Causal Inference in high dimensions: A marriage between Bayesian modeling and good frequentist properties

Joseph Antonelli, Georgia Papadogeorgou, and Francesca Dominici

Abstract

We introduce a framework for estimating causal effects of binary and continuous treatments in high dimensions. The proposed framework extends many existing estimators introduced in the causal inference literature to high-dimensional settings. We discuss how posterior distributions of any treatment and outcome model can be used together with any causal estimator that is defined as a function of data, and treatment and outcome models (e.g. inverse probability weighted or doubly robust estimator). We propose an approach to uncertainty quantification of causal estimators that utilizes posterior distributions of model parameters that (1) results in good frequentist properties in small samples, (2) is based on a single MCMC, and (3) improves over frequentist measures of uncertainty which rely on asymptotic properties. We show that, for the doubly robust estimator, the posterior contraction rate is the product of the contraction rates of the treatment and outcome models. We consider a flexible framework for modeling the treatment and outcome processes within the Bayesian paradigm that reduces model dependence, accommodates nonlinearity, and achieves dimension reduction of the covariate space. We illustrate the ability of the proposed approach to flexibly estimate causal effects in high dimensions and appropriately quantify uncertainty, and show that it performs well relative to existing approaches. Finally, we estimate the effect of continuous environmental exposures on cholesterol and triglyceride levels. An R package is available at `github.com/jantonelli111/DoublyRobustHD`

1 Introduction

There has been a rapid growth in the interest of estimating the causal effect of a treatment (T) on an outcome (Y) when the dimension of the covariate space ($\boldsymbol{X}$) grows with the sample size. In high-dimensions, some form of dimension reduction or variable selection is required, and most approaches utilize both the treatment and outcome to reduce the dimension of the parameter space in a way that eliminates confounding bias. Recent work has focused on utilizing doubly robust approaches to achieve $\sqrt{n}$ consistent estimates of treatments effects that provide uniformly valid inference (Belloni *et al.* , 2014; Farrell, 2015; Chernozhukov *et al.* , 2016). In related work, Athey *et al.* (2018) achieve $\sqrt{n}$ consistent estimation of treatment effects by combining outcome regression models with weights that balance any remaining differences in covariates between treated and control units. Targeted maximum likelihood (TMLE, Van Der Laan & Rubin (2006); Van der Laan & Rose (2011)) can also be used to estimate treatment effects in this setting by using high-dimensional linear models to estimate the necessary nuisance functions. Other approaches have combined information from both the treatment and outcome to improve finite sample performance of treatment effect estimation and confounder selection. Antonelli *et al.* (2016) utilized the lasso (Tibshirani, 1996) to estimate a propensity and prognostic score, and showed that matching on both quantities leads to doubly robust estimates of treatment effects. Ertefaie *et al.* (2018) derived a penalization estimator that incorporates information from both the treatment and outcome to identify confounders. Shortreed & Ertefaie (2017) used the adaptive lasso to estimate propensity score models that reduce shrinkage of coefficients for covariates also associated with the outcome. Antonelli *et al.* (2017) used similar ideas but used information from the treatment model to reduce shrinkage of coefficients in an outcome model. Finally, Hahn *et al.* (2018) utilized horseshoe priors on a re-parameterized outcome and treatment model to tailor shrinkage of coefficients towards estimating treatment effects.

Nonlinear models have been adopted in the causal inference framework to flexibly estimate treatment effects. TMLE estimators allow for the use of super learners or flexible machine learning techniques to estimate functions of the covariates necessary to estimate treatment effects. Recently, flexible Bayesian methods such as Bayesian additive regression trees (BART) (Chipman *et al.* , 2010) have been utilized to flexibly model potential outcomes and estimate treatment effects (Hill, 2011). While these approaches do not rely on modeling assumptions in either the treatment or outcome model, they do not immediately extend to the high-dimensional regime where $p \geq n$.

In this paper we show how posterior distributions of the propensity score and/or outcome models can be used in combination with any causal estimator from a large class. Such estimators correspond to estimators of treatment effects and average potential outcomes of binary, continuous, or multi-valued treatments that are functions of data and the propensity score (for example, the inverse probability of treatment weighted estimator), or functions of data, the propensity score, and the outcome model (for example, the doubly robust estimator (Bang & Robins, 2005)). We use the posterior distributions for these models to acquire estimates of causal quantities. We introduce an approach to estimate the variance of the causal effect estimator and perform inference by combining posterior samples of propensity score and outcome model estimates with an efficient resampling procedure that only requires one MCMC run. This leads to confidence intervals that have good frequentist properties in finite samples. Building on previous results on doubly robust estimation in high dimensions (Belloni *et al.* , 2014; Farrell, 2015) we prove that the doubly robust estimator contracts at a rate that is the product of the posterior contraction rates for the treatment and outcome models. Lastly, we propose one approach to model the high-dimensional propensity and outcome models utilizing flexible Bayesian methods that relax modeling assumptions and incorporate sparsity inducing prior distributions.

2 Notation, estimands, and identifying assumptions

Let T and Y be the treatment and outcome of interest, respectively, while $\boldsymbol{X}$ is a $p-$dimensional vector of potential confounders. We observe an i.i.d sample $\boldsymbol{D}_i = (\boldsymbol{X}_i, T_i, Y_i)$ for $i = 1 \dots n$, and denote $\boldsymbol{D} = (\boldsymbol{D}_1, \boldsymbol{D}_2, \dots, \boldsymbol{D}_n)$. We will be working under the high-dimensional situation where the number of covariates exceeds the sample size, and is potentially growing with the sample size. Our framework will be quite general in the sense that the ideas presented will be applicable to a wide variety of causal estimands. For clarity, we will first focus on binary treatments and the average treatment effect (ATE) defined as $\Delta = E(Y(1) - Y(0))$, where $Y(t)$ is the potential outcome that would have been observed under treatment $T = t$.

For binary treatments, identification of the average treatment effect based on the observed data relies on the stable unit treatment value assumption (SUTVA) (Little & Rubin, 2000), unconfoundedness, and positivity. SUTVA implies that for each unit the same treatment cannot lead to different outcomes. Unconfoundedness and positivity are:

Unconfoundedness: $T \perp\!\!\!\perp Y(t) | \boldsymbol{X}$ for $t = 0, 1$

Positivity: There exist $\delta \in (0, 1)$ such that $0 < \delta < P(T = 1|\boldsymbol{X}) < 1 - \delta < 1$.

where $P(T = 1|\boldsymbol{X})$ denotes the propensity score (Rosenbaum & Rubin, 1983). There exists analagous assumptions when estimating the effect of a continuous treatment on an outcome, though we refer readers to previous literature on the topic for more details (Gill & Robins, 2001; Hirano & Imbens, 2004; Kennedy *et al.* , 2017).

Even though the ideas presented below are applicable to any estimator that is a function of data, and at least one of the propensity score or outcome models, we focus for illustration purposes on doubly robust estimators for binary and continuous treatments. For instance, if $\boldsymbol{\Psi}$ represents the parameters of the propensity score and outcome models, and $p_{ti} = P(T_i = t|\boldsymbol{X_i})$, $m_{ti} = E(Y_i|T_i = t, \boldsymbol{X}_i)$ represent the fitted values of the treatment and outcome models based on the parameters $\boldsymbol{\Psi}$, a doubly robust estimator of the ATE for binary treatments is

$$\Delta(\boldsymbol{D}, \boldsymbol{\Psi}) = \frac{1}{n}\left[\sum_{i=1}^{n} \frac{T_i Y_i}{p_{1i}} - \frac{T_i - p_{1i}}{p_{1i}} m_{1i}\right] - \frac{1}{n}\left[\sum_{i=1}^{n} \frac{(1-T_i) Y_i}{p_{0i}} + \frac{T_i - p_{1i}}{p_{0i}} m_{0i}\right]. \qquad (1)$$

3 Combining causal estimators with propensity score and outcome model posterior distributions

Parameter values $\boldsymbol{\Psi}$ are typically not known and must be estimated. Since Bayesian modeling approaches provide a natural way of uncertainty quantification, assume that propensity score and outcome models are estimated within a Bayesian framework from which posterior distributions for $\boldsymbol{\Psi}$ are acquired. Even though we provide one such modeling approach in Section 4, the framework presented here allows for any Bayesian modeling technique.

When the propensity score and outcome model parameters are estimated within the Bayesian framework, questions arise on how to use the posterior distribution of these models to (1) acquire estimates of treatment effects, and (2) perform inference with good frequentist properties. Even though we focus on the doubly robust estimator in Equation 1, these ideas apply to any estimator that is a function of data, treatment and outcome model parameters.

$$\begin{tikzcd}
\begin{pmatrix} \Delta(\boldsymbol{D}^{(1)}, \boldsymbol{\Psi}^{(1)}) & \Delta(\boldsymbol{D}^{(1)}, \boldsymbol{\Psi}^{(2)}) & \dots & \Delta(\boldsymbol{D}^{(1)}, \boldsymbol{\Psi}^{(B)}) \\ \Delta(\boldsymbol{D}^{(2)}, \boldsymbol{\Psi}^{(1)}) & \ddots & & \Delta(\boldsymbol{D}^{(2)}, \boldsymbol{\Psi}^{(B)}) \\ \vdots & & \ddots & \vdots \\ \Delta(\boldsymbol{D}^{(M)}, \boldsymbol{\Psi}^{(1)}) & \Delta(\boldsymbol{D}^{(M)}, \boldsymbol{\Psi}^{(2)}) & \dots & \Delta(\boldsymbol{D}^{(M)}, \boldsymbol{\Psi}^{(B)}) \end{pmatrix} \arrow[r] & \begin{pmatrix} E_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{D}^{(1)}, \boldsymbol{\Psi})] \\ E_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{D}^{(2)}, \boldsymbol{\Psi})] \\ \vdots \\ E_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{D}^{(M)}, \boldsymbol{\Psi})] \end{pmatrix} \arrow[d] & , & \begin{pmatrix} \mathrm{Var}_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{D}^{(1)}, \boldsymbol{\Psi})] \\ \mathrm{Var}_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{D}^{(2)}, \boldsymbol{\Psi})] \\ \vdots \\ \mathrm{Var}_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{D}^{(M)}, \boldsymbol{\Psi})] \end{pmatrix} \arrow[dd] \\
& \mathrm{Var}_{\boldsymbol{D}^{(m)}}\{E_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{D}^{(m)}, \boldsymbol{\Psi})]\} & & \\
& & & E_{\boldsymbol{D}^{(m)}}\{\mathrm{Var}_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{D}^{(m)}, \boldsymbol{\Psi})]\}
\end{tikzcd}$$

Figure 1: Values of estimator for different combinations of resampled data sets and posterior samples

3.1 The doubly-robust estimator using posterior distributions

In frequentist settings, the doubly robust estimator in (1) is often evaluated using the MLE of the parameters $\boldsymbol{\Psi}$. However, when model estimation is performed within the Bayesian paradigm, investigators have access to a posterior distribution of $\boldsymbol{\Psi}$. Let $\{\boldsymbol{\Psi}^{(b)}\}_{b=1}^{B}$ represent B draws from the posterior distribution of $\boldsymbol{\Psi}$. We consider an estimator $\widehat{\Delta}$ defined as:

$$\widehat{\Delta} = E_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{D}, \boldsymbol{\Psi})] \approx \frac{1}{B}\sum_{b=1}^{B} \Delta(\boldsymbol{D}, \boldsymbol{\Psi}^{(b)}),$$

where $\Delta(\boldsymbol{D}, \boldsymbol{\Psi}^{(b)})$ is the quantity in Equation 1 evaluated using the observed data $\boldsymbol{D}$ and parameter values $\boldsymbol{\Psi}^{(b)}$.

3.2 Inference with good frequentist properties

While it is straightforward to use the posterior distribution of $\boldsymbol{\Psi}$ (or samples from it) to acquire a point estimate, performing inference is less clear. Typically in Bayesian inference, posterior draws are sufficient for uncertainty quantification of *model parameters* or *functionals of these parameters*. However, performing inference using the variance or quantiles of $\{\Delta(\boldsymbol{D}, \boldsymbol{\Psi}^{(b)})\}_{b=1}^{B}$ does not fully account for uncertainty in our causal estimate, since the estimator in Equation 1 is a function of both the parameters, *and* the observed data $\boldsymbol{D}$.

We note that frequentist operating characteristics (such as interval coverage) are focused around the estimator's sampling distribution, the distribution of the point estimate over different data sets. The variance of our estimator's sampling distribution can be written as

$$\mathrm{Var}_{\boldsymbol{D}}\big(\widehat{\Delta}\big) = \mathrm{Var}_{\boldsymbol{D}}\{E_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{D}, \boldsymbol{\Psi})]\}. \tag{2}$$

In an ideal world, we would find this variance by repeatedly resampling from the distribution of $\boldsymbol{D}$, calculate $E_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{D}, \boldsymbol{\Psi})]$ for each data set, and take the variance of these estimates. We can't do this for two reasons: 1) We don't know the distribution of the data (though it can be approximated by the empirical distribution), and 2) even if we did, it would be computationally prohibitive to estimate the posterior mean for each new data set.

