X}_i)\right\}\\
=:&\boldsymbol{V}(\boldsymbol{\beta})+\boldsymbol{\Delta}_a+\boldsymbol{\Delta}_b+\boldsymbol{\Delta}_c.
\end{aligned}\tag{A1}$$

By Assumption 3, independence between $\widehat{\omega}^{[\text{-}k]}(\cdot)$ and data from $\mathcal{I}_k$ or data from the target population, and using the central limit theorem (CLT), we have that: for each k,

$$\begin{aligned}
&\frac{K}{n}\sum_{i\in\mathcal{I}_k}\{\widehat{\omega}^{[\text{-}k]}(\boldsymbol{X}_i)-\bar{\omega}(\boldsymbol{X}_i)\}^2-\mathbb{E}_1\{\widehat{\omega}^{[\text{-}k]}(\boldsymbol{X})-\bar{\omega}(\boldsymbol{X})\}^2=o_p(n^{-1/2});\\
&\frac{K}{n}\sum_{i\in\mathcal{I}_k}\{\widehat{m}^{[\text{-}k]}(\boldsymbol{X}_i)-\bar{m}(\boldsymbol{X}_i)\}^2-\mathbb{E}_1\{\widehat{m}^{[\text{-}k]}(\boldsymbol{X})-\bar{m}(\boldsymbol{X})\}^2=o_p(n^{-1/2});\\
&\frac{1}{N}\sum_{i=n+1}^{N+n}\{\widehat{m}(\boldsymbol{X}_i)-\bar{m}(\boldsymbol{X}_i)\}^2-\mathbb{E}_0\{\widehat{m}(\boldsymbol{X})-\bar{m}(\boldsymbol{X})\}^2=o_p(n^{-1/2})
\end{aligned}$$

Also, by Assumption 3 and Assumption 1, we have that: for each k,

$$\begin{aligned}
&\mathbb{E}_1\{\widehat{\omega}^{[\text{-}k]}(\boldsymbol{X})-\bar{\omega}(\boldsymbol{X})\}^2=\mathbb{E}_1\left[\bar{\omega}(\boldsymbol{X})\left\{\frac{\widehat{\omega}^{[\text{-}k]}(\boldsymbol{X})}{\bar{\omega}(\boldsymbol{X})}-1\right\}^2\right]\\
=&\mathbb{E}_1\left[\bar{\omega}^2(\boldsymbol{X})\left(\|\boldsymbol{\Psi}\|_2^2\|\widehat{\boldsymbol{\alpha}}^{[\text{-}k]}-\bar{\boldsymbol{\alpha}}\|_2^2+\left\{\widehat{h}^{[\text{-}k]}(\boldsymbol{Z})-\bar{h}(\boldsymbol{Z})\right\}^2+\|\boldsymbol{\Psi}\|_2^4\|\widehat{\boldsymbol{\alpha}}^{[\text{-}k]}-\bar{\boldsymbol{\alpha}}\|_2^4+\left\{\widehat{h}^{[\text{-}k]}(\boldsymbol{Z})-\bar{h}(\boldsymbol{Z})\right\}^4\right)\right]\\
\leq&\mathbb{E}_1\left[\{\bar{\omega}^4(\boldsymbol{X})+\|\boldsymbol{\Psi}\|_2^4+\|\boldsymbol{\Psi}\|_2^8+O_p(n^{-1})\}\right]\|\widehat{\boldsymbol{\alpha}}^{[\text{-}k]}-\bar{\boldsymbol{\alpha}}\|_2^2+\{1+o_p(1)\}\mathbb{E}_1\left[\bar{\omega}^2(\boldsymbol{X})\{\widehat{h}^{[\text{-}k]}(\boldsymbol{Z})-\bar{h}(\boldsymbol{Z})\}^2\right]\\
=&O_p\left(\mathbb{E}_1\left[\bar{\omega}^2(\boldsymbol{X})\{\widehat{h}^{[\text{-}k]}(\boldsymbol{Z})-\bar{h}(\boldsymbol{Z})\}^2\right]+n^{-1}\right)=o_p(n^{-1/2}),
\end{aligned}$$

and that each $j \in \{0,1\}$,

$$\begin{aligned}
&\mathbb{E}_j\{\widehat{m}^{[\text{-}k]}(\boldsymbol{X}) - \bar{m}(\boldsymbol{X})\}^2 \\
=&\mathbb{E}_1\Big[\breve{g}^2\{\bar{m}(\boldsymbol{X})\}\left(\|\boldsymbol{\Phi}\|_2^2\|\widehat{\boldsymbol{\gamma}}^{[\text{-}k]} - \bar{\boldsymbol{\gamma}}\|_2^2 + \left\{\widehat{r}^{[\text{-}k]}(\boldsymbol{Z}) - \bar{r}(\boldsymbol{Z})\right\}^2\right) \\
&+ C_L^2\left(\|\boldsymbol{\Phi}\|_2^4\|\widehat{\boldsymbol{\gamma}}^{[\text{-}k]} - \bar{\boldsymbol{\gamma}}\|_2^4 + \left\{\widehat{r}^{[\text{-}k]}(\boldsymbol{Z}) - \bar{r}(\boldsymbol{Z})\right\}^4\right)\Big] \\
=&O_p\left(\mathbb{E}_1\left[\breve{g}_m^2\{\bar{m}(\boldsymbol{X})\}\{\widehat{r}^{[\text{-}k]}(\boldsymbol{Z}) - \bar{r}(\boldsymbol{Z})\}^2\right] + n^{-1}\right) = o_p(n^{-1/2}).
\end{aligned}$$

Thus, we have $\frac{K}{n}\sum_{i\in\mathcal{I}_k}\{\widehat{\omega}^{[\text{-}k]}(\boldsymbol{X}_i) - \bar{\omega}(\boldsymbol{X}_i)\}^2 = o_p(n^{-1/2})$, $\frac{K}{n}\sum_{i\in\mathcal{I}_k}\{\widehat{m}^{[\text{-}k]}(\boldsymbol{X}_i) - \bar{m}(\boldsymbol{X}_i)\}^2 = o_p(n^{-1/2})$ and $\frac{1}{N}\sum_{i=n+1}^{N+n}\{\widehat{m}(\boldsymbol{X}_i) - \bar{m}(\boldsymbol{X}_i)\}^2 = o_p(n^{-1/2})$. Combining these with Assumption 1, we have that

$$\begin{aligned}
\|\boldsymbol{\Delta}_a\|_\infty \leq& n^{-1}\max_i\|\boldsymbol{A}_i\|_\infty\sum_{k=1}^K\sum_{i\in\mathcal{I}_k}\{\widehat{\omega}^{[\text{-}k]}(\boldsymbol{X}_i) - \bar{\omega}(\boldsymbol{X}_i)\}^2 + \{\widehat{m}^{[\text{-}k]}(\boldsymbol{X}_i) - \bar{m}(\boldsymbol{X}_i)\}^2 = o_p(n^{-1/2}); \\
\|\boldsymbol{\Delta}_b\|_\infty \leq& \max_i\|\boldsymbol{A}_i\|_\infty\left[n^{-1}\sum_{k=1}^K\sum_{i\in\mathcal{I}_k}\bar{\omega}^2(\boldsymbol{X}_i)\right]^{\frac{1}{2}}\left[n^{-1}\sum_{k=1}^K\sum_{i\in\mathcal{I}_k}\{\widehat{m}(\boldsymbol{X}_i) - \bar{m}(\boldsymbol{X}_i)\}^2\right]^{\frac{1}{2}} \\
&+ \max_i\|\boldsymbol{A}_i\|_\infty\left[N^{-1}\sum_{i=n+1}^{N+n}\{\widehat{m}(\boldsymbol{X}_i) - \bar{m}(\boldsymbol{X}_i)\}^2\right]^{\frac{1}{2}} = o_p(n^{-1/4}); \\
\|\boldsymbol{\Delta}_c\|_\infty \leq& \max_i\|\boldsymbol{A}_i\|_\infty\left[n^{-1}\sum_{k=1}^K\sum_{i\in\mathcal{I}_k}Y_i^2 + \bar{m}^2(\boldsymbol{X}_i)\right]^{\frac{1}{2}}\left[n^{-1}\sum_{k=1}^K\sum_{i\in\mathcal{I}_k}\{\widehat{\omega}(\boldsymbol{X}_i) - \bar{\omega}(\boldsymbol{X}_i)\}^2\right]^{\frac{1}{2}} = o_p(n^{-1/4}).
\end{aligned}$$

Thus, $\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}}$ solves: $\boldsymbol{V}(\boldsymbol{\beta}) + o_p(n^{-1/4}) = \boldsymbol{0}$. Let the solution of $\mathbb{E}\boldsymbol{V}(\boldsymbol{\beta}) = \boldsymbol{0}$ be $\bar{\boldsymbol{\beta}}$. When $\bar{\omega}(\cdot) = \mathbb{w}(\cdot)$,

$$\begin{aligned}
\mathbb{E}\boldsymbol{V}(\boldsymbol{\beta}) =& \mathbb{E}_1\mathbb{w}(\boldsymbol{X})\boldsymbol{X}\{Y - g(\boldsymbol{A}^\intercal\boldsymbol{\beta})\} + [\mathbb{E}_1\mathbb{w}(\boldsymbol{X})\{g(\boldsymbol{A}^\intercal\boldsymbol{\beta}) - \bar{m}(\boldsymbol{X})\} - \mathbb{E}_0\{g(\boldsymbol{A}^\intercal\boldsymbol{\beta}) - \bar{m}(\boldsymbol{X})\}] \\
=& \mathbb{E}_0\boldsymbol{X}\{Y - g(\boldsymbol{A}^\intercal\boldsymbol{\beta})\} + \boldsymbol{0}.
\end{aligned}$$

As $\bar{m}(\cdot) = \mu(\cdot)$, $\mathbb{E}\boldsymbol{V}(\boldsymbol{\beta}) = \boldsymbol{0} + \mathbb{E}_0\{\bar{\mu}(\boldsymbol{X}) - g(\boldsymbol{A}^\intercal\boldsymbol{\beta})\}$. Both cases lead to that $\boldsymbol{\beta}_0$ solves $\mathbb{E}\boldsymbol{V}(\boldsymbol{\beta}) = \boldsymbol{0}$. So under Assumption 2, we have $\bar{\boldsymbol{\beta}} = \boldsymbol{\beta}_0$. By Assumption 1, $\boldsymbol{V}(\boldsymbol{\beta})$ is continuous differential on $\boldsymbol{\beta}$. Then using Theorem 8.2 of Pollard (1990), we have $\|\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}} - \boldsymbol{\beta}_0\|_2 = o_p(n^{-1/4}) = o_p(1)$.

Then we consider the asymptotic expansion of $\boldsymbol{c}^\intercal\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}}$. Noting that $\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}}$ is consistent for $\boldsymbol{\beta}_0$, by Theorem 5.21 of Van der Vaart (2000), we expand (A1) with respect to $\boldsymbol{c}^\intercal\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}}$

as:

$$
\begin{aligned}
&\sqrt{n}(\boldsymbol{c}^{\intercal}\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}} - \boldsymbol{c}^{\intercal}\boldsymbol{\beta}_0)\\
=&n^{-\frac{1}{2}}\sum_{i=1}^{n}\bar{\omega}(\boldsymbol{X}_i)\boldsymbol{c}^{\intercal}\widehat{\boldsymbol{J}}_{\breve{\boldsymbol{\beta}}}^{-1}\boldsymbol{A}_i\left\{Y_i-\bar{m}(\boldsymbol{X}_i)\right\}+\frac{\sqrt{\rho}}{\sqrt{N}}\sum_{i=n+1}^{N+n}\boldsymbol{c}^{\intercal}\widehat{\boldsymbol{J}}_{\breve{\boldsymbol{\beta}}}^{-1}\boldsymbol{A}_i\{\bar{m}(\boldsymbol{X}_i)-g(\boldsymbol{A}_i^{\intercal}\boldsymbol{\beta}_0)\}\\
&+n^{-\frac{1}{2}}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\{\widehat{\omega}^{[\text{-}k]}(\boldsymbol{X}_i)-\bar{\omega}(\boldsymbol{X}_i)\}\boldsymbol{c}^{\intercal}\widehat{\boldsymbol{J}}_{\breve{\boldsymbol{\beta}}}^{-1}\boldsymbol{A}_i\left\{Y_i-\bar{m}(\boldsymbol{X}_i)\right\}\\
&+n^{-\frac{1}{2}}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\bar{\omega}(\boldsymbol{X}_i)\boldsymbol{c}^{\intercal}\widehat{\boldsymbol{J}}_{\breve{\boldsymbol{\beta}}}^{-1}\boldsymbol{A}_i\{\widehat{m}^{[\text{-}k]}(\boldsymbol{X}_i)-\bar{m}(\boldsymbol{X}_i)\}-\frac{n^{\frac{1}{2}}}{N}\sum_{i=n+1}^{N+n}\boldsymbol{c}^{\intercal}\widehat{\boldsymbol{J}}_{\breve{\boldsymbol{\beta}}}^{-1}\boldsymbol{A}_i\{\widehat{m}(\boldsymbol{X}_i)-\bar{m}(\boldsymbol{X}_i)\}\\
&+n^{-\frac{1}{2}}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\boldsymbol{c}^{\intercal}\widehat{\boldsymbol{J}}_{\breve{\boldsymbol{\beta}}}^{-1}\boldsymbol{A}_i\{\widehat{\omega}^{[\text{-}k]}(\boldsymbol{X}_i)-\bar{\omega}(\boldsymbol{X}_i)\}\{\widehat{m}^{[\text{-}k]}(\boldsymbol{X}_i)-\bar{m}(\boldsymbol{X}_i)\}\\
=:&V+\Xi_1+\Xi_2+\Delta_3,
\end{aligned}
\tag{A2}
$$

where $\breve{\boldsymbol{\beta}}$ is some vector lying between $\boldsymbol{\beta}_0$ and $\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}}$. First, we shall show that $\|\widehat{\boldsymbol{J}}_{\breve{\boldsymbol{\beta}}}^{-1}-\boldsymbol{J}_{\boldsymbol{\beta}_0}^{-1}\|_\infty=O_p(n^{-1/4})$. Since the dimensionality of $\boldsymbol{A}$, d is fixed, we have

$$
\left\|\widehat{\boldsymbol{J}}_{\breve{\boldsymbol{\beta}}}^{-1}-\boldsymbol{J}_{\boldsymbol{\beta}_0}^{-1}\right\|_\infty=\left\|\widehat{\boldsymbol{J}}_{\breve{\boldsymbol{\beta}}}^{-1}\boldsymbol{J}_{\boldsymbol{\beta}_0}^{-1}(\widehat{\boldsymbol{J}}_{\breve{\boldsymbol{\beta}}}-\boldsymbol{J}_{\boldsymbol{\beta}_0})\right\|_\infty\le d^3\left\|\widehat{\boldsymbol{J}}_{\breve{\boldsymbol{\beta}}}^{-1}\right\|_\infty\left\|\boldsymbol{J}_{\boldsymbol{\beta}_0}^{-1}\right\|_\infty\left\|\widehat{\boldsymbol{J}}_{\breve{\boldsymbol{\beta}}}-\boldsymbol{J}_{\boldsymbol{\beta}_0}\right\|_\infty.
$$

