Model averaged double robust estimation

Matthew Cefalu

Department of Biostatistics, Harvard School of Public Health

Francesca Dominici

Department of Biostatistics, Harvard School of Public Health

Giovanni Parmigiani

Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute

Department of Biostatistics, Harvard School of Public Health

Abstract

We propose a new class of estimators for the average causal effect, the model averaged double robust (MA-DR) estimators, that account for model uncertainty in both the propensity score and outcome model through the use of model averaging. The MA-DR estimator is defined as a weighted average of double robust estimators, where each double robust estimator corresponds to a specific choice for the outcome model and the propensity score, respectively. The MA-DR estimators extend the desirable double robustness property by achieving consistency under the much weaker assumption that either the true propensity score model or the true outcome model be within a specified, possibly large, class of models. We provide asymptotic results and conduct a large scale simulation study that indicates the MA-DR estimator has better finite sample behavior than the usual double robust estimator. We show that the MA-DR that a priori links the propensity score and the outcome model can have 90% less variance than a double robust estimator constructed via model selection for the propensity score and the outcome model separately. Importantly, our simulation suggests that our MA-DR estimator dramatically reduces mean squared error by the largest

percentage in the realistic situation where the outcome model is misspecified.

1 Introduction

Methods for causal inference are predicated on knowledge of the covariates necessary to satisfy the no unmeasured confounding assumption, but the exact set of covariates needed to control confounding is rarely known. Practical tools that acknowledge uncertainty in confounder selection and are robust to model misspecification are imperative for correct estimation of the average causal effect [29, 30].

Although the literature on causal inference is vast, existing methods do not account for the uncertainty in selection of confounders, C, or in the form of the model for the treatment, X [29]. For example, a wealth of methods that rely on specification of a propensity score model, $\mathrm{P}(X|C)$, for treatment assignment (e.g propensity score matching or inverse probability weighting estimators; see [13] for a review) typically assume that both the covariates to include and the functional form of the propensity score model are known a priori.

In addition to specification of $\mathrm{P}(X|C)$, a broad class of methods for causal inference rely on the additional specification of a model for potential outcomes $\mathrm{P}(Y(x)|C)$ [19, 20, 5], where $Y(1)$ and $Y(0)$ are the potential outcomes under each treatment. Included in this class are methods for inverse probability treatment weighted estimators that are often promoted for properties such as consistency and double robustness [22, 1, 26]. Within the class of double robust estimators, covariate and model selection is specified a priori for both the propensity score and the outcome models separately, presenting further challenges to acknowledging uncertainty with respect to the selection of the confounders and providing robustness to model misspecification. There are few tools or guidelines for model selection in double robust estimators, and many researchers take an ad-hoc approach.

One possible tool to formally account for model uncertainty in the adjustment for confounding is Bayesian model averaging [16, 6]. These methods are based on treating the indicators of whether each confounder is included in the model as a nuisance parameter, and it has been suggested that an effect estimate can be formed by weighting the model-specific estimates [9], where the weights are determined by the models' posterior probabilities.

In the context of a regression model, where the goal is the estimation of the effect of X on Y adjusting for measured confounders, the use of Bayesian model averaging with non-informative priors on the models has received some criticism [4, 29, 30]: variable selection based on the outcome model only prioritizes the Cs strongly associated with Y, and variable selection based on the propensity score model only prioritize the Cs that are strongly associated with X. Both these approaches can result in inefficient and biased inferences because they fail to identify the full set of necessary confounders [2, 23].

Wang et al. [30] propose a solution to this important problem for a continuous exposure X and with confounding adjustment made by introducing Cs into the regression model as covariates. Two regression equations are specified along with two vectors of inclusion indicators: (1) a linear regression model for Y given X and C (the outcome model); and (2) a linear regression model for X given C (the exposure model). They assume a priori that if a covariate C is highly predictive of the exposure X, then the same covariate C will have a large probability of being included into the outcome model. It is shown that the model averaged estimator of the effect of X on Y, obtained with this form of prior dependence between the outcome model and the exposure model has lower mean squared error than the model averaged estimator that assumes a priori that the two vectors of inclusions indicators are independent.

Accounting for model uncertainty in the context of causal inference is a widely unexplored topic. In this paper, we propose a new class of methods for estimating the average causal effect, which we call *the model averaged double robust estimators*, that formally account for model uncertainty through the use of model averaging while maintaining the desirable properties of consistency and double robustness. These methods provide valid estimation of the average causal effect that: 1) are robust to the misspecification of the model for the treatment assignment; 2) are robust to the misspecification of the model for outcome; and 3) account for the uncertainty in the selection of the confounders in both the propensity score model and in the outcome model. Importantly, we show that a model averaged double robust estimator that assumes dependence between the propensity score and the outcome model a priori and separates estimation of the model weights into two stages can reduce the mean squared error of the double robust estimator by more than 90% when compared to traditional model selection procedures.

2 Methods

2.1 A double robust estimator

Consider continuous potential outcomes $(Y(0), Y(1))$, binary treatment X, and a p-dimensional set of potential confounders C. Assume there is no unmeasured confounding [18] (also referred to as strong ignorable treatment assignment[19]), so that $(Y(0), Y(1)) \perp\!\!\!\perp X|C$. Let (Y_h, X_h, C_h) be independent observations for $h = 1, ..., n$. We are interested in estimating the average causal effect:

$$\Delta = E\left[Y(1) - Y(0)\right] = E\left\{E(Y|X = 1, C) - E(Y|X = 0, C)\right\} \tag{1}$$

Given a model for the propensity score, $\mathrm{P}(X = 1|C) = e(C)$, and a model for the outcome under each treatment, $E(Y|X = 1, C) = m_1(C)$ and $E(Y|X = 0, C) = m_0(C)$, we define the well known parametric ($\widehat{\Delta}^p$), inverse probability weighted ($\widehat{\Delta}^{IPW}$), and double robust ($\widehat{\Delta}^{DR}$) estimators as:

$$\begin{aligned}
\widehat{\Delta}^p &= \frac{1}{n}\sum_{h=1}^{n}\{\widehat{m}_{1h} - \widehat{m}_{0h}\} \\
\widehat{\Delta}^{IPW} &= \frac{1}{n}\sum_{h=1}^{n}\left\{\frac{Y_h X_h}{\widehat{e}_h} - \frac{Y_h(1 - X_h)}{1 - \widehat{e}_h}\right\} = \frac{1}{n}\sum_{i=1}^{n}\frac{X_h - \widehat{e}_h}{\widehat{e}_h(1 - \widehat{e}_h)}Y_h \\
\widehat{\Delta}^{DR} &= \frac{1}{n}\sum_{h=1}^{n}\left\{\frac{Y_h X_h - (X_h - \widehat{e}_h)\widehat{m}_{1h}}{\widehat{e}_h} - \frac{Y_h(1 - X_h) + (X_h - \widehat{e}_h)\widehat{m}_{0h}}{1 - \widehat{e}_h}\right\}
\end{aligned} \tag{2}$$

where $\widehat{m}_{1h}$, $\widehat{m}_{0h}$, and $\widehat{e}_h$ are the estimated outcomes and propensity score for individual h under model $m_1(C)$, $m_2(C)$, and $e(C)$, respectively. To simplify the model averaging arguments in the next section, note that $\widehat{\Delta}^{DR}$ can be decomposed into $\widehat{\Delta}^{IPW}$, $\widehat{\Delta}^p$, and a third estimator $\widehat{\Delta}^{PIPW}$.

$$\widehat{\Delta}^{DR} = \frac{1}{n}\sum_{h=1}^{n}\left[\widehat{m}_{1h} - \widehat{m}_{0h} + \frac{X_h - \widehat{e}_h}{\widehat{e}_h(1-\widehat{e}_h)}Y_h - \frac{X_h - \widehat{e}_h}{\widehat{e}_h(1-\widehat{e}_h)}\widehat{m}_{X_h}\right]$$
$$= \widehat{\Delta}^{p} + \widehat{\Delta}^{IPW} - \widehat{\Delta}^{PIPW}$$

where $\widehat{\Delta}^{PIPW}$ is a parametric inverse probability weighted estimator and $\widehat{m}_{X_h} = \widehat{m}_{1h}$ if $X_h = 1$ and $\widehat{m}_{X_h} = \widehat{m}_{0h}$ otherwise. Observe that $\widehat{\Delta}^{p}$ only depends on the outcome model, $\widehat{\Delta}^{IPW}$ only depends on the propensity score model, and $\widehat{\Delta}^{PIPW}$ depends on both.

