Multiply robust estimation for causal inference problems

Article · November 2016

3 authors, including:

Emma McCoy
Imperial College London
25 PUBLICATIONS   **609** CITATIONS

SEE PROFILE

Daniel Graham
Imperial College London
176 PUBLICATIONS   **5,321** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project   International Bus Benchmarking Group View project

Project   PRACT - Predicting Road Accidents: A transferable methodology across Europe View project

Multiply robust dose-response estimation for multivalued causal inference problems

Cian Naik[1], Emma J. McCoy[1], and Daniel J. Graham[2]

[1]Department of Mathematics, Imperial College London, London, UK
[2]Corresponding author: Department of Civil Engineering, Imperial College London, London, SW7 2AZ, UK. Email: d.j.graham@imperial.ac.uk

Abstract

This paper develops a multiply robust (MR) dose-response estimator for causal inference problems involving multivalued treatments. We combine a family of generalised propensity score (GPS) models and a family of outcome regression (OR) models to achieve an average potential outcomes estimator that is consistent if just one of the GPS or OR models in each family is correctly specified. We provide proofs and simulations that demonstrate multiple robustness in the context of multivalued causal inference problems.

Keywords: Multiply robust; causal inference; multivalued treatment; propensity score; outcome regression.

1 Introduction

In recent years a great deal of attention has been devoted to doubly robust (DR) estimators for causal inference which combine propensity score (PS) and outcome regression (OR) models to give an average treatment effect (ATE) estimator that is consistent and asymptotically normal under correct specification of just one of the two component models. The key advantage of the DR approach is that it affords the analyst two chances at correct inference. DR estimators were first introduced by Scharfstein et al. (1999) and have been studied extensively in theory and application in the binary treatment setting (e.g. Robins 2000, Robins et al. 2000, Robins and Rotnitzky 2001, van der Laan and Robins 2003, Lunceford and Davidian 2004, Bang and Robins 2005, Kang and Schafer 2007). A small number of papers have formulated DR models for dose-response estimation (e.g. Flores and Mitnik 2009, Zhang et al. 2016, Graham et al. 2016).

The recent missing data literature has developed multiply robust (MR) estimation approaches, which consistently estimate a population mean, given ignorable missingness, if just one of multiple combined models is correctly specified. Since causal inference methods essentially treat unobserved potential outcomes as missing data, there are opportunities for translation of missing data methods in the causal inference setting. In this paper we generalise and apply the binary MR missing data estimation approach to address causal inference problems for dose-response estimation involving multivalued treatments.

In so doing we aim to make two main contributions to the causal inference literature. First, we advance methods for dose-response estimation based on the generalised propensity score

(GPS) (e.g. Imbens 2000, Hirano and Imbens 2004), which have been understudied relative to binary treatment problems. Second, we contribute to the exiting literature on robust modelling for causal inference by demonstrating how an MR model can be formulated and applied in the causal setting.

Building on the work of Han and Wang (2013), Han (2014a) and Han (2014b) we formulate our MR approach by combining a family of GPS models, $\mathcal{P}$, and a family of OR models, $\mathcal{A}$, to construct an average potential outcome (APO) estimator that is consistent if either $\mathcal{P}$ contains a correctly specified PS model or $\mathcal{A}$ contains a correctly specified OR model. We provide proofs of MR properties for APO estimation of multivalued treatments and we present simulations which demonstrate that the approach is robust to problems of confounding and functional form misspecification which are routinely encountered in causal analyses.

The paper is structured as follows. Section two briefly outlines the theory of doubly robust estimation. Section three defined the concept of multiply robust estimation in the context of causal inference problems involving multivalued treatments, provides proofs of MR properties for estimation of average potential outcomes (APOs), and proposes an algorithm for numerical implementation. Simulation results are then presented in section four. Conclusions are drawn in the final section.

2   Doubly Robust Estimation

Causal inference models are designed to estimate the effect of a treatment on an outcome of interest. Typically, the set up is one in which the treatment is not randomly assigned, and consequently, confounding may be present. Let $z_i = (y_i, d_i, x_i)$, $i = (1, ..., n)$, be a random vector of observed data available for causal modelling, where for the i-th unit of observation y_i denotes an outcome (or response), d_i the treatment received, and x_i a vector of pre-treatment covariates that represent confounding. The treatment could be binary, taking values in $\mathcal{D} \in \{0, 1\}$; or multivalued, with Q treatment categories $\mathcal{D} \equiv (d_1, d_2, ..., d_Q)$. We use lower case d_q to denote specific potential, as distinct from observed, treatments.

For every level of treatment d_q we can define a potential outcome, $Y_i(d_q)$, with $\mathcal{Y}_i = \{Y_i(d_q) : d_q \in \mathcal{D}\}$ denoting the full set of potential outcomes. The estimands of interest in causal inference studies include average potential outcomes (APOs), $APO = \mathbb{E}[Y(d_q)]$, which measure the average response across the population under treatment level $D_i = d_q$; or average treatment effects (ATEs), $ATE = \mathbb{E}[Y(d_q) - Y(0)]$, which measure the difference in average response that would occur under treatment levels $D_i = d_q$ and $D_i = 0$.

To estimate a 'causal' APO or ATE we must adjust for relevant unit level confounding characteristics in order to capture the marginal effect of the treatment. Identification of causal estimands requires three key assumptions to hold. First, responses must be conditionally independent of treatment assignment given observed covariates; i.e. $(Y_i(0), Y_i(1)) \perp\!\!\!\perp I_1(D_i)|X_i$, in the case of binary treatments, and $Y_i(d_q) \perp\!\!\!\perp I_{d_q}(D_i)|X_i$ for all $d_q \in \mathcal{D}$ in the case of multivalued treatments; where $I_{d_q}(D_i)$ is the indicator function for receiving treatment level d_q. Second, conditional on covariates X_i, the probability of assignment to treatment must be strictly positive for all x and d, or at least this must hold within some region of treatment, $\mathcal{C} \subseteq \mathcal{D}$ say, referred to as the common support region. Third, the relationship between observed and potential outcomes must comply with the Stable Unit Treatment Value Assumption

(SUTVA)(e.g. Rubin 1978), which requires that the observed response under a given treatment allocation is equivalent to the potential response under that treatment allocation. For binary treatments the SUTVA implies $Y_i = I_1(D_i)Y_i(1) + (1 - I_1(D_i))Y_i(0)$ for all $i = 1, ..., N$, and for multivalued treatments $Y_i \equiv I_{d_q}(D_i)Y_i(d_q)$ for all $d_q \in \mathcal{D}$, for all $Y_i(d_q) \in \mathcal{Y}_i$, and for $i = 1, ..., N$.

