Demystifying a class of multiply robust estimators

BY WEI LI
School of Mathematical Sciences, Peking University, Beijing, 100871, P.R.C.
weylpeking@pku.edu.cn

YUWEN GU
Department of Statistics, University of Connecticut, Storrs, Connecticut, 06269, U.S.A.
yuwen.gu@uconn.edu

AND LAN LIU
School of Statistics, University of Minnesota, Minneapolis, Minnesota, 55455, U.S.A.
liux3771@umn.edu

SUMMARY

When estimating the population mean of a response variable subject to ignorable missingness, a new class of methods called the multiply robust procedures has been proposed. The advantage of the multiply robust procedures over the traditional doubly robust methods is that the former permit the use of multiple candidate models for both the propensity score and the outcome regression, and the multiply robust estimators are consistent if any one of the multiple models is correctly specified. Such a property is termed multiple robustness. Somewhat surprisingly, we show that these multiply robust estimators are special cases of the doubly robust estimators where the final propensity score and outcome regression models are certain combinations of the candidate models. To further improve model specifications in the doubly robust estimators, we adapt a model mixing procedure as an alternative method to combine multiple candidate models. We show that the multiple robustness property and asymptotic normality can be also achieved by our mixing-based doubly robust estimator. In addition, our estimator and the established theoretical properties are not confined to parametric models. Numerical examples further demonstrate that our proposed estimator is comparable to or outperforms existing multiply robust estimators.

Some key words: Double robustness; Ignorable missingness; Model mixing; Multiple robustness.

1. INTRODUCTION

The missing data problem is commonly encountered in biomedical and socioeconomic studies. Missing data can arise when patients fail to visit their physicians for an annual checkup, when respondents refuse to answer private survey questions, or when data is damaged. In this paper, we focus on the missingness of response data and our parameter of interest is the population mean of the response variable. Additionally, we assume the missingness mechanism is ignorable, i.e., the missingness can be fully accounted for by observed information (Little & Rubin, 2002).

Various methods have been proposed to estimate the population mean of a response subject to ignorable missingness. Amongst them, two commonly used methods are inverse probability weighting estimation (Horvitz & Thompson, 1952; Rosenbaum & Rubin, 1983; Robins et al.,

1994) and outcome regression estimation. These two methods model the missingness mechanism, also known as the propensity score, and the outcome, respectively, and provide consistent estimation only if the corresponding model is correctly specified.

Over the past two decades, doubly robust procedures have gained popularity due to their protection against model misspecification (Robins et al., 1994, 1995; Scharfstein et al., 1999; Bang & Robins, 2005; Tan, 2007, 2010; Tsiatis, 2006; van der Laan & Gruber, 2010; Kang & Schafer, 2007; Qin et al., 2008; Tsiatis et al., 2011; Rotnitzky et al., 2012). Robins et al. (1994, 1995) proposed augmented estimating equations to construct the doubly robust estimator. A nice feature of such an estimator is that it involves one propensity score and one outcome regression model, and provides consistent estimation of the population mean of the response when either the propensity score or the outcome regression is correctly specified but not necessarily both. Furthermore, the augmented estimator achieves semiparametric efficiency in the joint space of both propensity score and outcome regression models being correctly specified. Robins (2000, 2002), van der Laan & Robins (2003), and Kang & Schafer (2007) proposed a number of alternative doubly robust estimators. Those doubly robust estimators have different finite sample properties but they all achieve semiparametric efficiency when the parameter of interest is the population mean of a response variable.

Recently, multiply robust estimators have been proposed to gain additional protection against model misspecification (Han & Wang, 2013; Han, 2014a,b; Chan & Yam, 2014; Chan, 2013; Han, 2016; Duan, 2017; Chen & Haziza, 2017). In contrast to the doubly robust estimator which includes just one propensity score and one outcome regression model, Han & Wang (2013) proposed a weighting-based multiply robust estimator that includes multiple propensity score and outcome regression models. The motivation is that, with an unknown data-generating process, there is no guarantee that either of the two models in the doubly robust estimator is correctly specified. They showed that the proposed estimator has the multiple robustness property which guarantees estimation consistency if any one of the candidate models for the propensity score or outcome regression is correctly specified. To overcome the computational challenges such as multiple roots and nonconvergence of the weighting-based multiply robust estimator, Han (2014a,b) redefined the estimator through a convex minimization problem. Chan & Yam (2014) generalized the weighting method in Han & Wang (2013) and considered calibration weighting using generalized empirical likelihood. Chan (2013) proposed an outcome regression-based multiply robust estimator where the inverse of the estimated propensity scores are used as covariates in the outcome regression model.

In this paper, we demonstrate that the aforementioned multiply robust estimators are special cases of doubly robust estimators where the final propensity score and outcome regression models are combinations of those multiple candidate models. The intuition is straightforward: for ignorable missingness data, the likelihood function factorizes as the product of two variation-independent factors where one corresponds to the propensity score and the other corresponds to the outcome regression. The parameter of interest, as a function of the likelihood, can thus be viewed as a function of the above two likelihood factors. Hence, for both doubly and multiply robust estimators, the propensity score and outcome regression are the only two parts in the likelihood that are modeled, irrespective of whether each part is constructed from a single model or a combination of a variety of candidate models. Additionally, the multiple robustness property is based on the double robustness property. The final propensity score and outcome regression models are constructed such that when any of the multiple candidate models are correctly specified, at least one of the estimated final models is consistent for the corresponding true model. Along these lines, we adapt the model mixing procedures (Yang, 2000, 2001) to assemble multiple candidate models, so as to improve model specifications for both propensity score and outcome regression in doubly robust estimators. We show that, theoretically, the improved doubly robust estimators also possess the multiple robustness property, and empirically, these doubly robust estimators can rival or outperform the aforementioned multiply robust estimators.

2. Preliminaries

Let $X \in \mathbb{R}^d$ denote the fully observed covariates and let Y be the response variable that is subject to missingness. Let R denote the missingness indicator, that is, $R = 1$ if Y is observed and $R = 0$ if otherwise. We use lower case letters to denote realizations of these random variables. For example, x denotes any possible value that X could take. Suppose we observe n independent and identically distributed copies of (R, RY, X). Without loss of generality, we assume Y is observed in the first m out of a total of n observations. Let $\pi(x) = \text{pr}(R = 1|X = x)$ and $a(x) = E(Y|X = x)$ denote the propensity score and outcome regression models, respectively. The parameter of interest is the population mean of the response, i.e., $\mu = E(Y)$. Throughout the paper, we use $\mathbb{P}_n(U) = \sum_{i=1}^n U_i/n$ to denote the empirical mean of a generic random variable U. We make the following assumptions.

Assumption 1. Ignorability: $R \perp\!\!\!\perp Y|X$.

Assumption 2. Positivity: $\pi(x) = \text{pr}(R = 1|X = x) \geq \pi_{\min} > 0$ for all x and some $\pi_{\min}$.

Assumption 1 requires that, given observed covariates, the missingness indicator is independent of the response variable. Such an assumption requires the collection of all relevant covariates. Assumption 2 requires that, for every individual, the missingness of the outcome is not deterministic, i.e., there is a positive chance of observing the outcome. This condition is needed to recover the missing information from the observed data. Both assumptions are frequently made in the missing data literature.

When $\pi(x)$ is consistently estimated by $\hat{\pi}(x)$, the inverse probability weighting estimator $\hat{\mu}_{\text{ipw}} = \mathbb{P}_n\{RY/\hat{\pi}(X)\}$ is consistent for μ. Otherwise, $\hat{\mu}_{\text{ipw}}$ is inconsistent. Similarly, the outcome regression estimator $\hat{\mu}_{\text{reg}} = \mathbb{P}_n\{\hat{a}(X)\}$ is consistent for μ if $\hat{a}(x)$ consistently estimates $a(x)$. Otherwise, $\hat{\mu}_{\text{reg}}$ is inconsistent. To overcome the vulnerability of model misspecification in both $\hat{\mu}_{\text{ipw}}$ and $\hat{\mu}_{\text{reg}}$, the augmented inverse probability weighted estimator $\hat{\mu}_{\text{aipw}} = \mathbb{P}_n[RY/\hat{\pi}(X) + \{1 - R/\hat{\pi}(X)\}\hat{a}(X)]$ uses the outcome regression model as an augmentation term to the inverse probability weighting method. When the propensity score model $\pi(x)$ is consistently estimated, the first term in $\hat{\mu}_{\text{aipw}}$ is consistent for μ and the second term converges to zero in probability so that the summation is consistent for μ. When $\pi(x)$ is not consistently estimated but the outcome regression model $a(x)$ is, the first term is biased but the second term consistently estimates the bias of the first term so the summation is again consistent for μ. Hence, $\hat{\mu}_{\text{aipw}}$ is doubly robust. The propensity score and outcome regression in the augmented inverse probability weighted estimator can be estimated using parametric, semiparametric or nonparametric models.

