The International Journal of Biostatistics

Volume 8, Issue 2

2012

Article 4

CAUSAL INFERENCE IN HEALTH RESEARCH

Double-Robust Estimators: Slightly More Bayesian than Meets the Eye?

Paul Gustafson, *University of British Columbia*

Recommended Citation:

Gustafson, Paul (2012) "Double-Robust Estimators: Slightly More Bayesian than Meets the Eye?," *The International Journal of Biostatistics*: Vol. 8: Iss. 2, Article 4.

DOI: 10.2202/1557-4679.1349

Double-Robust Estimators: Slightly More Bayesian than Meets the Eye?

Paul Gustafson

Abstract

Consider the simple setting of point exposure, outcome and confounding variables, all of which are discrete. As is well known, parametric modeling of outcome given exposure and confounders and also exposure given confounders can yield a double-robust estimator. This has the property of being consistent as long as at least one of the two specified models is correct. Such an estimator can also be cast as arising from a compromise between the parametric outcome model and a nonparametric or saturated outcome model. This brings to mind an alternate compromise based on Bayesian model averaging, and prompts comparisons between the double-robust method and the Bayesian method.

KEYWORDS: Bayesian model averaging, causal inference, double-robust, observational data

Author Notes: Work supported by grants from the Natural Sciences and Engineering Research Council and the Canadian Institutes of Health Research (Funding Reference Number 62863). The author appreciates the helpful comments of two anonymous referees.

1 Introduction

Double-robust estimators, as pioneered about a decade ago (Scharfstein, Rotnitzky, and Robins, 1999, Robins, Rotnitzky, and van der Laan, 2000, Robins and Rotnitzky, 2001), have great appeal. The modeler need only get one of two model specifications correct in order to obtain consistent parameter estimation. These estimators have attracted great interest in the causal inference and missing data domains, with literature searches turning up hundreds of papers developing or using double-robust methods over the last decade. Some key references include van der Laan and Robins (2003), Lunceford and Davidian (2004), Bang and Robins (2005), Kang and Schafer (2007), and Tan (2010). In the causal inference domain, double-robust estimators require modeling the outcome variable given exposure and confounders as well as the exposure given confounders, with consistent estimation arising provided at least one of the specified models is correct.

Bayesians have been rather left out of the excitement surrounding double-robust estimation. A Bayesian is obliged to model all aspects of the observable data. Particularly, this calls for joint inference on parameters in an outcome model and parameters in an exposure model. However, typical targets of inference, such as average causal effects, depend only on outcome model parameters, along with the distribution of confounders. Thus, at least if exposure model parameters and outcome model parameters are judged to be independent *a priori*, a Bayesian's fortunes seem hitched to outcome modeling, with exposure model parameters uninvolved. The purpose of this paper is to point out that, at least in simple settings, Bayesian estimators and double-robust estimators are not quite as disparate as first meets the eye. Both can be cast as arising from compromises between parametric and nonparametric outcome models, and this provides a point of comparison between the two approaches.

2 Double-Robust and Bayesian Estimators

Consider binary counterfactual outcomes (Y_0, Y_1) and binary exposure X giving rise to observed outcome $Y = (1 - X)Y_0 + XY_1$. Assume that a collection of variables C suffice to control confounding, in that (Y_0, Y_1) and X are conditionally independent given C . Assume that C has finite support $\mathcal{C}$. Principally we have in mind the setting that C is comprised of r binary variables $C = (C_1, \dots, C_r)$, with all combinations of these being possible, i.e., $|\mathcal{C}| = 2^r$. We also assume the positivity condition $0 < Pr(X = 1|C = c) < 1$

for all $c \in \mathcal{C}$, whereby no combination of confounders completely determines exposure status. The observed data $d_{1:n}$ arise as n independent realizations $d_i = (y_i, x_i, c_i)$ of (Y, X, C) .

Taking the average causal effect $\Delta = E(Y_1 - Y_0)$ as the target of inference, it is well known that

$$\Delta = E\{E(Y|X = 1, C) - E(Y|X = 0, C)\}. \quad (1)$$

Thus fitting an outcome model for $(Y|X, C)$ and then averaging the effect of exposure across the empirical or fitted distribution of C is a route to estimating the target. Specifically, say the outcome regression model is parameterized by β , with $m_j(C; \beta) = E(Y|X = j, C)$. Then (1) motivates the estimator

$$\hat{\Delta}_P = \frac{1}{n} \sum_{i=1}^n \{m_1(c_i; \hat{\beta}) - m_0(c_i; \hat{\beta})\}, \quad (2)$$

where the subscript P reminds us that this estimator is based on a *parametric* outcome model.