If these three conditions hold then the estimands of interest are identified and estimation may proceed via a number of different approaches (for a review see Imbens and Rubin 2015). Two are of particular interest in this paper. First, we can proceed by positing an outcome regression (OR) model $\Psi^{-1}\{m(X_i, D_i; \beta)\}$ for the mean response $\mathbb{E}(Y_i|D_i, X_i)$, where Ψ is a known link function, $m()$ is a regression function, and β is an unknown parameter vector. If the OR model is correctly specified APOs and ATEs can be consistently estimated by averaging over the covariates for given values of d_q. Second, we can assume a model for $f_{D|X}(d_i|x_i)$, the conditional density of the treatment given the covariates, and use this model to estimate propensity scores (PSs), which we denote $\hat{\pi}(D_i|X_i; \hat{\alpha})$ for unknown parameter vector α. PS weighting estimators of the form attributed to Horvitz and Thompson (1952) can then be used to estimate ATEs consistently if the PS model is correctly specified.

Doubly robust (DR) estimation combine OR and PS models in such a way that consistent estimates of APOs and ATEs can be obtained if just one of two component models are correctly specified. This can be achieved by weighting the OR model with a function of the inverse PS values. Using the notation of Lunceford and Davidian (2004), the DR estimator for binary treatment APOs can be written

$$\hat{\mu}_{DR}(d_q) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{I_{d_q}(D_i)Y_i}{h(d_q|X_i; \alpha)} - \frac{I_{d_q}(D_i) - h(d_q|X_i; \alpha)}{h(d_q|X_i; \alpha)} \Psi^{-1}\{m(X_i, d_q; \hat{\beta})\} \right] \tag{1}$$

where

$$h(I_{d_q}(D_i)|X_i; \alpha) = I_{d_q}(D_i)\hat{\pi}(D_i|X_i; \hat{\alpha}) + [1 - I_{d_q}(D_i)]\{1 - \hat{\pi}(D_i|X_i; \hat{\alpha})\}$$

A DR estimator for the ATE follows as $\hat{\tau}_{DR}(1) = \hat{\mu}_{DR}(1) - \hat{\mu}_{DR}(0)$ (for a proof see Lunceford and Davidian 2004).

For multivalued treatments, we can extend the binary approach as follows

$$\hat{\mu}_{DR}(d_q) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{I_{d_q}(D_i)Y_i}{\hat{\pi}(d_q|X_i; \hat{\alpha})} - \frac{I_{d_q}(D_i) - \hat{\pi}(d_q|X_i; \hat{\alpha})}{\hat{\pi}(d_q|X_i; \hat{\alpha})} \Psi^{-1}\{m(X_i, d_q; \hat{\beta})\} \right] \tag{2}$$

Where $\hat{\pi}(d_q|X_i; \hat{\alpha})$ is a model for

$$\pi(d_q|X_i) = \mathbb{P}(D_i = d_q|X_i) = \mathbb{E}(I_{d_q}(D_i)|X_i)$$

which we form by assuming a p.m.f. $f_{D|X}(d_q|x_i, \alpha)$ for the GPS $\pi(d_q|x_i)$, and estimating the parameter $\hat{\alpha}$ from a regression model using the observed treatment doses D_i and covariates X_i (for details see Imbens 2000). A proofs of the DR properties of the multivalued dose-response estimator shown in equation (2), which to our knowledge has not previously appeared in the literature, is given in the appendix.

3 A multiply robust dose-response estimator

Having discussed DR estimators, the question naturally arises as to whether estimators can be constructed which are 'more' than DR, and what this 'multiple' robustness would entail.

A number of recent papers have explored the issue of multiple robustness in the context of missing data problems including Chan (2013) and Chan and Yam (2014). Here we build on the work of Han and Wang (2013), Han (2014a) and Han (2014b) which develop the empirical likelihood based approach of Qin and Zhang (2007) to construct an MR estimation approach for missing data that consistently estimate a population mean given ignorable missingness. We generalise this work to the multivalued setting and derive a type of MR estimator for causal inference dose-response estimation problems in the following sense.

Definition 1. *(Multiple-robustness) We postulate multiple PS models $\mathcal{P} = \{\pi^j(x; \alpha^j) : j = 1, ..., J\}$ and multiple OR models $\mathcal{A} = \{a^k(x, d; \beta^k) : k = 1, ...K\}$. An estimator of an ATE or APO combining all of these models is multiply robust if it is consistent when either $\mathcal{P}$ contains a correctly specified model for the propensity score or $\mathcal{A}$ contains a correctly specified outcome regression model.*

We develop our approach for multivalued treatments noting that results for binary treatments follow directly as a special case.

For a multivalued treatment in a causal inference problem, we denote the number of observed responses with treatment level d_q as m_q with $\mathcal{M}_q = \{i : D_i = d_q\}$, $q = 1, ..., Q$ for Q levels of treatment, and $m_q = |\mathcal{M}_q|$. We specify a family of PS models $\mathcal{P} = \{\pi^j(x; \alpha^j) : j = 1, ..., J\}$ to estimate α from $f_{D|X}(d|x_i, \alpha)$, using the observed treatment doses D_i and covariates X_i. From this we calculate estimates $\hat{\pi}^j(d|X_i; \hat{\alpha}^j)$ for $\pi(d|X_i; \alpha)$. We next postulate a family of OR models $\mathcal{A} = \{a^k(x, d; \beta^k) : k = 1, ...K\}$. Using $\mathcal{P}$ and $\mathcal{A}$ we can now define our MR estimator for a multivalued treatment.

Definition 2. *(Multiply-Robust estimator for multivalued treatment) For possible treatment levels $d_1, ..., d_Q$ we define $\hat{\mu}_{MR}(d_q)$ by*

$$\hat{\mu}_{MR}(d_q) = \sum_{i \in \mathcal{M}_q} \hat{w}_{i,d_q} Y_i \tag{3}$$

We define the $\hat{w}_{i,d_q}$ as follows:

1. *Postulate multiple OR models $\mathcal{A} = \{a^k(x, d; \beta^k) : k = 1, ..., K\}$ for $a(d, x)$ and estimate the parameters $\hat{\beta}^k$ from the regression coefficients.*

2. *Assume a p.m.f. $f_{D|X}(d|x_i, \alpha)$ for the GPS and postulate multiple PS models $\mathcal{P} = \{\pi^j(x, d; \alpha^j) : j = 1, ..., J\}$ to estimate the parameters $\hat{\alpha}^j$. From this, calculate the estimates $\hat{\pi}^j(d_q|X_i; \hat{\alpha}^j)$ for $j = 1, ..., J$. We can think of this more simply as postulating a family $\mathcal{P}' = \{\hat{\pi}^j(d_q|X_i; \alpha^j) : j = 1, ..., J\}$ of models for $\pi(d_q|x)$, and then estimating the parameters $\hat{\alpha}^j$ from the data.*

