Statistica Sinica Preprint No: SS-2022-0089	
Title	Model-Assisted Inference for Covariate-Specific Treatment Effects with High-dimensional Data
Manuscript ID	SS-2022-0089
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0089
Complete List of Authors	Peng Wu, Zhiqiang Tan, Wenjie Hu and Xiao-Hua Zhou
Corresponding Authors	Xiao-Hua Zhou
E-mails	azhou@math.pku.edu.cn
Notice: Accepted version subject to English editing.	

Statistica Sinica

Model-Assisted Inference for Covariate-Specific Treatment Effects with High-dimensional Data

Peng Wu[a], Zhiqiang Tan[b], Wenjie Hu[c] and Xiao-Hua Zhou[c,d]*

[a]*Beijing Technology and Business University*

[b]Rutgers University, [c]Peking University, [d]Pazhou Lab,

Abstract: Covariate-specific treatment effects (CSTEs) represent heterogeneous treatment effects across subpopulations defined by certain selected covariates. In this article, we consider marginal structural models where CSTEs are linearly represented using a set of basis functions of the selected covariates. We develop a new approach in high-dimensional settings to obtain not only doubly robust point estimators of CSTEs, but also model-assisted confidence intervals, which are valid when a propensity score model is correctly specified but an outcome regression model may be misspecified. With a linear outcome model and subpopulations defined by discrete covariates, both point estimators and confidence intervals are doubly robust for CSTEs. In contrast, confidence intervals from existing high-dimensional methods are valid only when both the propensity score and outcome models are correctly specified. We establish asymptotic properties of

*correspond to: azhou@math.pku.edu.cn. This research was supported by the state key research program (No. 2021YFF0901400) and the National Natural Science Foundation of China [No.11971064 and 12171374].

the proposed point estimators and the associated confidence intervals. Simulation studies demonstrate the advantages of the proposed method compared with competing ones. We apply the proposed methods to a large psoriasis clinical dataset from a national registry in China, Psoriasis Center Data Platform (PCDP), exploring the effects of biologics versus the 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 taken to be the estimand of interest for simplicity and interpretation. However, researchers and policymakers can also be interested in the effects of treatments (or policies) at various subpopulation levels (Lee et al., 2017; Chernozhukov et al., 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 subpopulations. Define (Y^0, Y^1) as the potential outcomes under treatment arms 0 and 1 respectively. Of interest in this paper is 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 the conventional therapies on Psoriasis Area and Severity (PASI) improvement in different subpopulations defined by patient's age, baseline Dermatology Life Quality Index (DLQI), and baseline PASI. In clinical settings, CSTEs are useful in precision medicine for the discovery of optimal treatment regimes that can be tailored to individual's characteristics (Chakraborty and Moodie, 2013).

For observational studies, a large set of covariates are often included, possibly with nonlinear and interaction terms, in statistical analysis 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 obtain the identification of 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. Conditioning on a low-dimensional covariate, $\tau(z)$ is easier to interpret and communicate in practice. Moreover, estimation of $\tau(z)$ can be more manageable and less affected by modeling assumptions in statistical analysis. It is known to be difficult to obtain asymptotic normality and valid confidence intervals for $\tau(x)$ due to the high dimensionality of X, unless some restrictive assump-

tions are imposed (Tian et al., 2014; Dukes and Vansteelandt, 2020; Guo et al., 2021).

There has been increasing interest in estimating CSTEs in recent years. Abrevaya et al. (2015) derived an inverse probability weighting (IPW) estimator of $\tau(z)$ using kernel smoothing with continuous Z, Lee et al. (2017) proposed an AIPW (augmented IPW) estimator based on kernel smoothing, and Lechner (2019) proposed algorithms to construct causal random forests. These three approaches estimate $\tau(z)$ in low-dimensional settings. Fan et al. (2021), Zimmert and Lechner (2019) and Semenova and Chernozhukov (2021) extended the method of Lee et al. (2017) to high-dimensional settings. The authors adopted machine learning algorithms to mitigate model specification for nuisance parameters (PS and OR models) and used sample splitting (or cross-fitting) technique to reduce the impact of nuisance parameters estimation on the resulting estimator of $\tau(z)$. A limitation of these existing high-dimensional methods is that the confidence intervals are shown to be valid only when both PS and OR models are correctly specified. This is because Neyman orthogonality condition (Chernozhukov et al., 2018) cannot ensure the negligibility of first-order approximation error of $\tau(z)$ when only one of the PS and OR models is correctly specified. Further discussion is provided in Section 2.2.

In this paper, we consider three desired properties: (a) the point estimator is doubly robust, which is consistent if either the PS model or the OR model is correctly specified; (b) the confidence intervals are valid if the PS model is correctly specified but the OR model may be misspecified; (c) the confidence intervals are valid if the OR model is correctly specified but the PS model may be misspecified. If either property (b) or (c) is met, then the confidence intervals are called model-assisted (Tan, 2020a). If properties (2) and (3) are satisfied, then the confidence intervals are doubly robust. This article develops a new approach for CSTEs in high-dimensional settings that possesses the properties (1) and (2) for continuous Z. Furthermore, with a linear OR model and discrete Z, the proposed method possesses the properties (1), (2), and (3) simultaneously. To the best of our knowledge, there is no method for estimating CSTEs that possesses model-assisted or doubly robust confidence intervals, while retaining double robustness of the point estimator.

