Bayesian Analysis (2023) **18**, Number 2, pp. 639–694

Causal Inference Under Mis-Specification: Adjustment Based on the Propensity Score (with Discussion)*

David A. Stephens[†], Widemberg S. Nobre[‡], Erica E. M. Moodie[§], and Alexandra M. Schmidt[¶]

Abstract. We study Bayesian approaches to causal inference via propensity score regression. Much of Bayesian methodology relies on parametric and distributional assumptions, with presumed correct specification, whereas the extant propensity score methods in Bayesian literature have relied on approaches that cannot be viewed as fully Bayesian in the context of conventional 'likelihood times prior' posterior inference. We emphasize that causal inference is typically carried out in settings of mis-specification, and develop strategies for fully Bayesian inference that reflect this. We focus on methods based on decision-theoretic arguments, and show how inference based on loss-minimization can give valid and fully Bayesian inference. We propose a computational approach to inference based on the Bayesian bootstrap which has good Bayesian and frequentist properties.

Keywords: Bayesian bootstrap, Bayesian causal inference, de Finetti's representation, model mis-specification, propensity score adjustment.

1 Introduction

In the study of the causal relationship between an exposure (or treatment) and an outcome, bias in the estimation of the exposure effect may occur due to confounding if the exposure is not an experimental intervention. Confounding exists whenever the exposure assignment is dependent on predictors that also influence the outcome. If the dependence of outcome on exposure and predictors is modelled correctly, standard regression is adequate to obtain correct inference about the exposure effect. When correct specification cannot be guaranteed, the *propensity score* can be used to break the dependence between confounders and exposure, to create *balance* in the distribution of confounders across exposure groups, and facilitate correct inference. This paper studies how the propensity score can be deployed in a Bayesian causal analysis.

*Stephens, Moodie and Schmidt acknowledge support from the Natural Sciences and Engineering Research Council of Canada (NSERC). Nobre acknowledges support from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil, and Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), Brazil. Moodie holds a Canada Research Chair (Tier 1) in Statistical Methods for Precision Medicine, and further acknowledges support from a chercheur de mérite career award from the Fonds de recherche du Québec-Santé.

[†]Department of Mathematics and Statistics, McGill University, Canada, david.stephens@mcgill.ca

[‡]Departamento de Métodos Estatísticos, Universidade Federal do Rio de Janeiro, Brazil, widemberg@dme.ufrj.br

[§]Department of Epidemiology and Biostatistics, McGill University, Canada, erica.moodie@mcgill.ca

[¶]Department of Epidemiology and Biostatistics, McGill University, Canada, alexandra.schmidt@mcgill.ca

 https://doi.org/10.1214/22-BA1322

Adjustment via the propensity score can be carried out using regression, inverse weighting, stratification or matching. In regression settings, parametric models are proposed to represent the propensity score and the (expected) outcome given the propensity score. In frequentist approaches, adjustment is carried out by estimating parameters in the propensity score and the outcome models separately. In a fully Bayesian framework, such a two-step analysis is uncommon; it would be more natural to fit a single *joint* model for the treatment and outcome. This has led to discussion as to how Bayesian methods can be used in the causal setting, and even whether Bayesian methods are valid. There is a growing literature on sophisticated procedures for performing Bayesian causal analysis, but in a fully Bayesian framework, some aspects of the methodologies deployed appear non-standard and not justified via Bayesian logic.

We address these issues in this paper. Section 2 recaps the regression approach to causal estimation, and Section 3 describes how the key to valid Bayesian causal inference results from the assumption of exchangeability of the observable quantities to be modelled, which can be derived through de Finetti's representation, and a review of Bayesian adjustments using the propensity score. Section 4 describes Bayesian decision-theoretic inference which gives the framework for inference under mis-specification, and Section 5 gives the non-parametric computational strategy that we deploy. We provide simulation studies in Section 6, and conclude with a discussion in Section 7.

We note here that Bayesian methods that do not rely on the propensity score are also quite widely used: these methods utilize flexible parametric or non-parametric procedures to represent the outcome model as a function of the treatment and other predictors and attempt to avoid mis-specification. These methods are certainly useful, and the inferential theory supporting such one-stage analyses is more straightforward. However, such flexible outcome regression models cannot estimate the causal effect of interest in all cases, such as those where a more general target of inference is defined. These models are not the primary focus of this paper. Similarly, we will not discuss Bayesian matching methods in detail, although some comments are given in Section 7.

2 Background

To formulate causal inference estimation, *potential* or *counterfactual outcomes* are often used. Potential outcomes, $\{Y(z)\}$ for z in some putative treatment set, represent the outcomes that would be observed if treatment level Z was set to z. If exposure Z takes two levels labelled $\{0, 1\}$, the potential outcomes represent values of outcome Y that would be observed had exposure been set by intervention to $z = 0, 1$ respectively (Neyman, 1923; Rubin, 1974, 1985; Holland, 1986). We consider n subjects, and for the ith subject, let Y_i be the outcome of interest, Z_i be the exposure, and $X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^\top$ be a p-dimensional vector of confounders. We denote the data observation space by $\mathcal{X}$.

2.1 The average treatment effect (ATE)

The *Average Treatment Effect* (ATE) for a binary treatment is defined by

$$\tau = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]. \tag{2.1}$$

If Z is assigned *independently* of covariates X, $Z \perp X$, then the assumption of no unmeasured confounding or ignorability is trivially satisfied, implying that $\{Y(z)\} \perp Z|X$, and so the ATE is given by

$$\tau = \mathbb{E}_{Y|Z}[Y|Z=1] - \mathbb{E}_{Y|Z}[Y|Z=0] = \mathbb{E}_X[\mathbb{E}_{Y|X,Z}[Y|X,Z=1] - \mathbb{E}_{Y|X,Z}[Y|X,Z=0]]. \tag{2.2}$$

This definition differs notationally from the formulation via *counterfactual* or *potential outcomes* (Rubin, 1974) or the *do-operator* (Pearl, 2009), but under the independence assumption is equivalent. Equation (2.2) defines a marginal (over X) estimand, although conditional (subset-specific) estimands may also be defined. The calculation in (2.2) can be mimicked in the observed data to yield the estimate

$$\widehat{\tau} = \frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}_{Y|X,Z}[Y|X=x_i,Z=1] - \mathbb{E}_{Y|X,Z}[Y|X=x_i,Z=0]) \tag{2.3}$$

but this requires knowledge of the conditional expectation $\mathbb{E}_{Y|X,Z}[Y|X=x,Z=z]$ for all (x,z), and the assumption of no unmeasured confounding. Typically, this expectation would be represented using a regression model. If the regression model is misspecified, this approach can lead to incorrect inference about τ when the independence assumption does not hold and X is also associated with Y (i.e. in the presence of confounding).

2.2 The role of the propensity score

To estimate the ATE in the presence of confounding, Rosenbaum and Rubin (1983) showed that if the exposure assignment is ignorable and $b(X)$ is a *balancing score*, defined so that $X \perp Z|b(X)$, the ATE can be evaluated by averaging conditional means given Z and $b(X)$. If Z is binary, a typical choice for the balancing score is the propensity score, where $b(X) = \Pr[Z=1|X]$. Conditioning on the propensity score allows estimation of τ in the presence of confounding when the conditional model for Y given X and Z is not correctly specified by breaking the dependence between X and Z.

Propensity score regression represents the expected outcome conditioned on the exposure, confounders and propensity score. The ATE τ from (2.2) can be evaluated as

$$\tau = \mathbb{E}_X\left\{\mathbb{E}_{Y|X,B,Z}[Y|X,b(X),Z=1] - \mathbb{E}_{Y|X,B,Z}[Y|X,b(X),Z=0]\right\} \tag{2.4}$$

as $b(X)$ is a balancing score, using a model for $\mathbb{E}_{Y|X,B,Z}[Y|X=x_i,B=b(x_i),Z=z]$ for a modified version of (2.3) – see Rosenbaum and Rubin (1983) or Rubin (1985).

Typically $b(.)$ is represented using a parametric model, $b(x) \equiv b(x;\gamma)$, with γ estimated from the observed Z and X data. However, the balancing result $X \perp Z \mid B$ only holds when $b(X;\gamma)$ correctly characterizes the probability that $Z=1$ for any given X; this corresponds to the existence of a true value γ_0 of γ which defines the function precisely. For $\gamma \neq \gamma_0$, the method of proof of Rosenbaum and Rubin (1983) does not work to establish balance; see Supplement A.1 (Stephens et al., 2022) for a summary of the argument. Therefore, in a correctly specified parametric formulation of the propensity score, to yield balance, we **must** identify a single point in the parameter space, and use

that to define the propensity score. If γ_0 is not known, we must resort to substituting a consistent estimator $\widehat{\gamma}$ for γ_0, and then the required balancing result will hold asymptotically. In this paper we focus on parametric representations for the propensity score, but briefly discuss extensions in Section 7.

2.3 An illustrative model

Suppose the observed outcome data are generated according to the structural model

$$Y_i = X_{0i}\xi + Z_i\tau + \epsilon_i, \tag{2.5}$$

where for p-dimensional parameter ξ the term $X_{0i}\xi$ defines the true treatment-free mean model, and τ defines the ATE. If a regression model matching this specification is fitted using least squares, then the resulting estimator for τ is consistent. Similarly, if Z is assigned independently of X, then the estimator for τ is consistent even if the treatment-free mean model is mis-specified. However, if the model is mis-specified and Z and X are not independent, then the estimator of τ is in general inconsistent due to confounding. As demonstrated by Robins et al. (1992) the regression model

$$Y_i = b(X_i)\phi + Z_i\tau + \epsilon_i, \tag{2.6}$$

where ϕ is a scalar parameter, yields a consistent estimator of τ, albeit one whose variance is at least as large as the variance of the estimator arising from the correctly specified model. An 'augmented' model that contains an additional 'prognostic' linear predictor term $X_i\beta$ involving nuisance parameter β, that is, with

$$Y_i = X_i\beta + b(X_i)\phi + Z_i\tau + \epsilon_i \tag{2.7}$$

can be fitted as an attempt to reduce the variance for $\widehat{\tau}$; note, however, that the inclusion of this augmenting term is not necessary for consistent estimation of τ provided the propensity score model is correctly specified. Robins et al. (1992) provides a foundational treatment of the use of mis-specified regression models in the causal literature.

If a parametric model $b(x) = b(x;\gamma)$ is used, then parameter γ must be consistently estimated for the adequate adjustment. A plug-in estimation procedure, where γ is replaced by $\widehat{\gamma}$, and the regression utilizes $b(x;\widehat{\gamma})$ is typically used and corresponds to the 'feasible' E-estimator of Robins et al. (1992). It is justified in part by the asymptotic independence of $\widehat{\gamma}$ and $(\widehat{\phi},\widehat{\tau})$ (Henmi and Eguchi, 2004). The extended model (2.7) has the advantage that provided the $X_i\beta$ component is correctly specified (i.e. reflects the data-generating mechanism), the estimator of τ is consistent even if the propensity score is not correctly specified. This is known as *double robustness*. If the data generating model contains a more general treatment effect structure, the propensity score regression approach must be modified. For example, if the model takes the form

$$Y_i = X_{0i}\xi + Z_iM_{0i}\psi + \epsilon_i, \tag{2.8}$$

where ψ is a $q \times 1$ vector parameter and M_{0i} is a $1 \times q$ vector of predictors, the ATE is $\mathbb{E}[M_{0i}]\psi$. This quantity (and the parameters ψ) can be consistently estimated using the propensity score regression approach based on the model

$$Y_i = X_i\beta + b(X_i)M_{0i}\phi + Z_iM_{0i}\psi + \epsilon_i, \tag{2.9}$$

where now ϕ is a $q \times 1$ parameter, that is, with an interaction term involving the propensity score configured to match the treatment effect structure. This construction is necessary to ensure that confounding via the open paths that involve the interaction terms is also removed by conditioning on the propensity score. Further modifications are necessary if the structural model is extended beyond the linear case, or if multi-stage treatments are considered; see Supplement A.2.

3 Bayesian inference under exchangeability

The key construction for any Bayesian inference problem to be solved under an assumption of exchangeability of the observable quantities is de Finetti's representation, which leads to the standard definitions of likelihood, prior, parameters and the notion of 'correct specification'. If $\{O_i\}_{i=1}^{\infty}$ is a sequence of exchangeable observable quantities, where each O_i takes values on $\mathcal{X}$, the de Finetti representation for any collection of size $n \geq 1$ of the observables defines their joint density of $p_O(o_{1:n})$, which in turn defines the posterior distribution, $\pi_n(\theta)$ where

$$\pi_n(\theta) = \frac{\prod_{i=1}^{n} f_O(o_i;\theta)\pi_0(\theta)}{\int \prod_{i=1}^{n} f_O(o_i;t)\pi_0(t)dt} = \frac{\prod_{i=1}^{n} f_O(o_i;\theta)\pi_0(\theta)}{p_O(o_{1:n})}$$

say, where $\pi_0(\theta)$ is the prior distribution on parameter $\theta \in \Theta$. In the causal setting, θ indexes the joint distribution of the observables $o_{1:n} = (x_{1:n}, y_{1:n}, z_{1:n})$. We have

$$p_O(o_{1:n}) = p_X(x_{1:n})p_{Z|X}(z_{1:n}|x_{1:n})p_{Y|Z,X}(y_{1:n}|x_{1:n}, z_{1:n}). \tag{3.1}$$

Decomposing $\theta = (\eta, \gamma, \zeta)$, and assuming independent prior structure, we require that the three components in (3.1) each admit a de Finetti representation based on what we term *conditional* exchangeability assumptions (Saarela et al., 2022). For $n \geq 1$, the triples (X_i, Y_i, Z_i), $i = 1, \ldots, n$, are assumed to be conditionally independent given θ, and the Bayesian specification is completed after defining a probability distribution $\pi_0(\theta) \equiv \pi_0(\eta)\pi_0(\gamma)\pi_0(\zeta)$. Specifically

$$\begin{aligned} p_X(x_{1:n}) &= \int \prod_{i=1}^{n} f_X(x_i;\zeta)\pi_0(\zeta)d\zeta, \\ p_{Z|X}(z_{1:n}|x_{1:n}) &= \int \prod_{i=1}^{n} f_{Z|X}(z_i|x_i;\gamma)\pi_0(\gamma)d\gamma, \\ p_{Y|X,Z}(y_{1:n}|x_{1:n}, z_{1:n}) &= \int \prod_{i=1}^{n} f_{Y|X,Z}(y_i|x_i, z_i;\eta)\pi_0(\eta)d\eta. \end{aligned} \tag{3.2}$$

This formulation proposes that in the data generating model the Y_is are conditionally independent given the (X_i, Z_i) pairs and parameter η. This is a standard assumption in

the frequentist parametric sequel, and would hold in any conventional regression model. The full probability model for observables and unobservables can be decomposed as

$$f_X(x_{1:n};\zeta)f_{Z|X}(z_{1:n}|x_{1:n};\gamma)f_{Y|Z,X}(y_{1:n}|x_{1:n},z_{1:n};\eta)\pi_0(\zeta)\pi_0(\gamma)\pi_0(\eta)$$

with the usual conditional independence decompositions of the 'likelihood' terms. The prior independence assumption is natural in light of the conditional exchangeability formulation in (3.1). This leads to the posterior distribution $\pi_n(\eta,\gamma,\zeta)$ in the usual way. Under standard assumptions, the posterior distribution converges as $n \longrightarrow \infty$ to a unique degenerate limit at a single point $(\eta_0,\gamma_0,\zeta_0)$, and the data generating model is in fact factorized $f_X(x;\zeta_0)f_{Z|X}(z|x;\gamma_0)f_{Y|X,Z}(y|x,z;\eta_0)$. The Bayesian model is considered correctly specified if this limiting behaviour holds.

The formulation above is parametric, but extensions to the non-parametric case where θ is infinite dimensional are straightforward. We regard a valid Bayesian approach as one which relies on the de Finetti representation for observable quantities in the data generating model, with inference following a decision-theoretic argument, as outlined in Section 4. Note that under exchangeability, the de Finetti representation defines (up to the choice of the prior) the complete probabilistic specification for the model, whether or not we opt to depend on it for inference. Furthermore, it determines the frequentist characteristics of Bayesian inference procedures.

In the context of adjustment using the propensity score, some modifications to the Bayesian methodology are necessary for the reasons described in Section 2.2. Specifically, the outcome component $f_{Y|Z,X}(y|x,z;\eta)$ in (3.2) is not explicitly used; instead an outcome model that conditions on the propensity score is deployed, and for correct confounding adjustment this model needs to rely on the true value of the parameter γ_0. This issue has been approached in several ways that we describe in the next section.

3.1 Existing approaches to Bayesian causal inference

A parametric Bayesian analysis based on the true model (2.5) or proposed model (2.6) would proceed in a standard fashion. The marginal posterior distributions for τ derived from (2.5) and (2.6) are in general different. However, model (2.5) is essentially an 'oracle' model to which we do not have access. In this case, it is relatively straightforward to show that as n increases, the posterior distribution for τ derived from (2.6) becomes concentrated at the true (data generating) value of the ATE present in the structural model, despite the mis-specification present in (2.6). This asymptotic calculation hypothesizes an increasingly large sample of data drawn from the same probability model. These arguments hold for the extended model (2.7).

If $b(x)$ is replaced by $b(x;\gamma)$ in (2.6) or (2.7), and γ is treated as an unknown parameter, the question arises whether this log-likelihood, coupled with the log-likelihood for γ itself, should be used as the basis of a three-parameter posterior in the parameters (ϕ,τ,γ). It is not evident on first inspection whether this posterior, or the bivariate posterior based on (ϕ,τ) for some plug-in value $\widehat{\gamma}$ as in the frequentist approach is justified in a formal Bayesian inference setting. Zigler (2016) emphasizes that model

mis-specification is a key issue, studies the most commonly used approaches, and describes directions in which the Bayesian formulation may be developed productively. We summarize some of the key elements below.

Joint Bayesian modelling: The Bayesian propensity score model proposed by McCandless et al. (2009) assumes a joint parametric model and for (2.7), the joint model considers conditional models $f_{Z|X}(z|x) = f_{Z|X}(z|x;\gamma)$ and $f_{Y|X,Z,B}(y|x,z,b(x)) = f_{Y|X,Z,B}(y|x,z,b(x;\gamma);\beta)$. This leads to a joint likelihood function for (γ,β,ϕ,τ):

$$\mathcal{L}_n(\beta,\phi,\tau,\gamma) = \prod_{i=1}^{n} f_{Z|X}(z_i|x_i;\gamma) f_{Y|X,Z}(y_i|z_i,x_i,b(x_i;\gamma);\beta,\phi,\tau), \tag{3.3}$$

with inference carried out using Markov chain Monte Carlo (MCMC) – specifically the Gibbs sampler – by sampling recursively from the two full conditional distributions $\pi(\gamma|\beta,\tau,y_{1:n},z_{1:n},\mathrm{x}_{1:n})$ and $\pi(\beta,\phi,\tau|\gamma,y_{1:n},z_{1:n},\mathrm{x}_{1:n})$, along with any additional parameters that appear in the proposed models.

