Doubly-robust evaluation of high-dimensional surrogate markers

Denis Agniel, Boris Hejblum, Rodolphe Thiébaut, Layla Parast

Doubly-robust evaluation of high-dimensional surrogate markers

Denis Agniel, Layla Parast, Boris Hejblum

December 4, 2020

Abstract

When evaluating the effectiveness of a treatment, policy, or intervention, the desired measure of effectiveness may be expensive to collect, not routinely available, or may take a long time to occur. In these cases, it is sometimes possible to identify a surrogate outcome that can more easily/quickly/cheaply capture the effect of interest. Theory and methods for evaluating the strength of surrogate markers have been well studied in the context of a single surrogate marker measured in the course of a randomized clinical study. However, methods are lacking for quantifying the utility of surrogate markers when the dimension of the surrogate grows and/or when study data are observational. We propose an efficient nonparametric method for evaluating high-dimensional surrogate markers in studies where the treatment need not be randomized. Our approach draws on a connection between quantifying the utility of a surrogate marker and the most fundamental tools of causal inference – namely, methods for estimating the average treatment effect. We show that recently developed methods for incorporating machine learning methods into the estimation of average treatment effects can be used for evaluating surrogate markers. This allows us to derive limiting asymptotic distributions for key quantities, and we demonstrate their good performance in simulation.

1 Introduction

When evaluating the effectiveness of a treatment, policy, or intervention, the desired measure of effectiveness may be expensive to collect, not routinely available, or may take a long time to occur. In these cases, it is sometimes possible to identify a surrogate outcome that can more easily/quickly/cheaply capture the effect of interest. For example, when evaluating an intervention designed to delay dementia onset, the time required to observe enough dementia diagnoses is often very long and surrogates that have been considered for intervention evaluation include mild cognitive impairment, adiponectin levels, neuroimages, amyloid plaques and neurofibrillary tangles [27, 11, 30, 12, 13]. Similarly, in studies evaluating treatments to prevent diabetes, surrogate measures for diabetes onset have included changes in body weight, fasting plasma glucose and hemoglobin A1c [4, 9, 6, 18].

Theory and methods for evaluating the strength of surrogate markers have been well studied in the context of a single surrogate marker measured in the course of a randomized clinical study. In particular, robust nonparametric methods have been proposed in [21, 20]. However, fully nonparametric methods are not available or not reliable when the numbers of markers is more than one or two. In these cases, an initial model may be used to reduce the dimensionality of the

surrogate [1, 17, 20], but an inappropriate initial model can produce badly biased estimates of the utility of the surrogate. Additionally, as the dimension of the surrogate nears or exceeds the sample size, initial parametric models may not have adequate properties or might not even be possible. Furthermore, assessments of surrogate markers are not solely limited to studies of randomized treatment [7, 16]. In fact, using surrogate outcomes based on complex observational data is finding purchase in all corners of science [33].

In this work, we propose an efficient nonparametric method for evaluating high-dimensional surrogate markers in studies where the treatment need not be randomized. Our approach draws on a connection between quantifying the utility of a surrogate marker and the most fundamental tools of causal inference – namely, methods for estimating the average treatment effect. We take advantage of state-of-the-art methods for incorporating flexible machine learning and/or sparse high-dimensional models into estimation of treatment effects via the cross-fitting approach of [5]. Cross-fitting allows us to derive asymptotic distributions and compute confidence intervals for key quantities while putting minimal restrictions on what types of estimators are used. Furthermore, our estimator and results may be used regardless of the dimension of the surrogates.

The structure of the rest of the paper is as follows. In Section 2, we lay out notation and the setting in which we are working, and we motivate the importance of evaluating surrogate markers in light of recent advances in estimating surrogates from high-dimensional data. In Section 3, we detail assumptions necessary for identifying and interpreting parameters of interest, as well as our approach to estimating the quality of a surrogate marker. We derive the asymptotic behavior of our estimator in Section 4 and discuss variance estimation and inference. We evaluate the performance of the proposed approach using a simulation study in Section 5. We give final remarks and draw connections to other methods for evaluating surrogate markers in Section 6.

2 Evaluation of Surrogate Markers

2.1 Notation

Let A denote a binary treatment, and let the primary outcome of the study be Y. Let there be a vector $\mathbf{S}$ of potential surrogate information, and let $\mathbf{X}$ be a vector of pre-treatment covariates. The primary quantity of interest is the treatment effect on the outcome

$$\Delta = \mathbb{E}\{Y^{(1)} - Y^{(0)}\}$$

where $Y^{(a)}$ is the potential or counterfactual outcome that would have been observed if the observed value of treatment were $A = a$, possibly contrary to fact. Similarly let $\mathbf{S}^{(a)}$ be the potential/counterfactual value the vector of surrogates would take if $A = a$. We desire to approximate Δ via a difference in some function of the surrogates $\mathbf{S}$ and covariates $\mathbf{X}$. Let $\psi_a(\mathbf{X}, \mathbf{S})$ be such a scalar function that is here indexed by treatment a but need not depend on treatment. Let the data observed in the current study be $\mathcal{O} = (\mathbf{X}_i, A_i, \mathbf{S}_i, Y_i)_{i=1,\ldots,n}$, n iid realizations of $\mathbf{O} = (\mathbf{X}, A, \mathbf{S}, Y)$. And let $\mathcal{F} = (\check{\mathbf{X}}_i, \check{A}_i, \check{\mathbf{S}}_i)_{i=1,\ldots,n^*}$ be n^* iid realization of $\mathbf{O}$ in a future study. The function $\psi_a(\mathbf{x}, \mathbf{s})$ is useful and is of particular interest because for a future study, one could then use $\psi_i = \psi_1(\check{\mathbf{X}}_i, \check{\mathbf{S}}_i)I\{\check{A}_i = 1\} + \psi_0(\check{\mathbf{X}}_i, \check{\mathbf{S}}_i)I\{\check{A}_i = 0\}$ as the outcome

instead of Y. The true treatment effect Δ could be approximated by estimating the effect of $\check{A}_i$ on ψ_i conditional on $\check{\mathbf{X}}_i$ via regression, weighting, matching, or any other reasonable approach.

2.2 Importance of quantifying surrogate strength

Understanding the strength of a potential surrogate is important for many reasons. In practice, information about surrogate strength will inform decisions regarding whether to measure the surrogate in a future study (especially if it is costly or invasive to measure) and/or whether to use the surrogate to assess treatment effectiveness in a future study. In addition, a number of novel statistical methods for using surrogates in future studies can only be applied when the surrogate is strong. For example, Parast *et al.*[19] propose a robust nonparametric procedure to test for a treatment effect using surrogate marker information measured prior to the end of the study in a time-to-event outcome setting, but they rely on the assumption that the surrogate is sufficiently strong. As another example, Price *et al.*[23] propose constructing a function $\psi_a(\cdot)$ of the surrogates that leads to some optimality properties, but it can be shown that this method should only be used with a strong surrogate.

Specifically, they construct an optimal $\psi_a(\cdot)$ in terms of minimizing the following mean squared error (MSE)

$$\mathcal{M}_\psi = \mathbb{E}\left[e(\mathbf{X})\left\{Y^{(1)} - \psi_1(\mathbf{X}, \mathbf{S}^{(1)})\right\}^2\right] + \mathbb{E}\left[\{1 - e(\mathbf{X})\}\left\{Y^{(0)} - \psi_0(\mathbf{X}, \mathbf{S}^{(0)})\right\}^2\right], \qquad e(\mathbf{x}) = P(A = 1|\mathbf{X}) \tag{1}$$

under the constraint that it satisfies the so-called Prentice definition, i.e.,

$$E(Y^{(1)} - Y^{(0)}) = 0 \text{ if and only if } E\left\{\psi_1(\mathbf{X}, \mathbf{S}^{(1)}) - \psi_0(\mathbf{X}, \mathbf{S}^{(0)})\right\} = 0. \tag{2}$$

The optimal transformations in this case are shown to be $\psi_a(\mathbf{x}, \mathbf{s}) = \mathbb{E}\{Y^{(a)}|\mathbf{X} = \mathbf{x}, \mathbf{S}^{(a)} = \mathbf{s}\}$. In addition to this appealing optimality property, this proposal also resolves the so-called "surrogate paradox." The paradox states that the treatment could have a positive causal effect on a univariate surrogate S which could have a positive correlation with the outcome, but the treatment could yet have a negative effect on the outcome. The Price surrogate resolves the surrogate paradox by definition. The treatment effect on the surrogate $\mathbb{E}(\check{\Delta})$ is by definition equivalent to the treatment effect on the outcome: $\mathbb{E}(\check{\Delta}) = \mathbb{E}\{\psi_1(\mathbf{X}, \mathbf{S}^{(1)})\} - \mathbb{E}\{\psi_0(\mathbf{X}, \mathbf{S}^{(0)})\} = \Delta$.

