A Bayesian view of doubly robust causal inference

O. Saarela[*], L. R. Belzile[†] AND D. A. Stephens[‡]

Abstract

In causal inference the effect of confounding may be controlled using regression adjustment in an outcome model, propensity score adjustment, inverse probability of treatment weighting or a combination of these. Approaches based on modelling the treatment assignment mechanism, along with their doubly robust extensions, have been difficult to motivate using formal likelihood-based or Bayesian arguments, as the treatment assignment model plays no part in inferences concerning the expected outcomes. On the other hand, forcing dependency between the outcome and treatment assignment models by allowing the former to be misspecified results in loss of the balancing property of the propensity scores and the loss of any double robustness. In this paper, we explain in the framework of misspecified models why doubly robust inferences cannot arise from purely likelihood-based arguments. As an alternative to Bayesian propensity score analysis, we propose a Bayesian posterior predictive approach for constructing doubly robust estimation procedures by incorporating the inverse treatment assignment probabilities as importance sampling weights in Monte Carlo integration.

1. INTRODUCTION

In causal inference contexts, adjustment for confounding is most often carried out using one of two approaches: first, by conditioning on the confounders in a regression model for the outcome in an outcome regression model; secondly, by modelling the treatment assignment mechanism to obtain so-called propensity score values, and then using these scores to construct strata within the observed sample, or a pseudo-sample from a hypothetical population within which the treatment-outcome relationship is not confounded. This pseudo-sample can be obtained via inverse probability of treatment weighting of the original sample, analogously to survey sampling procedures. The outcome regression adjustment method requires correct specification of the regression function in order to obtain consistent inference; this may be achieved in practice using flexible regression strategies, which may involve complex functions of typically a large number of covariates. The propensity score adjustment methods focus principally on the specification of the treatment assignment model, which may be similarly flexible or complex. It is tempting, given that the validity of each model in turn depends on unverifiable assumptions, to use doubly robust estimators. The latter combine outcome regression adjustment and propensity score adjustment so that confounding is removed if either of the models is correctly specified.

Adjustments that depend on the propensity score using regression or reweighting are not easy to interpret from the Bayesian perspective, since Bayesian inferences are naturally based on modelling of the outcome, with modelling of the treatment assignment playing no role in inference relating to the outcome/treatment relationship (Robins and Ritov, 1997). Nevertheless, there has been work studying Bayesian versions of propensity score adjustment to control for confounding (Hoshino, 2008; McCandless et al., 2009a,b, 2010, 2012; An, 2010; Kaplan and Chen, 2012; Zigler et al., 2013; Chen and Kaplan, 2015; Graham et al., 2016). These approaches require one of two kinds of compromises: a parametrization that makes the outcome and treatment assignment models dependent or else cutting feedback between the exposure and outcome models. In the first case the balancing property of the propensity score

[*] Dalla Lana School of Public Health, University of Toronto, 155 College Street, Toronto, Ontario M5T 3M7, Canada. `olli.saarela@utoronto.ca`

[†] École Polytechnique Fédérale de Lausanne, EPFL-SB-MATHAA-STAT, Station 8, CH-1015 Lausanne, Switzerland. `leo.belzile@epfl.ch`

[‡] Department of Mathematics and Statistics, McGill University, Montreal, Quebec H3A 2K6, Canada. `d.stephens@math.mcgill.ca`

is lost, while the second results in inferences that are no longer Bayesian. In an alternative approach, Wang et al. (2012) and Zigler and Dominici (2014) have suggested connecting the outcome and treatment assignment models through the prior distribution in order to incorporate the uncertainty in confounder selection. Gustafson (2012) suggested a Bayesian interpretation as a compromise between a saturated outcome model and a parametric one; however, the treatment assignment model did not feature in this interpretation. Herein we do not consider model selection or uncertainty, but rather concentrate on inferences with a priori specified outcome and treatment assignment models.

The purpose of this paper is twofold. We aim to clarify the theoretical and practical motivations for Bayesian propensity score adjustment and the relationships between the different methods proposed for this, which have not been fully explored previously. We also propose an alternative approach to Bayesian propensity score adjustment. We address these issues in the context of double adjustment for both the potential confounders and the propensity score, and argue that the problem cannot be properly understood without considering it in the framework of misspecified models. It is important to distinguish between misspecification due to omission of relevant covariates from the outcome model, and misspecification of the functional form of the dependency between the covariates and the outcome. The frequentist propensity score adjusted outcome regression is robust against the former type of model misspecification, but this property is lost in Bayesian estimation if the misspecified outcome model is allowed to feed back to the estimation of the propensity scores. While the feedback issue has been documented in the literature (e.g., McCandless et al., 2009b; Zigler et al., 2013), and the reasons behind this were stated by Robins and Ritov (1997), here we point out that cutting this feedback in a two-step Bayesian estimation procedure unnecessarily inflates the posterior variance estimates. To provide a different approach, we propose deriving Bayesian versions of various inverse probability of treatment weighted estimators, including inverse probability of treatment weighted outcome regression and the semi-parametric double robust estimator, through posterior predictive expectations, with the weights introduced as importance sampling weights in Monte Carlo integration. All proofs are relegated to the Supplementary Material.

2. PRELIMINARIES

2. *Notation and assumptions*

We consider a single binary treatment, since this enables comparisons between different propensity-score adjusted and inverse-probability weighted estimation approaches. Let the random vectors X_i represent a vector of pre-treatment covariate measurements, Z_i a binary treatment allocation indicator, and Y_i an outcome for individual i, measured after enough time has passed since administering the treatment. We adopt for convenience the standard potential outcome framework, using counterfactuals: for individual i, the observed outcome is related to the two possible potential outcomes (Y_{0i}, Y_{1i}) by $Y_i = (1 - Z_i)Y_{0i} + Z_i Y_{1i}$. We assume ignorable treatment assignment (cf., Rosenbaum and Rubin, 1983, p. 43), meaning that X_i includes a set of covariates adequate to control for confounding in the sense that assignment to treatment is conditionally independent of the potential outcome, denoted $Z_i \perp\!\!\!\perp (Y_{0i}, Y_{1i}) \mid X_i$. The propensity score $e(X_i) \equiv \mathrm{pr}(Z_i = 1 \mid X_i)$ has the balancing property $Z_i \perp\!\!\!\perp X_i \mid e(X_i)$, which also implies that $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp Z_i \mid e(X_i)$. This scalar score is therefore useful in controlling for confounding. If the covariate space X_i is high-dimensional, the task of controlling for confounding often involves some covariate selection. One can either select for the features that predict the outcome or for those that predict the treatment assignment. To represent this, let S_i and B_i denote a priori selected subsets of the all observed features X_i, so that the latter can be partitioned as $X_i = (S_i, R_i)$ or $X_i = (B_i, C_i)$, where possibly $R_i = \emptyset$ or $C_i = \emptyset$. If the selected set of features S_i captures all relevant prognostic information, then $Y_{0i} \perp\!\!\!\perp X_i \mid S_i$ (Hansen, 2008). For the remainder of the paper, we consider the stronger condition (i) $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp X_i \mid S_i$, which requires that S_i also captures all relevant information about possible effect modification. In this case, conditioning on S_i adequately controls for confounding. If, on the other hand, the selected set of features B_i has the balancing property (ii) $Z_i \perp\!\!\!\perp X_i \mid B_i$, it is adequate to control for confounding. We are interested in estimation procedures which are valid when either $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp X_i \mid S_i$ or $Z_i \perp\!\!\!\perp X_i \mid B_i$.

We are interested in an average causal contrast such as the risk difference $\mathrm{E}(Y_{1i}) - \mathrm{E}(Y_{0i})$. Assuming there are no unmeasured confounders, it follows that (e.g., Hernán and Robins, 2006, p. 43)

$$\mathrm{E}(Y_{1i}) - \mathrm{E}(Y_{0i}) = \int_{x_i} \{\mathrm{E}(Y_i \mid Z_i = 1, x_i) - \mathrm{E}(Y_i \mid Z_i = 0, x_i)\}\, p(x_i)\, \mathrm{d}x_i.$$

The natural frequentist estimator of this quantity obtained by replacing $p(x)$ by its empirical counterpart is consistent provided the conditional expectation $\mathrm{E}(Y_i \mid Z_i = z, x_i)$ is correctly specified. We address the situation in which correct specification cannot be guaranteed.

2. *Bayesian model formulation for outcome regression*

The de Finetti representation for the joint distribution of exchangeable random variables $v_i \equiv (x_i, y_i, z_i)$ $(i = 1, 2, \ldots)$ may be written

$$p(v) = \int_{\phi,\gamma,\psi} \prod_{i=1}^{n} \{p(y_i \mid z_i, x_i; \phi) p(z_i \mid x_i; \gamma) p(x_i; \psi)\} \, \pi_0(\phi, \gamma, \psi) \, \mathrm{d}\phi \, \mathrm{d}\gamma \, \mathrm{d}\psi, \tag{1}$$

where $v \equiv (v_1, \ldots, v_n)$, implying the existence of the parametric sampling models and the prior density $\pi_0(\phi, \gamma, \psi)$. Since the representation theorem is not constructive, and does not specify the models implicit in (1), inferences about a given finite-dimensional parametrization involves the assumption of correctly specified models, that is $p(y_i \mid z_i, x_i; \phi_0) = f(y_i \mid z_i, x_i)$, $p(z_i \mid x_i; \gamma_0) = f(z_i \mid x_i)$ and $p(x_i; \psi_0) = f(x_i)$, where $(\phi_0, \gamma_0, \psi_0)$ is the limiting value of the posterior $\pi_n(\phi, \gamma, \psi) \equiv p(\phi, \gamma, \psi \mid v)$ in the sense of van der Vaart (1998, p. 139), and where the fs represent the true sampling distributions. We might further assume that the parameters are a priori independent, so that $\pi_0(\phi, \gamma, \psi) = \pi_0(\phi)\pi_0(\gamma)\pi_0(\psi)$. In this case, the posterior density factorizes, meaning $\pi_n(\phi, \gamma, \psi) = \pi_n(\phi)\pi_n(\gamma)\pi_n(\psi)$ (e.g., Gelman et al., 2004, p. 354–355).

In practice, the marginal covariate distribution $p(x_i; \psi)$ can be specified nonparametrically as in the Bayesian bootstrap (Rubin, 1981), leading to a Bayesian estimator of the average causal contrast,

$$\int_{\phi} \int_{\psi} \sum_{i=1}^{n} \psi_i \{m(1, x_i; \phi) - m(0, x_i; \phi)\} \, \pi_n(\phi) \pi_n(\psi) \, \mathrm{d}\phi \, \mathrm{d}\psi. \tag{2}$$

Here

$$\pi_n(\phi) \propto \prod_{i=1}^{n} p(y_i \mid z_i, x_i; \phi) \pi_0(\phi) \, \mathrm{d}\phi, \quad m(z, x; \phi) \equiv \textstyle\int y p(y \mid z, x; \phi) \, \mathrm{d}y,$$

and $\psi \equiv (\psi_1, \ldots, \psi_n)$, with $\pi_n(\psi)$ taken to be the uniform Dirichlet distribution.

Estimator (2) is the Bayesian version of the direct standardization or g-formula, and can be motivated without the use of potential outcomes notation through posterior predictive expectations for a new observation under a hypothetical completely randomized setting; see the Supplementary Material. Estimator (2) does not involve the treatment assignment model; rather, the posterior predictive approach for estimating the marginal causal contrast depends entirely on correct specification of the distribution $p(y_i \mid z_i, x_i; \phi)$. We discuss Bayesian alternatives to (2) in §6.

3. FREQUENTIST APPROACHES

3. *Inclusion of propensity scores into outcome regression*

We briefly review some common frequentist approaches for combining outcome regression and propensity score adjustment. Because of the balancing property, it is tempting to specify a propensity score $e(b_i; \gamma) \equiv \mathrm{pr}(Z_i = 1 \mid b_i; \gamma)$ and use a statistical model such as $p\{y_i \mid z_i, e(b_i), s_i; \phi\}$, in the hope that, if the prognostic model is misspecified, adjusting for the propensity score would still adequately control for any residual confounding. For simplicity, we take the parameters ϕ to specify also the functional dependence between the propensity score and the outcome; to model this dependency, it is advisable to use flexible formulations such as splines (e.g., Zhang and Little, 2009). Using such an outcome model, the marginal causal contrast would then be estimated by

$$\frac{1}{n} \sum_{i=1}^{n} \left[m\left\{1, e(b_i; \widehat{\gamma}), s_i; \widehat{\phi}\right\} - m\left\{0, e(b_i; \widehat{\gamma}), s_i; \widehat{\phi}\right\} \right], \tag{3}$$

where $m\{z, e(b; \gamma), s; \phi\} \equiv \int y p\{y \mid z, e(b; \gamma), s; \phi\} \, \mathrm{d}y$, and where $\widehat{\phi}$ and $\widehat{\gamma}$ are the maximum likelihood estimators for the parameters in the outcome regression and treatment assignment model, respectively. The motivation for such a double adjustment is that it is adequate to control for confounding if either $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp X_i \mid S_i$ or $Z_i \perp\!\!\!\perp X_i \mid B_i$ applies. We summarize this property in the following theorem.

