Statistica Sinica **34** (2024), 459-479
doi:https://doi.org/10.5705/ss.202022.0089

# MODEL-ASSISTED INFERENCE FOR COVARIATE-SPECIFIC TREATMENT EFFECTS WITH HIGH-DIMENSIONAL DATA

Peng Wu[1], Zhiqiang Tan[2], Wenjie Hu[3] and Xiao-Hua Zhou*[3,4]

[1]*Beijing Technology and Business University,* [2]*Rutgers University,*
[3]*Peking University and* [4]*Pazhou Lab*

*Abstract:* Covariate-specific treatment effects (CSTEs) are heterogeneous treatment effects across subpopulations defined by certain selected covariates. In this study, we consider marginal structural models in which CSTEs are represented linearly using a set of basis functions of the selected covariates. We develop a new approach for high-dimensional settings to obtain not only doubly robust point estimators of CSTEs, but also model-assisted confidence intervals, which are valid when the propensity score model is specified correctly, but the outcome regression model may be misspecified. With a linear outcome model and subpopulations defined by discrete covariates, both the point estimators and the confidence intervals are doubly robust for CSTEs. In contrast, the confidence intervals from existing high-dimensional methods are valid only when both the propensity score and the outcome models are specified correctly. We also establish several asymptotic properties of the proposed point estimators and the associated confidence intervals. The results of our simulation studies demonstrate the advantages of the proposed method over existing methods. Lastly, we apply the proposed method to a large clinical data set on psoriasis from a national registry in China, the Psoriasis Center Data Platform, to explore the effects of biologics versus those of conventional therapies across different subpopulations.

*Key words and phrases:* Covariate-specific treatment effect, doubly robust confidence interval, doubly robust point estimator, high-dimensional data, model-assisted confidence interval.

## 1. Introduction

When analyzing the causal effect of an intervention, the average treatment effect (ATE) is often used as the estimand of interest for simplicity and interpretability. However, researchers and policymakers may also be interested in the effects of treatments (or policies) at various subpopulation levels (Lee, Okui and Whang (2017); Chernozhukov, Fernández-Val and Luo (2018); Semenova and Chernozhukov (2021)). Specifically, let $Y$ be an outcome variable, $T$ be a treatment variable taking values in $\{0, 1\}$, and $Z$ be the covariates used to define the subpopulations. Define $(Y^0, Y^1)$ as the potential outcomes under treatment

*Corresponding author.

arms zero and one, respectively. Here, we focus on the covariate-specific treatment effect (CSTE) $\tau(z)$, defined by $E(Y^1 - Y^0 \mid Z = z)$ for possible values $z$ of $Z$. For example, in our empirical application, we study the effects of biologics versus conventional therapies on psoriasis area and severity (PASI) improvement in subpopulations defined by a patient's age, baseline dermatology life quality index (DLQI), and baseline PASI. CSTEs are also useful in precision medicine for determining an optimal treatment regime based on an individual's characteristics (Chakraborty and Moodie (2013)).

Observational studies often include a large set of covariates, possibly with nonlinear and interaction terms, to reduce confounding bias and enhance the credibility of causal inference. Thus, we introduce auxiliary covariates $V$, allowing $V$ to be high-dimensional, and posit that the unconfoundedness holds, conditioning on all covariates $X \equiv (Z, V)$ to identify the CSTEs.

The CSTE $\tau(z)$ is, in general, different from $\tau(x) \equiv E(Y^1 - Y^0 \mid X = x)$, the conditional treatment effect given the full covariates. By conditioning on a low-dimensional covariate, $\tau(z)$ is easier to interpret in practice. Moreover, estimating $\tau(z)$ is more manageable and less affected by modeling assumptions in statistical analysis. It is difficult to obtain asymptotic normality and valid confidence intervals (CIs) for $\tau(x)$, owing to the high dimensionality of $X$, unless some restrictive assumptions are imposed (Tian et al. (2014); Dukes and Vansteelandt (2020); Guo, Zhou and Ma (2021)).

There is increasing interest in estimating CSTEs. Abrevaya, Hus and Lieli (2015) derive an inverse probability weighting (IPW) estimator of $\tau(z)$ using kernel smoothing with continuous $Z$, Lee, Okui and Whang (2017) proposes an augmented IPW (AIPW) estimator based on kernel smoothing, and Lechner (2019) proposes algorithms for constructing causal random forests. The aforementioned approaches estimate $\tau(z)$ in low-dimensional settings. Fan et al. (2021), Zimmert and Lechner (2019), and Semenova and Chernozhukov (2021) extend the method of Lee, Okui and Whang (2017) to high-dimensional settings, using machine learning algorithms to mitigate the model specification for nuisance parameters (propensity score (PS) and outcome regression (OR) models), and a sample-splitting (or cross-fitting) technique to reduce the effect of having to estimate the nuisance parameters on the resulting estimator of $\tau(z)$. A limitation of these existing high-dimensional methods is that the CIs are valid only when both the PS and the OR models are specified correctly. This is because the Neyman orthogonality condition (Chernozhukov et al. (2018)) cannot ensure the negligibility of the first-order approximation error of $\tau(z)$ when only one of the models is specified correctly. Further discussion is provided in Section 2.2.

We consider three properties: (a) the point estimator is doubly robust, which is consistent if either the PS model or the OR model is specified correctly; (b) the CIs are valid if the PS model is specified correctly, but the OR model may be misspecified; and (c) the CIs are valid if the OR model is specified correctly,

but the PS model may be misspecified. If either property (b) or (c) is met, then we have model-assisted CIs (Tan (2020a)). If properties (b) and (c) are satisfied, then the CIs are doubly robust. We develop a new approach for CSTEs in high-dimensional settings that possesses properties (a) and (b) for continuous $Z$. Furthermore, with a linear OR model and discrete $Z$, the proposed method possesses all the properties (a), (b), and (c). To the best of our knowledge, no existing methods for estimating CSTEs possess model-assisted or doubly robust CIs, while retaining the double robustness of the point estimator.

Our proposed method is motivated by the work of Tan (2020a), who was the first to estimate ATEs and ATEs on treated. The work was recently extended to estimate local ATEs in high-dimensional settings (Sun and Tan (2021)). Here, we extend the method further to tackle the estimation of CSTEs. When $Z$ is discrete with finite support, the proposed method is closely related to the stratified analysis of Tan (2020a), which first splits the sample using $Z$, and then applies the author's method to estimate ATE separately within each stratum. However, a stratified analysis is troublesome if $Z$ takes many possible categories and we have a high-dimensional auxiliary covariate vector $V$, where different tuning parameters need to be selected using separate cross-validations. In contrast, the proposed method does not split the sample and is numerically more tractable, with only two lasso tuning parameters for the PS and OR models. See Section 3.3 for further discussion. For continuous $Z$, a direct extension of Tan's approach cannot guarantee the model-assisted property. However, our approach does provide model-assisted CIs.

The proposed method relies on similar sparsity conditions to those in Tan (2020a). For example, for a logistic PS model and a linear OR model with coefficients $\gamma$ and $\alpha_1$, respectively, suppose the estimators $\hat{\gamma}$ and $\hat{\alpha}_1$ converge to the target values of $\bar{\gamma}$ and $\bar{\alpha}_1$. With possible model misspecification, the point estimator is doubly robust provided that $(|S_{\bar{\gamma}}| + |S_{\bar{\alpha}_1}|)\log(p) = o(n)$, and the CIs are model assisted for continuous $Z$ and doubly robust for discrete $Z$, provided that $(|S_{\bar{\gamma}}| + |S_{\bar{\alpha}_1}|)\log(p) = o(n^{1/2})$. The sparsity requirements are comparable with those in Belloni, Chernozhukov and Hansen (2014); Belloni et al. (2017), Farrell (2015), and Chernozhukov et al. (2018) for ATEs based on commonly penalized PS and OR models, and those in Athey, Imbens and Wager (2018), Bradic, Wager and Zhu (2019), Ning, Peng and Imai (2020), and Wang and Shah (2020), which allow for model misspecification. See Smucler, Rotnitzky and Robins (2019) for a discussion on model doubly robust and rate doubly robust estimations in high-dimensional settings.

The remainder of this paper is structured as follows. In Section 2, we state the problem and discuss some existing methods for solving it. Section 3 presents our estimation procedures, and Section 4.2 provides the asymptotic results and discusses why the proposed methods work. In Section 5, we use extensive simulations to evaluate the finite sample performance of the proposed

methods. Section 6 illustrates how to use our methods by examining an empirical example. A brief discussion is presented in Section 7.

## 2. Background

### 2.1. Setup

Suppose that $\{(Y_i, T_i, X_i) : i = 1, \ldots, n\}$ is an independent and identically distributed (i.i.d.) sample of $n$ observations, where $Y$ is an outcome variable, $T$ is a binary treatment variable, and $X = (V^T, Z^T)^T$ is a vector of measured covariates, where $Z$ denotes the covariates used to define the subpopulations, and $V$ denotes the auxiliary covariates. In the potential outcomes framework (Rubin (1974); Neyman (1990)), let $(Y^0, Y^1)$ be the potential outcomes under treatment arms zero and one, respectively. By the consistency assumption, $Y = (1 - T)Y^0 + TY^1$. The causal parameter of interest is the CSTE defined by $\tau(z) = E(Y^1 - Y^0 \mid Z = z) = \mu^1(z) - \mu^0(z)$, with $\mu^t(z) = E(Y^t \mid Z = z)$ for $t = 0, 1$. For identification, we impose the following Assumption 1.