However, the Price surrogate's resolution of the surrogate paradox is in some sense too good. The surrogate paradox is thus resolved for every potential set of surrogates, good and bad. Even if the surrogates $\mathbf{S}$ are completely unrelated to the outcome Y, $\mathbb{E}(\check{\Delta}) = \Delta$. In fact, the power to detect a treatment effect in a future study using $\check{\Delta}$ actually increases as the surrogate becomes weaker (explaining less of the treatment effect), if the Prentice definition does not hold. Consider the toy example depicted in Figure 1 (see Appendix 7.1 for details of the data generation). As the strength of the surrogate increases, $\psi_a(\cdot)$ becomes more like the true outcome $Y_i^{(a)}$. As the surrogate becomes weaker, $\psi_a(\cdot)$ becomes more like $\bar{Y}_a$ the mean outcome in treatment group $A = a$ from the first study. This means that a weak surrogate ensures that the treatment effect in the second study will be identical to the treatment effect in the first study because the distribution of $\psi_a(\check{\mathbf{X}}_i, \check{\mathbf{S}}_i)$ will cluster closely around the estimate of $E(Y^{(a)})$ from the first study (see specifically,

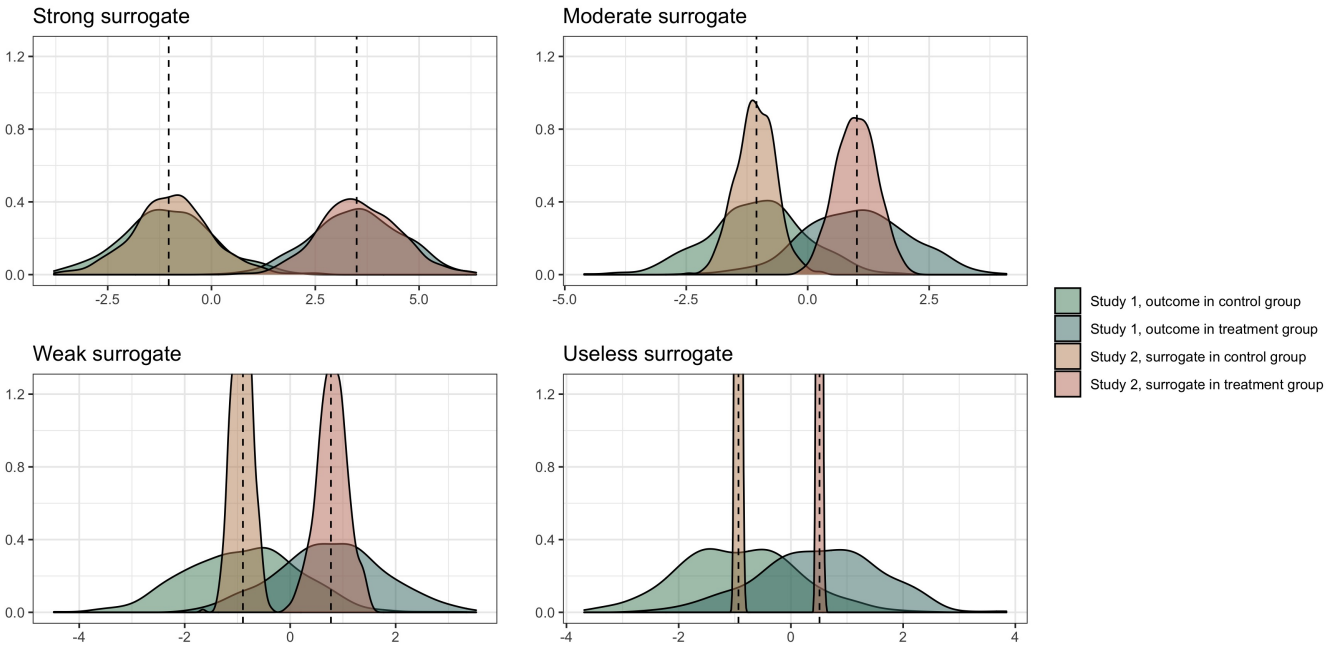

Figure 1: Distribution of outcomes Y in an initial study (Study 1) and surrogate functions (approximating the outcome) ψ_1 (surrogate in treatment group) and ψ_0 (surrogate in control group) in a future study (Study 2) based on simulated data. Vertical dotted lines correspond to the means of the treatment groups in the first study. See Appendix 7.1 for simulation details. Surrogate functions ψ_0, ψ_1 were estimated via a correctly specified linear regression in the first study and then applied to the second study population. When the surrogate was strong, the distribution of the surrogate in the second study was quite close to the distribution of the outcome in the first study. However, when the surrogate was weak, the distribution of the surrogate in the second study clustered around the group mean from the first study.

the fourth panel of Figure 7.1). In this scenario, the second study is providing no new information about the treatment. Thus, the Price surrogate should only be used with a strong $\mathbf{S}$, and its use with a weak or possibly even moderately strong set of surrogates may not be advisable.

Therefore, for both decision-making purposes and to use statistical methods that take advantage of strong surrogates, methods are needed to rigorously quantify the strength of the proposed set of surrogates before using (functions of) the surrogates in practice. In the following section, we propose a model-free method for estimating the strength of a possibly high-dimensional set of surrogates in observational studies.

2.3 Proposed method to evaluate surrogate strength

In this section we propose a general method for evaluating the usefulness of a set of surrogates. This can be used to evaluate the suitability of the Price surrogate but can also be used with other surrogate transformations, and it is appropriate for $\mathbf{S}$ of any dimension. To evaluate the surrogates' usefulness, we follow [21, 20, 1] in evaluating the *residual treatment effect*

$$\Delta_{\mathbf{S}} = \mathbb{E}\left\{\psi_1(\mathbf{X}, \mathbf{S}) - \psi_0(\mathbf{X}, \mathbf{S})\right\}, \tag{3}$$

where $\psi_a(\mathbf{x}, \mathbf{s}) = E(Y^{(a)}|\mathbf{X} = \mathbf{x}, \mathbf{S}^{(a)} = \mathbf{s})$, as defined above. This quantity corresponds to the treatment effect conditional on the distribution of $\mathbf{S}^{(0)}$ and $\mathbf{S}^{(1)}$ both being equal to the distribution of $\mathbf{S}$ (all conditional on $\mathbf{X}$). If $\mathbf{S}^{(0)}, \mathbf{S}^{(1)}$ are independent of $Y^{(0)}, Y^{(1)}$ conditional on $\mathbf{X}$, then $\Delta_{\mathbf{S}} = \Delta$. In contrast, if all of the treatment effect can be attributed to $\mathbf{S}$, then $\Delta_{\mathbf{S}}$ should be close to 0. We may use the residual treatment effect to quantify *how much* of the treatment effect the surrogates account for via the *proportion of treatment effect explained* (PTE)

$$R_{\mathbf{S}} = \frac{\Delta - \Delta_{\mathbf{S}}}{\Delta} = 1 - \frac{\Delta_{\mathbf{S}}}{\Delta}. \tag{4}$$

The quantity $\Delta_{\mathbf{S}}$ is defined such that it aims to quantify the effect of the treatment on the outcome that is *not* captured by the effect of the treatment on the surrogate, by forcing the conditional distribution of the potential surrogate in both treatment groups to be the same. Based on this definition, $\Delta - \Delta_{\mathbf{S}}$ then parallels the natural indirect effect quantity [10, 32] which is often used to assess the proportion mediated in a mediation analysis. However, there is an important distinction between assessing a variable (or variables) for surrogacy versus mediation. As described by [32], the aim of identifying a mediator is determining whether the effect of treatment operates through the mediator itself e.g., through some biological pathway. Often, a good surrogate marker is similarly conceptualized as a variable through which the treatment operates, but this is not necessarily required; a variable can be a good surrogate if it captures the treatment effect on the outcome, even if the treatment effect does not operate through the variable itself.

Remark 1. *While R_S is easy to understand and interpret, and it has analogues in the mediation literature, it is not guaranteed to lie between 0 and 1. We give conditions in Section 3.2 that guarantee that $R_S \in (0, 1)$. In situations when these conditions are not or are unlikely to be met, one may desire a secondary definition of the PTE that can be*

estimated in all situations similar to [34]. Conceptually, here we would consider a good surrogate to be one whereby the surrogate transformation substantially improves the MSE for predicting $Y^{(a)}$ over a null transformation that does not use $\mathbf{S}$*. The MSE of the null transformation would be defined as:*

$$\mathcal{M}_0 = \mathbb{E}\left(e(\mathbf{X})\left[Y^{(1)} - m_1\{\mathbf{X}\}\right]^2\right) + \mathbb{E}\left(\{1-e(\mathbf{X})\}\left[Y^{(0)} - m_0\{\mathbf{X}\}\right]^2\right)$$

where $m_a(\mathbf{x}) = \mathbb{E}(Y^{(a)}|\mathbf{X}=\mathbf{x})$*. The secondary MSE-based definition of PTE could then be defined as the proportion of the null MSE that is reduced by controlling for the surrogates:*

$$R_M = \frac{\mathcal{M}_0 - \mathcal{M}_\psi}{\mathcal{M}_0} = 1 - \mathcal{M}_\psi/\mathcal{M}_0 \tag{5}$$

where $\mathcal{M}_\psi$ *is defined in* (1)*.*

3 Assumptions and Estimation

3.1 Identifying assumptions

We make the three typical assumptions of treatment effect estimation: consistency, positivity, and no unmeasured confounding. Specifically, we assume that the observed values of $\mathbf{S}$ and Y when $A=a$ are identical to the counterfactuals $\mathbf{S}^{(a)}$ and $Y^{(a)}$ such that $\mathbf{S} = \mathbf{S}^{(1)}A + \mathbf{S}^{(0)}(1-A)$ and $Y = Y^{(1)}A + Y^{(0)}(1-A)$. We furthermore assume that $\mathbf{X}$ contains all confounders of the effects of A on the surrogates and the outcome, such that the treatment A is as good as randomized conditional on the covariates $\mathbf{X}$

$$\{\mathbf{S}^{(0)}, \mathbf{S}^{(1)}, Y^{(0)}, Y^{(1)}\} \perp\!\!\!\perp A \Big| \mathbf{X}. \tag{6}$$

In addition, we assume two forms of positivity, which ensures that individuals in the two study arms are not too different from one another. First, the usual positive probability of receiving either treatment for some $\epsilon_1 > 0$

$$P\{\epsilon_1 < e(\mathbf{X}) < 1-\epsilon_1\} = 1,$$

and a related assumption that further conditions on the surrogates

$$P\{\epsilon_2 < \pi(\mathbf{X}, \mathbf{S}) < 1-\epsilon_2\} = 1,$$

for some $\epsilon_2 > 0$ where $\pi(\mathbf{S}, \mathbf{X}) = P(A=1|\mathbf{S}, \mathbf{X})$ is the *surrogate score* – akin to the propensity score [25] – which has been proposed for a different purpose in [2]. Notably, because $\pi(\mathbf{x}, \mathbf{s}) = \frac{P(\mathbf{S}=\mathbf{s}|\mathbf{X}=\mathbf{x}, A=1)e(\mathbf{x})}{P(\mathbf{S}=\mathbf{s}|\mathbf{X}=\mathbf{x})} = \frac{P(\mathbf{S}^{(1)}=\mathbf{s}|\mathbf{X}=\mathbf{x})e(\mathbf{x})}{P(\mathbf{S}^{(1)}=\mathbf{s}|\mathbf{X}=\mathbf{x})e(\mathbf{x}) + P(\mathbf{S}^{(0)}=\mathbf{s}|\mathbf{X}=\mathbf{x})\{1-e(\mathbf{x})\}}$, these two positivity conditions ensure that the conditional distribution of the counterfactual surrogates under treatment and control cannot be too different from one another, i.e., ensuring overlap.