Apart from the augmented inverse probability weighted estimator, various doubly robust estimators have been proposed including the weighting-based and regression-based doubly robust estimators. Let $\tilde{\pi}(x;\delta)$ and $\tilde{a}(x;\psi)$ denote some generic parametric propensity score and outcome regression models. Based on $\hat{\mu}_{\text{ipw}}$, Rotnitzky & Robins (1995) proposed the weighting-based doubly robust estimator. Recall that the parametric inverse probability weighting estimator is $\hat{\mu}_{\text{ipw}} = \mathbb{P}_n\{RY/\tilde{\pi}(X;\hat{\delta})\}$, where $\hat{\delta}$ can be obtained by solving the estimating equation

$$\mathbb{P}_n\left[\left\{\frac{R}{\tilde{\pi}(X;\delta)} - 1\right\}t(X)\right] = 0, \qquad (1)$$

and $t(x)$ is a user-specified vector-valued function with dimension no less than that of δ. When one of the components in $t(x)$ is $\tilde{a}(x;\hat{\psi})$, where $\tilde{a}(x;\hat{\psi})$ consistently estimates $a(x)$, the inverse probability weighting estimator $\hat{\mu}_{\text{ipw}}$ is consistent for μ even when the propensity score model is misspecified (Rotnitzky & Robins, 1995). We denote this weighting-based doubly robust estimator by $\hat{\mu}_{\text{dr–ipw}}$. Under (1), $\hat{\mu}_{\text{dr–ipw}}$ can be written as $\hat{\mu}_{\text{dr–ipw}} = \mathbb{P}_n\{RY/\tilde{\pi}(X;\hat{\delta})\} = \mathbb{P}_n[RY/\tilde{\pi}(X;\hat{\delta}) + \{1 - R/\tilde{\pi}(X;\hat{\delta})\}\tilde{a}(X;\hat{\psi})]$, i.e., $\hat{\mu}_{\text{dr–ipw}}$ is in the form of an augmented inverse probability weighted estimator. Hence, it shares the double robustness property. More generally, if the linear space spanned by $t(x)$ contains $\tilde{a}(x;\hat{\psi})$, then the double robustness property of $\hat{\mu}_{\text{dr–ipw}}$ still holds. For notational convenience, we use $\hat{\mu}_{\text{dr–ipw}}$ to denote all such weighting-based doubly robust estimators. Thus, if there exists a constant vector ξ such that $\tilde{a}(x;\hat{\psi}) = \xi^{\mathrm{T}}t(x)$, then $\hat{\mu}_{\text{dr–ipw}} = \mathbb{P}_n\{RY/\tilde{\pi}(X;\hat{\delta})\}$ with $\hat{\delta}$ obtained from (1), is consistent for μ if either $\tilde{\pi}(x;\hat{\delta})$ consistently estimates $\pi(x)$ or $t(x)$ includes consistent estimator(s) of $a(x)$ up to a linear combination.

Another doubly robust estimator based on $\hat{\mu}_{\text{reg}}$ is proposed by Scharfstein et al. (1999). Also, see Bang & Robins (2005). Scharfstein et al. (1999) showed that to obtain a doubly robust estimator, it suffices to model the outcome regression as $\tilde{a}(x;\psi) = b\{\zeta(x;\psi_1) + \psi_2 \cdot 1/\tilde{\pi}(x;\hat{\delta})\}$, where $\psi = (\psi_1^{\mathrm{T}}, \psi_2)^{\mathrm{T}}$, $b(\cdot)$ is a link function, $\zeta(x;\psi_1)$ is a known function of x with parameter ψ_1, and $1/\tilde{\pi}(x;\hat{\delta})$ denotes the inverse of the estimated propensity score. They then defined the regression-based doubly robust estimator as $\hat{\mu}_{\text{dr–reg}} = \mathbb{P}_n\{\tilde{a}(X;\hat{\psi})\}$, where $\hat{\psi}$ is obtained by solving $\mathbb{P}_n[R\{Y - \tilde{a}(X;\psi)\}\{q(X)^{\mathrm{T}}, 1/\tilde{\pi}(X;\hat{\delta})\}^{\mathrm{T}}] = 0$, and $q(x)$ is a user-specified vector-valued function with dimension no less than that of ψ_1. Under the above constraint, $\hat{\mu}_{\text{dr–reg}}$ can be rewritten as $\hat{\mu}_{\text{dr–reg}} = \mathbb{P}_n\{\tilde{a}(X;\hat{\psi})\} = \mathbb{P}_n[\tilde{a}(X;\hat{\psi}) + R\{Y - \tilde{a}(X;\hat{\psi})\}/\tilde{\pi}(X;\hat{\delta})] = \mathbb{P}_n[RY/\tilde{\pi}(X;\hat{\delta}) + \{1 - R/\tilde{\pi}(X;\hat{\delta})\}\tilde{a}(X;\hat{\psi})]$, that is, $\hat{\mu}_{\text{dr–reg}}$ is in the form of an augmented inverse probability weighted estimator. Based on a similar argument which shows the consistency of $\hat{\mu}_{\text{aipw}}$, we can also establish the double robustness property of $\hat{\mu}_{\text{dr–reg}}$ if either $\tilde{\pi}(x;\delta)$ or $\tilde{a}(x;\psi)$ is correctly specified. More generally, one could include in the outcome regression model a vector-valued function $l(x)$, that is, one can set $\tilde{a}(x;\psi) = b\{\zeta(x;\psi_1) + \psi_2^{\mathrm{T}}l(x)\}$ and solve $\psi = (\psi_1^{\mathrm{T}}, \psi_2^{\mathrm{T}})^{\mathrm{T}}$ from the estimating equation

$$\mathbb{P}_n\big[R\{Y - \tilde{a}(X;\psi)\}\{q(X)^{\mathrm{T}}, l(X)^{\mathrm{T}}\}^{\mathrm{T}}\big] = 0. \qquad (2)$$

If the linear space spanned by $l(x)$ contains the inverse of the correctly specified propensity score estimate $1/\tilde{\pi}(x;\hat{\delta})$, then following the same argument as above, we can also show that the corresponding estimator achieves the double robustness property.

3. Multiply Robust Estimators as Special Cases of Doubly Robust Estimators

3·1. *Demystifying weighting-based multiply robust estimator*

To provide additional protection against model misspecification, Han & Wang (2013), Chan & Yam (2014) and Chan (2013) postulated multiple candidate parametric models for the propensity score and outcome regression and proposed different versions of multiply robust estimators of μ. Interestingly, we show that these estimators can be written as special cases of doubly robust estimators. Furthermore, the multiple robustness property is actually a result of the double robustness property.

Han & Wang (2013) postulated multiple candidate parametric models $\mathcal{P} = \{\pi^j(x;\alpha^j) : j = 1,\ldots,J\}$ for $\pi(x)$ and multiple candidate parametric models $\mathcal{A} = \{a^k(x;\gamma^k) : k = 1,\ldots,K\}$

for $a(x)$, where α^j and γ^k are the parameters in the jth propensity score and kth outcome regression models, respectively. Let $\hat{\alpha}^j$ and $\hat{\gamma}^k$ denote the estimates of α^j and γ^k, respectively. Han & Wang (2013) proposed to estimate μ via a weighted average of the responses $\sum_{i=1}^m \hat{\omega}_i Y_i$, where the following constraints are imposed on the weights $\hat{\omega}_i$ $(i = 1, \cdots, m)$:

$$\begin{gathered}\mathbb{P}_n\{nR\omega\} = 1, \quad \mathbb{P}_n\{nR\omega\pi^j(X;\hat{\alpha}^j)\} = \mathbb{P}_n\{\pi^j(X;\hat{\alpha}^j)\} \quad (j = 1, \dots, J), \\ \mathbb{P}_n\{nR\omega a^k(X;\hat{\gamma}^k)\} = \mathbb{P}_n\{a^k(X;\hat{\gamma}^k)\} \quad (k = 1, \dots, K).\end{gathered} \tag{3}$$