Cutting feedback: The joint model based on (3.3) does not create the required balance, or correct appropriately for confounding, due the presence of what is termed *feedback*, and the marginal posterior for τ does not concentrate at the true value. To overcome this, McCandless et al. (2010) proposed that the full conditional distribution of γ should be independent from the rightmost term of the likelihood in equation (3.3);

$$\pi_n(\gamma) \propto f_{Z|X}(z_{1:n}|\mathrm{x}_{1:n};\gamma)\pi_0(\gamma). \tag{3.4}$$

This is known as the *cutting feedback* approach which can be implemented as follows: a sample of size L of $\pi_n(\gamma)$ is produced, and then used to construct propensity score sampled values $b_i^{(l)} = \Pr[Z_i = 1|X_i = x_i;\gamma^{(l)}]$, where $\gamma^{(l)}$ denotes the l-th sample from $\pi_n(\gamma)$. Then, a sample of size L is obtained for the outcome parameters, with the l-th sample, for $l = 1,\cdots,L$ being generated from

$$\pi_n^{(l)}(\beta,\phi,\tau) \propto f_{Y|X,Z,B}(y_{1:n}|\mathrm{x}_{1:n},z_{1:n},b_{1:n}^{(l)};\beta,\phi,\tau)\pi_0(\beta,\phi,\tau). \tag{3.5}$$

Two-step inference: A *two-step* procedure (Zigler et al., 2013) assumes complete separation between the exposure and outcome models. First, a point estimate of γ is obtained from $\pi_n(\gamma)$ computed via (3.4). This point estimate is then used to construct an estimate of the propensity score, $\widehat{b}_i = f_{Z|X}(1|x;\widehat{\gamma})$, which is then plugged into the outcome model. A posterior sample is then obtained from

$$\pi_n(\beta,\phi,\tau) \propto f_{Y|X,Z,B}(y_{1:n}|z_{1:n},\mathrm{x}_{1:n},\widehat{b}_{1:n};\beta,\phi,\tau)\pi_0(\beta,\phi,\tau). \tag{3.6}$$

In the cutting feedback and two-step approaches, it is not immediately clear how the inferential uncertainty concerning γ in the estimation of τ should be handled. Several methods to evaluate the variance of the posterior distribution of τ have been proposed; see for example Kaplan and Chen (2012). The cutting feedback approach attempts

to account for the uncertainty in the estimation of γ by direct sampling from $\pi_n(\gamma)$ in (3.4) with posterior computation for the remaining parameters being carried out conditionally on each sampled value of γ; the two-step approach as described above ignores the uncertainty in γ, but an adjustment based on Taylor expansions around $\widehat{\gamma}$ can be implemented (Graham et al., 2016).

Several of the applications described above do not use regression on the propensity score itself, but instead attempt adjustment based on a stratification procedure by defining a factor predictor using quantiles of the observed propensity score distribution. Each observed $b(x_i)$ is converted into a categorical variable corresponding to the quantile interval within which $b(x_i)$ lies. This method can be very effective in practice, but we note that it can only provide approximate correction for confounding, and will yield consistent estimation only if the number of categories used (and hence the number of associated coefficients) grows with n. In effect, the stratification approach provides an effective way of approximating the true data generating mean model rather than adjusting for confounding in the same way that the Robins et al. (1992) model operates.

The methods outlined in this section have elements that do not follow the classical Bayesian formulation outlined above, and therefore there is debate as to whether they can be formally classed as fully Bayesian procedures. However, it transpires that they can in fact be viewed as 'generalized' Bayesian procedures (Bissiri et al., 2016) that use a specific decision-theoretic framework; we present a detailed exposition in Supplement B, and discuss other related formulations in Section 4.2. In this paper, we present a valid conventional Bayesian approach that relies on a non-parametric formulation based on consciously mis-specified models.

3.2 Current literature

It is not universally accepted that fully Bayesian inference is possible using the fitted propensity score in a regression as in Robins et al. (1992), or via other methods such as inverse probability weighting (see Supplement A.2), as such methods involve a plug-in strategy that is not fully Bayesian; see the discussion of Saarela et al. (2015). For example, it is contended that if the propensity model is unknown and must be estimated, the plug-in estimation of $b(x)$ is contrary to conventional Bayesian inference based on a 'likelihood times prior' formulation.

Despite such objections, there has been a marked increase in research on Bayesian methods for causal quantities based on propensity score adjustment (see, for example, Adhikari et al., 2020; Comment et al., 2019; Geneletti et al., 2019; Samartsidis et al., 2020; Nethery et al., 2020; Liu et al., 2020). While sharing a common goal of adjusting for bias due to confounding with a Bayesian lens, it is clear that consensus has not been reached on how to perform inference with propensity score-based approaches. For instance, Comment et al. (2019), Nethery et al. (2020), and Liao and Zigler (2020) all use an approach that succeeds in cutting feedback, using the propensity score to create a matched sample; these authors view the matching step as part of a 'design' rather than analytic phase of the analysis. Bornn et al. (2019) use a form of joint modelling of the treatment and outcome, as do Ray and van der Vaart (2020). Two-step approaches are

widely used, although there is no agreement in the literature on whether to plug in fixed quantities (such as a posterior mean or mode) or random (draws from the posterior). For instance, Vegetabile et al. (2020) use a Bayesian non-parametric approach to estimate the propensity score which is then plugged into a standard (frequentist) estimator of the average treatment effect. Some authors try multiple approaches; for example, Wang et al. (2012) use both a joint modelling and a two-step approach, where the latter is employed as a device to cut feedback. Wang and Rosner (2019) use propensity score regression, conditioning on the expected value of the propensity score. In contrast, Xu et al. (2018) take a propensity regression approach to estimate the quantile (rather than average) treatment effect, conditioning on draws from the posterior distribution of the propensity score. Hahn et al. (2020) sample the estimated propensity score's posterior distribution, incorporating the samples into a nonlinear regression model for the outcome (including heterogeneous treatment effects) using additive regression trees. Liu et al. (2020) use inverse weighting in a two-step procedure and propagate uncertainty using the Bayesian bootstrap; see also Graham et al. (2016). Other authors have combined aspects of Bayesian and frequentist modelling to address complex models. Davis et al. (2019) use approximate Bayesian methods to estimate both a propensity score and an outcome model, and then combine predictions from these into a frequentist doubly-robust estimator in a spatial modelling context. Antonelli et al. (2022) consider the high-dimensional case, also using Bayesian methods to estimate both a propensity score and an outcome model and computing a doubly-robust estimator by averaging over draws from the posterior distribution of the parameters of these models.

Models (2.6) or (2.7) are simple compared to some of the approaches described above, but serve to illustrate the relevant theoretical issues. Flexible models that attempt to model the outcome directly can be extremely useful in capturing the causal relationship by overcoming issues of mis-specification, or to represent the propensity score and, despite some drawbacks, such models can be effective. The methods described in this paper are relevant to any form of propensity score analysis, that is, the use of a propensity score in the analytic stage, rather than in a design phase, such as when it is employed to perform matching to create an analytic dataset.

4 Bayesian decision-theoretic inference

In this section, we recap details of Bayesian decision-theoretic procedures to build a framework in which to consider the setting where data are generated according to some likelihood model $f_O(.\,;\theta_0)$ which we cannot and do not need to specify correctly. Rather, we will focus on the consequences for inference in a second, *alternative* model with density f, acknowledging that this density is mis-specified. As noted above, causal inference is most often employed when the functional form of the dependence of the outcome on the treatment and covariates is not known, such that reliance on the propensity score is required to break the confounding that biases the estimated effect of treatment.

The Bayes estimate is a function of the observed data that minimizes the Bayes risk, or the posterior expected loss for some loss function $\ell(t,\theta) : \Theta \times \Theta \longrightarrow \mathbb{R}^+$, that is

$$\widehat{\theta} = \arg\min_{t\in\Theta} \mathbb{E}_{\pi_n}[\ell(t,\theta)] = \arg\min_{t\in\Theta} \int \ell(t,\theta)\pi_n(\theta)\ d\theta.$$

If the loss function can be written

$$\ell(t,\theta) = \int u(s,t) f_O(s;\theta)\ ds = \mathbb{E}_{f_O}[u(S,t);\theta] \tag{4.1}$$

for some function $u(s,t) : \mathcal{X} \times \Theta \longrightarrow \mathbb{R}^+$, then the estimation problem can be rewritten

$$\widehat{\theta} = \arg\min_{t\in\Theta} \int u(s,t) \left\{ \int f_O(s;\theta)\pi_n(\theta)\ d\theta \right\} ds = \arg\min_{t\in\Theta} \mathbb{E}_{p_n}[u(S,t)], \tag{4.2}$$

where $p_n(s)$ is the posterior predictive distribution implied by the Bayesian specification. For example, if, for $t \in \Theta$, $u(s,t) = -\log f_O(s;t)$, (see Bernardo (1979)) we have that

$$\widehat{\theta} = \arg\max_{t\in\Theta} \int \left\{ \int \log f_O(s;t) f_O(s;\theta)\ ds \right\} \pi_n(\theta)\ d\theta. \tag{4.3}$$

For example, in the Normal model with $f_O(s;t) \equiv Normal(t,1)$, the calculation becomes

$$\arg\min_{t\in\Theta} \iint (s-t)^2 \phi(s-\theta)\ ds\pi_n(\theta)\ d\theta = \int \left\{ \int s\phi(s-\theta)\ ds \right\} \pi_n(\theta)\ d\theta = \int \theta\pi_n(\theta)\ d\theta,$$

where $\phi(.)$ is the standard Normal pdf, that is, the estimate is the posterior mean. Equation (4.2) indicates that Bayesian parameter estimation can be formulated as a prediction problem if an appropriate loss function is defined. Equation (4.1) depends on an integral over a single variable s that can be taken to be a single 'future' variate drawn from $f_O(.;\theta)$, but the formulation extends to m independent 'future' variates, and can be expressed via the m-fold posterior predictive.

4.1 Bayesian inference under mis-specification

Broadly, mis-specification of a Bayesian model arises either if the 'likelihood' model – the conditional density of the observables given the parameters – does not match f_O, or if the true value θ_0 does not lie in the support of the prior. In such cases, there is no guarantee of reliable statistical behaviour. However, certain mis-specified models can have utility; for example, the model in (2.6) is not the data generating model, and yet can provide consistent frequentist inference provided the propensity score model is correctly specified. In this section we examine some aspects of mis-specification.

Suppose initially we retain the data generating likelihood model $f_O(.;\theta_0)$, but consider the implications for inference in an alternative model with density f with support $\mathcal{X}$, parameterized by $\vartheta \in \Theta'$. That is, while assuming the data are generated by f_O, we wish to perform inference for ϑ acknowledging that f is mis-specified. Conventional Bayesian inference for ϑ can be performed using a likelihood based on f, but it is difficult to justify the resulting posterior as the focus of inference since the model is mis-specified; see, for example, Walker (2013) and its discussion. The decision theoretic framework can still be deployed, however: define loss function $\ell(t',\theta) : \Theta' \times \Theta \longrightarrow \mathbb{R}^+$ by

$$\ell(t',\theta) = \mathcal{K}(f_O(.;\theta), f(.;t')) = \int \log\left(\frac{f_O(s;\theta)}{f(s;t')}\right) f_O(s;\theta)\ ds = \mathbb{E}_{f_O}[u_\theta(S,t');\theta],$$

where $u_\theta(s, t') = \log\left(f_O(s;\theta)/f(s;t')\right)$, which extends the calculation in (4.1) to allow the function $u(.,.)$ to depend on θ – note that the resulting optimization over t' may still not depend on θ. By arguments equivalent to those leading to (4.3), we have that

$$\widehat{\vartheta} = \arg\max_{t' \in \Theta'} \int \left\{ \int \log f(s;t') f_O(s;\theta)\ ds \right\} \pi_n(\theta)\ d\theta. \tag{4.4}$$

If standard Bayesian theory is used to compute the posterior for θ, then the posterior for ϑ also may be computed (at least numerically) via the deterministic transformation implied by (4.4). For example, to compute the posterior distribution for ϑ, we may use a simulation-based strategy: if a single sampled variate $\theta^{(l)}$ is generated from $\pi_n(\theta)$, then we may convert this into a sampled variate $\vartheta^{(l)}$ from the posterior for ϑ by performing the transformation

$$\vartheta^{(l)} = \arg\max_{t' \in \Theta'} \int \log f(s;t') f_O(s;\theta^{(l)})\ ds \tag{4.5}$$

and then replicate this for $l = 1, \ldots, L$. In general, the integral with respect to s may not be analytically tractable, but can be approximated using Monte Carlo by sampling $s_k, k = 1, \ldots, N$ from $f_O(.\,;\theta)$, and computing

$$\vartheta^{(l)} = \arg\max_{t' \in \Theta'} \sum_{k=1}^{N} \log f(s_k;t').$$

The Kullback-Leibler loss can be modified to reflect quantitative statements about ϑ in the alternative model. For example, we may specify

$$u_\theta(s, t') = \log\left(\frac{f_O(s;\theta)}{f(s;t')}\right) + \log u_0(t') \tag{4.6}$$

for some non-negative function $u_0(.)$ with domain Θ' that does not depend on θ or s. This additional term essentially functions as (minus) a log prior distribution on ϑ, although as we explicitly acknowledge that the model f is mis-specified, and ϑ has no real-world interpretation, this interpretation may be problematic for some Bayesians. In any case, the maximizations leading to the estimate $\widehat{\vartheta}$ in (4.4) and sampled variate $\vartheta^{(l)}$ in (4.5) can be modified accordingly.

4.2 Conscious mis-specification and modularization

The formulation of inference under mis-specification is inspired by the reasoning that inference concerning an alternative target model may be of interest in its own right (for example, simplicity of interpretation). In addition, note that the calculation in (4.2) does not require explicit computation of the posterior $\pi_n(\theta)$, so in principle a representation of, approximation to, or samples drawn directly from the posterior predictive distribution $p_n(s)$ can be used to compute the estimate or posterior sample for ϑ. Such a strategy would be useful if complex models such as flexible Bayesian models or artificial neural networks were used to construct prediction techniques. In the causal inference

setting, the parameters of interest are not defined in the actual data generating model, but rather are quantities defined with respect to some hypothetical data generating process where confounding is not present. It is possible to construct examples where even a correctly specified regression model, say, cannot yield consistent estimators of the causal effect of interest, although these examples typically need to have more complex structural forms than those in (2.5), involving multiple treatments. We return to these examples in Supplement A.2.

Such 'conscious' mis-specification has direct relevance in the causal setting, but it has also been argued that similar calculations, where the data generating model does not correspond to the inference model, may be relevant in Bayesian calculations more generally. Bayarri et al. (2009) argue for a form of Bayesian inference based on 'modularization' of the model, where a form of stagewise analysis in complex models is used. Motivated by formulations based on Bayesian mis-specification, Jacob et al. (2017) provide extensive evidence that such modularized inference can be advantageous in Bayesian settings, including a study of the empirical properties of propensity score stratification estimators using the methods from Section 3.1. The approaches described in Jacob et al. (2017) are entirely parametric in nature, and therefore approach the issue of mis-specification from a different perspective, by deliberately using mis-specified models for statistical advantage; see also Pompe and Jacob (2021). Finally, in the causal setting of Section 2.3, a very specific approach to modularized inference must be adopted. We discuss this in more detail below.

5 Bayesian non-parametric formulation

We now implement the decision-theoretic ideas from Section 4 in the causal setting using a non-parametric model.

5.1 The Dirichlet process model

In order to weaken the parametric assumption concerning f_O, we allow θ to become an infinite dimensional parameter describing the distribution of O. Suppose that $F_O(.)$ parameterizes unknown distribution function of the data with true value F_0, such that in reality $O_1, \ldots, O_n \sim F_0(.)$ are independent; this interpretation is consistent with the de Finetti formulation, with the $F_0(.)$ interpreted as the limiting empirical cdf derived from the exchangeable sequence. The Dirichlet process model $DP(\alpha, G)$ is a probability measure on the set of distribution functions with countable support, with probabilities $\omega_j, j = 1, 2, \ldots$ at locations $x_j, j = 1, 2, \ldots \in \mathcal{X}$, and the $DP(\alpha, G)$ model induces randomness by drawing the ω_js via a probabilistic algorithm that depends on α – commonly the so-called 'stick-breaking' algorithm is used – and the x_j independently from G. In the most common form of Bayesian non-parametric analysis, the Dirichlet process acts as a prior for parameter F_O; hyperparameter $\alpha > 0$ acts as a concentration parameter, and $G(.)$ is a prior (base) distribution with domain $\mathcal{X}$. In light of data $o_{1:n}$, the resulting posterior distribution is also a Dirichlet process $DP(\alpha_n, G_n)$ where $\alpha_n = \alpha + n$ and $G_n(.) = w_n G(.) + (1 - w_n)\widehat{F}_n(.)$, where $w_n = \alpha/(\alpha + n)$ and $\widehat{F}_n(.)$ is the empirical measure derived from $o_{1:n}$.

It is straightforward to generate samples from $DP(\alpha_n, G_n)$ (that is, randomly generated distributions that represent sampled versions of 'parameter' F_O) and also from the implied model for the observable quantities in light of the data (that is, a randomly generated posterior predictive distribution). Furthermore, the Dirichlet process posterior becomes concentrated at the data generating model F_0 in the limit as $n \longrightarrow \infty$ (Ghosal and van der Vaart, 2017, section 4.7), and provides a consistent estimation procedure.

With this relaxation of the parametric assumption about the data generating model, the calculations from Section 4.1 can be reproduced. The Bayes estimate again results from a minimum loss calculation based on the posterior predictive distribution. When the posterior distribution is the $DP(\alpha_n, G_n)$ distribution, we have, for example replicates $\vartheta^{(l)}, l = 1, \ldots, L$ sampled from the posterior for ϑ given by

$$\vartheta^{(l)} = \arg\max_{t' \in \Theta'} \sum_{j=1}^{\infty} \omega_j^{(l)} \log f(s_j^{(l)}; t'), \tag{5.1}$$

where $\{\omega_j^{(l)}, j = 1, 2, \ldots\}$ are a sample of probabilities drawn by, say, stick-breaking with parameter α_n, and $\{s_j^{(l)}, j = 1, 2, \ldots\}$ are drawn independently from G_n. In practice, the infinite sum is truncated by machine accuracy, as the ω_j values decrease in expectation as j increases. The $\{\omega_j\}$ may also be drawn such that they are decreasing in magnitude, rendering the truncation straightforward to implement.

5.2 The Bayesian bootstrap

The Bayesian bootstrap (Rubin, 1981) posits a multinomial likelihood on the finite set $\mathcal{O} = \{o_1, \ldots, o_n\}$ with unknown probabilities $\varpi = (\varpi_1, \ldots, \varpi_n)$ attached to each element, and combines this with a $Dirichlet(\alpha, \ldots, \alpha)$ prior to yield the posterior distribution for ϖ to be $Dirichlet(\alpha + 1, \ldots, \alpha + 1)$. Taking $\alpha \longrightarrow 0$ yields the Bayesian bootstrap, in which the predictive distribution is represented

$$p_n(o) = \sum_{i=1}^{n} \omega_i \delta_{o_i}(o), \tag{5.2}$$

where $\omega \sim Dirichlet(1, \ldots, 1)$, identical to the posterior distribution.

Originally the Bayesian bootstrap was proposed as a heuristic strategy, but its theoretical properties have since been widely studied; see for example Lo (1987); Cheng and Huang (2010) and Ghosal and van der Vaart (2017). The argument confirming that this strategy was in fact producing approximate Bayesian posterior statements was formalized by Newton and Raftery (1994). The procedure was used in the Bayesian causal settings in Saarela et al. (2015) and Saarela et al. (2016): in those papers, the utility argument is made explicit, and the log-density utility is justified by considering a hypothetical *experimental* data generating mechanism that is explicitly misspecified (compared to the *observational* data generating model). See Chamberlain and Imbens (2003), Graham et al. (2016) and Lyddon et al. (2019) for examples and generalizations.