We additionally give regularity conditions facilitating estimation in Appendix 7.2.

3.2 Required assumptions to ensure interpretation of $R_{\mathbf{S}}$ as a proportion

The interpretation of $R_{\mathbf{S}}$ as the PTE depends on it actually being a proportion, lying between 0 and 1. We adapt conditions in [35] and [1] which ensure that $0 \leq R_{\mathbf{S}} \leq 1$. These conditions are as follows (assuming without loss of generality that $\Delta > 0$):

$$\int \psi_1(\mathbf{x},\mathbf{s}) d\left\{F_{\mathbf{X},\mathbf{S}^{(1)}}(\mathbf{x},\mathbf{s}) - F_{\mathbf{X},\mathbf{S}}(\mathbf{x},\mathbf{s})\right\} \geq \int \psi_0(\mathbf{x},\mathbf{s}) d\left\{F_{\mathbf{X},\mathbf{S}^{(0)}}(\mathbf{x},\mathbf{s}) - F_{\mathbf{X},\mathbf{S}}(\mathbf{x},\mathbf{s})\right\} \text{ and} \tag{7}$$

$$\int \left\{\psi_1(\mathbf{x},\mathbf{s}) - \psi_0(\mathbf{x},\mathbf{s})\right\} dF_{\mathbf{X},\mathbf{S}}(\mathbf{x},\mathbf{s}) \geq 0. \tag{8}$$

If $\Delta < 0$, then one would require (7) and (8) to hold with inequalities reversed.

Because $F_{\mathbf{X},\mathbf{S}}$ is a mixture distribution $F_{\mathbf{X},\mathbf{S}}(\mathbf{x},\mathbf{s}) = e(\mathbf{x})F_{\mathbf{X},\mathbf{S}^{(1)}}(\mathbf{x},\mathbf{s}) + \{1-e(\mathbf{x})\}F_{\mathbf{X},\mathbf{S}^{(0)}}(\mathbf{x},\mathbf{s})$, the first condition (7) may be rewritten:

$$\int \left\{e(\mathbf{x})\psi_1(\mathbf{x},\mathbf{s}) + \{1-e(\mathbf{x})\}\psi_0(\mathbf{x},\mathbf{s})\right\} d\left\{F_{\mathbf{X},\mathbf{S}^{(1)}}(\mathbf{x},\mathbf{s}) - F_{\mathbf{X},\mathbf{S}^{(0)}}(\mathbf{x},\mathbf{s})\right\} \geq 0. \tag{9}$$

The condition (8) ensures that $\Delta_S \geq 0$, i.e., that the residual treatment effect is in the same direction as the overall treatment effect Δ. Condition (9), which ensures that $\Delta \geq \Delta_S$, requires that, roughly speaking, a propensity-weighted mixture of the two conditional mean functions $e(\mathbf{x})\psi_1(\mathbf{x},\mathbf{s}) + \{1-e(\mathbf{x})\}\psi_0(\mathbf{x},\mathbf{s})$ is larger when $\mathbf{s}$ takes values from the distribution of the counterfactual surrogates under treatment than if it took values from the distribution under control. These two conditions are analogues of conditions (A6) and (A7) in [1].

3.3 Identification

In this section, we show how $\Delta_{\mathbf{S}}$ may be characterized via functionals of the observed data distribution. Specifically,

$$\begin{aligned}
\Delta_{\mathbf{S}} &= \mathbb{E}\left\{\mu_1(\mathbf{X},\mathbf{S}) - \mu_0(\mathbf{X},\mathbf{S})\right\} && (10)\\
&= \mathbb{E}\left[\frac{AY}{\pi(\mathbf{X},\mathbf{S})} - \frac{(1-A)Y}{\{1-\pi(\mathbf{X},\mathbf{S})\}}\right] && (11)\\
&= \mathbb{E}\left[\frac{AY - \{A-\pi(\mathbf{X},\mathbf{S})\}\mu_1(\mathbf{X},\mathbf{S})}{\pi(\mathbf{X},\mathbf{S})} - \frac{(1-A)Y + (A-\pi(\mathbf{X},\mathbf{S}))\mu_0(\mathbf{X},\mathbf{S})}{1-\pi(\mathbf{X},\mathbf{S})}\right], && (12)
\end{aligned}$$

where $\mu_a(\mathbf{x},\mathbf{s}) = \mathbb{E}(Y|\mathbf{X}=\mathbf{x}, A=a, \mathbf{S}=\mathbf{s})$. The result (10) follows from the definition (3) and the fact that $\psi_a(\mathbf{X},\mathbf{S}^{(a)}) = \mu_a(\mathbf{X},\mathbf{S})$ because of (6), and the other results follow shortly thereafter, following familiar paths as arguments for identification of the average treatment effect in other contexts. These results show that the residual treatment effect may be identified without knowledge of the mean function for the outcome via (11), and gives an augmented inverse probability weighting [3] version of the estimand (12) .Δ may also be identified using similar functionals with

$\mu_a(\mathbf{X}, \mathbf{S})$ replaced by $m_a(\mathbf{X})$ and $\pi(\mathbf{X}, \mathbf{S})$ replaced by $e(\mathbf{X})$:

$$\Delta = \mathbb{E}\left[\frac{AY - \{A - e(\mathbf{X})\}m_1(\mathbf{X})}{e(\mathbf{X})} - \frac{(1-A)Y + (A - e(\mathbf{X}))m_0(\mathbf{X})}{1 - e(\mathbf{X})}\right] \tag{13}$$

where $m_a(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x}, A = a)$. In Section 3.4, we propose an estimation approach using the expressions in (12) and (13).

Remark 2. *If one wished to pursue the alternative PTE definition using $\mathcal{M}_\psi$ and $\mathcal{M}_0$, these quantities may also be identified from the observed data. Let $\eta_a^2(\mathbf{x}, \mathbf{s}) = \mathbb{E}\{Y - \mu_a(\mathbf{x}, \mathbf{s})\}^2$ be the conditional (on $\mathbf{X}$ and $\mathbf{S}$) variance of Y in treatment arm $A = a$. Further define $\sigma_a^2(\mathbf{x}) = \mathbb{E}\left\{\eta_a^2(\mathbf{x}, \mathbf{S})|\mathbf{X} = \mathbf{x}, A = a\right\}$ to be the expectation of $\eta_a^2(\mathbf{x}, \mathbf{S})$ over the conditional distribution of $\mathbf{S}$ given $A = a, \mathbf{X} = \mathbf{x}$. Then using a similar argument as above, we have that*

$$\begin{aligned}
\mathcal{M}_\psi &= \mathbb{E}\left[e(\mathbf{X})\sigma_1^2(\mathbf{X}) + \{1 - e(\mathbf{X})\}\sigma_0^2(\mathbf{X})\right] \\
&= \mathbb{E}\left(A\eta_1(\mathbf{X}, \mathbf{S}) + (1-A)\eta_0(\mathbf{X}, \mathbf{S})\right) \\
&= \mathbb{E}\left(A\eta_1(\mathbf{X}, \mathbf{S}) - \{A - e(\mathbf{X})\}\sigma_1^2(\mathbf{X}) + (1-A)\eta_0(\mathbf{X}, \mathbf{S}) + (A - e(\mathbf{X}))\sigma_0^2(\mathbf{X})\right)
\end{aligned}$$

with similar expressions holding for $\mathcal{M}_0$ mutatis mutandis. The first equality is analogous to an outcome-model-type identification, as in (10)*, the second equality is an inverse-probability-type identification, as in* (11)*, and the third is a doubly-robust-type identification, as in* (12)*. One would need to modify slightly the cross-fitting procedure outlined in the next section to incorporate $\sigma_a^2(\mathbf{x})$, which would require two disjoint folds of data for estimation: one to estimate $\eta_a^2(\mathbf{x}, \mathbf{s})$ and a second to estimate $\sigma_a^2(\mathbf{x})$.*

3.4 Estimation

We propose an estimation approach utilizing the augmented inverse probability weighting versions of the estimands in (12) and (13). The traditional form of these estimators would take the form (for Δ_S)

$$\widehat{\Delta}_{\mathbf{S}}(\mathcal{O}, \widehat{\pi}, \widehat{\mu}_0, \widehat{\mu}_1) = n^{-1}\sum_{i=1}^{n}\frac{A_iY_i - \{A_i - \widehat{\pi}(\mathbf{X}_i, \mathbf{S}_i)\}\widehat{\mu}_1(\mathbf{X}_i, \mathbf{S}_i)}{\widehat{\pi}(\mathbf{X}_i, \mathbf{S}_i)} - n^{-1}\sum_{i=1}^{n}\frac{(1-A_i)Y_i + \{A_i - \widehat{\pi}(\mathbf{X}_i, \mathbf{S}_i)\}\widehat{\mu}_0(\mathbf{X}_i, \mathbf{S}_i)}{1 - \widehat{\pi}(\mathbf{X}_i, \mathbf{S}_i)}$$

with $\widehat{\mu}_a(\mathbf{x}, \mathbf{s}) = \widehat{E}\{Y|\mathbf{X} = \mathbf{x}, A = a, \mathbf{S} = \mathbf{s}\}$ an estimate of the conditional outcome function and $\widehat{\pi}(\mathbf{x}, \mathbf{s}) = \widehat{P}(A = 1|\mathbf{S} = \mathbf{s}, \mathbf{X} = \mathbf{x})$, and (for Δ)

$$\widehat{\Delta}(\mathcal{O}, \widehat{e}, \widehat{m}_0, \widehat{m}_1) = n^{-1}\sum_{i=1}^{n}\frac{A_iY_i - \{A_i - \widehat{e}(\mathbf{X}_i)\}\widehat{m}_1(\mathbf{X}_i)}{\widehat{e}(\mathbf{X}_i)} - n^{-1}\sum_{i=1}^{n}\frac{(1-A_i)Y_i + \{A_i - \widehat{e}(\mathbf{X}_i)\}\widehat{m}_0(\mathbf{X}_i)}{1 - \widehat{e}(\mathbf{X}_i)}$$

with $\widehat{m}_a(\mathbf{x}) = \widehat{E}\{Y|\mathbf{X} = \mathbf{x}, A = a\}$ and $\widehat{e}(\mathbf{x}) = \widehat{P}(A = 1|\mathbf{X} = \mathbf{x})$ estimates of the corresponding mean functions and propensity score. We discuss the doubly robust property further in Section 6.