Instead, we detail an approach that combines the samples of the parameter posterior distribution based on our observed data with an efficient resampling procedure to obtain confidence intervals with good frequentist operating characteristics. Our strategy is to create M new datasets, $\boldsymbol{D}^{(1)}, \dots, \boldsymbol{D}^{(M)}$, by sampling with replacement from the empirical distribution of the data. Note that while we are using the nonparametric bootstrap here, other resampling techniques such as the Bayesian bootstrap could be applied as well. For all possible combinations of resampled data sets and posterior samples, we calculate $\Delta(\boldsymbol{D}^{(m)}, \boldsymbol{\Psi}^{(b)})$ for $m = 1, \dots, M$, and $b = 1, \dots, B$. The values $\Delta(\boldsymbol{D}^{(m)}, \boldsymbol{\Psi}^{(b)})$ can be arranged in a matrix where rows correspond to data sets, and columns correspond to posterior samples, as shown in Figure 1.

First, we acquire the mean within each row corresponding to $E_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{D}^{(m)}, \boldsymbol{\Psi})]$ for $m = 1, \dots, M$. The variance of these M values is an estimate of $\mathrm{Var}_{\boldsymbol{D}^{(m)}}\{E_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{D}^{(m)}, \boldsymbol{\Psi})]\}$, which resembles the target variance in (2) but is not equal to it for two reasons. The first is that we are drawing new data sets from the empirical distribution of the data instead of the true joint distribution. However, this is acceptable in many settings and is the main idea behind the bootstrap.

The second and perhaps most important reason is that that distribution used in the outer moment $(\boldsymbol{D}^{(m)})$ does not agree with the one in the inner moment $(\boldsymbol{\Psi}|\boldsymbol{D})$, stemming from calculating the estimator using the posterior distribution from the original data $\boldsymbol{\Psi}|\boldsymbol{D}$ instead of the posterior distribution

$\boldsymbol{\Psi}|\boldsymbol{D}^{(m)}$. Therefore, $\text{Var}_{\boldsymbol{D}^{(m)}}\{E_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{D}^{(m)}, \boldsymbol{\Psi})]\}$ accounts for the uncertainty in the estimator for a fixed posterior distribution (the one corresponding to the observed data) and ignores uncertainty stemming from the fact that the posterior distribution $\boldsymbol{\Psi}|\boldsymbol{D}^{(m)}$ might be different across data sets. For that reason, inference based on this quantity will achieve close to nominal level only when the variability of the posterior distribution relative to the variability of the data is low, and is likely to be anti-conservative in settings where there is a lot of uncertainty in estimating $\boldsymbol{\Psi}$ (see also Appendix H).

Therefore we correct the variance estimate $\text{Var}_{\boldsymbol{D}^{(m)}}\{E_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{D}^{(m)}, \boldsymbol{\Psi})]\}$ by adding a term that explicitly targets the variability of the posterior distribution over different data sets. Specifically, our variance estimate is

$$\text{Var}_{\boldsymbol{D}^{(m)}}\{E_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{D}^{(m)}, \boldsymbol{\Psi})]\} + \text{E}_{\boldsymbol{D}^{(m)}}\{\text{Var}_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{D}^{(m)}, \boldsymbol{\Psi})]\}.$$

We can estimate the second term in the same manner as the first term by using the M by B matrix of estimates in Figure 1. To better understand why the additional term $\text{E}_{\boldsymbol{D}^{(m)}}\{\text{Var}_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{D}^{(m)}, \boldsymbol{\Psi})]\}$ targets the error in the initial variance estimate, we focus on $\text{Var}_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{D}^{(m)}, \boldsymbol{\Psi})]$ where $\boldsymbol{D}^{(m)}$ is fixed, and the randomness in $\Delta(\boldsymbol{D}^{(m)}, \boldsymbol{\Psi})$ is only due to randomness in $\boldsymbol{\Psi}$. Further, (Freedman, 1999) shows that the posterior distribution of a function of $\boldsymbol{\Psi}$ is expected to resemble the sampling distribution of its posterior mean, which is the basis for using credible intervals in order to achieve nominal frequentist confidence interval coverage. With that in mind, we can see that $\text{Var}_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{D}^{(m)}, \boldsymbol{\Psi})] \approx \text{Var}_{\boldsymbol{D}}\{E_{\boldsymbol{\Psi}|\boldsymbol{D}}[\Delta(\boldsymbol{D}^{(m)}, \boldsymbol{\Psi})]\}$, which is the variance in our estimator that stems just from uncertainty in $\boldsymbol{D}$ and propogates through $\boldsymbol{\Psi}|\boldsymbol{D}$. This is exactly the variance that our initial variance estimator was ignoring, and therefore adding in this second term removes the error from only estimating one posterior distribution instead of M. We refer readers to Appendix C for more details.

In Section 6 we empirically show that this variance estimate accurately approximates the Monte Carlo variance of the estimator under various scenarios, even when both models are misspecified.

4 Modeling framework in high dimensions

While this approach to combining posterior samples with resampled data sets works in general, it is most useful in high-dimensions when accounting for uncertainty in parameter estimation can be quite difficult. Since many well-known estimators depend on the propensity score and outcome model fitted values, we posit Bayesian high-dimensional treatment and outcome models as

$$h_y^{-1}(E(Y_i|T_i, \boldsymbol{X}_i)) = \beta_0 + f_t(T_i) + \sum_{j=1}^{p} f_j(X_{ji}) \tag{3}$$

$$h_t^{-1}(E(T_i|\boldsymbol{X}_i)) = \alpha_0 + \sum_{j=1}^{p} g_j(X_{ji}), \tag{4}$$

where $h_y()$ and $h_t()$ are suitable link functions. For estimators that rely on only one of the two models, such as the inverse probability weighted estimator, the other model can be dropped. Note that these models only specify the mean of each random variable. In cases where the treatment or outcome are not binary, additional parameters and modeling assumptions are required. If either of them is a continuous random variable, we specify a normal distribution with the mean defined above (with the identity link function) and variance σ^2.

These models assume that the relationship of the covariates with the treatment and outcome are additive, which helps reduce the complexity of estimation in high-dimensions. We stress here that this is not a requirement for the uncertainty quantification presented in Section 3, and alternative high-dimensional models that allow for interactions (see for example Linero & Yang (2018)) can be accommodated. We explore this possibility via simulation in Appendix F.

4.1 Guassian process prior specification

For now the functional form of the relationships between the covariates and the treatment or outcome is unspecified. Here, we present the prior specification for the outcome model only, but analogous representations are used for the treatment model. We adopt Gaussian process priors for the unknown regression functions, $f_j()$ and $g_j()$ for $j = 1, \ldots, p$. We use $\boldsymbol{X}_j$ to denote the vector of values for the j^{th} covariate across all units, $(X_{j1}, \ldots, X_{jn})$. Since we only need to evaluate $f_j()$ at the n observed locations (denoted by $f_j(\boldsymbol{X_j})$), we can represent our prior as follows:

$$f_j(\boldsymbol{X_j}) \sim (1 - \gamma_j)\delta_{\boldsymbol{0}} + \gamma_j \mathcal{N}(\boldsymbol{0}_n, \sigma^2 \tau_j^2 \boldsymbol{\Sigma}_j)$$

$$\gamma_j \sim \text{Bernoulli}(\theta) \qquad\qquad \theta \sim \mathcal{B}(a_\theta, b_\theta)$$
$$\tau_j^2 \sim \text{Gamma}(1/2, 1/2) \qquad\qquad \sigma^2 \sim \text{InvGamma}(a_{\sigma^2}, b_{\sigma^2}).$$

Here, σ^2 is the residual variance of the model when the outcome is normally distributed, otherwise it is fixed to 1. We utilize a latent variable, γ_j, which indicates whether variable j is important for predicting the outcome. If $\gamma_j = 0$, the predictor X_j is eliminated from the model. We assume a gamma$(1/2, 1/2)$ prior on the variance τ_j^2 similarly to Mitra & Dunson (2010). Finally, the (i, i') entry of the covariance matrix $\boldsymbol{\Sigma}_j$ is $K(X_{ji}, X_{ji'})$, where $K(\cdot, \cdot)$ is the kernel function of the Gaussian process. Throughout we will proceed with $K(z, z') = \exp\{-\frac{|z-z'|}{\phi}\}$, where ϕ is a bandwidth parameter that must be chosen.

The formulation above allows for flexible modeling of the response functions $f_j()$. One criticism of using Gaussian processes is that they can be very computationally burdensome, particularly as the sample size increases, because at each iteration of an MCMC one must invert an n by n matrix. Reich *et al.* (2009) showed that this can be avoided by using a singular value decomposition on the kernel covariance matrices before running the MCMC. This allows us to utilize Gaussian processes in reasonably sized data sets, but the computation can still be slow for large sample sizes. Details of this can be found in their paper or in Appendix B.

4.2 Basis expansion specification

The computational burden of the Gaussian process prior can be greatly alleviated using basis functions, such as cubic splines. This reduces the amount of flexibility in estimation of $f_j(\boldsymbol{X_j})$ as we are restricting the class of functions permitted by our prior specification, but greatly reduces the computational complexity and allows us to model much larger data sets. To do this, we must introduce some additional notation. Let $\widetilde{\boldsymbol{X}}_j$ represent an n by q matrix of basis functions. Then, we write $f_j(\boldsymbol{X_j}) = \widetilde{\boldsymbol{X}}_j \boldsymbol{\beta}_j$ and assume:

$$(\boldsymbol{\beta}_j | \gamma_j) \sim (1 - \gamma_j)\delta_{\mathbf{0}} + \gamma_j \psi(\boldsymbol{\beta}_j)$$
$$\gamma_j \sim \text{Bernoulli}(\theta) \qquad \theta \sim \text{Beta}(a_\theta, b_\theta) \qquad \sigma^2 \sim \text{InvGamma}(a_{\sigma^2}, b_{\sigma^2}).$$

This specification places a multivariate spike and slab prior on the group of coefficients, $\boldsymbol{\beta}_j$, that will force all coefficients to zero and eliminate covariate j from the model if $\gamma_j = 0$. If $\gamma_j = 1$, then all elements of $\boldsymbol{\beta}_j$ will be nonzero with prior distribution $\psi(\boldsymbol{\beta}_j)$ specified as a mean$-\mathbf{0}$ multivariate normal distribution with covariance $\sigma^2 \sigma^2_{\boldsymbol{\beta}} I_n$. A value for $\sigma^2_{\boldsymbol{\beta}}$ can be selected either via empirical Bayes or by placing a hyper prior on $\sigma^2_{\boldsymbol{\beta}}$.

5 Contraction rate of the doubly robust estimator

In high-dimensional settings, achieving $\sqrt{n}-$consistency of treatment effect estimators is not trivial. We restrict attention to the doubly robust estimator, which is of particular interest in high-dimensional settings because it allows us to achieve better rates of contraction than approaches based on a single model. In the semi-parametric causal inference literature this has allowed the use of machine learning approaches that converge at $n^{1/4}$ rates or high-dimensional models that converge at $\sqrt{n/\log p}$ rates (Chernozhukov *et al.* , 2016; Farrell, 2015).

Here, we establish that these same ideas carry over into posterior contraction rates for the posterior distribution of $\Delta(\boldsymbol{D}, \boldsymbol{\Psi})$. Note that when treating $\boldsymbol{D}$ as fixed, $\Delta(\boldsymbol{D}, \boldsymbol{\Psi})$ is a functional of only $\boldsymbol{\Psi}$ and it has a posterior distribution. Throughout the rest of this section when we refer to a posterior distribution, it is this that we are referring to. It might initially appear counterintuitive to prove posterior contraction rates when our approach is not fully Bayesian. However, deriving contraction rates for the posterior distribution of $\Delta(\boldsymbol{D}, \boldsymbol{\Psi})$ also establishes convergence rates of specific features of the distribution such as the posterior mean which is our estimator.

5.1 Notation and assumptions

Let $p_{ti} = P(T_i = t | \boldsymbol{X}_i)$, $m_{ti} = E(Y(t) | \boldsymbol{X}_i)$, and $E(Y(t)) = \mu_t$, where each of these can be estimated using the parameters in our model specification above. We will denote their true values as p^*_{ti}, m^*_{ti}, and μ^*_t, respectively. Throughout this section, we will utilize the subscripts n and P_0 to represent moments with respect to the posterior distribution and true data generating process, respectively. In particular $\mathbb{P}_n$ represents the posterior distribution given a sample of n observations, and E_{P_0} is the expected value with respect to P_0, the true data generating process. Before we detail our result on posterior contraction, we must highlight a few assumptions.

Assumption 1 (Data generating process)

(a) $\{(Y_i, T_i, \boldsymbol{X}_i)\}_{i=1}^n$ are i.i.d samples from P_0

(b) The covariates X_j have bounded support, in that there exists $K_x < \infty$ such that $|X_j| < K_x$ with probability 1 for all j.

(c) $\sup\limits_{P_0} E_{P_0}((Y - m_{ti}^*)^2) \leq K_y < \infty$.

Assumption 1a restricts analysis to the cross-sectional setting. Assumption 1b is likely to be satisfied in real applications as nearly all underlying variables are naturally bounded. Assumption 1c ensures that the residual variance of the outcome is bounded, which again should be satisfied in most applications.