**Assumption 1.** *$T \perp Y^0 \mid X$ and $T \perp Y^1 \mid X$ (Rubin (1976)); $0 < \pi^*(x) < 1$, for all $x$, where $\pi^*(x) = P(T = 1 \mid X = x)$ is the PS (Rosenbaum and Rubin (1983)).*

Under Assumption 1, letting $m_t^*(X) = E(Y \mid T = t, X)$, we have

$$\mu^1(z) = E\left[\frac{TY}{\pi^*(X)} - \left(\frac{T}{\pi^*(X)} - 1\right)m_1^*(X) \mid Z = z\right]. \tag{2.1}$$

Similar equations can be derived for $\mu^0(z)$ and $\tau(z)$. Then, $(\mu^0(z), \mu^1(z))$ and $\tau(z)$ can be estimated by imposing additional modeling assumptions on the OR function $m_t^*(X)$ or the PS $\pi^*(X)$. Here, we estimate $\mu^1(z)$; for a discussion on $\mu^0(z)$ and $\tau(z)$, see the Supplementary Material.

### 2.2. Existing doubly robust estimators and their limitations

Consider a conditional mean working model for the OR in the treated group,

$$E(Y \mid T = 1, X) = m_1(X; \alpha_1) = \psi\{\alpha_1^T g(X)\}, \tag{2.2}$$

and a logistic regression working model for the PS,

$$P(T = 1 \mid X) = \pi(X; \gamma) = [1 + \exp\{-\gamma^T f(X)\}]^{-1}, \tag{2.3}$$

where $g(X) = \{1, g_1(X), \ldots, g_q(X)\}^T$ and $f(X) = \{1, f_1(X), \ldots, f_p(X)\}^T$ are vectors of known functions, and $\psi(\cdot)$ is a known inverse link function. In high-dimensional settings, let $\hat{\alpha}_{1,RML}$ and $\hat{\gamma}_{RML}$ be lasso-regularized maximum likelihood estimators (Tibshirani (1996)) of $\alpha_1$ and $\gamma$, respectively. Denote $\hat{m}_{1,RML}(X) = m_1(X; \hat{\alpha}_{1,RML})$ and $\hat{\pi}_{RML}(X) = \pi(X; \hat{\gamma}_{RML})$. Let

$$\varphi(Y_i, T_i, X_i; \hat{m}_{1,RML}, \hat{\pi}_{RML}) = \frac{T_i Y_i}{\hat{\pi}_{RML}(X_i)} - \left(\frac{T_i}{\hat{\pi}_{RML}(X_i)} - 1\right)\hat{m}_{1,RML}(X_i). \quad (2.4)$$

Equation (2.1) implies that the doubly robust AIPW estimator of $\mu^1(z)$ can be obtained by regressing $\varphi(Y_i, T_i, X_i; \hat{m}_{1,RML}, \hat{\pi}_{RML})$ on $Z$; see Lee, Okui and Whang (2017) for low-dimensional settings, and Fan et al. (2021), Zimmert and Lechner (2019), and Semenova and Chernozhukov (2021) for high-dimensional settings. For instance, for a continuous covariate $Z$, a local constant estimator of $\mu^1(z)$ is

$$\hat{\mu}^1(z; \hat{m}_{1,RML}, \hat{\pi}_{RML}) = \frac{\sum_{i=1}^n K_h(Z_i - z)\varphi(Y_i, T_i, X_i; \hat{m}_{1,RML}, \hat{\pi}_{RML})}{\sum_{i=1}^n K_h(Z_i - z)},$$

where $K_h(t) = K(t/h)/h$, $K(t)$ is a kernel function, and $h$ is a bandwidth. The aforementioned works use machine learning algorithms to fit flexible PS and OR models, and use a sample-splitting technique to reduce the effect of estimating parameters in PS and OR models on the resulting estimator of $\mu^1(z)$.

According to Fan et al. (2021), if models (2.2) and (2.3) are both specified correctly or exhibit negligible bias, then $\hat{\mu}^1(z; \hat{m}_{1,RML}, \hat{\pi}_{RML})$ converges to $\mu^1(z)$ at rate $O_p\{(nh)^{-1/2}\}$, and admits the asymptotic expansion

$$\hat{\mu}^1(z; \hat{m}_{1,RML}, \hat{\pi}_{RML}) = \frac{\sum_{i=1}^n K_h(Z_i - z)\varphi(Y_i, T_i, X_i; m_1^*, \pi^*)}{\sum_{i=1}^n K_h(Z_i - z) + R_n(z)},$$

where $R_n(z) = o_p\{(nh)^{-1/2}\}$. However, when only one of the models (2.2) or (2.3) is specified correctly, the asymptotic expansion or the associated CI for $\mu^1(z)$ does not hold in general.

## 3. Methods

We develop new methods to obtain both doubly robust point estimators and model-assisted CIs for $(\mu^1(z), \mu^0(z))$ and $\tau(z)$. We first discuss the estimation of $\mu^1(z)$. Let $\Phi(z) = (\phi_1(z), \ldots, \phi_K(z))^T$ be a vector of basis functions, excluding the constant. Consider a marginal structural model (Robins (1999); Tan (2010)), where $\mu^1(z)$ is represented linearly as

$$\mu^1(z) = \beta_0^* + \beta_1^{*T}\Phi(z), \quad (3.1)$$

where $\beta^* = (\beta_0^*, \beta_1^{*T})^T$ is a vector of parameters. Here, $\Phi(z)$ can be chosen to accommodate different data types of the covariates $Z$, as follows:

(i) $Z$ is a binary variable: Let $\Phi(z) = z$; then, model (3.1) is saturated.

(ii) $Z$ is a categorical variable taking multiple values: For example, suppose that $Z$ is a trichotomous variable encoded as two dummy variables $(Z_1, Z_2)$. Let $\Phi(z) = (z_1, z_2)^T$; then, model (3.1) saturated.

(iii) $Z$ consists of multiple binary variables: Suppose that $Z = (Z_1, Z_2)$, where $Z_1$ and $Z_2$ are binary variables. Let $\Phi(z) = (z_1, z_2, z_1 z_2)^T$; then, model (3.1) is saturated. Importantly, when $Z$ consists of multiple discrete variables, it can be encoded as multiple binary variables.

(iv) $Z$ is a continuous variable: In this case, $\Phi(z)$ can be specified using a spline basis (Schumaker (2007)) or a Fourier basis (Ramsay and Silverman (2005)), similarly to the nonparametric estimation of a regression curve.

(v) $Z$ is a discrete variable with infinite support; for example, it follows a Poisson distribution: Here, $\Phi(z)$ can be specified using the same basis functions as in the case of the continuous $Z$.

(vi) $Z$ is a combination of discrete and continuous variables; for example, $Z = (Z_1, Z_2)$, where $Z_1$ is a binary variable and $Z_2$ is a continuous variable. Then, we can set $\Phi(z) = (z_1, B^T(z_2), z_1 B^T(z_2))^T$, where $B(z_2)$ consists of basis functions of $Z_2$.

Model (3.1) can be made to be saturated by a proper choice of $\Phi(z)$ for a discrete $Z$ with finite support. However, for a continuous $Z$ or discrete $Z$ with infinite support, model (3.1) with a fixed set of basis functions may not hold exactly, that is, $\mu^1(z)$ may not fall in the working model class $\{\beta_0 + \beta_1^T \Phi(z) : (\beta_0, \beta_1) \in \mathbb{R}^{K+1}\}$. In this case, model (3.1) can be interpreted such that $\beta_0^* + \beta_1^{*T}\Phi(z)$ gives the best linear approximation of $\mu^1(z)$ using the basis functions $(1, \Phi(z))$, where

$$(\beta_0^*, \beta_1^*) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} E\{\mu^1(Z) - \beta_0 - \beta_1^T \Phi(Z)\}^2. \tag{3.2}$$

As shown in our simulation study (Section 5), the proposed method performs well when $\beta_0^* + \beta_1^{*T}\Phi(z)$ provides a sufficiently accurate approximation of $\mu^1(z)$.