The Bayesian bootstrap results from a Dirichlet process specification for the observed data $o_{1:n}$ in the limiting case $\alpha \longrightarrow 0$. Sampling from the posterior predictive coincides with the Bayesian bootstrap; if $\omega = (\omega_1, \ldots, \omega_n) \sim \mathit{Dirichlet}(1, 1, \ldots, 1)$, (5.2) yields the estimation procedure

$$\vartheta = \arg\max_{t' \in \Theta'} \sum_{i=1}^{n} \omega_i \log f(o_i; t') \tag{5.3}$$

with ϑ now being a random quantity as ω is random. The summation in this expression is a *deterministic* function of ω for every fixed t'; therefore the corresponding ϑ is also a deterministic function of ω. Hence, once we have sampled the weights in the Dirichlet process formulation, a transformation yields ϑ, and thus ϑ is simply a functional of the Dirichlet process posterior on F_O. Therefore the posterior sample formed by repeatedly sampling the Dirichlet weights to yield $\omega^{(1)}, \ldots, \omega^{(L)}$, with subsequent transformations to yield $\vartheta^{(1)}, \ldots, \vartheta^{(L)}$ is an exact sample from the posterior distribution for ϑ. Such inference is a fully Bayesian expression of posterior beliefs concerning the target of inference under the Bayesian non-parametric formulation. A proper prior $\pi_0(\vartheta)$ can be incorporated by modifying the specified utility function as in (4.6).

5.3 Bayesian inference for the structured causal model

For the causal inference problem with observed data $o_{1:n} = (x_{1:n}, y_{1:n}, z_{1:n})$, for a parametric analysis, we may compute the posterior distribution for $\theta = (\eta, \gamma, \zeta)$ using a factorization of the full model as in (3.2). We can also define the alternative model to respect the entire factorization, or target some component of interest. For example, a conditional model for Y given (X, Z) might be targeted, with $u_\theta(o, \vartheta) = -\log f(y|x, z; \vartheta)$ for some conditional density $f(.\,; x, z; \vartheta)$. Then, by sampling the posterior for θ, or the posterior predictive distribution, the method of Section 4 can be deployed.

For the illustrative model of Section 2.3, let $\theta = (\xi, \tau)$ and $\vartheta = (\phi, \tau)$ be the parameters in the data generating and alternative models respectively. In this parametric setting, assuming Normally distributed residual errors in both models, $\pi_n(\theta)$ is readily computable, and using the methods described in Section 4 we can obtain a sample from the posterior distribution and estimate for ϑ in the alternative model. Specifically, from the model (2.6), we have for $b(.)$ known $u_\theta(o, \vartheta) = ((y - b(x)\phi - z\tau)/\lambda)^2$. In this case the parameter of interest τ is identical in the two models, and the posterior computed for $\pi_n(\theta)$ yields correct inference under the presumed correct specification of the conditional model. The posterior for τ as a component of ϑ would still concentrate at true value τ_0, but in finite sample the posterior variance would be larger than that computed from the correctly specified model that led to $\pi_n(\theta)$.

To relax the assumption of Normal residual errors in the data generating model, we may use the Bayesian bootstrap, and obtain a sampled variate from the posterior as

$$(\phi^{(l)}, \tau^{(l)}) = \arg\min_{(\phi,\tau)} \sum_{i=1}^{n} \omega_j^{(l)} (y_i - b(x_i)\phi - z\tau)^2 \tag{5.4}$$

for which the minimization can be achieved analytically for $l = 1, \ldots, L$.

In (5.4), the Bayesian bootstrap is being used to sample the Dirichlet process posterior for the entire unknown joint distribution of the observables, but in the alternative parametric model only the conditional distribution for Y given X and Z is studied – the joint distribution does correspond to an implied conditional distribution. This possibility of partial specification of the model of interest is an advantage of the formulation from Section 4. In addition, if the utility is modified to be

$$u_\theta(o, \vartheta) = -\log f_1(y|x, z; \vartheta_1) - \log f_2(z|x; \vartheta_2)$$

for proposed conditional densities f_1 and f_2. Estimation or posterior sampling of ϑ_1 and ϑ_2 using the parametric or non-parametric algorithms can proceed by the obvious extension, and in this separable loss function the two optimizations can be carried out separately. However, in the inference problem for (2.6) with propensity score unknown, a modification of the loss function is required for optimal inference. Suppose that

$$u_\theta(o, \vartheta) = -\log f_1(y|x, z; \vartheta_1, \vartheta_2^{\mathrm{OPT}}) - \log f_2(z|x; \vartheta_2), \tag{5.5}$$

where $\vartheta_2^{\mathrm{OPT}}$ is the loss minimizing value of ϑ_2 obtained by considering the second term only. This utility reflects the estimation task in the causal problem based on (2.6); the outcome model based on f_1 is adjusted using the fitted propensity score computed using the best estimate of the data generating parameter in the model f_2.

Taking (2.6) or (2.7) as the alternative model, inference for $\vartheta_1 = (\beta, \phi, \tau)$ will be correct (specifically consistent for, and with the posterior concentrated at, true value τ_0) provided the propensity score model encapsulated in model f_2 is itself correctly specified with $\vartheta_2 \equiv \gamma$, so that the estimated propensity score based on the posterior mode $f_2(z|x; \widehat{\vartheta}_2)$ consistently estimates the true propensity score.

It is clear that the form in (5.5) represents a form of modularized inference akin to the "cut" posteriors described in Jacob et al. (2017) and Pompe and Jacob (2021), which do include propensity score regression examples. However, the specific application of the methodology that is required to solve the causal problem does not deploy the cut posterior in the conventional sense, due to the presence of $\vartheta_2^{\mathrm{OPT}}$ in the first term – recall that this is necessary to achieve removal of confounding. The two optimizations are linked by the presence of **common** generated (x, y, z) values in the two terms, which are (re)sampled from the joint Dirichlet process posterior.

6 Simulation studies

We examine the performance of the conventional Bayesian computational methods described in Section 3.1 with the decision-theoretic and non-parametric methods from Sections 4 and 5.

If the treatments are conditionally Normally distributed, then identical logic applies in the balancing argument (see for example Imai and van Dyk (2004)), and we may use the (fitted) conditional mean in a linear regression model for Z as the balancing score. In this simulation, the data generating mechanism assumes $p = 3$ confounders,

with $x = (x_1, x_2, x_3)^\top \sim Normal((-1, 2, 0.5)^\top, \Sigma)$, with $\Sigma_{ij} = \text{Cov}(X_i, X_j) = 0.8^{|i-j|}$, for $i, j = 1, 2, 3$. We consider sample sizes $n = 200, 500, 1000$ and 2000, and simulate a continuous treatment Z_i and continuous outcome Y_i from Normal distributions with unit variance and means

$$\mu_Z = 1 - x_1 + x_2 + 2x_3 - x_1x_2 + 2x_2x_3,$$
$$\mu_Y = 1 + 5z + x_1 + x_2 + x_3 + 5x_2x_3,$$

respectively. For each sample size, we generate 1000 datasets under the above scheme. For the exposure model, we fit the mean model $\mu_Z = \widetilde{x}\gamma$, where the linear predictor is based on row vector $\widetilde{x} = (1, x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3, x_1x_2x_3)$, using linear regression.

Conventional Bayesian methods

We fitted several parametric models under the assumption of Normal errors. In the cutting feedback models, $\widetilde{b}_i = \widetilde{x}_i\widetilde{\gamma}$ with $\widetilde{\gamma}$ being the sampled value of γ in a Gibbs sampler procedure, and in the two-step models $\widehat{b}_i = \widetilde{x}_i\widehat{\gamma}$, where $\widehat{\gamma}$ is the Bayesian estimator of γ obtained from the fitted exposure model.

- 'Unadjusted (UN)': unadjusted for confounding;

$$\text{UN}: \quad \beta_0 + \tau z$$
$$\text{UN} - \text{ext}: \quad \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + \tau z$$

- 'Joint (JT)': the joint model from equation (3.3);

$$\text{JT}: \quad \beta_0 + \phi\widetilde{x}\gamma + \tau z$$
$$\text{JT} - \text{ext}: \quad \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + \phi\widetilde{x}\gamma + \tau z$$

- 'Cutting feedback (CF)': the cut feedback approach via equation (3.4)

$$\text{CF}: \quad \beta_0 + \phi\widetilde{b} + \tau z$$
$$\text{CF} - \text{ext}: \quad \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + \tau z + \phi\widetilde{b}$$

- 'Two-step (2S)':

$$\text{2S}: \quad \beta_0 + \phi\widehat{b} + \tau z$$
$$\text{2S} - \text{ext}: \quad \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + \phi\widehat{b} + \tau z$$

- 'Correct': a correct specification of the linear regression model.

Table 1 contains the estimated bias and root mean square error (RMSE) the posterior estimates (means), and coverage of the 95% credible interval for τ. The unadjusted and joint models perform poorly as theory suggests. Estimation based on cutting feedback yields a small amount of bias, which decreases as the sample size n increases. The two-step approaches yield unbiased estimators. However, in all cases the coverage of the Bayesian credible intervals is not adequate when the outcome model is mis-specified, even though coverage at the nominal level can be obtained using a correct specification.

		n			
	Outcome	200	500	1000	2000
Bias	UN	2.084	2.092	2.093	2.089
	UN-ext	2.401	2.448	2.444	2.444
	JT	−0.355	−0.345	−0.344	−0.345
	JT-ext	−0.092	−0.088	−0.089	−0.090
	CF	0.059	0.027	0.013	0.006
	CF-ext	0.045	0.021	0.011	0.005
	2S	−0.002	0.001	0.001	0.000
	2S-ext	−0.002	0.001	0.001	0.000
	Correct	−0.002	0.001	−0.001	0.000
RMSE	UN	2.086	0.093	2.093	2.089
	UN-ext	2.416	2.454	2.447	2.445
	JT	0.365	0.349	0.346	0.346
	JT-ext	0.117	0.100	0.095	0.093
	CF	0.092	0.054	0.035	0.024
	CF-ext	0.084	0.051	0.034	0.023
	2S	0.071	0.047	0.033	0.023
	2S-ext	0.071	0.047	0.033	0.023
	Correct	0.056	0.036	0.025	0.018
Coverage	UN	0.0	0.0	0.0	0.0
	UN-ext	0.0	0.0	0.0	0.0
	JT	0.1	0.0	0.0	0.0
	JT-ext	75.0	49.7	19.8	2.1
	CF	100.0	100.0	100.0	100.0
	CF-ext	100.0	100.0	100.0	100.0
	2S	100.0	100.0	100.0	100.0
	2S-ext	100.0	100.0	100.0	100.0
	Correct	94.1	94.5	94.1	94.0

Table 1: Summary of the conventional Bayesian estimates of τ under a normal exposure. The rows correspond to mean bias of the point estimates, RMSE and the coverage rates of the posterior 95% credible intervals of τ. Results over 1000 replicate data sets.

Here, we have focused on a Normally distributed exposure represented using a linear model. Additional related results can be found in Supplement C.1. The binary exposure case can be found in Supplement C.2.

Estimation via the Bayesian bootstrap

The results demonstrate that model mis-specification disrupts parametric Bayesian inference. We repeated the analysis using the Bayesian bootstrap approach, restricting attention to the cutting feedback and two-step estimation procedures. To implement the cutting feedback procedure, recall that the Bayesian bootstrap produces a sample from the posterior for a target parameter. In our analysis, we assume correct specification for

the treatment assignment model, and so for the posterior for $\pi_n(\gamma)$, we may either use the exact posterior computed under a Normal assumption, or the Bayesian bootstrap. Having obtained a sample of size L from this posterior, we then use the Bayesian bootstrap to generate L posterior samples for τ, conditioning on the fitted value $\widetilde{b}_i = \widetilde{x}_i\widetilde{\gamma}$. For the two-step method, we may proceed in the same fashion, but instead use $\widehat{b}_i = \widetilde{x}_i\widehat{\gamma}$, where $\widehat{\gamma}$ is the posterior mean derived from $\pi_n(\gamma)$.

These methods follow the conventional approach of separating the posteriors from the two parts of the model. However, following the argument leading to (5.5), the correct Bayesian approach retains the linkage of the two models via the common Dirichlet weights noted in (5.3); that is, a **single** draw of weights ω is used in the optimization over γ and the consequent optimization over (β, ϕ, τ). This linkage reflects a Bayesian non-parametric assumption concerning the full joint distribution of the observables.

For the treatment assignment model, we carry out analysis using the ***True*** propensity score, and then compute $\pi_n(\gamma)$ using a ***Parametric*** (logistic regression) analysis, using the Bayesian bootstrap in an ***Unlinked*** fashion (via independent Dirichlet weights in the two components of expression (5.5)), and in a ***Linked*** fashion using a single Dirichlet draw. For the outcome model, we use a least-squares optimization for the Bayesian bootstrap sampling of (β, ϕ, τ). The analyses were conducted in 1000 replicate data sets, using 1000 Bayesian bootstrap draws for each replicate. For each data replicate, we compute the RMSE of the Bayesian posterior estimates; coverage rates were computed by constructing, for each replicate data set, posterior sample quantiles.

Results are presented in Table 2, for the same data generation and estimation procedures as described above. All of the methods were unbiased in large sample, although the CF method showed a small bias as discussed in Supplement B.3 when n was small, and also larger variability. In terms of RMSE, the two-step methods generally performed best. Coverage at the nominal level was recovered for the two-step method in a Linked analysis, as suggested by the theory studied in Section 5.

7 Discussion

When causal inference is the aim of a statistical analysis, control of confounding is an essential consideration. If an outcome model can be correctly specified or flexibly approximated, causal inferences may follow with or without the use of propensity score methods. However, when it is not possible to correctly capture the outcome process, propensity score methods can be very valuable, particularly when the treatment allocation process is easier to characterize. A joint modelling approach to the estimation of the propensity score and outcome model parameters can result in feedback from the outcome into the propensity score which prevents the estimated propensity score from providing balance, thus resulting in biased estimators of the treatment effect. Techniques aimed at cutting feedback have been suggested; we recap the reasoning as to why a Bayesian two-step approach, rather than one that cuts feedback is the correct approach to pursue, even if in large samples, a cutting feedback approach can provide adequate results. We demonstrated that the standard Bayesian two-step estimator results in poor frequentist performance, but shown that this can be rectified by using the

			n			
	Outcome	$\pi_n(\gamma)$	200	500	1000	2000
RMSE	PS	True	0.417	0.272	0.194	0.132
	PS-ext	True	0.214	0.143	0.096	0.069
	CF	Parametric	0.093	0.056	0.035	0.024
	CF-ext	Parametric	0.084	0.052	0.035	0.023
	2S	Parametric	0.073	0.048	0.032	0.023
	2S-ext	Parametric	0.072	0.047	0.032	0.022
	CF	Unlinked BB	5.487	3.518	2.532	1.757
	CF-ext	Unlinked BB	0.083	0.052	0.034	0.023
	2S	Unlinked BB	0.078	0.050	0.033	0.022
	2S-ext	Unlinked BB	0.072	0.048	0.032	0.022
	2S	Linked BB	0.071	0.047	0.032	0.022
	2S-ext	Linked BB	0.071	0.047	0.032	0.022
Coverage	PS	True	94.2	94.0	95.0	96.0
	PS-ext	True	93.1	92.8	94.1	94.8
	CF	Parametric	100.0	100.0	100.0	100.0
	CF-ext	Parametric	100.0	100.0	100.0	100.0
	2S	Parametric	100.0	100.0	100.0	100.0
	2S-ext	Parametric	100.0	100.0	100.0	100.0
	CF	Unlinked BB	96.5	95.3	94.1	95.1
	CF-ext	Unlinked BB	100.0	100.0	100.0	100.0
	2S	Unlinked BB	100.0	100.0	100.0	100.0
	2S-ext	Unlinked BB	100.0	100.0	100.0	100.0
	2S	Linked BB	94.2	92.8	94.7	94.1
	2S-ext	Linked BB	94.2	92.8	94.7	94.1

Table 2: Summary of the estimates of τ under a normal exposure using the Bayesian bootstrap in the outcome model, and different approaches to the propensity score model parameters posterior: True indicates the true value of γ is used in the propensity score (PS) regression model; Parametric indicates a parametric Normal model is used; Unlinked indicates that the posteriors for γ and (β, ϕ, τ) were computed using separate Bayesian bootstrap computations and different Dirichlet weights (Unlinked Bayesian bootstrap, UBB); Linked (LBB) indicates that common Dirichlet weights were used in the two model components. Rows correspond to RMSE and the coverage rates of the posterior 95% credible intervals. Results over 1000 replicate data sets, using the same data generation and models for estimation as in Table 1.

Bayesian bootstrap with linkage between the two component models, yielding a fully Bayesian procedure with good frequentist properties.

Our argument is based on the realization that the causal analysis is carried out under conscious mis-specification of the Bayesian model, and develop the framework reflecting the literature on Bayesian analysis under mis-specification (Walker, 2013) in the causal problem. The causal setting gives a concrete example where inference under

a mis-specified model – that is, where the target of inference is not a parameter in the data generating model – is actually the objective. Methods that posit the capability of recovering the correct components of the outcome model using flexible modelling without reference to the propensity score also provide valid routes to inference about this target, but these methods often carry a heavier computational burden. There are also links to modularized Bayesian inference (Bayarri et al., 2009; Jacob et al., 2017) which also depend on a 'conscious mis-specification' formulation, and in the causal setting (the main examples and the examples in Supplement A.2) existing frequentist semiparametric theory can give insight into the operating characteristics of such Bayesian analyses; see Pompe and Jacob (2021) for initial explorations in this direction.

The Bayesian bootstrap described in Section 5 relies on the limiting Dirichlet process specification with $\alpha \longrightarrow 0$, although equation (5.1) indicates that a more general model with $\alpha > 0$ can be deployed. In the inference methodology described in Section 4.1, the requirement is simply to be able to sample independently from the posterior predictive distribution, where that distribution is consistent for the data generating process; this can be achieved by statistical procedures beyond those based on the Dirichlet process. In this paper, we have considered loss functions derived from parametric target models, corresponding to the loss functions in (5.5). The propensity score (nuisance) model is parameterized by γ, and for consistent estimation using propensity score regression, this model must be correctly specified up to parametric form. If this correct specification holds along with the standard causal assumptions, frequentist theory confirms that estimation of this nuisance model does not perturb the consistency of the Bayesian estimator, or the approximate Normality of the large sample posterior distribution, for ψ under the Dirichlet process model for the observational data process. The result is more complicated if the nuisance model (loss function) is represented non-parametrically, and care must be taken to ensure that convergence of the posterior for ψ is preserved if the nuisance model is estimated at a slower than parametric rate. Results in this direction are available following the procedures laid out in Ghosal and van der Vaart (2017), and we will report on them elsewhere.

We have not discussed propensity score matching methods in detail. Such methods have been deployed successfully (Liao and Zigler, 2020) by using the propensity score to create a matched sample of treated and untreated individuals. The principles outlined in this paper suggest that matching on an estimated propensity score, rather than averaging over the posterior distribution of the propensity score parameters, would provide superior inference, although this would arguably depend on the matching criterion used. This is an interesting direction for future research.