To give an explicit form of the estimates $\hat{\omega}_i$'s, let $\alpha^{\mathrm{T}} = \{(\alpha^1)^{\mathrm{T}}, \dots, (\alpha^J)^{\mathrm{T}}\}$, $\gamma^{\mathrm{T}} = \{(\gamma^1)^{\mathrm{T}}, \dots, (\gamma^K)^{\mathrm{T}}\}$, and denote the estimates of α and γ by $\hat{\alpha}$ and $\hat{\gamma}$, respectively. Moreover, let $g(x;\hat{\alpha},\hat{\gamma}) = \{\pi^1(x;\hat{\alpha}^1) - \hat{\theta}^1, \dots, \pi^J(x;\hat{\alpha}^J) - \hat{\theta}^J, a^1(x;\hat{\gamma}^1) - \hat{\eta}^1, \dots, a^K(x;\hat{\gamma}^K) - \hat{\eta}^K\}^{\mathrm{T}}$, where $\hat{\theta}^j = \mathbb{P}_n\{\pi^j(X;\hat{\alpha}^j)\}$ and $\hat{\eta}^k = \mathbb{P}_n\{a^k(X;\hat{\gamma}^k)\}$. Han & Wang (2013) gave the estimated weights as

$$\hat{\omega}_i = \frac{1}{m}\frac{1}{1+\hat{\rho}^{\mathrm{T}}g(X_i;\hat{\alpha},\hat{\gamma})} \Big/ \left\{\frac{1}{m}\sum_{i=1}^m \frac{1}{1+\hat{\rho}^{\mathrm{T}}g(X_i;\hat{\alpha},\hat{\gamma})}\right\} \quad (i = 1, \dots, m), \tag{4}$$

where $\hat{\rho}$ is the estimate of a $(J + K)$-dimensional parameter satisfying

$$\sum_{i=1}^m \frac{g(X_i;\hat{\alpha},\hat{\gamma})}{1+\hat{\rho}^{\mathrm{T}}g(X_i;\hat{\alpha},\hat{\gamma})} = 0. \tag{5}$$

To facilitate numerical computation and guarantee uniqueness of $\hat{\rho}$, Han (2014a,b) further imposed the condition $1 + \rho^{\mathrm{T}}g(X_i;\hat{\alpha},\hat{\gamma}) > 0$ for $i = 1, \cdots, m$, and solved ρ from the minimization of a convex function $F(\rho) = -\mathbb{P}_n[R\log\{1 + \rho^{\mathrm{T}}g(X;\hat{\alpha},\hat{\gamma})\}]$. We denote the weighting-based multiply robust estimator by $\hat{\mu}_{\text{mr–ipw}} := \sum_{i=1}^m \hat{\omega}_i Y_i$.

In the sequel, we show that $\hat{\mu}_{\text{mr–ipw}}$ can be written as the weighting-based doubly robust estimator $\hat{\mu}_{\text{dr–ipw}}$ with a specific propensity score model $\tilde{\pi}(x;\delta)$, where δ can be estimated from (1) with a specific choice of $t(x)$. We rewrite the second and third constraints in (3) as $\mathbb{P}_n[\{nR\omega - 1\}h^s(X;\hat{\alpha},\hat{\gamma})] = 0$, where $h^s(x;\hat{\alpha},\hat{\gamma}) = \pi^s(x;\hat{\alpha}^s)$ for $s = 1, \dots, J$ and $h^s(x;\hat{\alpha},\hat{\gamma}) = a^{s-J}(x;\hat{\gamma}^{s-J})$ for $s = J+1, \dots, J+K$. Comparing estimator $\hat{\mu}_{\text{mr–ipw}}$ with $\hat{\mu}_{\text{dr–ipw}}$, and the constraint rewritten above with (1), we set

$$\tilde{\pi}(x;\delta) = \left\{\frac{1}{n}\sum_{i=1}^n \frac{R_i}{1+\delta^{\mathrm{T}}g(X_i;\hat{\alpha},\hat{\gamma})}\right\}\{1 + \delta^{\mathrm{T}}g(x;\hat{\alpha},\hat{\gamma})\}, \tag{6}$$

where δ is estimated from

$$\mathbb{P}_n\left[\left\{\frac{R}{\tilde{\pi}(X;\delta)} - 1\right\}h^s(X;\hat{\alpha},\hat{\gamma})\right] = 0 \quad (s = 1, \dots, J+K). \tag{7}$$

A comparison between (1) and (7) reveals that (7) is a special case of (1), where the base function $t(x)$ is chosen as

$$t(x) = \{\pi^1(x;\hat{\alpha}^1), \dots, \pi^J(x;\hat{\alpha}^J), a^1(x;\hat{\gamma}^1), \dots, a^K(x;\hat{\gamma}^K)\}^{\mathrm{T}}. \tag{8}$$

Consequently, a similar constraint on δ as that in (5) for ρ can be derived from (7). See the supplementary material for details. If we further assume that $1 + \delta^{\mathrm{T}}g(X_i;\hat{\alpha},\hat{\gamma}) > 0$ for $i = 1, \cdots, m$, then we can also obtain a unique $\hat{\delta}$ satisfying (7) and we have $\hat{\delta} = \hat{\rho}$. Additionally, for $i = 1, \ldots, m$, we have

$$\tilde{\pi}(X_i; \hat{\delta}) = \frac{1}{n}\left\{\sum_{i=1}^{m} \frac{1}{1 + \hat{\delta}^{\mathrm{T}} g(X_i; \hat{\alpha}, \hat{\gamma})}\right\}\{1 + \hat{\delta}^{\mathrm{T}} g(X_i; \hat{\alpha}, \hat{\gamma})\} = \frac{1}{n\hat{\omega}_i}.$$

This implies that $\hat{\mu}_{\text{mr–ipw}} = \sum_{i=1}^{m} \hat{\omega}_i Y_i = \sum_{i=1}^{m} Y_i / \{n\tilde{\pi}(X_i; \hat{\delta})\} = \mathbb{P}_n\{RY/\tilde{\pi}(X; \hat{\delta})\} = \hat{\mu}_{\text{dr–ipw}}$. Thus, the multiply robust estimator $\hat{\mu}_{\text{mr–ipw}}$ can be written as a doubly robust estimator $\hat{\mu}_{\text{dr–ipw}}$, where the form of the final propensity score $\tilde{\pi}(x; \delta)$ is given in (6) and δ is estimated from (7).

Now we show that the multiple robustness property of $\hat{\mu}_{\text{mr–ipw}}$ is actually a result of the double robustness property of $\hat{\mu}_{\text{dr–ipw}}$. Specifically, the multiple robustness property of $\hat{\mu}_{\text{mr–ipw}}$ is achieved by the construction of a final propensity score model and a final outcome regression model in $\hat{\mu}_{\text{dr–ipw}}$ through (6)–(7). Recall that $\hat{\mu}_{\text{dr–ipw}}$ is consistent if the outcome regression model is correctly specified up to a linear combination of $t(x)$. Such a property has the potential to achieve multiple robustness with multiple outcome regression models. However, this has never been fully leveraged until Han & Wang (2013). Furthermore, they proposed a new way to gain multiple robustness with multiple candidate propensity score models. The empirical likelihood methodology they considered motivated us to construct a special propensity score model in (6). As we show below, such a construction makes at least one of the two final models consistent when any of the multiple candidate models is correctly specified.

Indeed, if one of the K candidate models for the outcome is correctly specified, then the span of $t(x)$ in (8) contains the correct model, and additionally, the estimator $\hat{\mu}_{\text{dr–ipw}}$, and consequently $\hat{\mu}_{\text{mr–ipw}}$ are both consistent. Otherwise, it suffices to show that when one of the J candidate models for the propensity score is correctly specified, $\tilde{\pi}(x; \hat{\delta})$ is consistent for the true propensity score, and hence the estimator $\hat{\mu}_{\text{mr–ipw}}$ is also consistent. Without loss of generality, assume $\pi^1(x; \alpha^1)$ is correctly specified, i.e., $\pi(x) = \pi^1(x; \alpha_0)$ for some α_0. Let $\theta^1 = E\{\pi^1(X; \alpha_0)\}$. Following a similar derivation in Han & Wang (2013) and Han (2014a,b), one has $\hat{\delta} \to (1/\theta^1, 0, \ldots, 0)^{\mathrm{T}}$ in probability as $n \to \infty$. Thus, we have

$$\begin{aligned}
\tilde{\pi}(x; \hat{\delta}) &= \left\{\frac{1}{n}\sum_{i=1}^{n} \frac{R_i}{1 + \hat{\delta}^{\mathrm{T}} g(X_i; \hat{\alpha}, \hat{\gamma})}\right\}\{1 + \hat{\delta}^{\mathrm{T}} g(x; \hat{\alpha}, \hat{\gamma})\} \\
&\to E\left[\frac{R}{1 + \{\pi^1(X; \alpha_0) - \theta^1\}/\theta^1}\right][1 + \{\pi^1(x; \alpha_0) - \theta^1\}/\theta^1] \\
&= \theta^1[1 + \{\pi^1(x; \alpha_0) - \theta^1\}/\theta^1] = \pi^1(x; \alpha_0),
\end{aligned}$$

which holds in probability. Therefore, we have demonstrated that the weighting-based multiply robust estimator is indeed a special case of the doubly robust estimator.