### 3.1. Regularized calibrated estimation

Instead of using the regularized likelihood estimation discussed in Section 2.2, we adopt a regularized calibrated (RCAL) estimator of $\gamma$ and a regularized weighted likelihood (RWL) estimator of $\alpha_1$ (Tan (2020a,b)). For the PS model (2.3), the RCAL estimator $\hat{\gamma}_{RCAL}$ is defined as a minimizer of

$$L_{RCAL}(\gamma) = L_{CAL}(\gamma) + \lambda ||\gamma_{1:p}||_1, \tag{3.3}$$

where $L_{CAL}(\gamma) = \tilde{E}[T \exp\{-\gamma^T f(X)\} + (1-T)\gamma^T f(X)]$, $\tilde{E}(\cdot)$ denotes the sample average, $||\cdot||_1$ denotes the $L_1$-norm, $\gamma_{1:p}$ is $\gamma$ excluding the intercept, and $\lambda \geq 0$ is a tuning parameter. For the OR model (2.2), the RWL estimator $\hat{\alpha}_{1,RWL}$ is defined as a minimizer of

$$L_{RWL}(\alpha_1; \hat{\gamma}_{RCAL}) = L_{WL}(\alpha_1; \hat{\gamma}_{RCAL}) + \lambda ||(\alpha_1)_{1:q}||_1, \tag{3.4}$$

where $L_{WL}(\alpha_1;\hat{\gamma}_{RCAL}) = \tilde{E}(Tw(X;\hat{\gamma}_{RCAL})[-Y\alpha_1^T g(X) + \Psi\{\alpha_1^T g(X)\}])$, $\Psi(u) = \int_0^u \psi(u')\mathrm{d}u'$, and $w(X;\gamma) = \{1-\pi(X;\gamma)\}/\pi(X;\gamma) = \exp\{-\gamma^T f(X)\}$. Let $\hat{\pi}_{RCAL}(X) = \pi(X;\hat{\gamma}_{RCAL})$ and $\hat{m}_{1,RWL}(X) = m_1(X;\hat{\alpha}_{1,RWL})$ be the fitted PS and OR functions, respectively; several interesting properties algebraically associated with $\hat{\pi}_{RCAL}(X)$ and $\hat{m}_{1,RWL}(X)$ are presented in the Supplementary Material. As indicated by (3.4), $\hat{m}_{1,RWL}(X)$ depends on $\hat{\pi}_{RCAL}(X)$, in contrast to the recent works of Fan et al. (2021) and Semenova and Chernozhukov (2021), in which the PS and OR functions are estimated separately.

### 3.2. Model-assisted CIs of $\mu^1(z)$

For ease of exposition hereafter, we let $\hat{\gamma} = \hat{\gamma}_{RCAL}$, $\hat{\alpha}_1 = \hat{\alpha}_{1,RWL}$, $\hat{\pi} = \hat{\pi}_{RCAL}(X)$, $\hat{m}_1 = \hat{m}_{1,RWL}(X)$, $\hat{\varphi} = \varphi(Y,T,X;\hat{m}_1,\hat{\pi})$, $\varphi^* = \varphi(Y,T,X;m_1^*,\pi^*)$, and $\Phi^\dagger(z) = (1,\Phi(z)^T)^T$. By the identity (2.1) for $\mu^1(z)$ and the expression (3.2) for $(\beta_0^*,\beta_1^*)$, a natural estimator of $\beta^*$ is $\hat{\beta} = (\hat{\beta}_0,\hat{\beta}_1^T)^T = \tilde{E}^{-1}\{\Phi^\dagger(Z)\Phi^\dagger(Z)^T\}\tilde{E}\{\Phi^\dagger(Z)\hat{\varphi}\}$. The corresponding estimator of $\mu^1(z)$ is

$$\hat{\mu}^1(z;\hat{m}_1,\hat{\pi}) = \hat{\beta}^T\Phi^\dagger(z), \tag{3.5}$$

which is easily shown to be a doubly robust point estimator of $\mu^1(z)$ (Tan (2010)).

Remarkably, model-assisted CIs can be derived by a careful specification of $g(X)$ when fitting the OR model (2.2). Define $f(X)\otimes\Phi(Z)$ as the vector of all interactions between $f(X)$ and $\Phi(Z)$. To obtain model-assisted CIs, we set

$$g(X) = (f(X)^T, \{f(X)\otimes\Phi(Z)\}^T)^T. \tag{3.6}$$

Some functions may be repeated in $g(X)$. In that case, we let $g(X)$ be the vector $(f(X)^T,\{f(X)\otimes\Phi(Z)\}^T)^T$, after excluding the duplicated elements. The choice of $f(X)$ is flexible. For instance, it is possible to include full interactions between $V$ and $\Phi(Z)$ in $f(X)$, that is, $f(X) = (1,V^T,\Phi(Z)^T,\{V\otimes\Phi(Z)\}^T)^T$. Interestingly, we can use this choice of $f(X)$ to construct doubly robust CIs for $\mu^1(z)$ with discrete $Z$, as shown in Section 3.3. Furthermore, it is possible to include additional covariates, such as nonlinear terms of $V$, in $f(X)$. These additional terms are easily accommodated under sparsity conditions.

We provide a high-dimensional analysis of $\hat{\mu}^1(z;\hat{m}_1,\hat{\pi})$ in (3.5), allowing for possible model misspecification. Define $\bar{\gamma} = \mathrm{argmin}_\gamma E\{L_{CAL}(\gamma)\}$ and $\bar{\alpha}_1 = \mathrm{argmin}_{\alpha_1} E\{L_{WL}(\alpha_1;\bar{\gamma})\}$. Let $\bar{\pi} = \pi(X;\bar{\gamma})$, $\bar{m}_1 = m(X;\bar{\alpha}_1)$, and $\bar{\varphi} = \varphi(Y,T,X;\bar{m}_1,\bar{\pi})$. We define $\bar{\gamma}$ and $\bar{\alpha}_1$ to discuss the asymptotic properties of the proposed estimator when the PS model or the OR model is *misspecified*. By the definitions, $\hat{\gamma}$ and $\hat{\alpha}_1$ always converge to $\bar{\gamma}$ and $\bar{\alpha}_1$, respectively, regardless of whether or not the working models (2.3) and (2.2) are specified correctly. In addition, if model (2.3) is specified correctly, then $\bar{\pi} = \pi^*$; otherwise, $\bar{\pi} \neq \pi^*$. Similarly, if model (2.2) is specified correctly, then $\bar{m}_1 = m_1^*$; $\bar{m}_1 \neq m_1^*$ otherwise. Let $\bar{\beta} = (\bar{\beta}_0,\bar{\beta}_1^T)^T = \tilde{E}^{-1}\{\Phi^\dagger(Z)\Phi^\dagger(Z)^T\}\tilde{E}\{\Phi^\dagger(Z)\bar{\varphi}\}$, and $\hat{\mu}^1(z;\bar{m}_1,\bar{\pi}) = \bar{\beta}^T\Phi^\dagger(z)$.

Our main result shows that under some regularity conditions,

$$\hat{\mu}^1(z;\hat{m}_1,\hat{\pi}) = \hat{\mu}^1(z;\bar{m}_1,\bar{\pi}) + R_n(z), \tag{3.7}$$

with $|R_n(z)| = o_p(n^{-1/2})$ for both discrete $Z$ and continuous $Z$. For a vector $b = (b_0, b_1, \ldots, b_p)^T$, denote $S_b = \{0\} \cup \{j : b_j \neq 0, j = 1, \ldots, p\}$ and the size of the set $S_b$ as $|S_b|$.

**Proposition 1 (Model-assisted CIs).** *Suppose that regularity Assumptions* $1 - 2$ *in the Supplementary Material hold,* $g(X)$ *is chosen as in* (3.6)*, and* $(|S_{\bar{\gamma}}| + |S_{\bar{\alpha}_1}|)\log(q) = o(n^{1/2})$*. If the PS model* (2.3) *is specified correctly, then asymptotic expansion* (3.7) *is valid. Furthermore, for any given* $z_0$*, the following results hold:*

*(i)* $n^{1/2}\{\hat{\mu}^1(z_0;\hat{m}_1,\hat{\pi}) - \mu^1(z_0)\} \xrightarrow{D} N\{0, V(z_0)\}$*, where*

$$V(z_0) = var[\Phi^\dagger(z_0)^T E^{-1}\{\Phi^\dagger(Z)\Phi^\dagger(Z)^T\}\Phi^\dagger(Z)\varphi(Y,T,X;\bar{m}_1,\bar{\pi})].$$

*(ii) a consistent estimator of* $V(z_0)$ *is*

$$\hat{V}(z_0) = \frac{\Phi^\dagger(z_0)^T M^{-1}\hat{G}M^{-1}\Phi^\dagger(z_0)}{n},$$

*where* $\hat{G} = n^{-1}\sum_{i=1}^n[\Phi^\dagger(Z_i)\Phi^\dagger(Z_i)^T\{\varphi(Y_i,T_i,X_i;\hat{m}_1,\hat{\pi}) - \hat{\beta}^T\Phi^\dagger(Z_i)\}^2]$*, and* $M = \tilde{E}\{\Phi^\dagger(Z)\Phi^\dagger(Z)^T\}$*. That is, we obtain a model-assisted CI for* $\mu^1(z_0)$*.*

For simplicity, the preceding result is stated under model (3.1). If model (3.1) does not hold exactly, then the model-assisted CI remains valid when evaluated against the approximate value $\tilde{\mu}^1(z) = \beta_0^* + \beta_1^{*T}\Phi(z)$ for $(\beta_0^*, \beta_1^*)$ defined in (3.2). In Section 5, the results of our simulation study show that the approximate CIs perform very well.