It is also well known that the average causal effect can be alternately expressed as

$$\Delta = E\left\{Y \left(\frac{X}{Pr(X = 1|C)} - \frac{1 - X}{1 - Pr(X = 1|C)} \right)\right\}. \quad (3)$$

Upon fitting a parametric exposure or ‘propensity’ model for $(X|C)$ of the form $\pi(C; \alpha) = Pr(X = 1|C)$, (3) motivates an inverse-probability-weighted estimator,

$$\hat{\Delta}_{IPW} = \frac{1}{n} \sum_{i=1}^n y_i \left\{ \frac{x_i}{\pi(c_i; \hat{\alpha})} - \frac{(1 - x_i)}{(1 - \pi(c_i; \hat{\alpha}))} \right\}. \quad (4)$$

While (4) is helpful conceptually, a version with stabilized weights is typically much more efficient. However, (4) suffices to develop the arguments of this paper.

While numerous double-robust estimators have been proposed for the causal inference setting, Tan (2010) points out that many of them have the form

$$\hat{\Delta}_D = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{y_i x_i - (x_i - \hat{\pi}(c_i)) \hat{m}_1(c_i)}{\hat{\pi}(c_i)} - \frac{y_i(1 - x_i) + (x_i - \hat{\pi}(c_i)) \hat{m}_0(c_i)}{(1 - \hat{\pi}(c_i))} \right\}. \quad (5)$$

Here a standard and simple choice of estimator for the exposure model fits would be $\hat{\pi}(c_i) = \pi(c_i; \hat{\alpha})$ as in (4), and similarly $\hat{m}_j(c_i) = m_j(c_i; \hat{\beta})$ for the

outcome model fits, as in (2). These choices correspond to the double-robust estimator that Lunceford and Davidian (2004) compared to other estimators. As emphasized by Tan (2010), however, more intricate double-robust estimators arise by using estimators of $m_j()$ that depend in part on $\hat{\pi}()$, or iterative estimates with $\hat{m}_j()$ and $\hat{\pi}()$ informing one another. Tan (2010) reviews a large number of such estimators and their properties. It is straightforward to verify that (5) is consistent if one or both of the outcome model and the exposure model is correctly specified, i.e., if at least one of $\pi()$ and $\{m_0(), m_1()\}$ are consistently estimated.

A Bayesian seems to have little to say about the estimator (5). The rigors of combining a prior and a likelihood to form a posterior cannot produce something of this form. Working with a likelihood that factors into a term from the outcome model (which is free of α) and a term from the exposure model (which is free of β), we have that *a priori* independence of α and β leads to *a posteriori* independence. Moreover, since the target of interest can be expressed as a function of β and the distribution of C , inference on the target is the same whether the $(X|C)$ relationship is modeled or unmodeled. The situation becomes more nuanced if α and β are judged to be dependent *a priori*, as investigated by Wang, Parmigiani, and Dominici (2011). The relationship between their work and ours is commented on in Section 5.

The gulf between the double-robust world and the Bayesian world appears to diminish somewhat if we simply rewrite (5). Recall we are in the simple setting where the n observations of (Y, X, C) can be cross-classified into a $2 \times 2 \times |\mathcal{C}|$ table, and denote cell counts as $n_{cxy} = \sum_{i=1}^n I\{c_i = c, x_i = x, y_i = y\}$. A ‘dot’ notation indicates marginal counts, e.g., $n_{cx\bullet} = n_{cx0} + n_{cx1}$ and $n_{c\bullet\bullet} = n_{c0\bullet} + n_{c1\bullet}$. Further, let $\hat{P}r$ denote sample proportions with respect to the data table. The data are said to obey *sample positivity* if for all C values realized in the sample there is at least one unexposed subject and at least one exposed subject, i.e., $\hat{P}r(C = c) > 0$ implies $0 < \hat{P}r(X = 1|C = c) < 1$. If sample positivity holds, then (5) can be recast as follows. Consider the first part of (5) which estimates $E(Y_1)$. This can be rewritten according to

$$\frac{1}{n} \sum_{i=1}^n \left[\frac{y_i x_i}{\hat{\pi}(c_i)} - \left\{ \frac{x_i}{\hat{\pi}(c_i)} - 1 \right\} \hat{m}_1(c_i) \right] = \sum_{c \in \mathcal{C}} \frac{n_{c\bullet\bullet}}{n} h_1(c),$$

where

$$\begin{aligned} h_1(c) &= \frac{1}{n_{c\bullet\bullet}} \sum_{i:c_i=c} \left[\frac{y_i x_i}{\hat{\pi}(c)} - \left\{ \frac{x_i}{\hat{\pi}(c)} - 1 \right\} \hat{m}_1(c) \right] \\ &= \frac{n_{c11}}{n_{c\bullet\bullet} \hat{\pi}(c)} + \left(1 - \frac{n_{c1\bullet}}{n_{c\bullet\bullet} \hat{\pi}(c)} \right) \hat{m}_1(c). \end{aligned}$$