THEOREM 1. *Suppose that the outcome model is correctly specified in the sense that* $m\{z, e(b; \gamma_0), s; \phi_0\} = \int y f(y \mid z, e(b), s) \, dy$, *and that the outcome model parameters can be consistently estimated so that*

$\widehat{\phi} \to \phi_0$ in probability as $n \to \infty$. Suppose further that the treatment assignment model is correctly specified in the sense that $p(z_i \mid b_i; \gamma_0) = f(z_i \mid b_i)$, and that the treatment assignment model parameters can be consistently estimated so that $\widehat{\gamma} \to \gamma_0$ in probability as $n \to \infty$. Then, if either $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp X_i \mid S_i$ or $Z_i \perp\!\!\!\perp X_i \mid B_i$, (3) is a consistent estimator of $\mathrm{E}(Y_{1i}) - \mathrm{E}(Y_{0i})$.

The estimator (3) may be considered doubly robust in terms of the covariate selection in the sense that only one of the sets S_i and B_i needs to be correctly specified. It still relies however on a correct parametric specification of the model for the expected outcome conditional on $\{Z_i, e(B_i), S_i\}$. This should be contrasted to the semi-parametric double robustness property, discussed next in § 3.2. The latter does not require correct parametric specification of the outcome model if the treatment assignment model is correctly specified.

3. *The clever covariate and augmented outcome regression*

The estimator discussed in §3.1 did not specify which function of the propensity score should be added to the regression model. Scharfstein et al. (1999, p. 1141–1142) and Bang and Robins (2005, p. 964–965) drew a connection between propensity score regression adjustment and doubly robust estimator

$$\frac{1}{n}\sum_{i=1}^{n} \frac{y_i z_i - \{z_i - e(b_i;\widehat{\gamma})\} m(1, s_i; \widehat{\phi})}{e(b_i;\widehat{\gamma})} - \frac{1}{n}\sum_{i=1}^{n} \frac{y_i(1-z_i) - [(1-z_i) - \{1 - e(b_i;\widehat{\gamma})\}] m(0, s_i; \widehat{\phi})}{1 - e(b_i;\widehat{\gamma})},$$

which can be equivalently represented as

$$\frac{1}{n}\sum_{i=1}^{n} \left\{ \frac{z_i}{e(b_i;\widehat{\gamma})} - \frac{1-z_i}{1-e(b_i;\widehat{\gamma})} \right\} \left\{ y_i - m(z_i, s_i; \widehat{\phi}) \right\} + \frac{1}{n}\sum_{i=1}^{n} \left\{ m(1, s_i; \widehat{\phi}) - m(0, s_i; \widehat{\phi}) \right\}. \quad (4)$$

On considering the score equation derived from a regression of Y_i on Z_i and S_i with mean function $m(z, s; \phi)$, this form suggests incorporating the derived covariate $c(z_i, b_i) = z_i/e(b_i) - (1 - z_i)/\{1 - e(b_i)\}$, termed the clever covariate by Rose and van der Laan (2008, p. 8), additively into the outcome regression, that is, for example

$$m(z, s; \phi) = \phi_0 + \phi_1 z + \phi_2^{\top} s + \phi_3 c(z, b). \quad (5)$$

The first term in (4) is then zero through the maximum likelihood score equation, leaving only the last term, which is the model-based estimator of the marginal treatment effect. Thus with the clever covariate in the outcome model, the doubly robust estimator is equivalent to the model-based estimator. In the special case of model (5), this becomes

$$\frac{1}{n}\sum_{i=1}^{n} \left\{ m(1, x_i; \widehat{\phi}) - m(0, x_i; \widehat{\phi}) \right\} = \widehat{\phi}_1 + \widehat{\phi}_3 \frac{1}{n}\sum_{i=1}^{n} \left\{ \frac{1}{e(b_i;\widehat{\gamma})} - \frac{1}{1-e(b_i;\widehat{\gamma})} \right\}.$$

A potential drawback of using this covariate is that it may lead to extreme variability for the resulting mean difference estimator, even compared to inverse probability of treatment weighted estimators of the form

$$\frac{1}{n}\sum_{i=1}^{n} \left\{ \frac{y_i z_i}{e(b_i;\widehat{\gamma})} - \frac{y_i(1-z_i)}{1-e(b_i;\widehat{\gamma})} \right\}. \quad (6)$$

To see why this is the case, the distribution of $c(z_i, b_i)$ in itself can be very skewed to the right, but this becomes even more pronounced in the model-based estimator where the clever covariate has to evaluated at both $c(1, b_i)$ and $c(0, b_i)$ for each $i = 1, \ldots, n$. In contrast, the inverse probability of treatment weighted estimator (6) involves only the probabilities of treatments that were actually assigned.

The approximate Bayesian double robust approach proposed by Graham et al. (2016) involves replacing $m(z, x; \phi)$ in (2) with a linear predictor augmented with the clever covariate. We take this to be a special case of the two-step Bayesian methods to be discussed in § 4, and thus do not consider it separately in the present paper. However, we will show in §6.2 how the form (4) may be derived through posterior predictive expectations and importance sampling.

3. *Inverse probability of treatment weighted outcome regression*

Another estimator of the marginal causal contrast is

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\mathrm{E}(Y_{1i}\,|\,s_i;\widehat{\phi}) - \mathrm{E}(Y_{0i}\,|\,s_i;\widehat{\phi})\right\}, \tag{7}$$

where the parameters ϕ in the model for the potential outcomes (Y_{1i}, Y_{0i}) are estimated using an inverse probability of treatment weighted estimating function $l(\phi) = \sum_{i=1}^{n} l_i(\phi)$, where

$$l_i(\phi) \equiv \sum_{a=0}^{1} \mathrm{I}_{\{z_i=a\}} \frac{\log p(y_{ai}\,|\,s_i;\phi)}{\mathrm{pr}(Z_i = a\,|\,b_i)}. \tag{8}$$

The corresponding estimating equation is $u(\phi) = \sum_{i=1}^{n} u_i(\phi) = 0$, with the pseudo-score function given by $u_i = \partial l_i/\partial\phi$. Here the treatment assignment probabilities $\mathrm{pr}(Z_i = a\,|\,b_i)$ would in practice be replaced with estimates $\mathrm{pr}(Z_i = a\,|\,b_i;\widehat{\gamma})$. The properties of estimators derived from (8) are established in the following theorem.

THEOREM 2. *Suppose that the model for the potential outcomes is correctly specified in the sense that $p(y_{ai}\,|\,s_i;\phi_0) = f(y_{ai}\,|\,s_i)$. Then, if either $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp X_i\,|\,S_i$ or $Z_i \perp\!\!\!\perp X_i\,|\,B_i$ holds, the estimating equation $u(\phi) = 0$ defined by* (8) *is unbiased.*

If we further assume the consistency of the estimator for ϕ, as well as consistency of $\widehat{\gamma}$ when the weights are correctly specified, then (7) consistently estimates the marginal causal contrast. In § 6.1 we demonstrate that an estimator of the form (7) can also be motivated from Bayesian arguments.

4. TWO-STEP ESTIMATION WHEN THE PROPENSITY SCORE IS UNKNOWN

In observational settings the function $e(B_i)$ is unknown and has to be estimated. When using an estimator of the form (3), a central question from a Bayesian perspective then is how the uncertainty in the estimation of the parameters γ is incorporated in the inference of the marginal causal contrast. A Bayesian approach could be motivated by writing the posterior predictive expectation of the potential outcome as

$$\int_{\psi,\gamma,\phi}\int_{x_i} m\{a, e(b_i;\gamma), s_i;\phi\} p(x_i;\psi)\pi_n(\phi\,|\,\gamma)\pi_n(\gamma)\pi_n(\psi)\,\mathrm{d}x_i\,\mathrm{d}\phi\,\mathrm{d}\gamma\,\mathrm{d}\psi, \tag{9}$$

where

$$\pi_n(\phi\,|\,\gamma) \propto \prod_{i=1}^{n} p\{y_i\,|\,z_i, e(b_i;\gamma), s_i;\phi\}\pi_0(\phi), \quad \pi_n(\gamma) \propto \prod_{i=1}^{n} p(z_i\,|\,b_i;\gamma)\pi_0(\gamma). \tag{10}$$

The integrals of the form (9) could be evaluated by Monte Carlo integration, by forward sampling first from $\pi_n(\gamma)$ and given the current value γ, from the conditional posterior $\pi_n(\phi\,|\,\gamma)$. However, the product of the posterior distributions $\pi_n(\phi\,|\,\gamma)$ and $\pi_n(\gamma)$ in (10) may not correspond to any well-defined joint posterior, except when $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp X_i\,|\,S_i$, in which case the outcome does not depend on the propensity score, and further adjustment for the latter is redundant. We summarize this in the following theorem.

THEOREM 3. *If the outcome model is correctly specified in the sense that $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp X_i\,|\,S_i$, then the joint posterior distribution $\pi_n(\phi,\gamma)$ factorizes as $\pi_n(\phi,\gamma) = \pi_n(\phi)\pi_n(\gamma)$, where $\pi_n(\phi) \propto \prod_{i=1}^{n} p(y_i\,|\,z_i, s_i;\phi)\pi_0(\phi)$ and $\pi_n(\gamma) \propto \prod_{i=1}^{n} p(z_i\,|\,b_i;\gamma)\pi_0(\gamma)$.*

Because the specifications in (10) do not necessarily correspond to any joint posterior, the two-step approach outlined above is not proper Bayesian, and would have to be evaluated on its frequency-based properties. It may also be that the two-step Bayesian approach does not result in correct frequency-based inferences. Consider for example the outcome model $m\{z, e(b;\gamma), s;\phi\} = \phi_0 + \phi_1 z + \phi_2^\top s + \phi_3^\top g\{e(b;\gamma)\}$, where g is some appropriate spline basis transformation of the propensity score. With this model the estimator based on (9) for the average causal contrast $\mathrm{E}(Y_{1i})-\mathrm{E}(Y_{0i})$ reduces to $\int_{\phi,\gamma} \phi_1 \pi_n(\phi\,|\,\gamma)\pi_n(\gamma)\,\mathrm{d}\phi\,\mathrm{d}\gamma$, that is, to an estimator of the posterior mean $\mathrm{E}_{\Gamma\,|\,X,Z}\,\{\mathrm{E}(\Phi_1\,|\,v;\gamma)\}$. This estimator can in turn be approximated with $\sum_{j=1}^{m}\widehat{\phi}_1(\gamma^{(j)})/m$, where $(\gamma^{(1)},\ldots,\gamma^{(m)})$ is a Monte Carlo sample from $\pi_n(\gamma)$ (cf.

Kaplan and Chen, 2012, p. 589). We note first that this estimator has the same asymptotic distribution as $\widehat{\phi}_1(\widehat{\gamma})$, where the treatment assignment model parameters have been fixed to their maximum likelihood estimates, and further show in the Supplementary Material that $\mathrm{avar}\{\widehat{\phi}_1(\widehat{\gamma})\} \leq \mathrm{avar}\{\widehat{\phi}_1(\gamma_0)\}$, where $\widehat{\phi}_1(\gamma_0)$ is the estimator given the true propensity scores. Thus, with the propensity score adjusted outcome model specification, a variance adjustment due to estimating the propensity scores should reduce the asymptotic variance of the resulting treatment effect estimator compared to a hypothetical situation where the true propensity scores are known (cf. Henmi and Eguchi, 2004). In contrast, Kaplan and Chen (2012, p. 592) and Graham et al. (2016, p. 11) propose variance estimators based on the variance decomposition formula

$$\mathrm{var}(\Phi_1 \mid v) = \mathrm{E}_{\Gamma \mid X,Z}\left\{\mathrm{var}(\Phi_1 \mid v; \gamma)\right\} + \mathrm{var}_{\Gamma \mid X,Z}\left\{\mathrm{E}(\Phi_1 \mid v; \gamma)\right\}, \tag{11}$$

which appears to add a further variance component. An explanation for the discrepancy is that with the correctly specified models in the representation (1), $p(\phi_1 \mid v; \gamma) = p(\phi_1 \mid v)$, and the second variance component becomes zero. The incompatibility of the posterior distributions in (10) can be contrasted to multiple imputation (e.g., Rubin, 1996), which also makes use of the variance decomposition formula. If the imputation model and the analysis model are compatible, the two-step imputation procedure does produce posterior inferences, and often results in a reasonable approximation even if the models are not exactly the same. However, in the present context there is little motivation to use an approach which adds a component to the posterior variance, given that estimation of the propensity scores should reduce the asymptotic variance of the treatment effect estimator.