### 3.3. Doubly robust CIs of $\mu^1(z)$ with discrete $Z$

In this section, we derive doubly robust CIs for $\mu^1(z)$ with discrete $Z$ when using a linear OR model. Consider the linear OR working model

$$E(Y \mid T = 1, X) = m_1(X;\alpha_1) = \alpha_1^T g(X) \tag{3.8}$$

and the PS working model (2.3). Remarkably, doubly robust CIs for $\mu^1(z)$ can be obtained merely by including full interactions between $V$ and $\Phi(Z)$ in $f(X)$, that is, setting

$$f(X) = (1, V^T, \Phi(Z)^T, \{V \otimes \Phi(Z)\}^T)^T, \quad g(X) = (f(X)^T, \{f(X) \otimes \Phi(Z)\}^T)^T. \tag{3.9}$$

The following are specific forms of $f(X)$ and $g(X)$ for different types of discrete $Z$:

(i) $Z$ is a binary variable: $f(X) = g(X) = (1, V^T, Z, V^T Z)^T$;

(ii) $Z$ is trichotomous variable encoded as two dummy variables $(Z_1, Z_2)$; $f(X) = g(X) = (1, V^T, Z_1, Z_2, V^T Z_1, V^T Z_2)^T$;

(iii) $Z$ consists of two binary variables $Z_1$ and $Z_2$: $f(X) = g(X) = (1, V^T, Z_1, Z_2, Z_1 Z_2, V^T Z_1, V^T Z_2, V^T Z_1 Z_2)^T$.

The configuration of (3.9) makes the dimension of $f(X)$ the same as that of $g(X)$. In addition, the proposed setup of $f(X)$ is intuitively sensible, in the sense that the OR and PS models should include interaction terms between $V$ and $Z$. Proposition 2 presents tje large-sample properties of $\hat{\mu}^1(z_0; \hat{m}_1, \hat{\pi})$ for discrete $Z$.

**Proposition 2 (Doubly robust CIs).** *Suppose that regularity Assumptions* 1–2 *in the Supplementary Material hold, $f(X)$ and $g(X)$ are chosen as in* (3.9)*, and $(|S_{\bar{\gamma}}| + |S_{\bar{\alpha}_1}|)\log(q) = o(n^{1/2})$. Then, the asymptotic expansion* (3.7) *is valid. Moreover, if either the PS model* (2.3) *or the linear OR model* (3.8) *is specified correctly, then for any given $z_0$, the following results hold for discrete $Z$:*

$$n^{1/2}\{\hat{\mu}^1(z_0; \hat{m}_1, \hat{\pi}) - \mu^1(z_0)\} \xrightarrow{D} N\{0, V(z_0)\},$$

*and a consistent estimator of $V(z_0)$ is $\hat{V}(z_0)$, where $V(z_0)$ and $\hat{V}(z_0)$ are as in Proposition* 1*. That is, we obtain a doubly robust CI for $\mu^1(z_0)$.*

Note that the asymptotic expansion (3.7) holds in Proposition 2 without needing a correctly specified PS model (2.3), but does not hold in Proposition 1. The reasons for this phenomenon involve essential ideas about why the proposed methods work. A heuristic interpretation is given in Section 4.1. The results presented in Propositions 1 and 2 focus on estimating $\mu^1(z)$, we extend these results to estimate $\mu^0(z)$ and $\tau(z)$ in the Supplementary Material.

For discrete $Z$, a stratified analysis is used to estimate $\mu^1(z)$ (Abrevaya, Hus and Lieli (2015)). This method splits the sample by $Z$. Then, it estimates $\hat{m}_1$ and $\hat{\pi}$ for each subclass and uses the sample average of $\hat{\varphi}$ as the estimator of $\mu^1(z)$. Next, we show the connections between the proposed method and stratified analysis for discrete $Z$, and discuss the advantages of the proposed approach. Without loss of generality, consider the case of binary $Z$, and take $f(X) = g(X) = (1, V^T, Z, V^T Z)^T$, according to (3.9). We rewrite $f(X)$ as the equivalent expression $f(X) = g(X) = (I\{Z = 0\}, I\{Z = 0\}V^T, I\{Z = 1\}, I\{Z = 1\}V^T)^T$. Then, by setting the gradient of $L_{CAL}(\gamma)$ and $L_{WL}(\alpha_1)$ to zero yields that

$$\tilde{E}[\{T\pi^{-1}(X; \gamma) - 1\}f(X)] = 0, \tag{3.10}$$

$$\tilde{E}[T\{1 - \pi(X; \hat{\gamma})\}\pi^{-1}(X; \hat{\gamma})\{Y - \alpha_1^T f(X)\}f(X)] = 0, \tag{3.11}$$

which are the sample estimating equations for $\gamma$ and $\alpha_1$, respectively (up to

the lasso penalties in high-dimensional settings). We focus on analyzing equation (3.10); equation (3.11) can be discussed similarly. Equation (3.10) can be divided into two equations,

$$\tilde{E}\{(T[1+\exp\{-\gamma_0^T f_0(X)\}]-1)f_0(X)\}=0, \tag{3.12}$$

$$\tilde{E}\{(T[1+\exp\{-\gamma_1^T f_1(X)\}]-1)f_1(X)\}=0, \tag{3.13}$$

where $f_0(X)=I(Z=0)(1,V^T)^T$, $f_1(X)=I(Z=1)(1,V^T)^T$, and $\gamma=(\gamma_0^T,\gamma_1^T)^T$ that satisfies $\gamma^T f(X)=\gamma_0^T f_0(X)+\gamma_1^T f_1(X)$. Here, (3.12) and (3.13) are the sample estimating equations in stratified analysis. However, if there are multiple categories, stratified analysis is troublesome, especially in high-dimensional settings, where it may select different tuning parameters for the lasso penalties and different covariates in the strata. The proposed method is numerically more tractable, with only two lasso tuning parameters for the PS and OR models, while still allowing different covariates to be selected in different strata.

## 4. Asymptotic Properties

### 4.1. Heuristic discussion

We first present the basic ideas underlying the construction of the estimators $\hat{\gamma}$ and $\hat{\alpha}_1$, and show why we need a careful specification of $f(X)$ and $g(X)$ in (3.6) or (3.9), such that the estimator $\hat{\mu}^1(z_0;\hat{m}_1,\hat{\pi})$ satisfies the asymptotic expansion (3.7) *under possible model misspecification.* The discussion here is heuristic. For a given $z_0$, $\hat{\mu}^1(z_0;\hat{m}_1,\hat{\pi})=\hat{\mu}^1(z_0;\bar{m}_1,\bar{\pi})+\Phi^\dagger(z_0)^T\tilde{E}^{-1}\{\Phi^\dagger(Z)\Phi^\dagger(Z)^T\}\tilde{E}\{\Phi^\dagger(Z)(\hat{\varphi}-\bar{\varphi})\}$. For (3.7) to hold, it is sufficient to show that

$$\tilde{E}\{\Phi^\dagger(Z)(\hat{\varphi}-\bar{\varphi})\}=o_p(n^{-1/2}). \tag{4.1}$$

By a Taylor expansion, $\tilde{E}\{\Phi^\dagger(Z)\hat{\varphi}\}=\tilde{E}\{\Phi^\dagger(Z)\bar{\varphi}\}+(\hat{\alpha}_1-\bar{\alpha}_1)^T\Delta_1+(\hat{\gamma}-\bar{\gamma})^T\Delta_2+o_p(n^{-1/2})$, where the remainder is taken to be $o_p(n^{-1/2})$ under suitable conditions, and

$$\Delta_1=\frac{\partial}{\partial\alpha_1}\tilde{E}\{\Phi^\dagger(Z)\varphi(Y,T,X;\alpha_1,\gamma)\}\Big|_{(\alpha_1,\gamma)=(\bar{\alpha}_1,\bar{\gamma})},$$

$$\Delta_2=\frac{\partial}{\partial\gamma}\tilde{E}\{\Phi^\dagger(Z)\varphi(Y,T,X;\alpha_1,\gamma)\}\Big|_{(\alpha_1,\gamma)=(\bar{\alpha}_1,\bar{\gamma})}.$$

To show (4.1), it suffices to show that $(\hat{\alpha}_1-\bar{\alpha}_1)^T\Delta_1=o_p(n^{-1/2})$ and $(\hat{\gamma}-\bar{\gamma})^T\Delta_2=o_p(n^{-1/2})$, *with possible model misspecification.* In general, $\hat{\alpha}_1-\bar{\alpha}_1$ and $\hat{\gamma}-\bar{\gamma}$ are no smaller than $O_p(n^{-1/2})$ in low- or high-dimensional settings. To obtain the desired convergence rates, the crucial point is that $\Delta_1$ and $\Delta_2$ should be $o_p(1)$, and their corresponding population versions should satisfy

$$\frac{\partial}{\partial \alpha_1} E\{\Phi^\dagger(Z)\varphi(Y,T,X;\alpha_1,\gamma)\}\Big|_{(\alpha_1,\gamma)=(\bar{\alpha}_1,\bar{\gamma})} = 0, \tag{4.2}$$

$$\frac{\partial}{\partial \gamma} E\{\Phi^\dagger(Z)\varphi(Y,T,X;\alpha_1,\gamma)\}\Big|_{(\alpha_1,\gamma)=(\bar{\alpha}_1,\bar{\gamma})} = 0. \tag{4.3}$$

Hence, a natural approach is to solve (4.2) and (4.3) in low-dimensional settings, and to add lasso penalties in high-dimensional settings. Nevertheless, this method encounters a basic problem: there are more equations than parameters. It is easy to see that (4.2) includes $(K+1)(q+1)$ equations, and (4.3) contains $(K+1)(p+1)$ equations, whereas the dimensions of $\gamma$ and $\alpha_1$ are $p+1$ and $q+1$, respectively. Therefore, the coefficients $\gamma$ and $\alpha_1$ cannot be identified by solving (4.2) and (4.3) without further consideration. Fortunately, this difficulty can be overcome by a careful specification of $f(X)$ and $g(X)$.

