Bayesian Analysis (2016) **11**, Number 1, pp. 47–69

Approximate Bayesian Inference for Doubly Robust Estimation

Daniel J. Graham[*], Emma J. McCoy[†], and David A. Stephens[‡]

Abstract. Doubly robust estimators are typically constructed by combining outcome regression and propensity score models to satisfy moment restrictions that ensure consistent estimation of causal quantities provided at least one of the component models is correctly specified. Standard Bayesian methods are difficult to apply because restricted moment models do not imply fully specified likelihood functions. This paper proposes a Bayesian bootstrap approach to derive approximate posterior predictive distributions that are doubly robust for estimation of causal quantities. Simulations show that the approach performs well under various sources of misspecification of the outcome regression or propensity score models. The estimator is applied in a case study of the effect of area deprivation on the incidence of child pedestrian casualties in British cities.

Keywords: approximate bayes, doubly robust, propensity score, treatment effect.

1 Introduction

Typical targets of inference in causal studies include average potential outcomes (APOs) and average treatment effects (ATEs). The former measure average responses under given treatment regimes and the latter measure differences in average responses under different treatment regimes. A variety of approaches can be used to estimate such quantities including those that proceed via specification of an outcome regression (OR) model or a propensity score (PS) model. Doubly robust (DR) approaches combine both OR and PS models, often via PS weighting or augmentation of the OR model, such that valid causal estimates can be obtained when only one of the two models is correctly specified.

Bayesian interest in DR estimation has been subdued. A fully Bayesian interpretation of the DR approach is challenging because the relevant targets of inference ultimately depend only on the outcome model, and simultaneous modelling of the exposure and outcome models is problematic (for a recent discussion see Gustafson, 2012). A further impediment to the use of standard Bayesian methods arises because DR estimators are typically expressed as solutions to estimating equations based on a set of moment restrictions, and restricted moment models do not generally provide fully specified likelihood functions.

[*] Corresponding author: Department of Civil Engineering, Imperial College London, UK, d.j.graham@imperial.ac.uk

[†] Department of Mathematics, Imperial College London, UK, e.mccoy@imperial.ac.uk

[‡] Department of Mathematics and Statistics, McGill University, Montreal, Canada, dstephens@math.mcgill.ca

DOI: 10.1214/14-BA928

In this paper we develop an approximate Bayesian approach for DR estimation via the Bayesian bootstrap (Rubin, 1981). We derive approximate Bayesian inference for the parameters of the DR model and use these to construct posterior predictive distributions for ATE and APO estimation. We estimate a model for the conditional distribution of treatment given covariates and use the parameter estimates to calculate PS values. Following Scharfstein et al. (1999), we then specify an OR model, augmented with inverse PS covariates that is consistent with the DR moment restrictions. We assume a parametric form for the augmented OR (AOR) model and place an improper Dirichlet prior distribution on the parameters of that model. By repeatedly estimating the AOR model weighted by standardised sets of iid unit exponential random variables, we approximate the posterior distribution of model parameters and use these to form posterior predictive distribution for ATEs and APOs.

While our approach cannot offer the coherent framework for inference that a fully Bayesian analysis would provide, it does still offer two features of Bayesian inference which are particularly useful for causal modelling. First, it provides a natural framework for prediction: estimation of ATEs and APOs necessarily involves prediction over unobserved data and our posterior predictive distributions incorporate randomness originating both from estimation of the parameters of the DR model itself and from the random nature of the unobserved observations used to make predictions. Second, by generating posterior predictive densities, rather than point estimates, we can make probability statements about the causal quantities of interest (i.e. ATEs and APOs) giving greater flexibility in presenting results. For instance, we can discuss findings in relation to specific hypotheses or in terms of credible intervals which can offer a more intuitive understanding of treatment effects for practical interpretation.

The paper is structured as follows. Section 2 provides an overview of DR estimation and explains the principles of inference via the Bayesian bootstrap. Section 3 presents our Bayesian DR model and explains how to approximate samples from the posterior predictive distribution for ATEs or APOs. Simulation results are presented in Section 4. Section 5 applies the Bayesian DR model in a case study of the effect of area deprivation on the incidence of child pedestrian casualties in British cities. Conclusions are drawn in the final section.

2 Doubly robust estimation and the Bayesian bootstrap

2.1 Doubly robust estimation

The principles underpinning DR estimation have been reviewed extensively in the literature (e.g. van der Laan and Robins, 2003; Lunceford and Davidian, 2004; Bang and Robins, 2005; Kang and Schafer, 2007; Tsiatis and Davidian, 2007). Here we provide only a brief summary, and focus on the case of a single exposure and outcome.

In causal inference problems the data available for estimation are realisations of a random vector, $Z_i = (Y_i, D_i, X_i)$, $i = 1, \ldots, n$, where for the ith unit of observation Y_i denotes a response, D_i the treatment (or exposure) received, and X_i a vector of pre-treatment covariates. The treatment can be binary, multi-valued or continuous but crucially it is not assigned randomly. This means that simple comparisons of mean responses across different treatment groups will not in general reveal a 'causal' effect due to potential for confounding.

Confounding can be addressed if the vector of covariates X_i is sufficient to ensure *unconfoundedness*, or conditional independence of potential responses and treatment assignment. In the context of binary treatments, the conditional independence assumption requires that $(Y_i(0), Y_i(1)) \perp\!\!\!\perp I_1(D_i)|X_i$, where $I_1(D_i)$ is the indicator function for receiving the treatment and $Y(1)$ and $Y(0)$ indicate potential outcomes under treated or control status, respectively. For continuous or multi-valued treatments weak conditional independence must hold, which requires that $Y_i(d) \perp\!\!\!\perp I_d(D_i)|X_i$ for all $d \in \mathcal{D}$, where $I_d(D_i)$ is the indicator function for receiving dose d and $Y_i(d)$ is the potential outcome associated with that dose (see Imbens, 2000; Hirano and Imbens, 2004). An additional requirement for valid causal inference is that, conditional on covariates X_i, the probability of assignment to treatment is strictly positive for all x and d. In practice this may hold only within some region of treatment $\mathcal{C} \subseteq \mathcal{D}$, referred to as the common support region. A sufficient condition is that for any subset of $\mathcal{C}$, say $\mathcal{A} \subseteq \mathcal{C}$, $\Pr(D_i \in \mathcal{A}|X_i = x) > 0$ for all x and $\mathcal{A} \subseteq \mathcal{C}$.

In the case of binary treatments, the APOs of interest are $\mu(1) = \mathbb{E}[Y_i(1)]$ and $\mu(0) = \mathbb{E}[Y_i(0)]$, and the ATE is defined as $\tau(1) = \mu(1) - \mu(0)$. For multi-valued or continuous treatments $\mu(d) = \mathbb{E}[Y_i(d)]$ denotes the APO under treatment level d and the ATE is $\tau(d) = \mu(d) - \mu(0)$.

With a covariate vector sufficient to ensure conditional independence several estimators for ATEs are available, but three are of particular interest here. First, we could model the expectation of the conditional density of the response given the covariates and treatment, $\mathbb{E}[Y_i|X_i, D_i]$, using an OR model $\Psi^{-1}\{m(X_i, D_i; \xi)\}$, for known link function Ψ, regression function $m()$, and unknown parameter vector ξ. If the OR model is correctly specified for $\mathbb{E}[Y_i|X_i, D_i]$ it can be used to generate consistent estimates of ATEs. Second, we could assume a model for $f_{D|X}(d|x)$, the conditional density of the treatment given the covariates and use this model to estimate propensity scores, which we denote $\widehat{\pi}(D_i|X_i; \widehat{\alpha})$ with 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. Finally, we could assume both an OR and PS model and construct a DR estimator which yields a consistent estimate of ATEs provided either the OR or PS model is correctly specified.

The DR property requires one of two moment restrictions to hold. Either the OR model $\Psi^{-1}\{m(X_i, D_i; \xi)\}$ consistently estimates $\mathbb{E}[Y_i|D_i, X_i]$; or the PS estimator $\widehat{\pi}(D_i|X_i; \widehat{\alpha})$ is consistent for the true PS, $\mathbb{E}\left[I_d(D_i)|X_i = x\right]$ for each x. These two moment restrictions can be successfully combined for DR estimation by either weighting or augmenting the OR model with a function of the inverse PS values. In this paper we use augmented outcome regression because it allows us to derive an approximate Bayesian DR estimator with relative ease.