The model for the propensity score and the outcome under each treatment can be selected in any number of ways. A researcher may rely on expert knowledge to decide both the functional form and the covariates to include in each model, or may rely on a model selection procedure that chooses the best model from a set of candidate models. For the remainder of this paper, we will refer to $\widehat{\Delta}^{MS}_{DR}$ as the "model selected double robust estimator" in which both the propensity score and the outcome model have been selected independently using BIC [24] as a model selection procedure.

2.2 Model averaged double robust estimator

Let $\mathcal{M}^{ps} = \left\{\mathcal{M}^{ps}_1, \mathcal{M}^{ps}_2, ..., \mathcal{M}^{ps}_{M_{ps}}\right\}$, $\mathcal{M}^0 = \left\{\mathcal{M}^0_1, \mathcal{M}^0_2, ..., \mathcal{M}^0_{M_0}\right\}$, and $\mathcal{M}^1 = \left\{\mathcal{M}^1_1, \mathcal{M}^1_2, ..., \mathcal{M}^1_{M_1}\right\}$ be finite collections of models for $\mathrm{P}(X = 1|C)$, $E(Y|X = 0, C)$, and $E(Y|X = 1, C)$, respectively. For example, the collection of models for the propensity score $\mathcal{M}^{ps}$ could consist of logistic regression models with all subsets of C as linear predictors. Let $\mathcal{M}^{om} = \mathcal{M}^1 \times \mathcal{M}^0$ denote all combinations of models in $\mathcal{M}^1$ and $\mathcal{M}^0$. Further, define $\widehat{\Delta}^{DR}_{ij}$ as the double robust estimator corresponding to the models $\mathcal{M}^{ps}_i$ and $\mathcal{M}^{om}_j$. We define the model average double robust estimator as:

$$\widehat{\Delta}^{MA}_{DR} = \sum_{ij} p_{ij}\widehat{\Delta}^{DR}_{ij} \tag{3}$$

where $p_{ij} = \mathrm{P}(\mathcal{M}_i^{ps}, \mathcal{M}_j^{om}|\mathcal{D})$ is the joint posterior probability of models $\mathcal{M}_i^{ps}$ and $\mathcal{M}_j^{om}$. We expand the estimator based on the decomposition in the previous section. Let $\widehat{\Delta}_i^{IPW}$, $\widehat{\Delta}_j^{p}$, and $\widehat{\Delta}_{ij}^{PIPW}$ be the inverse probability weighted estimator, the parametric estimator, and the parametric inverse probability weighted estimator for the corresponding models $\mathcal{M}_i^{ps}$ and $\mathcal{M}_j^{om}$. Then,

$$\widehat{\Delta}_{DR}^{MA} = \sum_i p_{i\bullet}\widehat{\Delta}_i^{IPW} + \sum_j p_{\bullet j}\widehat{\Delta}_j^{p} - \sum_{ij} p_{ij}\widehat{\Delta}_{ij}^{PIPW} \tag{4}$$

where $p_{i\bullet} = \sum_j p_{ij}$ and $p_{\bullet j} = \sum_i p_{ij}$. Note that Equation 4 has a model averaged term for the inverse probability weighted, parametric, and parametric inverse probability weighted estimators. The variance of the model averaged double robust estimator can be estimated using standard resampling methods (e.g. bootstrap; see [7]).

2.3 Prior and posterior model probabilities

To complete the specification of the model averaged double robust estimator, a prior distribution on the model class must be assumed. We will return to choices of priors momentarily, but first let A_i be the prior odds of $\mathcal{M}_i$ versus some other model $\mathcal{M}_1$ that both belong to some model class $\mathcal{M}$. Then, the posterior probability of model $\mathcal{M}_i$ is given by:

$$\mathrm{P}(\mathcal{M}_i|\mathcal{D}) = \frac{A_i B_{i1}}{\sum_{j:\mathcal{M}_j \in \mathcal{M}} A_j B_{j1}} \tag{5}$$

where B_{i1} is the Bayes factor for model $\mathcal{M}_i$ against another model $\mathcal{M}_1$. Bayes factors and their estimates are well studied, and there is extensive literature on the subject [25, 14, 11, 12]. Among the properties of Bayes factors is consistency for model selection, which is a necessary component for consistency of $\widehat{\Delta}_{DR}^{MA}$ as seen in Section 2.4. A well known and popular estimate of Bayes factors is based on BIC [24] and allows us to estimate posterior model probabilities with ease.

Returning to the specification of a prior distribution for the model space, the simplest choice is to assume

that all models are equally likely a priori. This corresponds to assuming that the prior odds of each model is 1, and that the form of the propensity score and the outcome model are independent. Therefore, the resulting model averaged double robust estimator is given by:

$$\widehat{\Delta}_{DR}^{MA-i} = \sum_i p_i \widehat{\Delta}_i^{IPW} + \sum_j q_j \widehat{\Delta}_j^{p} - \sum_{ij} p_i q_j \widehat{\Delta}_{ij}^{PIPW} \tag{6}$$

where $p_i = \mathrm{P}(\mathcal{M}_i^{ps}|\mathcal{D})$ and $q_j = \mathrm{P}(\mathcal{M}_j^{om}|\mathcal{D})$. Notice that because of the prior independence, the posterior probabilities of the propensity score and outcome models are also independent. Therefore, $\mathrm{P}(\mathcal{M}_i^{ps}|\mathcal{D})$ and $\mathrm{P}(\mathcal{M}_j^{om}|\mathcal{D})$ can be computed separately using readily available software, and the model averaged double robust estimator is straightforward to calculate as given in Equation 6.

However, efficiency can be gained through the use of a prior on the model space that identifies confounders (Cs that are associated with both treatment and outcome) for use in the propensity score model. Under the prior independence assumption, the posterior model probability of the propensity score only prioritizes models in which the Cs are strongly associated with X and ignores all relationships with Y. The current literature in causal inference suggests that inclusion of covariates that are only related to the exposure into a propensity score model adds to the variance of the resulting double robust estimator [21, 2].

In this light, we propose an alternative formulation of the prior distribution on the model space that links the propensity score model to the outcome model through prior model dependence. First, let the prior odds of propensity score model $\mathcal{M}_i^{ps}$ to $\mathcal{M}_1^{ps}$ conditional on the outcome model $\mathcal{M}_j^{om}$ be such that:

$$\frac{\mathrm{P}(\mathcal{M}_i^{ps}|\mathcal{M}_j^{om})}{\mathrm{P}(\mathcal{M}_1^{ps}|\mathcal{M}_j^{om})} = \begin{cases} 1, & \text{if } \mathcal{M}_i^{ps} \subset \mathcal{M}_j^{om} \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where $\mathcal{M}_i^{ps} \subset \mathcal{M}_j^{om}$ indicates that the systematic component of $\mathcal{M}_i^{ps}$ is a subset of the systematic component of $\mathcal{M}_j^{om}$. We refer to the 'systematic component' of a model as the specification of its linear predictor, so that there is no issue with the exposure being binary while the outcome is continuous, and we

assume $\mathcal{M}^{ps}$ and $\mathcal{M}^{om}$ contain models with the same nested systematic components. We then choose a reference propensity score model $\mathcal{M}_1^{ps}$ such that $\mathcal{M}_1^{ps} \subset \mathcal{M}_j^{om}$ for all j. The reference model is either a null model or a model that includes confounders that are strictly required regardless of the inclusion of the other confounders. Further, assume that the prior distribution on the outcome model space is uniform. We will denote the model averaged double robust estimator using this prior as $\widehat{\Delta}_{DR}^{MA-d}$. The posterior model probabilities used in the construction of $\widehat{\Delta}_{DR}^{MA-d}$ can be estimated by first finding the prior odds of each model combination under the prior model dependence given by Equation 7 and then using Equation 5 to find the posterior model probabilities.