Our proposed method is motivated by Tan (2020a), which is first proposed to estimate ATEs and average treatment effects on treated, and recently extended to estimate local average treatment effects in high-dimensional settings (Sun and Tan, 2021). In this article, we further extend the method to tackle the estimation of CSTEs. When Z is discrete with finite support,

the proposed method is closely related to stratified analysis based on Tan (2020a), which first splits the sample by Z, and then applies Tan's method for ATE estimation separately within each stratum. However, stratified analysis is troublesome if Z takes many possible categories, together with a high-dimensional auxiliary covariate vector V, where different tuning parameters need to be selected by separate cross-validations. In comparison, without splitting the sample, the proposed method 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, the direct extension of Tan's approach can't guarantee the model-assisted property. However, our approach can get the model-assisted confidence intervals.

The proposed method relies on similar sparsity conditions as in Tan (2020a). For example, for logistic PS model and linear OR model with coefficients γ and α_1, suppose the estimators of $\hat{\gamma}$ 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 confidence intervals 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 to those in Belloni et al. (2014, 2017), Farrell (2015) and Chernozhukov et al. (2018) for ATE based on commonly penalized PS

and OR models, and those in Athey et al. (2018), Bradic et al. (2019), Ning et al. (2020) and Wang and Shah (2020) allowing model misspecification. See Smucler et al. (2019) for a discussion which distinguishes model doubly robust and rate doubly robust estimations in high-dimensional settings.

The rest of the article is structured as follows. In Section 2, we state the setup of the problem interested and discuss some existing methods. Section 3 presents our estimation procedures in details. Section 4.2 shows the asymptotic results and elucidates why the proposed methods work. In Section 5, extensive simulations are conducted to evaluate the finite sample performance of the proposed methods. Section 6 illustrates our methods with an empirical example. A brief discussion is presented in Section 7.

2. Background

2.1 Setup

Suppose that $\{(Y_i, T_i, X_i) : i = 1, ..., n\}$ is an independent and identically distributed 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 is the covariates used to define subpopulations, V is auxiliary covariates. In the potential outcomes framework (Rubin, 1974; Neyman, 1990), let (Y^0, Y^1) be the potential outcomes under the treatment

arms 0 and 1 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, Assumption 1 is imposed throughout:

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 called propensity score (Rosenbaum and Rubin, 1983).

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

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

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 outcome regression (OR) function $m_t^*(X)$ or the propensity score (PS) $\pi^*(X)$. We mainly discuss estimation of $\mu^1(z)$ and put the discussion about $\mu^0(z)$ and $\tau(z)$ in Supplementary Material.

2.2 Existing doubly robust estimators and theirs limitations

Consider a conditional mean working model for 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 PS

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

where $g(X)=\{1,g_1(X),...,g_q(X)\}^T$ and $f(X)=\{1,f_1(X),...,f_p(X)\}^T$ are two vectors of known functions, $\psi(\cdot)$ is a known inverse link function. In high-dimensional settings, let $\hat{\alpha}_{1,RML}$ and $\hat{\gamma}_{RML}$ are lasso regularized maximum likelihood estimators (Tibshirani, 1996) of α_1 and γ. Denote $\hat{m}_{1,RML}(X)=m_1(X;\hat{\alpha}_{1,RML})$ and $\hat{\pi}_{RML}(X)=\pi(X;\hat{\gamma}_{RML})$. Let

$$\varphi(Y_i,T_i,X_i;\hat{m}_{1,RML},\hat{\pi}_{RML})=T_iY_i/\hat{\pi}_{RML}(X_i)-(T_i/\hat{\pi}_{RML}(X_i)-1)\hat{m}_{1,RML}(X_i). \tag{2.4}$$

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

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

where $K_h(t)=K(t/h)/h$, $K(t)$ is a kernel function and h is a bandwidth. These authors also adopted machine learning algorithms to fit flexible PS and OR models, and used sample splitting technique to reduce the impact

of parameter estimation in PS and OR models on the resulting estimator of $\mu^1(z)$.

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

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

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

3. Methods

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

$$\mu^1(z) = \beta_0^* + \beta_1^{*T}\Phi(z), \tag{3.5}$$

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

follows:

(i) Z is a binary variable. Let $\Phi(z) = z$, then model (3.5) 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 (3.5) saturated.

(iii) Z consists of multiple binary variables. Suppose that $Z = (Z_1, Z_2)$, where Z_1 and Z_2 are two binary variables. Let $\Phi(z) = (z_1, z_2, z_1 z_2)^T$, then model (3.5) is saturated. Importantly, when Z consists of multiple discrete variables, it can be encoded as multiple binary variables.

(iv) Z is a continuous variable. Then $\Phi(z)$ can be specified using spline basis (Schumaker, 2007) and Fourier basis (Ramsay and Silverman, 2005) similarly as in the nonparametric estimation of a regression curve.