5. JOINT ESTIMATION OF OUTCOME AND TREATMENT ASSIGNMENT MODELS

As discussed in § 4, if the outcome model is correctly specified, the treatment assignment model plays no part in the inferences, since the corresponding posterior predictive estimator is then (2). However, the Bayesian propensity score approach proposed by McCandless et al. (2009a) specifies a parametrization making the outcome and treatment assignment models dependent and estimates the parameters jointly. Zigler et al. (2013) suggested that a similar approach could be used to obtain a Bayesian analogue to doubly robust inferences. Such an approach can be understood by assuming that there exists a de Finetti parametrization (ϕ^*, γ^*) for which

$$p(v) = \int_{\phi^*, \gamma^*, \psi} \prod_{i=1}^{n} \left\{p(y_i, z_i \mid x_i; \phi^*, \gamma^*) p(x_i; \psi)\right\} \pi_0(\phi^*)\pi_0(\gamma^*)\pi_0(\psi)\, \mathrm{d}\phi^*\, \mathrm{d}\gamma^*\, \mathrm{d}\psi,$$

where $p(y_i, z_i \mid x_i; \phi^*, \gamma^*) = p\{y_i \mid z_i, e(b_i; \gamma^*), s_i; \phi^*\} p(z_i \mid b_i; \gamma^*)$. Compared to (ϕ, γ) in (1), neither ϕ^* or γ^* retains the original interpretation, but now there is a well-defined joint posterior distribution

$$\pi_n(\phi^*, \gamma^*) \propto \prod_{i=1}^{n} \left[p\{y_i \mid z_i, e(b_i; \gamma^*), s_i; \phi^*\} p(z_i \mid b_i; \gamma^*)\right] \pi_0(\phi^*)\pi_0(\gamma^*).$$

Inferences could now be based on the posterior predictive expectations

$$\int_{\phi^*, \gamma^*, \psi} \int_{x_i} m\{a, e(b_i; \gamma^*), s_i; \phi^*\} p(x_i; \psi) \pi_n(\phi^*, \gamma^*) \pi_n(\psi)\, \mathrm{d}x_i\, \mathrm{d}\phi^*\, \mathrm{d}\gamma^*\, \mathrm{d}\psi. \tag{12}$$

At first sight, (12) would seem more natural than (9), since the specification (12) does not make use of incompatible models. However, the quantities $e(b_i; \gamma^*)$ do not possess the balancing properties of propensity scores, and thus the properties of the estimator given by (12) would be difficult to establish. To address the lack of balance, McCandless et al. (2010) suggested a Gibbs-sampler type approach similar to that of Lunn et al. (2009) to cut the feedback from the outcome model by successively drawing from the conditional posteriors $\pi_n(\gamma)$ and $\pi_n(\phi \mid \gamma)$ to approximate the joint posterior of (ϕ^*, γ^*). However, as discussed in § 4, these posteriors are incompatible and such a sampling procedure is not guaranteed to converge to any well-defined joint distribution. In fact, if the conditional posteriors can be sampled directly, or if the second sampling step is allowed to converge to the corresponding conditional distribution, the inferences based on the formulations (9) and (12) will be equivalent.

To sum up, trying to recover fully probabilistic inferences through sampling from a joint posterior of the outcome and treatment assignment model parameters loses the balancing property of the propensity scores. On the other hand, cutting the feedback in an attempt to recover the balancing property would mean that the inferences are no longer necessarily based on well-defined posterior distributions. Thus, in § 6, extending the approach outlined in Saarela et al. (2015b) to doubly robust estimation, we formulate alternative Bayesian estimators that are not based on Bayesian propensity score adjustment, and yet do possess the double robustness property.

6. POSTERIOR PREDICTIVE INFERENCES WITH IMPORTANCE SAMPLING

6. Inverse probability of treatment weighted outcome regression

The estimator (2) can be motivated without the use of potential outcomes notation as a posterior predictive expectation for a new observation under a hypothetical completely randomized treatment assignment mechanism or regime indexed by $\mathcal{E}$ under which $Z_i \perp\!\!\!\perp_{\mathcal{E}} X_i$ and the probabilities $\mathrm{pr}_{\mathcal{E}}(Z_i = a)$ are known constants; cf. the randomized trial measure of Røysland (2011). The data are observed under a regime indexed by $\mathcal{O}$, where $Z_i \not\perp\!\!\!\perp_{\mathcal{O}} X_i$, and causal inference then corresponds to inference across these regimes. It has been recognized by various authors (e.g., Røysland, 2011; Chakraborty and Moodie, 2013) that inverse probability of treatment weighting can be motivated through a change of probability measures, or equivalently, importance sampling. However, as far as we know, before Saarela et al. (2015b,c) this approach has not been used to formulate Bayesian causal inferences. Here we argue that it can resolve the paradoxes discussed in § 4 and § 5. While Saarela et al. (2015b,c) considered inverse probability weighted estimation of marginal structural models, in the context of estimating optimal dynamic treatment regimes Saarela et al. (2015a) suggested that the importance sampling procedure could also be used for weighted estimation of parametric outcome models, with the aim to achieve some protection against misspecification of the latter. We will further study this possibility in this section. In § 6.2 we use the importance sampling approach to motivate a semi-parametric doubly robust estimator.

Rather than trying to directly predict a new observation under the experimental setting $\mathcal{E}$, which we address in §6.2, we first consider parameter estimation under this regime. We approach this through a decision rule that maximizes an expected utility with respect to the parameters. We take the log-likelihood $l(\phi; v_i) \equiv \log p(y_i \mid z_i, s_i; \phi)$ for the outcome model parameters ϕ as the parametric utility function. A Bayesian estimator for ϕ can then be constructed by maximizing the expected utility $\mathrm{E}_{\mathcal{E}}\{l(\phi; V_i) \mid v\}$ with respect to ϕ, where the expectation is over a predicted new observation $v_i = (x_i, y_i, z_i)$, $i \notin \{1, \dots, n\}$.

Let ξ be a set of parameters characterizing the entire data-generating mechanism under the observational regime $\mathcal{O}$. We can write the expectation as $\mathrm{E}_{\mathcal{E}}\{l(\phi; V_i) \mid v\} = \mathrm{E}_{\Xi \mid V}\left[\mathrm{E}_{\mathcal{E}}\{l(\phi; V_i) \mid v; \xi\}\right]$, where, following Walker (2010, p. 26–27), we can consider the lower-dimensional task of maximizing the expected utility $\mathrm{E}_{\mathcal{E}}\{l(\phi; V_i) \mid v; \xi\}$ with respect to ϕ conditional on ξ. With a known regime $\mathcal{E}$ and the stability assumption of Dawid and Didelez (2010), $\arg\max_{\phi} \mathrm{E}_{\mathcal{E}}\{l(\phi; V_i) \mid v; \xi\}$ is a deterministic function of ξ. Thus, the uncertainty represented by the posterior distribution $\pi_n(\xi) \equiv p_{\mathcal{O}}(\xi \mid v)$ also then reflects the uncertainty about ϕ, providing a means to construct a posterior distribution for ϕ. This proceeds as follows; the inner expectation can be written as

$$\begin{aligned}
\mathrm{E}_{\mathcal{E}}\{l(\phi; V_i) \mid v; \xi\} &= \int_{v_i} l(\phi; v_i) p_{\mathcal{E}}(v_i \mid v; \xi)\, \mathrm{d}v_i \\
&= \int_{v_i} l(\phi; v_i) \frac{p_{\mathcal{E}}(v_i \mid v; \xi)}{p_{\mathcal{O}}(v_i \mid v; \xi)} p_{\mathcal{O}}(v_i \mid v; \xi)\, \mathrm{d}v_i \\
&= \int_{v_i} l(\phi; v_i) \frac{p_{\mathcal{E}}(y_i \mid z_i, x_i, v; \xi) p_{\mathcal{E}}(z_i) p_{\mathcal{E}}(x_i \mid v; \xi)}{p_{\mathcal{O}}(y_i \mid z_i, x_i, v; \xi) p_{\mathcal{O}}(z_i \mid x_i, v; \xi) p_{\mathcal{O}}(x_i \mid v; \xi)} p_{\mathcal{O}}(v_i \mid v; \xi)\, \mathrm{d}v_i \\
&= \int_{v_i} l(\phi; v_i) \frac{p_{\mathcal{E}}(z_i)}{p_{\mathcal{O}}(z_i \mid x_i, v; \xi)} p_{\mathcal{O}}(v_i \mid v; \xi)\, \mathrm{d}v_i,
\end{aligned}$$

using in the last equality the stability assumption, namely $p_{\mathcal{E}}(y_i \mid z_i, x_i; \phi) = p_{\mathcal{O}}(y_i \mid z_i, x_i; \phi)$ and $p_{\mathcal{E}}(x_i; \psi) = p_{\mathcal{O}}(x_i; \psi)$. We can replace the predictive distribution under $\mathcal{O}$ with the Bayesian bootstrap specification $p_{\mathcal{O}}(v_i \mid v; \xi) = \sum_{k=1}^{n} \xi_k \delta_{v_k}(v_i)$, where $\xi \equiv (\xi_1, \dots, \xi_n)$ and $\Xi \mid v \sim \text{Dirichlet}(1, \dots, 1)$. With $w_i(\xi) \equiv p_{\mathcal{E}}(z_i)/p_{\mathcal{O}}(z_i \mid x_i, v; \xi)$, the expected utility becomes

$$\mathrm{E}_{\mathcal{E}}\{l(\phi; V_i) \mid v; \xi\} = \int_{v_i} l(\phi; v_i) w_i(\xi) \sum_{k=1}^{n} \xi_k \delta_{v_k}(v_i)\, \mathrm{d}v_i = \sum_{k=1}^{n} w_k(\xi) \xi_k l(\phi; v_k), \qquad (13)$$

a weighted log-likelihood, motivating the estimator

$$\widehat{\phi}(\xi) \equiv \arg\max_{\phi} \sum_{k=1}^{n} w_k(\xi)\xi_k l(\phi; v_k).$$

This approach of creating a mapping between the non-parametric specification and a parametrization relevant to inferences is analogous to Newton and Raftery (1994) and Chamberlain and Imbens (2003), but multiplies the the Dirichlet weights by the importance sampling weights $w_i(\xi)$ in order to make inferences across the observational and experimental regimes. The importance-sampling weights add variability to the estimation, but provide some protection against misspecification of the outcome model in the sense of § 3.3. In principle, the weights are fully determined by the current realization of ξ under the non-parametric specification, but in practice parametric model specifications are needed for smoothing purposes, and we must link ξ and the treatment assignment model parameters γ. For this purpose γ can be estimated through the weighted likelihood bootstrap of Newton and Raftery (1994), which readily gives the deterministic function linking the two parametrizations; thus in (13) we choose $w_i(\xi) = p_{\mathcal{E}}(z_i)/p_{\mathcal{O}}\{z_i \mid b_i; \widehat{\gamma}(\xi)\}$, where $\widehat{\gamma}(\xi) \equiv \arg\max_{\gamma} \sum_{k=1}^{n} \xi_k \log p(z_k \mid b_k; \gamma)$. The probabilities $p_{\mathcal{E}}(z_i)$ are given by the chosen regime $\mathcal{E}$ that is the object of inference; in practice estimation is most efficient when we choose the target regime to be as close as possible to the observed regime $\mathcal{O}$; this can be achieved by fixing $p_{\mathcal{E}}(z_i)$ to the marginal treatment assignment probabilities under $\mathcal{O}$, which would result in the usual kind of stabilized inverse probability of treatment weights used in marginal structural modelling (Robins et al., 2000; Hernán et al., 2001; Cole and Hernán, 2008).