The prior model dependency given by Equation 7 forces the set of potential confounders included in the propensity score model to be a subset of those that are included in the outcome model. In other words, the prior probability of excluding a variable from the propensity score model given that it is excluded from the outcome model is one, and the prior probability of including a variable in the outcome model given that it is in the propensity score model is one. This type of restriction is supported by the current literature on propensity scores [21, 2], and is related to the priors for continuous exposure introduced by Wang et al. [30].

Our motivation for this prior distribution on the model space was to identify the set of potential confounders that should be included into the propensity score model based on the fact that they are associated with both treatment and outcome, instead of being associated with treatment only. In other words, the prior dependency given in Equation 7 gives zero weight a priori to propensity score models having a systematic component that is not included in the outcome model. However, the estimation of the joint posterior model probability $\mathrm{P}(\mathcal{M}_i^{ps}, \mathcal{M}_j^{om}|\mathcal{D})$ based on this prior has the additional undesirable property that it allows feedback from the propensity score into the outcome model. This feedback is such that the posterior model probabilities will favor outcome models that include any of the potential confounders that are associated with either X or Y, and some efficiency is lost by including potential confounders that are only associated with X into the outcome model.

We will cut the feedback from propensity score into the outcome model with the goal of improving the efficiency of the model averaged double robust estimator through the use of a two-stage approach for calculating the model weights. The two-stage approach for calculating the model weights and the resulting

model averaged double robust estimator proceeds as follows:

1. Estimate the marginal posterior of the outcome model, $q_j = \mathrm{P}(\mathcal{M}_j^{om}|\mathcal{D})$, assuming a uniform prior on the outcome model space and ignoring the specification of the propensity score model

2. Estimate the posterior of the propensity score model conditional on the outcome model, $\mathrm{P}(\mathcal{M}_i^{ps}|\mathcal{M}_j^{om}, \mathcal{D})$, using the prior model dependence given by Equation 7

3. Multiply the estimates from Stage 1 and 2 to find the joint model weight of the propensity score and outcome model, $p_{ij}^* = q_j \mathrm{P}(\mathcal{M}_i^{ps}|\mathcal{M}_j^{om}, \mathcal{D})$

4. Estimate the resulting model averaged double robust estimator $\widehat{\Delta}_{DR}^{MA-dII}$ using p_{ij}^* as the model weights. The notation II in the superscript of the estimator corresponds to the fact that we are calculating the model weights in two stages.

The model weights under this two-stage approach can be easily calculated because they are a transformation of the model probabilities assuming a uniform prior on the model space. First, the outcome model probabilities $q_j = \mathrm{P}(\mathcal{M}_j^{om}|\mathcal{D})$ in Step 1 are simply the model probabilities assuming a uniform prior on the model space. For the estimation of $\mathrm{P}(\mathcal{M}_i^{ps}|\mathcal{M}_j^{om}, \mathcal{D})$ in Step 2, note that conditional on each outcome model, the prior odds for the propensity score models are uniform for models that meet the restriction $\mathcal{M}_i^{ps} \subset \mathcal{M}_j^{om}$. Simple implementation transforming either Bayes' factors, BIC, or model probabilities under a uniform prior on the model space to model weights using the two-stage approach is available.

The difference between the two-stage model weights given by p_{ij}^* and the proper posterior model probabilities used in the estimator $\widehat{\Delta}_{DR}^{MA-d}$ is that the two-stage approach forces the marginal outcome model weights to be equal to the marginal posterior outcome model probabilities under a uniform prior on the model space. More specifically, the estimation of q_j in Stage 1 of the two-stage method does not correspond to the true marginal posterior $P(\mathcal{M}_j^{om}|\mathcal{D})$, while the estimation of $\mathrm{P}(\mathcal{M}_i^{ps}|\mathcal{M}_j^{om}, \mathcal{D})$ in Stage 2 does correspond to the true conditional posterior.

2.4 Asymptotic properties of $\widehat{\Delta}_{DR}^{MA}$

All of the results of this section require consistency of the posterior model probabilities. As stated in Section 2.3, the Bayes factor and its BIC approximations are consistent [11].

We will show that if either the true propensity score model is contained in $\mathcal{M}^{ps}$ <u>or</u> the true outcome models are contained in $\mathcal{M}^{om}$ <u>and</u> the posterior model probabilities are consistent for selecting the true models, then we have that $\widehat{\Delta}_{DR}^{MA}$ is consistent for the average casual effect defined in Equation 1. This result implies that we have added another layer of robustness to the double robust estimator, as we only need the true models to be in the collection of models. All $\rightarrow$ indicate limits as $n \rightarrow \infty$, and $\xrightarrow{p}$ indicates convergence in probability while $\xrightarrow{\mathcal{L}}$ indicates convergence in distribution.

Lemma 1. *Assume there is no unmeasured confounding, independent observations, and that $\mathcal{M}^{om}$ contains the true model, $\mathcal{M}_1^{om}$, for both $E(Y|X=1,C)$ and $E(Y|X=0,C)$. If $p_{\bullet 1} = \sum_i \mathrm{P}(\mathcal{M}_i^{ps}, \mathcal{M}_1^{om}|\mathcal{D}) \xrightarrow{p} 1$, then*

$$\widehat{\Delta}_{DR}^{MA} \xrightarrow{p} \Delta$$

Lemma 2. *Assume there is no unmeasured confounding, independent observations, and that $\mathcal{M}^{ps}$ contains the true model, $\mathcal{M}_1^{ps}$, for $\mathrm{P}(X=1|C)$. If $p_{1\bullet} = \sum_j \mathrm{P}(\mathcal{M}_1^{ps}, \mathcal{M}_j^{om}|\mathcal{D}) \xrightarrow{p} 1$, then*

$$\widehat{\Delta}_{DR}^{MA} \xrightarrow{p} \Delta$$

Theorem 1. *Assume there is no unmeasured confounding, independent observations, and let $\mathcal{M}^{om}$ and $\mathcal{M}^{ps}$ be collections of models. If,*

(1) $\mathcal{M}^{om}$ contains the true models, $\mathcal{M}_1^{om}$, for both $E(Y|X=1,C)$ and $E(Y|X=0,C)$, and $p_{\bullet 1} = \sum_i \mathrm{P}(\mathcal{M}_i^{ps}, \mathcal{M}_1^{om}|\mathcal{D}) \xrightarrow{p} 1$

<u>or</u>

(2) $\mathcal{M}^{ps}$ contains the true model, $\mathcal{M}_1^{ps}$, for $\mathrm{P}(X=1|C)$, and $p_{1\bullet} = \sum_j \mathrm{P}(\mathcal{M}_1^{ps}, \mathcal{M}_j^{om}|\mathcal{D}) \xrightarrow{p} 1$

Then,

$$\widehat{\Delta}_{DR}^{MA} \xrightarrow{p} \Delta$$

Proof. These results can be verified through the use of Slutsky's theorem and standard arguments utilizing the no unmeasured confounding assumption. See the Appendix for details. □

The consistency of $\widehat{\Delta}_{DR}^{MA}$ was shown here in relation to the true propensity score and outcome model. However, the requirement that $\mathcal{M}^{om}$ and $\mathcal{M}^{ps}$ contain the truth could be replaced with the requirement that $\mathcal{M}^{om}$ and $\mathcal{M}^{ps}$ contain a model that is sufficient to control confounding. No longer would the requirement be that posterior model probability be consistent for the truth, but only that the sum of the posterior model probabilities that adequately control confounding converges in probability to 1.