3. *Define $\hat{\theta}^j = \frac{1}{n} \sum_{i=1}^n \hat{\pi}^j(d_q|X_i; \hat{\alpha}^j)$ and $\hat{\eta}^k = \frac{1}{n} \sum_{i=1}^n a^k(X_i, d_q; \hat{\beta}^k)$*

4. *Letting $\hat{\alpha}^T = \{(\hat{\alpha}^1)^T, ..., (\hat{\alpha}^j)^T\}$ and $\hat{\beta}^T = \{(\hat{\beta}^1)^T, ..., (\hat{\beta}^K)^T\}$, define*

$$\begin{aligned} \hat{g}_i(\hat{\alpha}, \hat{\beta}) = \{\hat{\pi}^1(d_q|X_i; \hat{\alpha}^1) - \hat{\theta}^1, ..., \hat{\pi}^j(d_q|X_i; \hat{\alpha}^j) - \hat{\theta}^j, \\ a^1(X_i, d_q; \hat{\beta}^1) - \hat{\eta}^1, ..., a^K(X_i, d_q; \hat{\beta}^K) - \hat{\eta}^K\}^T \end{aligned} \tag{4}$$

5. *For $i \in \mathcal{M}_q$, set*

$$\hat{w}_{i,d_q} = \frac{1}{m_q} \frac{1}{1 + \hat{\rho}^T \hat{g}_i(\hat{\alpha}, \hat{\beta})} \bigg/ \left\{ \frac{1}{m_q} \sum_{i \in \mathcal{M}_q} \frac{1}{1 + \hat{\rho}^T \hat{g}_i(\hat{\alpha}, \hat{\beta})} \right\} \tag{5}$$

4

Where $\hat{\rho}^T = (\hat{\rho}_1, ..., \hat{\rho}_{J+K})$ is a $(J+K)$-dimensional vector satisfying:

$$\sum_{i \in \mathcal{M}_q} \frac{\hat{g}_i(\hat{\alpha}, \hat{\beta})}{1 + \hat{\rho}^T \hat{g}_i(\hat{\alpha}, \hat{\beta})} = 0 \tag{6}$$

Theorem 3. *The MR estimator has the multiply-robust property for estimating the APOs $\mu(d_q)$ in the multivalued treatment case.*

We start by considering the case in which $\mathcal{P}$ contains a correctly specified PS model, say $\pi^1(d|x; \alpha^1)$ without loss of generality. Let α_0^1 denote the true value of α^1 such that $\pi^1(d|x; \alpha_0^1) = \pi(d|x)$, and let $\hat{\alpha}^1$ be its estimator that we fit from our PS model.

Following Han and Wang (2013), we define the empirical probability of (Y_i, X_i) conditional on $D_i = d_q$ to be p_i for $i = 1, ..., m_q$. Here, we index the subjects with $D_i = d_q$ by $1, ..., m_q$ without loss of generality. Conditional on $D = d_q$ we then have the following lemma

Lemma 4.

$$\mathbb{E}\left[\frac{\pi^j(d_q|X; \alpha^j) - \mathbb{E}\{\pi^j(d_q|X; \alpha^j)\}}{\pi(d_q|X)} \middle| D = d_q\right] = 0 \qquad j = 1, ..., J,$$

$$\mathbb{E}\left[\frac{a^k(X, d_q; \beta^k) - \mathbb{E}\{a^k(X, d_q; \beta^k)\}}{\pi(d_q|X)} \middle| D = d_q\right] = 0 \qquad k = 1, ..., K \tag{7}$$

Proof. The case of $\pi^j(d_q|X; \alpha^j)$ for arbitrary j follows from the fact that

$$0 = \mathbb{E}\left(\frac{I_{d_q}(D)}{\pi(d_q|X)} [\pi^j(d_q|X; \alpha^j) - \mathbb{E}\{\pi^j(d_q|X; \alpha^j)\}]\right)$$

$$= \mathbb{P}(I_{d_q}(D) = 1)\mathbb{E}\left(\frac{I_{d_q}(D)}{\pi(d_q|X)} \left[\pi^j(d_q|X; \alpha^j) - \mathbb{E}\{\pi^j(d_q|X; \alpha^j)\}\right] \middle| I_{d_q}(D) = 1\right)$$

$$+ \mathbb{P}(I_{d_q}(D) = 0)\mathbb{E}\left(\frac{I_{d_q}(D)}{\pi(d_q|X)} \left[\pi^j(d_q|X; \alpha^j) - \mathbb{E}\{\pi^j(d_q|X; \alpha^j)\}\right] \middle| I_{d_q}(D) = 0\right)$$

$$= \mathbb{P}(I_{d_q}(D) = 1)\mathbb{E}\left(\frac{\pi^j(d_q|X; \alpha^j) - \mathbb{E}\{\pi^j(d_q|X; \alpha^j)\}}{\pi(d_q|X)} \middle| I_{d_q}(D) = 1\right)$$

$$= \mathbb{P}(D = d_q)\mathbb{E}\left(\frac{\pi^j(d_q|X; \alpha^j) - \mathbb{E}\{\pi^j(d_q|X; \alpha^j)\}}{\pi(d_q|X)} \middle| D = d_q\right)$$

This then proves the lemma, because the probabilistic assignment condition described in Section two implies that $\mathbb{P}(D = d_q) > 0$. The case of $a^k(X, d_q; \beta^k)$ with arbitrary k follows from a similar argument. $\square$

We then proceed as in the original proof. Lemma 4 implies that the most plausible value for

p_i should be defined through the constrained optimization problem

$$\max_{p_1,\ldots,p_{m_q}} \prod_{i=1}^{m_q} p_i \quad \text{subject to} \quad p_i \geq 0, \ (i=1,\ldots,m_q), \quad \sum_{i=1}^{m_q} p_i = 1$$

$$\sum_{i=1}^{m_q} p_i \frac{\pi^j(d_q|X_i;\alpha^j) - \hat{\theta}^j}{\pi^1(d_q|X_i;\alpha^1)} = 0$$

$$\sum_{i=1}^{m_q} p_i \frac{a^k(X_i,d_q;\beta^k) - \hat{\eta}^k}{\pi^1(d_q|X_i;\alpha^1)} = 0 \tag{8}$$

The first two constraints make the p_i empirical probabilities, and the last two are the empirical versions of the statements of Lemma 4, using the fact that $\pi^1(d_q|X;\alpha^1)$ is correctly specified.