Assumption 2 (Bounds on the error of posterior distributions)

(a) $\sup\limits_{P_0} E_{P_0} \mathrm{Var}_n \left(\frac{p_{ti} - p_{ti}^*}{p_{ti}} | \boldsymbol{D}_i \right) \leq K_p < \infty$

(b) $\sup\limits_{P_0} E_{P_0} \mathrm{Var}_n \left(m_{ti} - m_{ti}^* | \boldsymbol{D}_i \right) \leq K_m < \infty$

Assumption 2a effectively states that the posterior distribution of p_{ti} does not assign mass to neighborhoods of 0, and can be satisfied through prior distribution constraints. Assumption 2b states that the difference between the true conditional mean of the outcome and the corresponding posterior is bounded. This is a mild assumption in general, and is automatically satisfied if Y is categorical. Finally, we detail the assumptions required on either the treatment or outcome model:

Assumption 3 (Posterior contraction of treatment and outcome models) There exist two sequence of numbers $\epsilon_{nt} \to 0$ and $\epsilon_{ny} \to 0$, and constants $M_t > 0$ and $M_y > 0$ that are independent of ϵ_{nt} and ϵ_{ny}, respectively, such that

(a) $\sup\limits_{P_0} E_{P_0} \mathbb{P}_n \left(\frac{1}{\sqrt{n}} ||\boldsymbol{p}_t - \boldsymbol{p}_t^*|| > M_t \epsilon_{nt} | \boldsymbol{D} \right) \to 0$, and

(b) $\sup\limits_{P_0} E_{P_0} \mathbb{P}_n \left(\frac{1}{\sqrt{n}} ||\boldsymbol{m}_t - \boldsymbol{m}_t^*|| > M_y \epsilon_{ny} | \boldsymbol{D} \right) \to 0$,

where $||v|| = \sqrt{v_1^2 + \cdots + v_n^2}$.

Assumption 3a and 3b state that the posterior distribution of the treatment and outcome models contract at rates ϵ_{nt} and ϵ_{ny}, respectively. Achieving rates of posterior contraction such as these typically relies on their own set of assumptions, such as conditions on the design matrix $\boldsymbol{X}$ or sparsity. We will restrict discussion of these issues to relevant papers on posterior contraction in regression models (Castillo *et al.* , 2015; Yang *et al.* , 2015; Yoo *et al.* , 2016). Our goal is to examine the posterior of $\Delta(\boldsymbol{D}, \boldsymbol{\Psi})$, conditional on that either Assumption 3a or 3b is true. In particular, we show that we only need one of Assumptions 3a or 3b to hold in order to achieve posterior contraction of μ_t, though having both improves the rate of posterior contraction. Further, we show that this result holds uniformly over P_0.

5.2 Posterior contraction

We state the posterior contraction result for μ_t but the extension to $\mu_1 - \mu_2$ is trivial. We find a minimum possible sequence of numbers $\epsilon_n \to 0$, and a constant M which does not depend on ϵ_n, such that

$$\sup_{P_0} E_{P_0} \mathbb{P}_n(\boldsymbol{p}_t, \boldsymbol{m}_t : |\mu_t - \mu_t^*| > M\epsilon_n | \boldsymbol{D}) \to 0, \tag{5}$$

where ϵ_n defines the rate of contraction of the posterior. The faster ϵ_n converges to zero while maintaining this result, the faster our posterior distribution contracts.

Theorem 1: Assume positivity, no unmeasured confounding, SUTVA, and Assumptions 1 and 2 hold. If Assumptions 3a and 3b hold, Equation 5 is satisfied with $\epsilon_n = \max(n^{-1/2}, \epsilon_{nt}\epsilon_{ny})$. If only one of assumptions 3a or 3b hold, Equation 5 is satisfied with $\epsilon_n = \max(n^{-1/2}, \eta_n)$, where η_n is the contraction rate for the correctly specified model.

A proof of this result can be found in Appendix A. This results implies that posterior consistency is achieved as long as only one model is correctly specified (double robustness), regardless of the covariate dimension. Further, in high-dimensions where regression parameters will contract at slower rates than $n^{-1/2}$, we can still obtain a contraction rate of the treatment effect that is $n^{-1/2}$ if the product of their contraction rates is less than or equal to $n^{-1/2}$. For instance, in high-dimensional linear regression the posterior contraction rate of regression coefficients when using spike and slab priors is $\sqrt{\log p/n}$ (Castillo *et al.* , 2015). If both the treatment model and outcome model parameters contract at this rate, then the posterior distribution of $\Delta(\boldsymbol{D}, \boldsymbol{\Psi})$ still contracts at $n^{-1/2}$ as long as $\log p \leq n^{3/2}$. Our result has implications for low-dimensional models as well. If one is interested in using nonparametric priors, which contract more slowly than parametric models, but allow for highly flexible modeling of the regression models, $n^{-1/2}$ contraction can still be obtained under the same logic.

6 Simulation studies

6.1 Binary treatments

We set $n = 100$ and $p = 500$, and generate data as:

$$\boldsymbol{X}_i \sim N(\boldsymbol{0}_p, \boldsymbol{\Sigma}), \qquad T_i|\boldsymbol{X}_i \sim \text{Bernoulli}(p_i) \qquad Y_i|T_i, \boldsymbol{X}_i \sim \mathcal{N}(\mu_i, \boldsymbol{I}_n).$$

We set $\Sigma_{ij} = 1$ if $i = j$ and $\Sigma_{ij} = 0.3$ if $i \neq j$. We simulate data under two scenarios for the true propensity and outcome regressions:

$$\begin{aligned}
\textbf{Linear Simulation:} \quad & \mu_i = T_i + 0.75X_{1i} + X_{2i} + 0.6X_{3i} - 0.8X_{4i} - 0.7X_{5i} \\
& p_i = \Phi(0.15X_{1i} + 0.2X_{2i} - 0.4X_{5i}) \\
\textbf{Nonlinear Simulation:} \quad & \mu_i = T_i + 0.8X_{1i} + 0.4X_{2i}^3 + 0.25e^{|X_{2i}|} + 0.8X_{5i}^2 - 1.5\text{sin}(X_{5i}) \\
& p_i = \Phi(0.15X_{1i} - 0.4X_{2i} - 0.5X_{5i})
\end{aligned}$$

We estimate the average treatment effect using: a) double post selection regression (Double PS) introduced in Belloni *et al.* (2014); b) doubly robust estimators (lasso-DR) introduced in Farrell (2015); c) the residual de-biasing approach (De-biasing) of Athey *et al.* (2018); d) Targeted maximum likelihood (TMLE, Van Der Laan & Rubin (2006)) with lasso models; and e) the double machine learning approach (DML, Chernozhukov *et al.* (2016)) with lasso models. For each of these approaches, asymptotic standard errors were estimated, and confidence intervals were defined as the interval within 1.96 estimated standard errors of the point estimate. For the nonlinear scenarios we only compare with TMLE and DML, as the other approaches rely on linearity of models and do not immediately extend to nonlinear scenarios. For both of these approaches, we use an initial screening step from a group lasso model, and then fit nonlinear outcome models on the chosen covariates. Treatment models are specified to be linear. More details of our implementation of these approaches can be found in Appendix I. We implement our approach using the models in Section 4 and employing the doubly robust estimator. For the treatment and outcome models, we consider linear models, models using 3 degree of freedom splines for each covariate (Section 4.2), and models that use Gaussian process priors for each covariate (Section 4.1), and use the treatment and outcome model that minimizes WAIC ((Watanabe, 2010)). WAIC is a Bayesian analog of commonly used model selection criteria such as AIC or BIC. We will refer to this doubly robust estimator as Bayes-DR.

Figure 2 shows the results from the two simulation studies across both scenarios. The estimator proposed in this paper is in grey, while the existing approaches are in black. In the linear scenario, the models that estimate treatment effects using linear outcome models, the Double PS, TMLE, and DML do very well. Despite that, the Bayes-DR estimator is the only estimator that achieves interval coverages near the nominal level. In the nonlinear simulation, we see that our estimator obtains the lowest MSE of all approaches, and again achieves coverages close to the nominal level.

6.2 Continuous treatments

Here, we restrict attention to $n = 200$ and $p = 200$, and generate data as:

$$\boldsymbol{X}_i \sim N(\boldsymbol{0}_p, \boldsymbol{\Sigma}), \qquad T_i|\boldsymbol{X}_i \sim \mathcal{N}(\mu_i^t, 1), \qquad Y_i|T_i, \boldsymbol{X}_i \sim \mathcal{N}(\mu_i^y, 1)$$

where $\Sigma_{ij} = 1$ if $i = j$ and $\Sigma_{ij} = 0.3$ if $i \neq j$,

$$\begin{aligned}
\mu_i^t &= 0.6X_{1i}^2 + 0.6X_{1i} + \text{exp}(\text{abs}(0.65X_{2i})) - 0.8X_{3i}^2, \text{ and} \\
\mu_i^y &= 5 + 0.05T_i^3 - 0.1T_i^2 + 0.6X_{1i} + 0.4\text{exp}(X_{1i}) + \text{log}(\text{abs}(0.65X_{2i})) + 0.5(1 + X_{3i})^2.
\end{aligned}$$

Our estimand of interest is now the entire exposure response curve, therefore we will be estimating $E(Y(t))$ for all t in the support of T. To estimate this quantity for all t we use our treatment and

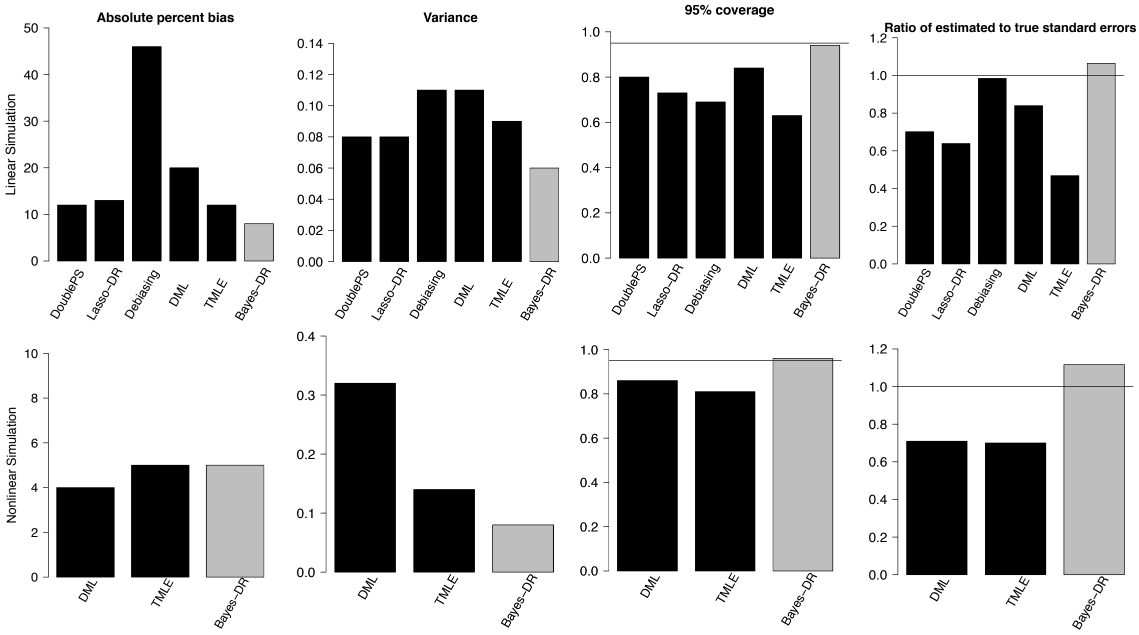

Figure 2: Results from simulations with binary treatments. The top panel shows results for the linear scenario, while the bottom panel shows results for the nonlinear scenario. The first column shows absolute bias, the second column shows the variance, the third column shows 95% interval coverages, while the fourth column is the ratio of estimated to Monte Carlo standard errors.

outcome models as described in Section 4 and apply either regression based estimators that only utilize the outcome model and marginalize over the covariate distribution, or the doubly robust estimator that was introduced in Kennedy *et al.* (2017). We will refer to the regression based estimators as Reg-1, Reg-3, and Reg-GP to denote the amount of nonlinearity allowed. The doubly robust estimator involves creating the pseudo-outcome:

$$\xi(\boldsymbol{D}_i, \boldsymbol{\Psi}) = \frac{Y_i - E(Y_i|T_i, \boldsymbol{X}_i)}{p(T_i = t|\boldsymbol{X}_i)} \int_{\mathcal{X}} p(T_i = t|\boldsymbol{X}_i) dP_n(\boldsymbol{X}) + \int_{\mathcal{X}} E(Y_i|T_i, \boldsymbol{X}_i) dP_n(\boldsymbol{X}), \tag{6}$$

where P_n is the empirical distribution of the data. Then, this pseudo-outcome is regressed against the treatment, potentially in a flexible, nonlinear manner. We use this estimator for treatment and outcome models built using the Bayesian machinery above to reduce the dimension of the covariate space, and perform inference using the resampling approach described in Section 3. Both the treatment and outcome models assume normality of the outcome. To assess the performance of the various methods at estimating the whole curve, we will evaluate the performance of each method at 20 distinct locations on the curve and average relevant metrics such as bias or interval coverage across the 20 locations. We will use cubic polynomials to model the exposure-response curve which encaptures the true curve, though any flexible approach could work here.