Here, we aim to specify our approach as generally as possible, such that in practice, one could choose any reasonable approach to obtaining estimates for $\widehat{\mu}_a(\mathbf{x}, \mathbf{s})$, $\widehat{\pi}(\mathbf{x}, \mathbf{s})$, $\widehat{m}_a(\mathbf{x})$, and $\widehat{e}(\mathbf{x})$ while maintaining the ability to make inference

about the PTE $R_{\mathbf{S}}$. Certainly, if the dimensions of $\mathbf{X}$ and $\mathbf{S}$ are low, one could use simple parametric functions or, if very low (e.g., X, S both scalar random variables), one could use nonparametric kernel smoothing. As we are specifically interested in settings where $\mathbf{S}$, and possibly even $\mathbf{X}$, may be high-dimensional, we do not restrict ourselves to settings where the above may be feasible, and we allow for machine learning and high-dimensional regression approaches to be used for estimating these functions. In our simulations, we specifically examine two versions of our proposed estimation approach: one that uses the Super Learner [31] and one that uses the relaxed lasso [14] to estimate each of the quantities: $\widehat{\mu}_a(\mathbf{x}, \mathbf{s})$, $\widehat{\pi}(\mathbf{x}, \mathbf{s})$, $\widehat{m}_a(\mathbf{x})$, and $\widehat{e}(\mathbf{x})$. Details are described in Section 5.

Importantly, in these settings when the dimensions of $\mathbf{X}$ and $\mathbf{S}$ may be high and/or when flexible machine learning models are used to estimate the quantities $\{\widehat{\mu}_0(\cdot), \widehat{\mu}_1(\cdot), \widehat{\pi}(\cdot), \widehat{m}_0(\cdot), \widehat{m}_1(\cdot), \widehat{e}(\cdot)\}$, additional steps that preclude overfitting and facilitate asymptotic analysis should be considered. We propose to use the *cross-fitting* approach of [5], where (for a given positive integer $K > 1$), the observed data are randomly partitioned into K sets $(\mathcal{O}_k)_{k=1,\ldots,K}, \bigcup_k \mathcal{O}_k = \mathcal{O}, \bigcap_k \mathcal{O}_k = \emptyset$, with the indices corresponding to $\mathcal{O}_k$ denoted I_k, such that each set $\mathcal{O}_k$ has $m = |I_k| = n/K$ observations in it (assuming n is a multiple of K). The estimator for $\Delta_{\mathbf{S}}$ then takes the form

$$\widehat{\Delta}_{\mathbf{S}} = K^{-1} \sum_{k=1}^{K} \widehat{\Delta}_{\mathbf{S},k} = K^{-1} \sum_{k=1}^{K} \widehat{\Delta}_{\mathbf{S}}(\mathcal{O}_k, \widehat{\pi}_{-k}, \widehat{\mu}_{0-k}, \widehat{\mu}_{1-k}) \tag{14}$$

where

$$\widehat{\Delta}_{\mathbf{S},k} = m^{-1} \sum_{i \in I_k} \frac{A_i Y_i - \{A_i - \widehat{\pi}_{-k}(\mathbf{X}_i, \mathbf{S}_i)\}\widehat{\mu}_{1-k}(\mathbf{X}_i, \mathbf{S}_i)}{\widehat{\pi}_{-k}(\mathbf{X}_i, \mathbf{S}_i)} - m^{-1} \sum_{i \in I_k} \frac{(1 - A_i) Y_i + \{A_i - \widehat{\pi}_{-k}(\mathbf{X}_i, \mathbf{S}_i)\}\widehat{\mu}_{0-k}(\mathbf{X}_i, \mathbf{S}_i)}{1 - \widehat{\pi}_{-k}(\mathbf{X}_i, \mathbf{S}_i)}$$

and $(\widehat{\pi}_{-k}, \widehat{\mu}_{0-k}, \widehat{\mu}_{1-k})$ are estimated on all of the data excluding $\mathcal{O}_k$: $\mathcal{O}_k^c = \mathcal{O}/\mathcal{O}_k$.

The estimator for Δ uses a similar cross-fitting procedure:

$$\widehat{\Delta} = K^{-1} \sum_{k=1}^{K} \widehat{\Delta}_k = K^{-1} \sum_{k=1}^{K} \widehat{\Delta}(\mathcal{O}_k, \widehat{e}_{-k}, \widehat{m}_{0-k}, \widehat{m}_{1-k}) \tag{15}$$

$$\widehat{\Delta}_k = m^{-1} \sum_{i \in I_k} \frac{A_i Y_i - \{A_i - \widehat{e}_{-k}(\mathbf{X}_i)\}\widehat{m}_{1-k}(\mathbf{X}_i)}{\widehat{e}_{-k}(\mathbf{X}_i)} - m^{-1} \sum_{i \in I_k} \frac{(1 - A_i) Y_i + \{A_i - \widehat{e}_{-k}(\mathbf{X}_i)\}\widehat{m}_{0-k}(\mathbf{X}_i)}{1 - \widehat{e}_{-k}(\mathbf{X}_i)} \tag{16}$$

where $\widehat{m}_{a-k}(\mathbf{x}) = \widehat{E}(Y|A = a, \mathbf{X} = \mathbf{x})$ is the outcome model conditional only on the pre-treatment covariates $\mathbf{X}$ and $\widehat{e}_{-k}(\mathbf{x}) = \widehat{P}(A = 1|\mathbf{X} = \mathbf{x})$ is the estimated propensity score, both estimated on $\mathcal{O}_k^c$. Finally, the PTE may be estimated as

$$\widehat{R}_{\mathbf{S}} = \frac{\widehat{\Delta} - \widehat{\Delta}_{\mathbf{S}}}{\widehat{\Delta}} = 1 - \frac{\widehat{\Delta}_{\mathbf{S}}}{\widehat{\Delta}}.$$

4 Asymptotic Behavior, Inference, and Variance Estimation

In this section, we show that our approach (and specifically, the use of cross-fitting) yields a convenient asymptotic distribution for the PTE estimator (given certain regularity conditions we outline), and we demonstrate how this may be used for making inference on $\widehat{R}$.

The key conditions are on the rate of convergence of the nuisance functions $\{\widehat{\mu}_0(\cdot), \widehat{\mu}_1(\cdot), \widehat{\pi}(\cdot), \widehat{m}_0(\cdot), \widehat{m}_1(\cdot),\ \widehat{e}(\cdot)\}$. These rates of convergence are allowed to be quite slow. Specifically, each estimated nuisance function is allowed to converge to the true nuisance function at a rate as slow as $n^{-1/4}$. Furthermore, because of sample-splitting, the form of the asymptotic variance and its plug-in estimate are convenient to use. We now state the main result more formally

Proposition 1. *Let*

$$\phi_1(\mathbf{O}; \Delta, \eta_1) = \frac{AY - \{A - e(\mathbf{X})\}m_1(\mathbf{X})}{e(\mathbf{X})} - \frac{(1-A)Y + \{A - e(\mathbf{X})\}m_0(\mathbf{X})}{1 - e(\mathbf{X})} - \Delta, \tag{17}$$

and

$$\phi_2(\mathbf{O}; \Delta_{\mathbf{S}}, \pi, \mu_0, \mu_1) = \frac{AY - \{A - \pi(\mathbf{X}, \mathbf{S})\}\mu_1(\mathbf{X}, \mathbf{S})}{\pi(\mathbf{X}, \mathbf{S})} - \frac{(1-A)Y + \{A - \pi(\mathbf{X}, \mathbf{S})\}\mu_0(\mathbf{X}, \mathbf{S})}{1 - \pi(\mathbf{X}, \mathbf{S})} - \Delta_{\mathbf{S}}. \tag{18}$$