Denote by $\boldsymbol{A}_i=(A_{1i},\ldots,A_{di})^{\intercal}$. By Assumption 1 and CLT, there exists a constant $C>0$ such that for $j,\ell\in\{1,\ldots,d\}$,

$$
\begin{aligned}
&\left|N^{-1}\sum_{i=n+1}^{n+N}A_{ji}A_{\ell i}\dot{g}(\boldsymbol{A}_i^{\intercal}\breve{\boldsymbol{\beta}})-\mathbb{E}_0A_{ji}A_{\ell i}\dot{g}(\boldsymbol{A}_i^{\intercal}\boldsymbol{\beta}_0)\right|\\
\le&\left|N^{-1}\sum_{i=n+1}^{n+N}A_{ji}A_{\ell i}\{\dot{g}(\boldsymbol{A}_i^{\intercal}\breve{\boldsymbol{\beta}})-\dot{g}(\boldsymbol{A}_i^{\intercal}\boldsymbol{\beta}_0)\}\right|+\left|N^{-1}\sum_{i=n+1}^{n+N}A_{ji}A_{\ell i}\dot{g}(\boldsymbol{A}_i^{\intercal}\boldsymbol{\beta}_0)-\mathbb{E}_0A_{ji}A_{\ell i}\dot{g}(\boldsymbol{A}_i^{\intercal}\boldsymbol{\beta}_0)\right|\\
\le&\left|N^{-1}\sum_{i=n+1}^{n+N}|A_{ji}A_{\ell i}|C_L|\boldsymbol{A}_i^{\intercal}\breve{\boldsymbol{\beta}}-\boldsymbol{A}_i^{\intercal}\boldsymbol{\beta}_0|\right|+O_p(n^{-1/2})\le C\|\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}}-\boldsymbol{\beta}_0\|_2+O_p(n^{-1/2})=o_p(n^{-1/4}).
\end{aligned}
$$

Also noting that $\|\boldsymbol{J}_{\boldsymbol{\beta}_0}^{-1}\|_\infty$ is bounded by Assumption 1, we have

$$
\left\|\widehat{\boldsymbol{J}}_{\breve{\boldsymbol{\beta}}}^{-1}-\boldsymbol{J}_{\boldsymbol{\beta}_0}^{-1}\right\|_\infty=o_p(n^{-1/4}).
\tag{A3}
$$

Under Assumption 2, and similar to the deduction above, the expectation of

$$
n^{-\frac{1}{2}}\sum_{i=1}^{n}\bar{\omega}(\boldsymbol{X}_i)\boldsymbol{A}_i\left\{Y_i-\bar{m}(\boldsymbol{X}_i)\right\}+\frac{\sqrt{\rho}}{\sqrt{N}}\sum_{i=n+1}^{N+n}\boldsymbol{A}_i\{\bar{m}(\boldsymbol{X}_i)-g(\boldsymbol{A}_i^{\intercal}\boldsymbol{\beta}_0)\}
$$

is $\mathbf{0}$. So by Assumption 1, equation (A3), CLT, and Slutsky's Theorem, we have that V weakly converges to $N(0, \sigma^2)$ where σ^2 represents the asymptotic variance of V and is order 1. We then consider the remaining terms separately. First, we have

$$\begin{aligned}\Xi_1 =& n^{-\frac{1}{2}} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \bar{\omega}(\boldsymbol{X}_i) \boldsymbol{c}^{\intercal} \widehat{\boldsymbol{J}}_{\breve{\boldsymbol{\beta}}}^{-1} \boldsymbol{A}_i \left[Y_i - g_m\{\boldsymbol{\Phi}^{\intercal}\bar{\boldsymbol{\gamma}} + \bar{r}(\boldsymbol{Z})\}\right] \left[\boldsymbol{\Psi}_i^{\intercal}(\widehat{\boldsymbol{\alpha}}^{[\text{-}k]} - \bar{\boldsymbol{\alpha}}) + O_p(\{\boldsymbol{\Psi}_i^{\intercal}(\widehat{\boldsymbol{\alpha}}^{[\text{-}k]} - \bar{\boldsymbol{\alpha}})\}^2)\right] \\ &+ n^{-\frac{1}{2}} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \bar{\omega}(\boldsymbol{X}_i) \boldsymbol{\kappa}_{i,\boldsymbol{\beta}_0} \left[Y_i - g_m\{\boldsymbol{\Phi}^{\intercal}\bar{\boldsymbol{\gamma}} + \bar{r}(\boldsymbol{Z})\}\right] \Delta h^{[\text{-}k]}(\boldsymbol{z}_j) \\ &+ n^{-\frac{1}{2}} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \bar{\omega}(\boldsymbol{X}_i) \boldsymbol{c}^{\intercal} (\widehat{\boldsymbol{J}}_{\breve{\boldsymbol{\beta}}}^{-1} - \boldsymbol{J}_{\boldsymbol{\beta}_0}^{-1}) \boldsymbol{A}_i \left[Y_i - g_m\{\boldsymbol{\Phi}^{\intercal}\bar{\boldsymbol{\gamma}} + \bar{r}(\boldsymbol{Z})\}\right] \Delta h^{[\text{-}k]}(\boldsymbol{z}_j) \\ =:& U_1 + \Delta_{11} + \Delta_{12},\end{aligned} \tag{A4}$$

where $\Delta h^{[\text{-}k]}(\boldsymbol{z}_j) = \widehat{h}^{[\text{-}k]}(\boldsymbol{Z}_i) - \bar{h}(\boldsymbol{Z}_i) + O_p(\{\widehat{h}^{[\text{-}k]}(\boldsymbol{Z}_i) - \bar{h}(\boldsymbol{Z}_i)\}^2)$. Recall that

$$\boldsymbol{\zeta}_{\alpha} = \mathbb{E}_1 \bar{\omega}(\boldsymbol{X}) \boldsymbol{\kappa}_{\boldsymbol{\beta}_0} \left[Y - g_m\{\boldsymbol{\Phi}^{\intercal}\bar{\boldsymbol{\gamma}} + \bar{r}(\boldsymbol{Z})\}\right] \boldsymbol{\Psi}.$$

Again using (A3) and Assumption 1, we have that

$$n^{-1} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \bar{\omega}(\boldsymbol{X}_i) \boldsymbol{c}^{\intercal} \widehat{\boldsymbol{J}}_{\breve{\boldsymbol{\beta}}}^{-1} \boldsymbol{A}_i \left[Y_i - g_m\{\boldsymbol{\Phi}^{\intercal}\bar{\boldsymbol{\gamma}} + \bar{r}(\boldsymbol{Z})\}\right] \xrightarrow{p} \boldsymbol{\zeta}_{\alpha}.$$

Combining this with Assumption 1, Assumption 3 that $\sqrt{n}(\widehat{\boldsymbol{\alpha}}^{[\text{-}k]} - \bar{\boldsymbol{\alpha}})$ is asymptotic normal with mean 0 and covariance of order 1, and using Slutsky's Theorem, we have that U_1 is asymptotically equivalent with $\sqrt{n}\boldsymbol{\zeta}_{\alpha}^{\intercal}(\widehat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}})$, which weakly converges to normal distribution with mean 0 and variance of order 1.

For Δ_{11}, by Assumption 2, the moment condition:

$$\mathbb{E}_1 \left[\bar{\omega}(\boldsymbol{X}) \boldsymbol{\kappa}_{\boldsymbol{\beta}_0} \left(Y - g_m\{\boldsymbol{\Phi}^{\intercal}\bar{\boldsymbol{\gamma}} + \bar{r}(\boldsymbol{Z})\}\right) \middle| \boldsymbol{Z}\right] = 0$$

holds because under Assumption 2(i), both limiting parameters $\omega^*(\cdot) = \bar{\omega}(\cdot) = \omega(\cdot)$ and $\bar{r}(\cdot)$ solves (7) while under 2(ii), $\mathbb{E}_1[Y|\boldsymbol{X}] = g_m\{\boldsymbol{\Phi}^{\intercal}\bar{\boldsymbol{\gamma}} + \bar{r}(\boldsymbol{Z})\}$, leading to

$$\mathbb{E}_1 \left[\bar{\omega}(\boldsymbol{X}) \boldsymbol{\kappa}_{\boldsymbol{\beta}_0} \left(Y - g_m\{\boldsymbol{\Phi}^{\intercal}\bar{\boldsymbol{\gamma}} + \bar{r}(\boldsymbol{Z})\}\right) \middle| \boldsymbol{X}\right] = 0.$$

Combining this with the fact that $\widehat{h}^{[\text{-}k]}(\cdot)$ is independent of the data in $\mathcal{I}_k$ due to the use of cross-fitting, we have $\mathbb{E}_1 \Delta_{11} = \mathbb{E}_1[\Delta_{11} \mid \widehat{h}^{[\text{-}k]}(\cdot)] = 0 + n^{1/2} O_p(\{\widehat{h}^{[\text{-}k]}(\boldsymbol{Z}_i) - \bar{h}(\boldsymbol{Z}_i)\}^2)$. By Assumptions 1 and 3(ii), we have that

$$\begin{aligned}&\mathrm{Var}_1 \left(\bar{\omega}(\boldsymbol{X}_i) \boldsymbol{\kappa}_{i,\boldsymbol{\beta}_0} \left[Y_i - g_m\{\boldsymbol{\Phi}^{\intercal}\bar{\boldsymbol{\gamma}} + \bar{r}(\boldsymbol{Z})\}\right] \{\widehat{h}^{[\text{-}k]}(\boldsymbol{Z}_i) - \bar{h}(\boldsymbol{Z}_i)\} \middle| \widehat{h}^{[\text{-}k]}(\cdot)\right) \\ =& O(\mathbb{E}_1[\bar{\omega}^2(\boldsymbol{X}_i) + Y_i^2 + \bar{m}^2(\boldsymbol{X}_i)]) \cdot o_p(1) = o_p(1),\end{aligned}$$

where Var_1 and Var_0 represent the variance operator of the source and target population respectively. Then by CLT and Assumption 3(ii), we have that

$$\Delta_{11} = \Big(\Delta_{11} - \mathbb{E}_1[\Delta_{11}|\widehat{h}^{[\text{-}k]}(\cdot)]\Big) + \mathbb{E}_1[\Delta_{11}|\widehat{h}^{[\text{-}k]}(\cdot)] = o_p(1) + n^{1/2} O_p(\{\widehat{h}^{[\text{-}k]}(\boldsymbol{Z}_i) - \bar{h}(\boldsymbol{Z}_i)\}^2) = o_p(1).$$

For term Δ_{12}, by (A3) and Assumptions 1 and 3, there exists constant $C_{12} > 0$ such that

$$|\Delta_{12}| \leq_p C_{12} \max_i \|\boldsymbol{A}_i\|_\infty \left\|\widehat{\boldsymbol{J}}_{\breve{\boldsymbol{\beta}}}^{-1} - \boldsymbol{J}_{\boldsymbol{\beta}_0}^{-1}\right\|_\infty \left[n^{-1} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \bar{\varpi}^2(\boldsymbol{X}_i)\{\widehat{h}^{[\text{-}k]}(\boldsymbol{Z}_i) - \bar{h}(\boldsymbol{Z}_i)\}^2\right]^{\frac{1}{2}} = o_p(1).$$

Therefore, we come to that Ξ_1 is asymptotically equivalent with $\sqrt{n}\boldsymbol{\zeta}_\alpha^\intercal(\widehat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}})$. Similarly, we write the term Ξ_2 as

$$\begin{aligned}
\Xi_2 =& n^{-\frac{1}{2}} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \bar{\varpi}(\boldsymbol{X}_i)\boldsymbol{c}^\intercal \widehat{\boldsymbol{J}}_{\breve{\boldsymbol{\beta}}}^{-1} \boldsymbol{A}_i \breve{g}_m\{\bar{m}(\boldsymbol{X}_i)\} \left[\boldsymbol{\Phi}_i^\intercal(\widehat{\boldsymbol{\gamma}}^{[\text{-}k]} - \bar{\boldsymbol{\gamma}}) + O_p(\{\boldsymbol{\Phi}_i^\intercal(\widehat{\boldsymbol{\gamma}}^{[\text{-}k]} - \bar{\boldsymbol{\gamma}})\}^2)\right] \\
& - \frac{n^{\frac{1}{2}}}{N} \sum_{i=n+1}^{N+n} \boldsymbol{c}^\intercal \widehat{\boldsymbol{J}}_{\breve{\boldsymbol{\beta}}}^{-1} \boldsymbol{A}_i \breve{g}_m\{\bar{m}(\boldsymbol{X}_i)\} \left[K^{-1} \sum_{k=1}^{K} \boldsymbol{\Phi}_i^\intercal(\widehat{\boldsymbol{\gamma}}^{[\text{-}k]} - \bar{\boldsymbol{\gamma}}) + O_p(\{\boldsymbol{\Phi}_i^\intercal(\widehat{\boldsymbol{\gamma}}^{[\text{-}k]} - \bar{\boldsymbol{\gamma}})\}^2)\right] \\
& + n^{-\frac{1}{2}} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \bar{\varpi}(\boldsymbol{X}_i)\boldsymbol{\kappa}_{i,\boldsymbol{\beta}_0} \breve{g}_m\{\bar{m}(\boldsymbol{X}_i)\}\Delta r^{[\text{-}k]}(\boldsymbol{Z}_i) - \frac{n^{\frac{1}{2}}}{N} \sum_{i=n+1}^{N+n} \boldsymbol{\kappa}_{i,\boldsymbol{\beta}_0} \breve{g}_m\{\bar{m}(\boldsymbol{X}_i)\}\Delta r(\boldsymbol{Z}_i) \\
& + n^{-\frac{1}{2}} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \bar{\varpi}(\boldsymbol{X}_i)\boldsymbol{c}^\intercal \left[\widehat{\boldsymbol{J}}_{\breve{\boldsymbol{\beta}}}^{-1} - \boldsymbol{J}_{\boldsymbol{\beta}_0}^{-1}\right] \boldsymbol{A}_i \breve{g}_m\{\bar{m}(\boldsymbol{X}_i)\}\Delta r^{[\text{-}k]}(\boldsymbol{Z}_i) \\
& - \frac{n^{\frac{1}{2}}}{N} \sum_{i=n+1}^{N+n} \boldsymbol{c}^\intercal \left[\widehat{\boldsymbol{J}}_{\breve{\boldsymbol{\beta}}}^{-1} - \boldsymbol{J}_{\boldsymbol{\beta}_0}^{-1}\right] \boldsymbol{A}_i \breve{g}_m\{\bar{m}(\boldsymbol{X}_i)\}\Delta r(\boldsymbol{Z}_i) \\
=:& U_2 + \Delta_{21} + \Delta_{22},
\end{aligned} \tag{A5}$$

where $\Delta r^{[\text{-}k]}(\boldsymbol{Z}_i) = \widehat{r}^{[\text{-}k]}(\boldsymbol{Z}_i) - \bar{r}(\boldsymbol{Z}_i) + O_p(\{\widehat{r}^{[\text{-}k]}(\boldsymbol{Z}_i) - \bar{r}(\boldsymbol{Z}_i)\}^2)$, $\Delta r(\boldsymbol{Z}_i) = K^{-1}\sum_{k=1}^{K} \Delta r^{[\text{-}k]}(\boldsymbol{Z}_i)$, U_2 represents the difference of the first two terms, and Δ_{22} represents the difference of the last two terms. Similar to U_1, by (A3) and Assumption 1,

$$\frac{1}{n} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \bar{\varpi}(\boldsymbol{X}_i)\boldsymbol{c}^\intercal \widehat{\boldsymbol{J}}_{\breve{\boldsymbol{\beta}}}^{-1} \boldsymbol{A}_i \breve{g}_m\{\bar{m}(\boldsymbol{X}_i)\}\boldsymbol{\Phi}_i - \frac{1}{N} \sum_{i=n+1}^{N+n} \boldsymbol{c}^\intercal \widehat{\boldsymbol{J}}_{\breve{\boldsymbol{\beta}}}^{-1} \boldsymbol{A}_i \breve{g}_m\{\bar{m}(\boldsymbol{X}_i)\}\boldsymbol{\Phi}_i \xrightarrow{p} \boldsymbol{\zeta}_\gamma.$$

Again, combining this with Assumptions 1 and Assumption 3, and using Slutsky's Theorem, we have that U_2 is asymptotically equivalent with $\sqrt{n}\boldsymbol{\zeta}_\gamma^\intercal(\widehat{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}})$, which weakly converges to normal distribution with mean 0 and variance of order 1.