Supplementary Material

Supplementary Material for Causal inference under mis-specification: adjustment based on the propensity score (DOI: 10.1214/22-BA1322SUPP; .pdf).

References

Adhikari, S., Rose, S., and Normand, S.-L. (2020). "Nonparametric Bayesian instrumental variable analysis: Evaluating heterogeneous effects of coronary arterial access site strategies." *Journal of the American Statistical Association*, 115(532): 1635–1644. MR4189743. doi: https://doi.org/10.1080/01621459.2019.1688663. 646

Antonelli, J., Papadogeorgou, G., and Dominici, F. (2022). "Causal inference in high dimensions: A marriage between Bayesian modeling and good frequentist properties." *Biometrics*, 78(1): 100–114. MR4408573. doi: https://doi.org/10.1111/biom.13417. 647

Bayarri, M. J., Berger, J. O., and Liu, F. (2009). "Modularization in Bayesian analysis, with emphasis on analysis of computer models." *Bayesian Analysis*, 4(1): 119–150. MR2486241. doi: https://doi.org/10.1214/09-BA404. 650, 658

Bernardo, J. M. (1979). "Expected information as expected utility." *The Annals of Statistics*, 7(3): 686–690. MR0527503. 648

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). "A general framework for updating belief distributions." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5): 1103–1130. MR3557191. doi: https://doi.org/10.1111/rssb.12158. 646

Bornn, L., Shephard, N., and Solgi, R. (2019). "Moment conditions and Bayesian nonparametrics." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(1): 5–43. MR3904778. doi: https://doi.org/10.1111/rssb.12294. 647

Chamberlain, G. and Imbens, G. W. (2003). "Nonparametric applications of Bayesian inference." *Journal of Business & Economic Statistics*, 21(1): 12–18. MR1973803. doi: https://doi.org/10.1198/073500102288618711. 652

Cheng, G. and Huang, J. Z. (2010). "Bootstrap consistency for general semiparametric M-estimation." *The Annals of Statistics*, 38(5): 2884–2915. MR2722459. doi: https://doi.org/10.1214/10-AOS809. 652

Comment, L., Mealli, F., Haneuse, S., and Zigler, C. M. (2019). "Survivor average causal effects for continuous time: A principal stratification approach to causal inference with semicompeting risks." *arXiv preprint* arXiv:1902.09304. 646

Davis, M. L., Neelon, B., Nietert, P. J., Hunt, K. J., Burgette, L. F., Lawson, A. B., and Egede, L. E. (2019). "Addressing geographic confounding through spatial propensity scores: A study of racial disparities in diabetes." *Statistical Methods in Medical Research*, 28(3): 734–748. MR3922886. doi: https://doi.org/10.1177/0962280217735700. 647

Geneletti, S., Ricciardi, F., O'Keeffe, A. G., and Baio, G. (2019). "Bayesian modelling for binary outcomes in the regression discontinuity design." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(3): 983–1002. MR3955506. doi: https://doi.org/10.1111/rssa.12440. 646

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian In-*

ference. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. MR3587782. doi: https://doi.org/10.1017/9781139029834. 651, 652, 658

Graham, D. J., McCoy, E. J., and Stephens, D. A. (2016). "Approximate Bayesian inference for doubly robust estimation." *Bayesian Analysis*, 11(1): 47–69. MR3447091. doi: https://doi.org/10.1214/14-BA928. 646, 647, 652

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). "Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects." *Bayesian Analysis*, 15(3): 965–1056. MR4154846. doi: https://doi.org/10.1214/19-BA1195. 647

Henmi, M. and Eguchi, S. (2004). "A paradox concerning nuisance parameters and projected estimating functions." *Biometrika*, 91(4): 929–941. MR2126042. doi: https://doi.org/10.1093/biomet/91.4.929. 642

Holland, P. W. (1986). "Statistics and causal inference." *Journal of the American Statistical Association*, 81(396): 945–960. MR0867618. 640

Imai, K. and van Dyk, D. A. (2004). "Causal inference with general treatment regimes." *Journal of the American Statistical Association*, 99(467): 854–866. MR2090918. doi: https://doi.org/10.1198/016214504000001187. 654

Jacob, P. E., Murray, L. M., Holmes, C. C., and Robert, C. P. (2017). "Better together? Statistical learning in models made of modules." *arXiv preprint* arXiv:1708.08719. 650, 653, 658

Kaplan, D. and Chen, J. (2012). "A two-step Bayesian approach for propensity score analysis: Simulations and case study." *Psychometrika*, 77(3): 581–609. MR2943114. doi: https://doi.org/10.1007/s11336-012-9262-8. 646

Liao, S. X. and Zigler, C. M. (2020). "Uncertainty in the design stage of two-stage Bayesian propensity score analysis." *Statistics in Medicine*, 39(17): 2265–2290. MR4119731. doi: https://doi.org/10.1002/sim.8486. 646, 658

Liu, K., Saarela, O., Feldman, B. M., and Pullenayegum, E. (2020). "Estimation of causal effects with repeatedly measured outcomes in a Bayesian framework." *Statistical Methods in Medical Research*, 29(9): 2507–2519. MR4129426. doi: https://doi.org/10.1177/0962280219900362. 646, 647

Lo, A. Y. (1987). "A large sample study of the Bayesian bootstrap." *The Annals of Statistics*, 15(1): 360–375. MR0885742. doi: https://doi.org/10.1214/aos/1176350271. 652

Lyddon, S. P., Holmes, C. C., and Walker, S. G. (2019). "General Bayesian updating and the loss-likelihood bootstrap." *Biometrika*, 106(2): 465–478. MR3949315. doi: https://doi.org/10.1093/biomet/asz006. 652

McCandless, L. C., Douglas, I. J., Evans, S. J., and Smeeth, L. (2010). "Cutting feedback in Bayesian regression adjustment for the propensity score." *The International Jour-*

nal of Biostatistics, 6(2). MR2602559. doi: https://doi.org/10.2202/1557-4679.1205. 645

McCandless, L. C., Gustafson, P., and Austin, P. C. (2009). "Bayesian propensity score analysis for observational data." *Statistics in Medicine*, 28(1): 94–112. MR2655553. doi: https://doi.org/10.1002/sim.3460. 645

Nethery, R. C., Yang, Y., Brown, A. J., and Dominici, F. (2020). "A causal inference framework for cancer cluster investigations using publicly available data." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(3): 1253–1272. MR4114485. 646

Newton, M. A. and Raftery, A. E. (1994). "Approximate Bayesian inference with the weighted likelihood bootstrap." *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(1): 3–48. MR1257793. 652

Neyman, J. (1923). "On the application of probability theory to agricultural experiments. Essay in principles. Section 9 (Translation published in 1990)." *Statistical Science*, 5: 465–472. MR1092986. 640

Pearl, J. (2009). *Causality*. Cambridge University Press, 2nd edition. MR2548166. doi: https://doi.org/10.1017/CBO9780511803161. 641

Pompe, E. and Jacob, P. E. (2021). "Asymptotics of cut distributions and robust modular inference using Posterior Bootstrap." *arXiv preprint* arXiv:2110.11149. 650, 653, 658

Ray, K. and van der Vaart, A. (2020). "Semiparametric Bayesian causal inference." *The Annals of Statistics*, 48(5): 2999–3020. MR4152632. doi: https://doi.org/10.1214/19-AOS1919. 647

Robins, J. M., Mark, S. D., and Newey, W. K. (1992). "Estimating exposure effects by modelling the expectation of exposure conditional on confounders." *Biometrics*, 48(2): 479–495. MR1173493. doi: https://doi.org/10.2307/2532304. 642, 646

Rosenbaum, P. R. and Rubin, D. B. (1983). "The central role of the propensity score in observational studies for causal effects." *Biometrika*, 70(1): 41–55. MR0742974. doi: https://doi.org/10.1093/biomet/70.1.41. 641, 642

Rubin, D. B. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology*, 65(5): 688–701. 640, 641

Rubin, D. B. (1981). "The Bayesian Bootstrap." *The Annals of Statistics*, 9(1): 130–134. MR0600538. 651

Rubin, D. B. (1985). "The Use of Propensity Scores in Applied Bayesian Inference." In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics, 2*, 463–472. North Holland. MR0862481. 640, 641

Saarela, O., Belzile, L. R., and Stephens, D. A. (2016). "A Bayesian view of doubly robust causal inference." *Biometrika*, 103(3): 667–681. MR3551791. doi: https://doi.org/10.1093/biomet/asw025. 652

Saarela, O., Stephens, D. A., and Moodie, E. E. M. (2022). "The role of exchangeability in causal inference." *arXiv preprint* arXiv:2006.01799. 643

Saarela, O., Stephens, D. A., Moodie, E. E. M., and Klein, M. B. (2015). "On Bayesian estimation of marginal structural models." *Biometrics*, 71(2): 279–288. MR3366229. doi: https://doi.org/10.1111/biom.12269. 646, 652

Samartsidis, P., Seaman, S. R., Montagna, S., Charlett, A., Hickman, M., and Angelis, D. D. (2020). "A Bayesian multivariate factor analysis model for evaluating an intervention by using observational time series data on multiple outcomes." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(4): 1437–1459. MR4157820. 646

Stephens, D. A., Nobre, W. S., Moodie, E. E. M., and Schmidt, A. M. (2022). "Supplementary Material for Causal inference under mis-specification: adjustment based on the propensity score." *Bayesian Analysis*. doi: https://doi.org/10.1214/22-BA1322SUPP. 642

Vegetabile, B. G., Gillen, D. L., and Stern, H. S. (2020). "Optimally balanced Gaussian process propensity scores for estimating treatment effects." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(1): 355–377. MR4049667. 647

Walker, S. G. (2013). "Bayesian inference with misspecified models." *Journal of Statistical Planning and Inference*, 143(10): 1621–1633. MR3082220. doi: https://doi.org/10.1016/j.jspi.2013.05.013. 649, 658

Wang, C., Parmigiani, G., and Dominici, F. (2012). "Bayesian effect estimation accounting for adjustment uncertainty." *Biometrics*, 68: 661–686. MR3055168. doi: https://doi.org/10.1111/j.1541-0420.2011.01731.x. 647

Wang, C. and Rosner, G. L. (2019). "A Bayesian nonparametric causal inference model for synthesizing randomized clinical trial and real-world evidence." *Statistics in Medicine*, 38(14): 2573–2588. MR3962129. doi: https://doi.org/10.1002/sim.8134. 647

Xu, D., Daniels, M. J., and Winterstein, A. G. (2018). "A Bayesian nonparametric approach to causal inference on quantiles." *Biometrics*, 74(3): 986–996. MR3860719. doi: https://doi.org/10.1111/biom.12863. 647

Zigler, C. M. (2016). "The central role of Bayes' Theorem for joint estimation of causal effects and propensity scores." *The American Statistician*, 70(1): 47–54. MR3480670. doi: https://doi.org/10.1080/00031305.2015.1111260. 645

Zigler, C. M., Watts, K., Yeh, R. W., Wang, Y., Coull, B. A., and Dominici, F. (2013). "Model feedback in Bayesian propensity score estimation." *Biometrics*, 69(1): 263–273. MR3058073. doi: https://doi.org/10.1111/j.1541-0420.2012.01830.x. 645

Acknowledgments

The authors thank the Editor, Associate Editor and two referees for suggestions which have improved the paper considerably. The authors would also like to thank Pierre Jacob, Yu Luo and Vivian Meng for comments on earlier drafts.

Invited Discussion

Pierre E. Jacob* and Christian P. Robert†,‡

We congratulate the authors for a stimulating article that touches upon many timely topics. Two-step procedures are standard in econometrics (Pagan, 1984; Murphy and Topel, 1985) but have only been taken seriously more recently in Bayesian inference (Liu et al., 2009) and remain surprisingly controversial, possibly because they clash with the plan of setting up a joint distribution on all observed and latent quantities and then simply conditioning on all observed data. Let us re-introduce a concrete example from Section 6 of the article. Triplets $(x_{i,1}, x_{i,2}, x_{i,3})$ are sampled independently for all $i = 1, \ldots, n$, from a Normal distribution, and rows of covariates are constructed as $x_i = (1, x_{i,1}, x_{i,2}, x_{i,3}, x_{i,1} \cdot x_{i,2}, x_{i,1} \cdot x_{i,3}, x_{i,2} \cdot x_{i,3}, x_{i,1} \cdot x_{i,2} \cdot x_{i,3})$. The two modeling stages are:

$$\begin{aligned} &\text{(first stage)} \quad z_i = x_i\gamma + \epsilon_{z,i} \quad \text{for} \quad \gamma \in \mathbb{R}^8, \quad \epsilon_{z,i} \sim \text{Normal}(0,1), \\ &\text{(second stage)} \quad y_i = \beta_0 + \phi\,(x_i\gamma) + \tau\, z_i + \epsilon_{y,i} \quad \text{for} \quad \beta_0, \phi, \tau \in \mathbb{R}, \quad \epsilon_{y,i} \sim \text{Normal}(0,1). \end{aligned}$$

The coefficient of interest τ can be interpreted as the average treatment effect under adequate assumptions. In the experiments the variable z is generated at the first stage using a certain parameter γ^0, and then y is generated through $y_i = 1 + 5z_i + x_1 + x_2 + x_3 + 5x_2x_3 + \text{Normal}(0,1)$ so that the second stage is generally misspecified. The article proposes an interesting comparison of the following approaches to infer τ.

1. Two-step without uncertainty propagation: at the first stage, inference on γ is made through a distribution $\pi_n(\gamma)$, for example using the Weighted Likelihood Bootstrap (Newton and Raftery, 1994) which amounts to drawing Exponential weights and solving a weighted least squares program. Then the distribution $\pi_n(\gamma)$ is summarized into the point estimator $\hat{\gamma}$ defined as the mean of $\pi_n(\gamma)$. At the second stage, inference is performed conditionally on $\{\gamma = \hat{\gamma}\}$, yielding a distribution $\pi_n(\tau|\hat{\gamma})$.

2. Two-step with uncertainty propagation (*cutting feedback*): at the first stage, γ is inferred through $\pi_n(\gamma)$. Then draws $\gamma^{(\ell)}$ for $\ell = 1, \ldots, L$ are obtained from $\pi_n(\gamma)$ and inference about τ in the second stage is conducted conditionally on each $\gamma^{(\ell)}$. The overall inference on τ is obtained by averaging over the L draws of τ, which approximates $\int \pi_n(\tau|\gamma)\pi_n(d\gamma)$. This is different from a standard "joint model" Bayesian treatment because, here, γ is obtained independently of the outcome y.

*ESSEC Business School, Cergy, France, pierre.jacob@essec.edu
†CEREMADE, CNRS, UMR 7534, Université Paris-Dauphine, PSL University, Paris, France, xian@ceremade.dauphine.fr
‡Department of Statistics, University of Warwick, Coventry, United Kingdom

Advantage of not propagating uncertainty? The quantity of interest τ can be interpreted as the average treatment effect as long as the scores $b_i = x_i\gamma$ used in the second stage are "balanced": the treatment level z_i must be independent of the covariates x_i conditionally on b_i. It is the case here if γ equals the data-generating γ^0, but we can only estimate γ^0. The authors rather surprisingly advocate the use of two-step *without* uncertainty propagation, for example in Supplement B.3 of the article ("*Finite sample bias of the cutting feedback approach*"), by making the argument that the variance of draws from $\pi_n(\gamma)$ translates into a biased estimation of τ, as with generated regressors. In passing, the authors do not justify the choice of the mean as a point estimator to summarize $\pi_n(\gamma)$. Is there a justification for using the mean rather than the result of a genuine decision-making process?

Both two-step strategies are based on plugging some γ vectors that are close to γ^0 as $n \to \infty$. The posterior mean $\hat{\gamma}$ of the first stage would typically fluctuate with variance of order n^{-1}, and the variance of the posterior distribution $\pi_n(\gamma)$ is of the same order. It appears important to take (first stage) sampling variability into account in order to draw a principled comparison between the two-step approaches with respect to the bias that they induce at the second stage. This is especially relevant from a Bayesian perspective, whose main justification stands in a coherent representation of uncertainties associated with parameters and other quantities of interest. Informally this first stage variability is expected to translate into a bias of order n^{-1} for τ, which becomes negligible in the overall mean squared error at the second stage as $n \to \infty$. Is this also expected by the authors?

Prior information An appeal of a Bayesian approach is the possibility of using prior information to improve performance in small samples. Informative priors could reduce the variability of the inferred γ at the first stage, and thus mitigate bias in the second stage. There might also be some available information on sign and magnitude of τ. The possibility of using prior information could be a reason to go for Bayesian causal inference over more traditional methods. Li et al. (2023) discuss the choice of prior in general causal inference settings. It would be interesting to read the authors' view in the specific setting of this article, e.g. on the interplay between prior choices and finite sample bias. In the numerical experiments it appears that the authors have experimented mostly with fairly vague Normal/Inverse Gamma or flat priors.

Weighted Likelihood Bootstrap As described in the article and in Pompe and Jacob (2021), the implementation of *cutting feedback* in a Weighted Likelihood Bootstrap manner (Newton and Raftery, 1994) involves choosing whether to use the same random weights at both stages or to re-sample them independently (*Linked* or *Unlinked Bayesian Bootstrap* in the article). Pompe and Jacob (2021) study the sampling variability of such schemes and support the use of the same weights (*Linked Bayesian Bootstrap*) here, as the data in both stages concern the same n individuals. It is unclear whether *Unlinked Bayesian Bootstrap* should be considered here. Furthermore, the techniques described in e.g. Lyddon et al. (2018, 2019); Pompe (2021) can accommodate prior information. Pompe (2021) employs Edgeworth expansions to identify the impact of the prior penalization and provides concrete guidelines for its choice; see also Section 3.3 in Pompe and Jacob (2021).

References

Li, F., Ding, P., and Mealli, F. (2023). "Bayesian causal inference: a critical review." *Philosophical Transactions of the Royal Society A*, 381(2247): 20220153. 664

Liu, F., Bayarri, M. J., and Berger, J. O. (2009). "Modularization in Bayesian analysis, with emphasis on analysis of computer models." *Bayesian Analysis*, 4(1): 119–150. MR2486241. doi: https://doi.org/10.1214/09-BA404. 663

Lyddon, S., Walker, S., and Holmes, C. C. (2018). "Nonparametric learning from Bayesian models with randomized objective functions." *Advances in neural information processing systems*, 31. 664

Lyddon, S. P., Holmes, C., and Walker, S. (2019). "General Bayesian updating and the loss-likelihood bootstrap." *Biometrika*, 106(2): 465–478. MR3949315. doi: https://doi.org/10.1093/biomet/asz006. 664

Murphy, K. M. and Topel, R. H. (1985). "Estimation and inference in two-step econometric models." *Journal of Business & Economic Statistics*, 3(4): 88–97. MR1940632. doi: https://doi.org/10.1198/073500102753410417. 663

Newton, M. A. and Raftery, A. E. (1994). "Approximate Bayesian Inference with the Weighted Likelihood Bootstrap." *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(1). MR1257793. doi: https://doi.org/10.2307/2346025. 663, 664

Pagan, A. (1984). "Econometric issues in the analysis of regressions with generated regressors." *International Economic Review*, 221–247. MR0741926. doi: https://doi.org/10.2307/2648877. 663

Pompe, E. (2021). "Introducing prior information in Weighted Likelihood Bootstrap with applications to model misspecification." *arXiv preprint* arXiv:2103.14445. 664

Pompe, E. and Jacob, P. E. (2021). "Asymptotics of cut distributions and robust modular inference using Posterior Bootstrap." *arXiv preprint* arXiv:2110.11149v2. 664

Invited Discussion

Joseph Antonelli*

I would like to congratulate Stephens, Nobre, Moodie, and Schmidt (hereafter SNMS) on a very interesting article that addresses many issues relevant to the Bayesian analysis of propensity score approaches. This is an area of research that has seen extensive attention in the recent literature as a Bayesian treatment of propensity scores comes with additional complexities relative to analogous frequentist counterparts. Various complications such as model feedback between the treatment and outcome model, whether and how to account for propensity score uncertainty, or the difficulty in performing Bayesian inference for estimators that are not likelihood-based have been studied, among others (McCandless et al., 2010; Kaplan and Chen, 2012; Zigler et al., 2013; Saarela et al., 2015; Liao and Zigler, 2020). For a recent review of these issues, see Antonelli (2023). SNMS do a very nice job of reviewing these important issues, and subsequently study how to perform Bayesian inference when one includes the propensity score in an outcome model. I think their solution to this problem is elegant, and helped to clarify certain issues that arise in the Bayesian analysis of propensity scores. I was particularly intrigued to see that existing, intuitive approaches such as the two-stage or cutting feedback approaches can lead to poor inferential properties. The authors nicely address this drawback with a justifiable and well-motivated approach to inference.