If, in addition to the assumptions in Section 3.1 and Appendix 7.2, we also have, for a sequence of positive constants $\delta_n, n = 1, ..., \infty$ *that*

$$\begin{aligned}
&E\left[\{\widehat{\mu}_{a-k}(\mathbf{X}, \mathbf{S}) - \mu_a(\mathbf{X}, \mathbf{S})\}^2\right] \times E\left[\{\widehat{\pi}_{-k}(\mathbf{X}, \mathbf{S}) - \pi(\mathbf{X}, \mathbf{S})\}^2\right] \leq \delta_n n^{-\frac{1}{2}}, \\
&E\left[\{\widehat{\mu}_{a-k}(\mathbf{X}, \mathbf{S}) - \mu_a(\mathbf{X}, \mathbf{S})\}^2\right] + E\left[\{\widehat{\pi}_{-k}(\mathbf{X}, \mathbf{S}) - \pi(\mathbf{X}, \mathbf{S})\}^2\right] \leq \delta_n, \\
&P\{\epsilon_2 < \widehat{\pi}_{-k}(\mathbf{S}, \mathbf{X}) < 1 - \epsilon_2\} = 1
\end{aligned}$$

and

$$\begin{aligned}
&E\left[\{\widehat{m}_{a-k}(\mathbf{X}) - m_a(\mathbf{X})\}^2\right] \times E\left[\{\widehat{e}_{-k}(\mathbf{X}) - e(\mathbf{X})\}^2\right] \leq \delta_n n^{-\frac{1}{2}}, \\
&E\left[\{\widehat{m}_{a-k}(\mathbf{X}) - m_a(\mathbf{X})\}^2\right] + E\left[\{\widehat{e}_{-k}(\mathbf{X}) - e(\mathbf{X})\}^2\right] \leq \delta_n, \\
&P\{\epsilon_1 < \widehat{e}_{-k}(\mathbf{X}) < 1 - \epsilon_1\} = 1,
\end{aligned}$$

then

$$\begin{aligned}
&\sqrt{n}\sigma^{-1}(\widehat{R}_{\mathbf{S}} - R_{\mathbf{S}}) \to N(0, 1), \quad \textit{where} \\
&\sigma^2 = \Delta^{-2} E\left\{\phi_1(\mathbf{O}, \Delta, e, m_0, m_1)^2\right\} + \Delta_{\mathbf{S}}\Delta^{-4} E\left\{\phi_2(\mathbf{O}, \Delta_{\mathbf{S}}, \pi, \mu_0, \mu_1)^2\right\} - \\
&\quad 2\Delta_{\mathbf{S}}\Delta^{-3} E\left\{\phi_1(\mathbf{O}, \Delta, e, m_0, m_1)\phi_2(\mathbf{O}, \Delta_{\mathbf{S}}, \pi, \mu_0, \mu_1)\right\}.
\end{aligned}$$

The consequence of this result is that a straightforward $(1-\alpha)\%$ confidence interval for $R_{\mathbf{S}}$ can be obtained as

$$\mathcal{C} = (\widehat{R}_{\mathbf{S}} - z_{1-\alpha/2}\widehat{\sigma}, \widehat{R}_{\mathbf{S}} + z_{1-\alpha/2}\widehat{\sigma}), \tag{19}$$

with z_q the qth quantile of the standard normal distribution and

$$\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}\left(\widehat{\Delta}^{-2}\widehat{\phi}_{1i}^2 + \widehat{\Delta}_{\mathbf{S}}^2\widehat{\Delta}^{-4}\widehat{\phi}_{2i}^2 - 2\widehat{\Delta}_{\mathbf{S}}\widehat{\Delta}^{-3}\widehat{\phi}_{1i}\widehat{\phi}_{2i}\right).$$

where

$$\widehat{\phi}_{1i} = \phi_1(\mathbf{O}_i, \widehat{\Delta}, \widehat{e}_{-k_i}, \widehat{m}_{0-k_i}, \widehat{m}_{1-k_i}), \widehat{\phi}_{2i} = \phi_2(\mathbf{O}_i, \widehat{\Delta}_{\mathbf{S}}, \widehat{\pi}_{-k_i}, \widehat{\mu}_{0-k_i}, \widehat{\mu}_{1-k_i})\}^{\intercal}$$

and $k_i = I\{k : i \in I_k\}$. In addition, it follows that $P(R_{\mathbf{S}} \in \mathcal{C}) \to 1-\alpha$ because, under these assumptions, based on the arguments in [5],

$$\sqrt{n}\Sigma^{-\frac{1}{2}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to N(\mathbf{0}, I) \tag{20}$$

where $\widehat{\boldsymbol{\theta}} = (\widehat{\Delta}, \widehat{\Delta}_{\mathbf{S}})^{\intercal}, \boldsymbol{\theta} = (\Delta, \Delta_{\mathbf{S}})^{\intercal}, \Sigma = \mathbb{E}(\boldsymbol{\phi}\boldsymbol{\phi}^{\intercal})$ and this holds when replacing Σ by $\widehat{\Sigma} = n^{-1}\sum_{i=1}^{n}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^{\intercal}$, $\widehat{\boldsymbol{\phi}}_i = (\phi_{1i}, \phi_{2i})^{\intercal}$.

5 Simulations

5.1 Simulation Overview

We used two approaches to estimate the six functions $\pi(\mathbf{X}, \mathbf{S}), \mu_1(\mathbf{X}, \mathbf{S}), \mu_0(\mathbf{X}, \mathbf{S}), \pi(\mathbf{X}), m_1(\mathbf{X}), m_0(\mathbf{X})$, yielding two different versions of our proposed estimator. In one, we used the Super Learner [31], which finds an optimal combination of a set of candidate models or learners, as implemented in the `SuperLearner` package [22] in `R`. We denote this estimator "DR-SL" in what follows. In the second version, which we call "DR-lasso" we use the relaxed lasso [14], as implemented in the `glmnet` package [8], to estimate all needed functions. For all simulations, we truncated estimates of the propensity and surrogate scores so that $0.01 \leq \widehat{\pi}_{-k}(\mathbf{X}_i, \mathbf{S}_i), \widehat{e}_{-k}(\mathbf{X}_i) \leq 0.99$ for all i.

It is our understanding that there are no currently available methods to estimate the PTE of a high-dimensional surrogate. However, since there are available methods to measure the strength of high-dimensional mediators, we compare our proposed approach to these available methods. While the goals of mediation analysis and surrogate markers analysis are different and the assumptions necessary differ, this allows us to offer some reasonable comparison to methodology that is currently available, rather than no comparison at all. Thus, to fairly compare, we purposefully set up our simulations such that $\mathbf{S}$ is a mediator in that it lies on the causal pathway between A and Y. While mediation methods are attempting to estimate a distinct quantity from the PTE we are interested in, in the following simulations, the estimation of the "proportion mediated" (which is used in mediation) and the PTE that we are interested in are, in fact, the same.

Specifically, in our simulations we compare our proposed approach to three methods for high-dimensional mediation

analysis: HIMA (High-dimensional mediation analysis, [37]) as implemented in the `HIMA` package [38]; BAMA (Bayesian mediation analysis, [29]) as implemented in the `bama` package [24]; and HILMA (High-dimensional linear mediation analysis, [39]) as implemented in the `freebird` package. All three of these methods propose $p+1$ linear models: p models for the surrogates/mediators

$$S_{ij} = \alpha_{0j} + \alpha_j A_i + \sum_{k=1}^{q} \boldsymbol{\gamma}^*_{jk} X_{ik} + e_{ij}, j = 1, ..., p; i = 1, ..., n,$$

and one outcome model

$$Y_i = \beta_0 + \Delta_S A_i + \sum_{j=1}^{p} \beta_j S_{ij} + \sum_{k=1}^{q} \gamma_k X_{ik} + \epsilon_i,$$

though the implementation of freebird does not allow for the inclusion of covariates X_{ik}. Using these models, the overall treatment effect can be identified as $\Delta = \Delta_S + \sum_{j=1}^{p} \alpha_j \beta_j$, and as above the PTE (or proportion mediated) may be identified as $R = 1 - \Delta_S/\Delta$. The HIMA procedure employs sure independence screening and penalized least squares with the minimax concave penalty [36] to set some values of $\alpha_j \beta_j$ to be exactly 0. Similarly, BAMA employs Bayesian shrinkage estimation to estimate all parameters (including some at exactly 0) under normality assumptions on the error terms. Finally, the freebird approach debiases a pilot scaled lasso estimator.

We computed 95% confidence intervals (CIs) for R from the results given in Section 4 for the two proposed approaches. We computed 95% credible intervals for the BAMA estimator from the 2.5% and 97.5% quantiles of the posterior distribution of $1 - \frac{\Delta_S}{\Delta_S + \sum_{j=1}^{p} \alpha_j \beta_j}$. CIs were computed for the HIMA and freebird implementations because they are not available.

5.2 Simulation Setup

We constructed two sets of simulations for a total of 10 simulation settings to assess the performance of our proposed approach in both low- and high-dimensional settings.

For the first set of simulations, the data-generating mechanism for $\mathbf{S}^{(a)}$ was linear in $\mathbf{X}$, the data-generating mechanism for $Y^{(a)}$ was linear in $\mathbf{X}$ and $\mathbf{S}^{(a)}$, and the propensity score was linear on the log-odds scale. Given this data-generating mechanism, we would expect that all methods (proposed and comparisons) should perform reasonably well. We set the dimension of $\mathbf{S}$ (p) and of $\mathbf{X}$ (q) to be 100. We let $X_{ij} \sim N(0,1)$, $A_i \sim \text{Bernoulli}\{\pi(\mathbf{X}_i)\}$, $\pi(\mathbf{X}_i) = \text{logit}(\boldsymbol{\gamma}^\intercal \mathbf{X}_i)$, where $\boldsymbol{\gamma} \sim N(0,1)$. The surrogates were generated as $\mathbf{S}_i^{(a)} = \boldsymbol{\alpha}_a + \boldsymbol{\beta}_a^\intercal \mathbf{X}_i + \mathbf{e}_i$, where $\alpha_{11} = 0.75, \alpha_{12} = 0.25, \alpha_{01} = \alpha_{02} = 0, \alpha_{1j} \sim U(0,1), \alpha_{0j} \sim U(-0.5, 0.5), j = 3, ...p$. And $(\beta_{11}, \beta_{12}, \beta_{13}, \beta_{14}, \beta_{15}) = (-1, -0.5, 0, 0.5, 1), (\beta_{01}, \beta_{02}, \beta_{03}, \beta_{04}, \beta_{05}) = (-2, -1.5, -1, -0.5, 0), \beta_{aj} = 0, a = 0, 1, j > 5$. The outcome counterfactuals were given by $Y_i^{(a)} = a\Delta_S + \sum_{j=1}^{25} X_{ij} + S_{i1}^{(a)} + S_{i2}^{(a)} + \epsilon_i$. The errors e_{ij} and ϵ_i were $N(0, \sigma^2)$. We let the sample size vary between $n = 100$ and $n = 500$ and the level of noise between $\sigma = 0.1$ and $\sigma = 0.5$, and we set $R = 0.5$.