Figure 3 shows the results averaged across 1000 simulations. The Reg-1 estimator does very poorly in terms of MSE and interval coverages, which is expected because it assumes linearity, when the true model is highly nonlinear. The Reg-3, Reg-GP, and Bayes-DR approaches all allow for nonlinear relationships between the covariates and treatment/outcome, and therefore these approaches perform well with respect to all metrics. Again, the Bayes-DR estimator achieves interval coverages at or near the nominal level of 95%. The right panel of Figure 3 shows that the Bayes-DR estimator generally estimates the entire curve well, with very few simulations deviating from the true shape.

6.3 Summary of additional simulation results

In Appendices D-H we present additional simulation results using different data generating mechanisms, different p/n ratios, misspecified models, models that do not assume additivity, and bootstrap inference for the competing approaches. We find similar results under different data generating mechanisms that show that our approach does well in terms of MSE and finite sample interval coverage. As we increase the sample size and p/n ratio, this difference disappears as the asymptotic standard errors of existing approaches perform much better. Even though the bootstrap is not theoretically

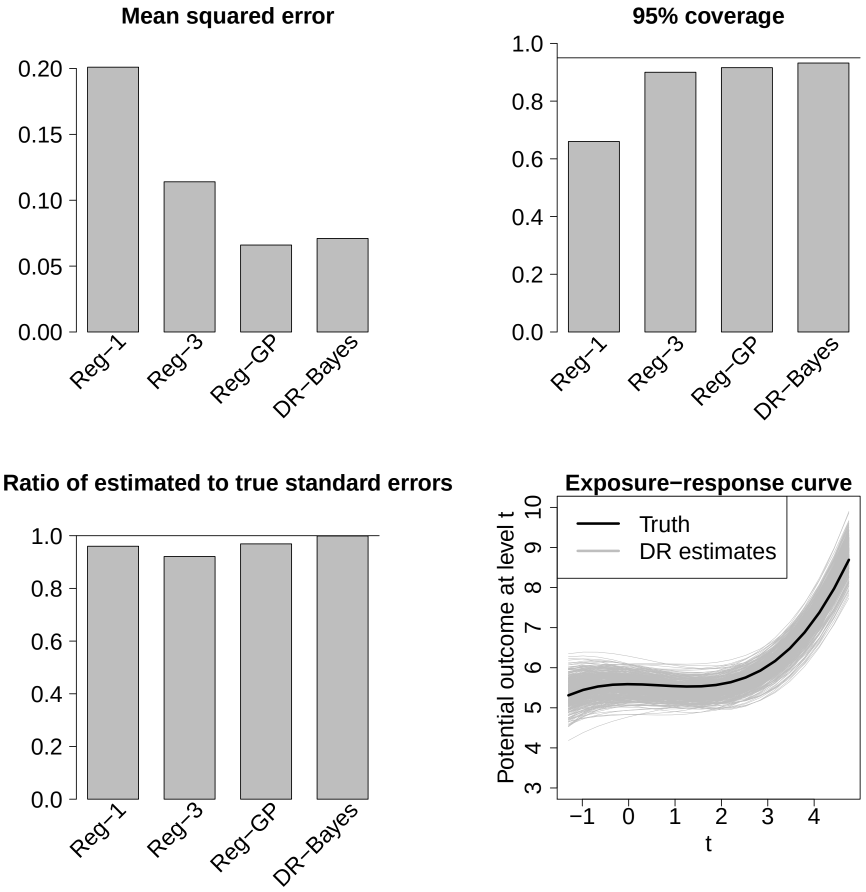

Figure 3: Simulation results for continuous treatments. The top left panel presents the mean squared error, the top right panel shows the 95% credible interval coverage, the bottom left panel shows the ratio of estimated to Monte Carlo standard errors, and the bottom right panel shows the estimates of the exposure-response curve across the 1000 simulations for the doubly robust estimator.

justified for every competing approach, we applied it to assess if our approach to inference performs better only due to the fact that it was the only approach using resampling. We found that bootstrap intervals were excessively large due to the erratic nature of the estimators in small samples, leading to interval coverages of 100% and very little power to detect signals. As the sample size increased, this problem also disappears and the bootstrap intervals perform very well, though only in scenarios when the asymptotic intervals also perform well.

7 Application to EWAS

Environmental wide association studies (EWAS) have been increasingly common in recent years as scientists attempt to gain a better understanding of how various chemicals and toxins affect the biological processes in the human body (Wild, 2005; Patel & Ioannidis, 2014). In particular, EWAS look to study the effects of a large number of exposures that humans are invariably exposed to on disease or other functions in the body. The National Health and Nutrition Examination Survey (NHANES) is a cross-sectional data source made publicly available by the Centers for Disease Control and Prevention (CDC). We will restrict attention to the 1999-2000, 2001-2002, 2003-2004, and 2005-2006 surveys, and we will aim to estimate the effects of environmental exposures on three different outcomes: HDL cholesterol levels, LDL cholesterol levels, and triglyceride levels in humans. We will use the data in Wilson *et al.* (2018), which studied the impact of environmental agents from the NHANES data. The study contains a large number of potential confounders as participants fill out questionnaires regarding their health status, and receive clinical and laboratory tests that contain information on environmental factors such as pollutants, allergens, bacterial/viral organisms, chemical toxicants, and nutrients. In previous work (Patel *et al.* , 2012), the environmental agents whose effect we estimate were separated into groups containing similar agents that might act through similar biological pathways. We will look at the effects of 14 different environmental agent groups on the three outcomes, leading to 42 different analyses. Each exposure we look at is defined as the average exposure level across all agents within the same grouping. In the NHANES data, different subjects had different environmental agents measured, leading to different populations, covariate dimensions, and sample sizes for each of the 14 exposures. We apply our Bayesian models in conjuction with the doubly robust estimator of Kennedy *et al.* (2017) to estimate the exposure response curves in each of the 42 analyses. Both p and n vary for each data set and there is a wide range of p/n ratios from 0.08 to 0.51, with a mean of 0.25.

7.1 Differing levels of nonlinearity and sparsity

To analyze the data we fit a treatment model and an outcome model under each of the three levels of flexibility that we used in the simulation study. This includes a linear function of the covariates, three degree of freedom splines, and Gaussian processes. For each data set we looked at the WAIC of both the treatment and outcome model, and used the model with the minimum WAIC. Figure 4 shows histograms of the ratio of the WAIC values with the minimum WAIC within a given dataset across the three models. A value of one indicates that a particular model had the best WAIC, while larger

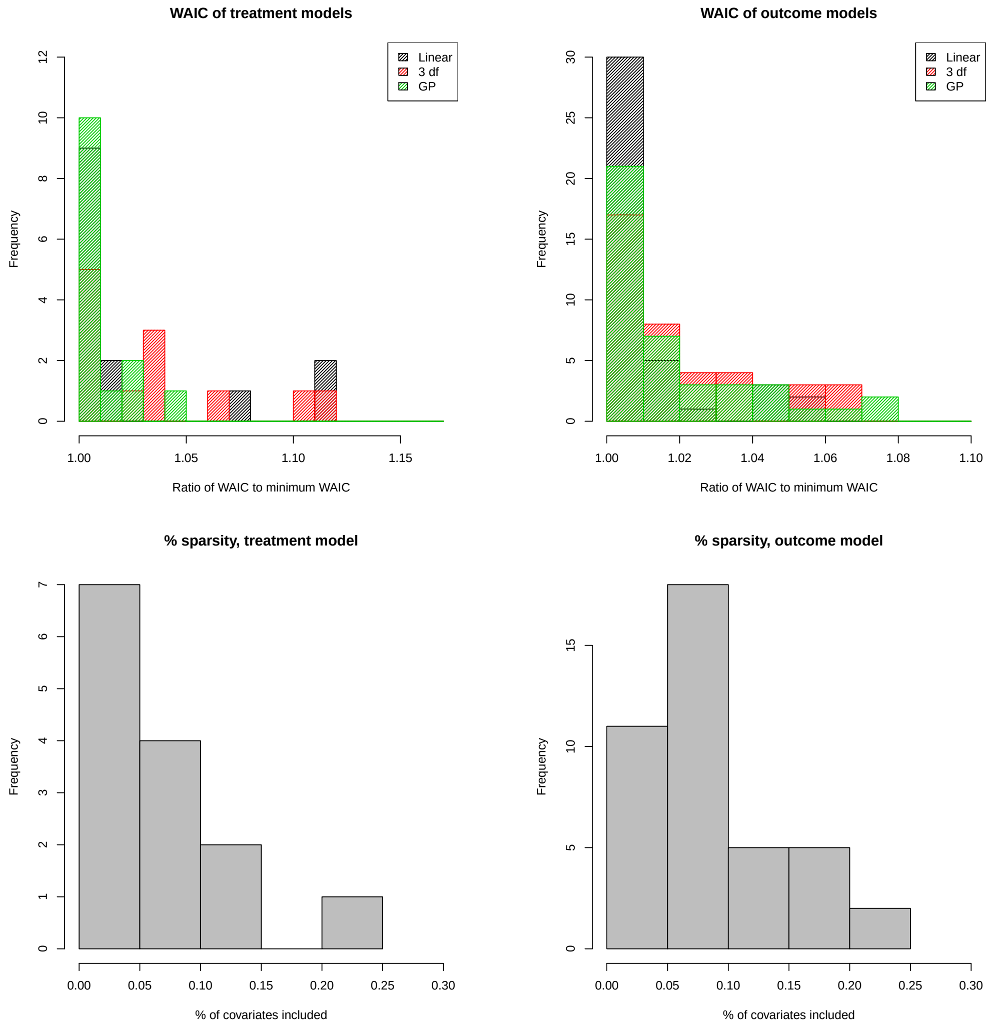

Figure 4: The top panel presents the ratio of WAIC values to the minimum values for each of the three models considered. The top left panel shows the treatment model WAIC values, while the top right panel shows the WAIC for the outcome models. The bottom panel shows the percentage of covariates included in the chosen treatment and outcome model.

values indicate worse fits to the data. We see that for the treatment model, the Gaussian process prior is selected more than any other model and most of the values in the histogram are less than 1.05. Linear models do the next best and have the lowest WAIC for a number of datasets. For the outcome model, the linear model does best, followed by the Gaussian process prior and spline model, which do similarly well. Overall, these plots suggest that different amounts of flexibility were required in these analyses, and our flexible approach might be more accurately depicting the true data generating processes.

We also examine the extent to which our sparsity inducing priors reduced the dimension of the covariate space. While our models provide posterior inclusion probabilities, we can report a binary assessment of a variable's importance by reporting whether the posterior inclusion probability is greater than 0.5. Figure 4 shows the percentage of covariates that have a posterior inclusion probability greater than 0.5 in the treatment and outcome models. It is clear that the spike and slab priors greatly reduce the number of covariates in both models as all datasets have less than 30% of the covariates in the models, and many less than 10%. Not shown in the figure is that there are even fewer covariates included in both models, indicating that there is not a lot of strong confounding within these datasets. This is further supported by the fact that many of the estimated exposure response curves are very similar to the curves one would get by not controlling for any covariates.

7.2 Exposure response curves

We highlight the estimation of the exposure response curves for three of the exposures in the analysis: Dioxins, Organochlorine pesticides, and Diakyl. The p/n ratio for these three analyses was 0.41, 0.18, and 0.34, respectively. Figure 5 shows the doubly robust estimate of the exposure response curve along with the naive curve one would get by not including any covariates in the analysis. The two estimated curves are fairly similar with a couple of exceptions. The effect of OC pesticides on Triglycerides has a much smaller slope when adjusting for covariates, and the effect of Diakyl on Triglycerides is much larger at lower levels of exposure when adjusting for covariates. In some areas of the curves there is less uncertainty in the doubly robust estimate, however, in general the naive curves have tighter uncertainty intervals. This is not entirely surprising as the doubly robust estimators are adjusting for a large number of covariates, which can decrease efficiency unless the covariates are highly predictive of the outcome. Importantly, the confidence intervals of the doubly robust estimator are not much wider than the naive curves, indicating that the dimension reduction from the spike and slab priors improves efficiency.

8 Discussion

We have introduced an approach for causal inference that has a number of desirable features. Our approach can be applied to semiparametric estimators of causal effects that rely on a treatment or outcome model, in the context of binary, categorical, or continuous treatments. This is particularly important as the literature on estimating the causal effect curve for continuous treatments is small, and has not been extended to high-dimensional scenarios. We showed our approach maintains asymptotic properties such as double robustness and posterior contraction rates, while showing improved performance in finite samples. In particular, our approach to inference is able to capture all of the uncertainty in the data, leading to nominal interval coverages when frequentist counterparts that rely on asymptotics have decreased interval coverage. Further, flexible Bayesian methods allow our approach to adapt to nonlinear relationships in the treatment and outcome models, reducing the impact of model misspecification. Our approach has widespread applicability, as many causal estimators can be written as functions of data and treatment and outcome models, and the ideas seen here will apply directly. This allows users to estimate causal effects using many desirable Bayesian tools such as nonparametric priors and spike and slab priors. While we focused on high-dimensional scenarios with spike and slab priors in this paper, the ideas presented apply to any type of modeling framework for the treatment and outcome models. Throughout, we have assumed that the treatment effect is homogeneous, i.e. that the treatment effect is constant across all levels of the covariate space. This assumption can easily be removed by using more flexible Bayesian methods for the outcome model.