Specifically, with the PS model (2.3) and the linear OR model (3.8), $\Delta_1$ and $\Delta_2$ reduce to

$$\Delta_1 = \tilde{E}[\{T\bar{\pi}^{-1}(X) - 1\}g(X) \otimes \Phi^\dagger(Z)] \text{ and}$$
$$\Delta_2 = \tilde{E}[T\{1-\bar{\pi}(X)\}\bar{\pi}^{-1}(X)\{Y - \bar{\alpha}_1^T g(X)\}f(X) \otimes \Phi^\dagger(Z)],$$

respectively. If $g(X)$ satisfies the form of (3.6), then according to the definition of $\bar{\alpha}_1$, (4.3) holds, regardless of whether or not the OR model is specified correctly. In addition, (4.2) holds, provided that the PS model (2.3) is specified correctly, even if the OR model (3.8) is misspecified, which explains why Proposition 1 can be derived. Furthermore, if $f(X)$ and $g(X)$ are specified as in (3.9), then $\Delta_1$ and $\Delta_2$ have a simpler form with discrete $Z$:

$$\Delta_1 = \tilde{E}[\{T\bar{\pi}^{-1}(X) - 1\}f(X)] \text{ and}$$
$$\Delta_2 = \tilde{E}[T\{1-\bar{\pi}(X)\}\bar{\pi}^{-1}(X)\{Y - \bar{\alpha}_1^T g(X)\}g(X)],$$

respectively, which are the gradients of $L_{CAL}(\bar{\gamma})$ and $L_{WL}(\alpha_1;\bar{\gamma})$, respectively. In this case, (4.2) and (4.3) hold by the definitions of $\bar{\gamma}$ and $\bar{\alpha}_1$, irrespective of the PS and OR model specifications, which explains why Proposition 2 can be obtained.

### 4.2. Theoretical analysis

Suppose that the lasso tuning parameters are specified as $A_0\lambda_0$ for $\hat{\gamma}$ and $A_1\lambda_1$ for $\hat{\alpha}_1$, where $A_0$ and $A_1$ are two sufficiently large positive constants, and $(\lambda_0,\lambda_1)$ is set as $\lambda_0 = [\log\{(1+p)/\epsilon\}/n]^{1/2}$, and $\lambda_1 = [\log\{(1+q)/\epsilon\}/n]^{1/2}$ $(\geq \lambda_0)$, where $0 < \epsilon < 1$ is a tail probability for the error bound. For example, $\lambda_0 = \{2\log(1+p)/n\}^{1/2}$ by taking $\epsilon = 1/(1+p)$. Tan (2020a) shows that the convergence rates for $(\hat{\gamma},\hat{\alpha}_1)$ are $||\hat{\gamma}-\bar{\gamma}||_1 = O_p(1)\cdot|S_{\bar{\gamma}}|\{\log(p)/n\}^{1/2}$, and $||\hat{\alpha}_1 - \bar{\alpha}_1||_1 = O_p(1)\cdot(|S_{\bar{\gamma}}| + |S_{\bar{\alpha}_1}|)\{\log(q)/n\}^{1/2}$. Theorem 1 presents the large-sample properties of $\hat{\mu}^1(z;\hat{m}_1,\hat{\pi})$ for discrete $Z$.

**Theorem 1 (doubly robust CIs).** *Suppose that regularity Assumptions* 1–2 *in the Supplementary Material hold. If the linear OR model* (3.8) *is used,* $f(X)$ *and* $g(X)$ *are specified as in* (3.9)*. Then, for any given* $z_0$ *of discrete* $Z$*,*

*(a) we have with probability at least* $1 - c_0\epsilon$*,*

$$\left|\hat{\mu}^1(z_0; \hat{m}_1, \hat{\pi}) - \hat{\mu}^1(z_0; \bar{m}_1, \bar{\pi})\right| \leq M_0(|S_{\bar{\gamma}}|\lambda_0 + |S_{\bar{\alpha}_1}|\lambda_1)\lambda_1, \tag{4.4}$$

*where* $c_0$ *and* $M_0$ *are positive constants;*

*(a) if* $(|S_{\bar{\gamma}} + |S_{\bar{\alpha}_1}|)\{\log(q)\}^{1/2} = o(n^{1/2})$*, we have with probability at least* $1 - (c_0 + 4)\epsilon$*,*

$$\hat{V}(z_0) - V(z_0) = o_p(1), \tag{4.5}$$

*where* $V(z_0)$ *and* $\hat{V}(z_0)$ *are defined in Proposition* 2*.*

Because $\hat{\mu}^1(z; \bar{m}_1, \bar{\pi})$ is a doubly robust point estimator of $\mu^1(z)$, $\hat{\mu}^1(z; \hat{m}_1, \hat{\pi})$ is also a doubly robust point estimator of $\mu^1(z)$, provided that $(|S_{\bar{\gamma}}|\lambda_0 + |S_{\bar{\alpha}_1}|\lambda_1)\lambda_1 = o(1)$, that is, $(|S_{\bar{\gamma}}| + |S_{\bar{\alpha}_1}|)\log(q) = o(n)$. In addition, obtaining a valid CI requires that the asymptotic expansion (3.7) holds, which implies that $(|S_{\bar{\gamma}}|\lambda_0 + |S_{\bar{\alpha}_1}|\lambda_1)\lambda_1 = o(n^{-1/2})$, that is, $(|S_{\bar{\gamma}}| + |S_{\bar{\alpha}_1}|)\log(q) = o(n^{1/2})$. In summary, Theorem 1 shows that, for discrete $Z$ with the linear OR model (3.8) and $f(X)$ and $g(X)$ specified as in (3.9), the proposed method obtains both doubly robust point estimators and doubly robust CIs for $\mu^1(z)$, provided that $(|S_{\bar{\gamma}}| + |S_{\bar{\alpha}_1}|)\log(q) = o(n^{1/2})$, which leads to Proposition 2. Similarly to Theorem 1, Theorem 2 implies the results presented in Proposition 1.

**Theorem 2 (Model-assisted CIs).** *Suppose that regularity Assumptions* 1–2 *in the Supplementary Material hold. If the OR model* (2.2) *is used,* $g(X)$ *is specified as in* (3.6)*, and the PS model* (2.3) *is specified correctly. Then, for a given value* $z_0$ *of discrete/continuous* $Z$*,*

*(a) we have with probability at least* $1 - (c_0 + 8)\epsilon$*,*

$$\left|\hat{\mu}^1(z_0; \hat{m}_1, \hat{\pi}) - \hat{\mu}^1(z_0; \bar{m}_1, \bar{\pi})\right| \leq M_1(|S_{\bar{\gamma}}|\lambda_0 + |S_{\bar{\alpha}_1}|\lambda_1)\lambda_1, \tag{4.6}$$

*where* $M_1$ *is a positive constant;*

*(b) if* $(|S_{\bar{\gamma}} + |S_{\bar{\alpha}_1}|)\{\log(q)\}^{1/2} = o(n^{1/2})$*, we have with probability at least* $1 - (c_0 + 12)\epsilon$*,*

$$\hat{V}(z_0) - V(z_0) = o_p(1), \tag{4.7}$$

*where* $V(z_0)$ *and* $\hat{V}(z_0)$ *are defined in Proposition* 1*.*

The preceding theoretical analysis focuses on $\hat{\mu}^1(z; \hat{m}_1, \hat{\pi})$. Similar results can be derived for $\hat{\mu}^0(z; \hat{m}_0, \hat{\pi}_0)$ and $\hat{\mu}^1(z; \hat{m}_1, \hat{\pi}) - \hat{\mu}^0(z; \hat{m}_0, \hat{\pi}_0)$ by analogous arguments.