In the second set of simulations, the data-generating mechanism was less linear, including interactions between covariates in both the propensity score and the model for $\mathbf{S}$, though the outcome model was still a simple linear combination of $\mathbf{X}$ and $\mathbf{S}$. This set of simulations should mimic what may happen in practice since all models are

typically subject to some amount of mis-specification. We expected the nonlinearity to induce bias in the competing methods (which require linear models to hold). There were two pre-treatment covariates $X_1 \sim \text{Uniform}(-2, 5)$ and $X_2 \sim \text{Bernoulli}(0.5)$. The probability of treatment was determined by $P(A = 1|\mathbf{X}) = \text{logit}(-X1 + 2X_1X_2)$. The surrogates under treatment were given by $\mathbf{S}^{(1)} = 1.5+(X_1+X_2)\mathcal{I}_{10}+\mathbf{e}$ and under control by $\mathbf{S}^{(0)} = 2+X_2\mathcal{I}_{10}-X_1X_2+\mathbf{e}$ where $\mathcal{I}_{10}$ is a $p \times p$ matrix of 0s except it has 1s on the first 10 diagonal elements and $\mathbf{e} \sim N(0, \Sigma)$ and $\Sigma = 0.4\mathbf{1}\mathbf{1}^\intercal + 0.6I_p$ is a covariance matrix with common covariance 0.4. Finally, the outcome counterfactuals were given by $Y^{(a)} = a\Delta_S + X_1 + X_2 + \sum_{j=1}^{15} S_j^{(a)} + \epsilon, \epsilon \sim N(0, 1)$. We set the sample size to be $n = 500$, and we varied the number of surrogates ($p = 50, 100, 500$), and the PTE (surrogates were either weak ($R = 0.41$) or strong ($R = 0.85$)).

5.3 Results for first data-generating mechanism

As expected, this data-generating mechanism was well approximated using linear models. When the sample size was large ($n = 500$) and $\sigma = 0.5$, all methods performed well, with relatively small bias (see Figure 2). The median $\widehat{R}$ estimates were 0.46 (BAMA), 0.48 (freebird), 0.49 (HIMA), 0.48 (DR-SL), and 0.49 (DR-lasso), all of which compare favorably to the true R value of 0.5. The distribution of estimates for the proposed approaches was a bit tighter than for the competing methods. The empirical 2.5% and 97.5% quantiles were 0.30 and 0.62 for DR-SL, 0.35 and 0.58 for DR-lasso, 0.15 and 0.67 for HIMA, 0.11 and 0.63 for BAMA, and -0.85 and 0.71 for freebird. Results were similar for the sample size with less noise ($\sigma = 0.1$), except for the freebird approach which began to estimate the PTE to be exactly 1 in almost all simulations: the empirical 2.5% and 97.5% quantiles of the $\widehat{R}$ distribution for freebird were exactly 1. At the lower sample size ($n = 100$), HIMA did not run, while BAMA estimates tended to be near 0 and freebird estimates were again clustered tightly around 1. Our proposed approaches had median $\widehat{R}$ values of 0.46 (DR-SL) and 0.51 (DR-lasso), though with more variability than when $n = 500$. The median absolute deviation ($|\widehat{R} - R|$) was smaller for both DR-SL and DR-lasso than any of the competing approaches for all settings.

CI coverage for the proposed estimators tended to be conservative in all settings: coverage for DR-SL was 100% ($n = 100, \sigma = 0.1$ and $n = 100, \sigma = 0.5$), 98% ($n = 500, \sigma = 0.1$), and 99% ($n = 500, \sigma = 0.5$), and for DR-lasso it was 100% in all settings. On the other hand, BAMA CIs only had nominal coverage at lower sample sizes: 95% ($n = 100, \sigma = 0.1$), 98% ($n = 100, \sigma = 0.5$), 88% ($n = 100, \sigma = 0.1$), and 84% ($n = 500, \sigma = 0.5$). Even when BAMA CIs had nominal coverage, they were about four times larger than the CIs for the proposed estimators: BAMA CI half-lengths were 1.38 ($n = 100, \sigma = 0.1$) and 1.64 ($n = 100, \sigma = 0.5$), while the corresponding half-lengths were 0.47 and 0.57 for DR-SL and 0.57 and 1.02 for DR-lasso.

5.4 Results for second data-generating mechanism

This data-generating mechanism was not as well approximated by linear models, in part because of the interactions between X_1 and X_2. Thus, the linear model-based estimates of HIMA, BAMA, and freebird all performed poorly (see Figure 3). HIMA and BAMA typically had $\widehat{R}$ values around 0 and lower, while freebird again estimated R to be 1 or above (the 2.5% quantile of the empricial distribution of freebird estimates was at 1 in all settings). On the other hand, the two proposed approaches gave more reasonable if imperfect estimates. When the dimensionality of the

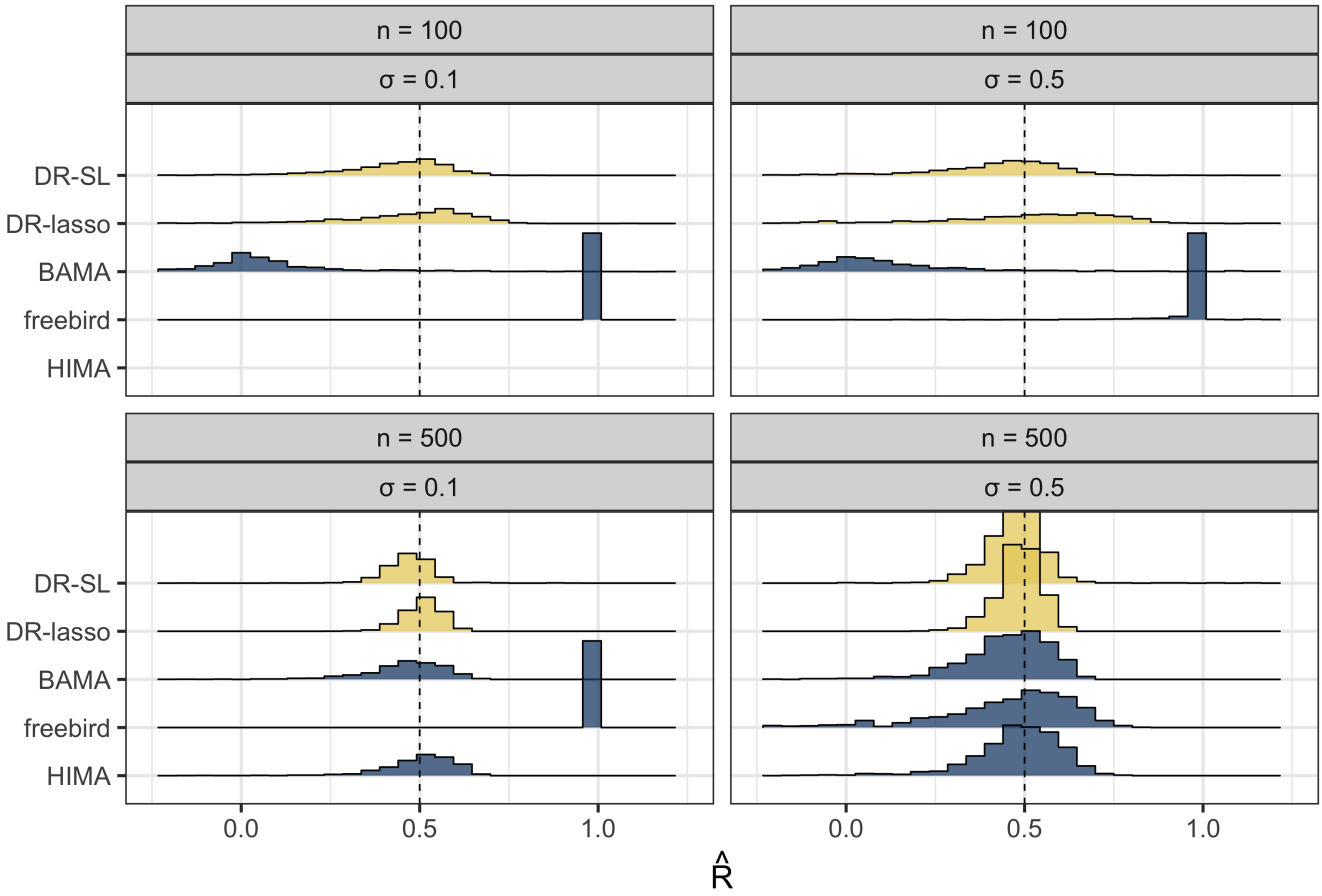

Figure 2: Distribution of estimates of $\widehat{R}$ in first data-generating mechanism. Lighter shaded regions represent the distribution of the proposed estimators ("DR-SL" and "DR-lasso"); darker shaded regions represent the distribution of the comparison estimators ("BAMA", "freebird", and "HIMA"). The true value of R is given as a vertical dotted line at 0.5.