Applying Lagrange multipliers gives

$$\hat{p}_i = \frac{1}{m_q} \frac{1}{1 + \hat{\lambda}^T \hat{g}_i(\hat{\alpha},\hat{\beta})/\pi^1(d_q|X_i;\alpha^1)}$$

Where $\hat{\lambda}^T = (\hat{\lambda}_1, \ldots, \hat{\lambda}_{J+K})$ is a $(J+K)$-dimensional vector satisfying:

$$\sum_{i=1}^{m_q} \frac{\hat{g}_i(\hat{\alpha},\hat{\beta})/\pi^1(d_q|X_i;\alpha^1)}{1 + \hat{\lambda}^T \hat{g}_i(\hat{\alpha},\hat{\beta})/\pi^1(d_q|X_i;\alpha^1)} = 0 \tag{9}$$

Comparing this to (6), we see that a solution to (6) is given by $\hat{\rho}^\pi$ with $\hat{\rho}_1^\pi = (\hat{\lambda}_1 + 1)/\hat{\theta}^1$ and $\hat{\rho}_l^\pi = \hat{\lambda}_l/\hat{\theta}^1$ for $l = 2, \ldots, J+K$. Inserting this into (5) gives

$$\hat{w}_{i,d_q} = \frac{1}{m_q} \frac{\hat{\theta}^1/\pi^1(d_q|X_i;\alpha^1)}{1 + \hat{\lambda}^T \hat{g}_i(\hat{\alpha},\hat{\beta})/\pi^1(d_q|X_i;\alpha^1)} \bigg/ \left\{ \frac{1}{m_q} \sum_{i=1}^{m_q} \frac{\hat{\theta}^1/\pi^1(d_q|X_i;\alpha^1)}{1 + \hat{\lambda}^T \hat{g}_i(\hat{\alpha},\hat{\beta})/\pi^1(d_q|X_i;\alpha^1)} \right\} \tag{10}$$

By the definition of $\hat{p}_i$, we then have that (when $\hat{w}_{i,d_q}$ is given by (5) and $\hat{\rho} = \hat{\rho}^\pi$)

$$\hat{w}_{i,d_q} = \frac{\hat{p}_i \hat{\theta}^1}{\pi^1(d_q|X_i;\alpha^1)} \bigg/ \sum_{i=1}^{m_q} \frac{\hat{p}_i \hat{\theta}^1}{\pi^1(d_q|X_i;\alpha^1)}$$

Using (13), we can show that $\sum_{i=1}^{m_q} \frac{\hat{p}_i \hat{\theta}^1}{\pi^1(d_q|X_i;\alpha^1)} = 1$. Thus $\hat{w}_{i,d_q} = \hat{p}_i \hat{\theta}^1/\pi^1(d_q|X_i;\alpha^1)$ and our estimator becomes

$$\hat{\mu}_{MR}(d_q) = \sum_{i=1}^{n} \frac{I_{d_q}(D_i)\hat{p}_i \hat{\theta}^1 Y_i}{\pi^1(d_q|X_i;\alpha^1)}$$

It is apparent that the weights $\hat{w}_{i,d_q}$ are positive. We also know as before that for $j = 1, \ldots, J$ and $k = 1, \ldots, K$, $\hat{\alpha}^j \to \alpha_*^j$ and $\hat{\beta}^k \to \beta_*^k$ in probability Here, α_*^j and β_*^k are the least false values that minimize the corresponding Kullback-Leibler distance between the probability distribution based on the postulated model and the one that generates the data. Since we assumed without loss of generality that $\pi^1(d_q|X_i;\alpha^1)$ was correctly specified, $\alpha_*^1 = \alpha_0^1$. Furthermore, $\hat{\theta}^j \to \theta_*^j$ and $\hat{\eta}^k \to \eta_*^k$ in probability under suitable regularity conditions, where $\theta_*^j = \mathbb{E}\{\pi^j(d_q|X;\alpha_*^j)\}$ and $\eta_*^k = \mathbb{E}\{a^k(X,d_q;\beta_*^k)\}$.

Han and Wang (2013) show that for their model, $\hat{\lambda} = O_p(n^{-1/2})$. Since we have found $\hat{\lambda}$ in the same way, this holds true for our model as well. Thus, since $\hat{w}_{i,d_q}$ is given by (10), simple algebra gives

$$\hat{\mu}_{MR}(d_q) = \sum_{i=1}^{n} \frac{I_{d_q}(D_i)Y_i}{\pi(d_q|X_i)} + o_p(1)$$

which tends to $\mu(d_q)$ in probability because $\pi(d_q|X) = \mathbb{E}(I_{d_q}(D)|X)$.

Hence we have the following result:

Lemma 5. *When $\mathcal{P}$ contains a correctly specified model for $\pi(d|x)$ and (5) is evaluated at $\hat{\rho} = \hat{\rho}^{\pi}$, $\hat{\mu}_{MR}(d_q)$ is a consistent estimator of $\mu(d_q)$ as $n \to \infty$*

We now consider the case that $\mathcal{A}$ contains a correctly specified OR model, say $a^1(x, d; \beta^1)$ without loss of generality. Let β_0^1 denote the true value of β^1 such that $a^1(x, d; \beta_0^1) = a(x, d)$. To prove the consistency of $\hat{\mu}_{MR}(d_q)$ in this case, suppose that there exists a solution $\hat{\rho}^a$ to (6) which has a finite probability limit ρ_*^a and that (5) is evaluated under $\hat{\rho}^a$. By the definition of $\hat{w}_{i,d_q}$ we have that, after some algebra