In this discussion I touch on three main issues relevant to the manuscript and the related literature. I will first describe different approaches to Bayesian causal inference and how they differ with the proposed work. I will then relate the methodology proposed in the manuscript to existing, frequentist resampling techniques to uncertainty quantification, which can provide insight into when and how the proposed procedures will be expected to work well. Lastly, I discuss different forms of model misspecification and their importance to Bayesian causal inference problems.

1 Different notions of Bayesian causal inference

Bayesian causal inference can mean a number of different things with respect to how Bayesian inference is utilized to solve causal inference problems, and I think it is important to first discuss their differences and relative merits. One approach to Bayesian causal inference that dates at least as far back as Rubin (1978) is to explicitly model the joint distribution of the potential outcomes and impute the missing potential outcomes from their posterior distribution. Related approaches have been useful for a wide range of estimands, such as those defined by principal strata, where a Bayesian approach to updating the joint distribution of potential intermediate variables is natural (Jin and Rubin, 2008; Schwartz et al., 2011). Another branch of Bayesian causal inference utilizes traditional frequentist estimators of causal effects, but incorporates Bayesian

*206 Griffin-Floyd Hall, P.O. Box 118545, Gainesville, FL 32611-8545, jantonelli@ufl.edu

modeling of certain unknown quantities such as the propensity score or outcome regression model, among others. Interest in this methodology stems from the desire to use uniquely Bayesian tools such as nonparametric prior distributions or Bayesian model averaging, among others. Additionally, Bayesian approaches to inference can be much easier in certain difficult settings, such as those with spatially correlated data or high-dimensional predictor spaces. Some approaches in this framework have aimed to perform fully Bayesian inference with these estimators (Saarela et al., 2016), while others have performed explicitly frequentist inference while trying to incorporate beneficial features of Bayesian modeling (Shin and Antonelli, 2023). The approach taken by SNMS represents a third distinct branch of Bayesian causal inference that is rooted in exchangeability of the observable quantities, where different estimators of causal effects are defined as those that minimize the Bayes risk under a particular loss function.

Operationally, each of these approaches to Bayesian causal inference (and others not discussed above) are quite different and it is difficult to make connections between them. Despite this, I believe that they are all similar in their core purpose, which is to utilize Bayesian machinery to obtain estimators of causal quantities with improved operating characteristics, such as root mean squared error (RMSE) or coverage. While critics have debated whether some of these approaches are truly Bayesian or not, ultimately it does not matter. Whether something is fully Bayesian, approximately Bayesian, or not Bayesian at all is irrelevant and not worth debating. All that matters is whether the proposed methodology leads to estimators with desirable and well-understood operating characteristics. In my view, Bayesian inference can help in this regard in many situations and it is therefore important to understand how Bayesian inference can help address problems in causal inference. I applaud SNMS in this regard, as they motivated an estimator through the Bayesian paradigm in order to come to an inferential strategy with good properties, which should be the main goal to begin with.

2 Connection to existing resampling techniques

One aspect of the article that I found particularly interesting was that even though the proposed framework was motivated entirely through Bayesian decision theory, it ultimately led to an inferential procedure with close ties to frequentist resampling techniques. The posterior distribution that one obtains from the linked Bayesian bootstrap procedure is extremely similar to the bootstrap distribution one would obtain if using frequentist inference and the nonparametric bootstrap. The main difference is in the construction of the weights assigned to each of the n data points. Whereas the Bayesian bootstrap assigns weights $(\omega_1, \ldots, \omega_n) \sim Dirichlet(1, \ldots, 1)$, the nonparametric bootstrap would assign weights $\omega_i = M_i/n$ with $(M_1, \ldots, M_n) \sim Multinomial(n, 1/n, \ldots, 1/n)$. For an in-depth discussion of the differences between these two, see Rubin (1981). Ultimately, however, these differences are minor as the weights have the same expectation, and inferences from the two should lead to similar conclusions. I also found the idea of using the Dirichlet process to approximate the distribution of the observed quantities very interesting, as this would represent something of a balance between the parametric and nonparametric bootstraps for inference, possibly inheriting the benefits of both.

This connection to frequentist resampling strategies leads one to consider when the proposed approach is most useful. The nonparametric bootstrap helps to perform inference in a wide range of statistical models, but can not be applied universally. The bootstrap does not work well for more complex situations such as certain high-dimensional or nonparametric models (El Karoui and Purdom, 2018). Would the proposed linked Bayesian bootstrap suffer in the same situations? My intuition is that it would, which is not necessarily a negative feature of the proposed approach as inference is inherently challenging in these settings. One of the benefits of the Bayesian paradigm, however, is to be able to provide inference in these more difficult settings through the posterior distribution of the unknown parameters. Additionally, many of the commonly used Bayesian modeling tools that are utilized in the causal inference literature are nonparametric Bayesian models that help to reduce model misspecification. It is unclear how these models could be incorporated within the framework presented by SNMS as they can not necessarily be solved as solutions to optimization problems for every sample of the linked Bayesian bootstrap. One question I have for the authors is whether the proposed framework only applies to fully parametric models, and if so, what advantages does the proposed framework provide over finding frequentist point estimates for these parametric models and performing inference with the nonparametric bootstrap, which also does not make any assumptions about the data generating model?

3 Role of misspecification and nonparametric modeling

Throughout the manuscript, SNMS refer to misspecification from a Bayesian modeling perspective where the data generating model is misspecified. This is useful in the manuscript and helps to elucidate inferential issues that arise when including a propensity score in a Bayesian analysis. This is a somewhat different notion of misspecification than what is typically seen in the causal inference literature, where misspecification would lead to inconsistent estimators of the estimand of interest. In the present manuscript, this would be a situation where the propensity score model is misspecified, and one would obtain biased estimates of the treatment effect regardless of the inferential strategy used. The estimation strategy proposed by SNMS separates the data generating model from the alternative target model, which helps to delineate these two different sources of misspecification. On one hand, the alternative target model can be misspecified, which would lead to inconsistent estimation. On the other hand, the data generating model can be misspecified, even if the alternative model leads to consistent estimation, which can lead to incorrect inference. This provides insight into why the cutting feedback and two-stage approaches over-estimated uncertainty in the simulation study and why those approaches don't lead to correct inferential procedures. I found this aspect of the manuscript interesting and insightful, and it naturally leads me to think about the role of flexible modeling in both of these sources of misspecification.

At various times in the manuscript, SNMS acknowledge that flexible outcome models are also useful for avoiding misspecification, though they state that these can be problematic at times due to a higher computational burden or the fact that they cannot estimate the causal effect of interest in all cases. While these statements are true, I do not think that the computational burden for these approaches is a huge concern, and

they can be used in a wide range of settings (Linero and Antonelli, 2023). Additionally, Bayesian nonparametric approaches need not focus solely on estimation of the conditional mean of the outcome, as they can be used for propensity score estimation (as SNMS notes) or modeling of the full joint distribution of observable quantities, which can be helpful for missing data imputation or estimation of more general treatment effects (Roy et al., 2018). Additionally, they have been shown to work well in recent data analysis competitions, likely because of their robustness to model misspecification (Dorie et al., 2019). The point I want to emphasize in this discussion is that I believe nonparametric Bayesian methods can be used to alleviate issues from both types of misspecification discussed above. Certainly, nonparametric Bayesian modeling of outcome regression or propensity score models would reduce bias due to model misspecification, but they also reduce the impacts of Bayesian misspecification as discussed by SNMS. The main issue discussed in the manuscript is that the target model used to estimate the causal effect does not represent the true data generating mechanism, and therefore a posterior distribution for that model that does not account for this conscious misspecification will lead to incorrect inference. If, however, our target model used to estimate and perform inference on the treatment effect is a nonparametric Bayesian model, then it would mostly avoid issues due to conscious misspecification as it would model the data generating mechanism directly. This would therefore lead to estimators with good estimation performance (bias, RMSE, etc.), but also good inferential performance as well (coverage).

4 Summary

The methodology developed by SNMS provides an elegant and simple solution to a difficult problem of including propensity scores into Bayesian causal inference approaches. Understanding the different ways in which misspecification can affect a causal analysis is an important question, and a better understanding of these issues will help alleviate misspecification in the future. Moving forward, I would be curious to see how the proposed methodology can be combined with existing, powerful Bayesian tools that have proven so useful for causal inference problems.

References

Antonelli, J. (2023). "Bayesian Propensity Score Methods and Related Approaches for Confounding Adjustment." In *Handbook of Matching and Weighting Adjustments for Causal Inference*, 501–528. Chapman and Hall/CRC. 666

Dorie, V., Hill, J., Shalit, U., Scott, M., and Cervone, D. (2019). "Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition." *Statistical Science*, 34(1): 43–68. MR3938963. doi: https://doi.org/10.1214/18-STS667. 669

El Karoui, N. and Purdom, E. (2018). "Can we trust the bootstrap in high-dimensions? The case of linear models." *The Journal of Machine Learning Research*, 19(1): 170–235. MR3862412. 668

Jin, H. and Rubin, D. B. (2008). "Principal stratification for causal inference with extended partial compliance." *Journal of the American Statistical Association*, 103(481): 101–111. MR2463484. doi: https://doi.org/10.1198/016214507000000347. 666

Kaplan, D. and Chen, J. (2012). "A two-step Bayesian approach for propensity score analysis: Simulations and case study." *Psychometrika*, 77: 581–609. MR2943114. doi: https://doi.org/10.1007/s11336-012-9262-8. 666

Liao, S. X. and Zigler, C. M. (2020). "Uncertainty in the design stage of two-stage Bayesian propensity score analysis." *Statistics in Medicine*, 39(17): 2265–2290. MR4119731. doi: https://doi.org/10.1002/sim.8486. 666

Linero, A. R. and Antonelli, J. L. (2023). "The how and why of Bayesian nonparametric causal inference." *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(1): e1583. MR4544390. 669

McCandless, L. C., Douglas, I. J., Evans, S. J., and Smeeth, L. (2010). "Cutting feedback in Bayesian regression adjustment for the propensity score." *The International Journal of Biostatistics*, 6(2). MR2602559. doi: https://doi.org/10.2202/1557-4679.1205. 666

Roy, J., Lum, K. J., Zeldow, B., Dworkin, J. D., Re III, V. L., and Daniels, M. J. (2018). "Bayesian nonparametric generative models for causal inference with missing at random covariates." *Biometrics*, 74(4): 1193–1202. MR3908137. doi: https://doi.org/10.1111/biom.12875. 669

Rubin, D. B. (1978). "Bayesian inference for causal effects: The role of randomization." *The Annals of Statistics*, 34–58. MR0472152. 666

Rubin, D. B. (1981). "The bayesian bootstrap." *The annals of Statistics*, 130–134. 667

Saarela, O., Belzile, L. R., and Stephens, D. A. (2016). "A Bayesian view of doubly robust causal inference." *Biometrika*, 103(3): 667–681. MR3551791. doi: https://doi.org/10.1093/biomet/asw025. 667

Saarela, O., Stephens, D. A., Moodie, E. E., and Klein, M. B. (2015). "On Bayesian estimation of marginal structural models." *Biometrics*, 71(2): 279–288. MR3366229. doi: https://doi.org/10.1111/biom.12269. 666

Schwartz, S. L., Li, F., and Mealli, F. (2011). "A Bayesian semiparametric approach to intermediate variables in causal inference." *Journal of the American Statistical Association*, 106(496): 1331–1344. MR2896839. doi: https://doi.org/10.1198/jasa.2011.ap10425. 666

Shin, H. and Antonelli, J. (2023). "Improved inference for doubly robust estimators of heterogeneous treatment effects." *Biometrics*. doi: https://doi.org/10.1111/biom.13837. 667

Zigler, C. M., Watts, K., Yeh, R. W., Wang, Y., Coull, B. A., and Dominici, F. (2013). "Model feedback in Bayesian propensity score estimation." *Biometrics*, 69(1): 263–273. MR3058073. doi: https://doi.org/10.1111/j.1541-0420.2012.01830.x. 666

Invited Discussion

P. Richard Hahn*, Andrew Herren*

1 Main critique

We thank the authors and the editors for the opportunity to comment on this manuscript and congratulate the authors on an ambitiously creative paper. Broadly, the authors consider the problem of how one ought to think about combining information from both an outcome regression model $Y \mid X, Z$ and a treatment, or propensity, model $Z \mid X$. For clarity, we will consider a parametric setting so that the previous two models have density functions $f(y \mid x, z, \theta, \tau)$ and $p(z \mid x, \gamma)$ respectively. As parametrized here, only the outcome model plays any role in estimating the treatment effect, supposing that the priors over (θ, τ) and γ are independent. However, as discussed in Zigler (2016) and Hahn et al. (2018), any reparametrization with shared parameters between the two likelihoods implies that the propensity model can indeed influence posterior inferences regarding the ATE. Why might this be desirable? Simply because a correctly specified joint likelihood provides more/stronger information about the unknown parameters than the conditional outcome model alone.

Of course, *correctly specified* is always a big ask. Much of the previous literature on this problem focuses on methods that prevent a misspecified propensity model from screwing up inferences when paired with a correct outcome model, while still retaining the efficiency benefit when the treatment model is specified correctly. However, it should be emphasized that a joint model is not *necessary* for ATE estimation, provided that the outcome model is acceptable. Given the availability of performant and highly flexible nonlinear regression models (Krantsevich et al., 2023; Hahn et al., 2020; Woody et al., 2020), this seems to us a worthy approach that was dismissed far too hastily in the paper.

Conversely, the present paper's animating concern is that the outcome model itself might be misspecified, while a correct propensity model is at hand; in that case, how might we proceed? In some ways this is an evergreen approach, and has motivated considerable work in the econometrics literature (Morgan and Winship (2015); Chernozhukov et al. (2018)). The same basic ideas underlie much of this work, and are implicit in the present paper as well. In the case that the true outcome model is linear in the treatment variable Z, it is enough to consider a linear regression on $E(Z \mid X) = b(x)$. The authors cite Robins et al. (1992), but it is instructive to pause and consider why this works. Throughout this note we will assume the following structural model, which is *linear* in Z:

$$\begin{aligned} Z_i &= b(X_i) + \varepsilon_i, \\ Y_i &= \mu(X_i) + \tau(X_i) Z_i + \epsilon_i, \end{aligned} \tag{1.1}$$

*Arizona State University, School of Mathematical and Statistical Sciences, Tempe, Arizona, prhahn@asu.edu

where $b(\cdot)$, $\mu(\cdot)$ and $\tau(\cdot)$ are possibly nonlinear functions of X and both error terms are exogenous (i.e. independent of all elements of X and Z as well as one another). First, rewrite the structural model as

$$Y_i = \mu(X_i) + \tau(X_i)b(X_i) + \tau(X_i)\left(Z_i - b(X_i)\right) + \epsilon_i,$$

noting that $\varepsilon_i = Z_i - b(X_i)$ is exogenous. Letting $\beta_0 + \beta_1 b(X)$ denote the linear projection of $E(Y \mid Z) = \mu(X_i) + \tau(X_i)b(X)$ onto $b(X)$, the model may again be rewritten, giving

$$Y_i = \beta_0 + \beta_1 b(X_i) + R(X_i) + \tau(X_i)\left(Z_i - b(X_i)\right) + \epsilon_i,$$

where $R(X) = \mu(X) + \tau(X)b(X) - \beta_0 - \beta_1 b(X)$ is orthogonal to $b(X)$ and mean-zero by construction.

Next, recall that a linear regression of Y on $b(X)$ and Z is equivalent to first separately regressing both Y and Z on $b(X)$ and then regressing $Y - \hat{Y}_b$ on $Z - \hat{Z}_b$ (Frisch and Waugh (1933), Yule (1907)). This implies that the resulting regression coefficient estimates

$$\frac{\mathrm{Cov}(\tau(X)(Z - b(X)) + R(X) + \epsilon, Z - b(X))}{\mathrm{Var}(Z - b(X))} = \frac{\mathrm{Cov}(\tau(X)\varepsilon + R(X), \varepsilon)}{\mathrm{Var}(\varepsilon)} = E(\tau(X)), \tag{1.2}$$

where an application of the law of total covariance, conditioning on $(\tau(X), R(X))$, gives the final equality.

The authors embrace this misspecified-but-consistent linear regression and proceed to develop a rather elaborate method for a pseudo-Bayesian estimator based upon it, bringing in ideas from Gibbs posteriors (Jiang and Tanner, 2008) and Bayesian bootstrap approximations (more on which to come). Though there is much interesting to discuss regarding those specific choices, the fundamental issues can be considered in a more traditional Bayesian modeling framework, which we lay out in the following subsection.

1.1 Towards Bayesian "double robustness"

Consider yet another reparametrization of our linear-in-Z structural model:

$$E(Y \mid X, Z, b) = \beta_0 + \beta_1 b(X) + \alpha(X) + \tau(X)Z, \tag{1.3}$$

where

$$\alpha(X) \equiv \mu(X) - \beta_0 + \beta_1 b(X).$$

With flat priors on β_0 and β_1, this model can estimate the ATE even if $\alpha(X)$ is aggressively shrunk towards zero and $\tau(X)$ is aggressively shrunk towards an unknown constant. Such a parametrization is something approaching Bayesian double-robustness, without the need for intentional model misspecification. In particular, when $\alpha(X)$ and $\tau(X)$ are extremely complex and $b(X)$ is relatively simple, the prior bias towards simpler forms may well dominate the outcome model and ATE estimation will implicitly rely primarily on the linear term. Conversely, if $b(X)$ itself is quite complex, it may be

regularized towards a constant and relatively simple estimate of $\alpha(X)$ and $\tau(X)$ may drive the estimation. In general, in highly confounded causal effect estimation problems, regularization can have an outsized impact on finite-sample performance, and the difficulty of estimating $b(x)$ is at least as challenging as estimating the outcome model. In this regard, the aesthetic appeal of a misspecified-but-consistent linear regression model may be statistical fool's gold.