Next we will show that if the collection of models $\mathcal{M}^{om}$ and $\mathcal{M}^{ps}$ contain the true models and the posterior model probabilities are $\sqrt{n}$-consistent for model selection, then $\widehat{\Delta}_{DR}^{MA}$ is asymptotically equivalent to $\widehat{\Delta}_{DR}$ when the true outcome and propensity score models are known a priori and achieves the semiparametric variance bound.

Theorem 2. *Consider $\widehat{\Delta}_{DR}^{MA}$ as described by Equation 4. Assume no unmeasured confounding and independent observations. Let $\mathcal{M}^{om}$ and $\mathcal{M}^{ps}$ be collections of models that contain the true models for $E(Y|X = 1, C)$, $E(Y|X = 0, C)$, and $Pr(X = 1|C)$. Let $\mathcal{M}_1^{om}$ and $\mathcal{M}_1^{ps}$ denote the true outcome and propensity score models, and let $\widehat{\Delta}_{11}^{DR}$ be the double robust estimator using $\mathcal{M}_1^{om}$ and $\mathcal{M}_1^{ps}$. Assume the usual regularity conditions so that $n^{1/2}(\widehat{\Delta}_{11}^{DR} - \Delta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{V})$ where $\mathcal{V}$ is the semiparametric variance bound. If $p_{11} = 1 - o_p(\frac{1}{\sqrt{n}})$, then*

$$n^{1/2}(\widehat{\Delta}_{DR}^{MA} - \Delta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathcal{V})$$

or

$$n^{1/2}(\widehat{\Delta}_{DR}^{MA} - \widehat{\Delta}_{11}^{DR}) = o_p(1)$$

Proof. This result can be verified by recognizing that $p_{11} = 1 - o_p(\frac{1}{\sqrt{n}})$ implies $p_{ij} = o_p(\frac{1}{\sqrt{n}})$ for $(i, j) \neq$

$(1, 1)$, and repeated application of Slutsky's Theorem. See Appendix for details. $\square$

The restriction on the posterior model probabilities ($p_{11} = 1 - o_p(\frac{1}{\sqrt{n}})$) for this result is quite strong and is not expected to be satisfied easily. Technically speaking, all of the previous results hold for any consistent model selection procedure, whether Bayesian or frequentist. This is not surprising, as model averaging and model selection are asymptotically equivalent. However, we expect that in finite samples model averaging and model selection will differ.

3 Simulations

3.1 Set up

It is not the purpose of these simulations to verify the asymptotic properties of $\widehat{\Delta}_{DR}^{MA}$, but instead to illustrate its finite sample behavior relative to the double robust estimator using model selection for both the propensity score and the outcome model ($\widehat{\Delta}_{DR}^{MS}$). We consider the model averaged double robust estimator assuming both prior model independence and prior model dependence. Let $\widehat{\Delta}_{DR}^{MA-i}$ denote the model averaged double robust estimator that assumes prior model independence given by Equation 6, let $\widehat{\Delta}_{DR}^{MA-d}$ denote the model averaged double robust estimator that assumes the prior model dependence given by Equation 7, and let $\widehat{\Delta}_{DR}^{MA-dII}$ be the model averaged double robust estimator that assumes the prior model dependence given by Equation 7 and uses the two-stage approach for calculating model weights. See Table 1 for a description of the estimators considered in these simulations.

We use two groups of simulations. In Group 1, all effects of the potential confounders are linear in both treatment and outcome. In Group 2, we allow for interactions and non-linearities in the confounder-outcome and confounder-treatment relationships. We consider a situation where we have 5 measured potential confounders. In all of our simulations, we restrict $\mathcal{M}^{ps}$ and $\mathcal{M}^{om}$ to only include linear combinations of the potential confounders so that there are $2^5 = 32$ models for both the propensity score and the outcome. Therefore, in Group 2, the true models are not contained in either $\mathcal{M}^{ps}$ or $\mathcal{M}^{om}$.

A full description of all scenarios can be found in Table 2 and Table 3. All simulations set $\beta = 1$ and use a sample size of 500 with 10,000 replications. In Group 1, we generate the data as follows: (1)

$C_1, ..., C_5 \overset{iid}{\sim} N(0,1)$; (2) $X \sim$ Bernoulli$(p = \text{expit}(C\alpha^{ps}))$; and (3) $Y \sim \text{N}(\beta X + C\alpha^{om}, 1)$. We consider different values of the unknown parameters α^{ps} and α^{om} to mimic different levels of confounding.

In Group 2, we generate the data in a similar manner, but with non-linearities in the propensity score or outcome models as follows: (1) $C_1, ..., C_5 \overset{iid}{\sim} N(0,1)$; (2) $X \sim$ Bernoulli$(p = \text{expit}(f(C))$; and (3) $Y \sim \text{N}(\beta X + g(C), 1)$, where $f(\cdot)$ and $g(\cdot)$ are polynomial functions of C. For example, Scenario 7 is linear in the propensity score model, with $f(C) = C_1 + C_2 + C_5$, but non-linear in the outcome, with $g(C) = 0.5 \sum_{i=1}^{5} \sum_{j=1}^{5} C_i C_j$. Additional simulation scenarios and sensitivity analyses are included in the Appendix.

3.2 Results

Recall that the "model selected double robust estimator", $\widehat{\Delta}_{DR}^{MS}$, refers to the double robust estimator in which both the propensity score and the outcome model have been selected independently using BIC based model selection. Table 4 provides the percent decrease in mean squared error of the three model averaged double robust estimators defined in Table 1 when compared to $\widehat{\Delta}_{DR}^{MS}$, for each simulation scenario when the sample size is 500. Strikingly, we observe the smallest mean squared error using $\widehat{\Delta}_{DR}^{MA-dII}$ across all simulation scenarios presented here.

We have found that utilizing model averaging strategies on double robust estimators can reduce mean squared error as compared to the model selected double robust estimator. Our simulations support this claim, as at least one of the model averaged estimators always has a smaller mean squared error when compared with the model selected double robust estimator. This holds even when the true model's functional form is not included in the model class considered.

In the Group 1 simulations, where all effects are linear in the potential confounders, model averaging assuming prior model independence ,$\widehat{\Delta}_{DR}^{MA-i}$, reduces mean squared error up to 1.39% compared to model selection. This is a very modest gain, but demonstrates that simply applying model averaging to account for model uncertainty has a benefit over using standard model selection procedures. The estimators $\widehat{\Delta}_{DR}^{MA-d}$ and $\widehat{\Delta}_{DR}^{MA-dII}$, where we assume prior model dependence, have reductions in mean squared error that are generally larger than those of $\widehat{\Delta}_{DR}^{MA-i}$. In fact, the estimator $\widehat{\Delta}_{DR}^{MA-dII}$ has reductions in mean squared error

that range from 0.36% to 59.33%.

Considering Scenario 3, $\widehat{\Delta}_{DR}^{MA-dII}$ has 59.33% smaller mean squared error when compared to the model selected double robust estimator $\widehat{\Delta}_{DR}^{MS}$. Most of this reduction is in the variance of the estimator, as both estimators have little to no bias. This indicates that even in the case where all potential confounders are linear in both the propensity score and the outcome model, model averaging can reduce the variance of the double robust estimator dramatically if we assume prior model dependence and use the two-stage approach to cut model feedback.