$$
\begin{aligned}
\hat{\mu}_{MR}(d_q) &= \sum_{i \in \mathcal{M}_q} \hat{w}_{i,d_q}\{Y_i - a^1(X_i, d_q; \hat{\beta}^1)\} + \frac{1}{n}\sum_{i=1}^{n} a^1(X_i, d_q; \hat{\beta}^1) \\
&= \sum_{i \in \mathcal{M}_q} \frac{\frac{1}{m_q}\frac{1}{1+\hat{\rho}^T\hat{g}_i(\hat{\alpha},\hat{\beta})}}{\frac{1}{m_q}\sum_{i \in \mathcal{M}_q}\frac{1}{1+\hat{\rho}^T\hat{g}_i(\hat{\alpha},\hat{\beta})}}\{Y_i - a^1(X_i, d_q; \hat{\beta}^1)\} + \frac{1}{n}\sum_{i=1}^{n} a^1(X_i, d_q; \hat{\beta}^1) \\
&= \sum_{i=1}^{n} \frac{\frac{I_{d_q}(D_i)}{1+\hat{\rho}^T\hat{g}_i(\hat{\alpha},\hat{\beta})}}{\sum_{i=1}^{n}\frac{I_{d_q}(D_i)}{1+\hat{\rho}^T\hat{g}_i(\hat{\alpha},\hat{\beta})}}\{Y_i - a^1(X_i, d_q; \hat{\beta}^1)\} + \frac{1}{n}\sum_{i=1}^{n} a^1(X_i, d_q; \hat{\beta}^1) \\
&= \left[\sum_{i=1}^{n} \frac{I_{d_q}(D_i)\{Y_i - a^1(X_i, d_q; \hat{\beta}^1)\}}{1+\hat{\rho}^T\hat{g}_i(\hat{\alpha},\hat{\beta})}\right] \bigg/ \left[\sum_{i=1}^{n} \frac{I_{d_q}(D_i)}{1+\hat{\rho}^T\hat{g}_i(\hat{\alpha},\hat{\beta})}\right] + \frac{1}{n}\sum_{i=1}^{n} a^1(X_i, d_q; \hat{\beta}^1)
\end{aligned}
$$

Then, since $a^1(x, d; \beta^1)$ is the correct OR model, under suitable regularity conditions we have that the final line tends in probability to

$$\xrightarrow{p} \mathbb{E}\left[\frac{I_{d_q}(D)\{Y - a(X, d_q)\}}{1 + (\rho_*^a)^T g(\alpha_*, \beta_*; X)}\right] \bigg/ \mathbb{E}\left[\frac{I_{d_q}(D)}{1 + (\rho_*^a)^T g(\alpha_*, \beta_*; X)}\right] + \mu(d_q) \tag{11}$$

where

$$
\begin{aligned}
g(\alpha_*, \beta_*; X) = \{&\hat{\pi}^1(d_q|X; \alpha_*^1) - \theta_*^1, ..., \hat{\pi}^j(d_q|X; \alpha_*^j) - \theta_*^j, \\
&a^1(X, d_q; \beta_*^1) - \eta_*^1, ..., a^K(X, d_q; \beta_*^K) - \eta_*^K\}^T
\end{aligned}
\tag{12}
$$

Recalling that $I_{d_q}(D)Y = Y(d_q)$ and thus

$$I_{d_q}(D)\{Y - a(X, d_q)\} = I_{d_q}(D)\{Y(d_q) - a(X, d_q)\} = 0$$

Hence (11) is equal to $\mu(d_q)$ and so

$$\hat{\mu}_{MR}(d_q) \xrightarrow{p} \mu(d_q)$$

Hence, we have the following result:

Lemma 6. *When $\mathcal{A}$ contains a correctly specified model for $a(x, d)$ and (5) is evaluated at $\hat{\rho} = \hat{\rho}^a$, where $\hat{\rho}^a$ solves (6) and has a finite probability limit, $\hat{\mu}_{MR}(d_q)$ is a consistent estimator of $\mu(d_q)$ as $n \to \infty$.*

Together, Lemmas 5 and 6 prove the result of Theorem 3, and thus we see that our constructed MR estimator does indeed have the MR property.

3.1 Numerical Implementation

An important considerations for numerical implementation is that in general, (6) has multiple roots, and not all of them will lead to MR estimators. The original approach of Han and Wang (2013) can lead to problems of convergence to an incorrect root, and in this case the estimator will not have the desired MR property. Han (2014a) and Han (2014b) solve this problem by utilising an alternative approach. Here, the root is found by solving a convex minimization problem which almost always converges to the correct root. When implementing our estimator in the following simulation study, we use a similar algorithm.

For each treatment level d_q our estimator is constructed as follows. With $\hat{g}_i(\hat{\alpha}, \hat{\beta})$ defined as in (4), we define $F_n(\rho)$ by

$$F_n(\rho) = -\frac{1}{n} \sum_{i \in \mathcal{M}_q} log(1 + \hat{\rho}^T \hat{g}_i(\hat{\alpha}, \hat{\beta}))$$

and let

$$\mathcal{D}_n = \{\rho : 1 + \hat{\rho}^T \hat{g}_i(\hat{\alpha}, \hat{\beta}) > 0, i \in \mathcal{M}_q\}$$

From here we follow the convex minimization algorithm of Han (2014a) and Han (2014b), where in the construction of Δ_1 and Δ_2 the sums are taken over $i \in \mathcal{M}_q$. We now show that the root found using the method of Han (2014b) satisfies the requirements for mutiple robustness in the multivalued setting.

Theorem 7. *The MR estimator has the multiply-robust property for estimating the APOs $\mu(d_q)$ in the multivalued treatment case, when the root of (6) is found using the convex optimization procedure.*

As in Han (2014a), the additional constraint from the original method is that the weights must be non-negative. As before, we index the subjects with $D_i = d_q$ by $1, ..., m_q$ without loss of generality. The estimator for $\hat{\mu}_{MR}(d_q)$ is then $\sum_{i=1}^{m_q} \hat{w}_{i,d_q} Y_i$. For ease of notation, we use w_i in place of $\hat{w}_{i,d_q}$ in what follows. The weights w_i are found by the optimization problem:

$$\max_{w_1, ..., w_{m_q}} \prod_{i=1}^{m_q} w_i \quad \text{subject to} \quad w_i \geq 0, \ (i = 1, ..., m_q), \quad \sum_{i=1}^{m_q} w_i = 1$$

$$\sum_{i=1}^{m_q} w_i \{\pi^j(d_q | X_i; \alpha^j) - \hat{\theta}^j\} = 0$$

$$\sum_{i=1}^{m_q} w_i \{a^k(X_i, d_q; \beta^k) - \hat{\eta}^k\} = 0 \tag{13}$$

8

Using the method of Lagrange multipliers gives the same weights as in (5), where $\hat{\rho}$ still solves (6). The additional constraint from the non-negativity of the w_i gives the following constraint on $\hat{\rho}$:

$$1 + \hat{\rho}^T \hat{g}_i(\hat{\alpha}, \hat{\beta}) > 0, i = 1, ..., m_q. \tag{14}$$

Having defined $F_n(\rho)$ and $\mathcal{D}_n$ as above, we have from Lemma 4 that $\mathbf{0}$ is inside the convex hull of $\{\hat{g}_i(\hat{\alpha}, \hat{\beta}), i = 1, ..., m_q\}$ when m_q is sufficiently large. Thus $\mathcal{D}_n$ is a bounded set, with the proof following as in Han (2014a). It is also easy to verify that it is an open convex set, and so is an open bounded convex polytope with its boundary consisting of some or all of the m_q hyperplanes $\{\rho : 1 + \hat{\rho}^T \hat{g}_i(\hat{\alpha}, \hat{\beta}) = 0\}$ $(i = 1, ..., m_q)$. Furthermore, $F_n(\rho)$ is a strictly convex function on $\mathcal{D}_n$, due to the fact that $-\log(.)$ is a strictly convex function.