3·2. *Demystifying other multiply robust estimators*

The calibration weighted multiply robust estimator, proposed by Chan & Yam (2014), aims to overcome the difficulty in numerical computation of the weighting-based multiply robust estimator by Han & Wang (2013). Let $\hat{\mu}_{\text{mr–cal}}$ denote the calibration-based multiply robust estimator. Following arguments similar to those in Section 3·1, $\hat{\mu}_{\text{mr–cal}}$ can be also interpreted as a special case of the doubly robust estimator.

Chan (2013) proposed a regression-based multiply robust estimator as an alternative estimator of μ. They fit a linear regression in which the response Y is regressed on the predictors $u(X; \hat{\alpha}, \hat{\gamma}) = \{1, a^1(X; \hat{\gamma}^1), \ldots, a^K(X; \hat{\gamma}^K), 1/\pi^1(X; \hat{\alpha}^1), \ldots, 1/\pi^J(X; \hat{\alpha}^1)\}^{\mathrm{T}}$ using the complete-case subsample. The regression-based estimator was then defined as $\hat{\mu}_{\text{mr–reg}} =$

$\mathbb{P}_n\{u(X;\hat{\alpha},\hat{\gamma})^{\mathrm{T}}\hat{\beta}\}$, where $\hat{\beta}$ is the least squares estimator from the linear regression, $\hat{\beta}=\{U(X;\hat{\alpha},\hat{\gamma})^{\mathrm{T}}U(X;\hat{\alpha},\hat{\gamma})\}^{-1}U(X;\hat{\alpha},\hat{\gamma})^{\mathrm{T}}Y_{\text{obs}}$, where $Y_{\text{obs}}=(Y_1,\dots,Y_m)^{\mathrm{T}}$ and $U(X;\hat{\alpha},\hat{\gamma})=\{u(X_1;\hat{\alpha},\hat{\gamma}),\dots,u(X_m;\hat{\alpha},\hat{\gamma})\}^{\mathrm{T}}$.

We show that $\hat{\mu}_{\text{mr–reg}}$ can be expressed as a form of the outcome regression-based doubly robust estimator $\hat{\mu}_{\text{dr–reg}}$. For ease of notation, let $u_1(x;\hat{\gamma})=\{1,a^1(x;\hat{\gamma}^1),\dots,a^K(x;\hat{\gamma}^K)\}^{\mathrm{T}}$, $u_2(x;\hat{\alpha})=\{1/\pi^1(X;\hat{\alpha}^1),\dots,1/\pi^J(X;\hat{\alpha}^1)\}^{\mathrm{T}}$, and $u(x;\hat{\alpha},\hat{\gamma})=\{u_1(x;\hat{\gamma})^{\mathrm{T}},u_2(x;\hat{\alpha})^{\mathrm{T}}\}^{\mathrm{T}}$. Recall that $\hat{\mu}_{\text{dr–reg}}=\mathbb{P}_n\{\tilde{a}(X;\hat{\psi})\}=\mathbb{P}_n[b\{\zeta(X;\hat{\psi}_1)+\hat{\psi}_2^{\mathrm{T}}l(X)\}]$, where $\hat{\psi}$ is estimated from estimating equation (2). If we specify in (2), $b(\cdot)$ to be the identity link, $q(x)=u_1(x;\hat{\gamma})$, $\zeta(x;\psi_1)=u_1(x;\hat{\gamma})^{\mathrm{T}}\psi_1$, and $l(x)=u_2(x;\hat{\alpha})$, then the estimator $\hat{\psi}$ obtained from (2) coincides with the least square estimator $\hat{\beta}$, i.e., $(\hat{\psi}_1^{\mathrm{T}},\hat{\psi}_2^{\mathrm{T}})^{\mathrm{T}}=\hat{\beta}$. Consequently, $\hat{\mu}_{\text{mr–reg}}=\mathbb{P}_n\{u(X;\hat{\alpha},\hat{\gamma})^{\mathrm{T}}\hat{\beta}\}=\mathbb{P}_n\{u_1(X;\hat{\gamma})^{\mathrm{T}}\hat{\psi}_1+u_2(X;\hat{\alpha})^{\mathrm{T}}\hat{\psi}_2\}=\mathbb{P}_n\{\zeta(X;\hat{\psi}_1)+\hat{\psi}_2^{\mathrm{T}}l(X)\}=\hat{\mu}_{\text{dr–reg}}$.

Similar to the weighting-based doubly robust estimator, $\hat{\mu}_{\text{dr–reg}}$ is also equipped with the multiple robustness property with multiple candidate propensity models. If one of the J candidate propensity score models is correctly specified, then the span of $l(x)$ contains the correct model, and additionally, according to the discussions below (2), the estimator $\hat{\mu}_{\text{dr–reg}}$ and consequently $\hat{\mu}_{\text{mr–reg}}$ are both consistent. Otherwise, it suffices to show that when one of the K outcome regression models is correctly specified, $\tilde{a}(x;\hat{\psi})$ is consistent for the true outcome regression. Without loss of generality, assume $a^1(x;\gamma^1)$ is correctly specified, i.e., $a^1(x;\hat{\gamma}^1)$ is a consistent estimator of $a(x)$. Note that $a^1(x;\hat{\gamma}^1)=\zeta(x;\psi_1^*)+\psi_2^{*\mathrm{T}}l(x)$, where $\psi_1^*=(0,1,0,\dots,0)^{\mathrm{T}}$ and $\psi_2^*=(0,\dots,0)^{\mathrm{T}}$. The estimators $(\hat{\psi}_1^{\mathrm{T}},\hat{\psi}_2^{\mathrm{T}})$ in $\tilde{a}(x;\hat{\psi})$ converges to $(\psi_1^{*\mathrm{T}},\psi_2^{*\mathrm{T}})$ in probability under constraint (2) with the pre-specified $q(x)$ and $l(x)$. This implies that $\tilde{a}(x;\hat{\psi})=\zeta(x;\hat{\psi}_1)+\hat{\psi}_2^{\mathrm{T}}l(x)$ converges in probability to the limit of $a^1(x;\hat{\gamma}^1)$, i.e., $\tilde{a}(x;\hat{\psi})$ consistently estimates $a(x)$. Therefore, the regression-based multiply robust estimator $\hat{\mu}_{\text{mr–reg}}$ is indeed a special case of the doubly robust estimator $\hat{\mu}_{\text{dr–reg}}$ by appropriately combining the multiple candidate models to obtain an estimate of the inverse of the final propensity score $\tilde{\pi}(x;\delta)$ through a linear span of $l(x)$, and an estimate of the final outcome regression $\tilde{a}(x;\psi)$.

4. Alternative Method to Improve Model Specification

As shown above, the multiple robustness property of existing multiply robust estimators results from the double robustness property through a specific combination of the multiple candidate models. This enlightens us to explore alternative ways to construct from these candidates a final propensity score and a final outcome regression for the improved doubly robust estimator so that it also possesses the multiple robustness property.

Yang (2000) and Yang (2001) proposed the adaptive classification/regression by mixing procedures for binary/continuous outcomes. Compared to other ensemble methods, a distinctive feature of such model mixing procedures is that the squared L_2 risks of the mixing estimators of the propensity score and outcome regression are almost as small as those of the best candidate estimators. This adaptation is achieved through a weighting scheme that exploits proper cumulative metrics, e.g., the cumulative predictive likelihood, to assess the performance of all candidates at each stage of the procedure. In what follows, we investigate the risk bounds of the final propensity score and outcome regression estimates using the mixing procedures as a foundation for developing the theoretical properties of our improved doubly robust estimator.