We can explain this reduction in variance in Scenario 3 by noting that only C_1 and C_2 are confounders, while C_3, C_4, and C_5 are strongly associated with the exposure only. Therefore, using model selection on the propensity score model independently of the outcome model will tend to choose models that include C_3, C_4, and C_5. These three potential confounders are unrelated to the outcome, so their inclusion in the propensity score model only adds to the variance of the estimator. By utilizing model averaging with the prior model dependence given by Equation 7, we effectively restrict the model space of the propensity score a priori to be those models that include only the potential confounders that are associated with the outcome (C_1 and C_2). Thus, C_3, C_4, and C_5 are excluded from consideration by the prior distribution because they are unrelated to the outcome, and the reduction in mean squared error can be attributed to the correct identification of the C's associated with both the outcome and the treatment for use in the propensity score model. It is important to note here that the 59.33% reduction in mean squared error occurs when we have both assumed prior model dependence and used the two-stage approach for calculating model weights. The benefit of the latter point is argued in Section 2.3, but in this specific example, the posterior model probabilities used to construct $\widehat{\Delta}_{DR}^{MA-d}$ will favor the outcome model that includes all five potential confounders. This is inefficient because only C_1 and C_2 are associated with the outcome and including C_3, C_4, C_5 - which are only associated with X - into the outcome model, will lead to a large increase in the variance of the estimator. Therefore, cutting the feedback from the propensity score model into the outcome model when calculating the model weights improves efficiency.

In Scenario 4, the estimators $\widehat{\Delta}_{DR}^{MA-i}$, $\widehat{\Delta}_{DR}^{MA-d}$, and $\widehat{\Delta}_{DR}^{MA-dII}$ each reduce the mean squared error by 0.36% when compared with $\widehat{\Delta}_{DR}^{MS}$. This occurs because each of the five potential confounders are associated

with both the exposure and the outcome, where those that are strongly associated with the outcome are moderately associated with the exposure and those that are strongly associated with the exposure are moderately associated with the outcome. Therefore, each method for estimating the posterior model weights will tend to select models that contain all five potential confounders.

Scenario 5 is a reproduction of Scenario 4, but with an additional 5 potential confounders that are unrelated to both the exposure and the outcome. In Scenarios 4 and 5, all the model averaged double robust estimators outperform the model selected double robust estimator. However, we note that when additional potential confounders are added (Scenario 5), the model averaged double robust estimators gain efficiency as compared to the model selected variety. This gain is expected to continue as more potential confounders are added, and since the estimator is scalable to a large number of potential confounders, the efficiency gain in using model averaging over model selection is likely to increase as the number of potential confounders grows.

Moving to Group 2 of the simulations, where the class of models considered is misspecified for either the propensity score or the outcome, $\widehat{\Delta}_{DR}^{MA-i}$ increases the mean squared error in 2 out of the 5 scenarios, the estimator $\widehat{\Delta}_{DR}^{MA-d}$ increases the mean square error in 3 out of 5 scenarios compared to $\widehat{\Delta}_{DR}^{MS}$, and no general conclusion about the comparison of model selection versus model averaging can be made. However, the estimator $\widehat{\Delta}_{DR}^{MA-dII}$ has a smaller mean squared error than $\widehat{\Delta}_{DR}^{MS}$ in all presented scenarios and reduces the mean squared error between 0.38% and 90.1%. In fact, $\widehat{\Delta}_{DR}^{MA-dII}$ appears to reduce mean squared error the most when the outcome model has been misspecified (Scenarios 6-8). In all of the Scenarios 6-8, $\widehat{\Delta}_{DR}^{MA-dII}$ has at least 25% smaller mean squared error than $\widehat{\Delta}_{DR}^{MS}$.

In Scenario 7, the use of model averaging has reduced the variance (again, the mean squared error approximates the variance due to little to no bias) by 90.1%. To put this into perspective, if the model selected double robust estimator had a variance of 10, then the model averaged double robust estimator assuming prior model dependence and using the two-stage approach for calculating model weights would have a variance of about 1.

To illustrate the why cutting the feedback between the propensity score model and outcome model is effective, Tables 5 and 6 provide the marginal outcome model and propensity score model weights, respec-

tively, used in the construction of $\widehat{\Delta}_{DR}^{MA-i}$, $\widehat{\Delta}_{DR}^{MA-d}$, and $\widehat{\Delta}_{DR}^{MA-dII}$ for Scenario 7 averaged over the 10,000 realizations. First, we will compare the model weights that are used in $\widehat{\Delta}_{DR}^{MA-i}$ and $\widehat{\Delta}_{DR}^{MA-dII}$ to describe why $\widehat{\Delta}_{DR}^{MA-i}$ increases the mean squared error by 12.5% while $\widehat{\Delta}_{DR}^{MA-dII}$ reduces mean square error by 90.1% when compared to the model selected double robust estimator. Note that the marginal outcome model weights used in $\widehat{\Delta}_{DR}^{MA-i}$ and $\widehat{\Delta}_{DR}^{MA-dII}$ are the same, so that the difference in the two estimators is due to the propensity score model weights. Referring to Table 6, the model weights used in construction of $\widehat{\Delta}_{DR}^{MA-i}$ assign 83.5% of the mass to the true propensity score model, while the weights used in construction of $\widehat{\Delta}_{DR}^{MA-dII}$ place the mass across many different propensity score models. The reduction in mean squared error can be attributed to the fact that when the outcome is non-linear in the potential confounders, it is unclear if adjusting for confounders linearly in the propensity score is optimal. The estimator $\widehat{\Delta}_{DR}^{MA-dII}$ captures this uncertainty, and the resulting model averaged double robust estimator averages over many different propensity score models resulting in a 90.1% reduction in mean squared error.

Next, we compare the model weights that are used in $\widehat{\Delta}_{DR}^{MA-d}$ and $\widehat{\Delta}_{DR}^{MA-dII}$ to describe why $\widehat{\Delta}_{DR}^{MA-d}$ increases the mean squared error by 61.78% while $\widehat{\Delta}_{DR}^{MA-dII}$ reduces mean square error by 90.1% when compared to $\widehat{\Delta}_{DR}^{MS}$. The prior model dependence forces a potential confounder that is included in the propensity score model to be included in the outcome model, and we see this through the marginal outcome model posterior probabilities used in the construction of $\widehat{\Delta}_{DR}^{MA-d}$. All of the mass is assigned to outcome models that include the three potential confounders that are associated with the treatment (C_1, C_2, and C_5) – only 4 of the 32 models have non-zero mass. The estimator $\widehat{\Delta}_{DR}^{MA-dII}$ distributes the outcome model weight more evenly across the model space, with all 32 outcome models receiving mass between 0.022 and 0.063. A similar result is observed in the marginal propensity score model weights, with $\widehat{\Delta}_{DR}^{MA-d}$ assigning 94.1% of the mass to the true propensity score model and $\widehat{\Delta}_{DR}^{MA-dII}$ distributing the weight across the model space. Therefore, the weights used in the estimator $\widehat{\Delta}_{DR}^{MA-d}$ tends to favor the potential confounders that are associated with the treatment in both the propensity score model and the outcome model. The data generating mechanism is non-linear in the potential confounders; using model selection or assigning most of the posterior weight to a few models that adjust for confounding linearly led to an inefficient estimate. This is important, as it emphasizes that model averaging provides the most gain in efficiency when there is non-linearities in the data

generating mechanism.