(v) Z is a discrete variable with infinite support such as being poisson distributed. Then $\Phi(z)$ can be specified using the same basis functions as 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 (3.5) can be made to be saturated by a proper choice of $\Phi(z)$

for a discrete Z with finite support. But for a continuous Z or discrete Z with infinite support, model (3.5) with a fixed set of basis functions may not hold exactly, i.e., $\mu^1(z)$ 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 (3.5) can be interpreted such that $\beta_0^* + \beta_1^{*T}\Phi(z)$ gives the best linear approximation of $\mu^1(z)$ using basis functions $(1, \Phi(z))$, where

$$(\beta_0^*, \beta_1^*) = \arg\min_{\beta_0,\beta_1} E(\mu^1(Z) - \beta_0 - \beta_1^T\Phi(Z))^2. \tag{3.6}$$

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 regularized likelihood estimation in Section 2.2, we adopt the regularized calibrated (RCAL) estimator of γ and regularized weighted likelihood (RWL)estimator of α_1 (Tan, 2020a,b). For PS model (2.3), the RCAL estimator $\hat{\gamma}_{RCAL}$ is defined as a minimizer of

$$L_{RCAL}(\gamma) = L_{CAL}(\gamma) + \lambda||\gamma_{1:p}||_1, \tag{3.7}$$

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 γ excluding the intercept,

$\lambda \geq 0$ is a tuning parameter. For OR model (2.2), the RWL estimator $\hat{\alpha}_{1,RWL}$ is defined as a minimizer of

$$L_{RWL}(\alpha_1; \hat{\gamma}_{RCAL}) = L_{WL}(\alpha_1; \hat{\gamma}_{RCAL}) + \lambda ||(\alpha_1)_{1:q}||_1, \tag{3.8}$$

where $L_{WL}(\alpha_1; \hat{\gamma}_{RCAL}) = \tilde{E}[Tw(X; \hat{\gamma}_{RCAL})\{-Y\alpha_1^T g(X) + \Psi(\alpha_1^T g(X))\}]$, $\Psi(u) = \int_0^u \psi(u')\mathrm{d}u'$, $w(X;\gamma) = \{1 - \pi(X;\gamma)\}/\pi(X;\gamma) = \exp\{-\gamma^T f(X)\}$. 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 Supplementary Material. As indicated by (3.8), $\hat{m}_{1,RWL}(X)$ depends on $\hat{\pi}_{RCAL}(X)$, in contrast with the recent papers of Fan et al. (2021) and Semenova and Chernozhukov (2021), where the propensity score and outcome regression functions are estimated separately.

3.2 Model-assisted confidence intervals 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 (3.6) for (β_0^*, β_1^*), A natural estimator of β^* 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

$$\hat{\mu}^1(z; \hat{m}_1, \hat{\pi}) = \hat{\beta}^T \Phi^\dagger(z), \tag{3.9}$$

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

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

$$g(X) = (f(X)^T, (f(X) \otimes \Phi(Z))^T)^T. \tag{3.10}$$

There may be same functions 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 $f(X)$ can be flexible. For instance, it is possible to include full interactions between V and $\Phi(Z)$ in $f(X)$, namely, $f(X) = (1, V^T, \Phi(Z)^T, (V \otimes \Phi(Z))^T)^T$. Interestingly, this choice of $f(X)$ can be applied to construct doubly robust confidence intervals for $\mu^1(z)$ with discrete Z, as shown in Section 3.3. In addition, it is possible to include more 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 (3.9), allowing for possible model misspecification. Define $\bar{\gamma} = \arg\min_\gamma E\{L_{CAL}(\gamma)\}$ and $\bar{\alpha}_1 = \arg\min_{\alpha_1} E\{L_{WL}(\alpha_1; \bar{\gamma})\}$. Let $\bar{\pi} = \pi(X; \bar{\gamma})$, $\bar{m}_1 = m(X; \bar{\alpha}_1)$ and $\bar{\varphi} = \varphi(Y, T, X; \bar{m}_1, \bar{\pi})$. The purpose of defining $\bar{\gamma}$ and $\bar{\alpha}_1$ is to facilitate

discussing the asymptotic properties of the proposed estimator when the propensity score model or outcome regression model is *misspecified*. By the definitions, $\hat{\gamma}$ and $\hat{\alpha}_1$ alway converge to $\bar{\gamma}$ and $\bar{\alpha}_1$ regardless of whether the working models (2.3) and (2.2) are correctly specified or not. In addition, If model (2.3) is correctly specified, then $\bar{\pi} = \pi^*$; otherwise, $\bar{\pi} \neq \pi^*$. Likewise, if model (2.2) is correctly specified, 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 regularity conditions,

$$\hat{\mu}^1(z; \hat{m}_1, \hat{\pi}) = \hat{\mu}^1(z; \bar{m}_1, \bar{\pi}) + R_n(z), \tag{3.11}$$

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

Proposition 1 (Model-assisted confidence intervals)**.** Suppose that regularity assumptions 1–2 in Supplementary Material hold, $g(X)$ is chosen as in (3.10), and $(|S_{\bar{\gamma}}| + |S_{\bar{\alpha}_1}|)\log(q) = o(n^{1/2})$. If PS model (2.3) is correctly specified, then asymptotic expansion (3.11) 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

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

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

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

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, a model-assisted confidence interval for $\mu^1(z_0)$ is obtained.

For simplicity, the preceding result is stated under model (3.5). If model (3.5) does not hold exactly, then the model-assisted confidence interval remains valid when evaluated against the approximate value $\tilde{\mu}^1(z) = \beta_0^* + \beta_1^{*T} \Phi(z)$ for (β_0^*, β_1^*) defined in (3.6). In Section 5, our simulation study shows that the approximate confidence intervals perform very well.