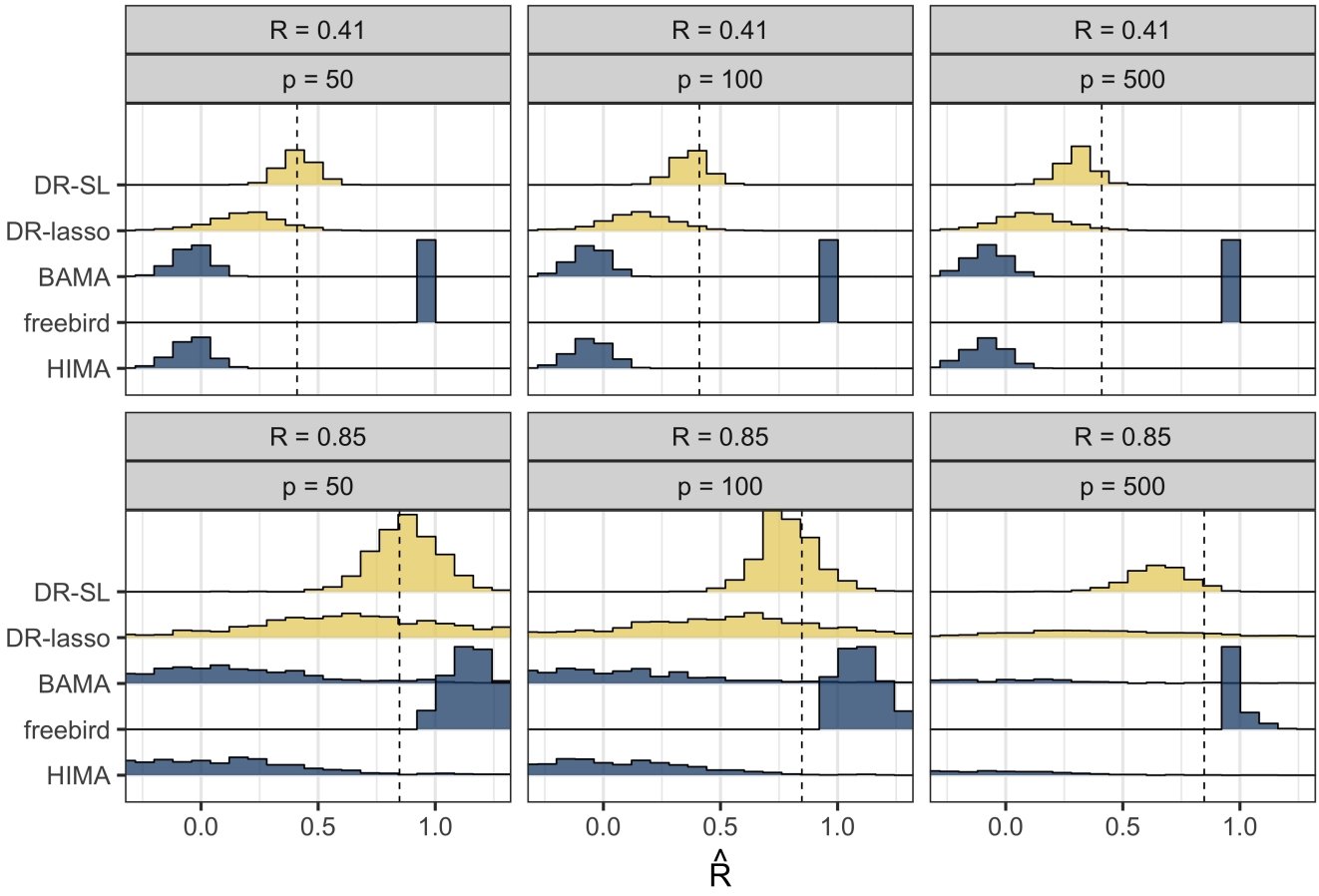

Figure 3: Distribution of estimates of $\widehat{R}$ in first data-generating mechanism. Lighter shaded histograms represent distribution of the proposed estimators ("DR-SL" and "DR-lasso"); darker shaded histograms represent the distribution of the comparison estimators ("BAMA", "freebird", and "HIMA"). The true value of R is given as a vertical dotted line.

surrogates was much smaller than the sample size ($p = 50$), the DR-SL estimator gave nearly unbiased results, with median $\widehat{R}$ values of 0.88 (when $R = 0.85$) and 0.41 (when $R = 0.45$), as it was able to adapt to the nonlinearity in the data-generating mechanism. As the dimensionality of the surrogate grew, the Super Learner estimates of nuisance functions began to lean more heavily on linear models (like the lasso), and performance degraded: median $\widehat{R}$ values were 0.78 ($R = 0.85, p = 100$), 0.37 ($R = 0.41, p = 100$), 0.65 ($R = 0.85, p = 500$), 0.31 ($R = 0.41, p = 500$). The DR-lasso estimator resulted in performance between HIMA/BAMA and DR-SL, typically underestimating R by more than DR-SL but less than HIMA/BAMA. CI coverage was strong for DR-SL when $R = 0.41$ (97%, 96%, 98% for $p = 50, 100, 500$, respectively) but was much weaker when $R = 0.85$ (91%, 74%, 69% for $p = 50, 100, 500$, respectively). This subpar coverage appeared to be due to the bias of the estimator and not due to the width of confidence intervals. When CIs were adjusted for the empirical bias of the DR-SL estimator (by subtracting the empirical bias from both sides of CIs), nominal coverage was observed (96%, 97%, 95% coverage for $p = 50, 100, 500$, respectively).

6 Discussion

We have proposed a very general approach to evaluating surrogate markers which can be applied in randomized experiments or in observational studies and can be used regardless of the dimensionality of either surrogates or any covariates needed to control for confounding. Our approach is nonparametric in that we have defined the proportion of the treatment effect explained by the surrogates without reference to any models, and we have shown how machine learning approaches like the super learner may be used to very flexibly estimate nuisance functions. Our simulation results suggest that our the Super Learner estimator (the estimator we called DR-SL) outperforms competing methods even when the underlying data-generating mechanism is linear and still gives reasonable results even when high dimensional linear models are mis-specified.

We note that the proposed estimator of the PTE, $\widehat{R}_{\mathbf{S}}$, is "doubly-robust" in the sense of *rate double robustness* discussed in [28], where the rates of two estimators (for us, the estimators of μ_a and π as well as the estimators of m_a and e) may compensate for one another – i.e., a slightly slower rate of convergence for π may be compensated for by a slightly faster rate of convergence for μ_a so long as their product is $o_p(n^{-\frac{1}{2}})$. In high dimensions, the approach of [28], based on ℓ_1-penalized regression, could be used to additionally obtain the slightly different property of *model double robustness* where one of the estimators may converge to a different value besides the true parameter. Our development here, which we chose to be quite general and applicable in both low- and high-dimensional settings, could easily be adapted to use the estimation technique of [28] in high dimensions to recover model double robustness.

Our approach here is intimately tied to previous approaches for evaluating surrogate markers. The approaches in [1] and in [20] can be seen to be similar to a version of our approach where the average treatment effect among controls is estimated under further assumptions. In [1], they further require strict randomization of treatment, and they assume that $\mathbf{S}$ is a realization of a smooth continuous function. [20] make similar assumptions but take S to be a scalar surrogate. They estimate a version of (10) among controls: $\Delta_{\mathbf{S}} = \mathbb{E}\left\{\mu_1(\mathbf{S}) - \mu_0(\mathbf{S}) | A = 0\right\}$ using kernel smoothing and taking advantage of the fact that treatment is randomized. Their estimates have the form

$$\widehat{\Delta}_{\mathbf{S}} = n_0^{-1} \sum_{A_i=0} \left\{\widehat{\mu}_1(\mathbf{S}_i) - Y_i\right\}$$

where $n_0 = \sum_{i=1}^n I\{A_i = 0\}$ and $\widehat{\mu}_1(\cdot)$ is estimated via kernel smoothing (possibly after dimension reduction). Our approach here could easily be adapted to estimate a similar quantity by using methods for doubly robust estimation of the average treatment effect on the treated [26, 15] by, for example,

$$\widehat{\Delta}_{\mathbf{S}} = n^{-1} \sum_{i=1}^n \left\{A_i \frac{1 - \widehat{\pi}(\mathbf{S}_i)}{\widehat{\pi}(\mathbf{S}_i)} - (1 - A_i)\right\} \left\{\widehat{\mu}_1(\mathbf{S}_i) - Y_i\right\}$$

where $\widehat{\pi}(\mathbf{s}) = \widehat{P}(A_i | \mathbf{S}_i = \mathbf{s})$ may be estimated using a model similar to the one used for $\widehat{\mu}_1(\mathbf{s})$. Furthermore, our cross-fitting approach could be used to facilitate the use of machine learning methods in the estimation of $\widehat{\mu}_1(\cdot), \widehat{\pi}(\cdot)$, to include covariates to control for confounding, or to simplify or strengthen asymptotic results – e.g., results for the

kernel estimator used in [1] obtain rates of convergence of $n^{-\frac{1}{2}}$ only under very limited technical conditions.

Our proposed robust approach to examine the strength of a high-dimensional surrogate will be useful in practice as there is increased attention on utilizing high-dimensional data that can be measured earlier or with less cost compared to primary outcomes, in an effort to make conclusions about treatments and interventions sooner. We include software to implement our proposed methods in the R package `crossurr` available at `github.com/denisagniel/crossurr`.

References

[1] Denis Agniel and Layla Parast. Evaluation of longitudinal surrogate markers. *Biometrics*, 2020.

[2] Susan Athey, Raj Chetty, Guido Imbens, and Hyunseung Kang. Estimating Treatment Effects using Multiple Surrogates: The Role of the Surrogate Score and the Surrogate Index. 2016.

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

[4] Erica J Caveney and Oren J Cohen. Diabetes and biomarkers. *Journal of diabetes science and technology*, 5(1):192–197, 2011.

[5] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/Neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–265, 2017.

[6] Sung Hee Choi, Tae Hyuk Kim, Soo Lim, Kyong Soo Park, Hak C Jang, and Nam H Cho. Hemoglobin a1c as a diagnostic tool for diabetes screening and new-onset diabetes prediction: a 6-year community-based prospective study. *Diabetes care*, 34(4):944–949, 2011.

[7] Eric de Groot, G Kees Hovingh, Albert Wiegman, Patrick Duriez, Andries J Smit, Jean-Charles Fruchart, and John JP Kastelein. Measurement of arterial wall thickness as a surrogate marker for atherosclerosis. *Circulation*, 109(23_suppl_1):III–33, 2004.

[8] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[9] Diabetes Prevention Program Research Group et al. 10-year follow-up of diabetes incidence and weight loss in the diabetes prevention program outcomes study. *The Lancet*, 374(9702):1677–1686, 2009.

[10] Marshall M Joffe and Tom Greene. Related causal frameworks for surrogate outcomes. *Biometrics*, 65(2):530–538, 2009.