Since

$$\mathrm{E}_{\mathcal{E}}(Y_i \mid Z_i = a, v; \xi) = \int_{x_i} m(a, x_i; \xi) p(x_i \mid v; \xi)\, \mathrm{d}x_i \tag{14}$$

is again a deterministic function of ξ, an approximate posterior distribution for the predictive means under $\mathcal{E}$ may be constructed by using the non-parametric specification $p(x_i \mid v; \xi) = \sum_{k=1}^{n} \xi_k \delta_{x_k}(x_i)$, repeatedly sampling ξ from the uniform Dirichlet distribution and substituting $m(a, x_i; \xi) = m\{a, s_i; \widehat{\phi}(\xi)\}$. In particular, for the posterior mean, we get the estimator

$$\begin{aligned}
\mathrm{E}_{\Xi \mid V}\{\mathrm{E}_{\mathcal{E}}(Y_i \mid Z_i = a, v; \xi)\} &= \int_{\xi} \int_{s_i} m\{a, s_i; \widehat{\phi}(\xi)\} \sum_{k=1}^{n} \xi_k \delta_{s_k}(s_i) \pi_n(\xi)\, \mathrm{d}s_i\, \mathrm{d}\xi \\
&= \int_{\xi} \sum_{k=1}^{n} \xi_k m\{a, s_k; \widehat{\phi}(\xi)\} \pi_n(\xi)\, \mathrm{d}\xi, \qquad (15)
\end{aligned}$$

which is the direct Bayesian analogue of (7), where the outcome model was estimated using inverse probability of treatment weighted regression. In fact, if we fix $\xi_k = 1/n$ $(k = 1, \ldots, n)$, instead of considering these as unknown parameters, the two estimators are equivalent. Thus, we conjecture that the estimator given by (15) has a similar double robustness property as (7). We support this through simulations in Section 7.

6. *Doubly robust estimation*

We now show how the semi-parametric doubly robust estimator (4) can be motivated through posterior predictive expectations. Under the non-parametric specification in terms of ξ, the posterior predictive causal contrast conditional on ξ may be written as

$$\begin{aligned}
&\mathrm{E}_{\mathcal{E}}(Y_i \mid Z_i = 1, v; \xi) - \mathrm{E}_{\mathcal{E}}(Y_i \mid Z_i = 0, v; \xi) \\
&\quad = \sum_{k=1}^{n} \xi_k \{y_k - m(z_k, x_k; \xi)\} \left\{ \frac{z_k}{\mathrm{pr}_{\mathcal{O}}(Z_k = 1 \mid x_k, v; \xi)} - \frac{1 - z_k}{\mathrm{pr}_{\mathcal{O}}(Z_k = 0 \mid x_k, v; \xi)} \right\} \\
&\qquad + \sum_{k=1}^{n} \xi_k \{m(1, x_k; \xi) - m(0, x_k; \xi)\}, \qquad (16)
\end{aligned}$$

which corresponds to formulation (4). Again, similarly to (14), this is a deterministic function of ξ, and an approximate posterior distribution for the causal contrast may be constructed by resampling ξ. The posterior mean is then estimated by

$$\mathrm{E}_{\Xi \mid V}\{\mathrm{E}_{\mathcal{E}}(Y_i \mid Z_i = 1, v; \xi) - \mathrm{E}_{\mathcal{E}}(Y_i \mid Z_i = 0, v; \xi)\}. \tag{17}$$

Since the non-parametric specification places no restrictions on the conditional distributions, in practice the non-parametrically specified quantities $m(z_k, x_k; \xi)$ and $\mathrm{pr}_{\mathcal{O}}(Z_k = a \mid x_k, v; \xi)$ would have to be replaced with the parametric versions $m\{z_k, s_k; \widehat{\phi}(\xi)\}$ and $\mathrm{pr}_{\mathcal{O}}\{Z_k = a \mid b_k; \widehat{\gamma}(\xi)\}$, estimated using the weighted likelihood bootstrap. It is straightforward to see that if the outcome model is correctly specified, the expression (16) reduces to the model-based mean difference appearing on the last line of (16). This reflects the fact that if we believe in our outcome model, the treatment assignment model does not play a part in the inferences. However, a Bayesian might want to use an estimator of the form (16) if being restricted by two parametric constraints, in terms of ϕ and γ, but not knowing which of these is correct. If either is correct, in the sense that they produce the same conditional mean outcomes or treatment assignment probabilities as the non-parametric specification, the resulting estimator will equal the posterior mean (17) under the non-parametric specification. We summarize this property in the following theorem.

THEOREM 4. *Suppose that either of the parametric specifications is correct in the sense that* $m\{z_k, s_k; \widehat{\phi}(\xi)\} = m(z_k, x_k; \xi)$ *or* $\mathrm{pr}_{\mathcal{O}}\{Z_k = a \mid b_k; \widehat{\gamma}(\xi)\} = \mathrm{pr}_{\mathcal{O}}(Z_k = a \mid x_k, v; \xi)$*. Then the estimator obtained by substituting the parametric specifications into expression* (16) *and taking the expectation over* ξ *is equal to the posterior mean* (17) *under the non-parametric specification.*

7. SIMULATION STUDY

Above we have made a distinction between model misspecification due to omission of relevant covariates, and misspecification of the parametric functional relationship between the outcome and the covariates. We noted that all the estimators discussed in Section 3 should be doubly robust against the former type of misspecification. However, in practice the consequences of these two types of misspecification will often be similar; they result in residual confounding. Therefore, we investigate how the different estimators perform when the covariate sets S_i and B_i are not only created by a partitioning, but also a transformation of the x_i's. For this purpose, we simulated covariates $X_{ij} \sim N(0, 1)$ $(j = 1, \ldots, 4)$ and transformed these as $u_{ij} = |x_{ij}|/(1 - 2/\pi)^{1/2}$. The true treatment assignment and outcome mechanisms were

$$Z_i \mid x_i \sim \mathrm{Ber}\left\{\mathrm{expit}(0.4u_{i1} + 0.4x_{i2} + 0.8x_{i3})\right\}, \qquad Y_i \mid z_i, x_i \sim N(z_i - u_{i1} - x_{i2} - x_{i4}, 1),$$

respectively. For estimation, we considered two scenarios: (I) misspecified outcome model and correctly specified treatment assignment model with $s_i \equiv (x_{i1}, x_{i2}, x_{i4})$ and $b_i \equiv (u_{i1}, x_{i2}, x_{i3})$, and (II) correctly specified outcome model and misspecified treatment assignment model with $s_i \equiv (u_{i1}, x_{i2}, x_{i4})$ and $b_i \equiv (x_{i1}, x_{i2}, x_{i3})$.

We are interested in the marginal causal contrast $\mathrm{E}(Y_{i1}) - \mathrm{E}(Y_{i0})$, with the true difference equal to 1. To estimate this, we applied the Bayesian estimators discussed in §§4, 5, 6. In the two-step estimation we attempted both forward sampling from the posterior distributions, and the variance decomposition formula (11). In the former, instead of Markov chain Monte Carlo, we applied normal approximations for the posterior distributions, of the form $\Phi \mid v; \gamma \sim N\{\widehat{\phi}(\gamma), S\}$, where $\widehat{\phi}(\gamma)$ is the maximum likelihood estimate and S its estimated variance-covariance matrix. The posterior distribution $\Gamma \mid x, z$ was approximated using the weighted likelihood bootstrap. In the joint estimation, we again used a normal approximation, centered at the joint maximum likelihood estimates $(\widehat{\phi}, \widehat{\gamma})$, and with covariance matrix given by the inverse of the observed information at the maximum likelihood point. In both two-step and joint estimation, the fitted models were specified as $m\{z_i, e(b_i; \gamma), s_i; \phi\} = \phi_0 + \phi_1 z_i + \phi_2^\top s_i + \phi_3^\top g\{e(b_i; \gamma)\}$, where g is a cubic polynomial basis, and $e(b_i; \gamma) = \mathrm{expit}(\gamma_0 + \gamma_1^\top b_i)$. In the importance sampling estimator proposed in § 6.1, and the doubly robust importance sampling estimator of § 6.2, the fitted treatment assignment model was the same, with the outcome model specified through $m(z_i, s_i; \phi) = \phi_0 + \phi_1 z_i + \phi_2^\top s_i$. For comparison to the Bayesian estimators, we also calculate naive unadjusted comparison, outcome regression adjusted for covariates s_i as well as the frequentist estimators of §3. To demonstrate the variance estimation issues discussed in §4, we calculated for estimator (3) both observed information-based standard errors and the adjusted sandwich-type standard errors discussed in the Supplementary Material. For the weighted estimators (4), (6) and (7), the standard errors were estimated through the frequentist nonparametric bootstrap (Efron, 1979).

The results over 5000 replications are shown in Table 1. Under the first scenario, all except the naive and adjusted estimators can correct for confounding, although the joint estimation approach produces a slight bias. The estimators based on propensity score adjusted outcome regression are the most efficient, with the inverse probability of treatment weighting based estimators losing slightly. As discussed in § 3.2, the clever covariate approach results in higher variability compared to the other doubly-adjusted estimators. In terms of variance estimation, the comparison between the unadjusted and adjusted standard errors for the propensity score adjusted outcome regression based estimator (3) suggests that under this simulation setting estimation of the propensity scores substantially reduces the variance, and not adjusting for this results in overcoverage. The resampling based variance estimators adjust for this automatically. However, the two-step approach to variance estimation performs poorly; as demonstrated in the Supplementary Material, the two-step point estimator has the same asymptotic variance as the other estimators based on (3), but the two-step variance estimators unnecessarily add a further variance component.

Under the second scenario, all the estimators except the inverse probability of treatment weighted estimator (6) are unbiased, which is expected based on their previously discussed theoretical properties. When the outcome model is correctly specified, there is also very little difference in the efficiencies of the various estimators.

The simulation results support the discussion in § 4 and § 5; the two-step and joint estimation do not seem to provide practical advantages in terms of their frequency-based properties. Rather, the two-step approach produces inflated posterior variance estimates, while the joint estimation seems to do worse in terms of bias compared to the corresponding frequentist propensity score adjusted estimator. On the other hand, the importance sampling-based Bayesian estimators produce very similar results to corresponding frequentist estimators (4) and (7).

8. DISCUSSION

We have demonstrated the feasibility of full Bayesian doubly robust causal inference. In the methods proposed in § 6, we used the non-parametric Bayesian bootstrap specification for the joint distribution of the observed data. As this places no restrictions on the resulting conditional distributions, we considered the double robustness property with respect to the non-parametric specification, which was taken to be the correct model. Because of the non-restricted conditional distributions, for estimation we considered mappings between parametric models and the Dirichlet weights in the non-parametric specification. Considering other non-parametric Bayesian modelling approaches in combination with the importance sampling estimators is a topic for further research. For example, if the joint distribution of the observed data was modelled through a Dirichlet process mixture model (e.g., Escobar and West, 1995), the required conditional distributions could be recovered directly from the Bayesian non-parametric model without the need for parametric specifications.

The disadvantage of the importance sampling approach is the same as in the corresponding frequentist inverse probability of treatment weighted inference procedures: the importance sampling weights add variability to the point estimator. In order to control this, a standard approach would be to truncate the weights (e.g., Xiao et al., 2013), which would also be possible in the importance sampling context (Ionides, 2008). Recently, Vehtari & Gelman (2015, arXiv:1507.02646v2) suggested probabilistic truncation of importance sampling weights; studying this in the present context is another topic for further research.

ACKNOWLEDGEMENT

The authors would like to thank the two Referees and the Editors for the constructive comments that helped us to improve the manuscript and acknowledge support of the Natural Sciences and Engineering Research Council of Canada.

Table 1: Estimates for the marginal causal contrast, with true value equal to 1, over 5000 simulation rounds

Scenario	Estimator	Point estimate	Relative bias (%)	SD ($\times 10$)	SE ($\times 10$)	Coverage (%)
(I)	Naive	0.34	−65.5	1.28	1.28	0.1
	Adjusted	0.67	−33.0	0.92	0.92	5.0
	IPTW	1.00	−0.0	1.34	1.37	94.1
	OR/PS (obs. information)	1.00	−0.1	0.71	0.95	99.3
	OR/PS (sandwich)	1.00	−0.1	0.71	0.93	99.1
	OR/PS (adj. sandwich)	1.00	−0.1	0.71	0.71	95.3
	DR	1.00	−0.2	0.88	0.89	94.7
	Clever covariate	1.02	2.4	1.11	1.11	93.3
	OR/IPTW	0.99	−1.0	0.83	0.83	94.5
	Two-step (forward sampling)	1.00	0.1	0.71	1.13	99.9
	Two-step (variance decomposition)	1.00	−0.4	0.71	1.12	99.9
	Joint estimation	1.05	4.8	0.71	0.71	89.1
	Importance sampling	0.99	−0.9	0.83	0.81	93.9
	Importance sampling/DR	1.00	−0.3	0.87	0.86	94.1
(II)	Naive	0.35	−65.5	1.28	1.28	0.1
	Adjusted	1.00	−0.1	0.66	0.67	95.6
	IPTW	0.63	−37.2	1.31	1.30	19.4
	OR/PS (obs. information)	1.00	−0.1	0.70	0.71	95.4
	OR/PS (sandwich)	1.00	−0.1	0.70	0.71	95.4
	OR/PS (adj. sandwich)	1.00	−0.1	0.70	0.71	95.4
	DR	1.00	−0.1	0.74	0.74	95.2
	Clever covariate	1.00	−0.1	0.75	0.75	95.1
	OR/IPTW	1.00	−0.1	0.73	0.74	95.1
	Two-step (forward sampling)	1.00	0.2	0.70	0.72	96.3
	Two-step (variance decomposition)	1.00	−0.1	0.70	0.71	95.5
	Joint estimation	1.00	−0.1	0.70	0.71	95.4
	Importance sampling	1.00	−0.1	0.74	0.72	94.9
	Importance sampling/DR	1.00	−0.1	0.74	0.73	94.8

SD, Monte Carlo standard deviation; SE, mean standard error estimate; Coverage, 95% confidence interval coverage probability; Scenario (I): Misspecified outcome model and correctly specified treatment assignment model, Scenario (II): Correctly specified outcome model, misspecified treatment assignment model, $s_i \equiv (u_{i1}, x_{i2}, x_{i4})$ and $b_i \equiv (x_{i1}, x_{i2}, x_{i3})$; Naive, naive unadjusted comparison; adjusted, outcome regression adjusted for covariates s_i; IPTW, estimator (6); DR, estimator (4); CC, clever covariate version of (4); OR/IPTW, estimator (7); OR/PS, estimator (3). The largest Monte Carlo error (batch means) of the mean point estimate in Column 1 is 0.002.