3.3 Doubly robust confidence intervals of $\mu^1(z)$ for discrete Z

We derive doubly robust confidence intervals for $\mu^1(z)$ with discrete Z when a linear OR model is used. Consider the linear OR working model

$$E(Y \mid T = 1, X) = m_1(X; \alpha_1) = \alpha_1^T g(X) \tag{3.12}$$

and the PS working model (2.3). Remarkably, doubly robust confidence intervals for $\mu^1(z)$ can be obtained merely by including full interactions

between V and $\Phi(Z)$ in $f(X)$, that is, setting

$$f(X) = (1, V^T, \Phi(Z)^T, (V \otimes \Phi(Z))^T)^T, \quad g(X) = (f(X)^T, (f(X) \otimes \Phi(Z))^T)^T. \tag{3.13}$$

We show some 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, $f(X) = g(X) = (1, V^T, Z_1, Z_2, Z_1 Z_2, V^T Z_1, V^T Z_2, V^T Z_1 Z_2)^T$.

It can be seen that the configuration of (3.13) will make the dimension of $f(X)$ the same as $g(X)$. 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 the large sample properties of $\hat{\mu}^1(z_0; \hat{m}_1, \hat{\pi})$ for discrete Z.

Proposition 2 (Doubly robust confidence intervals)**.** Suppose that regularity assumptions 1–2 in Supplementary Material hold, $f(X)$ and $g(X)$ are chosen as in (3.13), and $(|S_{\bar{\gamma}}| + |S_{\bar{\alpha}_1}|) \log(q) = o(n^{1/2})$. Then asymptotic expansion (3.11) is valid. Moreover, if either PS model (2.3) or linear OR model (3.12) is correctly specified, 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 the same as those in Proposition 1. That is, a doubly robust confidence interval for $\mu^1(z_0)$ is obtained.

It is noteworthy that asymptotic expansion (3.11) holds in Proposition 2 without the need for correctly specified PS model (2.3), while such a result 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 mainly focus on estimation of $\mu^1(z)$, we extend it to estimate $\mu^0(z)$ and $\tau(z)$ in Supplementary Material.

For discrete Z, stratified analysis is a routinely used method to estimate $\mu^1(z)$ (Abrevaya et al., 2015). It first splits the sample by Z, and then for each subclass, obtains the estimations of $\hat{m}_1$ and $\hat{\pi}$, 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 elucidate the advantages of the proposed approach. Without of generality, consider the case of binary Z, and take $f(X)=g(X)=(1,V^T,Z,V^TZ)^T$

according to (3.13). We rewrite $f(X)$ as its 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 gives that

$$\tilde{E}[\{T\pi^{-1}(X;\gamma) - 1\}f(X)] = 0, \tag{3.14}$$

$$\tilde{E}[T\{1 - \pi(X;\hat{\gamma})\}\pi^{-1}(X;\hat{\gamma})(Y - \alpha_1^T f(X))f(X)] = 0. \tag{3.15}$$

which are the sample estimating equations for γ and α_1 (up to the lasso penalties in high-dimensional settings). We focus on analyzing equation (3.14), and equation (3.15) can be discussed similarly. Equation (3.14) can be divided into two equations

$$\tilde{E}[\{T\{1 + \exp(-\gamma_0^T f_0(X))\} - 1\}f_0(X)] = 0, \tag{3.16}$$

$$\tilde{E}[\{T\{1 + \exp(-\gamma_1^T f_1(X))\} - 1\}f_1(X)] = 0, \tag{3.17}$$

where $f_0(X) = I\{Z = 0\}(1, V^T)^T$, $f_1(X) = I\{Z = 1\}(1, V^T)^T$, $\gamma = (\gamma_0^T, \gamma_1^T)^T$ that satisfies $\gamma^T f(X) = \gamma_0^T f_0(X) + \gamma_1^T f_1(X)$. (3.16) and (3.17) are the sample estimating equations in stratified analysis. However, if there are multiple categories, stratified analysis is troublesome, especially in high-dimensional settings, where stratified analysis may select different tuning parameters for lasso penalties and different covariates in different 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 selected in different strata.

4. Asymptotic properties

4.1 Heuristic discussion

We delineate basic ideas underlying the construction of the estimators $\hat{\gamma}$ and $\hat{\alpha}_1$, and point out why we need careful specification of $f(X)$ and $g(X)$ in (3.10) or (3.13), such that the estimator $\hat{\mu}^1(z_0; \hat{m}_1, \hat{\pi})$ satisfies asymptotic expansion (3.11) *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 (3.11) to hold, it is sufficient to show that

$$\tilde{E}\{\Phi^\dagger(Z)(\hat{\varphi} - \bar{\varphi})\} = o_p(n^{-1/2}). \tag{4.18}$$

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 (4.18), 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 get the desired convergence rates, the crucial point is that Δ_1 and Δ_2 should be $o_p(1)$, and their corresponding population version should satisfy

$$\frac{\partial}{\partial \alpha_1} E(\Phi^\dagger(Z)\varphi(Y,T,X;\alpha_1,\gamma))\Big|_{(\alpha_1,\gamma)=(\bar{\alpha}_1,\bar{\gamma})} = 0, \tag{4.19}$$

$$\frac{\partial}{\partial \gamma} E(\Phi^\dagger(Z)\varphi(Y,T,X;\alpha_1,\gamma))\Big|_{(\alpha_1,\gamma)=(\bar{\alpha}_1,\bar{\gamma})} = 0. \tag{4.20}$$

Hence a natural approach is to solve (4.19) and (4.20) being in low-dimensional settings and add lasso penalties in high-dimensional settings. Nevertheless, this method will encounter with a basic problem: there are more equations than parameters. It is easy to see that (4.19) includes $(K+1)(q+1)$ equations and (4.20) contains $(K+1)(p+1)$ equations, while the dimensions of γ and α_1 are $p+1$ and $q+1$, respectively. Therefore, the coefficients γ and α_1 cannot be identified by solving (4.19) and (4.20) without further consideration. Fortunately, this difficulty can be overcome by simply a careful specification of $f(X)$ and $g(X)$.