[11] Kejal Kantarci and Clifford R Jack. Quantitative magnetic resonance techniques as surrogate markers of alzheimer's disease. *NeuroRx*, 1(2):196–205, 2004.

[12] Allan Levey, James Lah, Felicia Goldstein, Kyle Steenland, and Donald Bliwise. Mild cognitive impairment: an opportunity to identify patients at high risk for progression to alzheimer's disease. *Clinical therapeutics*, 28(7):991–1001, 2006.

[13] Chester A Mathis, Brian J Lopresti, and William E Klunk. Impact of amyloid imaging on drug development in alzheimer's disease. *Nuclear medicine and biology*, 34(7):809–822, 2007.

[14] Nicolai Meinshausen. Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1):374–393, 2007.

[15] Erica EM Moodie, Olli Saarela, and David A Stephens. A doubly robust weighting estimator of the average treatment effect on the treated. *Stat*, 7(1):e205, 2018.

[16] Christian Obirikorang, Lawrence Quaye, and Isaac Acheampong. Total lymphocyte count as a surrogate marker for cd4 count in resource-limited settings. *BMC Infectious Diseases*, 12(1):1–5, 2012.

[17] Layla Parast, Tianxi Cai, and Lu Tian. Evaluating multiple surrogate markers with censored data. *Biometrics*.

[18] Layla Parast, Tianxi Cai, and Lu Tian. Evaluating surrogate marker information using censored data. *Statistics in Medicine*, 36(11):1767–1782, 2017.

[19] Layla Parast, Tianxi Cai, and Lu Tian. Using a surrogate marker for early testing of a treatment effect. *Biometrics*, 75(4):1253–1263, 2019.

[20] Layla Parast, Mary M McDermott, and Lu Tian. Robust estimation of the proportion of treatment effect explained by surrogate marker information. *Statistics in medicine*, 35(10):1637–1653, 2016.

[21] Layla Parast, Lu Tian, and Tianxi Cai. Assessing the value of a censored surrogate outcome. *Lifetime data analysis*, pages 1–21, 2019.

[22] Eric Polley, Erin LeDell, Chris Kennedy, Sam Lendle, and Mark van der Laan. Package 'superlearner', 2019.

[23] Brenda L Price, Peter B Gilbert, and Mark J van der Laan. Estimation of the optimal surrogate based on a randomized trial. *Biometrics*, 74(4):1271–1281, 2018.

[24] Alexander Rix and Yanyi Song. *bama: High Dimensional Bayesian Mediation Analysis*, 2020. R package version 1.0.1.

[25] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

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

[27] Gary W Small. Diagnostic issues in dementia: neuroimaging as a surrogate marker of disease. *Journal of geriatric psychiatry and neurology*, 19(3):180–185, 2006.

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

[29] Yanyi Song, Xiang Zhou, Min Zhang, Wei Zhao, Yongmei Liu, Sharon LR Kardia, Ana V Diez Roux, Belinda L Needham, Jennifer A Smith, and Bhramar Mukherjee. Bayesian shrinkage estimation of high dimensional causal mediation effects in omics studies. *Biometrics*, 2018.

[30] Antonio L Teixeira, Breno S Diniz, Alline C Campos, Aline S Miranda, Natalia P Rocha, Leda L Talib, Wagner F Gattaz, and Orestes V Forlenza. Decreased levels of circulating adiponectin in mild cognitive impairment and alzheimer's disease. *Neuromolecular medicine*, 15(1):115–121, 2013.

[31] Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

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

[33] Siruo Wang, Tyler H McCormick, and Jeffrey T Leek. Methods for correcting inference based on outcomes predicted by machine learning. *Proceedings of the National Academy of Sciences*, 2020.

[34] Xuan Wang, Layla Parast, Lu Tian, and Tianxi Cai. Model-free approach to quantifying the proportion of treatment effect explained by a surrogate marker. *Biometrika*, 107(1):107–122, 2020.

[35] Yue Wang and Jeremy MG Taylor. A measure of the proportion of treatment effect explained by a surrogate marker. *Biometrics*, 58(4):803–812, 2002.

[36] Cun-Hui Zhang et al. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.

[37] Haixiang Zhang, Yinan Zheng, Zhou Zhang, Tao Gao, Brian Joyce, Grace Yoon, Wei Zhang, Joel Schwartz, Allan Just, Elena Colicino, et al. Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics*, 32(20):3150–3154, 2016.

[38] Yinan Zheng, Haixiang Zhang, Lifang Hou, and Lei Liu. *HIMA: High-Dimensional Mediation Analysis*, 2018. R package version 1.0.7.

[39] Ruixuan Rachel Zhou, Liewei Wang, and Sihai Dave Zhao. Estimation and inference for the indirect effect in high-dimensional linear mediation models. *Biometrika*, 107(3):573–589, 2020.

7 Appendix

7.1 Details for example depicted in Figure 1

In this small simulation, we let $n = n^* = 1000$, and we allowed A to be randomized with $P(A = 1) = 0.5$. The rest of the data were generated as follows: $W = \delta A - 1 + \epsilon_w$, $S = \delta W + \epsilon_s, Y = 1.5A + W + \epsilon_y$. All three errors were

independently generated as $\epsilon_w \sim N(0,1), \epsilon_s \sim N(0,1), \epsilon_y \sim N(0, 0.25\sqrt{\delta+1})$. The surrogate $\psi_a(s)$ was taken to be an estimate of $E(Y|A=a, S=s)$ obtained as the predicted value from a least squares fit of Y on S in treatment group $A=a$ from study 1. Figure 1 depicts δ values of 0 (PTE 0), 0.25 (PTE 0.2), 0.5 (PTE 0.33), and 3 (PTE 0.75).

7.2 Regularity conditions

We follow the results of [5] and require minimal regularity conditions on the distribution of $\mathbf{O}$. Specifically, we require that for some $q_1, q_2 > 4$ $C_1, C_2 > 0$, $\widetilde{C} = \min\{C_1, C_2\}$, and $a = 0, 1$

$$[E\{|\mu_a(\mathbf{X},\mathbf{S})|^{q_1}\}]^{1/q_1} \leq C_1, [E\{|m_a(\mathbf{X})|^{q_2}\}]^{1/q_2} \leq C_2,$$

and further that $\{E(|Y|^{q_1})\}^{1/q_1} \leq \widetilde{C}$. We finally assume that errors are well behaved, for $c_1, c_2 > 0, a = 0, 1$:

$$\begin{aligned}
&E\left[\{Y-\mu_A(\mathbf{X},\mathbf{S})\}^2|\mathbf{X},\mathbf{S}\right] \leq C_1, E\left[\{Y-m_A(\mathbf{X})\}^2|\mathbf{X}\right] \leq C_2\\
&E\left[\{Y-\mu_A(\mathbf{X},\mathbf{S})\}^2\right] > c_1, E\left[\{Y-m_A(\mathbf{X})\}^2\right] > c_2\\
&E\left[\{A-\pi(\mathbf{X},\mathbf{S})\}^2\right] > c_1, E\left[\{A-e(\mathbf{X})\}^2\right] > c_2.
\end{aligned}$$

7.3 Justification of perturbation resampling

In this section, we will show that the limiting distribution (conditional on the observed data) of the resampled $\widehat{R}^*_{\mathbf{S}b}$ is the same as the limiting distribution of $\widehat{R}_{\mathbf{S}}$. Recall that

$$\begin{aligned}
\widehat{\Delta}^*_b &= n^{-1}\sum_{i=1}^n \mathcal{G}_{ib}\widehat{\phi}_{1i}\\
\widehat{\Delta}^*_{\mathbf{S}b} &= n^{-1}\sum_{i=1}^n \mathcal{G}_{ib}\widehat{\phi}_{2i},
\end{aligned}$$

and let $\widehat{\boldsymbol{\theta}}^*_b = (\widehat{\Delta}^*_b, \widehat{\Delta}^*_{\mathbf{S}b})^\intercal$. Therefore, $E(\widehat{\boldsymbol{\theta}}^*_b|\mathcal{O}) = \widehat{\boldsymbol{\theta}}$ and $\text{cov}(\widehat{\boldsymbol{\theta}}^*_b|\mathcal{O}) = n^{-2}\sum_{i=1}^n \widehat{\boldsymbol{\phi}}_i\widehat{\boldsymbol{\phi}}_i^\intercal = n^{-1}\widehat{\Sigma}$. Therefore,

$$\sqrt{n}\widehat{\Sigma}^{-\frac{1}{2}}(\widehat{\boldsymbol{\theta}}^*_b - \widehat{\boldsymbol{\theta}})\Big|\,\mathcal{O} \to N(\mathbf{0}, I),$$

and the delta method yields the result

$$\sqrt{n}\widehat{\sigma}^{-1}(\widehat{R}^*_{\mathbf{S}b} - \widehat{R}_{\mathbf{S}})\Big|\,\mathcal{O} \to N(0,1).$$

7.4 Implementation details for simulations

Here we provide further details on the implementation of the proposed estimator in simulation.

In using Super Learner to estimate nuisance functions, we specified the following candidate learners. For $\mu_1(\mathbf{X},\mathbf{S}), \mu_0(\mathbf{X},\mathbf{S})$, we included the following learners: the mean, the lasso, ridge regression, ordinary least squares, support vector ma-

chines (SVM), and random forests. For $\pi(\mathbf{X}, \mathbf{S})$, we included: the mean, the lasso, logistic regression, linear discriminant analysis, quadratic discriminant analysis, SVM, and random forests. For $m_1(\mathbf{X}), m_0(\mathbf{X})$, we additionally considered a generalized linear model including all interaction terms, step-wise regression including all interaction terms, and regression trees. And for $e(\mathbf{X})$, we additionally included logistic regression with all interaction terms, classification trees, and k-nearest neighbors. We took $K = 4$ for cross-fitting, and we used the default cross-validation procedure to select the tuning parameters for the candidate learners and the Super Learner.

For the DR-lasso estimator, we used the default cross-validation to select the tuning parameters.