Finally, we wish to emphasize that all of the above argumentation has concerned only the mean functions; indeed, this is what makes ordinary least squares (OLS) so powerful even in misspecified settings. (Notably, the above derivations work out even for binary Z.) However, we urge that there is substantial middle ground between fitting a Bayesian bootstrap to a linear regression optimization function and fitting a fully-specified nonparametric Bayesian model. Specifically, tree-based mean regression models with Gaussian errors have proved to be exceptionally capable in problems where misspecification is of concern. That said, we thank the authors for inspiring this line of thought about orthodox Bayesian double robust parametrizations and we plan to explore a linear-in-$b(x)$ offset in the next version of our software (Krantsevich et al., 2023).

1.2 Critique of the simulated example

Above, we claimed that the magnitude and complexity of $\tau(x)$ and/or $\mu(x)$ relative to $b(x)$ will determine whether the proposed approach outperforms alternatives. Let us examine this claim in light of the paper's simulated demonstration. The authors present a detailed simulation study comparing several of the approaches to treatment effect estimation discussed in the paper, including their proposed method that uses the Bayesian bootstrap. Their data generating process (DGP) is as follows

$$\begin{aligned}
X_1, X_2, X_3 &\sim \mathcal{N}\left(\begin{pmatrix}-1\\2\\0.5\end{pmatrix}, \begin{pmatrix}\rho^0 & \rho^1 & \rho^2\\ \rho^1 & \rho^0 & \rho^1\\ \rho^2 & \rho^1 & \rho^0\end{pmatrix}\right)\\
Z &\sim \mathcal{N}\left(b(X), 1\right)\\
Y &\sim \mathcal{N}\left(\mu(X) + Z\tau(X), 1\right)\\
b(X) &= 1 - X_1 + X_2 + 2X_3 - X_1X_2 + 2X_2X_3\\
\mu(X) &= 1 + X_1 + X_2 + X_3 + 5X_2X_3\\
\tau(X) &= 5
\end{aligned}$$

They set $\rho = 0.8$ so that the covariates are all highly correlated. One striking aspect of this DGP is the difference in magnitude of the components of the outcome model, where $\tau(X)Z$ swamps $\mu(X)$, as seen in Figure 1. In this case, the outcome component which is directly estimated in the authors' method, $E\left[\tau(X)\right]$, accounts for a large share of the total outcome variance and the simulations show that the method attains a low RMSE and high coverage.

Consider instead a scenario in which the magnitude of $\mu(X)$ is substantially larger than that of $\tau(X)Z$. This is especially plausible in health and social science settings

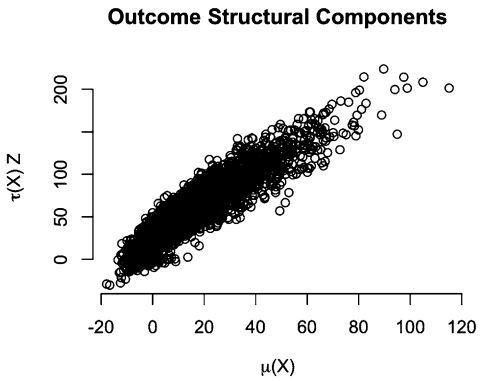

Figure 1: Comparison of $\mu(X)$ and $\tau(X)Z$.

where outcomes such as blood pressure, test scores, or T cell count are governed by many complex mechanisms. The OLS estimator $\hat{\tau}$ in the linear model

$$Y = b(X)\phi + Z\tau + \nu$$

identifies $E[\tau(X)]$ per the discussion in Section 1, but its variance depends on the variance of $\mu(X)$ which is omitted from the model. We see in simulation studies that the variance of $\hat{\tau}$, and thus its RMSE, increase in direct proportionality to the variance of $\mu(X)$. Eventually, concerns about the bias of a misspecified outcome model are outweighed by concerns about the variance of a procedure that refuses to model $\mu(X)$ or $\tau(X)$.

As a more general remark, we advocate that simulated data for evaluating causal inference methods be generated with respect to a set of full structural equations rather than in terms of a non-causal joint distribution; this aids in a degree of realism and allows finer calibration of the relative sizes of various competing effects (strength of confounding versus strength of the treatment effect, etc). It is also explicitly permits more exotic dependence structures that can cause problems for causal methods, such as the presence of colliders (Greenland, 2003).

1.3 Jeffrey updates and Gibbs posteriors

The OLS-based approach that underlies the authors' method is clearly well justified as a point estimate, but the uncertainty quantification is less transparent. The authors cite the very nice paper of Walker (2013) and provide some decision-theoretic arguments, but generally speaking the proper *interpretation* of pseudo-likelihood based posteriors is an open question. Here, we would like to point out some not-widely-known connections for readers interested in learning more about the conceptual challenges and some technical issues related to the use of loss-based likelihoods in a quasi-Bayesian setting.

There are many information-theoretic justifications of pseudo-likelihoods; see for example Zhang (1999) and Zhang (2006) in addition to Walker (2013). However, none of

these commonly-cited expositions speak to the interpretation of the resultant posteriors from the perspective of subjective probability. Surprisingly, work on this perspective on loss-based likelihoods originates instead in the philosophy literature, specifically the fascinating work of Richard C. Jeffrey. Jeffrey (1965) proposed a process for updating subjective probabilities that generalizes conditional probability. The procedure is to partition the space and to reweight each element of the partition while keeping the relative probabilities within each element unchanged. In effect, the idea is to decompose probabilities using the law of total probability and then to (subjectively) change the weights on each conditional term. In symbols,

$$P^*(A) = \sum_j P(A \mid E_j) P^*(E_j)$$

where P^* indicates the modified probability and $P(A \mid E_j)$ the original probabilities. When $P^*(E_j) = 1$ for some j, the Jeffrey's update just gives the usual conditional probability. For instance, if our probability was over a finite collection of cars of different makes, models, and years we might update our probability based on some information that made it twice as likely as before that a car was a sedan. In this example, E_j would index an exhaustive set of mutually exclusive car types (e.g. minivans, coupes, pickup trucks). Conditional on knowing the car type, the individual probabilities remain unchanged, but the various groups of cars are reweighted according to the new evaluations on the partition.

Jeffrey's idea spawned a number of interesting papers, from the philosophical (Skyrms (1987)) to the more mathematical (Diaconis and Zabell, 1982). However, his work appears not to have influenced applied data analysis so much, at least not explicitly. However, a Jeffrey's update turns out to underlie a number of applied methods; we claim without proof that Jeffreys' (*not* Jeffrey's!) substitution likelihood (Jeffreys (1998); Lavine (1995)) and Hoff's rank copula (Hoff (2007)) methods can both be recast as Jeffrey updates.

Here we will show that Jeffrey updates show up *twice* in the present work, first because the method of McCandless et al. (2010) is a form of Jeffrey update, and second because loss-based likelihoods, or *Gibbs posteriors* (Jiang and Tanner (2008)) are Jeffrey updates as well. We explain each in turn.

Consider the marginal posterior of target parameter θ in a model with nuisance parameter η, based on data $D = (X, Y, Z)$:

$$\begin{aligned} \pi(\theta \mid D) &= \int \pi(\theta, \eta \mid D) d\eta \\ &= \int \pi(\theta \mid \eta, D) \pi(\eta \mid D) d\eta. \end{aligned} \tag{1.4}$$

The approach of McCandless et al. (2010) is instead:

$$\pi^*(\theta \mid D) = \int \pi(\theta \mid \eta, D) \pi^*(\eta \mid X, Z) d\eta, \tag{1.5}$$

where $\pi^*(\eta \mid X, Z)$ comes from the posterior of a model without regard to the observed values of Y. The parallel with Jeffrey's update is patent, and the "cutting the feedback" justification seems pragmatic.

Gibbs posteriors, meanwhile, consider the sigma algebra induced over the parameter space by measureable function $L(\theta, Y)$ and fixed data Y. This can be achieved using the formulation given in Diaconis and Zabell (1982), which takes a familiar form:

$$\Pr^*(\theta \in A \mid Y) = \int \Pr(\theta \in A \mid L(\theta, Y) = r)\pi^*(r)dr,$$

where $\Pr(\theta \in A \mid L(\theta, Y) = r)$ is derived from an initial prior measure and $\pi^*(r)$ is a density function favoring lower loss evaluations. This is a canonical Gibbs posterior if $p^*(r) \propto \exp(-\psi r)$, but a transformation of r would yield a Gibbs posterior for a modified loss function in any case. Contrary to the previous example, the form of $\pi^*(r)$ here should give one pause, specifically the so-called "learning rate" parameter ψ; there is little consensus (Wu and Martin (2023)) as to how to calibrate it despite it having a large influence on the variance of the resulting posterior. In the present paper, λ plays a similar (reciprocal) role. Estimating such parameters *as if* they were parameters in a traditional data model is common, such as in the asymmetric Laplace distribution for quantile estimation (Yu and Moyeed (2001)), but the finite-sample implications of this strategy are unknown to the best of our knowledge.

1.4 The Bayesian Bootstrap for ATE estimation

The authors propose using the Bayesian bootstrap as an approximation to a posterior, incorporating sampled Dirichlet weights within an optimization algorithm. However, because the optimization is based on a pseudo-likelihood anyway, the goodness of this approximation is perhaps besides the point. Frankly, we remain unclear as to what the Bayesian bootstrap was even supposed to be approximating. The OLS estimation problem used an estimate of $b(x)$ that was obtained from a separately-fit propensity model, making their approach a non-standard hybrid. Do the standard justifying arguments work in this context? As a frequentist procedure there is little doubt, as the $b(x)$ estimate is presumed to be consistent and the OLS procedure is consistent for the ATE. But what of the inference? We discussed above that the uncertainty measures obtained from Gibbs posteriors are notoriously hard to parse, and here we have a Gibbs posterior, partial Bayesian bootstrap method.

In our own work, we appreciate the importance of "relaxing" fully coherent Bayesian models in the interest of fast estimators with favorable RMSE properties. However, it is precisely such knowledge and experience which gives us pause in our evaluation of the authors' proposed method. The authors dismiss outcome modeling and then proceed to construct their procedure without any clear exposition of the benefit of avoiding a classical Bayesian approach. We are receptive to the ideas presented in this work and glad to see more use of Gibbs posteriors in the literature, but we feel that this paper should have better justified its particular flavor of non-standard Bayesian treatment effect estimation.

In a different vein, we would like to take the opportunity to mention a less exotic role for the Bayesian bootstrap in causal inference, which is as a nonparametric way to estimate a population average treatment effect (PATE). For a known treatment effect function $\tau(x) = E(Y^1 - Y^0 \mid X = x)$, the ATE is defined as $\int \tau(x)p(x)dx$ where $p(x)$ indicates the density function defining some target population. When it is undesirable to estimate $p(x)$ — and when is it not? — the Bayesian bootstrap offers an alternative to simply falling back on the sample average treatment effect (SATE) as an unbiased estimator of the PATE. In particular, rather than recording $\sum_{i=1}^n \frac{1}{n}\tau^\ell(x_i)$ as the ℓ-th posterior sample of the PATE, you instead record $\sum_{i=1}^n \omega_i^\ell \tau^\ell(x_i)$, where $w^\ell = (\omega_1^\ell, \ldots, \omega_n^\ell)$ are vectors drawn independently from a unit Dirichlet distribution; the posterior mean remains the same, but the latter approach has increased posterior variance, as a (partial) reflection of the fact that $p(x)$ is unknown. This rationale was described to one of us by Antonio Linero in personal communications.

2 Final thoughts

In the end, composing this short comment took a great deal of time and thought, none of which was wasted! But wrestling with the details of this paper at length did drive home the fact that it involves several distinct "pivots", each of which could be having a separate effect on the method's performance. As we see it, the proposed method consists of making a series of choices:

1. How do I specify the propensity model?
2. How do I specify the outcome model?
3. How do I combine information from those two models to produce an estimator for the ATE?
4. How do I actually compute that estimator?

As we see it, the authors answer these questions as follows:

1. Assume the correct model specification is known.
2. Avoid specifying a direct connection between outcome and covariates. Instead construct a pseudo-model implied by a weighted least squares optimization problem.
3. Provide estimates of $b(x)$ to our pseudo-model and estimate the parameter of interest conditional on $b(x)$ and z.
4. The Bayesian bootstrap.

The comparisons in the simulation study provide a look at how different choices of the $b(x)$ model and different implementations of the Bayesian bootstrap performed, but not in a truly systematic way. This makes it hard to tell which of these nontrivial

decisions is responsible for any observed performance differences, nor if some of them work synergistically or antagonistically. As noted in Section 1.4, we feel it behooves the authors to articulate why many of the above choices were made, ideally with comparisons made to alternatives in the simulation study. We thank the authors once again for their work and look forward to their rejoinder.

References

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). "Double/debiased machine learning for treatment and structural parameters." MR3769544. doi: https://doi.org/10.1111/ectj.12097. 671

Diaconis, P. and Zabell, S. L. (1982). "Updating subjective probability." *Journal of the American Statistical Association*, 77(380): 822–830. MR0686405. 675, 676

Frisch, R. and Waugh, F. V. (1933). "Partial time regressions as compared with individual trends." *Econometrica: Journal of the Econometric Society*, 387–401. 672

Greenland, S. (2003). "Quantifying biases in causal models: classical confounding vs collider-stratification bias." *Epidemiology*, 300–306. 674

Hahn, P. R., Carvalho, C. M., Puelz, D., and He, J. (2018). "Regularization and confounding in linear regression for treatment effect estimation." *Bayesian Analysis*, 13(1): 163–182. MR3737947. doi: https://doi.org/10.1214/16-BA1044. 671

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). "Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects." *Bayesian Analysis*. MR4154846. doi: https://doi.org/10.1214/19-BA1195. 671

Hoff, P. (2007). "Extending the rank likelihood for semiparametric copula estimation." *Annals of Applied Statistics*, 1(1): 265–283. MR2393851. doi: https://doi.org/10.1214/07-AOAS107. 675

Jeffrey, R. C. (1965). *The Logic of Decision*. New York: McGraw-Hill. MR0233448. 675

Jeffreys, H. (1998). *The Theory of Probability*. Oxford University Press, 3rd edition. MR1647885. 675

Jiang, W. and Tanner, M. A. (2008). "Gibbs posterior for variable selection in high-dimensional classification and data mining." MR2458185. doi: https://doi.org/10.1214/07-AOS547. 672, 675

Krantsevich, N., He, J., and Hahn, P. R. (2023). "Stochastic tree ensembles for estimating heterogeneous effects." In *International Conference on Artificial Intelligence and Statistics*, 6120–6131. PMLR. 671, 673

Lavine, M. (1995). "On an approximate likelihood for quantiles." *Biometrika*, 82(1): 220–222. MR1332852. doi: https://doi.org/10.1093/biomet/82.1.220. 675

McCandless, L. C., Douglas, I. J., Evans, S. J., and Smeeth, L. (2010). "Cutting feedback in Bayesian regression adjustment for the propensity score." *The international journal*

of biostatistics, 6(2). MR2602559. doi: https://doi.org/10.2202/1557-4679.1205. 675

Morgan, S. L. and Winship, C. (2015). *Counterfactuals and causal inference*. Cambridge University Press. 671

Robins, J. M., Mark, S. D., and Newey, W. K. (1992). "Estimating exposure effects by modelling the expectation of exposure conditional on confounders." *Biometrics*, 479–495. MR1173493. doi: https://doi.org/10.2307/2532304. 671

Skyrms, B. (1987). "Dynamic coherence and probability kinematics." *Philosophy of Science*, 54(1): 1–20. MR0876027. doi: https://doi.org/10.1086/289350. 675

Walker, S. G. (2013). "Bayesian inference with misspecified models." *Journal of statistical planning and inference*, 143(10): 1621–1633. MR3082220. doi: https://doi.org/10.1016/j.jspi.2013.05.013. 674

Woody, S., Carvalho, C. M., Hahn, P. R., and Murray, J. S. (2020). "Estimating heterogeneous effects of continuous exposures using bayesian tree ensembles: revisiting the impact of abortion rates on crime." *arXiv preprint arXiv:2007.09845*. 671

Wu, P.-S. and Martin, R. (2023). "A comparison of learning rate selection methods in generalized Bayesian inference." *Bayesian Analysis*, 18(1): 105–132. MR4515727. doi: https://doi.org/10.1214/21-ba1302. 676

Yu, K. and Moyeed, R. A. (2001). "Bayesian quantile regression." *Statistics & Probability Letters*, 54(4): 437–447. MR1861390. doi: https://doi.org/10.1016/S0167-7152(01)00124-9. 676

Yule, G. U. (1907). "On the theory of correlation for any number of variables, treated by a new system of notation." *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 79(529): 182–193. 672

Zhang, T. (1999). "Theoretical analysis of a class of randomized regularization methods." In *Proceedings of the twelfth annual conference on Computational learning theory*, 156–163. MR1811611. doi: https://doi.org/10.1145/307400.307433. 674

Zhang, T. (2006). "Information-theoretic upper and lower bounds for statistical estimation." *IEEE Transactions on Information Theory*, 52(4): 1307–1321. MR2241190. doi: https://doi.org/10.1109/TIT.2005.864439. 674

Zigler, C. M. (2016). "The central role of Bayes' theorem for joint estimation of causal effects and propensity scores." *The American Statistician*, 70(1): 47–54. MR3480670. doi: https://doi.org/10.1080/00031305.2015.1111260. 671

Contributed Discussion*

Alejandra Avalos-Pacheco[†,‡], Veronica Ballerini[§], Matteo Pedone[¶], and Peter Müller[‖]

We would like to congratulate the authors for their work, which provides a fully Bayesian approach for causal inference using propensity scores. The authors proposed an interesting Bayesian decision-theoretic framework for inference under "conscious" misspecification. The authors show that their proposed Bayesian two-step approach, paired with a Bayesian bootstrap, provided a fully Bayesian procedure with good frequentist properties. The proposed approach fully accounts for uncertainty in the propensity score model, propagating it to the posterior distribution of the object of inference. This makes this contribution a promising framework for clinical trial design, particularly for externally controlled trials (ECTs).

ECTs are efficient and ethical alternatives to Randomized controlled trials (RCTs) that estimate the causal effect of a treatment by comparing the outcomes of an experimental arm to a synthetic control arm derived from external data sources. We refer to external data as non-concurrent sources of information, such as observational real-world data (Electronic Health Records and observational studies) or completed trials. Analyses deploying external data can significantly enhance the statistical power of inference on the treatment effect (Viele et al., 2014), and enable more timely and accurate decision-making, thus minimizing the risk of subjecting patients to ineffective or toxic treatments (Avalos-Pacheco et al., 2023). Furthermore, ECTs hold particular value in scenarios where RCTs are not feasible or not ethical, such as in rare diseases and emergency situations (Rahman et al., 2021). However, it is crucial to properly account for potential confounders in the statistical design of ECTs to avoid introducing bias into the evaluation of the experimental treatment (Ventz et al., 2022).

Several ECTs have been proposed in the frequentist paradigm. For instance, Ventz et al. (2019) introduced ECT designs using inverse probability weighting, matching, direct standardization, or marginal structural (regression) models. Inference assumes that the model is correctly specified. More recently in the Bayesian framework, Chandra et al. (2022) proposed an ECT design based on a nonparametric Bayesian (BNP) common atoms model. They set up two aligned mixture models for the treatment and external control cohorts, allowing density-free importance sampling to create equivalent populations, or alternatively nonparametric inference on the treatment effect. Inference has an interesting interpretation as including cluster-specific propensity scores which

*Veronica Ballerini and Matteo Pedone are supported by the European Union – Next GenerationEU, UNIFI Young Independent Researchers Call – BayesMeCOS Grant no. B008-P00634.