Specifically, with PS model (2.3) and linear OR model (3.12), Δ_1 and Δ_2 reduces to

$$\Delta_1 = \tilde{E}\{(T\bar{\pi}^{-1}(X) - 1)g(X) \otimes \Phi^\dagger(Z)\},$$

$$\Delta_2 = \tilde{E}\{T(1-\bar{\pi}(X))\bar{\pi}^{-1}(X)(Y - \bar{\alpha}_1^T g(X))f(X) \otimes \Phi^\dagger(Z)\}.$$

If $g(X)$ satisfies the form of (3.10), then according to the definition of $\bar{\alpha}_1$, (4.20) holds regardless of whether the OR model is specified correctly. In addition, (4.19) holds provided that PS model (2.3) is correctly specified but OR model (3.12) may be misspecified, which elucidates why Proposition 1 can be derived. Furthermore, if $f(X)$ and $g(X)$ are specified as in (3.13), then Δ_1 and Δ_2 have a simpler form with discrete Z:

$$\Delta_1 = \tilde{E}\{(T\bar{\pi}^{-1}(X) - 1)f(X)\},$$

$$\Delta_2 = \tilde{E}\{T(1 - \bar{\pi}(X))\bar{\pi}^{-1}(X)(Y - \bar{\alpha}_1^T g(X))g(X)\},$$

which exactly are the gradients of $L_{CAL}(\bar{\gamma})$ and $L_{WL}(\alpha_1; \bar{\gamma})$, respectively. In this case, (4.19) and (4.20) hold just by the definition of $\bar{\gamma}$ and $\bar{\alpha}_1$, irrespective of the model specifications for PS and OR, 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, (λ_0, λ_1) are set as $\lambda_0 = [\log\{(1+p)/\epsilon\}/n]^{1/2}$, $\lambda_1 = [\log\{(1+q)/\epsilon\}/n]^{1/2}$ $(\geq \lambda_0)$, 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) showed that the convergence rates for $(\hat{\gamma}, \hat{\alpha}_1)$, $||\hat{\gamma} - \bar{\gamma}||_1 = O_p(1) \cdot |S_{\bar{\gamma}}|\{\log(p)/n\}^{1/2}$,

$||\hat{\alpha}_1 - \bar{\alpha}_1||_1 = O_p(1) \cdot (|S_{\bar{\gamma}}| + |S_{\bar{\alpha}_1}|)\{\log(q)/n\}^{1/2}$. The following Theorem 1 presents the large sample properties of $\hat{\mu}^1(z; \hat{m}_1, \hat{\pi})$ for discrete Z.

Theorem 1 (doubly robust confidence intervals)**.** *Suppose that regularity assumptions 1–2 in Supplementary Material hold, if linear OR model (3.12) is used, $f(X)$ and $g(X)$ are specified as in (3.13), then for any given z_0 of discrete Z,*

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

$$\left|\hat{\mu}^1(z_0; \hat{m}_1, \hat{\pi}) - \hat{\mu}^1(z_0; \bar{m}_1, \bar{\pi})\right| \leq M_0(|S_{\bar{\gamma}}|\lambda_0 + |S_{\bar{\alpha}_1}|\lambda_1)\lambda_1, \tag{4.21}$$

where c_0 and M_0 are positive constants.

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

$$\hat{V}(z_0) - V(z_0) = o_p(1), \tag{4.22}$$

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

Since $\hat{\mu}^1(z; \bar{m}_1, \bar{\pi})$ 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, to obtain a valid confidence interval, it requires the asymptotic expansion (3.11) to hold, which implies that $(|S_{\bar{\gamma}}|\lambda_0 + |S_{\bar{\alpha}_1}|\lambda_1)\lambda_1 = o(n^{-1/2})$, namely,

$(|S_{\bar{\gamma}}| + |S_{\bar{\alpha}_1}|)\log(q) = o(n^{1/2})$. In summary, Theorem 1 shows that, for discrete Z with linear OR model (3.12) and specification of $f(X)$ and $g(X)$ as in (3.13), the proposed method obtains both doubly point estimators and doubly confidence intervals for $\mu^1(z)$, provided that $(|S_{\bar{\gamma}}| + |S_{\bar{\alpha}_1}|)\log(q) = o(n^{1/2})$, which leads to the Proposition 2. Similar to Theorem 1, the following Theorem 2 implies the results presented in Proposition 1.