To further emphasize this point, Figure 1 plots the model specific double robust estimators $\widehat{\Delta}_{ij}^{DR}$ versus their corresponding posterior weights p_{ij} used in the construction of $\widehat{\Delta}_{DR}^{MA-i}$, $\widehat{\Delta}_{DR}^{MA-d}$, and $\widehat{\Delta}_{DR}^{MA-dII}$ for a single realization of the data in Scenario 7. The vertical line is placed at the value of the corresponding model averaged estimator. It can be seen that when estimating both $\widehat{\Delta}_{DR}^{MA-i}$ and $\widehat{\Delta}_{DR}^{MA-d}$, the posterior mass is assigned to a few models that provide biased estimates of the true $\Delta = 1$. When estimating $\widehat{\Delta}_{DR}^{MA-dII}$, the posterior weight is spread across a different set of models that all provide a less biased estimate of Δ=1. As a reference, the model selected double robust estimate of Δ is $\widehat{\Delta}_{DR}^{MS} = 3.84$, which lies in the region of models that are assigned positive mass when estimating both $\widehat{\Delta}_{DR}^{MA-i}$ and $\widehat{\Delta}_{DR}^{MA-d}$.

Figure 1 provides results for a single realization of the data in Scenario 7, and as such, could be an artifact of randomness. To provide evidence of the contrary, Figure 2 provides a plot averaged over 10,000 realizations of the data in Scenario 7 that is constructed as follows: (1) for each simulated dataset, we round the model specific estimates $\widehat{\Delta}_{ij}^{DR}$ to the nearest whole number; (2) we assign each integer to the sum of the weights p_{ij} of the model specific double robust estimators that are mapped to that integer; and (3) we average the weights that are assigned to each integer over the 10,000 realizations of the data. In the estimation of $\widehat{\Delta}_{DR}^{MA-dII}$, approximately 80% of the posterior weight is assigned to models whose estimates round to the true value of $\Delta = 1$, while in the estimation of $\widehat{\Delta}_{DR}^{MA-i}$ and $\widehat{\Delta}_{DR}^{MA-d}$, only between 40% and 60% of the posterior mass is assigned to these same models. The model specific double robust estimators were rounded to the nearest integer to collapse the estimators based on the quality of the estimate within a given dataset. This allows us to summarize on average, how well do models that are assigned positive weight estimate $\Delta = 1$.

Putting the information from Figure 1 and 2 together, $\widehat{\Delta}_{DR}^{MA-dII}$ is a weighted average of model specific estimates that assigns most of the posterior weight to models that provide better estimates of $\Delta = 1$. This leads directly to $\widehat{\Delta}_{DR}^{MA-dII}$ reducing the mean squared error by 90.1% when compared to $\widehat{\Delta}_{DR}^{MS}$. In comparison, the estimators $\widehat{\Delta}_{DR}^{MA-i}$ and $\widehat{\Delta}_{DR}^{MA-d}$ fail to assign high posterior weight to models that provide good estimates of $\Delta = 1$, leading to more variable estimators.

The decision to compare the efficiency of the model averaged double robust estimator to that of the model

selected double robust estimator was made because in practice, one must always make a decision regarding the models to be used. Without relying on expert knowledge, the only other alternative is to include all of the potential confounders in both the propensity score and outcome models. A sensitivity analysis was performed that indicates the results of our simulations are not sensitive to the choice of using model selection in place of the kitchen sink approach. Additionally, if we allow the potential confounders to be generated in a non-i.i.d. fashion, similar results hold.

4 Discussion

We introduced model averaged double robust estimators, a new class of estimators for the average causal effect that account for model uncertainty. We proved that these estimators extend the popular double robustness property, by only requiring that the propensity score model or the outcome model be within a class of models. We also assessed small sample behavior: in realistic scenarios we showed substantial improvements over approaches that do not consider uncertainty in variable selection.

Our asymptotic results build on the most basic double robust estimator for the average causal effect. It has been demonstrated elsewhere that this double robust estimator can be biased especially when some of the estimated propensity scores are close to zero or are highly variable, and several adjustments to the estimator have been proposed (see [17, 3, 26] for discussion on this topic). The results on consistency of the model averaged double robust estimator will carry over to these other double robust estimators. Also, in the definition of the model averaged double robust estimator, we have assumed that the confounders' effect on the potential outcomes are the same between treatment groups, but this assumption is unnecessary. The methods presented in this paper can easily be extended to situations where the response surface differs between potential outcomes [31] by using separate models for the potential outcomes, and independently calculating the posterior model probabilities.

We devised a two-stage approach for calculating the weights of the propensity score and outcome models. This two-stage approach utilizes a prior distribution on the model space that assumes dependence between a confounder's inclusion in the propensity score and the outcome model while cutting feedback from the propensity score model into the outcome model. Different choices of priors on the model space could have

induced other desirable dependencies between the propensity score and the outcome model. For example, a similar dependence is implicit in the recent method proposed by Wang et al. [30], for continuous exposures in the context of linear regression. This work has recently been extended to binary exposures by Zigler and Dominici [32] in the context of stochastic search variable selections for propensity score models.

In our simulations, we have shown that through this two-stage approach for the calculation of model weights, one can reduce the mean squared error of the proposed estimator substantially —more than 90% in the most extreme cases. Reductions in mean squared error are largest in the likely situation when the outcome model is outside the model class considered and guaranteed to be misspecified. These results are not surprising, as there is a growing body of evidence that the use of non-informative priors for model selection in causal inference is not optimal [2, 4, 23, 29, 30]. While $\widehat{\Delta}_{DR}^{MA-dII}$ had the smallest mean squared error in the majority of the sensitivity analyses presented, in a few cases another estimator was more efficient. In these situations, the difference between the most efficient estimator and $\widehat{\Delta}_{DR}^{MA-dII}$ was minimal. It is unlikely that a researcher will correctly model the outcome; therefore, if a researcher chooses to use a doubly robust approach for estimation of the average causal effect, the two-stage model averaged double robust estimator with prior model dependence $\widehat{\Delta}_{DR}^{MA-dII}$ provides a very attractive implementation.

We restricted our class of models to be linear in the potential confounders, but even in the presence of non-linearities in the data generating mechanism, there were observed reductions in mean squared error as compared with the double robust estimator that performs model selection for the propensity score and the outcome model separately. Extension to nonlinear model classes would be conceptually straightforward.

Further work is needed to investigate whether these conclusions continue to hold when the set of potential confounders is larger and when the sample size is smaller (large p and small n). However, it is legitimate to conjecture that the improvements in efficiency should be greater in both directions, as both would emphasize the difference between model selection and model averaging. From this perspective, we expect that the gains presented in Section 3 should be conservative. For large model spaces, it is not feasible to explore every model combination as we did in our simulation study. However, one could implement Bayesian methods designed for model selection in the high dimensional data setting [8, 15, 10], and use the corresponding posterior model weights in a model averaged double robust estimator.

The methods described in this paper share some similarities with the targeted maximum likelihood super-learner of van der Laan and colleagues [27, 28]. The super-learner acknowledges that no single learner is optimal and attempts to combine learners in a fashion to minimize a loss function via cross-validation. In this sense, model averaging the double robust estimator achieves the same goal, but instead combines candidate estimators via their posterior model probabilities. To further distinguish the methods, one must recognize that in both cases a researcher needs to characterize some underlying part of the true data distribution (e.g. the propensity score), denoted $\mathcal{Q}$, to estimate the average causal effect. The super-learner attempts to find the best estimate of $\mathcal{Q}$ upfront, and then uses this estimate of $\mathcal{Q}$ to construct a single estimator of the average causal effect. In contrast, the model averaged double robust estimator constructs several estimates of the average causal effect based on different parametric models that fully characterize $\mathcal{Q}$, and then directly averages these model specific estimates based on the posterior support of each model.

Causal inference approaches are increasingly used to analyze large observational studies, such as administrative databases used in comparative effectiveness research or environmental epidemiology. In these applications, there seldom is a clear-cut way of determining a priori the precise set of confounders of scientific relevance. At the same time, improvements in computing speed and parallelization are creating the opportunity for a more systematic investigation of alternative specifications for confounding adjustment. In this scenario, the proposed model averaging strategy shows great promise as a data analysis tool to perform robust and consistent inferences with good small sample properties.