## 5. Simulation Studies

In this section, we present the simulation studies to evaluate the finite-sample performance of the proposed methods. We consider three scenarios for $Z$: a binary variable $Z$, a continuous variable $Z$, and $Z$ consisting of two binary variables $Z_1$ and $Z_2$. The RCAL estimation for the PS model and the RWL estimation for the OR model can be implemented using the `R` package **RCAL** (Tan and Sun (2019)); the corresponding tuning parameters are determined using five-fold cross-validation. Throughout this simulation, the data-generating processes of the covariates are as follows: $V = (V_1, \ldots, V_d) \sim N(0, \Sigma)$, with $\Sigma_{j,k} = 2^{-|j-k|}$, for $1 \le j, k \le d$, and independently, $Z \sim Ber(0.5)$ or $Z \sim Unif(-0.5, 0.5)$ for discrete or continuous $Z$. For $Z = (Z_1, Z_2)$, $Z_1$ and $Z_2$ are i.i.d. from $Ber(0.5)$. The error term is $\epsilon \sim N(0, 1)$. Let $\gamma = 0.5(1, -1, -1, 1, -1)^T$, $X = (Z^T, V^T)^T$, and $V_i$ be the $i$th element of $V$.

**Discrete** $Z$. We consider three scenarios, (C1)–(C3), with a binary $Z$, to assess the doubly robust properties of the point estimators and the CIs.

(C1) $Z$ is a binary variable; $P(T = 1|X) = \{1 + \exp(-(Z, V_1, V_2, V_3, V_4)^T\gamma)\}^{-1}$ and $Y^1 = 1 + Z + \sum_{i=1}^{4}\{V_i Z + 2V_i(1 - Z)\} + \epsilon$.

(C2) Generate $Z, V$, and $T$ as in case (C1); $Y^1 = 1 + Z + \sum_{i=1}^{4}\{V_i Z + 2V_i(1 - Z) + V_i^3/2^i\} + \epsilon$.

(C3) Generate $Z, V$ and $Y^1$ as in case (C1); $P(T = 1|X) = \{1+\exp(-(Z, V_1^2, V_2^2, V_3^2, V_4^2)^T\gamma)\}^{-1}$.

The three scenarios can be classified as follows: in (C1), both the PS and the OR models are specified correctly; in (C2), the PS model is specified correctly, but the OR model is misspecified; in (C3), the PS model is misspecified, but the OR model is specified correctly.

The true curve of $\mu^1(z)$ is $1 + z$ for all three cases of (C1)–(C3). We set $f(X) = g(X) = (1, V^T, Z, V^T Z)^T$, as discussed in Section 3.3. Each simulation study is based on 1,000 replicates, with a sample size $n = 500$. Bias and Var are the Monte Carlo bias and variance over the 1,000 simulations of the point estimates. EVar is the mean of the variance estimates. CP90 and CP95 are the coverage proportions of the 90% and 95% CIs, respectively. Table 1 summarizes the results of $\hat{\mu}^1(z)$ for scenarios (C1)–(C3).

As shown in Table 1, for all three cases, Bias is small, $\sqrt{\text{EVar}}$ is close to $\sqrt{\text{Var}}$, and the coverage proportions CP90 and CP95 are around the nominal levels of 0.90 and 0.95, respectively. Because case (C2) involves a misspecified OR model and case (C3) involves a misspecified PS model, the results for these cases show that both the point estimators and the CIs are doubly robust.

**Continuous** $Z$. We consider two data-generation mechanisms with continuous $Z$:

Table 1. Estimations of $\mu^1(z)$ for a binary variable $Z$.

| | | $n = 500$, $p = 200$ | | | | | $n = 500$, $p = 400$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\mu}^1(z)$ | Bias | $\sqrt{\text{Var}}$ | $\sqrt{\text{EVar}}$ | CP90 | CP95 | Bias | $\sqrt{\text{Var}}$ | $\sqrt{\text{EVar}}$ | CP90 | CP95 |
| (C1) | $\hat{\mu}^1(0)$ | -0.032 | 0.370 | 0.371 | 0.905 | 0.954 | -0.034 | 0.366 | 0.368 | 0.897 | 0.950 |
| | $\hat{\mu}^1(1)$ | -0.040 | 0.201 | 0.200 | 0.879 | 0.952 | -0.038 | 0.202 | 0.200 | 0.889 | 0.943 |
| (C2) | $\hat{\mu}^1(0)$ | -0.054 | 0.516 | 0.501 | 0.896 | 0.944 | -0.063 | 0.515 | 0.498 | 0.884 | 0.941 |
| | $\hat{\mu}^1(1)$ | -0.081 | 0.348 | 0.337 | 0.887 | 0.935 | -0.072 | 0.336 | 0.336 | 0.898 | 0.951 |
| (C3) | $\hat{\mu}^1(0)$ | 0.019 | 0.377 | 0.361 | 0.888 | 0.938 | 0.033 | 0.375 | 0.357 | 0.874 | 0.935 |
| | $\hat{\mu}^1(1)$ | -0.003 | 0.195 | 0.202 | 0.905 | 0.955 | -0.016 | 0.203 | 0.200 | 0.888 | 0.947 |

Note: the dimensions of $f(X)$ and $g(X)$ are both $p + 1$.

Table 2. Estimations of $\mu^1(z)$ for continuous $Z$, $n = 500$, $p = 60$, and $q = 420$.

| $\hat{\mu}^1(z)$ | Bias | $\sqrt{\text{Var}}$ | $\sqrt{\text{EVar}}$ | CP90 | CP95 | Bias | $\sqrt{\text{Var}}$ | $\sqrt{\text{EVar}}$ | CP90 | CP95 |
|---|---|---|---|---|---|---|---|---|---|---|
| | (C4) cor PS, cor OR | | | | | (C5) cor PS, mis OR | | | | |
| $\hat{\mu}^1(-0.4)$ | 0.011 | 0.409 | 0.400 | 0.886 | 0.942 | -0.027 | 0.201 | 0.199 | 0.882 | 0.942 |
| $\hat{\mu}^1(-0.2)$ | -0.035 | 0.359 | 0.341 | 0.881 | 0.936 | -0.020 | 0.172 | 0.170 | 0.884 | 0.944 |
| $\hat{\mu}^1(0.0)$ | -0.036 | 0.339 | 0.331 | 0.892 | 0.940 | -0.025 | 0.179 | 0.166 | 0.866 | 0.928 |
| $\hat{\mu}^1(0.2)$ | -0.013 | 0.341 | 0.342 | 0.899 | 0.948 | -0.036 | 0.173 | 0.177 | 0.894 | 0.948 |
| $\hat{\mu}^1(0.4)$ | -0.034 | 0.414 | 0.403 | 0.883 | 0.941 | -0.028 | 0.208 | 0.207 | 0.880 | 0.938 |

Note: the dimensions of $f(X)$ and $g(X)$ are $p + 1$ and $q + 1$, respectively.

(C4) $Z$ is a continuous variable; $Y^1 = Z + \sum_{i=1}^4 V_i + \epsilon$, $P(T = 1|X) = \{1 + \exp(-(Z, V_1, V_2, V_3, V_4)^T\gamma)\}^{-1}$. Both the PS and the OR models are specified correctly.

(C5) Generate $Z, V$, and $T$ as in case (C4); $Y^1 = Z(1+2Z)^2(Z-1)^2+\sum_{i=1}^4(V_i^2+V_i)/2^{i+1} + \epsilon$. The PS model is specified correctly, but the OR model is misspecified.

We set $f(X) = (1, V^T, Z)^T$, $g(X)$ is specified as in (3.6), and $\Phi(Z)$ are cubic spline basis functions with three knots selected using the 25%, 50%, and 75% sample quantiles of $Z$, which can be implemented using the `R` package **gam** (Hastie (2018)). Because $Z$ is continuous for cases (C4) and (C5), we report the simulation results at five representative points of $Z$: -0.4, -0.2, 0, 0.2, 0.4. Table 2 presents the numerical results of $\hat{\mu}^1(z)$ for cases (C4) and (C5), showing that both perform similarly to the cases of discrete $Z$.

We compare the proposed method with the AIPW methods of Fan et al. (2021) and Zimmert and Lechner (2019), discussed in Section 2.2 for continuous $Z$. The implementation details and associated results are given in Table S1 of Supplemental Material. Figure 1 presents the average values of CP90 and CP95 at five representative points for case (C5), suggesting that the competing methods *do not* enjoy the property of model-assisted CIs, whereas the proposed method *does*.

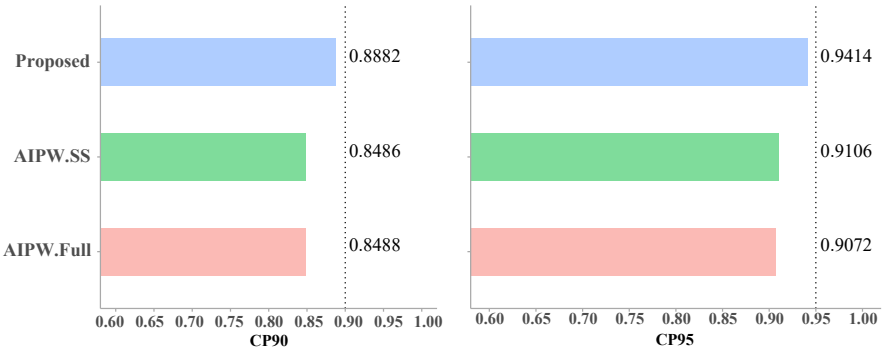


Figure 1. Average values of CP90 and CP95 at five representative points based on 1,000 simulations, where CP90 and CP95 are the coverage proportions of the 90% and 95% CIs, AIPW.Full refers to the AIPW method of Fan et al. (2021) with the full sample (without sample splitting), and AIPW.SS refers to the AIPW methods of Fan et al. (2021) and Zimmert and Lechner (2019) with four-fold cross-fitting (sample-splitting).