Theorem 2 (Model-assisted confidence intervals)**.** *Suppose that regularity assumptions 1–2 in Supplementary Material hold, if OR model (2.2) is used, $g(X)$ is specified as in (3.10), and PS model (2.3) is correctly specified, then for a given value z_0 of discrete/continuous Z,*

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

$$\left|\hat{\mu}^1(z_0;\hat{m}_1,\hat{\pi}) - \hat{\mu}^1(z_0;\bar{m}_1,\bar{\pi})\right| \leq M_1(|S_{\bar{\gamma}}|\lambda_0 + |S_{\bar{\alpha}_1}|\lambda_1)\lambda_1, \tag{4.23}$$

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 $1-(c_0+12)\epsilon$,

$$\hat{V}(z_0) - V(z_0) = o_p(1), \tag{4.24}$$

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

Extensive simulation studies are carried out to evaluate the finite sample performance of the proposed methods. We consider three scenarios of Z: binary variable Z, continuous variable Z, and Z consists of two binary variables Z_1 and Z_2. The RCAL estimation for the PS model and RWL estimation for OR model can be implemented by using R package **RCAL** (Tan, 2019), and the corresponding tuning parameters are determined via using 5-fold cross-validations. Throughout this simulation, the data generating processes of covariates are as follows: $V = (V_1, ..., V_d) \sim N(0, \Sigma)$ with $\Sigma_{j,k} = 2^{-|j-k|}$ for $1 \leq j, k \leq d$, 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 independent and identically distributed 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 i-th element of V.

Discrete Z. We consider three scenarios (C1)-(C3) with binary Z, which can help to assess the doubly robust properties for both point estimators and confidence intervals.

(C1) Z is binary variable, $P(T = 1|X) = \{1+\exp(-(Z, V_1, V_2, V_3, V_4)^T\gamma)\}^{-1}$,

$Y^1 = 1 + Z + \sum_{i=1}^4 \{V_i Z + 2V_i(1-Z)\} + \epsilon$.

(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: (C1), both PS and OR models are correctly specified; (C2), PS model is correctly specified, but OR model is misspecified; (C3), PS model is misspecified, but OR model is correctly specified.

The true curve of $\mu^1(z)$ is $1+z$ for all 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 1000 replicates with sample size $n = 500$. Bias and Var are the Monte Carlo bias and variance over the 1000 simulations of the points estimates. EVar is the mean of the variance estimates. CP90 and CP95 are the coverage proportions of the 90% and 95% confidence intervals, respectively. Table 1 summarizes the results of $\hat{\mu}^1(z)$ for scenarios (C1)-(C3).

As shown in Table 1, for all the cases (C1)-(C3), the Bias is small, $\sqrt{\text{EVar}}$ is close to $\sqrt{\text{Var}}$ and the coverage proportions CP90 and CP95 are around the nominal levels 0.90 and 0.95, respectively. Because case (C2) involves a misspecified OR model and case (C3) involves a misspecified

Table 1: Estimations of $\mu^1(z)$ for 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 $p + 1$.

PS model, the results of cases (C2) and (C3) justify that both the point estimators and confidence intervals are doubly robust.

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

(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 PS and OR models are correctly specified.

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

We set $f(X) = (1, V^T, Z)^T$, $g(X)$ is specified as in (3.10), and let $\Phi(Z)$ be cubic spline basis functions with three knots selected by the 25%, 50% and 75% sample quantiles of Z, which can be implemented using R package **gam** (Hastie, 2018). Since 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 shows the numerical results of $\hat{\mu}^1(z)$ for cases (C4) and (C5). Both of them have a similar performance to those of discrete Z.

We compare the proposed method with competing AIPW methods of

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

$\hat{\mu}^1(z)$	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.

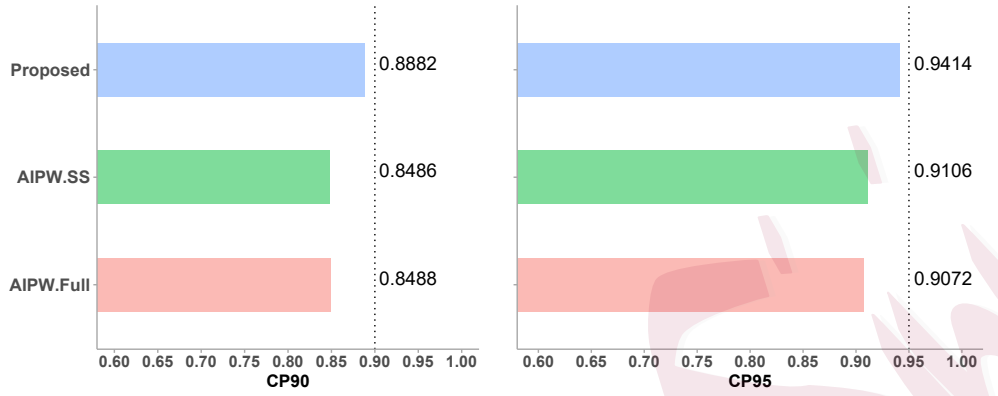

Figure 1: Average values of CP90 and CP95 at five representative points based on 1000 simulations, where CP90 and CP95 are the coverage proportions of the 90% and 95% confidence intervals, AIPW.Full means Fan et al. (2021)'s AIPW method with full sample (without sample splitting), and AIPW.SS represents Fan et al. (2021)'s and Zimmert and Lechner (2019)'s AIPW method with four-fold cross-fitting (sample splitting).

Fan et al. (2021) and Zimmert and Lechner (2019) discussed in Section 2.2 for continuous Z. The implement 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 confidence intervals, while the proposed method *does*.

The setups for all preceding numeric results are in exact sparsity set-

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

6. Application

Psoriasis is a chronic immune-mediated inflammatory disease that can damage patients quality of life severely and increase the burden on society substantially (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 (TNFα), interleukin-17 (IL-17) and interleukin-23 (IL-23))(Park et al., 2005). The new generation of biologics IL-17 inhibitors (secukinumab, ixekizumab, and brodalumab) and IL-23 inhibitors (guselkumab, risankizumab, and tildrakizumab) were developed successively. In 2020, Secukinumab has been added into the national drug reimbursement list (NDRL) of China for treat-

ing psoriasis. Since then clinical use of secukinumab increased significantly.