For Δ_{21}, by Assumptions 2 and 3, as well as the use of cross-fitting, we have that

$$\begin{aligned}
&\mathbb{E}_1\left(\frac{1}{n} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \bar{\varpi}(\boldsymbol{X}_i)\boldsymbol{\kappa}_{i,\boldsymbol{\beta}_0} \breve{g}_m\{\bar{m}(\boldsymbol{X}_i)\}\Delta r^{[\text{-}k]}(\boldsymbol{Z}_i)\right) \\
&\quad - \mathbb{E}_0\left(\frac{1}{N} \sum_{i=n+1}^{N+n} \boldsymbol{\kappa}_{i,\boldsymbol{\beta}_0} \breve{g}_m\{\bar{m}(\boldsymbol{X}_i)\}\Delta r^{[\text{-}k]}(\boldsymbol{Z}_i)\right) = o_p(n^{-1/2}).
\end{aligned}$$

Here, we follow the same idea as that for Δ_{11}: if Assumption 2(i) holds, we have $\bar{\omega}(\cdot) = \mathbb{w}(\cdot)$ and

$$\mathbb{E}_1\left[\exp\{\boldsymbol{\Psi}^{\intercal}\bar{\boldsymbol{\alpha}} + \bar{h}(\boldsymbol{Z})\}\boldsymbol{\kappa}_{\boldsymbol{\beta}_0}\breve{g}_m\{\bar{m}(\boldsymbol{X})\}f(\boldsymbol{X})\right] = \mathbb{E}_0\left[\boldsymbol{\kappa}_{\boldsymbol{\beta}_0}\breve{g}_m\{\bar{m}(\boldsymbol{X})\}f(\boldsymbol{X})\right]$$

holds for all measurable function of $\boldsymbol{X}$, $f(\cdot)$; when Assumption 2(ii) holds, we have that $m^*(\cdot) = \bar{m}(\cdot) = \mu(\cdot)$ and thus $\bar{h}(\cdot)$ solves (8). Also, note that

$$\begin{aligned}
&\mathrm{Var}_1\left(\bar{\omega}(\boldsymbol{X}_i)\boldsymbol{\kappa}_{i,\boldsymbol{\beta}_0}\breve{g}_m\{\bar{m}(\boldsymbol{X}_i)\}\{\widehat{r}^{[\text{-}k]}(\boldsymbol{Z}_i) - \bar{r}(\boldsymbol{Z}_i)\}\Big|\widehat{r}^{[\text{-}k]}(\cdot)\right)\\
=&O(\mathbb{E}_1[\bar{\omega}^2(\boldsymbol{X}_i) + \breve{g}_m^2\{\bar{m}(\boldsymbol{X}_i)\}]) \cdot o_p(1) = o_p(1);\\
&\mathrm{Var}_0\left(\boldsymbol{\kappa}_{i,\boldsymbol{\beta}_0}\breve{g}_m\{\bar{m}(\boldsymbol{X}_i)\}\{\widehat{r}^{[\text{-}k]}(\boldsymbol{Z}_i) - \bar{r}(\boldsymbol{Z}_i)\}\Big|\widehat{r}^{[\text{-}k]}(\cdot)\right) = O(\mathbb{E}_1\breve{g}_m^2\{\bar{m}(\boldsymbol{X}_i)\}) \cdot o_p(1) = o_p(1);
\end{aligned}$$

Then similar to Δ_{12}, we come to $\Delta_{22} = o_p(1)$. Thus, the term Ξ_2 is asymptotically equivalent with $\sqrt{n}\boldsymbol{\zeta}_{\gamma}^{\intercal}(\widehat{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}})$, which weakly converges to the normal distribution with mean 0 and variance of order 1.

Finally, we consider Δ_3 in (A2). By Assumption 1, the boundness of $|\boldsymbol{c}^{\intercal}\widehat{\boldsymbol{J}}_{\breve{\boldsymbol{\beta}}}^{-1}\boldsymbol{A}_i|$ and our derived bounds for $n^{-1}\sum_{k=1}^K\sum_{i\in\mathcal{I}_k}\{\widehat{\omega}^{[\text{-}k]}(\boldsymbol{X}_i) - \bar{\omega}(\boldsymbol{X}_i)\}^2$ and $n^{-1}\sum_{k=1}^K\sum_{i\in\mathcal{I}_k}\{\widehat{m}^{[\text{-}k]}(\boldsymbol{X}_i) - \bar{m}(\boldsymbol{X}_i)\}^2$,

$$\begin{aligned}
|\Delta_3| =&O\left(n^{-\frac{1}{2}}\sum_{k=1}^K\sum_{i\in\mathcal{I}_k}|\widehat{\omega}^{[\text{-}k]}(\boldsymbol{X}_i) - \bar{\omega}(\boldsymbol{X}_i)||\widehat{m}^{[\text{-}k]}(\boldsymbol{X}_i) - \bar{m}(\boldsymbol{X}_i)|\right)\\
\leq&\sqrt{n}O\left(\left[n^{-1}\sum_{k=1}^K\sum_{i\in\mathcal{I}_k}\{\widehat{\omega}^{[\text{-}k]}(\boldsymbol{X}_i) - \bar{\omega}(\boldsymbol{X}_i)\}^2\right]^{\frac{1}{2}}\left[n^{-1}\sum_{k=1}^K\sum_{i\in\mathcal{I}_k}\{\widehat{m}^{[\text{-}k]}(\boldsymbol{X}_i) - \bar{m}(\boldsymbol{X}_i)\}^2\right]^{\frac{1}{2}}\right),
\end{aligned}$$

which is again $o_p(1)$. Combining this with the asymptotic properties derived for V, Ξ_1 and Ξ_2 and the expansion (A2), we can finish the proof for the asymptotic normality of $\sqrt{n}(\boldsymbol{c}^{\intercal}\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}} - \boldsymbol{c}^{\intercal}\boldsymbol{\beta}_0)$.

At last, for the purpose of uncertainty quantification, we derive the form of the asymptotic variance of $\sqrt{n}(\boldsymbol{c}^{\intercal}\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}} - \boldsymbol{c}^{\intercal}\boldsymbol{\beta}_0)$. We start with the expansion:

$$\sqrt{n}(\boldsymbol{c}^{\intercal}\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}} - \boldsymbol{c}^{\intercal}\boldsymbol{\beta}_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n\boldsymbol{c}^{\intercal}F_i^{\mathcal{S}} + \frac{\sqrt{n}}{N}\sum_{n+1}^{n+N}\boldsymbol{c}^{\intercal}F_i^{\mathcal{T}} + \sqrt{n}\boldsymbol{\zeta}_{\alpha}^{\intercal}(\widehat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}}) + \sqrt{n}\boldsymbol{\zeta}_{\gamma}^{\intercal}(\widehat{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}}) + o_p(1), \tag{A6}$$

where $F_i^{\mathcal{S}} = \bar{\omega}(\boldsymbol{X}_i)\boldsymbol{A}_i\{Y_i - \bar{m}(\boldsymbol{X}_i)\}$, $F_i^{\mathcal{T}} = \boldsymbol{A}_i\{\bar{m}(\boldsymbol{X}_i) - g(\boldsymbol{A}_i^{\intercal}\boldsymbol{\beta})\}$,

$$\begin{aligned}
\boldsymbol{\zeta}_{\alpha} &= \mathbb{E}_1\bar{\omega}(\boldsymbol{X})\boldsymbol{\kappa}_{\boldsymbol{\beta}_0}\left[Y - g_m\{\boldsymbol{\Phi}^{\intercal}\bar{\boldsymbol{\gamma}} + \bar{r}(\boldsymbol{Z})\}\right]\boldsymbol{\Psi},\\
\boldsymbol{\zeta}_{\gamma} &= \mathbb{E}_1\bar{\omega}(\boldsymbol{X})\boldsymbol{\kappa}_{\boldsymbol{\beta}_0}\breve{g}_m\{\bar{m}(\boldsymbol{X})\}\boldsymbol{\Phi} - \mathbb{E}_0\boldsymbol{\kappa}_{\boldsymbol{\beta}_0}\breve{g}_m\{\bar{m}(\boldsymbol{X})\}\boldsymbol{\Phi},
\end{aligned}$$

$\widehat{\boldsymbol{\alpha}} = K^{-1}\sum_{k=1}^K\widehat{\boldsymbol{\alpha}}^{[\text{-}k]}$, and $\widehat{\boldsymbol{\gamma}} = K^{-1}\sum_{k=1}^K\widehat{\boldsymbol{\gamma}}^{[\text{-}k]}$. Let

$$\sqrt{n}(\widehat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}}) = \frac{1}{\sqrt{n}}\left(\sum_{i=1}^n F_i^{\mathcal{S},\boldsymbol{\alpha}} + \frac{n}{N}\sum_{i=n+1}^{n+N}F_i^{\mathcal{T},\boldsymbol{\alpha}}\right); \quad \sqrt{n}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) = \frac{1}{\sqrt{n}}\sum_{i=1}^n F_i^{\boldsymbol{\gamma}}.$$

Here $F_i^{\mathcal{S},\boldsymbol{\alpha}}$, $F_i^{\mathcal{T},\boldsymbol{\alpha}}$, and F_i^{γ} are dependent on sample i, with their specific forms depending on the estimating equations for $\widehat{\boldsymbol{\alpha}}^{[\text{-}k]}$ and $\widehat{\boldsymbol{\gamma}}^{[\text{-}k]}$ (which is flexible and upon one's choice in our framework). As an example, when $\widehat{\boldsymbol{\alpha}}^{[\text{-}k]} = \widetilde{\boldsymbol{\alpha}}^{[\text{-}k]}$ and $\widehat{\boldsymbol{\gamma}}^{[\text{-}k]} = \widetilde{\boldsymbol{\gamma}}^{[\text{-}k]}$ are estimated through equations (9) and (10), we will have

$$F_i^{\mathcal{S},\boldsymbol{\alpha}} = -\bar{\boldsymbol{J}}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{\Psi}_i \exp\left\{\boldsymbol{\Psi}_i^{\mathsf{T}}\boldsymbol{\alpha}^* + h^*(\boldsymbol{Z}_i)\right\},$$
$$F_i^{\mathcal{T},\boldsymbol{\alpha}} = \bar{\boldsymbol{J}}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{\Psi}_i, \quad , F_i^{\gamma} = \bar{\boldsymbol{J}}_{\boldsymbol{\gamma}}^{-1}\boldsymbol{\Phi}_i\left[Y_i - g_m\left\{\boldsymbol{\Phi}_i^{\mathsf{T}}\boldsymbol{\gamma}^* + r^*(\boldsymbol{Z}_i)\right\}\right],$$

where

$$\bar{\boldsymbol{J}}_{\boldsymbol{\alpha}} = \mathbb{E}_1 \exp\left\{\boldsymbol{\Psi}_i^{\mathsf{T}}\boldsymbol{\alpha}^* + h^*(\boldsymbol{Z}_i)\right\}\left(\boldsymbol{\Psi}_i - \frac{\mathbb{E}_1[\exp\left\{\boldsymbol{\Psi}_i^{\mathsf{T}}\boldsymbol{\alpha}^* + h^*(\boldsymbol{Z}_i)\right\}\boldsymbol{\Psi}_i|\boldsymbol{Z}_i]}{\mathbb{E}_1[\exp\left\{\boldsymbol{\Psi}_i^{\mathsf{T}}\boldsymbol{\alpha}^* + h^*(\boldsymbol{Z}_i)\right\}|\boldsymbol{Z}_i]}\right)^{\otimes 2},$$
$$\bar{\boldsymbol{J}}_{\boldsymbol{\gamma}} = \mathbb{E}_1\dot{g}_m\left\{\boldsymbol{\Phi}_i^{\mathsf{T}}\boldsymbol{\gamma}^* + r^*(\boldsymbol{Z}_i)\right\}\left(\boldsymbol{\Phi}_i - \frac{\mathbb{E}_1[\dot{g}_m\left\{\boldsymbol{\Phi}_i^{\mathsf{T}}\boldsymbol{\alpha}^* + r^*(\boldsymbol{Z}_i)\right\}\boldsymbol{\Phi}_i|\boldsymbol{Z}_i]}{\mathbb{E}_1[\dot{g}_m\left\{\boldsymbol{\Phi}_i^{\mathsf{T}}\boldsymbol{\alpha}^* + r^*(\boldsymbol{Z}_i)\right\}|\boldsymbol{Z}_i]}\right)^{\otimes 2},$$

and $\boldsymbol{u}^{\otimes 2} = \boldsymbol{u}\boldsymbol{u}^{\mathsf{T}}$. Plugging the expansions of $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\gamma}}$ into equation (A6), we have

$$\sqrt{n}(\boldsymbol{c}^{\mathsf{T}}\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}} - \boldsymbol{c}^{\mathsf{T}}\boldsymbol{\beta}_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\boldsymbol{c}^{\mathsf{T}}F_i^{\mathcal{S}} + \boldsymbol{\zeta}_{\alpha}^{\mathsf{T}}F_i^{\mathcal{S},\boldsymbol{\alpha}} + \boldsymbol{\zeta}_{\gamma}^{\mathsf{T}}F_i^{\gamma}) + \frac{\sqrt{n}}{N}\sum_{n+1}^{n+N}(\boldsymbol{c}^{\mathsf{T}}F_i^{\mathcal{T}} + \boldsymbol{\zeta}_{\alpha}^{\mathsf{T}}F_i^{\mathcal{T},\boldsymbol{\alpha}}) + o_p(1),$$

which implies the asymptotic variance of $\sqrt{n}(\boldsymbol{c}^{\mathsf{T}}\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}} - \boldsymbol{c}^{\mathsf{T}}\boldsymbol{\beta}_0)$ to be

$$\sigma^2 = \mathrm{Var}\left(\boldsymbol{c}^{\mathsf{T}}F_i^{\mathcal{S}} + \boldsymbol{\zeta}_{\alpha}^{\mathsf{T}}F_i^{\mathcal{S},\boldsymbol{\alpha}} + \boldsymbol{\zeta}_{\gamma}^{\mathsf{T}}F_i^{\gamma}\right) + \frac{n}{N}\mathrm{Var}\left(\boldsymbol{c}^{\mathsf{T}}F_i^{\mathcal{T}} + \boldsymbol{\zeta}_{\alpha}^{\mathsf{T}}F_i^{\mathcal{T},\boldsymbol{\alpha}}\right). \tag{A7}$$

Empirically, one could obtain a consistent estimation of σ^2 using the standard method of moments, with all the parameters replaced by their plug-in estimators. Clearly, by Slutsky's Theorem, the CIs constructed using the asymptotic normality of $\sqrt{n}(\boldsymbol{c}^{\mathsf{T}}\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}} - \boldsymbol{c}^{\mathsf{T}}\boldsymbol{\beta}_0)$ and the consistent estimator of σ^2 will be asymptotically valid.

■

Appendix B. Additional assumptions and justification of Proposition 1

In this section, we present the additional assumptions and justification for Proposition 1 that establishes the convergence rates and asymptotic behavior of our mainly studied nuisance estimators defined in Section 2.3. Our results are largely based on early literature of kernel and sieve methods like Fan et al. (1995), Newey (1997), Shen (1997), Carroll et al. (1998) and Chen (2007). We would also suggest that refined and sharper results could be potentially obtained inspired by recent literature like Belloni et al. (2015) and Cattaneo et al. (2020). For example, Belloni et al. (2015) substantially weakened the upper-bound condition on the dimension of the approximating functions for more general types of sieve estimators. For partitioning sieve estimation, Cattaneo et al. (2020) proposed an integrated mean squared error (IMSE) expansion method that can produce IMSE-optimal estimators. They also established bias-corrected inference procedures in this scheme.