Supplementary material

S1. Causal inference as a prediction problem

To motivate estimator (2) using the notations discussed in §6, we can write for $i \notin \{1, \ldots, n\}$

$$
\begin{aligned}
&\mathrm{E}_{\mathcal{E}}(Y_i \mid Z_i = a, v) \\
&\quad = \frac{\int_{\phi,\psi}\int_{x_i}\int_{y_i} y_i p(y_i \mid Z_i = a, x_i; \phi)\mathrm{pr}_{\mathcal{E}}(Z_i = a)p(x_i;\psi)\pi_n(\phi)\pi_n(\psi)\,\mathrm{d}y_i\,\mathrm{d}x_i\,\mathrm{d}\phi\,\mathrm{d}\psi}{\int_{\phi,\psi}\int_{x_i}\int_{y_i} p(y_i \mid Z_i = a, x_i; \phi)\mathrm{pr}_{\mathcal{E}}(Z_i = a)p(x_i;\psi)\pi_n(\phi)\pi_n(\psi)\,\mathrm{d}y_i\,\mathrm{d}x_i\,\mathrm{d}\phi\,\mathrm{d}\psi} \\
&\quad = \int_{\phi,\psi}\int_{x_i} m(a, x_i; \phi)p(x_i;\psi)\pi_n(\phi)\pi_n(\psi)\,\mathrm{d}x_i\,\mathrm{d}\phi\,\mathrm{d}\psi \qquad (18) \\
&\quad = \int_{\phi,\psi}\sum_{k=1}^{n} \psi_k m(a, x_k; \phi)\pi_n(\phi)\pi_n(\psi)\,\mathrm{d}\phi\,\mathrm{d}\psi, \qquad (19)
\end{aligned}
$$

where $\psi = (\psi_1, \dots, \psi_n)$ and $\pi_n(\psi)$ is taken to be the uniform Dirichlet distribution. The last form was obtained by using the non-parametric Bayesian bootstrap (Rubin, 1981) specification

$$\begin{aligned}\int_\psi p(x_i;\psi)\pi_n(\psi)\,\mathrm{d}\psi = p(x_i \mid x_1,\dots,x_n) &= \int_\psi p(x_i \mid x_1,\dots,x_n;\psi)\pi_n(\psi)\,\mathrm{d}\psi \\ &= \int_\psi \sum_{k=1}^n \psi_k \delta_{x_k}(x_i)\pi_n(\psi)\,\mathrm{d}\psi.\end{aligned}$$

Obtaining (18) also required assuming that $p_\mathcal{E}(y_i \mid z_i, x_i; \phi) = p_\mathcal{O}(y_i \mid z_i, x_i; \phi) \equiv p(y_i \mid z_i, x_i; \phi)$ and $p_\mathcal{E}(x_i;\psi) = p_\mathcal{O}(x_i;\psi) \equiv p(x_i;\psi)$, which corresponds to the stability assumption discussed by Dawid and Didelez (2010).

S2. ON THE FREQUENCY-BASED PROPERTIES OF THE TWO-STEP APPROACH

As in Theorem 3, if the outcome model is correctly specified, then $\Phi \perp\!\!\!\perp \Gamma \mid V$ and (9) reduces to (18). The interesting situations are naturally those where this is not the case. We denote $q_i(\phi;\gamma) = \log p\{y_i \mid z_i, e(b_i;\gamma), s_i;\phi)\}$ and $q(\phi;\gamma) \equiv \sum_{i=1}^n q_i(\phi;\gamma)$ and consider the quasi-maximum likelihood estimator $\widehat{\phi}(\widehat{\gamma}) \equiv \arg\max_\phi q(\phi;\widehat{\gamma})$. If the treatment assignment model is correctly specified, $\widehat{\gamma} \to \gamma_0$. In addition, we assume that with any fixed value of γ, $\widehat{\phi}(\gamma) \to \phi_0(\gamma)$, where $\phi_0(\gamma)$ is the parameter vector which minimizes the Kullback-Leibler divergence to the true outcome model (e.g. White, 1982, p. 4). Thus, by the law of large numbers and continuous mapping, we can write in the usual way that

$$\frac{1}{n}\sum_{i=1}^n [q_i(\phi;\widehat{\gamma}) - q_i\{\phi_0(\gamma_0);\gamma_0\}] \to E\left[q_i(\phi;\gamma_0) - q_i\{\phi_0(\gamma_0);\gamma_0\}\right],$$

where the right hand side is maximized at zero when $\phi = \phi_0(\gamma_0)$, at which point $\mathrm{E}\{Y_i \mid Z_i = a, e(b_i;\gamma_0), s_i; \phi_0(\gamma_0)\} = \mathrm{E}\{Y_{ia} \mid e(b_i;\gamma_0), s_i; \phi_0(\gamma_0)\}$. Since we also have that the posterior $p(\gamma \mid x, y) \to \delta_{\gamma_0}(\gamma)$ in distribution, we can then conjecture that posterior predictive inferences based on (9) will be asymptotically unconfounded.

With the definitions

$$\begin{aligned}U^\phi(\phi;\gamma) &\equiv \partial q(\phi;\gamma)/\partial\phi,\\ U^{\phi\phi}(\phi;\gamma) &\equiv \partial^2 q(\phi;\gamma)/\partial\phi^2,\\ U^{\phi\gamma}(\phi;\gamma) &\equiv \partial^2 q(\phi;\gamma)/\partial\phi\partial\gamma,\\ U^\gamma(\gamma) &\equiv \partial \textstyle\sum_{i=1}^n \log p(z_i \mid b_i;\gamma)/\partial\gamma,\\ U^{\gamma\gamma}(\gamma) &\equiv \partial^2 \textstyle\sum_{i=1}^n \log p(z_i \mid b_i;\gamma)/\partial\gamma^2,\end{aligned}$$

and noting that $U^\phi(\widehat{\phi};\gamma^{(j)}) = 0$ for each $\gamma^{(j)}$, $j = 1,\dots,m$, we can consider the first order Taylor expansion of $U^\phi(\widehat{\phi};\gamma^{(j)})$ around the true parameter values (ϕ_0,γ_0), which becomes

$$\begin{aligned}0 &= \frac{1}{n}U^\phi(\widehat{\phi};\gamma^{(j)})\\ &\approx \frac{1}{n}U^\phi(\phi_0;\gamma_0) + \frac{1}{n}U^{\phi\phi}(\phi_0;\gamma_0)\{\widehat{\phi}(\gamma^{(j)}) - \phi_0\} + \frac{1}{n}U^{\phi\gamma}(\phi_0;\gamma_0)(\gamma^{(j)} - \gamma_0)\\ &\approx \frac{1}{n}U^\phi(\phi_0;\gamma_0) + \mathrm{E}\{U_i^{\phi\phi}(\phi_0;\gamma_0)\}\{\widehat{\phi}(\gamma^{(j)}) - \phi_0\} + \mathrm{E}\{U_i^{\phi\gamma}(\phi_0;\gamma_0)\}(\gamma^{(j)} - \gamma_0),\end{aligned}$$

and further,

$$\begin{aligned}0 &= \frac{1}{m}\sum_{j=1}^m \frac{1}{n}U^\phi(\widehat{\phi};\gamma^{(j)})\\ &\approx \frac{1}{n}U^\phi(\phi_0;\gamma_0) + \mathrm{E}\{U_i^{\phi\phi}(\phi_0;\gamma_0)\}\left\{\frac{1}{m}\sum_{j=1}^m \widehat{\phi}(\gamma^{(j)}) - \phi_0\right\} + \mathrm{E}\{U_i^{\phi\gamma}(\phi_0;\gamma_0)\}\left(\widehat{\gamma} - \gamma_0\right),\end{aligned}$$

if $\frac{1}{m}\sum_{j=1}^{m}\gamma^{(j)} \approx \widehat{\gamma}$. Hence,

$$\begin{aligned} n^{1/2}\left\{\frac{1}{m}\sum_{j=1}^{m}\widehat{\phi}(\gamma^{(j)}) - \phi_0\right\} &\approx \mathrm{E}\{-U_i^{\phi\phi}(\phi_0;\gamma_0)\}^{-1} \\ &\quad\times\left[\frac{n^{1/2}}{n}U^{\phi}(\phi_0;\gamma_0) + \mathrm{E}\{U_i^{\phi\gamma}(\phi_0;\gamma_0)\}n^{1/2}(\widehat{\gamma}-\gamma_0)\right]. \end{aligned}$$

Here we have, by another first order expansion around γ_0, that

$$n^{1/2}(\widehat{\gamma}-\gamma_0) \approx \mathrm{E}\{-U_i^{\gamma\gamma}(\gamma_0)\}^{-1}\frac{n^{1/2}}{n}U^{\gamma}(\gamma_0),$$

so finally,

$$\begin{aligned} n^{1/2}\left\{\frac{1}{m}\sum_{j=1}^{m}\widehat{\phi}(\gamma^{(j)}) - \phi_0\right\} &\approx \mathrm{E}\{-U_i^{\phi\phi}(\phi_0;\gamma_0)\}^{-1} \\ &\quad\times\left(\frac{n^{1/2}}{n}\sum_{i=1}^{n}\left[U_i^{\phi}(\phi_0;\gamma_0) + \mathrm{E}\{U_i^{\phi\gamma}(\phi_0;\gamma_0)\}\mathrm{E}\{-U_i^{\gamma\gamma}(\gamma_0)\}^{-1}U_i^{\gamma}(\gamma_0)\right]\right). \end{aligned}$$

We may similarly expand $U^{\phi}(\widehat{\phi};\widehat{\gamma})$ where the parameters γ have been fixed to their maximum likelihood estimates around (ϕ_0,γ_0) as

$$\begin{aligned} 0 &= \frac{1}{n}U^{\phi}(\widehat{\phi};\widehat{\gamma}) \\ &\approx \frac{1}{n}U^{\phi}(\phi_0;\gamma_0) + \mathrm{E}\{U_i^{\phi\phi}(\phi_0;\gamma_0)\}(\widehat{\phi}(\widehat{\gamma})-\phi_0) + \mathrm{E}\{U_i^{\phi\gamma}(\phi_0;\gamma_0)\}(\widehat{\gamma}-\gamma_0) \end{aligned}$$

to find that

$$n^{1/2}\left\{\widehat{\phi}(\widehat{\gamma})-\phi_0\right\} \approx \mathrm{E}\{-U_i^{\phi\phi}(\phi_0;\gamma_0)\}^{-1}\frac{n^{1/2}}{n}\sum_{i=1}^{n}B_i(\phi_0;\gamma_0),$$

where

$$B_i(\phi_0;\gamma_0) \equiv U_i^{\phi}(\phi_0;\gamma_0) + \mathrm{E}\{U_i^{\phi\gamma}(\phi_0;\gamma_0)\}\mathrm{E}\{-U_i^{\gamma\gamma}(\gamma_0)\}^{-1}U_i^{\gamma}(\gamma_0).$$

Since the two estimators $\frac{1}{m}\sum_{j=1}^{m}\widehat{\phi}(\gamma^{(j)})$ and $\widehat{\phi}(\widehat{\gamma})$ have the same linear approximation which is a sum of independent terms, we conclude that they have the same asymptotic distribution,

$$n^{1/2}(\widehat{\phi}-\phi_0) \to \mathrm{N}\left[0, \mathrm{E}\{-U_i^{\phi\phi}(\phi_0;\gamma_0)\}^{-1}\mathrm{var}\{B_i(\phi_0;\gamma_0)\}\mathrm{E}\{-U^{\phi\phi}(\phi_0;\gamma_0)^{\top}\}^{-1}\right].$$