Assembling the pieces, we have $E(Y_1)$ estimated according to

$$\hat{E}(Y_1) = \sum_{c \in \mathcal{C}} \hat{Pr}(C = c) \{w_1(c)\hat{m}_1(c) + (1 - w_1(c))\hat{Pr}(Y = 1|X = 1, C = c)\} \quad (6)$$

where

$$w_1(c) = \frac{\hat{\pi}(c) - \hat{Pr}(X = 1|C = c)}{\hat{\pi}(c)}.$$

The analogous argument for Y_0 gives

$$\hat{E}(Y_0) = \sum_{c \in \mathcal{C}} \hat{Pr}(C = c) \{w_0(c)\hat{m}_0(c) + (1 - w_0(c))\hat{Pr}(Y = 1|X = 0, C = c)\}, \quad (7)$$

where

$$w_0(c) = \frac{\hat{Pr}(X = 1|C = c) - \hat{\pi}(c)}{1 - \hat{\pi}(c)}.$$

Thus $\hat{\Delta}_D$ can be viewed as the difference between (6) and (7), i.e.,

$$\begin{aligned} \hat{\Delta}_D = \sum_{c \in \mathcal{C}} \hat{Pr}(C = c) \{ & w_1(c)\hat{m}_1(c) + \\ & (1 - w_1(c))\hat{Pr}(Y = 1|X = 1, C = c) - \\ & w_0(c)\hat{m}_0(c) - \\ & (1 - w_0(c))\hat{Pr}(Y = 1|X = 0, C = c)\}. \end{aligned} \quad (8)$$

Despite the fact that they play a weighting role in (8), we refer to $w_0(c)$ and $w_1(c)$ as ‘discrepancies’ to avoid confusion with inverse-probability weights. Particularly, $w_0(c)$ and $w_1(c)$ reflect the discrepancy between parametric and nonparametric (or ‘saturated’) exposure model fits. The structure of (8) reveals that the estimator can be viewed as involving a weighted compromise between the model-based estimator for $(Y|X, C)$ and the saturated estimator for $(Y|X, C)$, with the provisos that (i), the weighting is different for estimating $E(Y_1)$ than for estimating $E(Y_0)$, i.e., $w_j(c)$ varies with j , (ii), the weighting is different within different C strata, i.e., $w_j(c)$ depends on c , and (iii), the weights $w_j(c)$ and $1 - w_j(c)$ sum to one, but are not constrained to be in the unit interval.

A casual user of double-robust estimators might be a little surprised to see such estimators cast in a form like (8). Whereas the double-robust intuition suggests a compromise between the parametric outcome model and the parametric exposure model, (8) appears to be more of a compromise between a parametric outcome model and a saturated outcome model. Noting that the exposure model residual $\hat{Pr}(X = 1|C = c) - \hat{\pi}(c)$ determines the numerators of $w_0(c)$ and $w_1(c)$, the nature of the compromise between the parametric and saturated outcome models is governed by the fit of the exposure model. Though it can't actually arise in practice, if all the discrepancies were set equal to one, i.e., $w_0(c) = w_1(c) = 1$ for all $c \in \mathcal{C}$, then (8) would reduce to (2), the estimator based on the parametric outcome model. On the other hand, if all the discrepancies were set to zero, (8) reduces to an estimator based only on the saturated outcome model. This situation can be approached in practice by making the exposure model bigger. That is, the saturated exposure model limit of the double-robust estimator and the saturated outcome model limit of the regression estimator coincide. As an aside, this is consistent with the point of Tan (2007), that the semiparametric information bound is approached from below by making the outcome model bigger and approached from above by making the exposure model bigger.