Denote by $G_m(x) = \int_{-\infty}^{x} g_m(t)dt$. Let Λ_{α^*}, Λ_{γ^*}, Λ_{h^*}, Λ_{r^*}, $\Lambda_{\bar{h}}$ and $\Lambda_{\bar{r}}$ represent the parameter space of α^*, γ^*, h^*, r^*, $\bar{h}$ and $\bar{r}$ respectively. Let $\mathcal{Z}$ be the domain of $\boldsymbol{Z} \in \mathbb{R}^{p_{\boldsymbol{z}}}$ and $\mathcal{C}^k(\mathcal{Z})$ represent all the k-times differentiable continuous functions on $\mathcal{Z}$. The Hölder (or ν-smooth) class $\Sigma(\nu, L)$ is defined as the set of functions $f \in \mathcal{C}^{[\nu]}(\mathcal{Z})$ with its $[\nu]$-times derivative satisfying

$$\sup_{\boldsymbol{z}_1,\boldsymbol{z}_2\in\mathcal{Z}} \frac{\|f^{([\nu])}(\boldsymbol{z}_1) - f^{([\nu])}(\boldsymbol{z}_2)\|_2}{\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_2} \leq L.$$

Assumption A1 *(i) $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$ and $\boldsymbol{Z}$ have compact domain and continuous differentiable probability density functions (as given for discrete variables).*

(ii) There exists $C_1 > 0$ that for all $\boldsymbol{z} \in \mathcal{Z}$,

$$\|\boldsymbol{\alpha}^*\|_\infty, \|\boldsymbol{\gamma}^*\|_\infty, |h^*(\boldsymbol{z})|, |r^*(\boldsymbol{z})|, |\bar{h}(\boldsymbol{z})|, |\bar{r}(\boldsymbol{z})| \leq C_1.$$

(iii) There exists $C_2 > 0$ such that

$$C_2^{-1} \leq \frac{\frac{\partial}{\partial\tau}\mathbb{E}_1 \exp\{\boldsymbol{\Psi}^\intercal[\boldsymbol{\alpha}_1 + \tau(\boldsymbol{\alpha}_2 - \boldsymbol{\alpha}_1)] + h_1(\boldsymbol{Z}) + \tau[h_2(\boldsymbol{Z}) - h_1(\boldsymbol{Z})]\}}{\|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\|_2^2 + \mathbb{E}_1[h_1(\boldsymbol{Z}) - h_2(\boldsymbol{Z})]^2} \leq C_2;$$
$$C_2^{-1} \leq \frac{\frac{\partial}{\partial\tau}\mathbb{E}_1 G_m\{\boldsymbol{\Phi}^\intercal[\boldsymbol{\gamma}_1 + \tau(\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_1)] + r_1(\boldsymbol{Z}) + \tau[r_2(\boldsymbol{Z}) - r_1(\boldsymbol{Z})]\}}{\|\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2\|_2^2 + \mathbb{E}_1[r_1(\boldsymbol{Z}) - r_2(\boldsymbol{Z})]^2} \leq C_2,$$

for any $\tau \in [0, 1]$, $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \Lambda_{\alpha^*}$, $h_1, h_2 \in \Lambda_{h^*}$, $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2 \in \Lambda_{\gamma^*}$, and $r_1, r_2 \in \Lambda_{r^*}$.

(iv) It holds that $\boldsymbol{\kappa}_{\boldsymbol{\beta}_0} \geq 0$ with probability 1. There exists $C_3 > 0$ that for all $\boldsymbol{z} \in \mathcal{Z}$,

$$C_3^{-1} \leq \left|h^{-p_{\boldsymbol{z}}}\mathbb{E}_1 K_h(\boldsymbol{Z} - \boldsymbol{z})\omega^*(\boldsymbol{X})\boldsymbol{\kappa}_{\boldsymbol{\beta}_0}\dot{g}_m\left\{\boldsymbol{\Phi}^\intercal\bar{\boldsymbol{\gamma}} + \bar{r}(\boldsymbol{z})\right\}\right| \leq C_3;$$
$$C_3^{-1} \leq \left|h^{-p_{\boldsymbol{z}}}\mathbb{E}_1 K_h(\boldsymbol{Z} - \boldsymbol{z})\exp(\boldsymbol{\Psi}^\intercal\bar{\boldsymbol{\alpha}})\boldsymbol{\kappa}_{\boldsymbol{\beta}_0}\breve{g}_m\{m^*(\boldsymbol{X})\}\exp\{\bar{h}(\boldsymbol{z})\}\right| \leq C_3.$$

Assumption A2 *There exists $\nu, L > 0$ such that all population-level nonparametric components $h^*(\boldsymbol{z}), r^*(\boldsymbol{z}), \bar{h}(\boldsymbol{z})$ and $\bar{r}(\boldsymbol{z})$ belong to the Hölder class $\Sigma(\nu, L)$ with the degree of smoothness ν satisfying $\nu > p_{\boldsymbol{z}}$.*

Assumption A3 (Specification of the sieve and kernel functions) *(i) The basis function $\boldsymbol{b}(\boldsymbol{Z})$ is taken as the tensor product of $\boldsymbol{b}_j(Z_j)$ for $j = 1, 2, \ldots, p_{\boldsymbol{z}}$, where each $\boldsymbol{b}_j(Z_j)$ is the Hermite polynomial basis of the univariate Z_j with its order $s \asymp n^{1/(p_z+\nu)}$. (ii) The kernel function K is symmetric, bounded, and of order $[\nu]$ and the bandwidth $h \asymp n^{-1/(p_z+2\nu)}$. The tuning parameters $\lambda_1, \lambda_2 = o(n^{-1/2})$.*

Remark A1 *Similar to Assumption 1 in the main paper, Assumptions A1(i) and A1(ii) are used to regular the distribution of $\boldsymbol{X}$ and the parameter spaces. Assumption A1(iii) is in a similar spirit of Condition 4.5 in Chen (2007), used to control the asymptotic variance of $\sqrt{n}(\widetilde{\boldsymbol{\alpha}}^{[\text{-}k]} - \boldsymbol{\alpha}^*)$ and $\sqrt{n}(\widetilde{\boldsymbol{\alpha}}^{[\text{-}k]} - \boldsymbol{\alpha}^*)$. Assumption A1(iv) requires the weighting term $\boldsymbol{\kappa}_{\boldsymbol{\beta}_0}$ to be positive-definite to ensure the regularity of the calibration equations. As we remark in Remark 5, this assumption can be granted by splitting the samples by the sign of $\boldsymbol{\kappa}_{\widetilde{\boldsymbol{\beta}}}$ when it is not always positive or always negative.*

Assumption A2 imposes the common smoothness conditions on the nuisance nonparametric components that are also used in semiparametric inference existing literature like Chen et al. (2008) and Rothe and Firpo (2015). Although $h^*(\cdot), r^*(\cdot), \bar{h}(\cdot)$ and $\bar{r}(\cdot)$ may come from integral equations with misspecified models, it is not hard to show that they will be smooth as long as the distribution function of $\boldsymbol{Z}$ and the true conditional models $\mathtt{w}(\cdot)$ and $\mu(\cdot)$ are smooth (on $\boldsymbol{Z}$). Thus, Assumption A2 could be implied by standard and reasonable smoothness assumptions on the true data-generation functions.

In Assumption A3, we choose the order of sieve of the preliminary nuisance estimators to be under-smoothed optimal since $\sqrt{n}$-consistency of the parametric part in these models is required. While the bandwidth h used in the calibrated estimating equation (11) can be rate-optimal since we do not need to estimate the parametric components in this step.

Proof [Proof of Proposition 1] Since we simply pick $\widehat{\boldsymbol{\alpha}}^{[\text{-}k]} = \widetilde{\boldsymbol{\alpha}}^{[\text{-}k]}$ and $\widehat{\boldsymbol{\gamma}}^{[\text{-}k]} = \widetilde{\boldsymbol{\gamma}}^{[\text{-}k]}$ in Section 2.3, Assumptions 1 and A1–A3 are sufficient for Assumption 3(i) by Lemma A3(b) presented and justified in this section. And Assumption 3(ii) is directly given by Lemma A4 that is proved based on Lemmas A1–A3.

■

Lemma A1 establishes the desirable convergence properties of the preliminary nuisance estimators based on the existing analysis of sieve M-estimation (Shen, 1997; Chen, 2007).

Lemma A1 ((Shen, 1997; Chen, 2007)) *Under Assumptions 1 and A1–A3, the preliminary nuisance estimators solved from equations (9) and (10) satisfy that:*
(a) For $j \in \{0, 1\}$,

$$\mathbb{E}_1\{\widetilde{r}^{[\text{-}k]}(\boldsymbol{Z}) - r^*(\boldsymbol{Z})\}^2 + \mathbb{E}_j\{\widetilde{h}^{[\text{-}k]}(\boldsymbol{Z}) - h^*(\boldsymbol{Z})\}^2 = o_p(n^{-1/2});$$
$$\sup_{\boldsymbol{z}\in\mathcal{Z}} |\widetilde{r}^{[\text{-}k]}(\boldsymbol{z}) - r^*(\boldsymbol{z})| + |\widetilde{h}^{[\text{-}k]}(\boldsymbol{z}) - h^*(\boldsymbol{z})| = o_p(1);$$

(b) $\sqrt{n}(\widetilde{\boldsymbol{\alpha}}^{[\text{-}k]} - \boldsymbol{\alpha}^*)$ and $\sqrt{n}(\widetilde{\boldsymbol{\alpha}}^{[\text{-}k]} - \boldsymbol{\alpha}^*)$ weakly converge to gaussian distributuon with mean zero and finite variance.

Proof We based on Theorem 3.5 of Chen (2007) to show (a) of Lemma A1. First, note that for both preliminary nuisance models, Conditions 3.9, 3.10, 3.11 and 3.13 of Chen (2007) are implied by Assumptions 1, A1(i) and A1(ii). Their Condition 3.12 is implied by Assumption A1(iii). Then by their Theorem 3.5, it holds that

$$\|\widetilde{\boldsymbol{\gamma}}^{[\text{-}k]} - \boldsymbol{\gamma}^*\|_2^2 + \mathbb{E}_1\{\widetilde{r}^{[\text{-}k]}(\boldsymbol{Z}) - r^*(\boldsymbol{Z})\}^2 = O_p\left(\frac{k_n}{n} + \rho_{2n}^2\right);$$
$$\|\widetilde{\boldsymbol{\alpha}}^{[\text{-}k]} - \boldsymbol{\alpha}^*\|_2^2 + \mathbb{E}_1\{\widetilde{h}^{[\text{-}k]}(\boldsymbol{Z}) - h^*(\boldsymbol{Z})\}^2 = O_p\left(\frac{k_n}{n} + \rho_{2n}^2\right),$$

where k_n and ρ_{2n}^2 respectively characterize the variance and approximation bias of sieve to be specified as follows. Inspired by Proposition 3.6 of Chen (2007), under our Assumptions A2 and A3(i), the specific rate of k_n and ρ_{2n}^2 is given by

$$k_n \asymp s^{p_z}, \quad \rho_{2n} \asymp s^{-\nu}, \text{ where } s \text{ is the order of each } \boldsymbol{b}_j(Z_j).$$

Then by Assumption A2 that $\nu > p_{\boldsymbol{z}}$ and Assumption A3(i) that $s \asymp n^{1/(p_z+\nu)}$, we have

$$\|\widetilde{\boldsymbol{\gamma}}^{[\text{-}k]} - \boldsymbol{\gamma}^*\|_2^2 + \mathbb{E}_1\{\widetilde{r}^{[\text{-}k]}(\boldsymbol{Z}) - r^*(\boldsymbol{Z})\}^2 = o_p(n^{-1/2});$$
$$\|\widetilde{\boldsymbol{\alpha}}^{[\text{-}k]} - \boldsymbol{\alpha}^*\|_2^2 + \mathbb{E}_1\{\widetilde{h}^{[\text{-}k]}(\boldsymbol{Z}) - h^*(\boldsymbol{Z})\}^2 = o_p(n^{-1/2}).$$

Similarly, it is not hard to justify that our Assumptions 1 and A1–A3 imply Conditions 3.1, 3.2, 3.4 and 3.5M of Chen (2007), which are sufficient for the consistency of sieve M-estimation according to their Remark 3.3, i.e.,

$$\sup_{\boldsymbol{z}\in\mathcal{Z}} |\widetilde{r}^{[\text{-}k]}(\boldsymbol{z}) - r^*(\boldsymbol{z})| + |\widetilde{h}^{[\text{-}k]}(\boldsymbol{z}) - h^*(\boldsymbol{z})| = o_p(1).$$

So we finish proving (a) of Lemma A1.

Next, we prove (b) based on (a) and using Theorem 4.3 of Chen (2007) (or early works like Shen (1997)). Their Conditions 4.1(iii) and 4.4 are as given in our standard non-linear M-estimation case. Since "$f(\theta)$" in Chen (2007) are simply the parametric parts $\boldsymbol{\gamma}$ or $\boldsymbol{\alpha}$ in our case, their Conditions 4.1(i) and 4.2(ii) are trivially satisfied. Their Condition 4.5 is implied by our Assumption A1(iii) that actually indicates $\sqrt{n}(\widetilde{\boldsymbol{\alpha}}^{[\text{-}k]} - \boldsymbol{\alpha}^*)$ and $\sqrt{n}(\widetilde{\boldsymbol{\alpha}}^{[\text{-}k]} - \boldsymbol{\alpha}^*)$ will have bounded asymptotic variance. And their Conditions 4.2' and 4.3' are implied by Assumption A1(i) and the continuity of the link function g. Therefore, we can combine our Lemma A1(a) and Theorem 4.3 of Chen (2007) to finishe the proof of Lemma A1(b).

■

Using Lemma A1 and that at least one nuisance model is correctly specified (i.e., Assumption 2), Lemma A2 establishes the $o_p(n^{-1/4})$ convergence of the preliminary estimator $\widetilde{\boldsymbol{\beta}}^{[\text{-}k]}$ to the true $\boldsymbol{\beta}_0$.

Lemma A2 *Under Assumptions 1, 2 and A1–A3,*

$$\mathbb{E}_j\{\widetilde{m}^{[\text{-}k]}(\boldsymbol{X}) - m^*(\boldsymbol{X})\}^2 + \mathbb{E}_1\{\widetilde{\omega}^{[\text{-}k]}(\boldsymbol{X}) - \omega^*(\boldsymbol{X})\}^2 + \|\widetilde{\boldsymbol{\beta}}^{[\text{-}k]} - \boldsymbol{\beta}_0\|_2^2 = o_p(n^{-1/2}).$$

Proof It immediately follows from Lemma A1 that

$$\mathbb{E}_j\{\widetilde{m}^{[\text{-}k]}(\boldsymbol{X})-m^*(\boldsymbol{X})\}^2+\mathbb{E}_1\{\widetilde{\omega}^{[\text{-}k]}(\boldsymbol{X})-\omega^*(\boldsymbol{X})\}^2=o_p(n^{-1/2}).$$

Then $\|\widetilde{\boldsymbol{\beta}}^{[\text{-}k]}-\boldsymbol{\beta}_0\|_2^2=o_p(n^{-1/2})$ can be proved by following the same proof procedures in Theorem 1 for analyzing the terms defined in (A1). ■

For each $\boldsymbol{z}\in\mathcal{Z}$, let the estimators $\breve{r}^{[\text{-}k]}(\boldsymbol{z})$ and $\breve{h}^{[\text{-}k]}(\boldsymbol{z})$ respectively solve:

$$\begin{aligned}&\frac{K}{n(K-1)h^{p_{\boldsymbol{z}}}}\sum_{i\in\mathcal{I}_{\text{-}k}}K_h(\boldsymbol{Z}_i-\boldsymbol{z})\omega^*(\boldsymbol{X}_i)\boldsymbol{\kappa}_{i,\boldsymbol{\beta}_0}\left[Y_i-g_m\left\{\boldsymbol{\Phi}_i^{\intercal}\bar{\boldsymbol{\gamma}}+r(\boldsymbol{z})\right\}\right]=0;\\&\frac{K}{n(K-1)h^{p_{\boldsymbol{z}}}}\sum_{i\in\mathcal{I}_{\text{-}k}}K_h(\boldsymbol{Z}_i-\boldsymbol{z})\exp(\boldsymbol{\Psi}_i^{\intercal}\bar{\boldsymbol{\alpha}})\boldsymbol{\kappa}_{i,\boldsymbol{\beta}_0}\breve{g}_m\{m^*(\boldsymbol{X}_i)\}\exp\{h(\boldsymbol{z})\}\\=&\frac{1}{Nh^{p_{\boldsymbol{z}}}}\sum_{i=n+1}^{n+N}K_h(\boldsymbol{Z}_i-\boldsymbol{z})\boldsymbol{\kappa}_{i,\boldsymbol{\beta}_0}\breve{g}_m\{m^*(\boldsymbol{X}_i)\},\end{aligned}\tag{A8}$$

i.e. the “oracle” version of the estimating equations in (11), obtained by replacing all the preliminary estimators plugged in (11) with their limits (true values). Also recall that $\bar{h}(\boldsymbol{z})$ and $\bar{r}(\boldsymbol{z})$ are defined as the solutions to equations (7) and (8).