We define

$$\hat{\rho}^{min} = arg \min_{\rho \in \mathcal{D}_n} F_n(\rho).$$

As in the original proof, the properties of $F_n(\rho)$ and $\mathcal{D}_n$ imply that $\hat{\rho}^{min}$ exists and is a root of (6). Therefore, the new estimator is formed by taking $\hat{\rho} = \hat{\rho}^{min}$ in the calculation of the weights.

As n increases, $F_n(\rho)$ converges to $F(\rho) = \mathbb{E}[I_{d_q}(D)\{1 + \rho^T g(\alpha_*, \beta_*)\}]$ in probability pointwise on $\mathcal{D}$ under suitable regularity conditions. Here, $g(\alpha_*, \beta_*)$ is as in the proof of Theorem 3, and

$$\mathcal{D} = \{\rho : 1 + \rho^T g_x(\alpha_*, \beta_*) > 0, \text{for any x that is a realization of } X \mid D = d_q\}$$

where $g_x(\alpha_*, \beta_*)$ is $g(\alpha_*, \beta_*)$ with $X = x$. Under some reasonable assumptions of the boundedness of $a^k(x, d_q; \beta_*^k)$ and $\hat{\pi}^j(d_q|x; \alpha_*^j)$, and the finiteness of η_*^k and θ_*^j, $\mathcal{D}$ is an open convex set. Furthermore, Lemma 4 implies that $\mathbb{E}[\frac{g(\alpha_*, \beta_*)}{\pi(d_q|X)}|D = d_q] = 0$. Thus $\sup_{x \sim X|D=d_q} \{-\tilde{\rho}^T g_x(\alpha_*, \beta_*)\} > 0$ for any $\tilde{\rho}$ with $||\tilde{\rho}|| = 1$, where $x \sim X \mid D = d_q$ means that x is a possible realization of $X \mid D = d_q$. Since $g(\alpha_*, \beta_*)$ is also bounded by assumption, we see that $\mathcal{D}$ is a bounded set, with the proof following as in Han (2014a). $F(\rho)$ is a strictly convex function because $-\log(.)$ is, and therefore when it has a minimum in $\mathcal{D}$, the minimizer is unique and the global minimizer.

From this we can see why $\hat{\rho}^{min}$ is the correct root for multiple-robustness. When $\pi(d|x)$ is correctly modelled, by $\pi^1(d|x; \alpha^1)$ without loss of generality, we have that

$$1 + (\hat{\rho}^\pi)^T \hat{g}_i(\hat{\alpha}, \hat{\beta}) = \frac{\pi^1(d_q|X_i; \hat{\alpha}^1)}{\hat{\theta}^1} + o_p(1) \quad (i = 1, ..., m_q).$$

Here, $\hat{\rho}^\pi$ is as defined in Lemma 5, and we use the fact that $\hat{\lambda} = O_p(n^{-1/2})$ as shown in the proof of the same lemma, where $\hat{\lambda}$ solves (9). Therefore $\hat{\rho}^\pi \in \mathcal{D}_n$ with probability approaching 1. Since $\hat{\rho}^\pi$ solves (6), $\frac{\partial F_n(\hat{\rho}^\pi)}{\partial \rho} = 0$. $F_n(\rho)$ is strictly convex, and thus $\hat{\rho}^\pi$ is the unique minimizer on $\mathcal{D}_n$ and so $\hat{\rho}^{min} = \hat{\rho}^\pi$ with probability approaching 1. The proof of Lemma 5 now follows through using $\hat{\rho}^{min}$ in place of $\hat{\rho}^\pi$.

If instead $a(x, d)$ is correctly specified, $\hat{\rho}^{min}$ is still uniquely identified. If we denote the minimizer of $F(\rho)$ in $\mathcal{D}$ by ρ_*, then ρ_* is finite due to the boundedness of $\mathcal{D}$. Furthermore, $\hat{\rho}^{min} \xrightarrow{p} \rho_*$, with the proof following through as in Han (2014a). The proof of Lemma 6 only

requires that the root of (6) used has a finite probability limit. Thus the proof follows through using $\hat{\rho}^{min}$ in place of $\hat{\rho}^a$. Finally, we see that calculating the weights using $\rho = \hat{\rho}^{min}$ leads to a multiply-robust estimator. Theorem 7 holds because the proofs of lemmas 5 and 6 follow with the new root $\hat{\rho}^{min}$ found using the convex optimization procedure.

4   Simulations

In this section we present simulations to demonstrate the robustness of our proposed estimator to problems of confounding and functional form misspecification, which are routinely encountered in causal analyses.

A multivalued treatment D with 4 levels is assigned as a function of covariates X. The outcome Y is quadratic in D with confounding from X. This simulation study is adapted from the one performed by Han and Wang (2013).

$$X \sim U[-2.5, 2.5]$$
$$D|X \sim Bin(3, \pi(X))$$
$$Y|X, D \sim \mathcal{N}(a(X, D), \sigma_Y^2 = 2)$$

where $logit\{\pi(x)\} = -0.5 + 0.1x - 0.2x^2$ and $a(x, d) = 1 + 2d - 0.35d^2 + 2x + 3x^2$

We propose two OR models, one correctly specified (a^1) and one incorrectly specified (a^2):

$$a^1(x, d; \beta^1) = \beta_0^1 + \beta_1^1 d + \beta_2^1 d^2 + \beta_3^1 x + \beta_4^1 x^2$$
$$a^2(x, d; \beta^2) = \beta_0^2 + \beta_1^2 d + \beta_2^2 x$$

We also propose a correctly specified (π^1) and incorrectly specified (π^2) model for $\pi(x)$

$$\pi^1(x; \alpha^1) = (\exp\{\alpha_0^1 + \alpha_1^1 x + \alpha_2^1 x^2\} + 1)^{-1}$$
$$\pi^2(x; \alpha^2) = 1 - \exp(-\exp\{\alpha_0^2 + \alpha_1^2 x + \alpha_2^2 e^x\})$$

The parameter estimates $\hat{\beta}^k$ are found as the regression coefficients of a generalized linear model for $a(x, d)$, and we can find the estimates $\hat{\alpha}^j$ as the maximizer of the binomial likelihood

$$\prod_{i=1}^n \{\pi^j(X_i; \alpha^j)\}^{D_i} \{1 - \pi^j(X_i; \alpha^j)\}^{3-D_i}$$

The PS model is then

$$\hat{\pi}(d_q|X) = \binom{N}{d_q} \hat{\pi}(X)^j (1 - \hat{\pi}(X))^{N-d_q}$$

where $N = 3$, $d_q = 0, ..., 3$ and $\hat{\pi}(X)$ is estimated by one of our two postulated models for $\pi(x)$.