More to the point for present purposes, Bayesians might perk up a bit upon seeing (8). It vaguely resembles an average of two estimators arising from two competing models: a smaller outcome model parameterized by β and a saturated or nonparametric outcome model. Such model-averaged estimation is a core tool in the Bayesian toolbox. However, the form of the compromise in (8) seems strange. More weight is given to the saturated outcome model relative to the parametric outcome model when the parametric exposure model fits the data better. It would seem much more natural to instead give more weight to the saturated outcome model when the parametric outcome model fits the data worse. An obvious tweak to try is then to replace the two C -dependent discrepancies, $w_0(c), w_1(c)$, with a single value w_B not varying with c . This results in an estimator of the form

$$\begin{aligned} \hat{\Delta}_B = & \sum_{c \in \mathcal{C}} \hat{Pr}(C = c) \left[w_B \{m_1(c; \hat{\beta}) - m_0(c; \hat{\beta})\} + \right. \\ & \left. (1 - w_B) \{ \hat{Pr}(Y = 1|X = 1, c) - \hat{Pr}(Y = 1|X = 0, c) \} \right]. \end{aligned} \quad (9)$$

Specifically, we can set w_B to be the posterior probability that the parametric outcome model is correct, relative to the two-model universe whereby either this model or the encompassing saturated model are the only possibilities. Then (9) is indeed a Bayesian model-averaged estimate of the target, reducing

to the parametric outcome model estimator in the case of overwhelming evidence in favor of this model ($w_B \rightarrow 1$), and reducing to the saturated outcome model estimator in the case of overwhelming evidence against the parametric model ($w_B \rightarrow 0$). To be clear, the exposure model plays no role in (9).

It is a point of natural curiosity to see how (9) compares to its double-robust ‘cousin’ (8). Whereas there are many comparisons of double-robust estimators with parametric outcome model estimators (i.e., comparisons of $\hat{\Delta}_D$ and $\hat{\Delta}_P$, see, for instance, Lunceford and Davidian 2004, Kang and Schafer 2007, Lefebvre and Gustafson 2010), their similar flavor as per (8) and (9) suggests that comparison of $\hat{\Delta}_D$ and $\hat{\Delta}_B$ is more pertinent.

3 Implementation Details

To formalize the implementation of the Bayesian model-averaged estimator (9) it is necessary to specify a Bayesian version of the saturated outcome model. Let q be the probability vector of length $|\mathcal{C}|$ giving the distribution of C via $q_c = \Pr(C = c)$, and assign the prior $q \sim \text{Dirichlet}(k_q, \dots, k_q)$. Similarly, let $p_{cx} = \Pr(Y = 1 | X = x, C = c)$ and assign the prior $p_{cx} \sim \text{Beta}(k_p, k_p)$ for each $c \in \mathcal{C}$ and $x \in \{0, 1\}$. Note that the hyperparameter choices $k_q = 1$ and $k_p = 1$ give rise to uniform prior distributions. Note also that the average causal effect can be expressed as

$$\Delta = \sum_{c \in \mathcal{C}} q_c (p_{c1} - p_{c0}).$$

Under this prior specification, q and each p_{cx} are *a posteriori* independent, with $(q | d_{1:n})$ distributed as Dirichlet with parameters $k_q + n_{c\bullet\bullet}$ for each $c \in \mathcal{C}$. Similarly, $(p_{cx} | d_{1:n}) \sim \text{Beta}(k_p + n_{cx1}, k_p + n_{cx0})$. Thus the posterior mean of the target, which we will refer to as the saturated outcome-model estimator, is

$$\hat{\Delta}_S = \sum_{c \in \mathcal{C}} \frac{k_q + n_{c\bullet\bullet}}{|\mathcal{C}|k_q + n} \left[\frac{k_p + n_{c11}}{2k_p + n_{c1\bullet}} - \frac{k_p + n_{c01}}{2k_p + n_{c0\bullet}} \right]. \quad (10)$$

Similarly, the posterior variance can be worked out.

We motivated consideration of Bayesian model averaging by writing the double-robust estimator in the form (8), which only applies to datasets obeying sample positivity. However, provided proper priors are chosen ($k_q > 0, k_p > 0$), the posterior distribution underlying (10) exists regardless of data sparsity. Thus one could apply (10) to a dataset arising under violations of positivity without triggering a ‘red flag.’ In contrast, an exposure model producing some

fitted values $\hat{\pi}(c_i)$ very near zero or one would be viewed with some alarm. Thus reliance on outcome modeling alone could possibly hide a problem from view. On the other hand, a dataset violating sample positivity would involve some zero values for $n_{c1\bullet}$ and/or $n_{c0\bullet}$, which in turn would inflate the posterior variance accompanying (10). Implicitly then, the limitations of sparse data are acknowledged. Note as well that for a dataset obeying sample positivity, in the improper prior limit of $k_q = k_p = 0$, (10) does indeed reduce to the intuitive saturated model estimator based strictly on cell proportions.