Fitting the regression model $y_i = \phi_0 + \phi_1 z_i + \phi_2^{\top}s_i + \phi_3^{\top}g\{e(b_i;\gamma)\} + \varepsilon_{1i}$ to estimate the parameter of interest ϕ_1 is numerically equivalent to fitting the sequence of regressions $y_i = \nu^{\top}s_i^*(\widehat{\gamma}) + \varepsilon_{2i}$, $z_i = \alpha^{\top}s_i^*(\widehat{\gamma}) + \varepsilon_{3i}$ and $\{y_i - \widehat{\nu}^{\top}s_i^*(\widehat{\gamma})\} = \phi_1\{z_i - \widehat{\alpha}^{\top}s_i^*(\widehat{\gamma})\} + \varepsilon_{4i}$, where $s_i^*(\gamma) \equiv [s_i, g\{e(b_i;\gamma)\}]$. Denoting the estimating function corresponding to the last regression as

$$U^{\phi_1}\{\phi_1,\widehat{\gamma},\widehat{\nu}(\widehat{\gamma}),\widehat{\alpha}(\widehat{\gamma})\} \equiv \sum_{i=1}^{n}\{z_i - \widehat{\alpha}^{\top}s_i^*(\widehat{\gamma})\}\left[\{y_i - \widehat{\nu}^{\top}s_i^*(\widehat{\gamma})\} - \phi_1\{z_i - \widehat{\alpha}^{\top}s_i^*(\widehat{\gamma})\}\right],$$

and the partial derivatives of this as e.g. $U^{\phi_1\gamma} \equiv \partial U^{\phi_1}/\partial\gamma$, we can expand this around $(\phi_{10},\gamma_0,\nu_0,\alpha_0)$, where $\nu_0 \equiv \nu_0(\gamma_0)$ and $\alpha_0 \equiv \alpha_0(\gamma_0)$ are the limiting values of the nuisance parameters, as

$$\begin{aligned} \frac{1}{n}U^{\phi_1}\{\phi_1,\widehat{\gamma},\widehat{\nu}(\widehat{\gamma}),\widehat{\alpha}(\widehat{\gamma})\} &\approx \frac{1}{n}U^{\phi_1}(\phi_{10},\gamma_0,\nu_0,\alpha_0) + \mathrm{E}\{U_i^{\phi_1\gamma}(\phi_{10},\gamma_0,\nu_0,\alpha_0)\}(\widehat{\gamma}-\gamma_0) \\ &\quad+ \mathrm{E}\{U_i^{\phi_1\nu}(\phi_{10},\gamma_0,\nu_0,\alpha_0)\}(\widehat{\nu}-\gamma_0) \\ &\quad+ \mathrm{E}\{U_i^{\phi_1\alpha}(\phi_{10},\gamma_0,\nu_0,\alpha_0)\}(\widehat{\alpha}-\gamma_0) \\ &= \frac{1}{n}U^{\phi_1}(\phi_{10},\gamma_0,\nu_0,\alpha_0) + \mathrm{E}\{U_i^{\phi_1\gamma}(\phi_{10},\gamma_0,\nu_0,\alpha_0)\}(\widehat{\gamma}-\gamma_0) \\ &\approx \frac{1}{n}U^{\phi_1}(\phi_1,\widehat{\gamma},\nu_0,\alpha_0), \end{aligned}$$

since here $\mathrm{E}\{U_i^{\phi_1\nu}(\phi_{10},\gamma_0,\nu_0,\alpha_0)\} = \mathrm{E}\{U_i^{\phi_1\alpha}(\phi_{10},\gamma_0,\nu_0,\alpha_0)\} = 0$. We can now see that the Theorem 1 of Henmi and Eguchi (2004, p. 935) applies to the last form here, implying that $\mathrm{avar}(\widehat{\phi}_1) \leq \mathrm{avar}(\tilde{\phi}_1)$, where $\widehat{\phi}_1$ is the solution to $U(\phi_1,\widehat{\gamma},\nu_0,\alpha_0) = 0$ and $\tilde{\phi}_1$ is the solution to $U(\phi_1,\gamma_0,\nu_0,\alpha_0) = 0$.

S3. THE DOUBLY ROBUST ESTIMATOR AS A POSTERIOR PREDICTIVE MEAN DIFFERENCE

We first note that because

$$
\begin{aligned}
&\int_{v_i} \mathrm{I}_{\{z_i=a\}} y_i p_{\mathcal{E}}(v_i \mid v;\xi)\,\mathrm{d}v_i \\
&\qquad = \int_{y_i,x_i} y_i p_{\mathcal{E}}(y_i \mid Z_i = a, x_i, v;\xi)\mathrm{pr}_{\mathcal{E}}(Z_i = a) p_{\mathcal{E}}(x_i \mid v;\xi)\,\mathrm{d}y_i\,\mathrm{d}x_i,
\end{aligned}
$$

we have that

$$
\mathrm{E}_{\mathcal{E}}(Y_i \mid Z_i = 1, v;\xi) = \mathrm{E}_{\mathcal{E}}\left\{ \frac{Z_i Y_i}{\mathrm{pr}_{\mathcal{E}}(Z_i = 1)} \,\middle|\, v;\xi \right\}
$$

and

$$
\mathrm{E}_{\mathcal{E}}(Y_i \mid Z_i = 0, v;\xi) = \mathrm{E}_{\mathcal{E}}\left\{ \frac{(1-Z_i) Y_i}{\mathrm{pr}_{\mathcal{E}}(Z_i = 0)} \,\middle|\, v;\xi \right\}.
$$

The usual IPT-weighted estimator for a marginal causal contrast may be derived through a posterior predictive argument as follows. First,

$$
\begin{aligned}
\mathrm{E}_{\mathcal{E}}[Z_i Y_i \mid v;\xi] &= \int_{v_i} z_i y_i p_{\mathcal{E}}(v_i \mid v;\xi)\,\mathrm{d}v_i \\
&= \int_{v_i} z_i y_i \frac{p_{\mathcal{E}}(v_i \mid v;\xi)}{p_{\mathcal{O}}(v_i \mid v;\xi)} p_{\mathcal{O}}(v_i \mid v;\xi)\,\mathrm{d}v_i \\
&= \int_{v_i} z_i y_i \frac{p_{\mathcal{E}}(y_i \mid z_i, x_i, v;\xi) p_{\mathcal{E}}(z_i) p_{\mathcal{E}}(x_i \mid v,\xi)}{p_{\mathcal{O}}(y_i \mid z_i, x_i, v;\xi) p_{\mathcal{O}}(z_i \mid x_i, v;\xi) p_{\mathcal{O}}(x_i \mid v;\xi)} p_{\mathcal{O}}(v_i \mid v;\xi)\,\mathrm{d}v_i \\
&= \int_{v_i} z_i y_i \frac{p_{\mathcal{E}}(z_i)}{p_{\mathcal{O}}(z_i \mid x_i, v;\xi)} p_{\mathcal{O}}(v_i \mid v;\xi)\,\mathrm{d}v_i \qquad (20)\\
&= \int_{v_i} z_i y_i \frac{p_{\mathcal{E}}(z_i)}{p_{\mathcal{O}}(z_i \mid x_i, v;\xi)} \sum_{k=1}^{n} \xi_k \delta_{v_k}(v_i)\,\mathrm{d}v_i \\
&= \sum_{k=1}^{n} \xi_k z_k y_k \frac{p_{\mathcal{E}}(z_k)}{p_{\mathcal{O}}(z_k \mid x_k, v;\xi)} \\
&= \mathrm{pr}_{\mathcal{E}}(Z_k = 1) \sum_{k=1}^{n} \xi_k \frac{z_k y_k}{\mathrm{pr}_{\mathcal{O}}(Z_k = 1 \mid x_k, v;\xi)},
\end{aligned}
$$

and thus,

$$
\mathrm{E}_{\mathcal{E}}\left\{ \frac{Z_i Y_i}{\mathrm{pr}_{\mathcal{E}}(Z_i = 1)} \,\middle|\, v;\xi \right\} = \sum_{k=1}^{n} \xi_k \frac{z_k y_k}{\mathrm{pr}_{\mathcal{O}}(Z_k = 1 \mid x_k, v;\xi)}.
$$

Similarly,

$$
\mathrm{E}_{\mathcal{E}}\left\{ \frac{(1-Z_i) Y_i}{\mathrm{pr}_{\mathcal{E}}(Z_i = 0)} \,\middle|\, v;\xi \right\} = \sum_{k=1}^{n} \xi_k \frac{(1-z_k) y_k}{\mathrm{pr}_{\mathcal{O}}(Z_k = 0 \mid x_k, v;\xi)},
$$

and

$$
\begin{aligned}
&\mathrm{E}_{\mathcal{E}}\left\{ \frac{Z_i Y_i}{\mathrm{pr}_{\mathcal{E}}(Z_i = 1)} - \frac{(1-Z_i) Y_i}{\mathrm{pr}_{\mathcal{E}}(Z_i = 0)} \,\middle|\, v;\xi \right\} \\
&\qquad = \sum_{k=1}^{n} \xi_k y_k \left\{ \frac{z_k}{\mathrm{pr}_{\mathcal{O}}(Z_k = 1 \mid x_k, v;\xi)} - \frac{1-z_k}{\mathrm{pr}_{\mathcal{O}}(Z_k = 0 \mid x_k, v;\xi)} \right\}.
\end{aligned}
$$

On the other hand, the usual outcome model based estimator may be motivated similarly as in § S1 through

$$\begin{aligned}
\mathrm{E}_{\mathcal{E}}(Y_i \mid Z_i = a, v; \xi) &= \int_{x_i} \left\{ \int_{y_i} y_i p_{\mathcal{O}}(y_i \mid Z_i = a, x_i, v; \xi) \, \mathrm{d}y_i \right\} p_{\mathcal{O}}(x_i \mid v; \xi) \, \mathrm{d}x_i \\
&= \int_{x_i} m(a, x_i; \xi) \sum_{k=1}^{n} \xi_k \delta_{x_k}(x_i) \, \mathrm{d}x_i \\
&= \sum_{k=1}^{n} \xi_k m(a, x_k; \xi),
\end{aligned}$$

and

$$\mathrm{E}_{\mathcal{E}}(Y_i \mid Z_i = 1, v; \xi) - \mathrm{E}_{\mathcal{E}}(Y_i \mid Z_i = 0, v; \xi) = \sum_{k=1}^{n} \xi_k \left\{ m(1, x_k; \xi) - m(0, x_k; \xi) \right\}.$$

Finally, we note that we can write (20) alternatively as

$$\begin{aligned}
&\mathrm{E}_{\mathcal{E}}(Z_i Y_i \mid v; \xi) \\
&\quad = \int_{v_i} z_i y_i \frac{p_{\mathcal{E}}(z_i)}{p_{\mathcal{O}}(z_i \mid x_i, v; \xi)} p_{\mathcal{O}}(v_i \mid v; \xi) \, \mathrm{d}v_i \\
&\quad = \int_{y_i, z_i, x_i} z_i y_i \frac{p_{\mathcal{E}}(z_i)}{p_{\mathcal{O}}(z_i \mid x_i, v; \xi)} p_{\mathcal{O}}(y_i \mid z_i, x_i, v; \xi) p_{\mathcal{O}}(z_i, x_i \mid v; \xi) \, \mathrm{d}y_i \, \mathrm{d}z_i \, \mathrm{d}x_i \\
&\quad = \int_{z_i, x_i} z_i m(z_i, x_i; \xi) \frac{p_{\mathcal{E}}(z_i)}{p_{\mathcal{O}}(z_i \mid x_i, v; \xi)} \sum_{k=1}^{n} \xi_k \delta_{(z_k, x_k)}(z_i, x_i) \, \mathrm{d}z_i \, \mathrm{d}x_i \\
&\quad = \sum_{k=1}^{n} \xi_k z_k m(z_k, x_k; \xi) \frac{p_{\mathcal{E}}(z_k)}{p_{\mathcal{O}}(z_k \mid x_k, v; \xi)} \\
&\quad = \mathrm{pr}_{\mathcal{E}}(Z_k = 1) \sum_{k=1}^{n} \xi_k \frac{z_k m(z_k, x_k; \xi)}{\mathrm{pr}_{\mathcal{O}}(Z_k = 1 \mid x_k, v; \xi)},
\end{aligned}$$

and therefore

$$\begin{aligned}
&\mathrm{E}_{\mathcal{E}} \left\{ \frac{Z_i Y_i}{\mathrm{pr}_{\mathcal{E}}(Z_i = 1)} - \frac{(1 - Z_i) Y_i}{\mathrm{pr}_{\mathcal{E}}(Z_i = 0)} \,\middle|\, v; \xi \right\} \\
&\quad = \sum_{k=1}^{n} \xi_k m(z_k, x_k; \xi) \left\{ \frac{z_k}{\mathrm{pr}_{\mathcal{O}}(Z_k = 1 \mid x_k, v; \xi)} - \frac{1 - z_k}{\mathrm{pr}_{\mathcal{O}}(Z_k = 0 \mid x_k, v; \xi)} \right\}.
\end{aligned}$$