Doubly robust estimation was first introduced in the Bayesian framework in Saarela *et al.* (2016), although there has been some debate about whether an estimate of counterfactual outcomes can utilize the propensity score within the Bayesian framework (Saarela *et al.* , 2015). Robins & Ritov (1997) showed that any Bayesian analysis honoring the likelihood principle can not utilize the propensity score. We take a somewhat different approach in this paper, as we do not attempt to address these concerns, nor are we concerned with whether or not the proposed approach is "fully" Bayesian. Our purpose is to show that Bayesian methods can be coupled with estimating equation based estimators

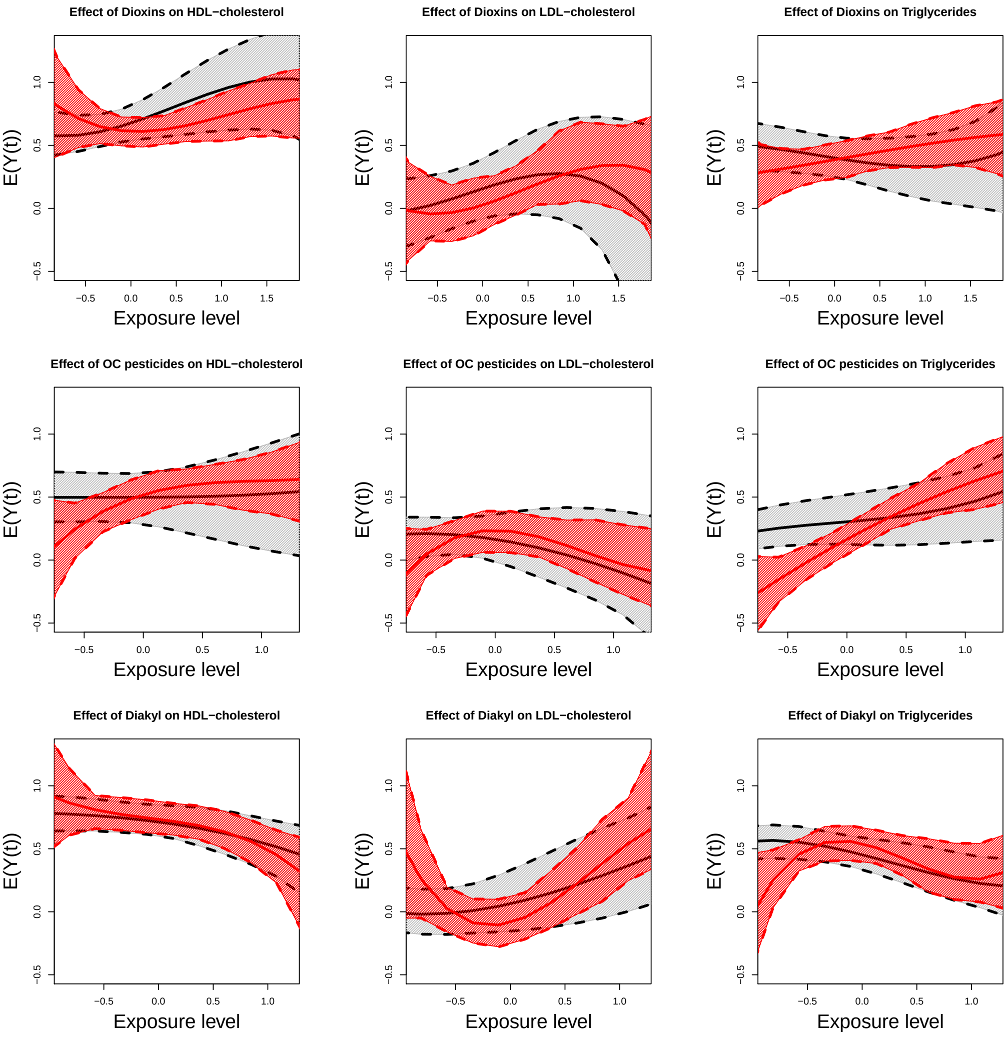

Figure 5: Estimated exposure response curves from the doubly robust estimator (black line) as well as the naive curve (red line), which does not adjust for any covariates.

in causal inference to provide flexible alternatives with desirable finite sample properties. This is even more important in high-dimensional scenarios where model uncertainty is higher and relying on asymptotics does not work well.

An important point is understanding why our approach to uncertainty estimation does better in finite samples than existing approaches rooted in asymptotic theory. Asymptotic approaches rely on certain terms vanishing as the sample size increases, and ignoring these additional terms can lead to anti-conservative inference in small samples. The sole manner in which our approach relies on asymptotics is that the bootstrap is only valid asymptotically. We account for parameter uncertainty through the posterior distribution which does not rely on asymptotics. Then, the bootstrap is used to account for the additional uncertainty in the estimators stemming from the observed data for fixed values of the parameters. When the parameters are fixed, the estimator has a simple form and the bootstrap should perform well in estimating uncertainty for simple quantities such as this. We have seen empirically that it works quite well and leads to nominal frequentist coverage. As the sample size increases, any differences in these two approaches to inference should dissipate.

Acknowledgement

The authors would like to thank Chirag Patel for help with the NHANES data, as well as Matthew Cefalu for helpful discussions on the manuscript. Funding for this work was provided by National Institutes of Health (R01MH118927, ES000002, ES024332, ES007142, ES026217, ES028033, P01CA134294, R01GM111339, R35CA197449, P50MD010428, DP2MD012722), The U.S Environmental Protection Agency (83615601, 83587201-0), and The Health Effects Institute (4953-RFA14-3/16-4).

Appendices

A Proof of posterior contraction rates

Proof: For simplicity we will drop the p_t, m_t notation in Equation 10 and write the quantity of interest as follows:

$$\begin{aligned}
\sup_{P_0} E_{P_0}\mathbb{P}_n(|\mu_t - \mu_t^*| > M\epsilon_n|\boldsymbol{D}) &= \sup_{P_0} E_{P_0}\mathbb{P}_n\left(\frac{1}{n}\left|\sum_{i=1}^n \frac{1(T_i = t)}{p_{ti}}(Y_i - m_{ti}) + m_{ti} - \mu_t^*\right| > M\epsilon_n|\boldsymbol{D}\right) \\
&= \sup_{P_0} E_{P_0}\mathbb{P}_n\left(\frac{1}{n}\left|\sum_{i=1}^n \frac{1(T_i = t)}{p_{ti}}(Y_i - m_{ti}) + m_{ti}\right.\right. \\
&- \frac{1(T_i = t)}{p_{ti}^*}(Y_i - m_{ti}^*) - m_{ti}^* \\
&+ \left.\left.\frac{1(T_i = t)}{p_{ti}^*}(Y_i - m_{ti}^*) + m_{ti}^* - \mu_t^*\right| > M\epsilon_n|\boldsymbol{D}\right) \\
&= \sup_{P_0} E_{P_0}\mathbb{P}_n\left(\left|A + B\right| > M\epsilon_n|\boldsymbol{D}\right).
\end{aligned}$$

where we can define A and B as follows:

$$\begin{aligned}
A &= \frac{1}{n}\sum_{i=1}^n \frac{1(T_i = t)}{p_{ti}}(Y_i - m_{ti}) + m_{ti} - \frac{1(T_i = t)}{p_{ti}^*}(Y_i - m_{ti}^*) - m_{ti}^* \\
B &= \frac{1}{n}\sum_{i=1}^n \frac{1(T_i = t)}{p_{ti}^*}(Y_i - m_{ti}^*) + m_{ti}^* - \mu_t^*
\end{aligned}$$

After re-arranging some terms, we can further decompose A into three separate parts such that $A = A_1 + A_2 + A_3$ where each are defined below:

$$\begin{aligned}
A_1 &= \frac{1}{n}\sum_{i=1}^n (m_{ti} - m_{ti}^*)\left(1 - \frac{1(T_i = t)}{p_{ti}^*}\right) \\
A_2 &= \frac{1}{n}\sum_{i=1}^n \frac{1(T_i = t)(p_{ti} - p_{ti}^*)(m_{ti}^* - Y_i)}{p_{ti}p_{ti}^*} \\
A_3 &= \frac{1}{n}\sum_{i=1}^n \frac{1(T_i = t)(p_{ti} - p_{ti}^*)(m_{ti} - m_{ti}^*)}{p_{ti}p_{ti}^*}.
\end{aligned}$$

We can now write the probability as

$$\begin{aligned}
\sup_{P_0} E_{P_0}\mathbb{P}_n(|\mu_t - \mu_t^*| > M\epsilon_n|\boldsymbol{D}) &= \sup_{P_0} E_{P_0}\mathbb{P}_n(|A_1 + A_2 + A_3 + B| > M\epsilon_n|\boldsymbol{D}) \\
&\leq \sup_{P_0} E_{P_0}\mathbb{P}_n(|A_1| > \frac{M}{4}\epsilon_n|\boldsymbol{D}) + \sup_{P_0} E_{P_0}\mathbb{P}_n(|A_2| > \frac{M}{4}\epsilon_n|\boldsymbol{D}) \\
&+ \sup_{P_0} E_{P_0}\mathbb{P}_n(|A_3| > \frac{M}{4}\epsilon_n|\boldsymbol{D}) + \sup_{P_0} E_{P_0}\mathbb{P}_n(|B| > \frac{M}{4}\epsilon_n|\boldsymbol{D}),
\end{aligned}$$

so it now suffices to show that each of the four components above contracts at the ϵ_n rate. We will begin with the B component, which does not depend on either the posterior distribution of the treatment or outcome model as it is simply the doubly robust estimator of μ_1 evaluated at the true values for the propensity score and outcome regression minus the parameter of interest.

$$\begin{aligned}
\sup_{P_0} E_{P_0}\mathbb{P}_n(|B| > \frac{M}{4}\epsilon_n|\boldsymbol{D}) &= \sup_{P_0} E_{P_0}\mathbb{P}_n\left(\left|\frac{1}{n}\sum_{i=1}^n \frac{1(T_i = t)}{p_{ti}^*}(Y_i - m_{ti}^*) + m_{ti}^* - \mu_t^*\right| > \frac{M}{4}\epsilon_n|\boldsymbol{D}\right) \\
&= \sup_{P_0} E_{P_0}1\left(\left|\frac{1}{n}\sum_{i=1}^n \frac{1(T_i = t)}{p_{ti}^*}(Y_i - m_{ti}^*) + m_{ti}^* - \mu_t^*\right| > \frac{M}{4}\epsilon_n\right)
\end{aligned}$$

$$= \sup_{P_0} P_{P_0}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\frac{1(T_i = t)}{p^*_{ti}}(Y_i - m^*_{ti}) + m^*_{ti} - \mu^*_t\right| > \frac{M}{4}\epsilon_n\right).$$

The second equality holds because all of the components of B are components of $\boldsymbol{D}$ and are therefore just constants when conditioning on $\boldsymbol{D}$. The quantity inside of the absolute values is easily shown to have expectation 0 with respect to P_0 and therefore we can apply Chebyshev's inequality to say:

$$\begin{aligned}\sup_{P_0} E_{P_0}\mathbb{P}_n(|B| > \frac{M}{4}\epsilon_n|\boldsymbol{D}) &\leq \sup_{P_0}\frac{16\text{Var}_{P_0}(B)}{M^2\epsilon_n^2}\\ &= \sup_{P_0}\frac{16\sigma_B^2}{M^2\epsilon_n^2 n}\end{aligned}$$

where $\sigma_B^2 = \text{Var}_{P_0}\left(\frac{1(T_i=t)}{p^*_{ti}}(Y_i - m^*_{ti})\right) \leq K_B < \infty$, for some constant K_B by assumption 1 and positivity. Clearly if $\epsilon_n > n^{-1/2}$ then this probability goes to zero and we have the desired result. Now, turning attention to A_1 let us first find $E_n(A_1|\boldsymbol{D})$, the posterior mean of A_1.

$$\begin{aligned}E_n(A_1|\boldsymbol{D}) &= E_n\left(\frac{1}{n}\sum_{i=1}^{n}(m_{ti} - m^*_{ti})\left(1 - \frac{1(T_i = t)}{p^*_{ti}}\right)|\boldsymbol{D}\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{1(T_i = t)}{p^*_{ti}}\right)E_n\left((m_{ti} - m^*_{ti}|\boldsymbol{D}\right).\end{aligned}$$

This quantity is nonzero and therefore we can not directly apply Chebyshev's inequality. We will add and subtract the posterior mean as follows:

$$\begin{aligned}\sup_{P_0} E_{P_0}\mathbb{P}_n(|A_1| > \frac{M}{4}\epsilon_n|\boldsymbol{D}) &= \sup_{P_0} E_{P_0}\mathbb{P}_n(|A_1 - E(A_1|D) + E(A_1|D)| > \frac{M}{4}\epsilon_n|\boldsymbol{D})\\ &\leq \sup_{P_0} E_{P_0}\mathbb{P}_n(|A_1 - E(A_1|D)| > \frac{M}{8}\epsilon_n|\boldsymbol{D})\\ &+ \sup_{P_0} E_{P_0}\mathbb{P}_n(|E(A_1|D)| > \frac{M}{8}\epsilon_n|\boldsymbol{D})\end{aligned}$$