For Bayesian model averaging we need to determine the marginal density of the data, i.e, with parameters integrated out. Since we only differ in the nonparametric versus parametric treatment of $(Y|X, C)$, for purposes of model comparison we can ignore the contributions of $(X|C)$ and (C) . The conjugate beta-binomial structure immediately gives

$$\begin{aligned} f_S(y|x, c) &= \int \prod_{i=1}^n f(y_i|x_i, c_i, p) f(p) dp \\ &= \prod_{x=0}^1 \prod_{c \in \mathcal{C}} \frac{\Gamma(n_{cx\bullet} + 1) \Gamma(n_{cx1} + k_p) \Gamma(n_{cx0} + k_p) \Gamma(2k_p)}{\Gamma(n_{cx\bullet} + 2k_p) \Gamma(n_{cx1} + 1) \Gamma(n_{cx0} + 1) \Gamma(k_p)^2}. \end{aligned} \quad (11)$$

Note that in the case of a uniform prior, $k_p = 1$, this reduces to $f_S(y|x, c) = \prod_{x=0}^1 \prod_{c \in \mathcal{C}} (n_{cx\bullet} + 1)^{-1}$. This suggests some stability of (11) in the non-sparse case that $Pr(X = x, C = c)$ is non-negligible for any combination of (x, c) , and n is large relative to $|\mathcal{C}|$. On the other hand, as $k_p \downarrow 0$, (11) tends to $\prod_{x=0}^1 \prod_{c \in \mathcal{C}} n_{cx\bullet} / \{n_{cx0} n_{cx1}\}$, which could also suffer from stability problems due to a rare outcome (or a rare non-outcome).

For the parametric outcome model parameterized by β and giving rise to $\hat{\Delta}_P$, we take a prior of the form $\beta \sim N(0, \Sigma)$, where typically Σ would be chosen to be very large, for the sake of comparability with a diffuse prior in the saturated outcome model. Then the marginal density can be approximated using the basic marginal density identity (Chib, 1995) and normal approximation. That is, for any choice of $\tilde{\beta}$,

$$\begin{aligned} f_P(y|x, c) &= \frac{f(y|x, c, \tilde{\beta}) \phi(\tilde{\beta}; 0, \Sigma)}{f(\tilde{\beta}|y, x, c)} \\ &\approx \frac{f(y|x, c, \tilde{\beta}) \phi(\tilde{\beta}; 0, \Sigma)}{\phi(\tilde{\beta}; \beta^*, \Sigma^*)}, \end{aligned} \quad (12)$$

where β^* and Σ^* are approximations to the posterior mean and variance of β . For instance, denoting the log-likelihood function as $l(\cdot)$ and the MLE as $\hat{\beta}$, we can take $(\Sigma^*)^{-1} = \Sigma^{-1} - l''(\hat{\beta})$ and $\beta^* = \Sigma^* \{-l''(\hat{\beta})\} \hat{\beta}$, and then set $\tilde{\beta} = \beta^*$.

The form of (12) reminds us that marginal densities are very sensitive to prior specification, via the second term in the numerator.

For a given dataset we can compute $f_S(y|x, c)$ according to (11) and $f_P(y|x, c)$ according to (12). With a prior assigning probability κ to the correctness of the parametric outcome model, and consequently probability $1 - \kappa$ to the saturated model, Bayes theorem gives the posterior probability assigned to the parametric model as

$$w_B = \frac{\kappa f_P(y|x, c)}{\kappa f_P(y|x, c) + (1 - \kappa) f_S(y|x, c)}.$$

In turn this value is used in determining the model-averaged estimator (9).

4 Simulation Comparisons

Some simulations are carried out to compare $\hat{\Delta}_D$ and $\hat{\Delta}_B$. We presume that C is comprised of $r = 4$ binary confounders, whose distribution arises from thresholding (at zero) four jointly normal variates with standardized marginals and all pairwise correlations being $\rho = 0.3$. The data-generating mechanisms for $(X|C)$ and $(Y|X, C)$ are taken as

$$\text{logit}Pr(X = 1|C) = -0.75 + C_1 + C_2, \quad (13)$$

$$\text{logit}Pr(Y = 1|X, C) = -1.25 + \gamma X + \lambda C_1 C_2 (1 - C_3), \quad (14)$$

for selected values of (γ, λ) . For a given setting the true value of Δ can be computed numerically via (1), upon noting that

$$Pr(C = c) = E \left(\Phi \left\{ (\rho^{-1} - 1)^{-1/2} Z \right\}^{r-g(c)} \left[1 - \Phi \left\{ (\rho^{-1} - 1)^{-1/2} Z \right\} \right]^{g(c)} \right),$$

where $g(c) = \sum_{j=1}^r c_j$, $Z \sim N(0, 1)$, and $\Phi()$ is the standard normal cdf.