We introduce Lemma A3 to give the consistency $o_p(n^{-1/4})$ convergence of $\breve{h}^{[\text{-}k]}(\boldsymbol{z})$ and $\breve{r}^{[\text{-}k]}(\boldsymbol{z})$ to $\bar{h}(\boldsymbol{z})$ and $\bar{r}(\boldsymbol{z})$, as a standard result of the higher–order kernel (or local polynomial) estimating equation (Fan et al., 1995).

Lemma A3 *Under Assumptions 1, 2 and A1–A3,*

$$\begin{aligned}&\mathbb{E}_1\{\breve{r}^{[\text{-}k]}(\boldsymbol{Z})-\bar{r}(\boldsymbol{Z})\}^2+\mathbb{E}_1\{\breve{h}^{[\text{-}k]}(\boldsymbol{Z})-\bar{h}(\boldsymbol{Z})\}^2=o_p(n^{-1/2});\\&\sup_{\boldsymbol{z}\in\mathcal{Z}}|\breve{r}^{[\text{-}k]}(\boldsymbol{z})-\bar{r}(\boldsymbol{z})|+|\breve{h}^{[\text{-}k]}(\boldsymbol{z})-\bar{h}(\boldsymbol{z})|=o_p(1).\end{aligned}$$

Proof By Assumption 2, at least one nuisance model is correctly specified. When the importance weighting model is correct, $w^*(\boldsymbol{x})=\bar{w}(\boldsymbol{x})=\mathbb{w}(\boldsymbol{x})$. So the first equation of (A8) is (asymptotically) valid for $\bar{r}(\boldsymbol{Z})$ that solves (7). Also, since $\mathbb{w}(\boldsymbol{x})=\exp(\psi^{\intercal}\boldsymbol{\alpha}_0+h_0(\boldsymbol{z}))$ and $\bar{\boldsymbol{\alpha}}=\boldsymbol{\alpha}_0$ when the importance weighting model is correct, the second equation of (A8) is valid for $\bar{h}(\boldsymbol{z})=h_0(\boldsymbol{z})$ that solves (8). So both equations in (A8) are valid. Similarly, this also holds when the imputation model is correct. Then by Assumptions 1, and A1–A3 and following Appendix A of Fan et al. (1995), we can derive that $\sup_{\boldsymbol{z}\in\mathcal{Z}}|\breve{r}^{[\text{-}k]}(\boldsymbol{z})-\bar{r}(\boldsymbol{z})|+|\breve{h}^{[\text{-}k]}(\boldsymbol{z})-\bar{h}(\boldsymbol{z})|=o_p(1)$ and

$$\mathbb{E}_1\{\breve{r}^{[\text{-}k]}(\boldsymbol{Z})-\bar{r}(\boldsymbol{Z})\}^2+\mathbb{E}_1\{\breve{h}^{[\text{-}k]}(\boldsymbol{Z})-\bar{h}(\boldsymbol{Z})\}^2=O_p\left(\frac{1}{nh^{p_{\boldsymbol{z}}}}+h^{2\nu}\right)=o_p(n^{-1/2}),$$

as the standard consistency and convergence results of kernel smoothing.

Note that (Fan et al., 1995) studied the local polynomial regression approach that is not exactly the same as our used $[\nu]$-th order kernel; see Assumption A3(ii). While the derivation of these two approaches are basically the same due to the orthogonality between a $[\nu]$-th order kernel function and the polynomial functions of the order up to $[\nu]$.

■

Finally, we come to Lemma A4 for the asymptotic properties of $\widehat{r}^{[\text{-}k]}(\boldsymbol{Z})$ and $\widehat{h}^{[\text{-}k]}(\boldsymbol{Z})$.

Lemma A4 *Under Assumptions 1, 2 and A1–A3, the calibrated nuisance estimators satisfy:*

$$\begin{aligned}
&\mathbb{E}_1\{\widehat{r}^{[\text{-}k]}(\boldsymbol{Z})-\bar{r}(\boldsymbol{Z})\}^2+\mathbb{E}_1\{\widehat{h}^{[\text{-}k]}(\boldsymbol{Z})-\bar{h}(\boldsymbol{Z})\}^2=o_p(n^{-1/2});\\
&\sup_{\boldsymbol{z}\in\mathcal{Z}}|\widehat{r}^{[\text{-}k]}(\boldsymbol{z})-\bar{r}(\boldsymbol{z})|+|\widehat{h}^{[\text{-}k]}(\boldsymbol{z})-\bar{h}(\boldsymbol{z})|=o_p(1).
\end{aligned}$$

Proof We compare the estimating equations in (11) with those in (A8) to analyze the additional errors incurred by the preliminary estimators in (11). By Assumption 1 and equation (A3) derived in the proof of Theorem 1, we have that for each $\boldsymbol{z}$,

$$\begin{aligned}
0=&\frac{K}{n(K-1)h^{p_{\boldsymbol{z}}}}\sum_{i\in\mathcal{I}_{\text{-}k}}K_h(\boldsymbol{Z}_i-\boldsymbol{z})\widetilde{\omega}^{[\text{-}k]}(\boldsymbol{X}_i)\boldsymbol{c}^{\intercal}\widehat{\boldsymbol{J}}^{-1}_{\widetilde{\boldsymbol{\beta}}^{[\text{-}k]}}\boldsymbol{A}_i\left[Y_i-g_m\left\{\boldsymbol{\Phi}_i^{\intercal}\widehat{\boldsymbol{\gamma}}^{[\text{-}k]}+\widehat{r}^{[\text{-}k]}(\boldsymbol{z})\right\}\right]\\
=&\frac{K}{n(K-1)h^{p_{\boldsymbol{z}}}}\sum_{i\in\mathcal{I}_{\text{-}k}}K_h(\boldsymbol{Z}_i-\boldsymbol{z})\omega^*(\boldsymbol{X}_i)\boldsymbol{\kappa}_{i,\boldsymbol{\beta}_0}\left[Y_i-g_m\left\{\boldsymbol{\Phi}_i\bar{\boldsymbol{\gamma}}+\widehat{r}^{[\text{-}k]}(\boldsymbol{z})\right\}\right]\\
&+\frac{K}{n(K-1)h^{p_{\boldsymbol{z}}}}\sum_{i\in\mathcal{I}_{\text{-}k}}K_h(\boldsymbol{Z}_i-\boldsymbol{z})\omega^*(\boldsymbol{X}_i)\boldsymbol{\kappa}_{i,\boldsymbol{\beta}_0}\left[g_m\left\{\boldsymbol{\Phi}_i^{\intercal}\bar{\boldsymbol{\gamma}}+\widehat{r}^{[\text{-}k]}(\boldsymbol{z})\right\}-g_m\left\{\boldsymbol{\Phi}_i^{\intercal}\widehat{\boldsymbol{\gamma}}^{[\text{-}k]}+\widehat{r}^{[\text{-}k]}(\boldsymbol{z})\right\}\right]\\
&+\frac{K}{n(K-1)h^{p_{\boldsymbol{z}}}}\sum_{i\in\mathcal{I}_{\text{-}k}}K_h(\boldsymbol{Z}_i-\boldsymbol{z})\omega^*(\boldsymbol{X}_i)\boldsymbol{c}^{\intercal}\left[\widehat{\boldsymbol{J}}^{-1}_{\widetilde{\boldsymbol{\beta}}^{[\text{-}k]}}-\boldsymbol{J}^{-1}_{\boldsymbol{\beta}_0}\right]\boldsymbol{A}_i\left[Y_i-g_m\left\{\boldsymbol{\Phi}_i^{\intercal}\widehat{\boldsymbol{\gamma}}^{[\text{-}k]}+\widehat{r}^{[\text{-}k]}(\boldsymbol{z})\right\}\right]\\
&+\frac{K}{n(K-1)h^{p_{\boldsymbol{z}}}}\sum_{i\in\mathcal{I}_{\text{-}k}}K_h(\boldsymbol{Z}_i-\boldsymbol{z})\{\widetilde{\omega}^{[\text{-}k]}(\boldsymbol{X}_i)-\omega^*(\boldsymbol{X}_i)\}\boldsymbol{c}^{\intercal}\widehat{\boldsymbol{J}}^{-1}_{\widetilde{\boldsymbol{\beta}}^{[\text{-}k]}}\boldsymbol{A}_i\left[Y_i-g_m\left\{\boldsymbol{\Phi}_i^{\intercal}\widehat{\boldsymbol{\gamma}}^{[\text{-}k]}+\widehat{r}^{[\text{-}k]}(\boldsymbol{z})\right\}\right]\\
=&\frac{K}{n(K-1)h^{p_{\boldsymbol{z}}}}\sum_{i\in\mathcal{I}_{\text{-}k}}K_h(\boldsymbol{Z}_i-\boldsymbol{z})\omega^*(\boldsymbol{X}_i)\boldsymbol{\kappa}_{i,\boldsymbol{\beta}_0}\left[Y_i-g_m\left\{\boldsymbol{\Phi}_i^{\intercal}\bar{\boldsymbol{\gamma}}+\widehat{r}^{[\text{-}k]}(\boldsymbol{z})\right\}\right]\\
&+O_p\left(\left[\mathbb{E}_1\{\widetilde{\omega}^{[\text{-}k]}(\boldsymbol{X})-\omega^*(\boldsymbol{X})\}^2\right]^{\frac{1}{2}}+\|\widetilde{\boldsymbol{\beta}}^{[\text{-}k]}-\boldsymbol{\beta}_0\|_2+\|\widehat{\boldsymbol{\gamma}}^{[\text{-}k]}-\bar{\boldsymbol{\gamma}}\|_2+n^{-1/2}\right)\\
=&\frac{K}{n(K-1)h^{p_{\boldsymbol{z}}}}\sum_{i\in\mathcal{I}_{\text{-}k}}K_h(\boldsymbol{Z}_i-\boldsymbol{z})\omega^*(\boldsymbol{X}_i)\boldsymbol{\kappa}_{i,\boldsymbol{\beta}_0}\left[Y_i-g_m\left\{\boldsymbol{\Phi}_i^{\intercal}\bar{\boldsymbol{\gamma}}+\widehat{r}^{[\text{-}k]}(\boldsymbol{z})\right\}\right]+o_p(n^{-1/4}),
\end{aligned}$$

Comparing this with the estimating equation (A8) for $\breve{r}^{[\text{-}k]}(\cdot)$, we have:

$$\frac{K}{n(K-1)h^{p_{\boldsymbol{z}}}}\sum_{i\in\mathcal{I}_{\text{-}k}}K_h(\boldsymbol{Z}_i-\boldsymbol{z})\omega^*(\boldsymbol{X}_i)\boldsymbol{\kappa}_{i,\boldsymbol{\beta}_0}\left[g_m\left\{\boldsymbol{\Phi}_i^{\intercal}\bar{\boldsymbol{\gamma}}+\breve{r}^{[\text{-}k]}(\boldsymbol{z})\right\}-g_m\left\{\boldsymbol{\Phi}_i^{\intercal}\bar{\boldsymbol{\gamma}}+\widehat{r}^{[\text{-}k]}(\boldsymbol{z})\right\}\right]=o_p(n^{-1/4}),$$

which combined with Assumption 1 that $\dot{g}(\cdot)$ is Lipsitz, leads to

$$\begin{aligned}
&\frac{K}{n(K-1)h^{p_{\boldsymbol{z}}}}\sum_{i\in\mathcal{I}_{\text{-}k}}K_h(\boldsymbol{Z}_i-\boldsymbol{z})\omega^*(\boldsymbol{X}_i)\boldsymbol{\kappa}_{i,\boldsymbol{\beta}_0}\dot{g}_m\left\{\boldsymbol{\Phi}_i^{\intercal}\bar{\boldsymbol{\gamma}}+\bar{r}(\boldsymbol{z})\right\}\left|\breve{r}^{[\text{-}k]}(\boldsymbol{z})-\widehat{r}^{[\text{-}k]}(\boldsymbol{z})\right|\\
=&o_p(n^{-1/4})+O_p\left([\widehat{r}^{[\text{-}k]}(\boldsymbol{z})-\bar{r}(\boldsymbol{z})]^2+[\breve{r}^{[\text{-}k]}(\boldsymbol{z})-\bar{r}(\boldsymbol{z})]^2\right).
\end{aligned}$$

Using Assumption 1(iv) and the weak law of large numbers, we can show that

$$\frac{K}{n(K-1)h^{p_{\boldsymbol{z}}}}\sum_{i\in\mathcal{I}_{\text{-}k}} K_h(\boldsymbol{Z}_i-\boldsymbol{z})\omega^*(\boldsymbol{X}_i)\boldsymbol{\kappa}_{i,\boldsymbol{\beta}_0}\dot{g}_m\left\{\boldsymbol{\Phi}_i^{\intercal}\bar{\boldsymbol{\gamma}}+\bar{r}(\boldsymbol{z})\right\}\asymp 1.$$

Then by Lemma A3, we conclude that $|\widehat{r}^{[\text{-}k]}(\boldsymbol{z})-\bar{r}(\boldsymbol{z})|=o_p(1)$ uniformly for all $\boldsymbol{z}\in\mathcal{Z}$, and $\mathbb{E}_1\{\widehat{r}^{[\text{-}k]}(\boldsymbol{Z})-\bar{r}(\boldsymbol{Z})\}^2=o_p(n^{-1/2})$.

For $\widehat{h}^{[\text{-}k]}(\cdot)$, we follow the same strategy to consider the difference between the second equation of (11) and equation (A8), to derive that

$$\begin{aligned}&\frac{K}{n(K-1)h^{p_{\boldsymbol{z}}}}\sum_{i\in\mathcal{I}_{\text{-}k}} K_h(\boldsymbol{Z}_i-\boldsymbol{z})\exp(\boldsymbol{\Psi}_i^{\intercal}\bar{\boldsymbol{\alpha}})\boldsymbol{\kappa}_{i,\boldsymbol{\beta}_0}\breve{g}_m\{m^*(\boldsymbol{X}_i)\}\exp\{\bar{h}(\boldsymbol{z})\}\left|\breve{h}^{[\text{-}k]}(\boldsymbol{z})-\widehat{h}^{[\text{-}k]}(\boldsymbol{z})\right|\\
=&O_p\left(\left[\mathbb{E}_1\{\widetilde{m}^{[\text{-}k]}(\boldsymbol{X})-m^*(\boldsymbol{X})\}^2\right]^{\frac{1}{2}}+\|\widetilde{\boldsymbol{\beta}}^{[\text{-}k]}-\boldsymbol{\beta}_0\|_2\right)+O_p\left([\widehat{h}^{[\text{-}k]}(\boldsymbol{z})-\bar{h}(\boldsymbol{z})]^2+[\breve{h}^{[\text{-}k]}(\boldsymbol{z})-\bar{h}(\boldsymbol{z})]^2\right)\\
=&o_p(n^{-1/4})+O_p\left([\widehat{h}^{[\text{-}k]}(\boldsymbol{z})-\bar{h}(\boldsymbol{z})]^2+[\breve{h}^{[\text{-}k]}(\boldsymbol{z})-\bar{h}(\boldsymbol{z})]^2\right).\end{aligned}$$

Again combining this with Assumption 1(iv) and Lemma A3, we can derive that

$$\sup_{\boldsymbol{z}\in\mathcal{Z}}|\widehat{h}^{[\text{-}k]}(\boldsymbol{z})-\bar{h}(\boldsymbol{z})|=o_p(1);\quad \mathbb{E}_1\{\widehat{h}^{[\text{-}k]}(\boldsymbol{Z})-\bar{h}(\boldsymbol{Z})\}^2=o_p(n^{-1/2}).$$

Thus we have finished proving Lemma A4.