Our goal is now to show that both of these quantities tend to 0 as $n \to \infty$. We can use Chebyshev's inequality on the first of the two probabilities to see that

$$\begin{aligned}\sup_{P_0} E_{P_0}\mathbb{P}_n(|A_1 - E(A_1|D)| > \frac{M}{8}\epsilon_n|\boldsymbol{D}) &\leq \sup_{P_0} E_{P_0}\frac{64\text{Var}_n(A_1|\boldsymbol{D})}{M^2\epsilon_n^2}\\ &= \sup_{P_0} E_{P_0}\frac{64\sigma^2_{A_{11}}}{M^2\epsilon_n^2 n}\end{aligned}$$

where $E_{P_0}(\sigma^2_{A_{11}}) = E_{P_0}\text{Var}_n\left((m_{ti} - m^*_{ti})\left(1 - \frac{1(T_i=t)}{p^*_{ti}}\right)|\boldsymbol{D}\right) \leq K_{A_{11}} < \infty$, for some constant $K_{A_{11}}$ by assumption 2 and positivity. Clearly, this limit goes to zero as long as $\epsilon_n > n^{-1/2}$. Now we can turn our attention to the second probability below:

$$\sup_{P_0} E_{P_0}\mathbb{P}_n(|E(A_1|D)| > \frac{M}{8}\epsilon_n|\boldsymbol{D}).$$

Notice, however, that $E(A_1|\boldsymbol{D})$ is constant with respect to the posterior distribution of the parameters given $\boldsymbol{D}$, and therefore this quantity simplifies to

$$\sup_{P_0} P_{P_0}(|E(A_1|D)| > \frac{M}{8}\epsilon_n) = \sup_{P_0} P_{P_0}\left(\Big|\frac{1}{n}\sum_{i=1}^{n}\Big(1 - \frac{1(T_i = t)}{p^*_{ti}}\Big)E_n\Big((m_{ti} - m^*_{ti}|\boldsymbol{D}\Big)\Big| > \frac{M}{8}\epsilon_n\right).$$

We will again use Chebyshev's inequality to bound this quantity, but first we need to show that the quantity inside the absolute value has expectation 0 with respect to P_0.

$$E_{P_0}\left(\frac{1}{n}\sum_{i=1}^{n}\Big(1 - \frac{1(T_i = t)}{p^*_{ti}}\Big)E_n\Big((m_{ti} - m^*_{ti})|\boldsymbol{D}\Big)\right)$$

$$
\begin{aligned}
&= E_{P_0}\left(\Big(1 - \frac{1(T_i = t)}{p^*_{ti}}\Big)E_n\Big((m_{ti} - m^*_{ti})|\boldsymbol{D}\Big)\right) \\
&= E_{Y(t),X}\left[E_{T|Y(t),X}\left(\Big(1 - \frac{1(T_i = t)}{p^*_{ti}}\Big)E_n\Big((m_{ti} - m^*_{ti})|\boldsymbol{D}\Big)\right)\right] \\
&= E_{Y(t),X}\left[E_n\Big((m_{ti} - m^*_{ti}|\boldsymbol{D}\Big)E_{T|Y(t),X}\left(1 - \frac{1(T_i = t)}{p^*_{ti}}\right)\right] \\
&= E_{Y(t),X}\left[E_n\Big((m_{ti} - m^*_{ti}|\boldsymbol{D}\Big)E_{T|X}\left(1 - \frac{1(T_i = t)}{p^*_{ti}}\right)\right] \\
&= E_{Y(t),X}\left[E_n\Big((m_{ti} - m^*_{ti}|\boldsymbol{D}\Big)\left(1 - \frac{p^*_{ti}}{p^*_{ti}}\right)\right] \\
&= 0.
\end{aligned}
$$

Where the fourth equality held due to the unconfoundedness assumption. Now we can apply Chebyshev's inequality to see that

$$
\sup_{P_0} P_{P_0}\left(\Big|\frac{1}{n}\sum_{i=1}^n \Big(1 - \frac{1(T_i = t)}{p^*_{ti}}\Big)E_n\Big((m_{ti} - m^*_{ti}|\boldsymbol{D}\Big)\Big| > \frac{M}{8}\epsilon_n\right) \leq \sup_{P_0} \frac{64\sigma^2_{A_{12}}}{M^2\epsilon_n^2 n},
$$

where $\sigma^2_{A_{12}} = \mathrm{Var}_{P_0}\left(\Big(1 - \frac{1(T_i=t)}{p^*_{ti}}\Big)E_n\Big((m_{ti} - m^*_{ti})|\boldsymbol{D}\Big)\right) \leq K_{A_{12}} < \infty$ for some constant $K_{A_{12}}$ by assumption 2 and positivity. This expression goes to zero as long as $\epsilon_n > n^{-1/2}$. Now we can look at A_2, which has a similar construction as A_1. Again, let us first calculate the posterior mean of A_2.

$$
\begin{aligned}
E_n(A_2|\boldsymbol{D}) &= E_n\left(\frac{1}{n}\sum_{i=1}^n \frac{1(T_i = t)(p_{ti} - p^*_{ti})(m^*_{ti} - Y_i)}{p_{ti}p^*_{ti}}\right)|\boldsymbol{D}\Big) \\
&= \frac{1}{n}\sum_{i=1}^n \left(\frac{1(T_i = t)(m^*_{ti} - Y_i)}{p^*_{ti}}\right)E_n\left(\frac{p_{ti} - p^*_{ti}}{p_{ti}}|\boldsymbol{D}\right).
\end{aligned}
$$

Again, this quantity does not necessarily have mean 0 and therefore we can not directly apply Chebyshev's inequality, but we can add and subtract the mean as before.

$$
\begin{aligned}
\sup_{P_0} E_{P_0}\mathbb{P}_n(|A_2| > \frac{M}{4}\epsilon_n|\boldsymbol{D}) &= \sup_{P_0} E_{P_0}\mathbb{P}_n(|A_2 - E(A_2|D) + E(A_2|D)| > \frac{M}{4}\epsilon_n|\boldsymbol{D}) \\
&\leq \sup_{P_0} E_{P_0}\mathbb{P}_n(|A_2 - E(A_2|D)| > \frac{M}{8}\epsilon_n|\boldsymbol{D}) \\
&\quad + \sup_{P_0} E_{P_0}\mathbb{P}_n(|E(A_2|D)| > \frac{M}{8}\epsilon_n|\boldsymbol{D}).
\end{aligned}
$$

Our goal is to show that both probabilities tend to zero and we have the desired result. For the first expression we can apply Chebyshev's inequality, since the quantity has mean 0:

$$
\begin{aligned}
\sup_{P_0} E_{P_0}\mathbb{P}_n(|A_2 - E(A_2|D)| > \frac{M}{8}\epsilon_n|\boldsymbol{D}) &\leq \sup_{P_0} E_{P_0}\frac{64\mathrm{Var}_n(A_2|\boldsymbol{D})}{M^2\epsilon_n^2} \\
&= \sup_{P_0} E_{P_0}\frac{64\sigma^2_{A_{21}}}{M^2\epsilon_n^2 n}
\end{aligned}
$$

where $E_{P_0}(\sigma^2_{A_{21}}) = E_{P_0}\mathrm{Var}_n\left(\frac{1(T_i=t)(p_{ti}-p^*_{ti})(m^*_{ti}-Y_i)}{p_{ti}p^*_{ti}}|\boldsymbol{D}\right) \leq K_{A_{21}} < \infty$ for some constant $K_{A_{21}}$ by assumptions 1, 2, and positivity. Clearly, this limit goes to zero as long as $\epsilon_n > n^{-1/2}$. Now we can turn our attention to the second probability below:

$$
\sup_{P_0} E_{P_0}\mathbb{P}_n(|E(A_2|D)| > \frac{M}{8}\epsilon_n|\boldsymbol{D}).
$$

Notice, however, that $E(A_2|\boldsymbol{D})$ is constant with respect to the posterior distribution of the parameters given $\boldsymbol{D}$, and therefore this quantity simplifies to

$$\sup_{P_0} P_{P_0}(|E(A_2|D)| > \frac{M}{8}\epsilon_n) = \sup_{P_0} P_{P_0}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1(T_i = t)(m^*_{ti} - Y_i)}{p^*_{ti}}\right)E_n\left(\frac{p_{ti} - p^*_{ti}}{p_{ti}}|\boldsymbol{D}\right) > \frac{M}{8}\epsilon_n\right).$$

We will again use Chebyshev's inequality to bound this quantity, but first we need to show that the quantity inside the absolute value has expectation 0 with respect to P_0.

$$\begin{aligned}
&E_{P_0}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1(T_i = t)(m^*_{ti} - Y_i)}{p^*_{ti}}\right)E_n\left(\frac{p_{ti} - p^*_{ti}}{p_{ti}}|\boldsymbol{D}\right)\right)\\
&= E_{P_0}\left(\left(\frac{1(T_i = t)(m^*_{ti} - Y_i)}{p^*_{ti}}\right)E_n\left(\frac{p_{ti} - p^*_{ti}}{p_{ti}}|\boldsymbol{D}\right)\right)\\
&= E_{T,X}\left[E_{Y|T,X}\left(\left(\frac{1(T_i = t)(m^*_{ti} - Y_i)}{p^*_{ti}}\right)E_n\left(\frac{p_{ti} - p^*_{ti}}{p_{ti}}|\boldsymbol{D}\right)\right)\right]\\
&= E_{T,X}\left[\left(\frac{1(T_i = t)}{p^*_{ti}}\right)E_n\left(\frac{p_{ti} - p^*_{ti}}{p_{ti}}|\boldsymbol{D}\right)E_{Y|T,X}(m^*_{ti} - Y_i)\right]\\
&= E_{T,X}\left[\left(\frac{1(T_i = t)}{p^*_{ti}}\right)E_n\left(\frac{p_{ti} - p^*_{ti}}{p_{ti}}|\boldsymbol{D}\right)(m^*_{ti} - m^*_{ti})\right]\\
&= 0
\end{aligned}$$

Now we can apply Chebyshev's inequality to see that

$$\sup_{P_0} P_{P_0}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1(T_i = t)(m^*_{ti} - Y_i)}{p^*_{ti}}\right)E_n\left(\frac{p_{ti} - p^*_{ti}}{p_{ti}}|\boldsymbol{D}\right)\right| > \frac{M}{8}\epsilon_n\right) \leq \sup_{P_0}\frac{64\sigma^2_{A_{22}}}{M^2\epsilon^2 n},$$

where $\sigma^2_{A_{22}} = \text{Var}_{P_0}\left(\left(\frac{1(T_i=t)(m^*_{ti}-Y_i)}{p^*_{ti}}\right)E_n\left(\frac{p_{ti}-p^*_{ti}}{p_{ti}}|\boldsymbol{D}\right)\right) \leq K_{A_{22}} < \infty$ for some constant $K_{A_{22}}$ by assumptions 1, 2, and positivity. This expression goes to zero as long as $\epsilon_n > n^{-1/2}$. Finally, we need to show contraction rates for A_3, which is where the double robustness property can be seen for the posterior distribution of μ_t.

$$\begin{aligned}
&\sup_{P_0} E_{P_0}\mathbb{P}_n(|A_3| > \frac{M}{4}\epsilon_n|\boldsymbol{D})\\
&= \sup_{P_0} E_{P_0}\mathbb{P}_n\left(\left|\frac{1}{n}\sum_{i=1}^{n}\frac{1(T_i = t)(p_{ti} - p^*_{ti})(m_{ti} - m^*_{ti})}{p_{ti}p^*_{ti}}\right| > \frac{M}{4}\epsilon_n|\boldsymbol{D}\right)\\
&\leq \sup_{P_0} E_{P_0}\mathbb{P}_n\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1(T_i = t)(p_{ti} - p^*_{ti})}{p_{ti}p^*_{ti}}\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_{ti} - m^*_{ti})^2} > \frac{M}{4}\epsilon_n|\boldsymbol{D}\right)\\
&\leq \sup_{P_0} E_{P_0}\mathbb{P}_n\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}K_{A_{31}}(p_{ti} - p^*_{ti})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_{ti} - m^*_{ti})^2} > \frac{M}{4}\epsilon_n|\boldsymbol{D}\right)\\
&= \sup_{P_0} E_{P_0}\mathbb{P}_n\left(\frac{1}{n}||\boldsymbol{p}_t - \boldsymbol{p}^*_t||\ ||\boldsymbol{m}_t - \boldsymbol{m}^*_t|| > \frac{M\epsilon_n}{4\sqrt{K_{A_{31}}}}|\boldsymbol{D}\right).
\end{aligned}$$

The first inequality comes from the Cauchy-Schwartz inequality, and the second inequality holds true for some constant $0 < K_{A_{31}} < \infty$ from assumption 2. Using the law of total probability we can separate this probability into scenarios when the outcome model contracts at rate $\epsilon_n^{\nu_1}$, and when it doesn't.