Under each condition we simulate 10,000 datasets of size $n = 1000$ each. This choice of n , in tandem with the data-generating mechanisms chosen above, leads to non-sparse (C, X) data. The parametric exposure and outcome models fitted are both logistic models with intercepts plus main effects only. Thus α and β have five and six components respectively. Note that the exposure model is correctly specified, but the outcome model is only correctly specified under the simulation condition $\lambda = 0$, since otherwise the true relationship involves a three-way interaction between confounders. Initially we simulate under the condition $\gamma = 0$, so that the true causal effect is null, i.e., $\Delta = 0$. To study an interesting range of settings where the data are equivocal

about the appropriateness of the fitted outcome model, some pilot simulations were used to decide on three values of λ (0.6, 1.0, 1.4), corresponding to differing levels of empirical detectability of the outcome model misspecification. Since a given simulated dataset can be cast as $2^{r+1} = 32$ binomial responses with appreciable sample sizes, the deviance-based goodness-of-fit test for a generalized linear model can assess the evidence against the fit of the parametric outcome model. The three values of λ are selected so that this test, at significance level 0.05, has power around 20%, 50%, and 80%. We also simulate under the condition $\lambda = 0$, in which case the fitted outcome model is correct. However, results under $\lambda = 0$ and $\lambda = 0.6$ are generally similar, so the former are not shown.

Only the simple and intuitive form of the double-robust estimator is considered, with $\hat{\pi}(c) = \pi(c; \hat{\alpha})$ and $\hat{m}_j(c) = m_j(c; \hat{\beta})$ in (5). The Bayesian estimator is implemented with hyperparameters $k_q = k_p = 1$, particularly in light of the increased stability of the marginal density compared to smaller k_p values, as mentioned earlier. Equal prior weight is given to the parametric and saturated outcome models, i.e., $\kappa = 0.5$. For the parametric model we use a very diffuse prior for β by setting $\Sigma = (30^2)I_6$. We do indeed find that the Bayesian model comparison is sensitive to this choice, and pragmatically this setting is chosen because it yields a rough calibration: the Bayesian model comparison tends to not overwhelmingly support the parametric outcome model when the goodness-of-fit test has medium power to detect the misspecification of this model.

In the case of a null effect ($\gamma = 0$) and medium power to detect outcome model misspecification ($\lambda = 1$), the left panel of Figure 1 plots the estimators $\hat{\Delta}_B$ and $\hat{\Delta}_D$ against one another. While they are highly correlated, it should be noted that $\hat{\Delta}_P$ and $\hat{\Delta}_S$ are mutually highly correlated with them as well. The plot in the right panel is more telling, as it compares w_B , the posterior weight on the parametric model in the Bayesian estimator, against a ‘reverse-engineered’ weight for the double-robust estimator,

$$w_D = \frac{\hat{\Delta}_D - \hat{\Delta}_S}{\hat{\Delta}_P - \hat{\Delta}_S}.$$

To be clear, w_D is not an ingredient in the construction of the double-robust estimator. It simply reflects where the estimate lies between the saturated outcome model estimate and the parametric outcome model estimate, to facilitate comparison with the Bayesian estimate. The plot shows that in fact the double-robust estimator and the Bayes estimator very often make quite different compromises between the parametric outcome model estimator $\hat{\Delta}_P$ and the saturated outcome model estimator $\hat{\Delta}_S$. Whereas w_B lies in $[0, 1]$ by

Figure 1: Double-robust and Bayes estimators applied to data simulated under the condition $\gamma = 0$, $\lambda = 1$. The two estimates are plotted against one another in the left panel. The right panel compares the posterior weight on the parametric model, w_B , with the ‘reverse-engineered’ weight for the double-robust estimator, $w_D = (\hat{\Delta}_D - \hat{\Delta}_S)/(\hat{\Delta}_P - \hat{\Delta}_S)$. Note that for about 20% of simulated datasets, w_D lies outside the plotted range of $(-1, 2)$. Without this exclusion it is very hard to visualize the lack of association between w_B and w_D .

construction, often w_D does not. Moreover, even restricting to instances where w_D lies in $[0, 1]$, a positive association with w_B is not evident. This agrees with the reality that the double-robust compromise via $\{w_0(c), w_1(c)\}$ is governed exclusively by the fit of the parametric exposure model, whereas the Bayesian compromise via w_B is governed exclusively by the fit of the parametric outcome model.