Thus, the posterior predictive mean difference can be written as

$$\begin{aligned}
&\mathrm{E}_{\mathcal{E}}(Y_i \mid Z_i = 1, v; \xi) - \mathrm{E}_{\mathcal{E}}(Y_i \mid Z_i = 0, v; \xi) \\
&\quad = \mathrm{E}_{\mathcal{E}} \left\{ \frac{Z_i Y_i}{\mathrm{pr}_{\mathcal{E}}(Z_i = 1)} - \frac{(1 - Z_i) Y_i}{\mathrm{pr}_{\mathcal{E}}(Z_i = 0)} \,\middle|\, v; \xi \right\} + \mathrm{E}_{\mathcal{E}}(Y_i \mid Z_i = 1, v; \xi) \\
&\qquad - \mathrm{E}_{\mathcal{E}} \left\{ \frac{Z_i Y_i}{\mathrm{pr}_{\mathcal{E}}(Z_i = 1)} - \frac{(1 - Z_i) Y_i}{\mathrm{pr}_{\mathcal{E}}(Z_i = 0)} \,\middle|\, v; \xi \right\} - \mathrm{E}_{\mathcal{E}}(Y_i \mid Z_i = 0, v; \xi) \\
&\quad = \sum_{k=1}^{n} \xi_k \{ y_k - m(z_k, x_k; \xi) \} \left\{ \frac{z_k}{\mathrm{pr}_{\mathcal{O}}(Z_k = 1 \mid x_k, v; \xi)} - \frac{1 - z_k}{\mathrm{pr}_{\mathcal{O}}(Z_k = 0 \mid x_k, v; \xi)} \right\} \\
&\qquad + \sum_{k=1}^{n} \xi_k \left\{ m(1, x_k; \xi) - m(0, x_k; \xi) \right\}. \qquad (21)
\end{aligned}$$

S4. PROOFS TO THEOREMS

Proof (to Theorem 1). If $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp X_i \mid S_i$, then also $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp e(B_i) \mid S_i$, and the propensity score adjustment does not add information. If on the other hand $Z_i \perp\!\!\!\perp X_i \mid B_i$ holds true, $\{e(B_i), S_i\}$

has jointly the balancing property $Z_i \perp\!\!\!\perp X_i \mid \{e(B_i), S_i\}$. This follows from Theorem 2 of Rosenbaum and Rubin (1983, p. 44) and also implies that $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp Z_i \mid \{e(B_i), S_i\}$ (Rosenbaum and Rubin, 1983, Theorem 3). Now

$$\begin{aligned}\mathrm{E}\{Y_i \mid Z_i, e(B_i), S_i\} &= \mathrm{E}\{(1 - Z_i)Y_{0i} \mid Z_i, e(B_i), S_i\} + \mathrm{E}\{Z_i Y_{1i} \mid Z_i, e(B_i), S_i\}\\ &= (1 - Z_i)\mathrm{E}\{Y_{0i} \mid Z_i, e(B_i), S_i\} + Z_i\mathrm{E}\{Y_{1i} \mid Z_i, e(B_i), S_i\}\\ &= (1 - Z_i)\mathrm{E}\{Y_{0i} \mid e(B_i), S_i\} + Z_i\mathrm{E}\{Y_{1i} \mid e(B_i), S_i\},\end{aligned}$$

and further,

$$\begin{aligned}&\int_{x_i} \mathrm{E}\{Y_i \mid Z_i = 1, e(b_i), s_i\}p(x_i)\,\mathrm{d}x_i - \int_{x_i} \mathrm{E}\{Y_i \mid Z_i = 0, e(b_i), s_i\}p(x_i)\,\mathrm{d}x_i \qquad (22)\\ &\quad = \int_{x_i} \mathrm{E}\{Y_{1i} \mid e(b_i), s_i\}p(x_i)\,\mathrm{d}x_i - \int_{x_i} \mathrm{E}\{Y_{0i} \mid e(b_i), s_i\}p(x_i)\,\mathrm{d}x_i\\ &\quad = \mathrm{E}(Y_{1i}) - \mathrm{E}(Y_{0i}).\end{aligned}$$

The consistency of the estimator (3) then relies on being able to consistently estimate the quantities in (22). □

Proof (to Theorem 2). Consider first the expectation of the estimating equation of (8) under the assumption that $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp X_i \mid S_i$. Now

$$\begin{aligned}&\mathrm{E}\left\{\mathrm{I}_{\{z_i=a\}}\frac{\frac{\partial}{\partial\phi}\log p(y_{ai} \mid s_i, \phi)}{\mathrm{pr}(Z_i = a \mid b_i)}\right\}\\ &= \int_{x_i}\int_{y_{ai}}\sum_{z_i}\mathrm{I}_{\{z_i=a\}}\frac{\frac{\partial}{\partial\phi}\log p(y_{ai} \mid s_i, \phi)}{\mathrm{pr}(Z_i = a \mid b_i)} f(y_{ai} \mid z_i, x_i)f(z_i \mid x_i)f(x_i)\,\mathrm{d}y_{ai}\,\mathrm{d}x_i\\ &= \int_{x_i}\left\{\int_{y_{ai}}\frac{\frac{\partial}{\partial\phi}p(y_{ai} \mid s_i; \phi)}{p(y_{ai} \mid s_i; \phi)} f(y_{ai} \mid x_i)\,\mathrm{d}y_{ai}\right\}\frac{\mathrm{pr}(Z_i = a \mid x_i)}{\mathrm{pr}(Z_i = a \mid b_i)} f(x_i)\,\mathrm{d}x_i\\ &= \int_{x_i}\left\{\frac{\partial}{\partial\phi}\int_{y_{ai}} p(y_{ai} \mid s_i; \phi)\,\mathrm{d}y_{ai}\right\}\frac{\mathrm{pr}(Z_i = a \mid x_i)}{\mathrm{pr}(Z_i = a \mid b_i)} f(x_i)\,\mathrm{d}x_i = 0,\end{aligned}$$

which followed because now $p(y_{ai} \mid s_i; \phi) = f(y_{ai} \mid x_i)$ at the true parameter value. Thus, the misspecified weights do not influence the estimation (in terms of bias) as long as the outcome model is correctly specified.

Under the assumption that $Z_i \perp\!\!\!\perp X_i \mid B_i$, we have in turn that

$$\begin{aligned}&\mathrm{E}\left\{\mathrm{I}_{\{z_i=a\}}\frac{\frac{\partial}{\partial\phi}\log p(y_{ai} \mid s_i; \phi)}{\mathrm{pr}(Z_i = a \mid b_i)}\right\}\\ &= \int_{x_i}\int_{y_{ai}}\sum_{z_i}\mathrm{I}_{\{z_i=a\}}\frac{\frac{\partial}{\partial\phi}\log p(y_{ai} \mid s_i; \phi)}{\mathrm{pr}(Z_i = a \mid b_i)} f(y_{ai} \mid z_i, x_i)f(z_i \mid x_i)f(x_i)\,\mathrm{d}y_{ai}\,\mathrm{d}x_i\\ &= \int_{x_i}\int_{y_{ai}}\frac{\partial}{\partial\phi}\log p(y_{ai} \mid s_i; \phi) f(y_{ai} \mid x_i)f(x_i)\,\mathrm{d}y_{ai}\,\mathrm{d}x_i,\end{aligned}$$

since now $p(z_i \mid b_i) = f(z_i \mid x_i)$. Using the partitioning $x_i = (s_i, r_i)$, we can write the last form in above as

$$\begin{aligned}&\int_{x_i}\int_{y_{ai}}\frac{\partial}{\partial\phi}\log p(y_{ai} \mid s_i; \phi) f(y_{ai} \mid x_i)f(x_i)\,\mathrm{d}y_{ai}\,\mathrm{d}x_i\\ &= \int_{s_i}\int_{r_i}\int_{y_{ai}}\frac{\partial}{\partial\phi}\log p(y_{ai} \mid s_i; \phi) f(y_{ai}, s_i, r_i)\,\mathrm{d}y_{ai}\,\mathrm{d}s_i\,\mathrm{d}r_i\\ &= \int_{s_i}\int_{y_{ai}}\frac{\frac{\partial}{\partial\phi}p(y_{ai} \mid s_i; \phi)}{p(y_{ai} \mid s_i; \phi)} f(y_{ai} \mid s_i)f(s_i)\,\mathrm{d}y_{ai}\,\mathrm{d}s_i\\ &= \int_{s_i}\left\{\frac{\partial}{\partial\phi}\int_{y_{ai}} p(y_{ai} \mid s_i; \phi)\,\mathrm{d}y_{ai}\right\} f(s_i)\,\mathrm{d}s_i = 0.\end{aligned}$$

Thus, even though the outcome regression does not include a sufficient set of confounders, through the IPT weighting we can still obtain valid estimates for the conditional distributions $p(y_{ai} \mid s_i; \phi)$. $\square$

Proof (to theorem 3). Write $\pi_n(\phi \mid \gamma) = \pi_n(\phi, \gamma)/\pi_n(\gamma)$. Here

$$\pi_n(\phi, \gamma) = \frac{\prod_{i=1}^n \left[p\{y_i \mid z_i, e(b_i; \gamma), s_i; \phi\} p(z_i \mid b_i; \gamma)\right] \pi_0(\phi)\pi_0(\gamma)}{\int_\phi \int_\gamma \prod_{i=1}^n \left[p\{y_i \mid z_i, e(b_i; \gamma), s_i; \phi\} p(z_i \mid b_i; \gamma)\right] \pi_0(\phi)\pi_0(\gamma) \, \mathrm{d}\phi \, \mathrm{d}\gamma}$$

and

$$\begin{aligned} \pi_n(\gamma) &= \int_\phi \pi_n(\phi, \gamma) \, \mathrm{d}\phi \\ &= \frac{\int_\phi \prod_{i=1}^n \left[p\{y_i \mid z_i, e(b_i; \gamma), s_i; \phi\} p(z_i \mid b_i; \gamma)\right] \pi_0(\phi)\pi_0(\gamma) \, \mathrm{d}\phi}{\int_\phi \int_\gamma \prod_{i=1}^n \left[p\{y_i \mid z_i, e(b_i; \gamma), s_i; \phi\} p(z_i \mid b_i; \gamma)\right] \pi_0(\phi)\pi_0(\gamma) \, \mathrm{d}\phi \, \mathrm{d}\gamma}. \end{aligned}$$

We note that generally $\pi_n(\gamma)$ will not be proportional to $\prod_{i=1}^n p(z_i \mid b_i; \gamma)\pi_0(\gamma)$. However, assume now that $Y_{ai} \perp\!\!\!\perp X_i \mid S_i$. Because $Y_{ai} \perp\!\!\!\perp Z_i \mid X_i$, it follows that $Y_{ai} \perp\!\!\!\perp X_i \mid (Z_i, S_i)$ and $Y_i \perp\!\!\!\perp X_i \mid (Z_i, S_i)$. Thus, the posteriors factorize as

$$\pi_n(\phi, \gamma) = \frac{\prod_{i=1}^n p\{y_i \mid z_i, s_i; \phi\}\pi_0(\phi)}{\int_\phi \prod_{i=1}^n p\{y_i \mid z_i, s_i; \phi\}\pi_0(\phi) \, \mathrm{d}\phi} \frac{\prod_{i=1}^n p(z_i \mid b_i; \gamma)\pi_0(\gamma)}{\int_\gamma \prod_{i=1}^n p(z_i \mid b_i; \gamma)\pi_0(\gamma) \, \mathrm{d}\gamma}$$

and

$$\pi_n(\gamma) = \frac{\int_\phi \prod_{i=1}^n p(y_i \mid z_i, s_i; \phi)\pi_0(\phi) \, \mathrm{d}\phi}{\int_\phi \prod_{i=1}^n p(y_i \mid z_i, s_i; \phi)\pi_0(\phi) \, \mathrm{d}\phi} \frac{\prod_{i=1}^n p(z_i \mid b_i; \gamma)\pi_0(\gamma)}{\int_\gamma \prod_{i=1}^n p(z_i \mid b_i; \gamma)\pi_0(\gamma) \, \mathrm{d}\gamma}. \qquad \square$$

Therefore, $\pi_n(\phi \mid \gamma) = \pi_n(\phi) \propto \prod_{i=1}^n p(y_i \mid z_i, s_i; \phi)\pi_0(\phi)$ and $\pi_n(\gamma) \propto \prod_{i=1}^n p(z_i \mid b_i; \gamma)\pi_0(\gamma)$.