The augmented outcome regression (AOR) approach adds 'inverse PS covariates' to the OR model to correct for bias from misspecification. In the binary setting, Scharfstein et al. (1999) include the reciprocal of the covariate

$$h\left(I_1(D_i)|X_i;\alpha\right) = I_1(D_i)\cdot\widehat{\pi}(D_i|X_i;\widehat{\alpha}) + [1-I_1(D_i)]\cdot\{1-\widehat{\pi}(D_i|X_i;\widehat{\alpha})\},$$

and derive a DR ATE estimator as

$$\widehat{\tau}_{DR}(1) = \frac{1}{n}\sum_{i=1}^{n}\left[\Psi^{-1}\left\{m(1,X_i;\widehat{\beta}) + \frac{\widehat{\varphi}}{h\left(1|X_i;\alpha\right)}\right\} - \Psi^{-1}\left\{m(0,X_i;\widehat{\beta}) + \frac{\widehat{\varphi}}{h\left(0|X_i;\alpha\right)}\right\}\right],$$

where φ is the coefficient of the inverse PS covariate. For continuous or multi-valued treatments Graham et al. (2012) include a set of inverse PS covariates to induce bias correction for distinct strata of the treatment. Defining Q, $q=(1,\ldots,Q)$, strata over the range of d, and using $\mathcal{D}_q$ to denote treatment stratum q ($\mathcal{D}_q \subset \mathcal{D} \subseteq \mathbb{R}$), they suggest a point estimate of the mean APO for each treatment stratum using a four step approach:

1. Use the observed data to estimate PS values and form a set of Q inverse PS covariates

$$\frac{I_q(D_i)}{\widehat{\pi}(D_i|X_i;\widehat{\alpha})}$$

where the indicator $I_q(D_i)$ denotes membership of treatment stratum q (i.e. $D_i \in \mathcal{D}_q$).

2. Estimate the AOR model

$$e\left\{D_i,X_i,I_q(D_i);\beta,\varphi\right\} = \Psi^{-1}\left\{m\left(D_i,X_i;\beta\right) + \sum_{q=1}^{Q}\frac{\varphi_q I_q(D_i)}{\widehat{\pi}(D_i|X_i;\widehat{\alpha})}\right\},$$

where $\varphi=(\varphi_1,\ldots,\varphi_Q)$ is a Q dimensional parameter vector for the inverse PS covariates.

3. For each distinct treatment d_{qj}, $j=(1,\ldots,J)$, in stratum q calculate the mean of the predicted values from the AOR model evaluated at d_{qj}:

$$\frac{1}{n}\sum_{i=1}^{n}\Psi^{-1}\left\{m(d_{qj},X_i;\widehat{\beta}) + \frac{\widehat{\varphi}_q}{\widehat{\pi}\left(d_{qj}|X_i;\widehat{\alpha}\right)}\right\}.$$

4. Take the average of these mean predicted values over all J treatment levels in q

$$\widehat{\mu}_{DR}(\mathcal{D}_q) = \frac{1}{J}\sum_{j=1}^{J}\left[\frac{1}{n}\sum_{i=1}^{n}\Psi^{-1}\left\{m(d_{qj},X_i;\widehat{\beta}) + \frac{\widehat{\varphi}_q}{\widehat{\pi}\left(d_{qj}|X_i;\widehat{\alpha}\right)}\right\}\right]$$

to obtain an estimate of the mean APO for treatment stratum q.

The AOR has a DR property when estimates of β and φ are obtained as solutions to estimating equations of the form

$$\sum_{i=1}^{n} \frac{\partial e\left\{D_i, X_i, I_q(D_i); \beta, \varphi\right\}}{\partial\left(\beta^{\mathsf{T}}, \varphi^{\mathsf{T}}\right)} \frac{1}{\phi}\left[Y_i - e\left\{D_i, X_i, I_q(D_i); \beta, \varphi\right\}\right] = 0 \tag{1}$$

where ϕ is a scale parameter in $\mathrm{Var}[Y_i|D_i, X_i, I_q(D_i)]$. The DR bias correction property arises via the Q score equations for the inverse PS covariates for which we have by construction

$$\sum_{i=1}^{n} \frac{I_q(d_i)}{\widehat{\pi}(d_i|x_i;\widehat{\alpha})} \cdot \left[y_i - \Psi^{-1}\left\{m(d_i, x_i; \widehat{\beta}) + \sum_{q=1}^{Q} \frac{\widehat{\varphi}_q I_q(d_i)}{\widehat{\pi}(d_i|x_i;\widehat{\alpha})}\right\}\right] = 0. \tag{2}$$

If the propensity score model is correctly specified, and the assumption of conditional independence holds, this procedure provides asymptotic bias correction for each stratum of the treatment under misspecification of the OR model; if the OR model is correctly specified, the parameter estimates $\widehat{\varphi}_q$ will converge to zero, and inclusion of inverse PS covariates simply adds noise to the AOR model without affecting its consistency properties. If the PS model is correct, but the OR is not, the augmented regression has a bias correction property which allows us to consistently estimate ATEs or APOs (for proofs see Scharfstein et al., 1999; Graham et al., 2012).

Many parametric or nonparametric optimisation routines could yield estimating equations consistent with the form shown in (1). In the parametric setting the score (or quasi-score) equations of Maximum Likelihood Estimation (MLE), Maximum Quasi-Likelihood (MQL), Restricted MLE (REML) for linear mixed models (LMMs), and Penalised Quasi-Likelihood (PQL) for generalised linear mixed models (GLMMs) all provide optimisation solutions consistent with the DR property.

In this paper we focus on the use of maximum likelihood based estimators for DR estimation, although the way we derive our approximate Bayesian inference could be easily adapted for other optimisation routines. The DR estimating equations are solved, and at least one of the moment restrictions satisfied, when the likelihood is maximised but not necessarily otherwise. Consequently, we do not have a fully specified likelihood function since parameter vectors other than that corresponding to the MLE estimates may not produce estimating equations of the form given in (1), and thus may not have the DR property. In the absence of a fully specified likelihood, Bayesian inference via the posterior distribution is not straightforward.

2.2 Model diagnostics

It should be emphasised that DR ATE estimates are asymptotically unbiased if one of the two models is misspecified, but not both. Ideally, we would conduct diagnostic tests for specification of the two component models, but standard goodness-of-fit based diagnostics may not be particularly informative for model selection. The key requirement of causal models is that they provide sufficient adjustment for confounding between treatment and response such that conditional independence holds. The fit of PS or OR models could be improved by conditioning on non-confounding covariates, but doing so can have adverse consequences for estimation of causal parameters in terms of both bias and efficiency (see, for example, Pearl, 2009, 2010). While the assumption of unconfoundedness is typically made in existing applied work, it is essentially untestable. Recent Bayesian work has focussed on sensitivity analysis approaches to adjustment in the presence of missing confounding variables (McCandless et al., 2012), rather than methods for diagnosing the possible influence of such variables.

In this paper we use the following diagnostic approaches for model specification

(i) Comparison of estimated parameters – Robins and Rotnitzky (2001) propose a goodness of fit test based on comparison of the DR, OR and PS weighting estimators. They argue that if the estimate of the parameter of interest from the DR model differs from the PS and OR estimates by much more than can be explained by sampling variation, it indicates that the PS and OR models have both been badly specified and thus all three estimators may be unreliable. If the DR and PS estimates are close, but the OR is not, it indicates that the OR model may be badly specified. Similarly, if the DR and OR estimates are close, but the PS is not, it indicates that the PS model may be badly specified. A formal test for parameter μ can be constructed as follows. Denote the empirical variances of $(\widehat{\mu}_{DR} - \widehat{\mu}_{OR})$ and $(\widehat{\mu}_{DR} - \widehat{\mu}_{PS})$ as s^2_{DR-OR} and s^2_{DR-PS} respectively calculated via bootstrap re-sampling. Then the tests with rejection regions: $|(\widehat{\mu}_{DR} - \widehat{\mu}_{OR})/s_{DR-OR}| > 1.96$ and $|(\widehat{\mu}_{DR} - \widehat{\mu}_{PS})/s_{DR-PS}| > 1.96$ are large sample 0.05 level tests of the null-hypotheses that the OR and PS models are correctly specified (for details see Bang and Robins, 2005). We implement Bayesian versions of these diagnostic checks.