Let $\pi^j(x)$ denote the jth candidate propensity score model and $a^k(x)$ denote the kth candidate outcome regression model, where $j\in\mathcal{J}=\{1,\cdots,J\}$ and $k\in\mathcal{K}=\{1,\cdots,K\}$. Let Z denote the random triplet (R,RY,X). Let $\hat{\pi}_i^j(x)$ and $\hat{a}_i^k(x)$ be the estimates of $\pi^j(x)$ and

$a^k(x)$, respectively, by fitting the corresponding candidate models on the first i observations $Z^i = (R_l, R_lY_l, X_l)_{l=1}^{i}$, where $i = 1, \ldots, n$, $j \in \mathcal{J}$ and $k \in \mathcal{K}$. The candidate models are allowed to be either parametric, semiparametric, or nonparametric.

In the mixing procedure for the propensity score model, we start by randomly partitioning the full data into two parts. Without loss of generality, we denote them by $Z^{(1)} = (R_i, R_iY_i, X_i)_{i=1}^{N_n}$ and $Z^{(2)} = (R_i, R_iY_i, X_i)_{i=N_n+1}^{n}$, respectively, where $N_n = \max(1, \lfloor cn \rfloor)$ for some $0 < c < 1$. In practice, one can take, e.g., $N_n = \lfloor n/2 \rfloor$. Next, we fit each candidate propensity score model on the training set $Z^{(1)}$ and calculate its predictive risk, based on the Bernoulli likelihood, using each observation in $Z^{(2)}$. The weights for the candidate models are obtained by accumulating the predictive likelihood on $Z^{(2)}$ in the manner of $\hat{\Lambda}_j = (n - N_n)^{-1} \sum_{i=N_n+1}^{n} \hat{\Lambda}_{j,i}$, $j \in \mathcal{J}$, where $\hat{\Lambda}_{j,N_n+1} = 1/J$,

$$\hat{\Lambda}_{j,i} = \frac{\prod_{l=N_n+1}^{i-1} \hat{\pi}_{N_n}^{j}(X_l)^{R_l} \{1 - \hat{\pi}_{N_n}^{j}(X_l)\}^{1-R_l}}{\sum_{j' \in \mathcal{J}} \prod_{l=N_n+1}^{i-1} \hat{\pi}_{N_n}^{j'}(X_l)^{R_l} \{1 - \hat{\pi}_{N_n}^{j'}(X_l)\}^{1-R_l}} \quad (N_n + 2 \leq i \leq n). \tag{9}$$

Here, the weights $\hat{\Lambda}_{j,i}$ are adapted from Yang (2000). These weights satisfy $\hat{\Lambda}_j \geq 0$ and $\sum_{j \in \mathcal{J}} \hat{\Lambda}_j = 1$. The final mixing estimator of the propensity score, $\hat{\pi}_{\text{mix}}$, is a weighted average of the candidate estimators, i.e., $\hat{\pi}_{\text{mix}}(x) = \sum_{j \in \mathcal{J}} \hat{\Lambda}_j \hat{\pi}_{N_n}^{j}(x)$. We summarize the mixing procedure for the final propensity score model in Algorithm 1.

Algorithm 1. Mixing procedure for the final propensity score model.

(1) Randomly partition the full data Z^n into two parts $Z^{(1)} = (R_i, R_iY_i, X_i)_{i=1}^{N_n}$ and $Z^{(2)} = (R_i, R_iY_i, X_i)_{i=N_n+1}^{n}$, where $N_n = \max(1, \lfloor cn \rfloor)$ for some $0 < c < 1$.
(2) Fit the candidate propensity score models to the subsample $Z^{(1)}$ to obtain the estimates $\hat{\pi}_{N_n}^{j}(x), j \in \mathcal{J}$.
(3) Calculate weights $\hat{\Lambda}_j = (n - N_n)^{-1} \sum_{i=N_n+1}^{n} \hat{\Lambda}_{j,i}$, where $\hat{\Lambda}_{j,i}$ is defined in (9).
(4) Return the mixing estimator for the propensity score $\hat{\pi}_{\text{mix}}(x) = \sum_{j \in \mathcal{J}} \hat{\Lambda}_j \hat{\pi}_{N_n}^{j}(x)$.

The following lemma shows that under mild conditions, the squared L_2 risk of $\hat{\pi}_{\text{mix}}(x)$ is bounded from above by the smallest risk of all candidates plus a small remainder term, where the remainder term typically vanishes at a rate no slower than the risk itself. We define the L_2 norm of a generic function f with respect to the distribution ν of X by $\|f\|^2 = \int f^2(x)\nu(\mathrm{d}x)$.

LEMMA 1. *Suppose for each* $j \in \mathcal{J}$*, there exists a constant* A_j *with* $0 < A_j < 1/2$*, such that* $A_j \leq \hat{\pi}_{N_n}^{j}(x) \leq 1 - A_j$ *for all* x*. Then*

$$E\big(\|\hat{\pi}_{\text{mix}} - \pi\|^2\big) \leq \inf_{j \in \mathcal{J}} \frac{2}{A_j^2} E\big(\|\hat{\pi}_{N_n}^{j} - \pi\|^2\big) + \frac{2\log(J)}{n - N_n}.$$

The remainder term, $2\log(J)/(n - N_n)$, vanishes at the rate $1/(n - N_n)$ or $1/n$. This is typically faster than the convergence rate of $\inf_{j \in \mathcal{J}} E(\|\hat{\pi}_{N_n}^{j} - \pi\|^2)$ when the candidate models are nonparametric. If one of the candidate models is parametric and it is the model that correctly specifies $\pi(x)$, then $\inf_{j \in \mathcal{J}} E(\|\hat{\pi}_{N_n}^{j} - \pi\|^2)$ converges at rate $1/n$. Therefore, the remainder term is negligible in terms of convergence rate compared to the statistical risks of the candidate models themselves.

Under ignorable missingness, $a(x) = \mathrm{E}(Y \mid X, R = 1)$, and hence one can similarly obtain the mixing procedure for a binary outcome, except that only the complete-case subsample can be used for fitting the candidate models. In the mixing procedure for a continuous outcome, a different criterion must be used to calculate the weights. Recall that the first m cases are assumed to be completely observed. We randomly partition Z^m into two parts $Z^{(3)} = (R_i, R_iY_i, X_i)_{i=1}^{N_m}$ and $Z^{(4)} = (R_i, R_iY_i, X_i)_{i=N_m+1}^{m}$, where $N_m = \max(1, \lfloor cm \rfloor)$ for some $c \in (0, 1)$. For example, one can take $N_m = \lfloor m/2 \rfloor$. After fitting the candidate models on $Z^{(3)}$, we calculate the weights $\hat{\Omega}_k = (m - N_m)^{-1} \sum_{i=N_m+1}^{m} \hat{\Omega}_{k,i}$, where $\hat{\Omega}_{k,N_m+1} = 1/K$ and

$$\hat{\Omega}_{k,i} = \frac{\exp\big[-\lambda \sum_{l=N_m+1}^{i-1} \{Y_l - \hat{a}^k_{N_m}(X_l)\}^2\big]}{\sum_{k' \in \mathcal{K}} \exp\big[-\lambda \sum_{l=N_m+1}^{i-1} \{Y_l - \hat{a}^{k'}_{N_m}(X_l)\}^2\big]} \quad (N_m + 2 \le i \le m), \qquad (10)$$

for $k \in \mathcal{K}$. As with $\hat{\Lambda}_{j,i}$ in the mixing procedure for the final propensity score model, the weights $\hat{\Omega}_{k,i}$ are modified from Yang (2004) to work for our specific situations. The mixing estimator of the outcome regression is then $\hat{a}_{\text{mix}}(x) = \sum_{k \in \mathcal{K}} \hat{\Omega}_k \hat{a}^k_{N_n}(x)$. The calculation of the weights here is based on exponential weighting of the cumulative predictive risks under the mean squared error of the candidate models. The parameter λ in (10) is a properly chosen positive constant that controls the effect of the performance of the candidate models on the weights. We recommend the use of cross-validation for the selection of λ in practice. See more discussions about this in Yang (2004) and Gu & Zou (2019). We summarize the mixing procedure for the final outcome model in Algorithm 2.

Algorithm 2. Mixing procedure for the final outcome regression model.

(1) Randomly partition the complete-case data Z^m into two parts $Z^{(3)} = (R_i, R_iY_i, X_i)_{i=1}^{N_m}$ and $Z^{(4)} = (R_i, R_iY_i, X_i)_{i=N_m+1}^{m}$, where $N_m = \max(1, \lfloor cm \rfloor)$ for some $c \in (0, 1)$.