5 Acknowledgments

Support for this research was provided by National Institute of Environmental Health Sciences grants 5T32ES007142 and R01-ES012054, National Cancer Institute grant P01-CA134294, Environmental Protection Agency grant RD-83479801, and a Health Effects Institute grant (Dominici). We would like to thank Eric Tchetgen Tchetgen for his useful discussions.

References

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

[2] M A Brookhart. Variable Selection for Propensity Score Models. *American Journal of Epidemiology*, 163(12):1149–1156, March 2006.

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

[4] Ciprian M. Crainiceanu, Francesca Dominici, and Giovanni Parmigiani. Adjustment uncertainty in effect estimation. *Biometrika*, 95(3):635–651, 2008.

[5] C Drake. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, 1993.

[6] David Draper. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):pp. 45–97, 1995.

[7] B Efron and R Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, NY, 1993.

[8] E.I. George and R.E. McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

[9] Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. Bayesian model averaging: a tutorial (with comments by m. clyde, david draper and e. i. george, and a rejoinder by the authors). *Statistical Science*, 14(4):382–417, 1999.

[10] V.E. Johnson and D. Rossell. Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107(498):649–660, 2012.

[11] R.E. Kass and A.E. Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

[12] Sadanori Konishi and Genshiro Kitagawa. Generalised information criteria in model selection. *Biometrika*, 83(4):875–890, 1996.

[13] JK Lunceford and M Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23(19):2937–2960, 2004.

[14] Ryuei Nishii. Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*, 12(2):758–765, 1984.

[15] R.B. O'Hara and M.J. Sillanpää. A review of bayesian variable selection methods: what, how and which. *Bayesian Analysis*, 4(1):85–117, 2009.

[16] A.E. Raftery, D. Madigan, and J.A. Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92:179–191, 1997.

[17] James Robins, Mariela Sued, Quanhong Lei-Gomez, and Andrea Rotnitzky. Comment: Performance of double-robust estimators when"inverse probability" weights are highly variable. *Statistical Science*, 22(4):544–559, 2007.

[18] JM Robins, MA Hernan, and B Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

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

[20] Paul R. Rosenbaum and Donald B. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79:516–524, 1984.

[21] D.B. Rubin et al. Estimating causal effects from large data sets using propensity scores. *Annals of internal medicine*, 127:757–763, 1997.

[22] Daniel O. Scharfstein, Andrea Rotnitzky, and James M. Robins. Reply to comments on "Adjusting for nonignorable drop-out using semiparametric nonresponse models". *Journal of the American Statistical Association*, 94:1135–1146, 1999.

[23] Sebastian Schneeweiss, Jeremy A Rassen, Robert J Glynn, Jerry Avorn, Helen Mogun, and M Alan Brookhart. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*, 20(4):512–22, Jul 2009.

[24] Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978.

[25] A.F.M. Smith and D.J. Spiegelhalter. Bayes factors and choice criteria for linear models. *RoyalStat*, B-42:213–220, 1980.

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

[27] Mark van der Laan, Eric Polley, and Alan Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007.

[28] M.J. van der Laan. Targeted maximum likelihood based causal inference: part i. *The International Journal of Biostatistics*, 6(2), 2010.

[29] Stijn Vansteelandt, Maarten Bekaert, and Gerda Claeskens. On model selection and model misspecification in causal inference. *Statistical methods in medical research*, November 2010.

[30] Chi Wang, Giovanni Parmigiani, and Francesca Dominici. Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics*, 2012.

[31] Chi Wang, Giovanni Parmigiani, and Francesca Dominici. Rejoinder: Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics*, 2012.

[32] Corwin Zigler and Francesca Dominici. Uncertainty in propensity score estimation: Bayesian methods for variable selection and model averaged causal effects. *Technical Report*, 2012.

Estimator	Description
$\widehat{\Delta}_{DR}^{MS}$	model selected double robust estimator that chooses propensity model and the outcome model based on the BIC
$\widehat{\Delta}_{DR}^{MA-i}$	MA-DR estimator assuming prior model independence
$\widehat{\Delta}_{DR}^{MA-d}$	MA-DR estimator assuming prior model dependence defined by Equation 7
$\widehat{\Delta}_{DR}^{MA-dII}$	MA-DR estimator assuming prior model dependence defined by Equation 7 and using the two-stage approach for calculating model weights

Table 1: Description of all estimators used in the simulations. Included is (1) the type of estimator; and (2) the choice of prior distribution for the model space. All Bayes factors are estimated using the BIC approximation.

Scenario	α^{ps} (PS model)	α^{om} (Outcome model)
1	(0.4,0.3,0.2,0.1,0)	(0,0,0,0,0)
2	(0.5,0.5,0.1,0,0)	(0.5,0,1,0.5,0)
3	(0.1,0.1,1,1,1)	(2,2,0,0,0)
4	(0.5,0.4,0.3,0.2,0.1)	(0.5,1,1.5,2,2.5)
5	(0.5,0.4,0.3,0.2,0.1,0,0,0,0,0)	(0.5,1,1.5,2,2.5,0,0,0,0,0)

Table 2: Simulation Group 1: All effects of confounders are linear on both the treatment and outcome. Data is generated as follows: (1) $C_1, ..., C_5 \overset{iid}{\sim} N(0,1)$; (2) $X \sim$ Bernoulli$(p = \text{expit}(C\alpha^{ps}))$; and (3) $Y \sim \text{N}(\beta X + C\alpha^{om}, 1)$.

Scenario	$f(C)$ (PS model)	$g(C)$ (Outcome model)
6	$.5C_1 + .5C_2 + .1C_3$	$C_3 + C_4 + C_5 + \sum_{ij} C_i C_j$
7	$C_1 + C_2 + C_5$	$\sum_{ij} 0.5 C_i C_j$
8	$.2C_1 + .2C_2 + .2C_5$	$.25C_3 + (C_1 + C_2)^2 - (C_1^2 - C_3)^2 + (C_4^2 - .5C_5)(C_3 - .5C_4)$
9	$C_3 + C_4 + C_5 + \sum_{ij} C_i C_j$	$.5C_1 + .5C_2 + .1C_3$
10	$(C_1 + C_2 + .5C_3)^2$	$.5C_1 + .5C_3 + .5C_4$

Table 3: Simulation Group 2: Effects of potential confounders are allowed to be non-linear on both the treatment and outcome. Data is generated as follows: (1) $C_1, ..., C_5 \overset{iid}{\sim} N(0,1)$; (2) $X \sim$ Bernoulli$(p = \text{expit}(f(C))$; and (3) $Y \sim \text{N}(\beta X + g(C), 1)$, where $f(\cdot)$ and $g(\cdot)$ are polynomial functions of C.

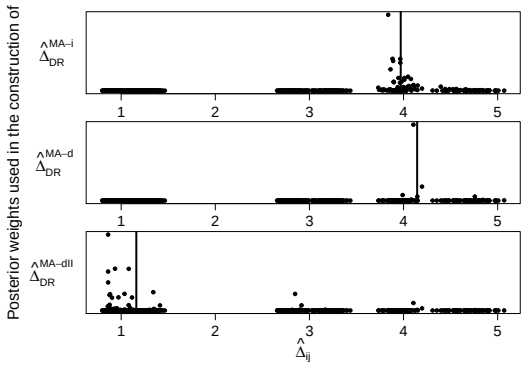

Figure 1: The model specific double robust estimators $\widehat{\Delta}_{ij}^{DR}$ versus their corresponding posterior weights p_{ij} used in the construction $\widehat{\Delta}_{DR}^{MA-i}$, $\widehat{\Delta}_{DR}^{MA-d}$, and $\widehat{\Delta}_{DR}^{MA-dII}$ of for a single realization of the data in Scenario 7. The vertical line is placed at the value of the corresponding model averaged estimator. The true value of Δ is 1. See Table 1 for definition of each estimator and Table 3 for a description of Scenario 7.