(ii) Test for balancing of the PS – under conditional independence, the PS has a balancing property in that in the binary case $X_i \perp\!\!\!\perp I_1(D_i)|\pi(D_i|X_i;\alpha)$, or for multi-valued or continuous treatments $X_i \perp\!\!\!\perp I_d(D_i)|\pi(D_i|X_i;\alpha)$ – see the appendix. The balancing property of the PS is testable in the observed data, and provides a useful diagnostic which we employ in the case study below.

(iii) Box plots for inverse PS covariates – if estimates of the coefficients of the inverse propensity score covariates (i.e. φ_q) are significantly different from zero, that indicate that the OR model may not provide a universally good model specification over all doses of interest. We examine samples from the posterior distributions for these parameters and construct box-plots to identify obvious misspecification of the OR model.

2.3 Approximate Bayesian inference via the Bayesian bootstrap

We use the Bayesian bootstrap approach introduced by Rubin (1981) to approximate the posterior density of parameters consistent with the DR estimating equations and then derive posterior predictive distributions for ATEs and APOs. The Bayesian bootstrap has been applied previously for likelihood models by Newton and Raftery (1994) and for instrumental variables and quantile regression by Chamberlain and Imbens (2003).

Let data z, with observations z_i, $i = (1, \ldots, n)$, be distributed according to a family of probability distributions that are regular for likelihood inference from the class $\mathcal{P} = \{f(z;\theta), z \in \mathcal{Z}, \theta \in \Omega_\theta\}$, where $\mathcal{Z} = \{z : f(z;\theta) > 0\}$ is the sample space in which the data lie. We define a_k, $k = (1, \ldots, K)$, as the possible discrete values that z can take and $\theta = (\theta_1, \ldots, \theta_K)$ as the associated probabilities for vector $a = (a_1, \ldots, a_K)$ such that $\Pr(Z = a_k|\theta) = \theta_k$, thus effectively treating the data as a sample from a multinomial distribution. If $n_k = \sum_{i=1}^{n} 1(z_i = a_k)$ is the number of observations equal to the kth distinct value of the data, the likelihood for z is

$$L(\theta) = \prod_{k=1}^{K} \theta_k^{n_k}.$$

Under an improper Dirichlet prior on the probability of observing each of the distinct values

$$\pi(\theta) \propto \prod_{k=1}^{K} \theta_k^{-1},$$

then the posterior density is also a Dirichlet distribution

$$p(\theta|v) \propto \prod_{k=1}^{K} \theta_k^{n_k - 1}.$$

For Bayesian inference for θ, we consider a weighted likelihood

$$\widetilde{L}(\theta) = \prod_{i=1}^{n} f(z_i;\theta)^{w_i},$$

in which the weights $w = (w_1, \ldots, w_n)$ are distributed according to the uniform Dirichlet distribution and simulated as n independent standard exponential (i.e. gamma(1,1)) variates and standardised. The weighted likelihood reduces to

$$\widetilde{L}(\theta) = \prod_{i=1}^{n} \left\{ \prod_{k=1}^{K} \theta_k^{I_k(z_i)} \right\}^{w_i} = \prod_{k=1}^{K} \theta_k^{\sum\limits_{i=1}^{n} w_i I_k(z_i)} = \prod_{k=1}^{K} \theta_k^{n\gamma_k},$$

say, where $n\gamma_k$ is the sum of the weights w_i for which $z_i = a_k$. Since the vector $\gamma = (\gamma_1, \ldots, \gamma_K)$ has a Dirichlet distribution with parameters $n_k = (n_1, \ldots, n_K)$,

$$p(\gamma) \propto \prod_{k=1}^{K} \gamma_k^{n_k - 1}$$

and since at the point of maximisation of $\widetilde{L}(\theta)$ is $\widetilde{\theta} = \gamma$, then the solutions to the maximised weighted likelihood function with repeatedly sampled uniform Dirichlet weights $w^{(l)}$ represent a sample from the posterior of θ under the improper prior $\prod_k \theta_k^{-1}$. Note that this is in effect a semiparametric model, since although we require a parametric form for the weighted likelihood, the Dirichlet prior imposes no restrictions on the distribution of z other than $\theta_k \geq 0$ and $\sum_{k=1}^K \theta_k = 1$. Newton and Raftery (1994) provide asymptotic results for weighted likelihood bootstrapped models. They show that this class of model are first order correct under rather general conditions and suggest improvements in accuracy through the application of sampling-importance resampling methods.

3 Approximate Bayesian doubly robust estimation

3.1 Approximate Bayesian inference

We now show how the Bayesian bootstrap approach can be used to generate posterior predictive distributions that have the DR property for estimation of ATEs and APOs. For the purposes of illustration we derive an approximate Bayesian DR model in the context of MLE for exponential family GLMs. Let $Z = (Y, X, D)$ be a random sample from the discrete distribution with support on observed data $z = (z_1, \dots, z_n)$. To generalize notation we write $\widehat{\kappa}_i(d, x)$ to denote inverse PS values estimated from observed data and evaluated at d. In the binary case there is a single inverse PS covariate that we use in the AOR model so that

$$\widehat{\kappa}_i(D_i, X_i) = \frac{I_1(D_i)}{\widehat{\pi}(D_i|X_i;\widehat{\alpha})} + \frac{[1 - I_1(D_i)]}{1 - \widehat{\pi}(D_i|X_i;\widehat{\alpha})},$$

and for multi-valued and continuous treatments $\widehat{\kappa}_i(D_i, X_i)$ is an $(n \times Q)$ matrix each column of which contains a covariate for treatment stratum q

$$\frac{I_q(D_i)}{\widehat{\pi}(D_i|X_i;\widehat{\alpha})}.$$

We specify a weighted AOR GLM as

$$\begin{aligned} e\left(D_i, X_i, \widehat{\kappa}_i(D_i, X_i); \xi\right) &= \Psi^{-1}\left\{ m\left(X_i, D_i; \beta\right) + \sum_{q=1}^{Q} \frac{\varphi_q I_q(D_i)}{\widehat{\pi}(D_i|X_i, \widehat{\alpha})} \right\} \\ &= \Psi^{-1}\left\{ m_A\left(X_i, D_i, \widehat{\kappa}_i(D_i, X_i); \xi\right) \right\} \end{aligned}$$

say, where $\xi = (\beta, \varphi)$ and in which the weights feature as a prior weight in the dispersion parameter $a(\phi) = \phi/w$. The maximiser of $\widetilde{L}(\xi)$, which we denote $\widetilde{\xi}$, implies a solution to

$$\sum_{i=1}^{n} w_i^{(l)} \frac{1}{\phi} \frac{\partial e\left(d_i, x_i, \widehat{\kappa}_i(d_i, x_i); \xi\right)}{\partial \xi^{\mathsf{T}}} \left[y_i - e\left(d_i, x_i, \widehat{\kappa}_i(d_i, x_i); \xi\right)\right] = 0, \tag{3}$$

which as noted in relation to (1) has a bias correction property via the inclusion of inverse PS covariates. We repeatedly draw sets of random weights $\{w_i^{(l)}\}_{i=1}^n$ as n standardised independent standard exponential variates and solve (3) to build up a posterior density of $\widetilde{\xi}$, denoted $p_n(\widetilde{\xi})$, from which the sampled values $\widetilde{\xi}^{(l)}$ are consistent with the DR estimating equations.

3.2 Approximate Bayesian prediction

Our final objective is to use the posterior distribution $p_n(\widetilde{\xi})$ to construct posterior predictive intervals for ATEs or APOs with the double robust property. An iid sample from the posterior predictive of interest can be obtained by repeatedly computing

$$\widehat{Y}^{(l)} = \Psi^{-1}\{m_A(d, x, \widehat{\kappa}_i(d, x); \widetilde{\xi}^{(l)})$$

for appropriately chosen d and x. The posterior predictive distribution for ATEs can be obtained via the same principle applying the calculations outlined in Section 2.1.