Proof (to Theorem 4). The expression obtained through substituting in the parametric models into (16) is

$$\begin{aligned} &\sum_{k=1}^n \xi_k \left[y_k - m\{z_k, s_k; \widehat{\phi}(\xi)\}\right] \left[\frac{z_k}{\mathrm{pr}_{\mathcal{O}}\{Z_k = 1 \mid b_k; \widehat{\gamma}(\xi)\}} - \frac{1 - z_k}{\mathrm{pr}_{\mathcal{O}}\{Z_k = 0 \mid b_k; \widehat{\gamma}(\xi)\}}\right] \\ &\quad + \sum_{k=1}^n \xi_k \left[m\{1, s_k; \widehat{\phi}(\xi)\} - m\{0, s_k; \widehat{\phi}(\xi)\}\right]. \end{aligned} \tag{23}$$

First, if the outcome model is correctly specified in the sense that $m\{z_k, s_k; \widehat{\phi}(\xi)\} = m(z_k, x_k; \xi)$, we get

$$\begin{aligned} &\mathrm{pr}_{\mathcal{E}}(Z_k = a) \sum_{k=1}^n \xi_k \frac{\mathrm{I}_{\{z_k = a\}} m\{z_k, s_k; \widehat{\phi}(\xi)\}}{\mathrm{pr}_{\mathcal{O}}\{Z_k = a \mid b_k; \widehat{\gamma}(\xi)\}} \\ &\quad = \sum_{k=1}^n \xi_k \mathrm{I}_{\{z_k = a\}} m\{z_k, s_k; \widehat{\phi}(\xi)\} \frac{p_{\mathcal{E}}(z_k)}{p_{\mathcal{O}}\{z_k \mid b_k; \widehat{\gamma}(\xi)\}} \\ &\quad = \int_{z_i, x_i} \mathrm{I}_{\{z_i = a\}} m\{z_i, s_i; \widehat{\phi}(\xi)\} \frac{p_{\mathcal{E}}(z_i)}{p_{\mathcal{O}}\{z_i \mid b_i; \widehat{\gamma}(\xi)\}} \sum_{k=1}^n \xi_k \delta_{(z_k, x_k)}(z_i, x_i) \, \mathrm{d}z_i \, \mathrm{d}x_i \\ &\quad = \int_{y_i, z_i, x_i} \mathrm{I}_{\{z_i = a\}} y_i \frac{p_{\mathcal{E}}(z_i)}{p_{\mathcal{O}}\{z_i \mid b_i; \widehat{\gamma}(\xi)\}} p_{\mathcal{O}}(y_i \mid z_i, x_i, v; \xi) p_{\mathcal{O}}(z_i, x_i \mid v; \xi) \, \mathrm{d}y_i \, \mathrm{d}z_i \, \mathrm{d}x_i \\ &\quad = \int_{v_i} \mathrm{I}_{\{z_i = a\}} y_i \frac{p_{\mathcal{E}}(z_i)}{p_{\mathcal{O}}\{z_i \mid b_i; \widehat{\gamma}(\xi)\}} p_{\mathcal{O}}(v_i \mid v; \xi) \, \mathrm{d}v_i \\ &\quad = \mathrm{pr}_{\mathcal{E}}(Z_k = a) \sum_{k=1}^n \xi_k \frac{\mathrm{I}_{\{z_k = a\}} y_k}{\mathrm{pr}_{\mathcal{O}}\{Z_k = a \mid b_k; \widehat{\gamma}(\xi)\}}, \end{aligned}$$

because the second to last form is equivalent to (20). Thus, the first summation term in (23) cancels out, leaving only the model based mean difference, which itself is equivalent to $\mathrm{E}_{\mathcal{E}}(Y_i \mid Z_i = 1, v; \xi) - \mathrm{E}_{\mathcal{E}}(Y_i \mid Z_i = 0, v; \xi)$.

On the other hand, if the treatment assignment model is correctly specified in the sense that $\mathrm{pr}_{\mathcal{O}}\{Z_k = a \mid b_k; \widehat{\gamma}(\xi)\} = \mathrm{pr}_{\mathcal{O}}(Z_k = a \mid x_k, v; \xi)$, we get

$$\begin{aligned}
&\mathrm{pr}_{\mathcal{E}}(Z_k = a) \sum_{k=1}^{n} \xi_k \frac{\mathrm{I}_{\{z_k = a\}} m\{z_k, s_k; \widehat{\phi}(\xi)\}}{\mathrm{pr}_{\mathcal{O}}\{Z_k = a \mid b_k; \widehat{\gamma}(\xi)\}} \\
&\quad = \sum_{k=1}^{n} \xi_k \mathrm{I}_{\{z_k = a\}} m\{z_k, s_k; \widehat{\phi}(\xi)\} \frac{p_{\mathcal{E}}(z_k)}{p_{\mathcal{O}}\{z_k \mid b_k; \widehat{\gamma}(\xi)\}} \\
&\quad = \int_{z_i, x_i} \mathrm{I}_{\{z_i = a\}} m\{z_i, s_i; \widehat{\phi}(\xi)\} \frac{p_{\mathcal{E}}(z_i)}{p_{\mathcal{O}}\{z_i \mid b_i; \widehat{\gamma}(\xi)\}} \sum_{k=1}^{n} \xi_k \delta_{(z_k, x_k)}(z_i, x_i) \,\mathrm{d}z_i \,\mathrm{d}x_i \\
&\quad = \int_{z_i, x_i} \mathrm{I}_{\{z_i = a\}} m\{z_i, s_i; \widehat{\phi}(\xi)\} \frac{p_{\mathcal{E}}(z_i)}{p_{\mathcal{O}}\{z_i \mid b_i; \widehat{\gamma}(\xi)\}} p_{\mathcal{O}}(z_i \mid x_i, v; \xi) p_{\mathcal{O}}(x_i \mid v; \xi) \,\mathrm{d}z_i \,\mathrm{d}x_i \\
&\quad = \int_{z_i, x_i} \mathrm{I}_{\{z_i = a\}} m\{z_i, s_i; \widehat{\phi}(\xi)\} p_{\mathcal{E}}(z_i) p_{\mathcal{O}}(x_i \mid v; \xi) \,\mathrm{d}z_i \,\mathrm{d}x_i \\
&\quad = \mathrm{pr}_{\mathcal{E}}(Z_i = a) \int_{x_i} m\{a, s_i; \widehat{\phi}(\xi)\} \sum_{k=1}^{n} \xi_k \delta_{(x_k)}(x_i) \,\mathrm{d}x_i \\
&\quad = \mathrm{pr}_{\mathcal{E}}(Z_k = a) \sum_{k=1}^{n} \xi_k m\{a, s_k; \widehat{\phi}(\xi)\}.
\end{aligned}$$

Therefore, expression (23) now reduces to

$$\sum_{k=1}^{n} \xi_k y_i \left[\frac{z_k}{\mathrm{pr}_{\mathcal{O}}\{Z_k = 1 \mid b_k; \widehat{\gamma}(\xi)\}} - \frac{1 - z_k}{\mathrm{pr}_{\mathcal{O}}\{Z_k = 0 \mid b_k; \widehat{\gamma}(\xi)\}} \right],$$

which is again equivalent to $\mathrm{E}_{\mathcal{E}}(Y_i \mid Z_i = 1, v; \xi) - \mathrm{E}_{\mathcal{E}}(Y_i \mid Z_i = 0, v; \xi)$; see § S3. □

REFERENCES

An, W. (2010). Bayesian propensity score estimators: incorporating uncertainties in propensity scores into causal inference. *Sociological Methodology*, 40:151–189.

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

Chakraborty, B. and Moodie, E. E. M. (2013). *Statistical Methods for Dynamic Treatment Regimes*. Springer-Verlag, New York.

Chamberlain, G. and Imbens, G. W. (2003). Nonparametric applications of Bayesian inference. *Journal of Business & Economic Statistics*, 21:12–18.

Chen, J. and Kaplan, D. (2015). Covariate balance in Bayesian propensity score approaches for observational studies. *Journal of Research on Educational Effectiveness*, 8:280–302.

Cole, S. R. and Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168:656–664.

Dawid, A. P. and Didelez, V. (2010). Identifying the consequences of dynamic treatment strategies: a decision-theoretic overview. *Statistical Surveys*, 4:184–231.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7:1–26.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC, Boca Raton, FL.

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

Gustafson, P. (2012). Double-robust estimators: slightly more Bayesian than meets the eye? *The International Journal of Biostatistics*, 8:Issue 2, Article 4.

Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, 95:481–488.

Henmi, M. and Eguchi, S. (2004). A paradox concerning nuisance parameters and projected estimating functions. *Biometrika*, 91:929–941.

Hernán, M. A., Brumback, B., and Robins, J. M. (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association*, 96:440–448.

Hernán, M. A. and Robins, J. M. (2006). Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health*, 60:578–586.

Hoshino, A. (2008). A Bayesian propensity score adjustment for latent variable modeling and MCMC algorithm. *Computational Statistics & Data Analysis*, 52:1413–1429.

Ionides, E. L. (2008). Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17:295–311.

Kaplan, D. and Chen, J. (2012). Two-step Bayesian approach for propensity score analysis: simulations and case study. *Psychometrika*, 77:581–609.

Lunn, D., Best, N., Spiegelhalter, D., Graham, G., and Neuenschwander, B. (2009). Combining MCMC with "sequential" PKPD modelling. *Journal of Pharmacokinetics and Pharmacodynamics*, 36:19–39.

McCandless, L. C., Douglas, I. J., Evans, S. J., and Smeeth, L. (2010). Cutting feedback in Bayesian regression adjustment for the propensity score. *The International Journal of Biostatistics*, 6:Issue 2, Article 16.

McCandless, L. C., Gustafson, P., and Austin, P. C. (2009a). Bayesian propensity score analysis for observational data. *Statistics in Medicine*, 28:94–112.

McCandless, L. C., Gustafson, P., Austin, P. C., and Levy, A. R. (2009b). Covariate balance in a Bayesian propensity score analysis of beta blocker therapy in heart failure patients. *Epidemiologic Perspectives & Innovations*, 6:Article 5.

McCandless, L. C., Richardson, S., and Best, N. (2012). Adjustment for missing confounders using external validation data and propensity scores. *Journal of the American Statistical Association*, 107:40–51.

Newton, M. A. and Raftery, A. E. (1994). Approximating Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society, Series B*, 56:3–48. With discussion and a reply by the authors.

Robins, J. M., Hernán, M. Á., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11:550–560.

Robins, J. M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine*, 16:285–319.

Rose, S. and van der Laan, M. J. (2008). Simple optimal weighting of cases and controls in case-control studies. *The International Journal of Biostatistics*, 4:Issue 1, Article 19.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 6:41–55.

Røysland, K. (2011). A martingale approach to continuous-time marginal structural models. *Bernoulli*, 17:895–915.

Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics*, 9:130–134.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91:473–489.

Saarela, O., Arjas, E., Stephens, D. A., and Moodie, E. E. M. (2015a). Predictive Bayesian inference and dynamic treatment regimes. *Biometrical Journal*, 57:941–958.

Saarela, O., Moodie, E. E. M., Stephens, D. A., and Klein, M. B. (2015b). On Bayesian estimation of marginal structural models. *Biometrics*, 71:279–288.

Saarela, O., Moodie, E. E. M., Stephens, D. A., and Klein, M. B. (2015c). Rejoinder: On Bayesian estimation of marginal structural models. *Biometrics*, 71:299–301.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94:1096–1146.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, New York, NY.

Walker, S. G. (2010). Bayesian nonparametric methods: motivation and ideas. In Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G., editors, *Bayesian Nonparametrics*, pages 22–34. Cambridge University Press, Cambridge, UK.

Wang, C., Parmigiani, G., and Dominici, F. (2012). Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics*, 68:661–671.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 115:1–25.

Xiao, Y., Moodie, E. E. M., and Abrahamowicz, M. (2013). Comparison of approaches to weight truncation for marginal structural Cox models. *Epidemiologic Methods*, 2:1–20.

Zhang, G. and Little, R. (2009). Extensions of the penalized spline of propensity prediction method of imputation. *Biometrics*, 65:911–918.

Zigler, C. M. and Dominici, F. (2014). Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects. *Journal of the American Statistical Association*, 109(505):95–107.

Zigler, C. M., Watts, K., Yeh, R. W., Wang, Y., Coull, B. A., and Dominici, F. (2013). Model feedback in Bayesian propensity score estimation. *Biometrics*, 69:263–273.