Having confirmed that $\hat{\Delta}_D$ and $\hat{\Delta}_B$ are making quite different compromises between $\hat{\Delta}_P$ and $\hat{\Delta}_S$, we proceed to evaluate the effectiveness of these compromises. The first two panels of Figure 2 compare root-mean-squared estimation error (RMSE) under the conditions $\gamma = 0$ and $\gamma = 0.7$, the latter constituting a fairly strong causal exposure-outcome association. Unsurprisingly, the RMSE of $\hat{\Delta}_B$ tends to fall between that of $\hat{\Delta}_P$ and $\hat{\Delta}_S$. Thereby $\hat{\Delta}_B$ mitigates possible outcome model misspecification without necessarily having to incur all the variance of the saturated model estimator. A point of curiosity is that $\hat{\Delta}_S$ performs very well when $\gamma = 0$ and $\lambda = 0$. Presumably this arises because the hyperparameter choices $k_p = k_q = 1$ induce some shrinkage toward the null value of $\Delta = 0$, which happens to be the correct value in this setting. The performance when $\gamma = 0.7$ is more predictable, with the RMSE ‘winner’ between $\hat{\Delta}_P$ and $\hat{\Delta}_S$ switching as the extent of the outcome model misspecification increases.

The performance of $\hat{\Delta}_D$ is not particularly good. It always has higher RMSE than $\hat{\Delta}_B$, and in fact in the presence of relatively detectable outcome model misspecification it tends to have higher RMSE than even $\hat{\Delta}_P$. The price paid in terms of added variance to obtain robustification against outcome-model misspecification is not favorable.

A third simulation condition is constructed by retaining the outcome generating mechanism (14), with $\gamma = 0.7$, but introducing exposure model misspecification by replacing (13) with $\text{logit}Pr(X = 1|C) = -0.75 + 2C_1C_2$. The results appear in the third panel of Figure 2. As is to be expected, $\hat{\Delta}_D$ does particularly poorly when both exposure and outcome models are misspecified. On the other hand, $\hat{\Delta}_B$ ‘learns’ the outcome model misspecification and has RMSE almost as low as $\hat{\Delta}_S$.

A fourth simulation condition is constructed by restoring (13) as the data-generating mechanism for X , but replacing (14) with

$$\text{logit}Pr(Y = 1|X, C) = -1.25 + \lambda XC_2(1 - C_3),$$

again with $\lambda \in \{0.6, 1.0, 1.4\}$. The misspecification now arises because the fitted outcome model cannot capture the effect modification which is present. Results, which appear in the final panel of Figure 2, are broadly compatible with those of the other conditions.

5 Discussion

To reiterate, double-robust estimators in the causal inference setting can be cast as compromise estimators, with the compromise being between the postulated parametric outcome model for $(Y|X, C)$ and a saturated model treatment of $(Y|X, C)$. The compromise, which differs across strata of C , is such that a smaller exposure model residual for a stratum gives rise to less weight on the parametric version of the outcome model and more weight on the saturated version of the outcome model. This scheme lacks logical coherence, and a much more obvious strategy is to give more weight to the parametric outcome model when it fits the data better, as per a Bayesian model-averaged estimator. Viewed in this light, the double-robust and Bayesian model-averaged estimators do appear to be ‘cousins.’ Their empirical performance suggests that perhaps they are not such close cousins though. At least in the simple simulation settings here, the Bayesian estimator performed better than the double-robust estimator.

Figure 2: Root-mean-squared error of estimators $\hat{\Delta}_P$, $\hat{\Delta}_S$, $\hat{\Delta}_D$, $\hat{\Delta}_B$. In (i) the true data generating mechanism uses $\gamma = 0$, while (ii) is based on $\gamma = 0.7$. In (iii) the exposure model is misspecified as the $(X|C)$ data generating mechanism involves interaction. In (iv) the $(Y|X, C)$ data generating mechanism involves effect modification. Each panel gives results for three values of λ , referenced by the corresponding power of the deviance goodness-of-fit test for the outcome model, as indicated on the abscissa. Monte Carlo error is reported via 95% confidence intervals arising from the 10,000 simulated datasets under each condition.

A clear limitation of the work presented is the restriction to the easy setting of point outcome, exposure, and confounding variables, all of which are discrete. Moreover, non-sparse data are assumed. This is a simple canvas on which to illustrate the form of a Bayesian model-averaged estimator compared to a double-robust estimator. Clearly the former does not scale up to hard problems as easily as the latter. Thus the findings are not intended to be generally critical of double-robust estimators. Rather, we are simply trying to view Bayesian and double-robust procedures through the same lens, in the hope of encouraging Bayesian thinking in the causal inference domain, and spurring further research.