[†]Applied Statistics Research Unit, Faculty of Mathematics and Geoinformation, TU Wien, Vienna, Austria, alejandra.avalos@tuwien.ac.at

[‡]Harvard-MIT Center for Regulatory Science, Harvard Medical School, Boston, MA, USA

[§]Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence, Florence, Italy, veronica.ballerini@unifi.it

[¶]Department of Statistics, Computer Science, Applications "G. Parenti", University of Florence, Florence, Italy, matteo.pedone@unifi.it

[‖]Department of Mathematics, UT Austin, Austin, TX, USA, pmueller@math.utexas.edu

can be estimated jointly with the desired treatment effect (since propensity scores are not used in an outcome model joint inference is valid). This and other BNP approaches avoid complications about model mis-specification because BNP models are "always right" (in the sense of full prior support and – usually – posterior consistency).

Instead of avoiding or excluding model mis-specification Stephens et al. (2022) embrace it and provide an inference framework to coherently account for it, allowing to leverage propensity score for inference in high-dimensional settings. Therefore, the approach proposed by Stephens et al. (2022) may be particularly appealing in the context of ECTs, and its implementation would be straightforward.

Let $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ be a p-dimensional vector of observed pre-treatment covariates for patient i, $i = 1, \ldots, n$. Let y_i denote the response to the treatment of patient i. We restrict here our focus on binary endpoints, such as survival outcomes at 12 months following enrollment, although other extensions, such as survival time outcomes, could be easily implemented. We also consider a binary indicator z_i to indicate if patient i is assigned to the experimental treatment, $z_i := 1$, or comes from the externally controlled arm, $z_i := 0$. We test the hypothesis:

$$H_0 : \tau \leq \delta, \quad \text{vs} \quad H_1 : \tau > \delta,$$

where τ represents the average treatment effect (ATE). We then model the data as:

$$\log\left(\frac{p(y_i = 1 \mid \boldsymbol{x}_i, z_i)}{p(y_i = 0 \mid \boldsymbol{x}_i, z_i)}\right) = \text{logit}(\pi_i) = \boldsymbol{x}_i\boldsymbol{\beta} + b(\boldsymbol{x}_i)\phi + \tau z_i, \tag{0.1}$$

where $\boldsymbol{\beta}$ is a p-dimensional vector of pre-treatment coefficients, $b(\cdot)$ is the propensity score estimated through a parametric model, and ϕ is the associated parameter. The inference on model (0.1) would proceed following the approaches presented in Section 3.1 of the paper. Specifically, we would adopt a Bayesian decision-theoretic approach that provides a framework for conducting inference under mis-specification.

Stephens et al. (2022) present a crucial contribution in the context of clinical trials by offering a treatment effect estimate that is *robust* to mis-specification of the outcome process, thereby avoiding the need to accurately capture the outcome model. Moreover, this approach provides a natural measure of uncertainty derived from Bayesian bootstrap methodology, key to control type I error rates. To this end, in-silico trials could be easily simulated under the null hypothesis using posterior predictive sampling from the entire model, including the marginal distribution of the pre-treatment covariates, under the bootstrapping scheme proposed by Stephens et al. (2022).

In addition, to address potential mis-specification in the outcome model, our proposed ECT implementation would ensure all of the benefits offered by the Bayesian approach in clinical trials, including:

(i) monitoring studies through predictive probabilities (Berger and Berry, 1988);

(ii) providing more flexible tools for study design and analysis;

(iii) incorporating expert knowledge and previous data through informative priors and hierarchical models to account for complex patterns in the covariates;

(iv) quantifying the uncertainty of estimands and parameters (Berry et al., 2010).

References

Avalos-Pacheco, A., Ventz, S., Arfè, A., Alexander, B. M., Rahman, R., Wen, P. Y., and Trippa, L. (2023). "Validation of Predictive Analyses for Interim Decisions in Clinical Trials." *JCO Precision Oncology*, 7: e2200606. 680

Berger, J. O. and Berry, D. A. (1988). "Statistical Analysis and the Illusion of Objectivity." *American Scientist*, 76(2): 159–165. 681

Berry, S. M., Carlin, B. P., Lee, J. J., and Muller, P. (2010). *Bayesian adaptive methods for clinical trials*. CRC press. MR2723582. 682

Chandra, N. K., Sarkar, A., de Groot, J., Yuan, Y., and Müller, P. (2022). "Bayesian Nonparametric Common Atoms Regression for Generating Synthetic Controls in Clinical Trials." *arXiv preprint* arXiv:2201.00068. 680

Rahman, R., Ventz, S., McDunn, J., Louv, B., Reyes-Rivera, I., Polley, M.-Y. C., Merchant, F., Abrey, L. E., Allen, J. E., Aguilar, L. K., et al. (2021). "Leveraging external data in the design and analysis of clinical trials in neuro-oncology." *The Lancet Oncology*, 22(10): e456–e465. 680

Stephens, D. A., Nobre, W. S., Moodie, E. E. M., and Schmidt, A. M. (2022). "Causal Inference Under Mis-Specification: Adjustment Based on the Propensity Score." *Bayesian Analysis*, 1(1). 681

Ventz, S., Khozin, S., Louv, B., Sands, J., Wen, P. Y., Rahman, R., Comment, L., Alexander, B. M., and Trippa, L. (2022). "The design and evaluation of hybrid controlled trials that leverage external data and randomization." *Nature communications*, 13(1): 5783–5783. 680

Ventz, S., Lai, A., Cloughesy, T. F., Wen, P. Y., Trippa, L., and Alexander, B. M. (2019). "Design and Evaluation of an External Control Arm Using Prior Clinical Trials and Real-World Data." *Clinical Cancer Research*, 25(16): 4993–5001. 680

Viele, K., Berry, S., Neuenschwander, B., Amzal, B., Chen, F., Enas, N., Hobbs, B., Ibrahim, J. G., Kinnersley, N., Lindborg, S., Micallef, S., Roychoudhury, S., and Thompson, L. (2014). "Use of historical control data for assessing treatment effects in clinical trials." *Pharmaceutical Statistics*, 13(1): 41–54. 680

Contributed Discussion

Francesco Bartolucci*, Stefano Peluso†, and Antonietta Mira‡

We read with great interest the paper by Stephens and his colleagues and we congratulate them on the high quality of the proposed methods and their potential usefulness for practitioners. One point that caught our attention is about the role of inverse weighting based on the propensity score within the framework discussed in the paper.

Suppose that Y is a count variable affected by a binary treatment Z in the presence of a covariate X and we want to perform Bayesian inference on the average treatment effect (ATE) under the usual conditions of propensity score methods. Adopting the standard notation, for the two potential outcomes $Y(0)$ and $Y(1)$ we assume a marginal model based on a Poisson distribution with parameters $\lambda_0 = \exp(\delta_0)$ and $\lambda_1 = \exp(\delta_1)$, respectively, so that the ATE is $\tau = \lambda_1 - \lambda_0$. For the conditional distribution of Z given X we adopt a logistic model with parameter vector γ and likelihood based on the observed data denoted by $L(\gamma)$. Finally, we assume that parameters γ, δ_0, and δ_1 are *a priori* independent with Normal distribution having mean 0 and large variance encoding a vague prior belief. Given these assumptions, we can conceive a Markov chain Monte Carlo (MCMC) algorithm that at the r-th iteration draws values of the parameters, denoted by $\gamma^{(r)}$, $\delta_0^{(r)}$, $\delta_1^{(r)}$, and $\tau^{(r)}$, as follows:

1. propose a new value for γ, denoted by γ^*, from a bivariate Normal distribution centered on $\gamma^{(r-1)}$ and with variances equal to 0.01 and accept it with probability

$$\alpha_\gamma = \min\left(1, \frac{\pi(\gamma^*)L(\gamma^*)}{\pi(\gamma^{(r-1)})L(\gamma^{(r-1)})}\right);$$

2. compute the propensity scores $e_{1:n}^{(r)}$ on the basis of $\gamma^{(r)}$ and the corresponding weights $w_{1:n}^{(r)}$, obtained by normalizing the quantities $z_i/e_i^{(r)} + (1-z_i)/(1-e_i^{(r)})$, $i = 1, \ldots, n$, so that their sum is equal to n;

3. propose a new value of δ_0, denoted by δ_0^*, from distribution $N(\delta^{(r-1)}, 0.0025)$ and similarly propose δ_1^*; the proposed values are accepted with probability

$$\alpha_\delta = \min\left(1, \frac{\pi(\delta_0^*)\pi(\delta_1^*)\prod_{i:z_i=0} p(y_i|\lambda_0^*)^{w_i^{(r)}} \prod_{i:z_i=1} p(y_i|\lambda_1^*)^{w_i^{(r)}}}{\pi(\delta_0^{(r-1)})\pi(\delta_1^{(r-1)})\prod_{i:z_i=0} p(y_i|\lambda_0^{(r-1)})^{w_i^{(r)}} \prod_{i:z_i=1} p(y_i|\lambda_1^{(r-1)})^{w_i^{(r)}}}\right);$$

4. compute $\tau^{(r)} = \lambda_1^{(r)} - \lambda_0^{(r)}$.

*Università degli Studi di Perugia (IT), francesco.bartolucci@unipg.it
†Università degli Studi di Milano-Bicocca (IT)
‡Università della Svizzera italiana (CH) and Università degli Studi dell'Insubria (IT)

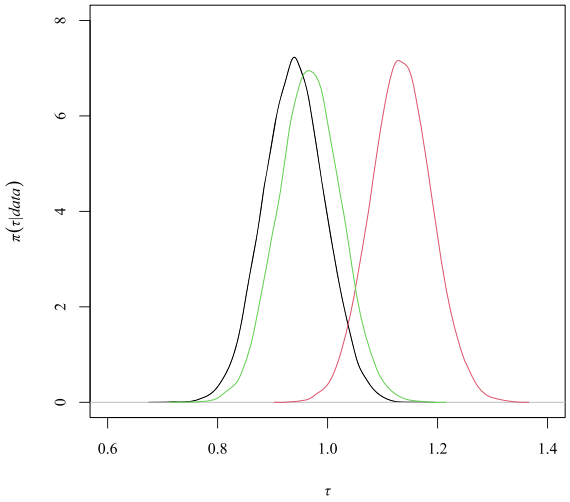

Figure 1: Estimated posterior distribution of the parameter τ given the sample obtained under assignment mechanism (i) with the "unweighted" MCMC algorithm (black curve) and for the sample obtained under mechanism (ii) with the "unweighted" (red curve) and "weighted" (green curve) MCMC algorithm.

We applied the above algorithm to data generated from a model in which, for $i = 1, \ldots, n$, x_i is drawn from a standard Normal distribution and $y_i(0)$ and $y_i(1)$ are drawn from two Poisson distributions with mean $\exp(x_i/5)$ and $\exp(x_i/5) + 1$, respectively. Moreover, we considered two different treatment assignment mechanisms which are: (i) pure randomization with each z_i drawn from a Bernoulli distribution with parameter 0.5; (ii) randomization conditional on x_i with each z_i drawn from a Bernoulli distribution with parameter $\exp(x_i)/[1 + \exp(x_i)]$. For a single sample of size $n = 2000$ obtained under design (i) we applied an "unweighted" version of the "weighted" MCMC algorithm illustrated above where all w_i are equal to 1; for the corresponding sample obtained under design (ii) we run both the "unweighted" and the "weighted" MCMC algorithm. The three estimated distributions of parameter τ are represented in Figure 1. It clearly emerges that, under assignment (ii), the "weighted" MCMC algorithm produces a distribution of τ which is very similar to that produced by its "unweighted" version under randomization, and then the first seems to properly correct the observable confounding due to dependence on the covariates of the potential outcomes and the binary indicator variable. We obtained similar results for other samples and scenarios.

Overall, the "naive" solution to perform Bayesian inference on the ATE here illustrated seems to work well in practice and to be of simple implementation; essentially, this scheme mimics that adopted in frequentist inference. We would appreciate a comment from the Authors on this solution in connection with their statements about inverse weighting within Bayesian inference discussed in Section 3.2 of their paper.

Rejoinder

D. A. Stephens*, W. S. Nobre†, E. E. M. Moodie‡, and A. M. Schmidt‡

We would like to thank the Editor, Professor Steel, for the invitation to make our paper a Discussion Paper in the journal. We would also like to thank all the Discussants for their insightful comments; indeed this paper has benefitted from the ideas of the former Editor, Professor Guindani, and the Associate Editor and Reviewers as it has passed through the review process.

At the outset, the principal goal of this work was simply to establish which of the several proposed methods for Bayesian causal inference that deployed the propensity score (described in Section 3.1 of the main paper) was most appropriate, and if possible, to verify that this procedure adhered fully to Bayesian principles. One very thought-provoking comment that came up during review questioned what was meant by a 'fully Bayesian procedure'. This is an issue that still exercises us (as Bayesians from three continents, albeit ones with strongly orthodox Eurocentric influences), but fundamentally we regard 'fully Bayesian' to imply inference that uses probabilistic arguments and prior-to-posterior updating using Bayes Theorem. This is relevant here as it is clear that two of the methods of Section 3.1 have elements that are unorthodox (the cut feedback, CF, and two-step, 2S, methods), whereas the apparently orthodox method (joint estimation) gets the wrong answer when deployed in its most common form. The consensus in the literature, arguably stemming from the work of Zigler et al. (2013) and Zigler and Dominici (2014), and emphasized in Zigler (2016), is that the two-step strategy that plugs in propensity score model parameters, γ, into the linear predictor is the most appropriate strategy when using propensity score regression. Our goal was to assess the legitimacy of this consensus, to examine its Bayesian credentials, and to establish the correct way to propagate uncertainty into inference (several procedures had been proposed to perform this propagation, and on this point, there did not appear to be any consensus). As noted at the end of Section 1, we consciously did not address other competing and successful methods for performing inference in the causal setting, such as flexible outcome regression (which attempts to avoid mis-specification altogether) and matching (which adjusts for confounding in a different way). These methods were not at all dismissed, but were simply not the focus of the paper. We note that there are situations that arise in which outcome regression alone is insufficient as a procedure for causal adjustment; we discuss this further below, in relation to the comments of Avalos-Pacheco et al., and to some other models.

Before turning to more specific comments, we are indebted to Hahn and Herren for their very interesting comments citing Jeffrey updating as a justification for the existing modularized approaches that underpin the widely used Bayesian-like and generalized

*Department of Mathematics and Statistics, McGill University, Canada, david.stephens@mcgill.ca
†Departamento de Métodos Estatísticos, Universidade Federal do Rio de Janeiro, Brazil, widemberg@dme.ufrj.br
‡Department of Epidemiology and Biostatistics, McGill University, Canada, erica.moodie@mcgill.ca; alexandra.schmidt@mcgill.ca

Bayesian inference procedures that our paper studies. To us, this is very insightful. To this point we have not found the terminology 'generalized Bayesian' attractive for methods such as the Gibbs posterior, as they lack a fully probabilistic implied model for the observable quantities. However, as Hahn and Herren indicate, since Jeffrey's rule is accepted by many as a generalization of Bayes's rule for updating probabilities, perhaps the terminology is acceptable after all. This connection is likely a fruitful avenue for much future research. We must admit that we are complete novices in the fields of Jeffrey updating, Dempster-Shafer theory, and the processing of imprecise probabilities, and therefore cannot comment intelligently on them. Instead, we simply re-iterate (and give details below) that – for better or worse – our proposed solution, which is inspired by a specific non-parametric analysis of a mis-specified model, falls fully within the conventional Bayesian inference domain, and that the posterior distributions constructed are standard Bayesian posteriors computed using Monte Carlo methods. It is puzzling to us to read of them being referred to as 'pseudo Bayesian'; perhaps the confusion arises due to the inclusion of material related to, and some discussion of, the Gibbs posterior approach in the appendix. We emphasize that although the Gibbs posterior can be a productive tool, it is not problem-free, and we do not regard it as fully Bayesian, for the reasons given above.

Antonelli also raises some important philosophical and practical questions concerning the nature of causal inference from a Bayesian viewpoint. It can be persuasively argued that modern interest in causal inference stemmed from the pioneering work of D. B. Rubin, and that this work was explicitly Bayesian. The Rubin causal model deploys the potential outcomes construction in a compelling formulation, and clearly lays out the targets of inference that are typically based on expected differences in potential outcomes under different treatment interventions. The construction, however, does not readily lend itself to inference using likelihood-based procedures without reference to missing data ideas. Although potential outcomes are useful for explanatory purposes, our preference is to consider the contrasting 'observational' and 'experimental' worlds. In the former, the data are generated according to a probability distribution with full dependence structure, including the confounding structures that render causal inferences impossible without further adjustment; in the latter, a hypothetical distribution that matches the observational distribution in many aspects but posits the independence of treatment allocation and previously confounding variables, is considered. Bayesian inference in the target (experimental) world, when the data are generated in the observational world, is therefore necessarily one that involves mis-specification. We do not quite see the trichotomy of Bayesian modelling approaches that Antonelli identifies, as the methods proposed in this paper align quite closely with those proposed in Saarela et al. (2015, 2016), although we do agree that the specifics – regression modelling as opposed to weighting – do differ a little.

Antonelli also asserts that the label attached to a particular type of analysis is not of primary importance, and

> "(w)hether something is fully Bayesian, approximately Bayesian, or not Bayesian at all is irrelevant and not worth debating. All that matters is whether the proposed methodology leads to estimators with desirable and well-understood operating characteristics."

This is perhaps somewhat provocative in a discussion of a paper in a journal entitled *Bayesian Analysis*, but on one level we do not disagree. In de Finetti (1974), the author makes the distinction between the Bayesian *viewpoint* as a foundation for inference, underpinned by the coherence embedded in subjective probabilistic formulation, and Bayesian *techniques*, that might be implemented in a specific analysis, and concludes the latter "are no more trustworthy than any other tool". That is, in the hands of an analyst with a particular data set in front of them, a Bayesian analysis is as fallible as analysis derived from any other sort of paradigm. Nevertheless, we contend that if an analysis can be viewed as following Bayesian orthodoxy, then at least the rules of probability are being followed, and certain well-established theoretical guarantees concerning optimal decision making are available *in finite sample* rather than merely asymptotically. A case in point in the propensity score setting is the handling of posterior uncertainty associated with estimation of propensity score parameters γ when reporting posterior credible intervals for average treatment effect τ.

Returning to the methods proposed in the paper, in terms of procedures that deploy the propensity score in regression, it is useful to recall the distinction between the existing methods. To recap, inspired initially by the joint model with posterior factorization $\pi_n(\theta, \gamma) = \pi_n(\theta \mid \gamma)\pi_n(\gamma)$, we may proceed by first sampling $\gamma^{(l)}$, $l = 1, 2, \ldots, L$ from $\pi_n(\gamma)$, and then

$$\begin{aligned} &\text{Cut feedback (CF):} && \text{Sample } \theta^{(l)}, l = 1, 2, \ldots, L \text{ from } \pi_n(\theta \mid \gamma^{(l)}) \\ &\text{Two step (2S):} && \text{Obtain summary } \widehat{\gamma} \\ & && \text{Sample } \theta^{(l)}, l = 1, 2, \ldots, L \text{ from } \pi_n(\theta \mid \widehat{\gamma}). \end{aligned}$$

This 'forward sampling' approach to the cut feedback settings is possible in this particular model, but may not be possible for all cut feedback approaches. We argue in the main paper that the joint model does not represent a correct implementation of the propensity score regression model, as the balancing property of the propensity score only holds at the true value γ_0 and not an arbitrary value of γ.