$$\begin{aligned}
&\sup_{P_0} E_{P_0}\mathbb{P}_n(|A_3| > \frac{M}{4}\epsilon_n|\boldsymbol{D})\\
&\leq \sup_{P_0} E_{P_0}\left[\mathbb{P}_n\left(\frac{1}{\sqrt{n}}||\boldsymbol{p}_t - \boldsymbol{p}^*_t|| > \frac{M\epsilon_n}{4\sqrt{K_{A_{31}}}\frac{1}{\sqrt{n}}||\boldsymbol{m}_t - \boldsymbol{m}^*_t||}|\boldsymbol{D}, \frac{1}{\sqrt{n}}||\boldsymbol{m}_t - \boldsymbol{m}^*_t|| < \epsilon_n^{\nu_1}\right)\right.\\
&\times \mathbb{P}_n\left(\frac{1}{\sqrt{n}}||\boldsymbol{m}_t - \boldsymbol{m}^*_t|| < \epsilon_n^{\nu_1}|\boldsymbol{D}\right)
\end{aligned}$$

$$+\mathbb{P}_n\left(\frac{1}{\sqrt{n}}||\boldsymbol{p}_t-\boldsymbol{p}_t^*||>\frac{M\epsilon_n}{4\sqrt{K_{A_{31}}\frac{1}{\sqrt{n}}||\boldsymbol{m}_t-\boldsymbol{m}_t^*||}}|\boldsymbol{D},\frac{1}{\sqrt{n}}||\boldsymbol{m}_t-\boldsymbol{m}_t^*||\geq\epsilon_n^{\nu_1}\right)$$
$$\times\mathbb{P}_n\left(\frac{1}{\sqrt{n}}||\boldsymbol{m}_t-\boldsymbol{m}_t^*||\geq\epsilon_n^{\nu_1}|\boldsymbol{D}\right)\Bigg]$$

First, let's assume that the outcome model contracts at the $\epsilon_n^{\nu_1}$ rate, i.e $\sup\limits_{P_0}E_{P_0}\mathbb{P}_n(\frac{1}{\sqrt{n}}||\boldsymbol{m}_t-\boldsymbol{m}_t^*||<\epsilon_n^{\nu_1}|\boldsymbol{D})\to 1$.

$$\begin{aligned}
&\sup_{P_0}E_{P_0}\mathbb{P}_n(|A_3|>\frac{M}{4}\epsilon_n|\boldsymbol{D})\\
&\leq\sup_{P_0}E_{P_0}\Bigg[\mathbb{P}_n\left(\frac{1}{\sqrt{n}}||\boldsymbol{p}_t-\boldsymbol{p}_t^*||>\frac{M\epsilon_n}{4\sqrt{K_{A_{31}}\frac{1}{\sqrt{n}}||\boldsymbol{m}_t-\boldsymbol{m}_t^*||}}|\boldsymbol{D},\frac{1}{\sqrt{n}}||\boldsymbol{m}_t-\boldsymbol{m}_t^*||<\epsilon_n^{\nu_1}\right)\\
&\times\mathbb{P}_n\left(\frac{1}{\sqrt{n}}||\boldsymbol{m}_t-\boldsymbol{m}_t^*||<\epsilon_n^{\nu_1}|\boldsymbol{D}\right)\Bigg]\\
&+\sup_{P_0}E_{P_0}\left[\mathbb{P}_n\left(\frac{1}{\sqrt{n}}||\boldsymbol{m}_t-\boldsymbol{m}_t^*||\geq\epsilon_n^{\nu_1}|\boldsymbol{D}\right)\right]\\
&\leq\sup_{P_0}E_{P_0}\left[\mathbb{P}_n\left(\frac{1}{\sqrt{n}}||\boldsymbol{p}_t-\boldsymbol{p}_t^*||>\frac{M\epsilon_n}{4\sqrt{K_{A_{31}}\frac{1}{\sqrt{n}}||\boldsymbol{m}_t-\boldsymbol{m}_t^*||}}|\boldsymbol{D},\frac{1}{\sqrt{n}}||\boldsymbol{m}_t-\boldsymbol{m}_t^*||<\epsilon_n^{\nu_1}\right)\right]\\
&+\sup_{P_0}E_{P_0}\left[\mathbb{P}_n\left(\frac{1}{\sqrt{n}}||\boldsymbol{m}_t-\boldsymbol{m}_t^*||\geq\epsilon_n^{\nu_1}|\boldsymbol{D}\right)\right]\\
&\leq\sup_{P_0}E_{P_0}\left[\mathbb{P}_n\left(\frac{1}{\sqrt{n}}||\boldsymbol{p}_t-\boldsymbol{p}_t^*||>\frac{M\epsilon_n}{4\sqrt{K_{A_{31}}\epsilon_n^{\nu_1}}}|\boldsymbol{D}\right)\right]+\sup_{P_0}E_{P_0}\left[\mathbb{P}_n\left(\frac{1}{\sqrt{n}}||\boldsymbol{m}_t-\boldsymbol{m}_t^*||\geq\epsilon_n^{\nu_1}|\boldsymbol{D}\right)\right]\\
&=\sup_{P_0}E_{P_0}\left[\mathbb{P}_n\left(\frac{1}{\sqrt{n}}||\boldsymbol{p}_t-\boldsymbol{p}_t^*||>\frac{M\epsilon_n^{1-\nu_1}}{4\sqrt{K_{A_{31}}}}|\boldsymbol{D}\right)\right]+\sup_{P_0}E_{P_0}\left[\mathbb{P}_n\left(\frac{1}{\sqrt{n}}||\boldsymbol{m}_t-\boldsymbol{m}_t^*||\geq\epsilon_n^{\nu_1}|\boldsymbol{D}\right)\right].
\end{aligned}$$

The first of these two expressions converges to 0 if the treatment model contracts at the $\epsilon_n^{1-\nu_1}$ rate. The second expression converges to 0 by the definition of posterior contraction for the outcome model. This shows that if both the treatment and outcome model contract, then the posterior of the average treatment effect contracts at a faster rate than either of the individual models. We saw earlier that A_1, A_2, and B contract at the $\epsilon_n=n^{-1/2}$ rate and therefore the average treatment effect can not contract any faster than $n^{-1/2}$. This result shows that our posterior contracts at the $n^{-1/2}$ rate if the product of the contraction rates for the treatment and outcome model is less than or equal to $n^{-1/2}$. If the product is larger than $n^{-1/2}$ then the posterior of the treatment effect converges at the product of the contraction rates for the treatment and outcome models. This is a Bayesian analog to results seen in Farrell (2015), and shows that we can use flexible or high-dimensional models and obtain fast posterior contraction rates due to the double robustness property. Now, let's examine the situation when the outcome model does not contract, i.e $\sup\limits_{P_0}E_{P_0}\mathbb{P}_n(\frac{1}{\sqrt{n}}||\boldsymbol{m}_t-\boldsymbol{m}_t^*||<\epsilon_n^{\nu_1}|\boldsymbol{D})\to 0$ for any ν_1 such that $\epsilon_n^{\nu_1}\to 0$.

$$\begin{aligned}
&\sup_{P_0}E_{P_0}\mathbb{P}_n(|A_3|>\frac{M}{4}\epsilon_n|\boldsymbol{D})\\
&\leq\sup_{P_0}E_{P_0}\Bigg[\mathbb{P}_n\left(\frac{1}{\sqrt{n}}||\boldsymbol{p}_t-\boldsymbol{p}_t^*||>\frac{M\epsilon_n}{4\sqrt{K_{A_{31}}\frac{1}{\sqrt{n}}||\boldsymbol{m}_t-\boldsymbol{m}_t^*||}}|\boldsymbol{D},\frac{1}{\sqrt{n}}||\boldsymbol{m}_t-\boldsymbol{m}_t^*||<\epsilon_n^{\nu_1}\right)\\
&\times\mathbb{P}_n\left(\frac{1}{\sqrt{n}}||\boldsymbol{m}_t-\boldsymbol{m}_t^*||<\epsilon_n^{\nu_1}|\boldsymbol{D}\right)\\
&+\mathbb{P}_n\left(\frac{1}{\sqrt{n}}||\boldsymbol{p}_t-\boldsymbol{p}_t^*||>\frac{M\epsilon_n}{4\sqrt{K_{A_{31}}\frac{1}{\sqrt{n}}||\boldsymbol{m}_t-\boldsymbol{m}_t^*||}}|\boldsymbol{D},\frac{1}{\sqrt{n}}||\boldsymbol{m}_t-\boldsymbol{m}_t^*||\geq\epsilon_n^{\nu_1}\right)\\
&\times\mathbb{P}_n\left(\frac{1}{\sqrt{n}}||\boldsymbol{m}_t-\boldsymbol{m}_t^*||\geq\epsilon_n^{\nu_1}|\boldsymbol{D}\right)\Bigg]\\
&\leq\sup_{P_0}E_{P_0}\left[\mathbb{P}_n\left(\frac{1}{\sqrt{n}}||\boldsymbol{m}_t-\boldsymbol{m}_t^*||<\epsilon_n^{\nu_1}|\boldsymbol{D}\right)\right]
\end{aligned}$$

$$+ \sup_{P_0} E_{P_0} \left[\mathbb{P}_n \left(\frac{1}{\sqrt{n}} ||\boldsymbol{p}_t - \boldsymbol{p}_t^*|| > \frac{M\epsilon_n}{4\sqrt{K_{A_{31}}} \frac{1}{\sqrt{n}} ||\boldsymbol{m}_t - \boldsymbol{m}_t^*||} |\boldsymbol{D}, \frac{1}{\sqrt{n}} ||\boldsymbol{m}_t - \boldsymbol{m}_t^*|| \geq \epsilon_n^{\nu_1} \right) \right]$$
$$\leq \sup_{P_0} E_{P_0} \left[\mathbb{P}_n \left(\frac{1}{\sqrt{n}} ||\boldsymbol{m}_t - \boldsymbol{m}_t^*|| < \epsilon_n^{\nu_1} |\boldsymbol{D} \right) \right] + \sup_{P_0} E_{P_0} \left[\mathbb{P}_n \left(\frac{1}{\sqrt{n}} ||\boldsymbol{p}_t - \boldsymbol{p}_t^*|| > \frac{M\epsilon_n}{4\sqrt{K_{A_{31}}} K_{A_{32}}} |\boldsymbol{D} \right) \right].$$

The last inequality holds for some constant $0 < K_{A_{32}} < \infty$ by assumption 2. The first expression goes to zero since the outcome model is misspecified and does not contract at any rate, while the second expression converges to zero if the treatment model contracts at rate ϵ_n. This shows that even if the outcome model is misspecified, we get contraction of the treatment effect at a rate of either $n^{-1/2}$ or ϵ_n, whichever is larger. This shows both the double robustness property and the contraction rate of the posterior when only the treatment model is correctly specified. An analogous proof holds for when the treatment model is misspecified and the outcome model contracts at a given rate, though we leave out the proof as it is nearly identical to the one above.

B Details of posterior sampling

Here we will present the details required for posterior sampling from both the semiparametric and nonparametric priors utilized. Throughout we will denote the full observed data as $\boldsymbol{D}_i = (Y_i, T_i, \boldsymbol{X}_i)$. First we will present the posterior sampling for the semiparametric prior that models the conditional associations between the treatment/outcome and covariates using splines with d degrees of freedom. We will be always be working with $\boldsymbol{X}$ being standardized to have mean zero and variance 1, which is crucial when using spike and slab priors. Throughout, we will show how to estimate the outcome model, but sampling from the treatment model is analagous with straightforward alterations. Finally, we will be working with the latent outcome Y_i^*, where in the case of continuous data, $Y_i^* = Y_i$. If Y_i is binary, then at every iteration of our MCMC we draw Y_i^* from a truncated normal distribution with mean set to $\beta_0 + f_t(T_i) + \sum_{j=1}^p f_j(X_{ji})$ and variance set to 1. If $Y_i = 1$ then this distribution is truncated below by 0 and if $Y_i = 0$ then it is truncated above by 0. Once we have obtained Y_i^*, then posterior sampling can continue using the latent outcome as if we had linear regression, even if the outcome is binary.

MCMC sampling for semiparametric prior

Below we detail the full conditional updates for all parameters in the model.