(2) Fit the candidate outcome regression models to the subsample $Z^{(3)}$ to obtain the estimates $\hat{a}^k_{N_m}(x), k \in \mathcal{K}$.

(3) Calculate weights $\hat{\Omega}_k = (m - N_m)^{-1} \sum_{i=N_m+1}^{m} \hat{\Omega}_{k,i}$, where $\hat{\Omega}_{k,i}$ is defined in (10).

(4) Return the mixing estimator of the outcome regression $\hat{a}_{\text{mix}}(x) = \sum_{k \in \mathcal{K}} \hat{\Omega}_k \hat{a}^k_{N_m}(x)$.

The L_2 risk of $\hat{a}_{\text{mix}}(x)$ is derived in the following lemma. Similar to the mixing estimator of the propensity score, the mixing estimator of the outcome regression is also bounded from above by the smallest risk of all candidates plus a small remainder term. To see this, we define the sub-exponential norm of a generic random variable U by $\sup_{\tau \ge 1} \tau^{-1} \{E(|U|^\tau)\}^{1/\tau}$. If this norm is finite, then U is called a sub-exponential random variable (Vershynin, 2010).

LEMMA 2. *Suppose there exist constants $C_1, C_2 > 0$ such that $\sup_{k \in \mathcal{K}} |\hat{a}^k_{N_m}(x) - a(x)| \le C_1$ for all x and the sub-exponential norm of $Y - a(x)$ given $X = x$ is bounded by C_2 from above for all x. Then*

$$E\big(\|\hat{a}_{\text{mix}} - a\|^2\big) \le \inf_{k \in \mathcal{K}} E\big(\|\hat{a}^k_{N_m} - a\|^2\big) + \frac{\log(K)}{\lambda(m - N_m)},$$

for

$$0 < \lambda \le \max\left[\frac{1}{16eC_1C_2}, \frac{\exp\{C_1(8eC_2)^{-1}\}}{4\mathcal{M}_2\{(4eC_2)^{-1}\} + 16C_1^2\mathcal{M}_0\{(4eC_2)^{-1}\}}\right],$$

where $\mathcal{M}_0(t) = 2\exp(2e^2C_2^2t^2)$, $\mathcal{M}_2(t) = 16\sqrt{2}C_2^2\exp(8e^4C_2^2t^2)$ and $e = \exp(1)$.

We now establish the theoretical properties for our proposed mixing-based doubly robust estimator $\hat{\mu}_{\text{mix}} = \mathbb{P}_n[RY/\hat{\pi}_{\text{mix}}(X) + \{1 - R/\hat{\pi}_{\text{mix}}(X)\}\hat{a}_{\text{mix}}(X)]$. For $j \in \mathcal{J}$, let $\bar{\pi}^j(x)$ be some non-stochastic function to which $\hat{\pi}^j_{N_n}(x)$ converges in the sense that $\|\hat{\pi}^j_{N_n} - \bar{\pi}^j\| = o_p(1)$. Similarly, assume $\|\hat{a}^k_{N_m} - \bar{a}^k\| = o_p(1)$ for some non-stochastic function $\bar{a}^k(x)$, where $k \in \mathcal{K}$. In what follows, we assume $m = \Theta(n)$, i.e., there exist constants $0 < c_0 < c_1 < 1$ such that $c_0 \leq m/n \leq c_1$ for any m, n.

THEOREM 1. *Suppose the conditions in Lemmas* 1 *and* 2 *hold. If (1) $\bar{\pi}^j = \pi$ for some $j \in \mathcal{J}$, or (2) $\bar{a}^k = a$ for some $k \in \mathcal{K}$, then $\hat{\mu}_{\text{mix}}$ is a consistent estimator of μ as $n \to \infty$.*

Theorem 1 indicates that if one of the propensity score or one of the outcome regression candidate models is correctly specified and consistently estimated, the proposed estimator $\hat{\mu}_{\text{mix}}$ is consistent for the true parameter μ. Hence, the multiple robustness property is achieved by our improved doubly robust estimator $\hat{\mu}_{\text{mix}}$ that utilizes the mixing propensity score and outcome regression models.

THEOREM 2. *Suppose the conditions in Lemmas* 1 *and* 2 *hold. We also assume that $1/\bar{\pi}^j(x)$, $\hat{a}^k_{N_m}(x)$ and $\bar{a}^k(x)$ are uniformly bounded for each $j \in \mathcal{J}$ and $k \in \mathcal{K}$. If (1) $\bar{\pi}^j = \pi$ for some $j \in \mathcal{J}$, or (2) $\bar{a}^k = a$ for some $k \in \mathcal{K}$, then*

$$|\hat{\mu}_{\text{mix}} - \mu| = O_p\big(n^{-1/2} + \|\hat{\pi}_{\text{mix}} - \pi\|\|\hat{a}_{\text{mix}} - a\|\big).$$

Theorem 2 characterizes the rate of convergence of the proposed estimator $\hat{\mu}_{\text{mix}}$ when either one of the propensity score or one of the outcome regression models is correctly specified. This rate mainly depends on the order of the product of the convergence rates of $\hat{\pi}_{\text{mix}}(x)$ and $\hat{a}_{\text{mix}}(x)$. When one of the propensity score models and one of the outcome models are correctly specified, we show that $\sqrt{n}(\hat{\mu}_{\text{mix}} - \mu)$ converges to a normal distribution as $n \to \infty$ and $\hat{\mu}_{\text{mix}}$ achieves semiparametric efficiency under some mild conditions.

THEOREM 3. *Suppose the conditions in Theorem* 2 *hold. If for some $j \in \mathcal{J}$ and some $k \in \mathcal{K}$, (1) $\bar{\pi}^j = \pi$, (2) $\bar{a}^k = a$, and (3) $\|\hat{\pi}^j_{N_n} - \pi\|\|\hat{a}^k_{N_m} - a\| = o_p(n^{-1/2})$, then $\sqrt{n}(\hat{\mu}_{\text{mix}} - \mu)$ converges in distribution to $N(0, \sigma^2)$, where*

$$\sigma^2 = E\left\{\frac{\text{var}(Y \mid X)}{\pi(X)}\right\} + \text{var}\{E(Y \mid X)\}.$$

Therefore, $\hat{\mu}_{\text{mix}}$ attains the semiparametric efficiency bound.

Condition (3) in the above theorem is a mild requirement and can be satisfied under many scenarios. For example, when $\pi^j(x)$ is a correctly specified parametric model for $\pi(x)$, it follows that $\|\hat{\pi}^j_{N_n} - \pi\| = O_p(n^{-1/2})$ under fairly common regularity conditions. Then condition (3) is satisfied as long as $\hat{a}^k_{N_m}(x)$ is consistent for $a(x)$, i.e., $\|\hat{a}^k_{N_m} - a\| = o_p(1)$, regardless of whether $a^k(x)$ is parametric or not. The same holds true when $a^k(x)$ is a correctly specified parametric model for $a(x)$ and we require minimal assumptions on $\pi^j(x)$. When $\pi^j(x)$ and $a^k(x)$ are nonparametric, as long as both $\hat{\pi}^j_{N_n}(x)$ and $\hat{a}^k_{N_m}(x)$ converge faster than $n^{-1/4}$, condition (3) is satisfied. Theorem 3 also implies that $\hat{\mu}_{\text{mix}}$ is a regular estimator for μ, which is formally stated in the following corollary.

COROLLARY 1. *Under the conditions of Theorem* 3, *$\hat{\mu}_{\text{mix}}$ is a regular asymptotic linear estimator of μ.*

It is worth pointing out that our mixing-based doubly robust estimator $\hat{\mu}_{\text{mix}}$ and its attached theoretical properties are not confined to parametric models. In addition, one can readily utilize existing variable selection and regularization techniques, such as the lasso (Tibshirani, 1996), adaptive lasso (Zou, 2006), elastic net (Zou & Hastie, 2005) and so on, to handle a large number of predictors, or equivalently, candidate models with different subsets of predictors. These methods guarantee consistent estimation and give reasonable rates of convergence for the propensity score and outcome regression models under high dimensionality when certain sparsity structures are assumed for the true models. As a consequence, $\hat{\mu}_{\text{mix}}$ remains multiply robust for high-dimensional data. For example, suppose that the estimators $\hat{\pi}^j_{N_n}$ for some $j \in \mathcal{J}$ and $\hat{a}^k_{N_m}$ for some $k \in \mathcal{K}$ are obtained from lasso. Under some regularity conditions, we have $\|\hat{\pi}^j_{N_n} - \pi\| = O_p\{(s_\pi \log d/n)^{1/2}\}$ (Van de Geer et al., 2008) and $\|\hat{a}^k_{N_m} - a\| = O_p\{(s_a \log d/m)^{1/2}\}$ (Bickel et al., 2009), where s_π and s_a denote the sparsity levels of the true propensity score and outcome regression models, respectively. Hence, as long as $s_\pi s_a (\log d)^2/n = o(1)$, the results in Theorem 3 still hold. Although existing multiply robust methods can be also applied to high-dimensional or nonparametric scenarios by adding a preceding regularization procedure or including nonparametric models in the constraints, the theoretical results for such cases may need to be further developed. Nevertheless, for fair comparison in the simulation studies of Section 5, we also conducted variable selection procedures before applying existing multiply robust methods.