We test the following models:

1. Four DR models - $\hat{\mu}_{DR\_1010}(d_q)$, $\hat{\mu}_{DRP\_1001}(d_q)$, $\hat{\mu}_{DR\_0110}(d_q)$ and $\hat{\mu}_{DR\_0101}(d_q)$. Here we combine the OR models with the PS models using the inverse weighting approach. With

10

two OR models and two PS models there are four possible combinations. The notation we use is as follows. The first two digits denote which PS model we are using, and the second two digits denote which OR model. For example $\hat{\mu}_{DR\_1001}(d_q)$ denotes the estimator using PS model π^1 and OR model a^2.

2. Five MR models - $\hat{\mu}_{MR\_1101}(d_q)$, $\hat{\mu}_{MR\_1110}(d_q)$, $\hat{\mu}_{MR\_1011}(d_q)$, $\hat{\mu}_{MR\_0111}(d_q)$ and $\hat{\mu}_{MR\_1111}(d_q)$. Here we combine the OR models with the PS models using the multiply-robust approach introduced in the previous chapter. For example $\hat{\mu}_{MR\_1101}(d_q)$ denotes the estimator using PS models π^1 and π^2 and OR model a^2.

In each case, point estimates of the APOs for each level of the treatment are presented, as well as empirical variance and bias estimates. In each case these are obtained by calculating the estimate 1000 times on simulated data sets of size 10000.

Table 4 shows the results for the simulations. We first note that for the DR models, when either the PS or the OR model is correctly specified, the estimators have small biases. However, when neither estimator is correctly specified the bias is several orders of magnitude larger. We also note that the variances for the DR models are small.

We see that the MR estimator is working well compared to the DR estimators, with similarly small biases and with variances similar in size to the DR estimates for either a correctly specified PS or OR model. The one case where the variance appears unusually high is in the estimate of the first treatment level with the MR_0111 estimator which does not include the correctly specified PS model. However, this is still of a comparable size to the variances of the other estimator, and the bias is still very small.

We conclude that the MR estimator is performing as expected, giving an accurate estimate of the APOs when either the family of PS or OR models contains a correctly specified model. Importantly, because we specify two PS and two OR models, we ensure against the case in DR estimation when both models are specified incorrectly. Thus our results confirm that, when there is significant uncertainty about the correct specification of the OR and PS model, it it of value to use a MR estimator. For example, in the case examined in the simulation where we have two OR and two PS models but do not know which is correctly specified, there is a chance of choosing a DR estimator that performs poorly (DR_0101 in this case). However, in the MR setting, we can use an estimator that includes all four models, and the resulting estimator performs well.

5   Conclusions

In this paper we have built on the work of Han and Wang (2013), Han (2014a) and Han (2014b) to formulate a multiply robust estimator for causal dose-response estimation. The model is multiply robust in the sense that consistent estimates of average potential outcomes can be obtained under misspecification of all but one of our multiple outcome regression or generalised propensity score models. We have shown that the MR approach can provide a good approximation to linear or nonlinear dose-response functions and is robust to problems of confounding or functional form misspecification.

Table 1: Mean Average Potential Outcomes for DR and MR Estimators

		Treatment level			
		0	1	2	3
Truth		7.253	8.903	9.853	10.103
DR_1010	Av Est	7.255	8.902	9.853	10.103
	Emp Var	0.008	0.006	0.005	0.008
	Bias	0.002	-0.001	0.000	0.000
DR_1001	Av Est	7.238	8.903	9.850	10.109
	Emp Var	0.073	0.015	0.012	0.071
	Bias	-0.015	0.001	-0.002	0.006
DR_0110	Av Est	7.253	8.901	9.853	10.103
	Emp Var	0.008	0.006	0.005	0.008
	Bias	0.000	-0.001	0.000	0.000
DR_0101	Av Est	7.010	8.784	9.969	10.479
	Emp Var	0.071	0.014	0.011	0.061
	Bias	-0.243	-0.118	0.117	0.377
MR_1101	Av Est	7.251	8.899	9.852	10.098
	Emp Var	0.009	0.006	0.005	0.009
	Bias	-0.002	-0.004	-0.001	-0.005
MR_1110	Av Est	7.247	8.899	9.848	10.099
	Emp Var	0.008	0.005	0.005	0.009
	Bias	-0.006	-0.004	-0.005	-0.006
MR_1011	Av Est	7.250	8.905	9.854	10.105
	Emp Var	0.008	0.006	0.005	0.009
	Bias	-0.003	0.002	0.002	0.002
MR_0111	Av Est	7.249	8.899	9.851	10.098
	Emp Var	0.008	0.005	0.005	0.008
	Bias	-0.004	-0.003	-0.002	-0.005
MR_1111	Av Est	7.248	8.899	9.850	10.096
	Emp Var	0.008	0.005	0.005	0.009
	Bias	-0.005	-0.004	-0.003	-0.007

Acknowledgements

Appendix

Proof of doubly-robust properties for multivalued treatments

Proof. By the WLLN, we have that

$$\hat{\mu}_{DR}(d_q) \xrightarrow{p} \mathbb{E}\left[\frac{I_{d_q}(D)Y}{\pi(d_q|X;\alpha)} - \frac{I_{d_q}(D) - \pi(d_q|X;\alpha)}{\pi(d_q|X;\alpha)}\Psi^{-1}\{m(X, d_q; \beta)\}\right]$$

Where $\pi(d_q|X;\alpha)$ is the postulated PS model for $\pi(d_q|X)$, $\Psi^{-1}\{m(X,d_q;\beta)\}$ is the postulated OR model and α and β are the true parameters of the PS and OR models respectively.