The marginal (over covariates) posterior predictive distribution of new response variable $\widehat{Y}$ computed for fixed exposure d is defined by

$$\widehat{p}_n(y|d, \mathrm{Data}) = \int \left\{ \int f(y|d, x, \widehat{\kappa}(d, x); \widetilde{\xi})\, p_n(\widetilde{\xi})\, \mathrm{d}\widetilde{\xi} \right\} \widehat{p}_n(x)\, \mathrm{d}x, \tag{4}$$

where $\widehat{p}_n(x)$ is a posterior predictive distribution on the covariates x; in a non-informative and non-parametric specification, this distribution can be taken to be the empirical distribution of the covariate values. We note that by iterated expectation, the marginal posterior predictive expectation may be computed as

$$\int \left\{ \int \Psi^{-1}\{m_A(d, x, \widehat{\kappa}(d, x); \widetilde{\xi})\}\, p_n(\widetilde{\xi})\, \mathrm{d}\widetilde{\xi} \right\} \widehat{p}_n(x)\, \mathrm{d}x$$

and also note that within this calculation, the order of integration may be reversed. In this calculation, we treat the quantity

$$\widehat{\kappa}(d, x) = \frac{I_q(d)}{\widehat{\pi}(d|x; \widehat{\alpha})}$$

as a function of (d, x) for fixed $\alpha = \widehat{\alpha}$. Alternatively, it is possible to make the calculations over the posterior distribution of α. Sections 3.3 and 3.4 provide a discussion of this point.

For different treatment types, we proceed as follows:

- For a binary treatment, we re-sample V values of our covariate vector uniformly over the observed values, and a single vector $\xi^{(l)}$, and form the averages

  $$\widehat{\mu}^{(l)}(1) = \frac{1}{V} \sum_{v=1}^{V} \Psi^{-1}\{m_A(1, x_v, \widehat{\kappa}_i(1, x_v); \widetilde{\xi}^{(l)})\},$$

  $$\widehat{\mu}^{(l)}(0) = \frac{1}{V} \sum_{v=1}^{V} \Psi^{-1}\{m_A(0, x_v, \widehat{\kappa}_i(0, x_v); \widetilde{\xi}^{(l)})\}.$$

  We then use these to form a sampled value of the ATE random variable

  $$\tau_{BDR}^{(l)}(1) = \widehat{\mu}^{(l)}(1) - \widehat{\mu}^{(l)}(0).$$

  We repeat this procedure this L times, $l = (1, \ldots, L)$, to obtain the predictive distribution of the ATE.

- For continuous (or multi-valued) treatments we randomly sample V values of
$$\widehat{y}_v^{(l)}(d_{qj}) = \Psi^{-1}\{m_A(d_{qj}, x_v, \widehat{\kappa}_i(d_{qj}, x_v); \widetilde{\xi}^{(l)})\}$$
for $j = (1, \ldots, J)$ fixed treatment levels in stratum q and form the random variable
$$\widehat{\mu}_{BDR}^{(l)}(\mathcal{D}_q) = \frac{1}{J}\sum_{j=1}^{J}\left[\frac{1}{V}\sum_{v=1}^{V}\widehat{y}_v^{(l)}(d_{qj})\right].$$
We do this L times to obtain the posterior predictive distribution of the mean APO for treatment stratum q. We repeat this process for all strata of interest.

3.3 Fixing the parameters of the PS model

The posterior inference described above is *conditional* on fixed estimates of $\alpha = \widehat{\alpha}$ and we avoid making joint inference on the parameters of the outcome and PS models. This conditional approach provides a first order approximation to the mean of the posterior predictive distribution of APOs, and is justified on the following grounds.

(i) Recall that the assumed conditional mean response model is $\Psi^{-1}\{m(d, x; \beta)\}$ rather than $\Psi^{-1}\{m_A(d, x, \widehat{\kappa}; \xi)\}$. Recall also that $\pi(d|x, \alpha)$ is a parametric representation, akin to the *parametric submodel* in the terminology of frequentist semiparametric inference (see, for example, Tsiatis, 2006) of the treatment assignment mechanism encapsulated in $\pi(d|x)$.

(ii) Under the unconfoundedness assumption, the response is conditionally independent of treatment allocation given the confounders. Consequently, in any likelihood-based inference, the observed treatment allocations cannot be informative about the parameters β from the OR model, and we should treat the estimated PS quantities $\widehat{\pi}(d|x)$ as fixed functions of d and x. The only exception to this case in a Bayesian calculation would arise if the parameters (α, β) were considered a priori dependent; however, in the absence of specific knowledge there is no compelling reason to make such an assumption.

(iii) In a fully Bayesian calculation, the terms $\widehat{\pi}(d|x)$ should themselves be computed as posterior predictive quantities, as in Bayesian density estimation: that is, we compute
$$\widehat{\pi}(d|x) = \int \pi(d|x, \alpha)p_n(\alpha)\, \mathrm{d}\alpha \tag{5}$$
where $p_n(\alpha)$ is the posterior distribution for α computed from the observed treatment and confounder data and an appropriate prior distribution. The quantity $\widehat{\pi}(d|x)$ is the posterior predictive expectation of $\pi(d|x, \alpha)$ computed with respect to $p_n(\alpha)$.

(iv) In the limit as $n \longrightarrow \infty$, under standard conditions, $p_n(\alpha)$ converges to a degenerate distribution at some value α^*, and for all (d, x),
$$\widehat{\pi}(d|x) \longrightarrow \pi(d|x, \alpha^*).$$

Therefore, a reasonable finite sample approximation to $\widehat{\pi}(d|x)$ is $\pi(d|x,\widehat{\alpha})$ for a suitable consistent estimator $\widehat{\alpha}$ of α such as the ML estimate, or Bayesian posterior mean.

3.4 Variance correction to accommodate uncertainty in estimation of the PS model

A drawback in fixing the parameters of the PS model, using MLE or Bayesian posterior means, is that such conditional inference does not account for uncertainty in estimation of the PS model (for a discussion of this problem in the context of parametric empirical Bayes models see Kass and Steffey, 1989). The posterior variance of the parameters of the AOR model given the data is

$$\mathrm{Var}(\xi|z_i) = \mathbb{E}_\alpha\left\{\mathrm{Var}(\xi|z_i,\alpha)\right\} + \mathrm{Var}_\alpha\left\{\mathbb{E}(\xi|z_i,\alpha)\right\}. \tag{6}$$

Applying the weighted likelihood approach to simulate the posterior of the AOR model, with fixed estimate $\widehat{\alpha}$ plugged in, yields an estimate of $\mathrm{Var}(\xi|z_i,\widehat{\alpha})$ for the posterior variance, which approximates only the first term in (6).

To incorporate uncertainty in estimation of the PS model we can make the weighted likelihood calculations for the AOR model over a sample from the posterior of α rather than for a fixed value. To do so we suggest the following approach.

(i) Use weighted likelihood to simulate the posterior of α, which we denote $p_n(\widetilde{\alpha})$, by repeatedly drawing random weights $\{w_i^{(m)}\}_{i=1}^n$ as n standardised independent exponential variates and estimating the weighted PS model (as explained in Section 2.3).

(ii) Repeatedly sample single vectors from $p_n(\widetilde{\alpha})$, $\widetilde{\alpha}^{(m)}$, and form

$$\widehat{\kappa}_i^{(m)}\left(d_i, x_i; \widetilde{\alpha}^{(m)}\right) = \frac{I_q(d_i)}{\widehat{\pi}\left(d_i|x_i;\widetilde{\alpha}^{(m)}\right)},$$

for each stratum q, thus generating sets of inverse PS covariates which vary across values of $\widetilde{\alpha}^{(m)}$ for fixed x_i.

(iii) Draw a single set of random weights $\{w_i^{(l)}\}_{i=1}^n$ for the AOR model

$$\Psi^{-1}\left\{m_A\left(d_i, x_i, \widehat{\kappa}_i^{(m)}\left(d_i, x_i; \widetilde{\alpha}^{(m)}\right); \xi^{(l)}\right)\right\}$$

and repeatedly estimate it for all values of $\widehat{\kappa}_i^{(m)}\left(d_i, x_i; \widetilde{\alpha}^{(m)}\right)$ generated in set (ii).

(iv) Repeatedly compute (iii) using new weights.

In this way we build up a posterior density for $\widetilde{\xi}$ which has variance within distinct values of α and between different values of α, thus approximating both components of (6). In the application section of the paper, we demonstrate the effects of this variance correction.

It is worth noting that while we are able to make such a correction to account for uncertainty in estimation of the PS model, for practical purposes the fixed version of the PS model with MLE estimates is still useful. In particular, in order to make viable comparisons of potential outcomes at different treatment levels we require common support, and this has to be evaluated empirically according to some defined criterion prior to specification of the outcome model. Here we we follow the convention in the causal literature of using the PS model $\pi(d|x, \widehat{\alpha})$ with MLE estimates to evaluate common support.