The strategy advocated in the paper is based on the Bayesian bootstrap, which itself is based on a Dirichlet process posterior; the procedure represents the inferential (epistemic) uncertainty concerning the unobservable *observational* distribution F_O in a nonparametric fashion, and maps it to the required uncertainty for the parameter of interest (the causal average treatment effect, ATE) in the *experimental* (unconfounded) world by means of a deterministic transformation. As the prior-to-posterior mapping for F_O is performed in a conventional, fully Bayesian fashion, the method can legitimately be regarded as fully Bayesian; the posterior for the ATE is merely computed as a functional of the posterior for F_O by transformation. For simplicity, recall equation (4.5) in the main paper, and imagine that the parametric posterior for θ can be sampled directly, and the integral with respect to s and the minimization over t' can be performed analytically; the calculation that yields the sampled variate $\vartheta^{(l)}$ is then merely a transformation.

The implementation of the proposed strategy in the context of the model of Robins et al. (1992) is instructive in several aspects, and we reproduce it for clarity here, as

well as to re-emphasize (in response to the comments of Hahn and Herren) precisely what it is that the posterior calculation computes. The Robins et al. (1992) model has special structure and what we learn from it may not be portable to other settings, but for this case the general version of the proposed algorithm is implemented as follows: for $l = 1, 2, \ldots, L$:

(I) Sample $F_O^{(l)} \sim DP(\alpha_n, G_n)$, represented as

$$F_O^{(l)}(B) = \sum_{j=1}^{\infty} \omega_j^{(l)} \mathbb{1}_B(o_j^{(l)})$$

parameterized by $\{(\omega_j^{(l)}, o_j^{(l)}), j = 1, 2, \ldots, \infty\}$, a countable collection of weights and ordinates, obtained by stick-breaking or other algorithms.

(II) Produce a variate $\gamma^{(l)}$ by solving the maximization problem

$$\gamma^{(l)} = \arg\max_{\gamma} \sum_{j=1}^{\infty} \omega_j^{(l)} \log f_2(z_j^{(l)} | x_j^{(l)}; \gamma)$$

based on the treatment allocation model.

(III) Form the 'fitted' propensity score values $b(x_j^{(l)}; \gamma^{(l)})$, $j = 1, \ldots, \infty$.

(IV) Produce variates $(\tau^{(l)}, \phi^{(l)})$ by solving the maximization problem

$$(\tau^{(l)}, \phi^{(l)}) = \arg\max_{(\tau,\phi)} \sum_{j=1}^{\infty} \omega_j^{(l)} \log f_1(y_j^{(l)} | z_j^{(l)}, b(x_j^{(l)}; \gamma^{(l)}); \tau, \phi)$$

based on the propensity score regression model.

This recommended 'Linked' algorithm retains a single set of probability weights sampled in Step (I) for the two optimizations. The 'Unlinked' analyses most closely correspond to the typical parametric analysis. In the limiting case of the Bayesian bootstrap, with $\alpha \longrightarrow 0$, all sums become finite, $\alpha_n = n$, and G_n is the empirical measure, and the procedure effectively coincides with the Newton and Raftery (1994) 'weighted likelihood bootstrap' algorithm. The emphasis of much work on the Newton & Raftery algorithm is on its large sample properties; in our paper, we emphasize that the procedure is an exact Bayesian procedure in finite samples (up to Monte Carlo sampling), albeit, for the Bayesian bootstrap, one that relies upon a non-informative prior specification.

As pointed out by Jacob and Robert, in our Table 2 the method listed as 'Linked 2S' can be described equally well as a cut feedback (CF) method. The output of Step (II) is a *sample* from, and not a *summary of*, the posterior arising from the propensity score module. The calculation matches that implied by equation (5.5) in the main paper, and is what we reported as the Linked version of the 2S procedure. It provides the correct 'linkage' between the two parts of the model, and correctly propagates the

inferential uncertainty originating in the DP posterior to the marginal for the parameter of interest, τ. The procedure has two steps of optimization that implement the deterministic mapping of $F_O^{(l)}$ to $\tau^{(l)}$ via plug-in, but it is perhaps misleading to refer to it as a '2S' method. Our use of this terminology arose from a difficulty in distinguishing between 'Linked' CF and 2S methods; this distinction is clear for parametric and 'Unlinked' approaches where separate Bayesian bootstraps can be implemented for the two model components. It is important to note this cut feedback 'sample' is not drawn independently of the sampled value from the target parameter at Step (II). This Linked version is our preferred method in all cases including the case of a binary treatment. In a parametric analysis, the 2S method is preferred for the binary treatment case, but it is certainly the case that a posterior mode derived from $\pi_n(\gamma)$ would often be superior plug-in quantity, at least in terms of finite sample bias; this is examined in some simulations in the arxiv version of the paper https://arxiv.org/pdf/2201.12831.

Jacob and Robert also raise the issue of prior specification. Under the Dirichlet Process formulation, there is an implied prior for all parameters in the targeted model that can be computed (at least, sampled) by replacing posterior $DP(\alpha_n, G_n)$ by prior $DP(\alpha_0, G_0)$ in Step (I) of the above algorithm. In the Bayesian bootstrap version, the prior is improper, which is not wholly satisfactory. Including extra terms in the utility function that act as log prior distributions is certainly possible. However, an issue here concerns appropriate scaling of the two contributing terms. For example, in the derivation of Section 5.3, the tailored function in equation (5.5)

$$u_\theta(o, \vartheta) = -\log f_1(y|x, z; \vartheta_1, \vartheta_2^{\text{OPT}}) - \log f_2(z|x; \vartheta_2)$$

may be augmented by a 'prior' term, $u_0(\vartheta) \equiv -\log \pi_0(\vartheta)$, reflecting prior opinion about ϑ. If f_1 and f_2 are densities in y and z respectively, then no further consideration of scaling is necessary. However, as indicated in equation (5.4), it may be attractive to base the optimization on the quadratic function $(y - b(x; \gamma^{\text{OPT}})\phi - z\tau)^2$ in which case it should be noted that this corresponds to a specific choice of f_1 (Normal) with a *specific* known scale parameter, that is, $\sigma_Y^2 = 1/2$. As indicated in the Supplement, it is possible to include the estimation of an unknown scale parameter in the estimation process: for example, we might take

$$\log f_1(y|x, z; \phi, \tau, \sigma_Y^2, \gamma^{\text{OPT}}) = -\frac{1}{2\sigma_Y^2}(y - b(x; \gamma^{\text{OPT}})\phi - z\tau)^2 - \frac{1}{2}\log \sigma_Y^2$$

for the first component representing the mis-specified model, and optimize simultaneously over (ϕ, τ) and σ_Y; in this case, analytical solution of the optimization problem is possible, and the posterior for σ_Y in the mis-specified model can be computed. If f_1 is replaced by a loss function that cannot be viewed as a density in y, then the scaling of the loss function is arbitrary, and the incorporation of a prior distribution is more complicated; this problem has now been extensively studied in the context of the Gibbs posterior; see for example Syring and Martin (2018).

Hahn and Herren comment on the choice of parameters for the simulation studies, and suggest that the magnitude of the treatment effect may lead to unrealistically good performance of the Bayesian methods. The parameters were not chosen with a great

degree of forethought (apart perhaps from the propensity score parameters in the binary treatment cases, to ensure that the propensity scores were not too extreme, and that there were no violations of positivity requirements) and were largely intended simply to illustrate the methodology. In retrospect, the effect size $\tau = 5$ may appear quite large, and the residual variances quite small. We note that Table C4 in the Supplement does contain a simulation for binary treatment with no treatment effect whatsoever ($\tau = 0$), which demonstrates that the performance of the linked Bayesian bootstrap approach is still good in this case, although several of the other methods also perform adequately. Additional comparisons are also contained in the arxiv version of the paper https://arxiv.org/pdf/2201.12831, including some studies of performance in the case of heterogeneous treatment effects. For a further check, we replicated the analyses for Example 1, first setting $\tau = 0$, and then setting $\tau = 1$ and increasing the residual variances for Y and Z to be 5^2. The conclusions did not change in any significant way.

Hahn and Herren also comment on simulation study design, and state that structural models should form the basis of such studies. We re-iterate that our objective is to draw causal inferences (where this is possible) from observational data; the underlying causal/structural/experimental model is only useful for identifying the target parameter. In our simulations, where the observational model is a fully specified joint distribution on all observables, the causal parameter matches precisely the parameter in the (linear) outcome model studied by Robins et al. (1992). When possible, we prefer to use 'plasmode' type simulations which generate synthetic outcomes (and potentially treatments) from a structural model formed using real joint data structures from existing data sets of confounders, as these represent more realistic test scenarios – see for example Alam et al. (2019).

Bartolucci et al. raise the issue of Bayesian causal inference using weighted likelihood, with the posterior for $\mathrm{z} = 0, 1$ derived as

$$\pi_n(\lambda_\mathrm{z}) \propto \pi_0(\lambda_\mathrm{z}) \prod_{i:Z_i=\mathrm{z}} [p(y_i \mid \lambda_\mathrm{z})]^{w_i},$$

where, in the weighted likelihood case, $w_i = z_i/e(x_i;\gamma)+(1-z_i)/(1-e(x_i;\gamma))$. Relatively diffuse prior distributions complete the Bayesian specification. Bartolucci et al. propose a cut feedback MCMC approach to updating the $(\gamma, \lambda_0, \lambda_1)$ parameters, and then the average treatment effect is computed as $\tau = \lambda_1 - \lambda_0$.

Weighted likelihood methods can be effective, as in the illustrative Poisson example, but their use as a standard 'fully' Bayesian method is perhaps questionable, as the weighted likelihood is not a proper likelihood. It is not straightforward to identify a joint probability model (in the de Finetti representation) that yields the weighted likelihood. Despite this, it is indeed the case that the proposed MCMC implementation of the misspecified Poisson model will produce a 'posterior' distribution that will concentrate at the true value of the causal parameter as the sample size grows. There are well-known drawbacks of inverse probability weighted approaches, the most significant being that the weights themselves can be highly variable, which arises when the propensity score values approach zero or one. However, the main advantage of inverse weighting over the propensity score regression strategy advanced by Robins et al. (1992) is that it does not

require correct specification of the model that quantifies the effect of treatment on the outcome, merely correct specification of the propensity score model.

We implemented the analysis of the Poisson model using the principles established in the main paper: we compare six approaches for four sample sizes $n = 200, 500, 1000, 2000$, replicating 1000 times an analysis that produces $L = 2000$ sampled values from the posterior for τ. The six methods are

I. Unweighted outcome model;

II. Weighted outcome model using a cut feedback approach (i.e., γ sampled from its posterior and used to form the weights w_i at each iteration);

III. Weighted outcome model using a two-step approach (i.e., γ sampled from its posterior, then a posterior modal estimate formed, and used to form the weights w_i for use in the second step analysis);

IV. Linked Bayesian bootstrap based on the weighted Poisson loss,

$$-\log f_1(y|x, z; \lambda_0, \lambda_1) = w\,[\lambda_z - y \log \lambda_z + \log y!]$$

and Dirichlet-weighted estimation of propensity score parameters, using a common set of Dirichlet weights;

V. Unlinked Bayesian bootstrap based on the weighted Poisson loss, and Dirichlet-weighted estimation of propensity score parameters, using a different set of Dirichlet weights;

VI. Correctly specified outcome model.

In each case, the same priors (as specified in the comment) are used.

The weighted models II to VI largely perform similarly in terms of bias and RMSE. The parametric analyses based on weighted Poisson outcome models produce coverage below the nominal level due to the mis-specification. The Bayesian bootstrap methods exhibit better coverage and in this example, the unlinked version performs well. Note also that in the linear outcome model without a heterogeneous treatment effect studied by Bartolucci et al., we may write the structural model as $\mathbb{E}[Y_z] = \lambda_0 + \tau z$ and thus propensity score regression can be deployed via our proposed non-parametric strategy. A more conventional log link would be slightly harder to implement using a regression approach, as discussed in the Supplement, but weighting methods are largely unaffected by a change of link.

Avalos-Pacheco et al. raised the utility of implementing Bayesian propensity score regression as presented in the main article in the context of externally controlled trials. Returning to the discussion of contrasting a propensity score based approach to flexible outcome regression, we note that while the latter approach can offer good performance (RMSE, coverage), it is well-known that it is more difficult to assess whether balance has been achieved – or positivity maintained – in an outcome regression analysis when

		n			
		200	500	1000	2000
Bias	I	0.259	0.256	0.262	0.257
	II	−0.014	−0.007	0.000	−0.001
	III	0.005	0.003	0.000	−0.004
	IV	0.012	0.003	−0.001	0.000
	V	0.002	0.000	0.001	0.003
	VI	0.001	0.004	−0.006	0.002
RMSE	I	0.295	0.272	0.270	0.261
	II	0.200	0.119	0.083	0.061
	III	0.177	0.119	0.082	0.059
	IV	0.178	0.120	0.084	0.065
	V	0.179	0.108	0.090	0.063
	VI	0.127	0.088	0.060	0.043
Coverage	I	0.481	0.122	0.006	0.000
	II	0.769	0.756	0.790	0.778
	III	0.711	0.712	0.697	0.672
	IV	0.912	0.922	0.932	0.945
	V	0.955	0.958	0.957	0.951
	VI	0.965	0.943	0.955	0.948

Table 1: Analysis of Poisson model using weighting methods, and six possible strategies for computing the posterior distribution. I: Unweighted, II: CF, III: 2S, IV: Linked BB, V: Unlinked BB, VI: Correct.

the dimension of the confounders is moderate or large. Propensity score methods also offer an advantage in settings where treatment is not uncommon, but the outcome is: in such settings, there may be insufficient degrees of freedom to model the outcome flexibly enough to avoid mis-specification, but sufficient numbers of both treated and untreated individuals so that many confounders can be adequately (and in a univariate fashion) summarized through the treatment model. Such a setting is typical of externally controlled trials, which may be employed in rare disease settings where trial sizes are limited. Natural extensions to this work to address loss to follow-up or other forms of systematic imbalance can equally be applied in both randomized and observational settings, as well as hybrid-settings such as externally controlled trials.

In addition to such situations where propensity modelling is advantageous, there are settings where outcome regression cannot be readily implemented to estimate causal effects. Once such setting is detailed in the Supplement to the main paper, where multiple treatments are given sequentially, with earlier treatments affecting later ones. To recap the example, two binary treatments (Z_1, Z_2) are generated as $X_1 \sim Normal(1,1)$, $Z_1 \sim Bernoulli(\text{expit}(-2+X_1))$, and $X_2 \sim Normal(-3+X_1+Z_1,1)$ and $Z_2 \sim Bernoulli(\text{expit}(2-X_2))$ at the second stage, with outcome model $Y \sim Normal(X_1 + Z_1 + X_2 + Z_2, 1)$. Then the causal effect of *intervening* to set $(Z_1, Z_2) = (z_1, z_2)$ can be captured by the linear model

$$\mathbb{E}[Y(z_1, z_2)] = -1 + 2z_1 + z_2 = \psi_0 + \psi_1 z_1 + \psi_2 z_2$$

say. However, the causal effect cannot be correctly estimated only by regressing the observed y on the observed (x_1, z_1, x_2, z_2), however flexibly. Weighting methods can be deployed, however, and were the focus of the Bayesian modelling in Saarela et al. (2015).

Ultimately, Hahn and Herren accurately characterize our proposed solutions to their four questions; our intended contribution lies primarily in the domains of questions 3 and 4, concerning construction and computation of appropriate procedures to facilitate Bayesian causal inference. The models we deploy are indeed simple and illustrative, and in practice it is surely the case that more sophisticated modelling of the different modules will be warranted. Antonelli raises the important issue of the behaviour of flexible and non-parametric approaches in high-dimensional settings. For example, the frequentist theory informs us that the estimation of the nuisance propensity model needs to be carried out adequately, with consistency achieved at a fast enough rate, for the plug-in strategy to yield consistency and asymptotic Normality of the estimator of the target parameter. These issues do not disappear in the Bayesian version, and there is a growing literature studying non- and semi-parametric Bayesian estimation of nuisance components that goes well beyond the parametric versions given in our paper. It is also the case that the Bayesian bootstrap implemented in the paper would require amendment in more complex settings (for example, hierarchical, clustered or dependent data). However, the principles that we attempted to advocate in the main paper that aim to facilitate Bayesian inference under mis-specification still hold; we require a model that can consistently estimate the data generating model, and then a procedure to link the data generating model to the target approximating model. We re-iterate that the causal setting, where data arising in the observational world are used to make inferences pertaining to a hypothetical experimental world, provides a canonical example of where this kind of approach is necessary.

Code for the examples in this Rejoinder, and in the Main paper and Supplement, will be available on the GitHub repository `https://github.com/mcgdas01/CIMBPS`, maintained by DAS.

References

Alam, S., Moodie, E. E. M., and Stephens, D. A. (2019). "Should a propensity score model be super? The utility of ensemble procedures for causal adjustment." *Statistics in Medicine*, 38(9): 1690–1702. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8075` MR3934814. doi: `https://doi.org/10.1002/sim.8075`. 690

de Finetti, B. (1974). "Bayesianism: its unifying role for both the foundations and applications of statistics." *International Statistical Review / Revue Internationale de Statistique*, 42(2): 117–130. MR0428528. doi: `https://doi.org/10.2307/1403075`. 687

Newton, M. A. and Raftery, A. E. (1994). "Approximate Bayesian inference with the weighted likelihood bootstrap." *Journal of the Royal Statistical Society. Series B (Methodological)*, 3–48. MR1257793. 688

Robins, J. M., Mark, S. D., and Newey, W. K. (1992). "Estimating exposure effects by

modelling the expectation of exposure conditional on confounders." *Biometrics*, 48(2): 479–495. MR1173493. doi: https://doi.org/10.2307/2532304. 687, 688, 690

Saarela, O., Belzile, L. R., and Stephens, D. A. (2016). "A Bayesian view of doubly robust causal inference." *Biometrika*, 103(3): 667–681. MR3551791. doi: https://doi.org/10.1093/biomet/asw025. 686

Saarela, O., Stephens, D. A., Moodie, E. E. M., and Klein, M. B. (2015). "On Bayesian estimation of marginal structural models." *Biometrics*, 71(2): 279–288. MR3366229. doi: https://doi.org/10.1111/biom.12269. 686, 693

Syring, N. and Martin, R. (2018). "Calibrating general posterior credible regions." *Biometrika*, 106(2): 479–486. MR3949316. doi: https://doi.org/10.1093/biomet/asy054. 689

Zigler, C. M. (2016). "The central role of Bayes' Theorem for joint estimation of causal effects and propensity scores." *The American Statistician*, 70(1): 47–54. MR3480670. doi: https://doi.org/10.1080/00031305.2015.1111260. 685

Zigler, C. M. and Dominici, F. (2014). "Uncertainty in propensity score estimation: Bayesian methods for variable selection and model-averaged causal effects." *Journal of the American Statistical Association*, 109(505): 95–107. MR3180549. doi: https://doi.org/10.1080/01621459.2013.869498. 685

Zigler, C. M., Watts, K., Yeh, R. W., Wang, Y., Coull, B. A., and Dominici, F. (2013). "Model feedback in Bayesian propensity score estimation." *Biometrics*, 69(1): 263–273. MR3058073. doi: https://doi.org/10.1111/j.1541-0420.2012.01830.x. 685