A different, but not completely unrelated, direction for the use of Bayesian model averaging in causal inference settings is pursued by Wang et al. (2011). As mentioned earlier, if the prior distribution treats outcome model parameters as independent of exposure model parameters, then the exposure model becomes irrelevant to inferring the average causal effect. Wang et al. (2011) give both conceptual and empirical support for treating the two sets of parameters as dependent *a priori*. They work in the context of large, but not necessarily saturated, outcome and exposure models, with priors that put some weight on zero values of coefficients. Prior dependence is introduced by asserting that for each potential confounding variable, having a non-zero coefficient in the exposure model implies a greater chance of having a non-zero coefficient in the outcome model. The model averaging ends up giving different weights to different subsets of the potential confounders being active in the outcome model, with these weights influenced in part by the fitted exposure model. This is a middle ground between model weights which are completely determined by the exposure model fit (as in $\hat{\Delta}_D$ here) and weights which are unaffected by an exposure model fit (as in $\hat{\Delta}_B$ here). The parallel is inexact though, as Wang et al. (2011) are compromising amongst an ensemble of different parametric outcome models, rather than between a single parametric outcome model and a saturated outcome model.

One direction to extend this work would be via more careful use of nonparametric Bayesian ideas to fit saturated outcome models in more complicated settings, with appropriate smoothing to deal with sparser data. In a related vein, Hill (2011) reports good results from using Bayesian tree-based models for flexible outcome modeling in the causal inference setting. More generally, there is renewed interest in parametric outcome modeling as a component of estimating causal effects via the g-formula (e.g. Taubman, Robins, Mittleman, and Hernán, 2009, Snowden, Rose, and Mortimer, 2011), and this

may provide another direction in which Bayesian ideas can usefully contribute to the causal inference milieu.

References

- Bang, H. and J. M. Robins (2005): “Doubly robust estimation in missing data and causal inference models,” *Biometrics*, 61, 692–972.
- Chib, S. (1995): “Marginal likelihood from the Gibbs output,” *Journal of the American Statistical Association*, 90, 1313–1321.
- Hill, J. (2011): “Bayesian nonparametric modeling for causal inference,” *Journal of Computational and Graphical Statistics*, 20, 217–240.
- Kang, J. D. Y. and J. L. Schafer (2007): “Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data,” *Statistical Science*, 22, 523–539.
- Lefebvre, G. and P. Gustafson (2010): “Impact of outcome model misspecification on regression and doubly-robust inverse probability weighting to estimate causal effect,” *International Journal of Biostatistics*, 6, issue 2 article 15.
- Lunceford, J. K. and M. Davidian (2004): “Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study,” *Statistics in Medicine*, 23, 2937–2960.
- Robins, J. M. and A. Rotnitzky (2001): “Comment on the Bickel and Kwon article: *Inference for semiparametric models: Some questions and an answer*,” *Statistica Sinica*, 11, 920–936.
- Robins, J. M., A. Rotnitzky, and M. van der Laan (2000): “Comment on the Murphy and van der Vaart article: *On profile likelihood*,” *Journal of the American Statistical Association*, 95, 431–435.
- Scharfstein, D. O., A. Rotnitzky, and J. M. Robins (1999): “Adjusting for non-ignorable drop-out using semiparametric non-response models (with discussion),” *Journal of the American Statistical Association*, 94, 1096–1120.
- Snowden, J. M., S. Rose, and K. M. Mortimer (2011): “Implementation of g-computation on a simulated data set: demonstration of a causal inference technique,” *American Journal of Epidemiology*, 173, 721–730.
- Tan, Z. (2007): “Understanding OR, PS and DR. Comment on the Kang and Schafer article: *Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data*,” *Statistical Science*, 22, 560–568.
- Tan, Z. (2010): “Bounded, efficient, and doubly robust estimation with inverse weighting,” *Biometrika*, 97, 661–682.

- Taubman, S. L., J. M. Robins, M. A. Mittleman, and M. A. Hernán (2009): “Intervening on risk factors for coronary heart disease: an application of the parametric g-formula,” *International Journal of Epidemiology*, 38, 1599–1611.
- van der Laan, M. J. and J. M. Robins (2003): *Unified Methods for Censored Longitudinal Data and Causality*, Springer.
- Wang, C., G. Parmigiani, and F. Dominici (2011): “Bayesian effect estimation accounting for adjustment uncertainty,” *Biometrics*, to appear.