4 Simulations

In this section we present simulations to demonstrate the DR properties of the approximate Bayesian approach. To provide comparison with frequentist results, we follow exactly the simulations of Graham et al. (2012) which test models to estimate the dose-response of a continuous treatment. Continuous treatment D is assigned as a function of covariates X_1, X_2, and U. The dose–response function is quadratic in D with confounding from X_1 and X_2.

$$\begin{aligned} X_1, X_2 &\sim \mathcal{N}(\mu_{X_1} = 4, \mu_{X_2} = 8, \sigma^2_{X_1} = 1, \sigma^2_{X_2} = 2, \rho = -0.5), \quad U \sim \mathcal{N}(10, 4), \\ D|X_1, X_2, U &\sim \mathcal{N}(0.5 + 0.5X_1 + 0.25X_2 + U, \sigma^2_D = 10), \\ Y|X_1, X_2, D &\sim \mathcal{N}(1 + 3D - 0.11D^2 + 0.5X_1 + 2X_2 - 0.5X_2^2, \sigma^2_Y = 4). \end{aligned}$$

The correct OR model is

$$\mathbb{E}[Y|D, X_1, X_2] = \beta_0 + \beta_1 D + \beta_2 D^2 + \beta_3 X_1 + \beta_4 X_2 + \beta_5 X_2^2,$$

and the correct PS model is

$$\widehat{\pi}_T^{-1} = \int_{D-\delta}^{D+\delta} \frac{1}{\sqrt{2\pi\widehat{\sigma}^2_D}} \exp\left(-\frac{1}{2\widehat{\sigma}^2_D}(t - \widehat{\mu}_D)^2\right) dt,$$

where μ_D and σ^2_D are estimated via the model with $\mathbb{E}[D|X_1, X_2] = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2$ that omits the non-confounder, U. The following models are tested:

1. $\widehat{\mu}(\mathcal{D}_q)_{ABOR1}$ – an approximate Bayesian OR model (ABOR). Our point estimate is taken at the mean of the APO posterior predictive distributions from the correctly specified OR model, with predicted values for doses (d_{qj}) averaged over $j = (1, \ldots, J)$ treatment levels within each stratum q, i.e.

$$\widehat{\mu}_{OR}(\mathcal{D}_q) = \frac{1}{L} \sum_{l=1}^{L} \left[\frac{1}{J} \sum_{j=1}^{J} \left[\frac{1}{V} \sum_{v=1}^{V} \Psi^{-1} \left\{ m(d_{qj}, x_v; \beta^{(l)}) \right\} \right] \right].$$

2. $\widehat{\mu}(\mathcal{D}_q)_{ABOR2}$ – same estimate as $\widehat{\mu}(\mathcal{D}_q)_{OR1}$ but based on an incorrectly specified OR model, with X_1 assumed as sole confounder.

3. $\widehat{\mu}(\mathcal{D}_q)_{ABDR1}$ – an approximate Bayesian DR estimator (ABDR) calculated at the mean of the posterior predictive distributions for APOs, i.e.

$$\widehat{\mu}^{(l)}_{BDR}(\mathcal{D}_q) = \frac{1}{L}\sum_{l=1}^{L}\left[\frac{1}{J}\sum_{j=1}^{J}\left[\frac{1}{V}\sum_{v=1}^{V}\Psi^{-1}\left\{m_A(d_{qj}, x_v, \widehat{\kappa}(d_{qj}, x_v); \widetilde{\xi}^{(l)})\right\}\right]\right].$$

The model has an incorrectly specified OR model (X_1 as sole confounder) augmented with correctly estimated inverse PS ($\widehat{\pi}_T$) covariates for defined strata of the treatment.

4. $\widehat{\mu}(\mathcal{D}_q)_{ABDR2}$ – an approximate Bayesian DR estimator based on a correctly specified OR model augmented with incorrectly estimated inverse PS covariates ($\widehat{\pi}_F$), with X_1 assumed as sole confounder, for defined strata of the treatment.

5. $\widehat{\mu}(\mathcal{D}_q)_{ABDR3}$ – an approximate Bayesian DR estimator based on an incorrectly specified OR model augmented with incorrectly estimated inverse PS covariates.

The results are derived from 1000 runs on generated datasets of size 10,000. The mean of d is 14.5 and the range approximately 1 to 30. Estimates are presented for the following treatment strata: (10,12], (12,14], (14,16], (16,18], (18,20]. Mean values and variances of the point estimates (i.e. means and variances of the APO distributions) obtained from the simulations and the mean squared error (MSE) are reported.

Table 1 shows results for the simulations. The correctly specified ABOR model, ABOR1, provides a good estimate of the quadratic dose–response as expected. The APO estimates from the incorrectly specified ABOR2 model produce a poor representation of the dose–response curve indicating a linear decreasing effect with relatively large bias and MSE. The ABDR1 model, which augments the incorrect OR model with inverse PS covariates, produces a good approximation to the true dose–response by correcting for bias from confounding and from functional misspecification of the treatment covariate. The ABDR2 model demonstrates that the inclusion of addition irrelevant PS covariates in the OR model does not induce bias, but it does appear to increase variance and MSE relative to ABOR1. The results for ABDR3 show large bias in estimation of the dose–response, demonstrating that at least one of the OR or PS models must be correct for the DR property to hold.

Table 2 compares parameter estimates for the correctly specified PS model, $\alpha = (\alpha_0, \alpha_1, \alpha_2)$, derived via MLE, the Bayesian Bootstrap, and a full Bayesian model for $d|x \sim \mathcal{N}(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2, \sigma^2_{d|x})$ with the following priors: $\alpha_k \sim \mathcal{N}(0, 1000^2)$ and $\sigma_{d|x} \sim Unif(0, 0.01)$. Mean estimates and variances of the estimates are very similar. Thus, the Bayesian Bootstrap estimates appear to be approximately normally distributed with mean $\widehat{\alpha} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}Y$ and variance $\sigma^2(X^{\mathsf{T}}X)^{-1}$, as we would expect in a Bayesian analysis of the normal linear model with noninformative priors for n large. The results demonstrate that with reasonably large n, and in the absence of specific prior information, the MLE estimates, mean Bayesian Bootstrap estimates, and Bayesian posterior means will offer similar approximations to $\widehat{\pi}(d|x)$.

		treatment intervals				
		(10,12]	(12,14]	(14,16]	(16,18]	(18,20]
Truth		5.919	6.416	6.037	4.782	2.646
ABOR1	Av Est	5.927	6.423	6.042	4.784	2.650
	Emp Var	0.011	0.010	0.010	0.010	0.012
	MSE	0.011	0.010	0.010	0.010	0.012
ABOR2	Av Est	5.856	5.156	4.457	3.757	3.057
	Emp Var	0.015	0.010	0.011	0.016	0.026
	MSE	0.019	1.598	2.451	1.067	0.195
ABDR1	Av Est	5.912	6.420	6.034	4.769	2.646
	Emp Var	0.052	0.039	0.038	0.046	0.077
	MSE	0.052	0.039	0.038	0.046	0.077
ABDR2	Av Est	5.921	6.416	6.031	4.783	2.655
	Emp Var	0.017	0.016	0.015	0.017	0.030
	MSE	0.017	0.016	0.015	0.017	0.030
ABDR3	Av Est	6.376	6.579	5.869	4.289	1.847
	Emp Var	0.051	0.037	0.039	0.052	0.091
	MSE	0.260	0.063	0.067	0.295	0.728

Table 1: Simulation results for Gaussian dose-response GLM with quadratic treatment effect.

	α_0	α_1	α_2
Likelihood Av Est	10.437	0.504	0.256
Likelihood Emp Var	1.377	0.020	0.009
Bayesian Bootstrap Av mean Est	10.436	0.504	0.256
Bayesian Bootstrap Emp Var	1.376	0.020	0.009
Full Bayes Av posterior mean	10.437	0.504	0.256
Full Bayes Emp Var	1.380	0.020	0.009

Table 2: A comparison of Likelihood, Full Bayes and Bayesian Bootstrap estimates of the Propensity Score model.