5. Simulation Studies

In this section, we present simulation studies to investigate the finite sample performance of our proposed estimator $\hat{\mu}_{\text{mix}}$, and compare it with existing multiply robust estimators $\hat{\mu}_{\text{mr–ipw}}$, $\hat{\mu}_{\text{mr–cal}}$, and $\hat{\mu}_{\text{mr–reg}}$.

The simulation setting follows that in Kang & Schafer (2007). Sample sizes for each simulation scenario are 200 and 1000. For each simulation, the data are generated with $X = \{X^{(1)}, \ldots, X^{(4)}\} \sim N_4(0, \mathrm{I}_4)$, $Y \mid X = x \sim N\{a(x), 1\}$, and $R \mid X = x \sim \text{Ber}\{\pi(x)\}$, where I_4 is a 4×4 identity matrix, $\pi(x) = [1 + \exp\{x^{(1)} - 0{\cdot}5x^{(2)} + 0{\cdot}25x^{(3)} + 0{\cdot}1x^{(4)}\}]^{-1}$ and $a(x) = 210 + 27{\cdot}4x^{(1)} + 13{\cdot}7x^{(2)} + 13{\cdot}7x^{(3)} + 13{\cdot}7x^{(4)}$. The missing data proportion is about 50%. As with Kang & Schafer (2007), we calculate $V^{(1)} = \exp\{X^{(1)}/2\}$, $V^{(2)} = X^{(2)}/[1 + \exp\{X^{(1)}\}] + 10$, $V^{(3)} = \{X^{(1)}X^{(3)}/25 + 0{\cdot}6\}^3$ and $V^{(4)} = \{X^{(2)} + X^{(4)} + 20\}^2$.

We consider four different scenarios: (i) one of the candidate propensity score models and one of the candidate outcome regression models are correctly specified, (ii) only one of the candidate outcome regression models is correctly specified, (iii) only one of the candidate propensity score models is correctly specified, and (iv) none of the candidate models for propensity score or outcome regression is correctly specified. The correctly specified propensity score and outcome regression models are fitted by using X as covariates, whereas we treat $V = \{V^{(1)}, \ldots, V^{(4)}\}$ to be the covariates instead of X in the misspecified models.

We carry out the simulation for scenario (i) in the following steps:

Step 1. We simulate a random sample of $n = 200$ or $n = 1000$ observations according to the Kang & Schafer (2007) setting.

Step 2. We choose the identity link function for the outcome model, and the commonly used logit, probit, and complementary log-log link functions for the candidate propensity score models. We apply the elastic net to select variables separately within each given link function where the predictors are composed of all first- and second-order interactions of the covariates X.

Step 3. We then conduct the model mixing procedures using the candidate models obtained under different link functions. We calculate $\hat{\mu}_{\text{mix}}$ based on the combined propensity score and outcome regression models.

Step 4. For a fair comparison, we also use the selected models by elastic net from *Step* 2 as candidate propensity score and outcome regression models for $\hat{\mu}_{\text{mr–ipw}}$, $\hat{\mu}_{\text{mr–cal}}$ and $\hat{\mu}_{\text{mr–reg}}$. Since $\hat{\mu}_{\text{mr–cal}}$ involves only one propensity score model, we choose the one picked by the elastic net under the logit link. Based on the selected candidate models, $\hat{\mu}_{\text{mr–ipw}}$, $\hat{\mu}_{\text{mr–cal}}$, and $\hat{\mu}_{\text{mr–reg}}$ are then calculated.

Step 5. We repeat *Step* 1–4 for 200 times.

The simulation studies for other scenarios are conducted similarly except that the covariates V are used in the misspecified models in *Step* 2. The results for all scenarios with sample sizes $n = 200$ and $n = 1000$ are reported in Figure 1. In almost all the scenarios, our porposed mixing-based doubly robust estimator $\hat{\mu}_{\text{mix}}$ has comparable bias and variance with the one that achieves the best performance among those existing multiply robust estimators $\hat{\mu}_{\text{mr–ipw}}$, $\hat{\mu}_{\text{mr–cal}}$, and $\hat{\mu}_{\text{mr–reg}}$. It is also observed that $\hat{\mu}_{\text{mr–reg}}$ performs slightly worse in some scenarios, compared with the other estimators. As sample size increases, all estimators have smaller biases in scenarios (i)—(iii), but no such trends are observed in scenario (iv). This is expected because only when at least one of the candidate models for the propensity score or outcome regression is correctly specified, are these estimators consistent.

The existing multiply robust estimators $\hat{\mu}_{\text{mr–ipw}}$, $\hat{\mu}_{\text{mr–cal}}$, and $\hat{\mu}_{\text{mr–reg}}$ all achieve satisfactory empirical performances through adding a preceding regularization procedure. However, their theoretical underpinnings need to be further developed. More specifically, the theoretical results of the existing multiply robust estimators such as their asymptotic normality are established in the case where both propensity score and outcome regression models are parametric and the parameters in these models are estimated via the maximum likelihood method. In contrast, we have developed theoretical results of our proposed mixing-based estimator without relying on parametric modeling assumptions. Any method that provides consistent estimates and guarantees fast enough rates of convergence for the candidate propensity score and outcome regression models can be incorporated into our estimator, as outlined in Theorems 1–3.

6. Discussion

The current literature on multiple robustness, e.g., Han & Wang (2013); Han (2014a,b); Chan (2013); Chan & Yam (2014), raises an important point that multiple opportunities should be highly appreciated to achieve the correct specification of both the propensity score and outcome regression models in missing data problems. Although the doubly robust estimator has two chances to achieve consistent estimation, there is no guarantee that either of the two models, i.e., the propensity score and outcome regression models, is correctly specified. Additionally, as pointed out by Kang & Schafer (2007), the doubly robust estimator could suffer from severe bias if both models are misspecified. Hence, the doubly robust estimator should not be regarded as a panacea, but rather an extra opportunity. In that sense, despite the robustness property of the estimating procedures, it is encouraged that more information about the underlying mechanism is collected and more careful scrutiny of modeling is incorporated.

In this paper, we mainly focus on model combination techniques to improve model specification of the final propensity score and outcome regression in the doubly robust estimator. However, we point out that model selection techniques can be also applied to the selection of candidate