The aforementioned numeric results are all in exact sparsity settings. A more common scenario in modern applications is approximate sparsity, that is, all covariates are relevant and associated with nonzero coefficients, but only a few are truly important with large coefficients. We also conduct numerical experiments to assess the finite-sample performance of the proposed methods under approximate sparsity settings. The corresponding results are similar to those of the exact sparsity settings and are presented in Table S2 of the Supplementary Material.

## 6. Application

Psoriasis is a chronic immune-mediated inflammatory disease that can damage a patient's quality of life and increase the burden on society (Griffiths and Barker (2007); Griffiths et al. (2021)). Immunological and genetic studies have discovered that the key drivers of psoriasis are pathogenesis proinflammatory cytokines (tumor necrosis factor-alpha (TNF$\alpha$), interleukin-17 (IL-17), and interleukin-23 (IL-23))(Park et al. (2005)). The new generation of biologics are IL-17 inhibitors (secukinumab, ixekizumab, and brodalumab) and IL-23 inhibitors (guselkumab, risankizumab, and tildrakizumab). In 2020, secukinumab was added to the national drug reimbursement list (NDRL) of China for treating psoriasis. Since then, clinical use of secukinumab has increased significantly.

In China, biologics are used only for patients who are unresponsive, intolerant, or contraindicated for systemic therapy to cure severe plaque psoriasis or arthropathic psoriasis (Menter et al. (2019); Comittee on Psoriasis Chinese Society of Dermatology (2019)). Evidence of the efficacy of biologics is limited, especially for mild-to-moderate psoriasis. In addition, with the increase of clinical usage, how to use biologics effectively and appropriately is becoming a major concern of therapists (Chen et al. (2020)). In this study, we explore the

heterogeneous effects of biologics versus those of conventional therapies across different subpopulations.

### 6.1. Data description

Data were collected from the Psoriasis Center Data Platform (PCDP), which is led by the National Clinical Research Center for Skin and Immune Disease, and covers patients of 237 tertiary hospitals in about 100 cities in mainland of China. In this study, the data are restricted to patients who enrolled between September 2020 and September 2021, and were diagnosed with plaque psoriasis and had at least one follow-up visit. In addition, we include patients treated with IL-inhibitor biologics (mainly the IL-17 inhibitor, secukinumab) or with conventional therapies (topic drugs, systemic medicines, or phototherapy), and eliminate data with other treatments. The use of biologics is divided into an induction period and a maintained period. Here, we assess the treatment heterogeneity in the maintained period, that is, excluding patients with a follow-up time of less than four weeks. The final analytical dataset contains 2,356 samples, where 708 (30.05%) use biologics and 1,648 (69.95%) do not.

The clinical benefit of a treatment is measured by an improvement in the psoriasis area and severity index (PASI). In this study, the outcome variable $Y$ (`PASI 80`) is the indicator of an 80% or more improvement from the baseline PASI (`baseline PASI`) in the first follow-up visit. The exposure variable $T = 1$ indicates that the patient was treated with IL-inhibitor biologics, and $T = 0$ means conventional therapies were used. The covariates $X$ include the patient's demographics, clinical characteristics, and interactions. Descriptions of these variables are provided in Table S3 of the Supplementary Material. The subpopulations of interest are defined by the covariates $Z$, which are taken to be `baseline PASI`, `baseline DLQI`, `Age`, `Employment`, `Martial status`, `Education`, `Insurance`, and `Sex`. The first three variables are continuous, and the last five are binary. Because only a few samples are available at extremely high values of `baseline PASI`, we set the `baseline PASI` values above 45 to 45. Similarly, we set `Age` above 75 to 75.

As suggested in Sections 3.3 and 3.2, we set $f(X) = g(X) = (1, V^T, Z, V^T Z)^T$ for binary $Z$, and $f(X) = (1, V^T, \Phi(Z)^T)^T$ and $g(X) = (1, V^T, \Phi(Z)^T, (V \otimes \Phi(Z))^T, (\Phi(Z) \otimes \Phi(Z))^T)^T$ for continuous $Z$, where $V$ denotes all the covariates, excluding the variable $Z$. All variables in $f(X)$ and $g(X)$ are standardized to have sample mean zero and sample variance one. As in our simulation studies, we select the lasso tuning parameter $\lambda$ using five-fold cross-validation.

### 6.2. Results

**Discrete** $Z$. Figure 2 presents the estimated causal effects of biologics in terms of the improvement of the PASI, conditional on different binary variables $Z$.

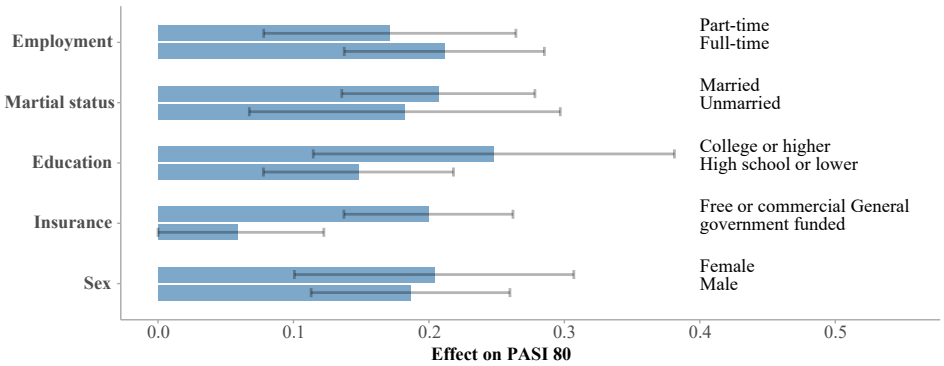


Figure 2. Estimated CSTE (the height of the bar plot) and the associated 95% CI (error bar) for `PASI 80` with discrete $Z$.

The results show that higher education and free or commercial insurance groups have larger biologics benefits than the corresponding lower education and general government-funded insurance groups. In addition, the causal effects exhibit no significant differences between subgroups with different values of `Employment`, `Marital status`, and `Sex`.

**Continuous** $Z$. To estimate a CSTE curve when $Z$ is continuous, we apply the proposed method using a cubic spline to approximate $\tau(z)$, and find the optimal number of knots by using a grid search with the Akaike information criterion (AIC) (Akaike (1974)) and the Bayesian information criterion (BIC) (Schwarz (2005)). Specifically, we first fix the number of knots as three, that is, $f(X) = (1, V^T, \Phi(Z)^T)^T$ and $g(X) = (1, V^T, \Phi(Z)^T, (V \otimes \Phi(Z))^T, (\Phi(Z) \otimes \Phi(Z))^T)^T$, with $\Phi(Z)$ being cubic spline basis functions with three knots. Then, we use $f(X)$ and $g(X)$ to estimate the PS and OR functions. Finally, we conduct a least squares regression of $\varphi(Y, T, X; \hat{m}_1, \hat{\pi}) - \varphi(Y, 1-T, X; \hat{m}_0, 1-\hat{\pi}_0)$ on $\tilde{\Phi}(Z)$ to get the values of AIC and BIC, where $\tilde{\Phi}(Z)$ are cubic spline basis functions with number of knots ranging from 1 to 10. The corresponding results are given in Table S4 of the Supplementary Material.

Figure 3 displays the estimated CSTE curves and corresponding 95% CIs for different continuous $Z$. It indicates that biologics have a positive effect over conventional therapies, and that heterogeneity is ubiquitous in different subpopulations. As `baseline PASI` increases from zero to five, the relative advantage of biological agents over conventional therapy increases; When the score exceeds five points, the CSTE is a V-shaped curve, with a trough near the PASI value of 10 (Figure 3A). Our results indicate that biologics are also effective for mild-to-moderate psoriasis (`baseline PASI` $\leq 10$). A higher value of the self-reported dermatology life quality index (`baseline DLQI`) means a worse quality of life, where DLQI = 0 means the life quality is not affected at all by psoriasis. As shown in Figure 3B, the CSTE is a V-shaped curve as the value of DLQI

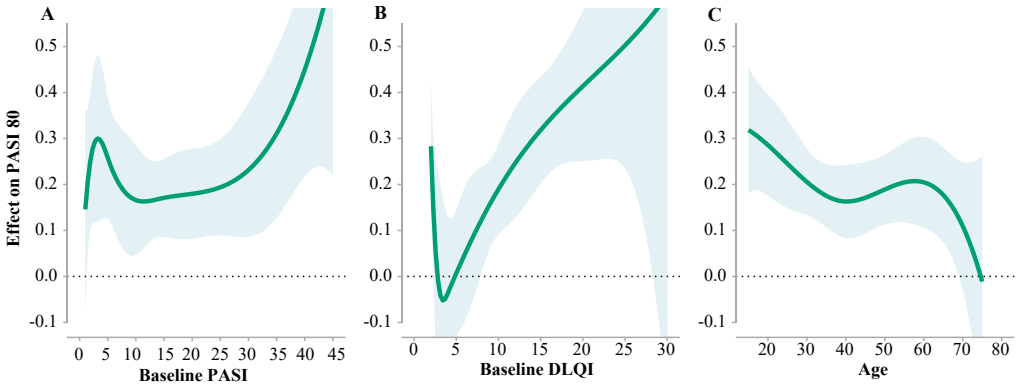


Figure 3. Estimated CSTE and the associated 95% CI for `PASI 80` with continuous $Z$.

increases. In addition, the effect first decreases slightly for ages between 12 and 30, flattens for ages 30 to 60, and then decreases rapidly for ages between 60 and 75. Overall, older patients show lower benefits of biologics.