5 Application: quantifying the marginal effect of socioeconomic deprivation on child pedestrian casualties

A positive association between socio-economic deprivation and the incidence of child pedestrian casualties (CPCs) has frequently been reported in the literature (for reviews see Christie, 1995; Graham and Stephens, 2008). Statistical work appears to show that children from deprived backgrounds have a substantially higher chance of being involved in a pedestrian accident. Graham et al. (2013) argue that the relationship is likely confounded and conduct a frequentist area-based analysis of the effect of exposure to area based socio-economic deprivation on the incidence of CPCs in British cities. In this section we apply the Bayesian DR estimator to the same British data to derive predictive posterior distributions of mean APOs for different strata of deprivation.

This case study does not represent a typical causal inference problem as deprivation is a composite measure depending on several factors, and the precise mechanisms via which these factors might affect CPC rates are not well understood. Nevertheless, it does share the same basic features in that the available data allow us to represent three key dimensions of our problem: response, exposure, and confounding covariates; and our objective is to quantify the *marginal* effect of the exposure on the response. It is certainly conceivable that one may devise an intervention to alter 'deprivation'. Hence our results will represent an unconfounded estimate of the impact of such an intervention.

The data and the logic underpinning covariate construction are described in full in Graham et al. (2013). Here, we provide only a brief summary. The response variable Y comprises annual counts of child (< 16) pedestrian casualties (CPCs) over the period 2001 to 2007 for small spatial units of British cities (based on census wards). The exposure variable D is a measure of poverty based deprivation for zones calculated as the natural logarithm of number of residents in receipt of Government benefits. The mean of D is 6.496, the minimum and maximum values are 4.007 and 8.738, and the standard deviation is 0.643.

We suspect that exposure to deprivation is confounded with area-based characteristics. Specifically, we hypothesise the following sources of confounding:

i. *Child population* – deprived families may tend to have more children creating a larger supply of potential victims. We include a covariate measuring resident child population.

ii. *Traffic generation potential and nature of the urbanised environment* – deprived zones may tend to experience greater volumes of traffic. We construct a form of 'gravity' trip generation model to represent potential traffic flows at the zone level.

iii. *Variation in the nature of the built environment* – we expect road safety and zone deprivation to be associated with the nature of land use and the degree of urbanisation. For instance, children living in suburban residential environments may be more affluent with less exposure to traffic risk than those living in dense inner city mixed use locations. To represent such factors we include measures of zone population and employment density.

iv. *Scale of the road network* – high capacity networks tend to depress land values which in turn will influence the socio-economic profile of the people that live in close proximity. Using GIS software we generated longitudinal data on network capacity for each zone including a breakdown by road type: A-road, B-road, minor road, and motorway.

v. *Road network density* – with the available GIS data we were also able to represent the road network density in each zone using a measure of the number of network nodes per unit of area. A network node is defined as the meeting point of two or more links. Deprived zones may tend to have more extensive dense networks.

In estimating the exposure model we found that a log transformation yields data that approximately follow a Gaussian distribution. We estimated the conditional density of exposure given covariates using a Gaussian Generalised Linear Model (GLM) estimated by MLE. We then used a calculated the PS values using a normal distribution.

To test our PS specification we check for balancing. In a similar manner to Hirano and Imbens (2004) and Flores et al. (2012), we regress the exposure on the covariates, the PS, and a set of indicator variables corresponding to a range discretisation of the exposure variable in ten strata. The BIC values obtained from the linear regression models with and without covariates are -35304 and -35367, respectively, indicating that the inclusion of covariates leads to a deterioration in model adequacy. We also conducted an F-test between the restricted (without covariates) and unrestricted models and obtained an F statistic (p-value) of 0.439 (0.508). These results suggest that the balancing property has been achieved for our PS specification.

The AOR model we use is a Poisson GLM estimated by MLE. To derive the posterior predictive distributions for APOs based on PS and OR model results we use the approach described in Section 3 and demonstrated in the simulations above. We use the following treatment discretisation for strata between 6.6 and 7.6 of width 0.1. The APOs are averages of mean predicted values from the AOL model over treatment levels within each stratum. Posterior variance estimates and credible intervals are calculated via the variance correction approach described in Section 3.4. Estimates of the mean and variance (corrected and uncorrected) of the posterior predictive distributions and 95% credible intervals for 10 treatment levels are shown in Table 3 along with Likelihood point estimates for comparison. Kernel density fits to posterior predictive distributions for four treatment levels are shown in Figure 1 below.

	Bayesian bootstrap				Likelihood bootstrap	
	posterior		uncorrected			
mean $d \in k$	mean	s.d.	s.d.	95% cred. int.	Est.	s.e.
6.61	2.061	0.137	0.123	(1.793, 2.330)	2.055	0.041
6.75	2.478	0.126	0.118	(2.232, 2.725)	2.466	0.039
6.85	2.879	0.120	0.110	(2.644, 3.114)	2.893	0.032
6.95	2.759	0.108	0.093	(2.548, 2.970)	2.755	0.028
7.05	2.845	0.105	0.085	(2.639, 3.051)	2.860	0.028
7.15	2.889	0.105	0.085	(2.683, 3.095)	2.919	0.026
7.25	3.190	0.124	0.086	(2.948, 3.432)	3.190	0.030
7.35	3.407	0.127	0.096	(3.158, 3.655)	3.399	0.036
7.45	3.809	0.133	0.112	(3.548, 4.071)	3.831	0.043
7.55	4.130	0.192	0.154	(3.754, 4.506)	4.150	0.062

Table 3: Bayesian and likelihood bootstrapped estimates of mean average potential outcomes by treatment level.

Diagnostic tests were conducted to identify obvious instances of model misspecification. Using the approach of Robins and Rotnitzky (2001) described in Section 2.2 above, we find little evidence of PS model misspecification for APO estimates at each dose of interest but we do find that the OR model may not provide a universally good specification over all doses of interest, although we cannot reject the null that the OR model is correctly specified for most doses. Similarly, box plots of our φ_q estimates indicate values significantly different from zero, again potentially indicative of some deficiency of the OR model. These diagnostic results underline the usefulness a DR approach in correcting for sources of model misspecification.

The results indicate a positive increasing effect of deprivation on CPCs having adjusted for measured confounders. Over the range of exposure considered, the predicted number of CPC is almost twice as large in zones exposed to the highest dose of deprivation than the lowest. We therefore find compelling evidence of a deprivation gradient. Note the flexibility in presentation of results offered via the approximate Bayesian approach. We are able to present our APOs as distributions rather than point estimates, and we can discuss our results in terms of central 95% credible intervals. If there was interest in some particular hypotheses regarding the regions within which the ATEs lie, it would also be possible to test these using our approach but not using the likelihood point estimates.

The dose–response estimates obtained from the Bayesian and Likelihood bootstrapped approaches are similar in magnitude, but the variance of the Bayesian bootstrap estimates is larger. As discussed in Section 3.4, accounting for uncertainty in estimation of the PS model inflates the variance of the posterior predictive distributions and this is illustrated in the comparison of corrected and uncorrected posterior variances. The kernel density fits shown in Figure 1 indicate that the posterior predictive distributions for ATEs are approximately normally distributed.

6 Conclusions

This paper has presented an approach that can be used to derive approximate Bayesian inference for doubly robust estimation of causal quantities. This is a useful extension to existing methods for two reasons. First, doubly robust ATE estimation typically involves prediction and extrapolation over unobserved covariate distributions and a Bayesian approach provides a natural framework for prediction in which both the unobserved covariates and the parameters have random status. Second, in constructing approximate posterior predictive densities our approach allows results to be presented in terms of probability statements about key causal quantities of interest, which can offer greater flexibility for practical interpretation than point estimates.

Our case study indicates a positive relationship between exposure to deprivation and the incidence of child pedestrian casualties, having adjusted for confounding via outcome regression and propensity score adjustment. Tests for model misspecification indicate that the outcome regression model may not provide a universally good specification for our case study analysis, but the propensity score model appears to perform well overall.