Using the SUTVA this quantity is equal to

$$\mathbb{E}\left[\frac{I_{d_q}(D)Y(d_q)}{\pi(d_q|X;\alpha)} - \frac{I_{d_q}(D) - \pi(d_q|X;\alpha)}{\pi(d_q|X;\alpha)}\Psi^{-1}\{m(X,d_q;\beta)\}.\right]$$

Rearranging, gives

$$\mathbb{E}\left[Y(d_q)\right] + \mathbb{E}\left[\frac{I_{d_q}(D) - \pi(d_q|X;\alpha)}{\pi(d_q|X;\alpha)}\left\{Y(d_q) - \Psi^{-1}\{m(X,d_q;\beta)\}\right\}\right] \tag{15}$$

The first term is our estimand, so all that remains is to show that the second term is 0 if either the PS or OR model is correctly specified. Let us first consider the case where the OR model is correctly specified, so that

$$\Psi^{-1}\{m(X,d_q;\beta)\} = \mathbb{E}(Y|D = d_q, X)$$

Then, using the above equality and the law of iterated expectation, the second term of (15) becomes

$$\mathbb{E}\left(\mathbb{E}\left[\frac{I_{d_q}(D) - \pi(d_q|X;\alpha)}{\pi(d_q|X;\alpha)}\left\{Y(d_q) - \mathbb{E}(Y|D = d_q, X)\right\}\Big| I_{d_q}(D), X\right]\right)$$

$$=\mathbb{E}\left(\frac{I_{d_q}(D) - \pi(d_q|X;\alpha)}{\pi(d_q|X;\alpha)}\mathbb{E}\left[\left\{Y(d_q) - \mathbb{E}(Y|D = d_q, X)\right\}\Big| I_{d_q}(D), X\right]\right)$$

$$=\mathbb{E}\left(\frac{I_{d_q}(D) - \pi(d_q|X;\alpha)}{\pi(d_q|X;\alpha)}\left\{\mathbb{E}[Y(d_q)|I_{d_q}(D), X] - \mathbb{E}(Y|D = d_q, X)\right\}\right)$$

$$=\mathbb{E}\left(\frac{I_{d_q}(D) - \pi(d_q|X;\alpha)}{\pi(d_q|X;\alpha)}\left\{\mathbb{E}[Y(d_q)|X] - \mathbb{E}(Y(d_q)|X)\right\}\right) = 0$$

where the last line follows from the weak conditional independence condition implied by the no unmeasured confounders assumption, i.e.

$$\mathbb{E}(Y|D = d_q, X) = \mathbb{E}(Y(d_q)|D = d_q, X) = \mathbb{E}(Y(d_q)|X) = \mathbb{E}(Y(d_q)|I_{d_q}(D), X)$$

Conversely, if the PS model is correctly specified then

$$\pi(d_q|X;\alpha) = \pi(d_q|X) = \mathbb{P}(D = d_q|X) = \mathbb{E}(I_{d_q}(D)|X)$$

Thus, using the above equality and the law of iterated expectation, the second term of (15) becomes

$$\mathbb{E}\left(\mathbb{E}\left[\frac{I_{d_q}(D) - \pi(d_q|X)}{\pi(d_q|X)}\left\{Y(d_q) - \Psi^{-1}\{m(X,d_q;\beta)\}\right\}\Big| Y(d_q), X\right]\right)$$

$$=\mathbb{E}\left(\left\{Y(d_q) - \Psi^{-1}\{m(X,d_q;\beta)\}\right\}\mathbb{E}\left[\frac{I_{d_q}(D) - \pi(d_q|X)}{\pi(d_q|X)}\Big| Y(d_q), X\right]\right)$$

$$=\mathbb{E}\left(\left\{Y(d_q) - \Psi^{-1}\{m(X,d_q;\beta)\}\right\}\frac{\mathbb{E}[I_{d_q}(D)|Y(d_q), X] - \pi(d_q|X)}{\pi(d_q|X)}\right)$$

$$=\mathbb{E}\left(\left\{Y(d_q) - \Psi^{-1}\{m(X,d_q;\beta)\}\right\}\frac{\mathbb{E}[I_{d_q}(D)|X] - \pi(d_q|X)}{\pi(d_q|X)}\right)$$

$$=\mathbb{E}\left(\left\{Y(d_q) - \Psi^{-1}\{m(X,d_q;\beta)\}\right\}\frac{\pi(d_q|X) - \pi(d_q|X)}{\pi(d_q|X)}\right) = 0$$

where the second last line follows from the no unmeasured confounders assumption as before.

Thus we see that $\hat{\mu}_{DR}(d_q)$ consistently estimates the APO $\mu(d_q)$ when either the OR model or PS model is correctly specified, in the case of multivalued treatment. □

References

Bang, H. and J. M. Robins (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics 61*, 962–972.

Chan, K. C. G. (2013). A simple multiply robust estimator for missing response problem. *Stat 2*(1), 143–149.

Chan, K. C. G. and S. C. P. Yam (2014). Oracle, multiple robust and multipurpose calibration in a missing response problem. *Statistical Science 29*(3), 380–396.

Flores, C. A. and O. A. Mitnik (2009). Evaluating nonexperimental estimators for multiple treatments: Evidence from experimental data. IZA Discussion Papers 4451, Institute for the Study of Labor (IZA).

Graham, D. J., E. J. McCoy, and D. A. Stephens (2016). Approximate bayesian inference for doubly robust estimation. *Bayesian Anal. 11*(1), 47–69.

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*(507), 1159–1173.

Han, P. and L. Wang (2013). Estimation with missing data: beyond double robustness. *Biometrika 100*(2), 417–430.

Hirano, K. and G. W. Imbens (2004). The propensity score with continuous treatments. In A. Gelman and X. L. Meng (Eds.), *Applied Bayesian modeling and causal inference from incomplete data perspectives*, pp. 73–84. New York: Wiley.

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

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika 87*(3), 706–710.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* New York, NY, USA: Cambridge University Press.

Kang, J. D. Y. and J. L. Schafer (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science 22*(4), 523–539.

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

Qin, J. and B. Zhang (2007). Empirical-likelihood-based inference in missing response problems and its application in observational studies. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(1), 101–122.

Robins, J. M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association, Section on Bayesian Statistical Science*, pp. 6–10. Alexandria, VA: American Statistical Association.

Robins, J. M. and A. Rotnitzky (2001). Comment on "Inference for semiparametric models: some questions and an answer". *Statistical sinica 11*, 920–936.

Robins, J. M., A. Rotnitzky, and M. J. van der Laan (2000). Comment on the Murphy and Van der Vaart article "On profile likelihood". *Journal of the American Statistical Association 95*, 431–435.

Rubin, D. B. (1978). Bayesian inference for causal effects: the role of randomization. *Annals of Statistics 6*(1), 34–58.

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins (1999). Adjusting for nonignorable dropout using semiparametric nonresponse models. *Journal of the American Statistical Association 94*(448), 1096–1120. (with rejoinder 1135-1146).

van der Laan, M. and J. M. Robins (2003). *Unified methods for censored longitudinal data and causality*. Berlin: Springer.

Zhang, Z., J. Zhou, W. Cao, and J. Zhang (2016). Causal inference with a quantitative exposure. *Statistical Methods in Medical Research 25*(1), 315–335.