■

Appendix C. Details of the extension discussed in Section 6

C.1 Sieve estimator

We consider $r(\boldsymbol{Z}) = \boldsymbol{\xi}^\intercal \boldsymbol{b}(\boldsymbol{Z})$ and $h(\boldsymbol{Z}) = \boldsymbol{\eta}^\intercal \boldsymbol{b}(\boldsymbol{Z})$ where $\boldsymbol{b}(\boldsymbol{Z})$ represents some prespecified basis function of $\boldsymbol{Z}$, e.g. natural spline or Hermite polynomials with diverging dimensionality, and $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ represent their coefficients to estimate. In analog to (11), we propose to estimate the coefficients $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ by solving

$$\frac{K}{n(K-1)}\sum_{i\in\mathcal{I}_{\text{-}k}} \widetilde{\omega}^{[\text{-}k]}(\boldsymbol{X}_i)\boldsymbol{c}^\intercal \widehat{\boldsymbol{J}}^{-1}_{\widetilde{\boldsymbol{\beta}}^{[\text{-}k]}}\boldsymbol{A}_i\boldsymbol{b}(\boldsymbol{Z}_i)\left[Y_i - g_m\left\{\boldsymbol{\Phi}_i^\intercal\widehat{\boldsymbol{\gamma}}^{[\text{-}k]}+\boldsymbol{\xi}^\intercal\boldsymbol{b}(\boldsymbol{Z}_i)\right\}\right]=\boldsymbol{0};$$

$$\begin{aligned}&\frac{K}{n(K-1)}\sum_{i\in\mathcal{I}_{\text{-}k}} \boldsymbol{c}^\intercal \widehat{\boldsymbol{J}}^{-1}_{\widetilde{\boldsymbol{\beta}}^{[\text{-}k]}}\boldsymbol{A}_i\breve{g}_m\{\widetilde{m}^{[\text{-}k]}(\boldsymbol{X}_i)\}\exp\{\boldsymbol{\Psi}_i^\intercal\widehat{\boldsymbol{\alpha}}^{[\text{-}k]}+\boldsymbol{\eta}^\intercal\boldsymbol{b}(\boldsymbol{Z}_i)\}\boldsymbol{b}(\boldsymbol{Z}_i)\\ =&\frac{1}{N}\sum_{i=n+1}^{n+N}\boldsymbol{c}^\intercal \widehat{\boldsymbol{J}}^{-1}_{\widetilde{\boldsymbol{\beta}}^{[\text{-}k]}}\boldsymbol{A}_i\breve{g}_m\{\widetilde{m}^{[\text{-}k]}(\boldsymbol{X}_i)\}\boldsymbol{b}(\boldsymbol{Z}_i).\end{aligned}$$

For one-dimensional $\boldsymbol{Z}_i$ occurring in our numerical studies, this sieve approach should have similar performance as kernel smoothing. While if $p_{\boldsymbol{z}} > 1$ and $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ip_{\boldsymbol{z}}})^\intercal$, classic nonparametric approaches like kernel smoothing and sieve could have poor performance due to the curse of dimensionality. One may use additive model of $Z_{i1}, \ldots, Z_{ip_{\boldsymbol{z}}}$ (constructed with the basis $\{\boldsymbol{b}^\intercal(Z_{i1}), \ldots, \boldsymbol{b}^\intercal(Z_{ip_{\boldsymbol{z}}})\}^\intercal$) instead of the fully nonparametric model for $\boldsymbol{Z}_i$, to avoid excessive model complexity.

C.2 General machine learning method

Given a response A, predictors $\boldsymbol{C}$, and an arbitrary black-box learning algorithm $\mathcal{L}$, we let $\widehat{\mathcal{E}}^{\mathcal{L}}[A \mid \boldsymbol{C}]$ and $\widehat{\mathcal{P}}^{\mathcal{L}}(A \mid \boldsymbol{C})$ denote the conditional expectation and conditional probability density (or mass) function of A on $\boldsymbol{C}$ estimated using the learning algorithm $\mathcal{L}$. Here, we neglect the index of training samples in our notation for simplicity while in general, one should follow the established work like Chernozhukov et al. (2018a), to adopt cross-fitting, and ensure that $\widehat{\mathcal{E}}^{\mathcal{L}}[A \mid \boldsymbol{C}]$ and $\widehat{\mathcal{P}}^{\mathcal{L}}(A \mid \boldsymbol{C})$ are estimated using training data independent with their plug-in samples.

Without loss of generality, we assume that knowing $\boldsymbol{X}$ is sufficient to identify $\boldsymbol{Z}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Psi}$. We propose novel procedures using $\mathcal{L}$ to estimate and calibrate the nuisance models. First, we regress Y on $\boldsymbol{X}$ on $\mathcal{S}$ using learning algorithm $\mathcal{L}$ to obtain $\widehat{\mathcal{E}}^{\mathcal{L}}[Y \mid \boldsymbol{X}]$, and regress S on $\boldsymbol{X}$ to obtain $\widehat{\mathcal{P}}^{\mathcal{L}}(S = 1 \mid \boldsymbol{X})$. Also, we use $\mathcal{L}$ to learn $\widehat{\mathcal{P}}^{\mathcal{L}}(\boldsymbol{X} \mid \boldsymbol{Z}, S = 1)$, i.e. the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{Z}$ on the source population. Then we solve:

$$\begin{aligned}&\frac{K}{n(K-1)}\sum_{i\in\mathcal{I}_{\text{-}k}}\boldsymbol{\Phi}_i\left\{\widehat{\mathcal{E}}^{\mathcal{L}}[Y_i \mid \boldsymbol{X}_i] - g_m[\boldsymbol{\Phi}_i^\intercal\boldsymbol{\gamma} + r(\boldsymbol{Z}_i)]\right\} = \boldsymbol{0},\\ &\int_{\boldsymbol{x}\in\mathcal{X}\cap\{\boldsymbol{z}\}}\widehat{\mathcal{P}}^{\mathcal{L}}(\boldsymbol{x} \mid \boldsymbol{Z} = \boldsymbol{z}, S = 1)\left\{\widehat{\mathcal{E}}^{\mathcal{L}}[Y \mid \boldsymbol{X} = \boldsymbol{x}] - g_m[\boldsymbol{\Phi}_i^\intercal\boldsymbol{\gamma} + r(\boldsymbol{z})]\right\}d\boldsymbol{x} = 0, \quad \text{for } \boldsymbol{z}\in\mathcal{Z},\end{aligned}\tag{A9}$$

to obtain the preliminary estimators $\widetilde{\boldsymbol{\gamma}}^{[\text{-}k]}$ and $\widetilde{r}^{[\text{-}k]}(\cdot)$, where $\boldsymbol{x} \in \mathcal{X} \cap \{\boldsymbol{z}\}$ represents the set of $\boldsymbol{X}$ belonging to its domain $\mathcal{X}$ and satisfying $\boldsymbol{Z} = \boldsymbol{z}$ for the fixed $\boldsymbol{z}$. To solve (A9) numerically, we adopt a Monte Carlo procedure introduced as follows. Let M be some prespecified number much larger than n, says $100n$. For each $i \in \mathcal{I}^{[\text{-}k]}$, sample $\boldsymbol{X}_{i,1}$, $\boldsymbol{X}_{i,2}$,..., $\boldsymbol{X}_{i,M}$ independently from the estimated $\widehat{\mathcal{P}}^{\mathcal{L}}(\boldsymbol{X}_i \mid \boldsymbol{Z}_i, S_i = 1)$ given $\boldsymbol{Z}_{i,m} = \boldsymbol{Z}_i$ for each $m \in \{1, \ldots, M\}$. Then solve the estimating equation:

$$\frac{K}{nM(K-1)}\sum_{i\in\mathcal{I}_{\text{-}k}}\sum_{m=1}^{M}\mathbf{\Phi}_{i,m}\left\{\widehat{\mathcal{E}}^{\mathcal{L}}[Y_{i,m}\mid \boldsymbol{X}_{i,m}] - g_m(\mathbf{\Phi}_{i,m}^{\intercal}\boldsymbol{\gamma}+r_i)\right\}=\mathbf{0},$$

$$\frac{1}{M}\sum_{m=1}^{M}\widehat{\mathcal{E}}^{\mathcal{L}}[Y_{i,m}\mid \boldsymbol{X}_{i,m}] - g_m(\mathbf{\Phi}_{i,m}^{\intercal}\boldsymbol{\gamma}+r_i)=\mathbf{0}, \quad \text{for } i \in \mathcal{I}^{[\text{-}k]},$$

to obtain the estimators $\widetilde{\boldsymbol{\gamma}}^{[\text{-}k]}$ and $\widetilde{r}_i$, and set $\widetilde{r}^{[\text{-}k]}(\boldsymbol{Z}_i) = \widetilde{r}_i$ for each $i \in \mathcal{I}^{[\text{-}k]}$. Based on these estimators, we construct the debiased estimator for $\boldsymbol{\gamma}$ generally satisfying Assumption 3(i). In specific, we use $\mathcal{L}$ to obtain the estimators $\widehat{\mathcal{E}}^{\mathcal{L}}[\mathbf{\Phi}\dot{g}_m\{(\widetilde{\boldsymbol{\gamma}}^{[\text{-}k]})^{\intercal}\mathbf{\Phi} + \widetilde{r}^{[\text{-}k]}(\boldsymbol{Z})\}|\boldsymbol{Z}, S = 1]$ and $\widehat{\mathcal{E}}^{\mathcal{L}}[g_m\{(\widetilde{\boldsymbol{\gamma}}^{[\text{-}k]})^{\intercal}\mathbf{\Phi} + \widetilde{r}^{[\text{-}k]}(\boldsymbol{Z})\}|\boldsymbol{Z}, S = 1]$. Then we let

$$\widetilde{\boldsymbol{\delta}}_i = (\widetilde{\delta}_{i1}, \ldots, \widetilde{\delta}_{ip_{\mathbf{\Phi}}})^{\intercal} = \mathbf{\Phi}_i - \frac{\widehat{\mathcal{E}}^{\mathcal{L}}[\mathbf{\Phi}_i\dot{g}_m\{(\widetilde{\boldsymbol{\gamma}}^{[\text{-}k]})^{\intercal}\mathbf{\Phi}_i + \widetilde{r}^{[\text{-}k]}(\boldsymbol{Z}_i)\}|\boldsymbol{Z}_i, S_i = 1]}{\widehat{\mathcal{E}}^{\mathcal{L}}[g_m\{(\widetilde{\boldsymbol{\gamma}}^{[\text{-}k]})^{\intercal}\mathbf{\Phi}_i + \widetilde{r}^{[\text{-}k]}(\boldsymbol{Z}_i)\}|\boldsymbol{Z}_i, S_i = 1]},$$

solve

$$\widetilde{\mathbf{w}}_j^{[\text{-}k]} = \min_{\mathbf{w}} \frac{K}{n(K-1)}\sum_{i\in\mathcal{I}_{\text{-}k}}\dot{g}_m\{(\widetilde{\boldsymbol{\gamma}}^{[\text{-}k]})^{\intercal}\mathbf{\Phi}_i + \widetilde{r}^{[\text{-}k]}(\boldsymbol{Z}_i)\}\left(\widetilde{\delta}_{ij} - \mathbf{w}^{\intercal}\widetilde{\boldsymbol{\delta}}_{i,\text{-}j}\right)^2,$$

for each $j \in \{1, \ldots, p_{\mathbf{\Phi}}\}$, and let $\widetilde{\boldsymbol{\varepsilon}}_i = (\widetilde{\epsilon}_{i1}, \ldots, \widetilde{\epsilon}_{ip_{\mathbf{\Phi}}})^{\intercal}$, where $\widetilde{\epsilon}_{ij} = \widetilde{\delta}_{ij} - (\widetilde{\mathbf{w}}_j^{[\text{-}k]})^{\intercal}\widetilde{\boldsymbol{\delta}}_{i,\text{-}j}$, and

$$\widetilde{\sigma}_j^2 = \frac{K}{n(K-1)}\sum_{i\in\mathcal{I}_{\text{-}k}}\widetilde{\epsilon}_{ij}^2\dot{g}_m\left\{(\widetilde{\boldsymbol{\gamma}}^{[\text{-}k]})^{\intercal}\mathbf{\Phi}_i + \widetilde{r}^{[\text{-}k]}(\boldsymbol{Z}_i)\right\}.$$

Then we construct the debiased estimator $\widehat{\boldsymbol{\gamma}}^{[\text{-}k]} = (\widehat{\gamma}_1^{[\text{-}k]}, \ldots, \widehat{\gamma}_{p_{\mathbf{\Phi}}}^{[\text{-}k]})^{\intercal}$ through:

$$\widehat{\gamma}_j^{[\text{-}k]} = \widetilde{\gamma}_j^{[\text{-}k]} + \frac{K}{n(K-1)}\sum_{i\in\mathcal{I}_{\text{-}k}}\frac{\widetilde{\epsilon}_{ij}}{\widetilde{\sigma}_j}\left[Y_i - g_m\{(\widetilde{\boldsymbol{\gamma}}^{[\text{-}k]})^{\intercal}\mathbf{\Phi}_i + \widetilde{r}^{[\text{-}k]}(\boldsymbol{Z}_i)\}\right]. \tag{A10}$$

Finally, the calibrated estimator of the nuisance component $r(\cdot)$ is obtained by solving $\widehat{r}_i$ from:

$$\frac{1}{M}\sum_{m=1}^{M}\widetilde{\omega}^{[\text{-}k]}(\boldsymbol{X}_{i,m})\boldsymbol{c}^{\intercal}\widehat{\boldsymbol{J}}_{\widetilde{\boldsymbol{\beta}}^{[\text{-}k]}}^{-1}\boldsymbol{A}_{i,m}\left[\widehat{\mathcal{E}}^{\mathcal{L}}[Y_{i,m}\mid \boldsymbol{X}_{i,m}] - g_m\left\{\mathbf{\Phi}_{i,M}^{\intercal}\widehat{\boldsymbol{\gamma}}^{[\text{-}k]} + r_i\right\}\right] = 0,$$

for each i, and set $\widehat{r}^{[\text{-}k]}(\boldsymbol{Z}_i) = \widehat{r}_i$, where $\widetilde{\boldsymbol{\beta}}^{[\text{-}k]}$ is again solved through:

$$\frac{K}{n(K-1)}\sum_{i\in\mathcal{I}_{\text{-}k}}\widetilde{\omega}^{[\text{-}k]}(\boldsymbol{X}_i)\boldsymbol{A}_i\{Y_i - \widetilde{m}^{[\text{-}k]}(\boldsymbol{X}_i)\} + \frac{1}{N}\sum_{i=n+1}^{N+n}\boldsymbol{A}_i\{\widetilde{m}^{[\text{-}k]}(\boldsymbol{X}_i) - g(\boldsymbol{A}_i^{\intercal}\boldsymbol{\beta})\} = \mathbf{0}.$$

Noting that our above-introduced procedure is applicable to any SNP M-estimation problem, so the preliminary estimator $\widetilde{\omega}^{[\text{-}k]}(\boldsymbol{X}_i)$ and the calibrated estimator for $\boldsymbol{\alpha}$ and $h(\cdot)$ can be obtained in the same way.