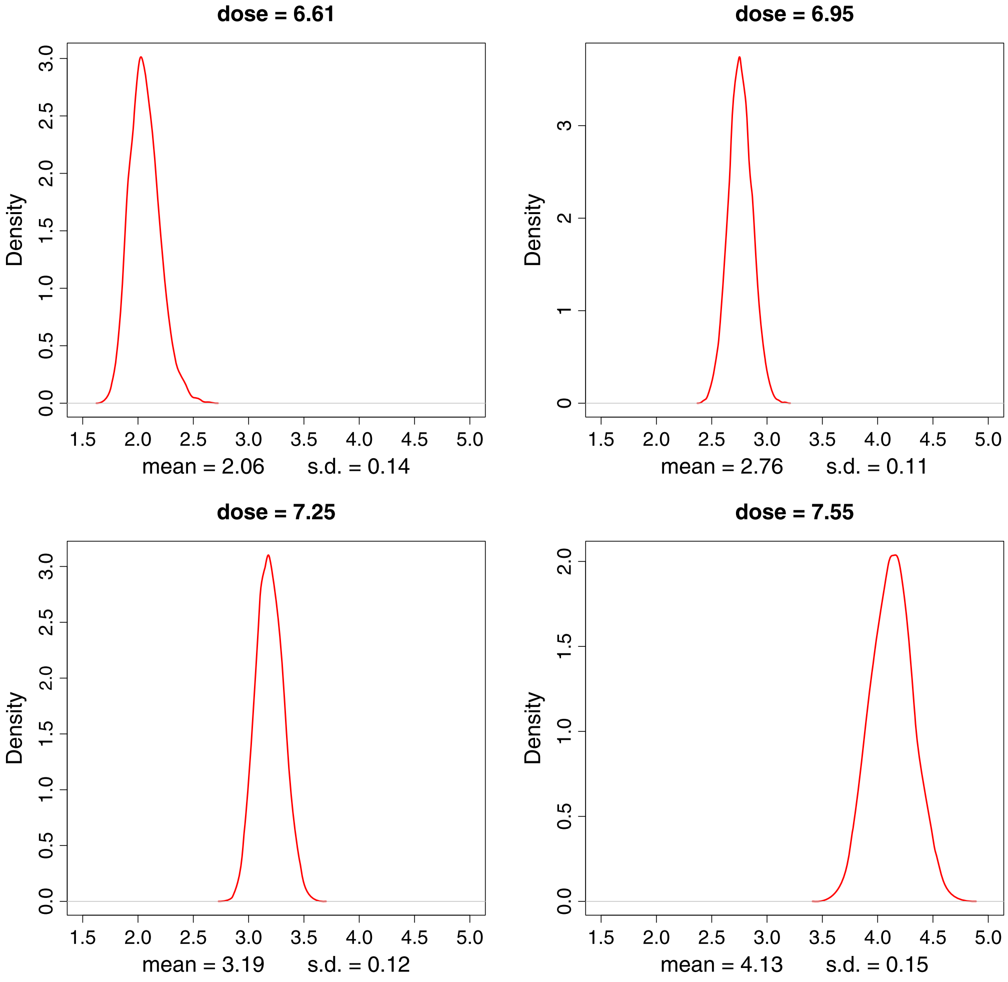

Figure 1: Predictive posterior densities for mean average potential outcomes at doses 6.61, 6.95, 7.25 and 7.55.

This underlines the usefulness of a doubly robust approach which combines the two model to adjust for sources of misspecification.

Appendix A: Balancing and conditional independence given Propensity scores

Propensity score (PS) estimators require that conditional independence holds given the PS (i.e. $Y_i(0), Y_i(1)) \perp\!\!\!\perp I_1(D_i)|\pi(D_i|X_i;\alpha)$ for binary treatment and $Y_i(d) \perp\!\!\!\perp I_d(D_i)| \pi(D_i|X_i;\alpha)$ for all $d \in \mathcal{D}$ for multi-valued or continuous treatments). A necessary condition for conditional independence to hold is that the PS has a balancing property: $X_i \perp\!\!\!\perp I_1(D_i)|\pi(D_i|X_i;\alpha)$ in the binary case and $X_i \perp\!\!\!\perp I_d(D_i)|\pi(D_i|X_i;\alpha)$ for multi-valued or continuous treatments. Proofs showing why balancing is needed are given below. For further details see Imbens (1999) and Hirano and Imbens (2004).

Balancing and conditional independence given the binary propensity score

Lemma 1. *(Balancing of pre-treatment covariates given the propensity score). If $\pi(D_i|X_i;\alpha)$ is the propensity score, then*

$$X_i \perp\!\!\!\perp I_1(D_i)|\pi(D_i|X_i;\alpha).$$

Proof. First, by the result for expectation of indicator functions and the pull-through property

$$\begin{aligned}\Pr[I_1(D_i)=1|X_i,\pi(D_i|X_i;\alpha)] &= \mathbb{E}[I_1(D_i)=1|X_i,\pi(D_i|X_i;\alpha)]\\ &= \mathbb{E}[I_1(D_i)=1|X_i] = \pi(D_i|X_i;\alpha).\end{aligned}$$

Second,

$$\begin{aligned}\Pr[I_1(D_i)=1|\pi(D_i|X_i;\alpha)] &= \mathbb{E}[I_1(D_i)=1|\pi(D_i|X_i;\alpha)]\\ &= \mathbb{E}_X\left[\mathbb{E}\left\{I_1(D_i)=1|X_i,\pi(D_i|X_i;\alpha)\right\}|\pi(D_i|X_i;\alpha)\right]\\ &= \mathbb{E}[\pi(D_i|X_i;\alpha)|\pi(D_i|X_i;\alpha)]\\ &= \pi(D_i|X_i;\alpha)\end{aligned}$$

Thus, $\Pr[I_1(D_i)=1|X_i,\pi(D_i|X_i;\alpha)] = \Pr[I_1(D_i)=1|\pi(D_i|X_i;\alpha)]$. □

Next we show that, given balancing, conditional independence can be established on the PS rather than covariate vector X_i.

Lemma 2. *(Conditional independence given the propensity score). Given $(Y_i(0),Y_i(1))\perp\!\!\!\perp I_1(D_i)|X_i$ and the propensity score $\pi(D_i|X_i;\alpha)$, then*

$$(Y_i(0),Y_i(1)) \perp\!\!\!\perp I_1(D_i)|\pi(D_i|X_i;\alpha).$$

Proof. First,

$$\begin{aligned}&\Pr[I_1(D_i)=1|Y_i(1),\pi(D_i|X_i;\alpha)]\\ &\quad= \mathbb{E}[I_1(D_i)=1|Y_i(1),\pi(D_i|X_i;\alpha)]\\ &\quad= \mathbb{E}\left[\mathbb{E}\left[I_1(D_i)=1|Y_i(1),X_i,\pi(D_i|X_i;\alpha)\right]|Y_i(1),\pi(D_i|X_i;\alpha)\right]\\ &\quad= \mathbb{E}[\pi(D_i|X_i;\alpha)|Y_i(1),\pi(D_i|X_i;\alpha)]\\ &\quad= \pi(D_i|X_i;\alpha)\end{aligned}$$

Second, from the proof for Lemma 1 we know that $\Pr[I_1(D_i)=1|\pi(D_i|X_i;\alpha)] = \pi(D_i|X_i;\alpha)$, and therefore $\Pr[I_1(D_i)=1|Y_i(1),\pi(D_i|X_i;\alpha)] = \Pr[I_1(D_i)=1|\pi(D_i|X_i;\alpha)]$ implying that $Y_i(1) \perp\!\!\!\perp I_1(D_i)=1|\pi(D_i|X_i;\alpha)$. The same logic yields an analogous proof under control rather than treated status. □

Balancing and conditional independence given the propensity score for multivalued and continuous treatments

Lemma 3. *(Balancing of pre-treatment covariates given the generalised propensity score). If $\pi(d, X_i; \alpha)$ is the PS defined with respect to the marginal distribution of X_i for fixed d, then*

$$I_{D_i}(d) \perp\!\!\!\perp X_i | \pi(d, X_i; \alpha).$$

This follows because $\pi(d, X_i; \alpha)$ is a function of X_i alone and so conditioning on X_i adds no additional information

$$\mathbb{E}\left[I_{D_i}(d) | X_i, \pi(d, X_i; \alpha)\right] = \mathbb{E}\left[I_{D_i}(d) | \pi(d, X_i; \alpha)\right].$$

Given balancing, we can establish *weak* conditional independence (i.e. conditional independence for each value of the treatment but not joint independence of all potential outcomes) using the PS.