In China, biologics are used only for patients 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 aim to explore the heterogeneous effects of biologics versus conventional therapies across different subpopulations.

6.1 Data description

Data were collected from Psoriasis Center Data Platform (PCDP), which was led by the National Clinical Research Center for Skin and Immune Disease and covered patients of 237 tertiary hospitals in about 100 cities in the mainland of China. In this study, data is restricted to patients who enrolled from Sep 2020 to Sep 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 IL-17 inhibitor, Secukinumab) or the conventional therapies (topic drugs, systemic medicines, or phototherapy),

and eliminate patients with other treatments. The use of biologics can be divided into the induction period and the maintained period, we assess the treatment heterogeneity in the maintained period, i.e., excluding patients with follow-up time less than 4 weeks. The final analytical dataset with a total of 2356 samples, where 708 (30.05%) use biologics and 1648 (69.95%) do not.

The clinical benefit is measured by the improvement on Psoriasis Area and Severity Index (PASI). In this study, the outcome variable Y is the indicator of 80% or more improvement from baseline PASI ($Y = \texttt{PASI 80}$) in the first follow-up visits. The exposure variable $T = 1$ represents that the patient was treated with IL-inhibitor biologics and $T = 0$ means the conventional therapies. The covariates X include the patients demographics, clinical characteristics, and theirs interactions. Description of these variables were given in Table S3 of Supplementary Material. The subpopulations of interest are defined by covariates Z, which are taken to be `baseline PASI`, `baseline DLQI`, `Age`, `Employment`, `Martial status`, `Education`, `Insurance` and `Sex`, respectively. The first three variables are continuous and the last five are binary. Since there are only a few samples available at extremely high values of `baseline PASI`, we set the `baseline`

`PASI` 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 $f(X) = (1, V^T, \Phi(Z)^T)^T$, $g(X) = (1, V^T, \Phi(Z)^T, (V \otimes \Phi(Z))^T, (\Phi(Z) \otimes \Phi(Z))^T)^T$ for continuous Z, where V is all the covariates excluding variable Z. All variables in $f(X)$ and $g(X)$ are standardized to have sample mean 0 and sample variance 1. As done in simulation studies, the lasso tuning parameter λ is selected by five-fold cross validation.

6.2 Results

Discrete Z. Figure 2 presents the estimated causal effects of biologics on the improvement of PASI conditional on different binary variables Z. The result shows that higher education and free or commercial insurance groups have larger biologics benefits than the corresponding lower education and general government funded insurance groups, respectively. In addition, the causal effects have no large difference among subgroups with different values of `Employment`, `Martial status`, and `Sex`.

Continuous Z. To estimate a CSTE curve when Z is continuous, we apply the propose method using cubic spline to approximate $\tau(z)$ and find the optimal number of knots by using grid search with Akaike information criterion (AIC) (Akaike, 1974) and Bayesian information criterion (BIC)

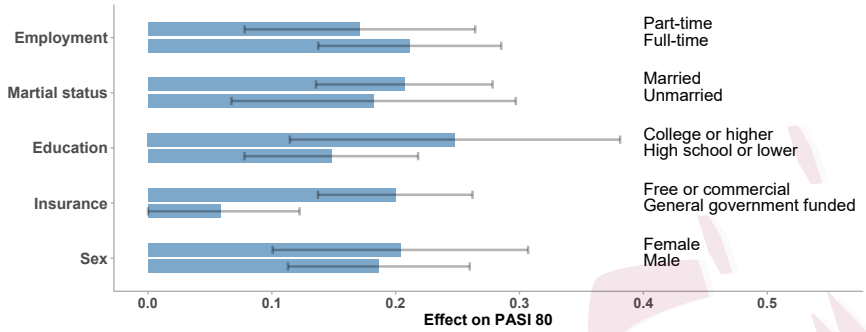

Figure 2: Estimated CSTE (the height of the bar plot) and the associated 95% CI (error bar) for `PASI 80` with discrete Z.

(Schwarz, 2005). Specifically, we first fix the number of knots is 3, that is, $f(X) = (1, V^T, \Phi(Z)^T)^T$, $g(X) = (1, V^T, \Phi(Z)^T, (V \otimes \Phi(Z))^T, (\Phi(Z) \otimes \Phi(Z))^T)^T$ with $\Phi(Z)$ being cubic spline basis functions with 3 knots. Then we use $f(X)$ and $g(X)$ to estimate propensity score and outcome regression functions. Finally, we conduct least squares by regressing $\varphi(Y, T, X; \hat{m}_1, \hat{\pi}) - \varphi(Y, 1-T, X; \hat{m}_0, 1-\hat{\pi}_0)$ on $\tilde{\Phi}(Z)$ to get the values of where $\tilde{\Phi}(Z)$ is cubic spline basis functions with number of knots ranging from 1 to 10. The corresponding results are given in Table S4 in Supplementary Material.