Remark A2 *Our construction procedure proposed in this section involves the estimation of the probability density function, which is typically more challenging than purely estimating the conditional mean for a machine learning method. Note that for linear, log-linear, and logistic models, one can avoid estimating probability density function to construct the doubly robust (or DML) estimators; see Dukes and Vansteelandt (2020); Ghosh and Tan (2020); Liu et al. (2021). Thus, when the link function $g(a) = a$, $g(a) = e^a$ or $g(a) = e^a/(1 + e^a)$, our construction actually does not require estimating the probability density function with $\mathcal{L}$.*

At last, we provide discussion and justification towards the $n^{1/2}$-consistency and asymptotic normality of the debiased estimator $\widehat{\boldsymbol{\gamma}}^{[\text{-}k]}$. In specific, we take $\bar{\boldsymbol{\gamma}} = \boldsymbol{\gamma}^*$, and write (A10) as:

$$\widehat{\gamma}_j^{[\text{-}k]} =\widetilde{\gamma}_j^{[\text{-}k]} + \frac{K}{n(K-1)} \sum_{i\in\mathcal{I}_{\text{-}k}} \frac{\widetilde{\epsilon}_{ij}}{\widetilde{\sigma}_j} \Bigg[Y_i - \mathbb{E}_1[Y_i \mid \boldsymbol{X}_i] + \mathbb{E}_1[Y_i \mid \boldsymbol{X}_i] - g_m\{(\boldsymbol{\gamma}^*)^\intercal \boldsymbol{\Phi}_i + r^*(\boldsymbol{Z}_i)\} \\ + g_m\{\bar{\boldsymbol{\gamma}}^\intercal \boldsymbol{\Phi}_i + r^*(\boldsymbol{Z}_i)\} - g_m\{(\widetilde{\boldsymbol{\gamma}}^{[\text{-}k]})^\intercal \boldsymbol{\Phi}_i + \widetilde{r}^{[\text{-}k]}(\boldsymbol{Z}_i)\}\Bigg].$$

Note that $Y_i - \mathbb{E}_1[Y_i \mid \boldsymbol{X}_i]$ is orthogonal to $\widetilde{\epsilon}_{ij}$ and its estimation error since the latter is deterministic on $\boldsymbol{X}_i$. According to our moment equation for $\boldsymbol{\gamma}^*$ and $r^*(\cdot)$, $\mathbb{E}_1[Y_i \mid \boldsymbol{X}_i] - g_m\{(\boldsymbol{\gamma}^*)^\intercal \boldsymbol{\Phi}_i + r^*(\boldsymbol{Z}_i)\}$ is orthogonal to arbitrary (regular) function of $\boldsymbol{Z}_i$ and linear function of $\boldsymbol{\Phi}_i$, so is also orthogonal to $\widetilde{\epsilon}_{ij}$ and its estimation error. In addition, through our construction,

$$\mathbb{E}_1\left(\boldsymbol{\Phi}_i - \frac{\mathbb{E}_1[\boldsymbol{\Phi}_i \dot{g}_m\{(\boldsymbol{\gamma}^*)^\intercal \boldsymbol{\Phi}_i + r^*(\boldsymbol{Z}_i)\} \mid \boldsymbol{Z}_i]}{\mathbb{E}_1[\dot{g}_m\{(\boldsymbol{\gamma}^*)^\intercal \boldsymbol{\Phi}_i + r^*(\boldsymbol{Z}_i)\} \mid \boldsymbol{Z}_i]}\right) = \boldsymbol{0},$$

and $\widetilde{\epsilon}_{ij}$ is orthogonal to any linear function of $\boldsymbol{\Phi}_{i,\text{-}j}$ and $\boldsymbol{\delta}_{i,\text{-}j}$. So the first order error in $g_m\{\bar{\boldsymbol{\gamma}}^\intercal \boldsymbol{\Phi}_i + r^*(\boldsymbol{Z}_i)\} - g_m\{(\widetilde{\boldsymbol{\gamma}}^{[\text{-}k]})^\intercal \boldsymbol{\Phi}_i + \widetilde{r}^{[\text{-}k]}(\boldsymbol{Z}_i)\}$, i.e. $\dot{g}_m\{\bar{\boldsymbol{\gamma}}^\intercal \boldsymbol{\Phi}_i + r^*(\boldsymbol{Z}_i)\}\{(\widetilde{\boldsymbol{\gamma}}^{[\text{-}k]} - \bar{\boldsymbol{\gamma}})^\intercal \boldsymbol{\Phi}_i + r^*(\boldsymbol{Z}_i) - \widetilde{r}^{[\text{-}k]}(\boldsymbol{Z}_i)\}$, is orthogonal to $\widetilde{\epsilon}_{ij}$ for each j. Thus, all the first-order error terms in $\widehat{\gamma}_j^{[\text{-}k]} - \bar{\boldsymbol{\gamma}}$ could be removed through our Neyman orthogonal construction.

Inspired by existing DML literature like Chernozhukov et al. (2018b) and Liu et al. (2021), when the mean squared error of machine learning algorithm $\mathcal{L}$ has the convergence rates $o_p(n^{-1/2})$ with respect to all the learning objectives included in this section, i.e. the rate double robustness property, the machine learning estimator $\widehat{r}^{[\text{-}k]}(\cdot)$ satisfies Assumption 3(ii). Also, the second order error of $\widehat{\gamma}_j^{[\text{-}k]} - \bar{\boldsymbol{\gamma}}$ could be removed asymptotically. And consequently, $\widehat{\boldsymbol{\gamma}}^{[\text{-}k]}$ satisfy Assumption 3(i). Again, these arguments are applicable to the nuisance estimators for $\boldsymbol{\alpha}$ and $h(\cdot)$ derived in the same way. Therefore, our proposed nuisance estimators introduced in this section tend to satisfy Assumption 3.

C.3 Intrinsic efficient construction

In this section, we introduce the intrinsic efficient construction of the imputation model under our framework. For simplicity, we consider a semi-supervised setting with n labeled source samples and $N \gg n$ unlabeled target samples. The augmentation approach proposed by Shu and Tan (2018) could be used for extending our method to the $N \asymp n$ case. For some given $h(\cdot)$, let the estimating equation of $\widetilde{\boldsymbol{\alpha}}^{[\text{-}k]}$ be

$$\sum_{i\in\{n+1,\ldots,n+N\}\cup\mathcal{I}_{\text{-}k}} \boldsymbol{S}\{\delta_i, \boldsymbol{X}_i; \boldsymbol{\alpha}, h(\cdot)\} = \boldsymbol{0},$$

with $\boldsymbol{S}\{\delta_i, \boldsymbol{X}_i; \boldsymbol{\alpha}, h(\cdot)\}$ representing the score function. For example, one can take

$$\boldsymbol{S}\{\delta_i, \boldsymbol{X}_i; \boldsymbol{\alpha}, h(\cdot)\} = \delta_i \exp\{\boldsymbol{\Psi}_i^{\mathsf{T}}\boldsymbol{\alpha} + h(\boldsymbol{Z}_i)\}\boldsymbol{\Psi}_i - |\mathcal{I}_{\text{-}k}|(1-\delta_i)\boldsymbol{\Psi}_i/N.$$

Denote that $\boldsymbol{S}_i = \boldsymbol{S}\{\delta_i, \boldsymbol{X}_i; \widetilde{\boldsymbol{\alpha}}^{[\text{-}k]}, \widetilde{h}^{[\text{-}k]}(\cdot)\}$ and let $\boldsymbol{\Pi}_{\mathcal{I}_{\text{-}k}}(\epsilon_i; \boldsymbol{S}_i)$ be the empirical projection operator of any variable ϵ_i to the space spanned by $\boldsymbol{S}_i$ on the samples $\mathcal{I}_{\text{-}k}$ and $\boldsymbol{\Pi}^{\perp}_{\mathcal{I}_{\text{-}k}}(\epsilon_i; \boldsymbol{S}_i) = \epsilon_i - \boldsymbol{\Pi}_{\mathcal{I}_{\text{-}k}}(\epsilon_i; \boldsymbol{S}_i)$. When the importance weight model is correctly specified and $N \gg n$, the empirical asymptotic variance for $\boldsymbol{c}^{\mathsf{T}}\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}}$ with nuisance parameters $\boldsymbol{\gamma}$ and $r(\cdot)$ can be expressed as

$$\frac{K}{n(K-1)}\sum_{i\in\mathcal{I}_{\text{-}k}}\left[\widetilde{\omega}^{[\text{-}k]}(\boldsymbol{X}_i)\boldsymbol{\Pi}^{\perp}_{\mathcal{I}_{\text{-}k}}\left(\boldsymbol{c}^{\mathsf{T}}\widehat{\boldsymbol{J}}^{-1}_{\widetilde{\boldsymbol{\beta}}^{[\text{-}k]}}\boldsymbol{A}_i[Y_i - g_m\{\boldsymbol{\Phi}_i^{\mathsf{T}}\boldsymbol{\gamma} + r(\boldsymbol{Z}_i)\}]; \boldsymbol{S}_i\right)\right]^2. \tag{A11}$$

Then the intrinsically efficient construction of the imputation model is given by minimizing (A11) subject to the moment constraint:

$$\frac{1}{|\mathcal{I}_{\text{-}k}\cap\mathcal{I}^a|}\sum_{i\in\mathcal{I}_{\text{-}k}\cap\mathcal{I}^a} K_h(\boldsymbol{Z}_i - \boldsymbol{z})\widetilde{\omega}^{[\text{-}k]}(\boldsymbol{X}_i)\boldsymbol{c}^{\mathsf{T}}\widehat{\boldsymbol{J}}^{-1}_{\widetilde{\boldsymbol{\beta}}^{[\text{-}k]}}\boldsymbol{A}_i\left[Y_i - g_m\left\{\boldsymbol{\Phi}_i^{\mathsf{T}}\boldsymbol{\gamma} + r(\boldsymbol{Z})\right\}\right] = 0,$$

which is the same as the first equation of (11) except that both $\boldsymbol{\gamma}$ and $r(\boldsymbol{Z})$ are unknown here. This optimization problem could be solved with methods like profile kernel and backfitting (Lin and Carroll, 2006). Alternatively and more conveniently, one could use the sieve estimation, as discussed in Appendix C.1, to model $r(\boldsymbol{Z}_i)$ and use a constrained least square regression: let $\boldsymbol{b}(\boldsymbol{Z})$ be some basis of $\boldsymbol{z}$ and solve

$$\begin{aligned}
\min_{\boldsymbol{\gamma},\boldsymbol{\xi}} \quad & \sum_{i\in\mathcal{I}_{\text{-}k}}\left[\widetilde{\omega}^{[\text{-}k]}(\boldsymbol{X}_i)\boldsymbol{\Pi}^{\perp}_{\mathcal{I}_{\text{-}k}}\left(\boldsymbol{c}^{\mathsf{T}}\widehat{\boldsymbol{J}}^{-1}_{\widetilde{\boldsymbol{\beta}}^{[\text{-}k]}}\boldsymbol{A}_i[Y_i - g_m\{\boldsymbol{\Phi}_i^{\mathsf{T}}\boldsymbol{\gamma} + \boldsymbol{b}^{\mathsf{T}}(\boldsymbol{Z}_i)\boldsymbol{\xi}\}]; \boldsymbol{S}_i\right)\right]^2;\\
\text{s.t.} \quad & \sum_{i\in\mathcal{I}_{\text{-}k}\cap\mathcal{I}^a}\boldsymbol{b}(\boldsymbol{Z}_i)\widetilde{\omega}^{[\text{-}k]}(\boldsymbol{X}_i)\boldsymbol{c}^{\mathsf{T}}\widehat{\boldsymbol{J}}^{-1}_{\widetilde{\boldsymbol{\beta}}^{[\text{-}k]}}\boldsymbol{A}_i\left[Y_i - g_m\left\{\boldsymbol{\Phi}_i^{\mathsf{T}}\boldsymbol{\gamma} + \boldsymbol{b}^{\mathsf{T}}(\boldsymbol{Z}_i)\boldsymbol{\xi}\right\}\right] = 0,
\end{aligned}$$

to obtain $\widetilde{\boldsymbol{\gamma}}^{[\text{-}k]}$ and $\widetilde{r}^{[\text{-}k]}(\boldsymbol{Z}) = \boldsymbol{b}^{\mathsf{T}}(\boldsymbol{Z})\widetilde{\boldsymbol{\xi}}^{[\text{-}k]}$ simultaneously. To get the intrinsic efficient estimator for a nonlinear but differentiable function $\ell(\boldsymbol{\beta}_0)$, with its gradient being $\dot{\ell}(\cdot)$, we first estimate the entries β_{0i} using our proposed method for every $i \in \{1, 2, \ldots, d\}$ and use them to form a preliminary $\sqrt{n}$-consistent estimator $\widehat{\boldsymbol{\beta}}_{(init)}$. Then we estimate the linear function $\boldsymbol{\beta}_0^{\mathsf{T}}\dot{\ell}\{\widehat{\boldsymbol{\beta}}_{(init)}\}$ with the intrinsically efficient estimator and utilize the expansion $\ell(\boldsymbol{\beta}_0) \approx \ell\{\widehat{\boldsymbol{\beta}}_{(init)}\} + \{\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}_{(init)}\}^{\mathsf{T}}\dot{\ell}\{\widehat{\boldsymbol{\beta}}_{(init)}\}$ for an one-step update.

Appendix D. Implementing details and additional results of simulation

We introduce the multiplier bootstrap procedure used to quantify the uncertainty of our estimator and construct confidence intervals for $\boldsymbol{c}^\intercal\boldsymbol{\beta}_0$ in our numerical studies. Let $\boldsymbol{\epsilon} = \{\epsilon_1, \ldots, \epsilon_n, \epsilon_{n+1}, \ldots, \epsilon_{n+N}\}$ be $N+n$ independent $\mathrm{N}(1,1)$ (or $\mathrm{Gamma}(1,1)$) random variables sampling independent of the data. We first solve the re-weighted equations:

$$\frac{K}{n(K-1)}\sum_{i\in\mathcal{I}_{\text{-}k}}\epsilon_i\boldsymbol{\Psi}_i^{\boldsymbol{b}}\exp(\boldsymbol{\theta}_w^\intercal\boldsymbol{\Psi}_i^{\boldsymbol{b}})+\lambda_1(0,\boldsymbol{\theta}_{w,\text{-}1}^\intercal)^\intercal=\frac{1}{N}\sum_{i=n+1}^{n+N}\epsilon_i\boldsymbol{\Psi}_i^{\boldsymbol{b}};\quad\text{with }\boldsymbol{\theta}_w=(\boldsymbol{\alpha}^\intercal,\boldsymbol{\eta}^\intercal)^\intercal$$
$$\frac{K}{n(K-1)}\sum_{i\in\mathcal{I}_{\text{-}k}}\epsilon_i\boldsymbol{\Phi}_i^{\boldsymbol{b}}\left\{Y_i-g_m(\boldsymbol{\theta}_m^\intercal\boldsymbol{\Phi}_i^{\boldsymbol{b}})\right\}+\lambda_2(0,\boldsymbol{\theta}_{m,\text{-}1}^\intercal)^\intercal=\boldsymbol{0},\qquad\text{with }\boldsymbol{\theta}_m=(\boldsymbol{\gamma}^\intercal,\boldsymbol{\xi}^\intercal)^\intercal$$

to obtain the estimators $\widetilde{\boldsymbol{\theta}}_{w,\boldsymbol{\epsilon}}^{[\text{-}k]}=(\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\epsilon}}^{[\text{-}k]\intercal},\widetilde{\boldsymbol{\eta}}_{\boldsymbol{\epsilon}}^{[\text{-}k]\intercal})^\intercal$, $\widetilde{\boldsymbol{\theta}}_{m,\boldsymbol{\epsilon}}^{[\text{-}k]}=(\widetilde{\boldsymbol{\gamma}}_{\boldsymbol{\epsilon}}^{[\text{-}k]\intercal},\widetilde{\boldsymbol{\xi}}_{\boldsymbol{\epsilon}}^{[\text{-}k]\intercal})^\intercal$. Then we take $\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\epsilon}}^{[\text{-}k]}=\widetilde{\boldsymbol{\alpha}}_{\boldsymbol{\epsilon}}^{[\text{-}k]}$, $\widehat{\boldsymbol{\gamma}}_{\boldsymbol{\epsilon}}^{[\text{-}k]}=\widetilde{\boldsymbol{\gamma}}_{\boldsymbol{\epsilon}}^{[\text{-}k]}$,

$$\widehat{\omega}_{\boldsymbol{\epsilon}}^{[\text{-}k]}(\boldsymbol{X}_i)=\exp\{\boldsymbol{\Psi}_i^\intercal\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\epsilon}}^{[\text{-}k]}+\widehat{h}^{[\text{-}k]}(\boldsymbol{Z}_i)\},\quad\widehat{m}_{\boldsymbol{\epsilon}}^{[\text{-}k]}(\boldsymbol{X}_i)=g_m\{\boldsymbol{\Phi}_i^\intercal\widehat{\boldsymbol{\gamma}}_{\boldsymbol{\epsilon}}^{[\text{-}k]}+\widehat{r}^{[\text{-}k]}(\boldsymbol{Z}_i)\},$$

and $\widehat{m}_{\boldsymbol{\epsilon}}(\boldsymbol{X}_i)=K^{-1}\sum_{k=1}^K\widehat{m}_{\boldsymbol{\epsilon}}^{[\text{-}k]}(\boldsymbol{X}_i)$. Note that we do not need to refit and replace the nonparametric components in $\widehat{\omega}$ and $\widehat{m}$ since they do not create first-order impact (on the asymptotic variance of our estimator as shown in Theorem 1).