## 7. Discussion

We have developed new methods for obtaining doubly robust point estimators and model-assisted CIs for conditional ATEs in high-dimensional settings. In addition, with a linear OR model and discrete $Z$, the CIs are also doubly robust. We establish the theoretical properties for the proposed methods with different data types for the outcome $Y$ and the covariates $Z$, and estimate the corresponding variances using a sandwich method.

Further work is required to extend our method and theory by relaxing the parametric structural model (3.1) to be nonparametric, subject to smoothness conditions, while allowing the basis functions $\Phi(z)$ to be chosen based on the data, rather than being prespecified. Another interesting question is whether we can derive doubly robust CIs for continuous $Z$. Here, a possible approach is to discretize $Z$. For example, for two knots $t_1 < t_2$, we can discretize $Z$ as $(Z_1, Z_2) = (I\{t_1 < Z \le t_2\}, I\{Z > t_2\})$ or $(Z_1, Z_2) = (I\{Z > t_1\}, I\{Z > t_2\})$. With either choice of $(Z_1, Z_2)$, the proposed method using $f(X) = (1, V^T, Z_1, Z_2, V^T Z_1, V^T Z_2)^T$ achieves the desired doubly robust CIs in the discretized model $\mu^1(Z) = E(Y^1|Z) = \beta_0 + (Z_1, Z_2)\beta_1$. The method can be extended easily to include multiple knots, corresponding to a piecewise constant model for $\mu^1(z)$. Then, various theoretical questions need to be investigated, such as studying the convergence and whether we can achieve doubly CIs, depending on the number of knots used.

Another extension is to consider $Z$ composed of multiple continuous variables. A possible strategy is to postulate an additive model (Hastie and Tibshirani (1990)). Alternatively, we may consider a single-index model (Guo, Zhou and Ma (2021)). These topics are left to future research.

## Supplementary Material

Supplementary Material available online includes technical proofs and additional numerical results from the simulation and application.

## Acknowledgments

The authors thank the assistant editor and the anonymous reviewers for their helpful comments and valuable suggestions. This research was supported by the State Key Research Program (No. 2021YFF0901400), the National Natural Science Foundation of China (Nos. 11971064, 12071015, and 12171374), and the Major Project of National Statistical Science Foundation of China (No. 2021LD01).

## References

Abrevaya, J., Hus, Y. C. and Lieli, R. P. (2015). Estimating conditional average treatment effect. *Journal of Business and Economic Statistics* **33**, 485–505.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723.

Athey, S., Imbens, G. W. and Wager, S. (2018). Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society, Series B* **80**, 597–623.

Belloni, A., Chernozhukov, V., Fernández-Val, I. and Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica* **85**, 233–298.

Belloni, A., Chernozhukov, V. and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies* **81**, 608–650.

Bradic, J., Wager, S. and Zhu, Y. (2019). Sparsity double robust inference of average treatment effects. *arXiv:1905.00744*.

Chakraborty, B. and Moodie, E. E. (2013). *Statistical Methods for Dynamic Treatment Regimes*. Springer, New York.

Chen, A.-J., Gao, X.-H., Gu, H. and et al. (2020). Chinese experts consensus on biologic therapy for psoriasis. *International Journal of Dermatology and Venereology* **3**, 76–85.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. et al. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal* **21**, C1–C68.

Chernozhukov, V., Fernández-Val, I. and Luo, Y. (2018). The sorted effects method: Discovering heterogeneous effects beyond their averages. *Econometrica* **86**, 1911–1938.

Comittee on Psoriasis Chinese Society of Dermatology (2019). Guideline for the diagnosis and treatment of psoriasis in China. *Chinese Journal of Dermatology* **52**, 667–710.

Dukes, O. and Vansteelandt, S. (2020). Inference for treatment effect parameters in potentially misspecified high-dimensional models. *Biometrika* **108**, 321–334.

Fan, Q., Hsu, Y. C., Lieli, R. P. and Zhang, Y. (2021). Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business and Economic Statistics* **40**, 313–327.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics* **189**, 1–23.

Griffiths, C. E. and Barker, J. N. (2007). Pathogenesis and clinical features of psoriasis. *The Lancet* **370**, 263–271.

Griffiths, C. E. M., Armstrong, A. W., Gudjonsson, J. E. and Barker, J. N. W. N. (2021). Psoriasis. *The Lancet* **397**, 1301–1315.

Guo, W., Zhou, X. H. and Ma, S. (2021). Estimation of optimal individualized treatment rules using a covariate-specific treatment effect curve with high-dimensional covariates. *Journal of the American Statistical Association* **116**, 309–321.

Hastie, T. (2018). *gam: Generalized Additive Models*. (version 1.22-2) Web: `https://CRAN.R-project.org/package=gam`.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Lechner, M. (2019). Modified causal forests for estimating heterogeneous causal effects. *arXiv:1812.09487*.

Lee, S., Okui, R. and Whang, Y. J. (2017). Doubly robust uniform confidence band for the conditional average treatment effect function. *Journal of Applied Econometrics* **32**, 1207–1225.

Menter, A., Strober, B. E., Kaplan, D. H., Kivelevitch, D., Prater, E. F., Stoff, B. et al. (2019). Joint AAD-NPF guidelines of care for the management and treatment of psoriasis with biologics. *Journal of the American Academy of Dermatology* **80**, 1029–1072.

Neyman, J. S. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science* **5**, 465–472.

Ning, Y., Peng, S. and Imai, K. (2020). Robust estimation of causal effects via a high-dimensional covariate balancing propensity score. *Biometrika* **107**, 533–554.

Park, H., Li, Z., Yang, X. O , Chang, S. H., Nurieva, R., Wang, Y.-H. et al. (2005). A distinct lineage of CD4 T cells regulates tissue inflammation by producing interleukin 17. *Nature Immunology* **6**, 1133–1141.

Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*. 2nd Edition. Springer.

Robins, J. M. (1999). Marginal structural models versus structural nested models as tools for causal inference In *Statistical Models in Epidemiology: The Environment and Clinical Trials*, 95-134. Springer, New York.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688–701.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.

Schumaker, L. L. (2007). *Spline Functions: Basic Theory*. 3rd Edition. Cambridge University Press.

Schwarz, G. (2005). Estimating the dimension of a model. *The Annals of Statistics* **6**, 15–18.

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

Smucler, E., Rotnitzky, A. and Robins, J. M. (2019). A unifying approach for doubly-robust $l_1$ regularized estimation of causal contrasts. *arXiv:1904.03737*.

Sun, B. and Tan, Z. (2021). High-dimensional model-assisted inference for local average treatment effects with instrumental variables. *Journal of Business and Economic Statistics* **40**, 1732–1744.

Tan, Z. (2010). Nonparametric likelihood and doubly robust estimating equations for marginal and nested structural models. *The Canadian Journal of Statistics* **38**, 609–632.

Tan, Z. and Sun, B.(2019). *RCAL: Regularized Calibrated Estimation.* (version 2.0) Web: `https://CRAN.R-project.org/package=RCAL`.

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

Tan, Z. (2020b). Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *Biometrika* **107**, 137–158.

Tian, L., Alizadeh, A. A., Gentles, A. J. and Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association* **109**, 1517–1532.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **58**, 267–288.

Wang, Y. and Shah, R. D. (2020). Debiased inverse propensity score weighting for estimation of average treatment effects with high-dimensional confounders. *arXiv:2011.08661.*

Zimmert, M. and Lechner, M. (2019). Nonparametric estimation of causal heterogeneity under high-dimensional confounding. *arXiv:1908.08779v1.*

Peng Wu

School of Mathematics and Statistics, Beijing Technology and Business University, Beijing, China.

E-mail: pengwu@btbu.edu.cn

Zhiqiang Tan

Department of Statistics, Rutgers University, Piscataway, NJ 08854, USA.

E-mail: ztan@stat.rutgers.edu

Wenjie Hu

Department of Probability and Statistics, Peking University, Beijing 100871, China.

E-mail: huwenjie@pku.edu.cn

Xiao-Hua Zhou

Department of Biostatistics, Peking University, Beijing 100871, China.

E-mail: azhou@math.pku.edu.cn

(Received March 2022; accepted July 2022)