Theorem 1. *(Weak conditional independence given the propensity score). If assignment to the treatment is weakly conditionally independent given pre-treatment characteristics X_i, then for all $D_i = d$*

$$Y_i(d) \perp\!\!\!\perp I_{D_i}(d) | \pi(d, X_i; \alpha).$$

Proof. To prove that $Y_i(d)$ is conditionally independent of $I_{D_i}(d)$, given the generalised propensity score $\pi(d, X_i; \alpha)$ and the assumption of weak conditional independence, it is sufficient to show that $f_{D|\pi,Y}\left(d | \pi(d, x_i; \alpha), Y_i(d)\right) = f_{D|\pi}\left(d | \pi(d, x_i; \alpha)\right)$. Let $\mathcal{X}$ be the sample space in which covariates X_i lie, then

$$\begin{aligned}
f_{D|\pi}\left(d|\pi(d, x_i; \alpha)\right) &= \int_{\mathcal{X}} f_{D,X|\pi}\left(d, x_i | \pi(d, x_i; \alpha)\right) dx_i \\
&= \int_{\mathcal{X}} f_{D|X,\pi}\left(d | x_i, \pi(d, x_i; \alpha)\right) f_{X|\pi}\left(x_i | \pi(d, x_i; \alpha)\right) dx_i \\
&= \int_{\mathcal{X}} f_{D|X}(d|x_i) f_{X|\pi}\left(x_i | \pi(d, x_i; \alpha)\right) dx_i \\
&= \int_{\mathcal{X}} \pi(d, x_i; \alpha) f_{X|\pi}\left(x_i | \pi(d, x_i; \alpha)\right) dx_i, \\
&= \pi(d, x_i; \alpha) = f_{D|X}(d|x_i).
\end{aligned}$$

Thus,$f_{D|\pi}\left(d|\pi(d, x_i; \alpha)\right) = f_{D|X}(d|x_i)$. Furthermore,

$$\begin{aligned}
& f_{D|\pi,Y}\left(d | \pi(d, x_i; \alpha), Y_i(d)\right) \\
&\quad = \int_{\mathcal{X}} f_{D,X|\pi,Y}\left(d, x_i | \pi(d, x_i; \alpha), Y_i(d)\right) dx_i \\
&\quad = \int_{\mathcal{X}} f_{D|X,\pi,Y}\left(d | x_i, \pi(d, x_i; \alpha), Y_i(d)\right) f_{X|\pi}\left(x_i | \pi(d, x_i; \alpha), Y_i(d)\right) dx_i \\
&\quad = \int_{\mathcal{X}} f_{D|X}(d|x_i) f_{X|\pi}\left(x_i | \pi(d, x_i; \alpha), Y_i(d)\right) dx_i
\end{aligned}$$

$$
\begin{aligned}
&= \int_{\mathcal{X}} \pi(d, x_i; \alpha) f_{X|\pi}\left(x_i | \pi(d, x_i; \alpha), Y_i(d)\right) dx_i \\
&= \pi(d, x_i; \alpha) = f_{D|X}(d|x_i).
\end{aligned}
$$

Therefore, for all d, $f_{D|\pi}(d|\pi(d, x_i; \alpha)) = f_{D|\pi,Y}\left(d|\pi(d, x_i; \alpha), Y_i(d)\right)$ and we have weak conditional independence given $\pi(d, X_i; \alpha)$. □

References

Bang, H. and Robins, J. M. (2005). "Doubly robust estimation in missing data and causal inference models." *Biometrics*, 61: 962–972. MR2216189. doi: http://dx.doi.org/10.1111/j.1541-0420.2005.00377.x. 48, 52

Chamberlain, G. and Imbens, G. W. (2003). "Nonparametric Applications of Bayesian Inference." *Journal of Business & Economic Statistics*, 21(1): 12–18. MR1973803. doi: http://dx.doi.org/10.1198/073500102288618711. 52

Christie, N. (1995). "Social, economic and environmental factors in child pedestrian accidents: a research overview." Technical Report 116, Transport Research Laboratory, Berkshire. 60

Flores, C. A., Flores-Lagunes, A., Gonzalez, A., and Neumann, T. C. (2012). "Estimating the Effects of Length of Exposure to Instruction in a Training Program: The Case of Job Corps." *The Review of Economics and Statistics*, 94(1): 153–171. doi: http://dx.doi.org/10.1162/REST_a_00177. 62

Graham, D. J., McCoy, E. J., and Stephens, D. A. (2012). "Semiparametric double-robust estimation for continuous treatment effects." *Paper Presented at the 2012 Joint Statistical Meetings, San Diego*. 50, 51, 58

— (2013). "Quantifying the effect of area deprivation on child pedestrian casualties using longitudinal mixed models to adjust for confounding, interference, and spatial dependence." *Journal of the Royal Statistical Society: Series A*, 176(4): 931–950. MR3120956. doi: http://dx.doi.org/10.1111/j.1467-985X.2012.01071.x. 60, 61

Graham, D. J. and Stephens, D. A. (2008). "Decomposing the impact of deprivation on child pedestrian casualties in England." *Accident Analysis & Prevention*, 40: 1351–1364. doi: http://dx.doi.org/10.1016/j.aap.2008.02.006. 60

Gustafson, P. (2012). "Double-robust estimators: slightly more Bayesian than meets the eye." *The International Journal of Biostatistics*, 8(2): 1–15. MR2925326. doi: http://dx.doi.org/10.2202/1557-4679.1349. 47

Hirano, K. and Imbens, G. W. (2004). "The propensity score with continuous treatments." In Gelman, A. and Meng, X. (eds.), *Applied Bayesian modeling and causal inference from incomplete data perspectives*, 73–84. New York: Wiley. MR2134803. doi: http://dx.doi.org/10.1002/0470090456.ch7. 49, 62, 65

Horvitz, D. G. and Thompson, D. J. (1952). "A generalization of sampling without replacement from a finite universe." *Journal of the American Statistical Association*, 47: 663–685. MR0053460. doi: http://dx.doi.org/10.1080/01621459.1952.10483446. 49

Imbens, G. W. (1999). "The role of the propensity score in estimating dose–response functions." *NBER Working Paper*, 237. 65

— (2000). "The role of the propensity score in estimating dose–response functions." *Biometrika*, 87(3): 706–710. MR1789821. doi: http://dx.doi.org/10.1093/biomet/87.3.706. 49

Kang, J. D. Y. and Schafer, J. L. (2007). "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data." *Statistical Science*, 22(4): 523–539. MR2420458. doi: http://dx.doi.org/10.1214/07-STS227. 48

Kass, R. E. and Steffey, D. (1989). "Approximate Bayesian Inference in Conditionally Independent Hierarchical Models (Parametric Empirical Bayes Models)." *Journal of the American Statistical Association*, 84(407): 717–726. MR1132587. 57

Lunceford, J. K. and Davidian, M. (2004). "Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study." *Statistics in Medicine*, 23: 2937–2960. doi: http://dx.doi.org/10.1002/sim.1903. 48

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(497): 40–51. MR2949340. doi: http://dx.doi.org/10.1080/01621459.2011.643739. 52

Newton, M. A. and Raftery, A. E. (1994). "Approximate Bayesian Inference with the Weighted Likelihood Bootstrap (with discussion)." *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(1): pp. 3–48. MR1257793. 52, 54

Pearl, J. (2009). *Causality – models, reasoning and inference*. Cambridge: Cambridge University Press, 2nd edition. MR2548166. doi: http://dx.doi.org/10.1017/CBO9780511803161. 52

— (2010). "On a Class of Bias-Amplifying Variables that Endanger Effect Estimates." In *Proceeding of the 25th Conference on Uncertainty in Artificial Intelligence (UAI 2010)*, 425–432. Corvallis: Association for Uncertainty in Artificial Intelligence. 52

Robins, J. M. and Rotnitzky, A. (2001). "Comment on "Inference for semiparametric models: some questions and an answer"." *Statistica Sinica*, 11: 920–936. MR1867326. 52, 63

Rubin, D. B. (1981). "The Bayesian Bootstrap." *The Annals of Statistics*, 9(1): 130–134. MR0600538. doi: http://dx.doi.org/10.1214/aos/1176345338. 48, 52

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(448): 1096–1120 (with rejoinder 1135–1146). MR1731478. doi: http://dx.doi.org/10.2307/2669923. 48, 49, 51

Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. Berlin: Springer. MR2233926. 56

Tsiatis, A. A. and Davidian, M. (2007). "Comment: Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data." *Statistical Science*, 22(4): 569–573. MR2420466. doi: http://dx.doi.org/10.1214/07-STS227B. 48

van der Laan, M. and Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*. Berlin: Springer. MR1958123. doi: http://dx.doi.org/10.1007/978-0-387-21700-0. 48

Acknowledgments

The authors wish to thank Robin Henderson, Nick Heard, the associate editor and two anonymous referees for helpful comments.