Based on these, we solve

$$\frac{1}{n}\sum_{k=1}^K\sum_{i\in\mathcal{I}_k}\epsilon_i\widehat{\omega}_{\boldsymbol{\epsilon}}^{[\text{-}k]}(\boldsymbol{X}_i)\boldsymbol{A}_i\left\{Y_i-\widehat{m}_{\boldsymbol{\epsilon}}^{[\text{-}k]}(\boldsymbol{X}_i)\right\}+\frac{1}{N}\sum_{i=n+1}^{N+n}\epsilon_i\boldsymbol{A}_i\{\widehat{m}_{\boldsymbol{\epsilon}}(\boldsymbol{X}_i)-g(\boldsymbol{A}_i^\intercal\boldsymbol{\beta})\}=\boldsymbol{0},$$

to obtain $\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}}^{\boldsymbol{\epsilon}}$ and take $\boldsymbol{c}^\intercal\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}}^{\boldsymbol{\epsilon}}$ as the bootstrap estimator of $\boldsymbol{c}^\intercal\boldsymbol{\beta}_0$. We repeat sampling $\boldsymbol{\epsilon}$ and computing $\boldsymbol{c}^\intercal\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}}^{\boldsymbol{\epsilon}}$, and use the standard deviation of the samplers $\boldsymbol{c}^\intercal\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}}^{\boldsymbol{\epsilon}}$ to estimate the standard error of $\boldsymbol{c}^\intercal\widehat{\boldsymbol{\beta}}_{\mathsf{ATReL}}$, which is consistent by the standard bootstrap theory.

To obtain the preliminary estimators $\widetilde{\omega}^{[\text{-}k]}(\cdot)$ and $\widetilde{m}^{[\text{-}k]}(\cdot)$ of our method, we use semiparametric logistic regression with covariates including the parametric basis and the natural splines of the nonparametric components Z with order $[n^{1/4}]$ for the imputation model and $[(N+n)^{1/4}]$ for the importance weight model. In this process, we add ridge penalty tuned by cross-validation with tuning parameter of order $n^{-2/3}$ (below the parametric rate) to enhance the training stability.

We set the loading vector $\boldsymbol{c}$ as $(1,0,0,0)^\intercal$, $(0,1,0,0)^\intercal$, $(0,0,1,0)^\intercal$, and $(0,0,0,1)^\intercal$ to estimate $\beta_0,\beta_1,\beta_2,\beta_3$ separately. For β_1,β_2,β_3, the weights $\boldsymbol{c}^\intercal\widehat{\boldsymbol{J}}_{\widetilde{\boldsymbol{\beta}}^{[\text{-}k]}}^{-1}\boldsymbol{A}_i$'s are not positive definite so we split the source and target samples as $\mathcal{I}^+=\{i:\boldsymbol{c}^\intercal\widehat{\boldsymbol{J}}_{\widetilde{\boldsymbol{\beta}}^{[\text{-}k]}}^{-1}\boldsymbol{A}_i\geq 0\}$ and $\mathcal{I}^-=\{i:\boldsymbol{c}^\intercal\widehat{\boldsymbol{J}}_{\widetilde{\boldsymbol{\beta}}^{[\text{-}k]}}^{-1}\boldsymbol{A}_i<0\}$ as introduced in Remark 5, and use (12) to estimate their nonparametric components. For β_0, we find that $\boldsymbol{c}^\intercal\widehat{\boldsymbol{J}}_{\widetilde{\boldsymbol{\beta}}^{[\text{-}k]}}^{-1}\boldsymbol{A}_i$ is nearly positive definite under all configurations but these weights are sometimes of high variation. So we also split the source/target samples by cutting the $\boldsymbol{c}^\intercal\widehat{\boldsymbol{J}}_{\widetilde{\boldsymbol{\beta}}^{[\text{-}k]}}^{-1}\boldsymbol{A}_i$'s with their median, to reduce the variance of weights at each fold and improve the effective sample size. We use cross-fitting with

$K = 5$ folds for our method and the two DML estimators. And all the tuning parameters including the bandwidth of our method and kernel machine and the coefficients of the penalty functions are selected by 5-folded cross-validation on the training samples. We present the estimation performance (mean square error, bias, and coverage probability) on each parameter in Tables A2–A5, for the four configurations separately.

At last, we carry out an additional simulation study to demonstrate the limitation of our method, i.e., its failure when both nuisance models are severely misspecified. We generate $\boldsymbol{X} = \widetilde{\boldsymbol{X}}$ and $\boldsymbol{W}$ following the same way as Configuration (iii) in Section 4, and take

$$\begin{aligned}\mathrm{P}(S = 1 \mid \boldsymbol{X}) =& g_m\{\mathbf{a}_w^{\intercal}\boldsymbol{W} + 2h_x(X_1) + 2h_x(X_2)\};\\ \mathrm{P}(Y = 1 \mid \boldsymbol{X}) =& g_m\{\mathbf{b}_w^{\intercal}\boldsymbol{W} + 2r_x(X_1) + 2r_x(X_2)\},\end{aligned}$$

with h_x and r_x set to be the same as Configuration (iii), $\mathbf{a}_w = (0.5, 0.5, 0.5, 0.3, 0.3, 0.2, 0.2)^{\intercal}$ and $\mathbf{b}_w = (-0.5, 0.5, 0.8, 0.3, -0.3, -0.2, 0.15, 0.15)^{\intercal}$. Implementation setups of the methods are also the same as in Section 4. Different from Configurations (i)–(iv) in Section 4, this data-generation mechanism violates the SNP model assumption on both nuisance models in ATReL, with non-linear effects from $\boldsymbol{X}$ and strong covariate shift bias caused by $2h_x(X_1) + 2h_x(X_2)$ and $2r_x(X_1) + 2r_x(X_2)$. This confoundedness coming from X_1 and X_2 could not be properly removed by our method since we set $Z = X_1$ only.

The RMSE and bias are presented in Table A1. When the non-linear terms have a strong impact and both the SNP nuisance models are wrong, both Parametric and ATReL fail to output a reasonable estimator. For example, the bias of our estimator on β_2 is as high as 1.3, almost equaling its RMSE, which means the asymptotic normality and unbiasedness property in Theorem 1 does not hold in this case. This is not unexpected considering the severe violation of our model assumption. Meanwhile, the DML estimator constructed with the fully nonparametric kernel machine shows better performance, e.g., a 0.05 bias on β_2 that is also much smaller compared to its RMSE.

Table A1: Estimation performance under the simulation setting described in Appendix D. Parametric: doubly robust estimator with parametric nuisance models; ATReL: our method; $\mathrm{DML}_{\mathsf{BE}}$: DML with flexible basis expansions; $\mathrm{DML}_{\mathsf{KM}}$: DML with kernel machine.

		Parametric	ATReL	$\mathrm{DML}_{\mathsf{BE}}$	$\mathrm{DML}_{\mathsf{KM}}$
β_0	RMSE	0.132	0.122	0.251	0.290
	Bias	0.078	0.059	0.201	−0.044
β_1	RMSE	0.576	0.215	0.195	0.351
	Bias	0.563	0.162	0.066	0.222
β_2	RMSE	1.304	1.445	0.532	0.359
	Bias	1.288	1.434	0.475	0.054
β_3	RMSE	0.108	0.111	0.162	0.264
	Bias	−0.003	−0.016	0.010	0.003

Table A2: Estimation performance of the methods on parameters $\beta_0, \beta_1, \beta_2, \beta_3$ under Configuration (i) described in Section 4. Parametric: doubly robust estimator with parametric nuisance models; ATReL: our proposed doubly robust estimator using SNP nuisance models; DML_{BE}: double machine learning with flexible basis expansions; DML_{KM}: double machine learning with kernel machine. RMSE: root mean square error; CP: coverage probability of the 95% confidence interval.

		Estimator			
Covariates		Parametric	ATReL	DML_{BE}	DML_{KM}
β_0	RMSE	0.102	0.110	0.168	0.116
	Bias	-0.007	0.0005	0.112	0.010
	CP	0.95	0.95	0.84	0.93
β_1	RMSE	0.181	0.124	0.160	0.198
	Bias	-0.146	-0.056	-0.104	-0.163
	CP	0.91	0.93	0.92	0.85
β_2	RMSE	0.133	0.126	0.191	0.134
	Bias	0.059	0.032	-0.109	-0.017
	CP	0.99	0.97	0.94	0.98
β_3	RMSE	0.137	0.133	0.195	0.150
	Bias	0.049	0.030	-0.108	-0.040
	CP	0.99	0.97	0.96	0.97

Table A3: Estimation performance of the methods on parameters $\beta_0, \beta_1, \beta_2, \beta_3$ under Configuration (ii) described in Section 4. Parametric: doubly robust estimator with parametric nuisance models; ATReL: our proposed doubly robust estimator using SNP nuisance models; DML_{BE}: double machine learning with flexible basis expansions; DML_{KM}: double machine learning with kernel machine. RMSE: root mean square error; CP: coverage probability of the 95% confidence interval.

		Estimator			
Covariates		Parametric	ATReL	DML_{BE}	DML_{KM}
β_0	RMSE	0.108	0.114	0.186	0.124
	Bias	−0.004	0.004	0.136	0.018
	CP	0.92	0.94	0.82	0.90
β_1	RMSE	0.107	0.118	0.144	0.122
	Bias	−0.001	−0.015	−0.062	−0.046
	CP	0.99	0.95	0.95	0.98
β_2	RMSE	0.129	0.131	0.209	0.166
	Bias	−0.006	−0.024	−0.136	−0.084
	CP	0.98	0.96	0.94	0.95
β_3	RMSE	0.124	0.128	0.200	0.171
	Bias	−0.008	−0.019	−0.123	−0.097
	CP	0.98	0.97	0.94	0.96

Table A4: Estimation performance of the methods on parameters $\beta_0, \beta_1, \beta_2, \beta_3$ under Configuration (iii) described in Section 4. Parametric: doubly robust estimator with parametric nuisance models; ATReL: our proposed doubly robust estimator using SNP nuisance models; $\text{DML}_{\textsf{BE}}$: double machine learning with flexible basis expansions; $\text{DML}_{\textsf{KM}}$: double machine learning with kernel machine. RMSE: root mean square error; CP: coverage probability of the 95% confidence interval.

		Estimator			
Covariates		Parametric	ATReL	$\text{DML}_{\textsf{BE}}$	$\text{DML}_{\textsf{KM}}$
β_0	RMSE	0.113	0.112	0.134	0.114
	Bias	−0.052	−0.014	−0.064	−0.026
	CP	0.93	0.95	0.93	0.95
β_1	RMSE	0.341	0.151	0.152	0.189
	Bias	−0.300	−0.047	−0.043	−0.135
	CP	0.82	0.93	0.95	0.86
β_2	RMSE	0.145	0.133	0.141	0.133
	Bias	−0.006	−0.011	−0.035	−0.054
	CP	0.95	0.94	0.95	0.91
β_3	RMSE	0.143	0.137	0.139	0.131
	Bias	−0.008	0.004	0.003	−0.033
	CP	0.94	0.95	0.95	0.91

Table A5: Estimation performance of the methods on parameters $\beta_0, \beta_1, \beta_2, \beta_3$ under Configuration (iv) described in Section 4. Parametric: doubly robust estimator with parametric nuisance models; ATReL: our proposed doubly robust estimator using SNP nuisance models; DML_{BE}: double machine learning with flexible basis expansions; DML_{KM}: double machine learning with kernel machine. RMSE: root mean square error; CP: coverage probability of the 95% confidence interval.

		Estimator			
Covariates		Parametric	ATReL	DML_{BE}	DML_{KM}
β_0	RMSE	0.103	0.107	0.189	0.109
	Bias	-0.003	0.010	0.151	0.027
	CP	0.95	0.95	0.73	0.95
β_1	RMSE	0.140	0.128	0.132	0.156
	Bias	-0.008	0.008	0.035	0.100
	CP	0.94	0.93	0.94	0.86
β_2	RMSE	0.137	0.126	0.127	0.121
	Bias	-0.004	-0.004	-0.025	0.000
	CP	0.96	0.96	0.95	0.90
β_3	RMSE	0.139	0.126	0.121	0.122
	Bias	0.005	0.015	0.022	0.050
	CP	0.95	0.97	0.96	0.93

Appendix E. Implementing details and additional results of real example

The specific nuisance model constructions are described as follows.

Method	Importance weighting	Imputation
Parametric	Logistic model with features $(\boldsymbol{X}^{\intercal}, X_1X_2, X_1X_3, X_2X_3)^{\intercal}$.	Logistic model with features taken as $\boldsymbol{X}$.
ATReL	Logistic model with $\boldsymbol{\Psi} = (\boldsymbol{X}_{\text{-}2}^{\intercal}, X_1X_2, X_1X_3, X_2X_3)^{\intercal}$ and set $Z = X_2$ for nonparametric modeling	Logistic model with $\boldsymbol{\Phi} = \boldsymbol{X}_{\text{-}2}$ and set $Z = X_2$ for nonparametric modeling
DML with flexible basis expansions	$\ell_1 + \ell_2$ regularized regression including basis terms: $\boldsymbol{X}$, natural splines of X_1, X_2 and X_6 of order 5 and interaction terms of these natural splines	$\ell_1 + \ell_2$ regularized regression including basis terms: $\boldsymbol{X}$, natural splines of X_1, X_2 and X_6 of order 5 and interaction terms of these natural splines
DML with kernel machine	Support vector machine with the radial basis function kernel	Support vector machine with the radial basis function kernel

We present the fitted coefficients of all the included approaches in Table A6.

Table A6: Estimators of the target model coefficients. $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ represent respectively the intercept, coefficient of the total healthcare utilization (X_1), coefficient of the log(NLP+1) of RA (X_2), coefficient of the indicator for NLP mention of tumor necrosis factor (TNF) inhibitor (X_3), and coefficient of the indicator for NLP mention of bone erosion (X_4). Parametric: doubly robust estimator with parametric nuisance models; ATReL: our proposed doubly robust estimator using SNP nuisance models; $\text{DML}_{\textsf{BE}}$: double machine learning with flexible basis expansions; $\text{DML}_{\textsf{KM}}$: double machine learning with kernel machine.

	Source	Parametric	ATReL	$\text{DML}_{\textsf{BE}}$	$\text{DML}_{\textsf{KM}}$	Target
β_0	-5.70	-5.08	-5.75	-8.88	-5.73	-5.03
β_1	0.03	0.12	-0.19	0.01	0.05	-0.31
β_2	1.73	1.39	1.56	2.64	1.61	1.35
β_3	0.69	0.62	0.78	0.77	0.66	0.94
β_4	0.60	0.62	0.44	0.62	0.35	0.14