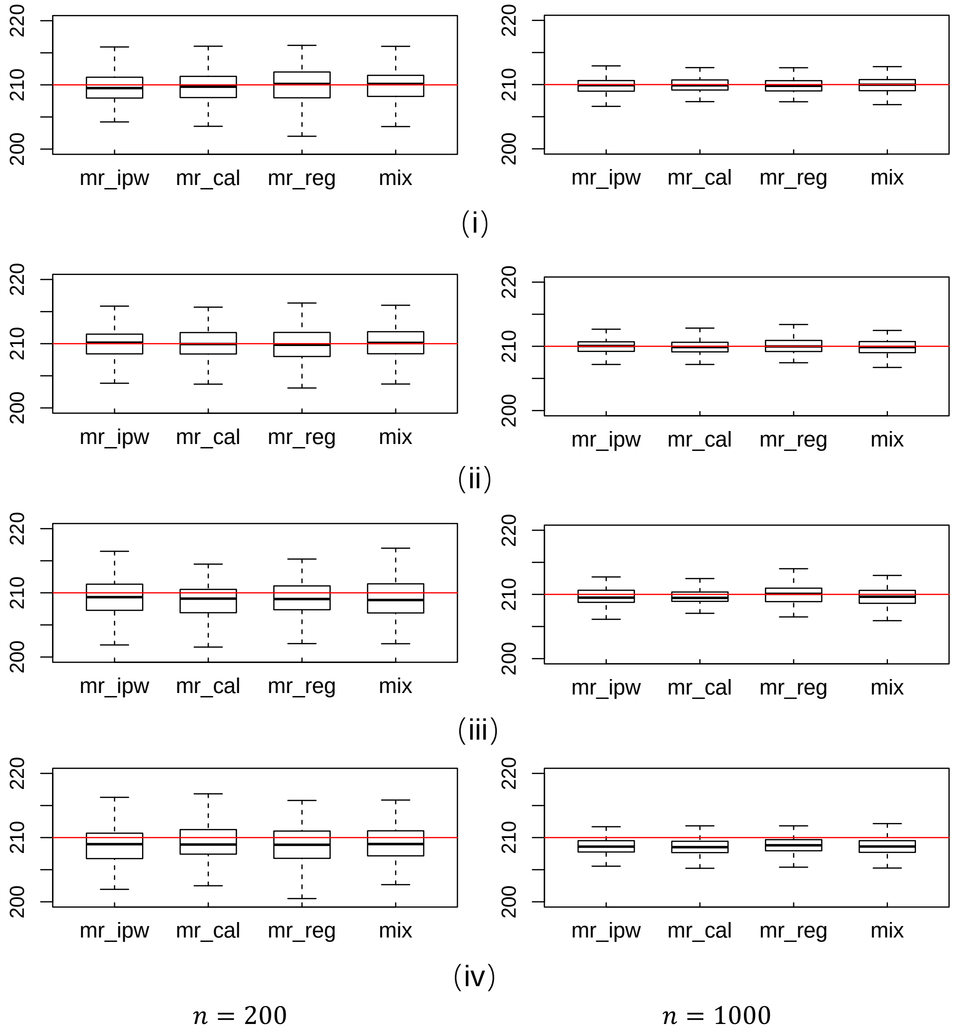

Fig. 1: Comparison between the proposed mixing-based doubly robust estimator $\hat{\mu}_{\text{mix}}$ and existing multiply robust estimators $\hat{\mu}_{\text{mr–ipw}}$, $\hat{\mu}_{\text{mr–cal}}$, and $\hat{\mu}_{\text{mr–reg}}$ for Kang & Schafer (2007) setting with sample sizes $n = 200$ and $n = 1000$. Four scenarios are considered: (i) one outcome regression candidate and one propensity score candidate are correct; (ii) only one outcome regression candidate is correct; (iii) only one propensity score candidate is correct; (iv) no candidates are correct. The x-axis shows different estimators and the y-axis shows the values of these estimators. The red line represents the true value of each scenario.

models for improved modeling accuracy. Our preliminary analyses using model selection techniques found that when there is one candidate model that has a dominating performance over the others, the model selection approach could outperform the model combination approach, and vice versa. We leave further comparisons for future exploration.

In this paper, we develop our methods under the ignorable missingness assumption. Of course, one should be more cautious about model misspecification with non-ignorable missingness data. For interested readers, we refer to Miao et al. (2015) and Miao & Tchetgen Tchetgen (2016) for doubly robust estimators of the mean of the response variable when the missingness is non-ignorable, and Han (2018) for discussion of a multiply robust method for non-ignorable missing data. We leave the generalization of our methods to those settings as a future research topic.

We end by pointing out that the multiply robust estimators discussed in this paper are different from other multiply robust estimators in the literature, though they bear the same name. For example, the multiply robust estimators proposed by Tchetgen Tchetgen & Shpitser (2012) model three rather than two parts of the likelihood function and they are consistent for their parameters of interest when two out of the three models are correctly specified. Those estimators cannot be interpreted as special cases of doubly robust estimators under our framework.

Acknowledgments

The authors thank Prof. Eric Tchetgen Tchetgen Prof. Yuhong Yang and Prof. James Robins for the very helpful discussions. We show special thanks to the editor, associate editor and reviewers for their helpful and insightful comments. Lan Liu's research is supported by NSF DMS 1916013.

Supplementary Material

Supplementary material includes a derivation of a constraint on δ from (7), and proofs of all lemmas, theorems and the corollary.

References

Bang, H. & Robins, J. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 962–973.

Bickel, P. J., Ritov, Y., Tsybakov, A. B. et al. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics* **37**, 1705–1732.

Chan, K. (2013). A simple multiply robust estimator for missing response problem. *Stat* **2**, 143–149.

Chan, K. & Yam, S. (2014). Oracle, multiple robust and multipurpose calibration in a missing response problem. *Statistical Science* **29**, 380–396.

Chen, S. & Haziza, D. (2017). Multiply robust imputation procedures for the treatment of item nonresponse in surveys. *Biometrika* **104**, 439–453.

Duan, X. Yin, G. (2017). Ensemble approaches to estimating thepopulation mean with missing response. *Scandinavian Journal of Statistics* **44**, 899–917.

Gu, Y. & Zou, H. (2019). Aggregated expectile regression by exponential weighting. *Statistica Sinica* **29**, 671–692.

Han, P. (2014a). A further study of the multiply robust estimator in missing data analysis. *Journal of Statistical Planning and Inference* **148**, 101–110.

Han, P. (2014b). Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association* **109**, 1159–1173.

Han, P. (2016). Combining inverse probability weighting and multiple imputation to improve robustness of estimation. *Scandinavian Journal of Statistics* **43**, 246–260.

Han, P. (2018). Calibration and multiple robustness when data are missing not at random. *Statistica Sinica* **28**, 1725–1740.

Han, P. & Wang, L. (2013). Estimation with missing data: beyond double robustness. *Biometrika* **100**, 417–430.

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685.

Kang, J. & Schafer, J. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* **22**, 523–539.

Little, R. & Rubin, D. (2002). *Statistical Analysis with Missing Data*. New York: Wiley, 2nd ed.

Miao, W., Tchetgen, E. T. & Geng, Z. (2015). Identification and doubly robust estimation of data missing not at random with a shadow variable. *arXiv preprint arXiv:1509.02556* .

Miao, W. & Tchetgen Tchetgen, E. (2016). On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika* **103**, 475–482.

QIN, J., SHAO, J. & ZHANG, B. (2008). Efficient and doubly robust imputation for covariate-dependent missing responses. *Journal of the American Statistical Association* **103**, 797–810.

ROBINS, J. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association Section on Bayesian Statistical Science*, vol. 1999.

ROBINS, J. (2002). Commentary on "using inverse weighting and predictive inference to estimate the effects of time-varying treatments on the discrete-time hazard". *Statistics in Medicine* **21**, 1663–1680.

ROBINS, J., ROTNITZKY, A. & ZHAO, L. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89**, 846–866.

ROBINS, J., ROTNITZKY, A. & ZHAO, L. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* **90**, 106–121.

ROSENBAUM, P. & RUBIN, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

ROTNITZKY, A., LEI, Q., SUED, M. & ROBINS, J. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika* **99**, 439–456.

ROTNITZKY, A. & ROBINS, J. (1995). Semiparametric regression estimation in the presence of dependent censoring. *Biometrika* **82**, 805–20.

SCHARFSTEIN, D., ROTNITZKY, A. & ROBINS, J. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* **94**, 1096–1120.

TAN, Z. (2007). Comment: Understanding OR, PS and DR. *Statistical Science* **22**, 560–568.

TAN, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika* **97**, 661–682.

TCHETGEN TCHETGEN, E. & SHPITSER, I. (2012). Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of Statistics* **40**, 1816–1845.

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

TSIATIS, A. (2006). *Semiparametric Theory and Missing Data*. Springer.

TSIATIS, A., DAVIDIAN, M. & CAO, W. (2011). Improved doubly robust estimation when data are monotonely coarsened, with application to longitudinal studies with dropout. *Biometrics* **67**, 536–545.

VAN DE GEER, S. A. et al. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics* **36**, 614–645.

VAN DER LAAN, M. & GRUBER, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *The International Journal of Biostatistics* **6**, 1–71.

VAN DER LAAN, M. & ROBINS, J. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer Verlag.

VERSHYNIN, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027v7* .

YANG, Y. (2000). Adaptive estimation in pattern recognition by combining different procedures. *Statistica Sinica* **10**, 1069–1089.

YANG, Y. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association* **96**, 574–588.

YANG, Y. (2004). Combining forecasting procedures: some theoretical results. *Econometric Theory* **20**, 176–222.

ZOU, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

ZOU, H. & HASTIE, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 301–320.