Figure 3 displays the estimated CSTE curves and corresponding 95% confidence intervals (CI) for different continuous Z. It indicates that biologics have a positive effect over conventional therapies and heterogeneity is ubiquitous in different subpopulations. As the `baseline PASI` increased

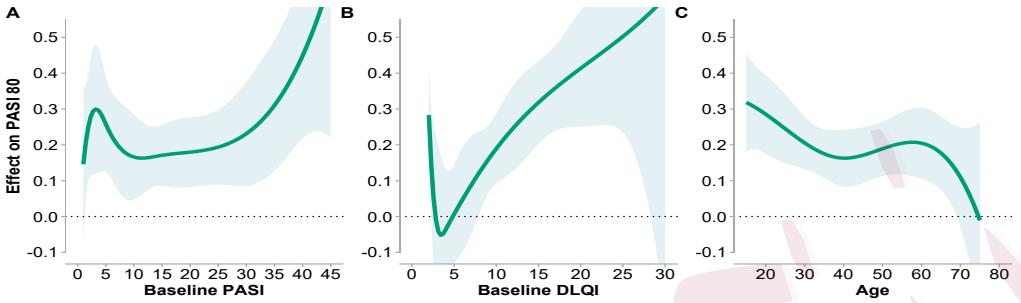

Figure 3: Estimated CSTE and the associated 95% CI for `PASI 80` with continuous Z.

from 0 to 5, the relative advantage of biological agents over conventional therapy increased; When the score exceeds 5 points, the CSTE is a V-shaped curve, with a trough near the PASI value of 10 (Figure 1A). Our results indicate that biologics also is effective in mild-to-moderate psoriasis (`baseline PASI` ≤ 10). A higher value 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 1B, the CSTE is a V-shaped curve as the value of DLQI increases. In addition, the effect first decreases slightly with age ranging from 12-30, then flattenes from 30-60, and then decreases rapidly from 60-75. Overall, patients with larger age indicate higher benefits of biologics.

7. Discussion

This article develops new methods to obtain both doubly robust point estimators and model-assisted confidence intervals for conditional average treatment effects in high-dimensional settings. In addition, with a linear OR model and discrete Z, the confidence intervals are also doubly robust. Theoretical properties are established for the proposed methods with different data types of outcome Y and covariates Z, and the corresponding variances can be estimated by a sandwich method.

Further work is desired to extend our method and theory by relaxing the parametric structural model (3.5) to be nonparametric subject to smoothness conditions, while allowing the basis functions $\Phi(z)$ to be data-adaptively chosen, instead of pre-specified. Another interesting question is that whether doubly robust confidence intervals can be derived for continuous Z. To deal with this question, a possible approach is to discretize Z. For example, for two knots $t_1 < t_2$, we can discretize Z as $(Z_1, Z_2) = (I\{t_1 < Z \leq t_2\}, I\{Z > t_2\})$ 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 desired doubly robust confidence intervals in the discretized model $\mu^1(Z) = E(Y^1|Z) = \beta_0 + (Z_1, Z_2)\beta_1$. The

method can be easily extended to multiple knots, corresponding to piecewise constant model for $\mu^1(z)$. Then various theoretical questions need to be investigated. For example, it is interesting to study convergence and whether doubly confidence intervals can be achieved, depending on the number of knots used.

Another extension is to consider the case that Z is 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 et al., 2021). It is interesting to study how to incorporate such strategies in future research.

Supplementary Materials

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

Acknowledgements

The authors thank AE and anonymous reviewers for their helpful comments and valuable suggestions on this article. This research was supported by

the state key research program [No. 2021YFF0901400] and the National Natural Science Foundation of China [No.11971064 and 12171374].

References

Abrevaya, J., Y. C. Hus, and R. P. Lieli (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., G. W. Imbens, and S. Wager (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., V. Chernozhukov, I. Fernández-Val, and C. Hansen (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica 85*, 233–298.

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

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

Chakraborty, B. and E. E. Moodie (2013). *Statistical methods for dynamic treatment regimes*. Springer, New York.

Chen, A.-J., X.-H. Gao, H. Gu, and et al. (2020). Chinese experts consensus on biologic therapy

for psoriasis. *International Journal of Dermatology and Venereology 3*, 76–85.

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

Chernozhukov, V., I. FernándezVal, and Y. Luo (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 S. Vansteelandt (2020). Inference for treatment effect parameters in potentially misspecified high-dimensional models. *Biometrika 00*, 1–14.

Fan, Q., Y. C. Hsu, R. P. Lieli, and Y. Zhang (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 J. N. Barker (2007). Pathogenesis and clinical features of psoriasis. *The Lancet 370*, 263–271.

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

REFERENCES

Guo, W., X. H. Zhou, and S. Ma (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*. https://CRAN.R-project.org/package=gam.

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

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

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

Menter, A., B. E. Strober, D. H. Kaplan, and 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., S. Peng, and K. Imai (2020). Robust estimation of causal effects via a high-dimensional covariate balancing propensity score. *Biometrika 107*, 533–554.

Park, H., Z. Li, X. O. Yang, and 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 B. Silverman (2005). *Functional Data Analysis* (Second ed.). Springer.

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

Rosenbaum, P. R. and D. B. Rubin (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* (Third ed.). Cambridge University Press.

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

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

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

Sun, B. and Z. Tan (2021). High-dimensional model-assisted inference for local average treat-

ment effects with instrumental variables. *Journal of Business and Economic Statistics*, (To Appear).

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. (2019). *RCAL: Regularized calibrated estimation.* 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., A. A. Alizadeh, A. J. Gentles, and R. Tibshirani (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 58*, 267–288.

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

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