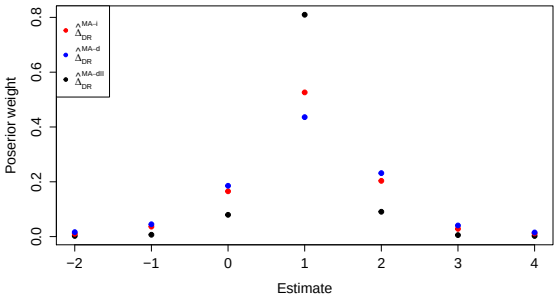

Figure 2: Summary of the posterior weights p_{ij} averaged over 10,000 realizations of the data in Scenario 7 that is constructed as follows: (1) for each simulated dataset, the model specific estimates $\widehat{\Delta}_{ij}^{DR}$ are rounded to the nearest whole number; (2) each integer is assigned the sum of the weights p_{ij} of the model specific double robust estimators that are mapped to that integer; and (3) average the weights that are assigned to each integer over the 10,000 realizations of the data. The horizontal axis is the value of the model specific double robust estimators that have been rounded to the nearest integer, and the vertical axis is the sum of the posterior weights of the corresponding model averaged double robust estimators that round to the specified integer averaged over the 10,000 realizations. The true value of Δ is 1. See Table 1 for definition of each estimator and Table 3 for a description of Scenario 7.

	Percent reduction in MSE		
Scenario	$\widehat{\Delta}_{DR}^{MA-i}$	$\widehat{\Delta}_{DR}^{MA-d}$	$\widehat{\Delta}_{DR}^{MA-dII}$
1	0.60	1.35	**5.42**
2	-0.03	-0.01	**5.23**
3	1.39	2.00	**59.33**
4	0.36	0.36	0.36
5	0.76	1.08	**1.27**
6	-1.16	-5.75	**29.1**
7	-12.5	-61.78	**90.1**
8	0.87	1.48	**29.9**
9	1.24	-0.65	**5.94**
10	0.19	0.47	**0.38**

Table 4: The percent reduction in mean square error as compared to the model selected double robust estimator when the sample size is 500 for various model averaged double robust estimators. See Table 1 for definition of each estimator and Tables 2 and 3 for descriptions of each scenario. Bold indicates estimator with smallest MSE.

Model	Systematic Component	$\widehat{\Delta}_{DR}^{MA-i}$	$\widehat{\Delta}_{DR}^{MA-d}$	$\widehat{\Delta}_{DR}^{MA-dII}$
1	C_1	0.043	0	0.043
2	C_2	0.043	0	0.043
3	C_3	0.045	0	0.045
4	C_4	0.046	0	0.046
5	C_5	0.043	0	0.043
6	$C_1 + C_2$	0.031	0	0.031
7	$C_1 + C_3$	0.032	0	0.032
8	$C_1 + C_4$	0.032	0	0.032
9	$C_1 + C_5$	0.031	0	0.031
10	$C_2 + C_3$	0.032	0	0.032
11	$C_2 + C_4$	0.032	0	0.032
12	$C_2 + C_5$	0.032	0	0.032
13	$C_3 + C_4$	0.035	0	0.035
14	$C_3 + C_5$	0.032	0	0.032
15	$C_4 + C_5$	0.032	0	0.032
16	$C_1 + C_2 + C_3$	0.025	0	0.025
17	$C_1 + C_2 + C_4$	0.025	0	0.025
18	$C_1 + C_2 + C_5$	0.027	0.437	0.027
19	$C_1 + C_3 + C_4$	0.026	0	0.026
20	$C_1 + C_3 + C_5$	0.025	0	0.025
21	$C_1 + C_4 + C_5$	0.025	0	0.025
22	$C_2 + C_3 + C_4$	0.026	0	0.026
23	$C_2 + C_3 + C_5$	0.024	0	0.024
24	$C_2 + C_4 + C_5$	0.025	0	0.025
25	$C_3 + C_4 + C_5$	0.025	0	0.025
26	$C_1 + C_2 + C_3 + C_4$	0.022	0	0.022
27	$C_1 + C_2 + C_3 + C_5$	0.025	0.216	0.025
28	$C_1 + C_2 + C_4 + C_5$	0.025	0.216	0.025
29	$C_1 + C_3 + C_4 + C_5$	0.022	0	0.022
30	$C_2 + C_3 + C_4 + C_5$	0.022	0	0.022
31	$C_1 + C_2 + C_3 + C_4 + C_5$	0.027	0.131	0.027
32	intercept only	0.063	0	0.063

Table 5: The marginal posterior outcome model weights used in the construction of $\widehat{\Delta}_{DR}^{MA-i}$, $\widehat{\Delta}_{DR}^{MA-d}$, and $\widehat{\Delta}_{DR}^{MA-dII}$ for each model in $\mathcal{M}^{om}$ in Scenario 7 and sample size 500 averaged over the 10,000 realizations. See Table 1 for a description of the estimators and Table 3 for a description of Scenario 7.

Model	Systematic Component	$\widehat{\Delta}_{DR}^{MA-i}$	$\widehat{\Delta}_{DR}^{MA-d}$	$\widehat{\Delta}_{DR}^{MA-dII}$
1	C_1	0	0	0.126
2	C_2	0	0	0.127
3	C_3	0	0	0.021
4	C_4	0	0	0.021
5	C_5	0	0	0.126
6	$C_1 + C_2$	0	0	0.099
7	$C_1 + C_3$	0	0	0.003
8	$C_1 + C_4$	0	0	0.003
9	$C_1 + C_5$	0	0	0.094
10	$C_2 + C_3$	0	0	0.004
11	$C_2 + C_4$	0	0	0.002
12	$C_2 + C_5$	0	0	0.097
13	$C_3 + C_4$	0	0	0
14	$C_3 + C_5$	0	0	0.003
15	$C_4 + C_5$	0	0	0.003
16	$C_1 + C_2 + C_3$	0	0	0.003
17	$C_1 + C_2 + C_4$	0	0	0.003
18	$C_1 + C_2 + C_5$	0.835	0.941	0.095
19	$C_1 + C_3 + C_4$	0	0	0
20	$C_1 + C_3 + C_5$	0	0	0.006
21	$C_1 + C_4 + C_5$	0	0	0.003
22	$C_2 + C_3 + C_4$	0	0	0
23	$C_2 + C_3 + C_5$	0	0	0.003
24	$C_2 + C_4 + C_5$	0	0	0.003
25	$C_3 + C_4 + C_5$	0	0	0
26	$C_1 + C_2 + C_3 + C_4$	0	0	0
27	$C_1 + C_2 + C_3 + C_5$	0.08	0.029	0.004
28	$C_1 + C_2 + C_4 + C_5$	0.078	0.029	0.004
29	$C_1 + C_3 + C_4 + C_5$	0	0	0
30	$C_2 + C_3 + C_4 + C_5$	0	0	0
31	$C_1 + C_2 + C_3 + C_4 + C_5$	0.007	0.001	0
32	intercept only	0	0	0.146

Table 6: The marginal posterior propensity score model weights used in the construction of $\widehat{\Delta}_{DR}^{MA-i}$, $\widehat{\Delta}_{DR}^{MA-d}$, and $\widehat{\Delta}_{DR}^{MA-dII}$ for each model in $\mathcal{M}^{om}$ in Scenario 7 and sample size 500 averaged over the 10,000 realizations. See Table 1 for a description of the estimators and Table 3 for a description of Scenario 7.