1. If Y_i is binary then set $\sigma^2 = 1$, and if the outcome is continuous draw σ^2 from an inverse-gamma distribution with parameters a^* and b^*, defined as:

$$a^* = a_{\sigma^2} + \frac{n}{2} + \frac{d \sum_{j=1}^p \gamma_j}{2}$$
$$b^* = b_{\sigma^2} + \frac{\sum_{i=1}^n \left(Y_i^* - \beta_0 - f_t(T_i) - \sum_{j=1}^p f_j(\boldsymbol{X_{ji}}) \right)^2}{2} + \sum_{j=1}^p \sum_{k=1}^d \frac{\beta_{jk}^2}{2\sigma_{\boldsymbol{\beta}}^2}$$

2. While not discussed in the main text, we will be placing a $\mathcal{IG}(a_{\sigma_{\boldsymbol{\beta}}^2}, b_{\sigma_{\boldsymbol{\beta}}^2})$ prior on $\sigma_{\boldsymbol{\beta}}^2$ and therefore we can update from the full conditional:

$$\sigma_{\boldsymbol{\beta}}^2 | \bullet \sim \mathcal{IG} \left(a_{\sigma_{\boldsymbol{\beta}}^2} + \frac{d \sum_{j=1}^p \gamma_j}{2}, b_{\sigma_{\boldsymbol{\beta}}^2} + \sum_{j=1}^p \sum_{k=1}^d \frac{\beta_{jk}^2}{2\sigma^2} \right)$$

3. Update θ from the full conditional:

$$\theta | \bullet \sim \mathcal{B} \left(a_\theta + \sum_{j=1}^p \gamma_j, b_\theta + \sum_{j=1}^p (1 - \gamma_j) \right)$$

4. To update γ_j for $j = 1 \ldots p$ we need to look at the conditional posterior that has marginalized over $\boldsymbol{\beta}_j$. Specifically, if we allow $\boldsymbol{\Lambda}$ to represent all parameters in the model except for $(\gamma_j, \boldsymbol{\beta}_j)$ then we can update γ_j from the following conditional distribution:

$$p(\gamma_j = 1 | \boldsymbol{D}, \boldsymbol{\Lambda}) = \frac{p(\boldsymbol{\beta_j} = \boldsymbol{0}, \gamma_j = 1 | \boldsymbol{D}, \boldsymbol{\Lambda})}{p(\boldsymbol{\beta_j} = \boldsymbol{0} | \gamma_j = 1, \boldsymbol{D}, \boldsymbol{\Lambda})}$$

$$
\begin{aligned}
&= \frac{p(\boldsymbol{D}, \boldsymbol{\Lambda}|\boldsymbol{\beta_j} = \boldsymbol{0}, \gamma_j = 1)p(\boldsymbol{\beta_j} = \boldsymbol{0}, \gamma_j = 1)}{p(\boldsymbol{D}, \boldsymbol{\Lambda})p(\boldsymbol{\beta_j} = \boldsymbol{0}|\gamma_j = 1, \boldsymbol{D}, \boldsymbol{\Lambda})} \\
&= \frac{p(\boldsymbol{D}, \boldsymbol{\Lambda}|\boldsymbol{\beta_j} = \boldsymbol{0})p(\boldsymbol{\beta_j} = \boldsymbol{0}, \gamma_j = 1)}{p(\boldsymbol{D}, \boldsymbol{\Lambda})p(\boldsymbol{\beta_j} = \boldsymbol{0}|\gamma_j = 1, \boldsymbol{D}, \boldsymbol{\Lambda})} \\
&\propto \frac{p(\boldsymbol{\beta_j} = \boldsymbol{0}, \gamma_j = 1)}{p(\boldsymbol{\beta_j} = \boldsymbol{0}|\gamma_j = 1, \boldsymbol{D}, \boldsymbol{\Lambda})} \\
&= \frac{\theta \ \Phi(\boldsymbol{0}; \boldsymbol{0}, \boldsymbol{\Sigma_\beta})}{\Phi(\boldsymbol{0}; \boldsymbol{M}, \boldsymbol{V})}
\end{aligned}
$$

where $\Phi()$ represents the multivariate normal density function. $\boldsymbol{M}$ and $\boldsymbol{V}$ represent the conditional posterior mean and variance for $\boldsymbol{\beta}_j$ when $\gamma_j = 1$ and can be defined as

$$
\boldsymbol{M} = \left(\frac{\widetilde{\boldsymbol{X}}_j^T \widetilde{\boldsymbol{X}}_j}{\sigma^2} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} \right)^{-1} \widetilde{\boldsymbol{X}}_j^T \widetilde{\boldsymbol{Y}}, \qquad \boldsymbol{V} = \left(\frac{\widetilde{\boldsymbol{X}}_j^T \widetilde{\boldsymbol{X}}_j}{\sigma^2} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} \right)^{-1},
$$

where $\widetilde{\boldsymbol{Y}} = \boldsymbol{Y}^* - \beta_0 - f_t(\boldsymbol{T}) - \sum_{k \neq p} f_k(\boldsymbol{X}_k)$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ is a $d-$dimensional diagonal matrix with $\sigma^2 \sigma^2_{\boldsymbol{\beta}}$ on the diagonals.

5. For $j = 1 \dots p$, if $\gamma_j = 1$ update $\boldsymbol{\beta}_j$ from a multivariate normal distribution with mean $\boldsymbol{M}$ and variance $\boldsymbol{V}$ as defined above. If $\gamma_j = 0$, then set $\boldsymbol{\beta}_j = \boldsymbol{0}$.
6. We will jointly update β_0 and $f_t(\boldsymbol{T})$. For now we will let $f_t(\boldsymbol{T}) = \beta_t T$, though the full conditional will take the same form even if we model $f_t(\boldsymbol{T})$ with polynomials or splines. Define $\boldsymbol{Z}_t = [\boldsymbol{1}', \boldsymbol{T}]$, then the full conditional is of the form

$$
(\beta_0, \beta_t)|\bullet \sim MVN \left(\left(\frac{\boldsymbol{Z}_t^T \boldsymbol{Z}_t}{\sigma^2} + \boldsymbol{\Sigma}_t^{-1} \right)^{-1} \boldsymbol{Z}_t^T \widetilde{\boldsymbol{Y}}, \left(\frac{\boldsymbol{Z}_t^T \boldsymbol{Z}_t}{\sigma^2} + \boldsymbol{\Sigma}_t^{-1} \right)^{-1} \right)
$$

where $\widetilde{\boldsymbol{Y}} = \boldsymbol{Y}^* - \sum_{j=1}^p f_j(\boldsymbol{X}_j)$ and $\boldsymbol{\Sigma_t}$ is a diagonal matrix with K on the diagonals, with K large so that the treatment effect is not heavily shrunk towards zero.

MCMC sampling with gaussian process priors

Now we will detail the posterior sampling for the model defined in Section 2.2.

1. Update $(\theta, \beta_0, \beta_t)$ using the same updates as above for the semiparametric prior specification.
2. To update γ_j for $j = 1 \dots p$ we need to look at the conditional posterior that has marginalized over $f_j(\boldsymbol{X}_j)$. Specifically, if we allow $\boldsymbol{\Lambda}$ to represent all parameters in the model except for $(\gamma_j, f_j(\boldsymbol{X}_j))$ then we can update γ_j from the following conditional distribution:

$$
\begin{aligned}
p(\gamma_j = 1|\boldsymbol{D}, \boldsymbol{\Lambda}) &= \frac{p(f_j(\boldsymbol{X}_j) = \boldsymbol{0}, \gamma_j = 1|\boldsymbol{D}, \boldsymbol{\Lambda})}{p(f_j(\boldsymbol{X}_j) = \boldsymbol{0}|\gamma_j = 1, \boldsymbol{D}, \boldsymbol{\Lambda})} \\
&= \frac{p(\boldsymbol{D}, \boldsymbol{\Lambda}|f_j(\boldsymbol{X}_j) = \boldsymbol{0}, \gamma_j = 1)p(f_j(\boldsymbol{X}_j) = \boldsymbol{0}, \gamma_j = 1)}{p(\boldsymbol{D}, \boldsymbol{\Lambda})p(f_j(\boldsymbol{X}_j) = \boldsymbol{0}|\gamma_j = 1, \boldsymbol{D}, \boldsymbol{\Lambda})} \\
&= \frac{p(\boldsymbol{D}, \boldsymbol{\Lambda}|f_j(\boldsymbol{X}_j) = \boldsymbol{0})p(f_j(\boldsymbol{X}_j) = \boldsymbol{0}, \gamma_j = 1)}{p(\boldsymbol{D}, \boldsymbol{\Lambda})p(f_j(\boldsymbol{X}_j) = \boldsymbol{0}|\gamma_j = 1, \boldsymbol{D}, \boldsymbol{\Lambda})} \\
&\propto \frac{p(f_j(\boldsymbol{X}_j) = \boldsymbol{0}, \gamma_j = 1)}{p(f_j(\boldsymbol{X}_j) = \boldsymbol{0}|\gamma_j = 1, \boldsymbol{D}, \boldsymbol{\Lambda})} \\
&= \frac{\theta \ \Phi(\boldsymbol{0}; \boldsymbol{0}, \sigma^2 \tau^2 \boldsymbol{\Sigma_j})}{\Phi(\boldsymbol{0}; \boldsymbol{M}, \boldsymbol{V})}
\end{aligned}
$$

where $\Phi()$ represents the multivariate normal density function. $\boldsymbol{M}$ and $\boldsymbol{V}$ represent the conditional posterior mean and variance for $f_j(\boldsymbol{X}_j)$ when $\gamma_j = 1$ and can be defined as

$$
\boldsymbol{M} = \left(\boldsymbol{I}_n + \frac{1}{\tau_j^2} \boldsymbol{\Sigma}_{\boldsymbol{j}}^{-1} \right)^{-1} \widetilde{\boldsymbol{Y}}, \qquad \boldsymbol{V} = \left(\boldsymbol{I}_n + \frac{1}{\tau_j^2} \boldsymbol{\Sigma}_{\boldsymbol{j}}^{-1} \right)^{-1},
$$

where $\widetilde{\boldsymbol{Y}} = \boldsymbol{Y}^* - \beta_0 - f_t(\boldsymbol{T}) - \sum_{k \neq p} f_k(\boldsymbol{X}_k)$.

3. For $j = 1 \dots p$, if $\gamma_j = 1$ update $f_j(\boldsymbol{X}_j)$ from a multivariate normal distribution with mean $\boldsymbol{M}$ and variance $\boldsymbol{V}$ as defined above. If $\gamma_j = 0$, then set $f_j(\boldsymbol{X}_j) = \boldsymbol{0}$.

4. If $\gamma_j = 0$, update τ_j^2 from it's prior distribution, which is a Gamma$(1/2, 1/2)$. If $\gamma_j = 1$, update τ_j^2 from the following distribution:

$$\mathcal{IG}\left(\frac{n+1}{2}, \frac{1}{2} + \frac{f_j(\boldsymbol{X}_j)^T \boldsymbol{\Sigma}_j^{-1} f_j(\boldsymbol{X}_j)}{2\sigma^2} \right)$$

5. If Y_i is binary then set $\sigma^2 = 1$, and if the outcome is continuous draw σ^2 from an inverse-gamma distribution with parameters a^* and b^* defined as:

$$a^* = a_{\sigma^2} + \frac{n(1 + \sum_{j=1}^p \gamma_j)}{2}$$

$$b^* = b_{\sigma^2} + \frac{\sum_{i=1}^n \left(Y_i^* - \beta_0 - f_t(T_i) - \sum_{j=1}^p f_j(\boldsymbol{X_{ji}}) \right)^2}{2} + \sum_{j=1}^p \frac{\gamma_j f_j(\boldsymbol{X}_j)^T \boldsymbol{\Sigma}_j^{-1} f_j(\boldsymbol{X}_j)}{2\tau_j^2}$$

One thing to note is that in the conditional updates for $(\gamma_j, f_j(\boldsymbol{X}_j))$, we must calculate $\left(\boldsymbol{I}_n + \frac{1}{\tau_j^2} \boldsymbol{\Sigma}_j^{-1} \right)^{-1}$, which means inverting an n by n matrix at every MCMC iteration. To avoid this, we can first compute the singular value decomposition, $\boldsymbol{\Sigma}_j = \boldsymbol{A}\boldsymbol{B}\boldsymbol{A}^T$, where $\boldsymbol{A}$ is a matrix of eigenvectors and $\boldsymbol{B}$ is a diagonal matrix of eigenvalues. From this, it can be shown that $\left(\boldsymbol{I}_n + \frac{1}{\tau_j^2} \boldsymbol{\Sigma}_{\boldsymbol{j}}^{-1} \right)^{-1} = \boldsymbol{A}\big(\boldsymbol{I}_n + \frac{\boldsymbol{B}^{-1}}{\tau_j^2}\big)^{-1} \boldsymbol{A}^T$, which only requires inverting a diagonal matrix and can be computed much faster.

C Understanding the variance estimate

Throughout we use the subscript $\boldsymbol{D}$ to denote the distribution of the observed data, $\boldsymbol{D}^{(m)}$ to denote the empirical distribution of the data from which we draw new data sets, and $\boldsymbol{D}^{obs}$ is the observed data.

In this section, we explicitly write out the error in using $\text{Var}_{\boldsymbol{D}^{(m)}}\big[E_{\boldsymbol{\Psi}|\boldsymbol{D}^{obs}}\big(\Delta(\boldsymbol{D}^{(m)}, \boldsymbol{\Psi})\big)\big]$ as the estimate for the variance $\text{Var}_{\boldsymbol{D}}\big[E_{\boldsymbol{\Psi}|\boldsymbol{D}}\big(\Delta(\boldsymbol{D}, \boldsymbol{\Psi})\big)\big]$. This will provide insight as to why the variance correction tends to recover the true variance or a conservative estimate, leading to nominal coverage rates. We can write out their difference as follows:

$$\begin{aligned}
\text{Difference} &= \text{Var}_{\boldsymbol{D}}\Big[E_{\boldsymbol{\Psi}|\boldsymbol{D}}\Big(\Delta(\boldsymbol{D}, \boldsymbol{\Psi})\Big)\Big] - \text{Var}_{\boldsymbol{D}^{(m)}}\Big[E_{\boldsymbol{\Psi}|\boldsymbol{D}^{obs}}\Big(\Delta(\boldsymbol{D}^{(m)}, \boldsymbol{\Psi})\Big)\Big] \\
&= E_{\boldsymbol{D}}\Big[E^2_{\boldsymbol{\Psi}|\boldsymbol{D}}\Big(\Delta(\boldsymbol{D}, \boldsymbol{\Psi})\Big)\Big] - E^2_{\boldsymbol{D}}\Big[E_{\boldsymbol{\Psi}|\boldsymbol{D}}\Big(\Delta(\boldsymbol{D}, \boldsymbol{\Psi})\Big)\Big] \\
&\quad - E_{\boldsymbol{D}^{(m)}}\Big[E^2_{\boldsymbol{\Psi}|\boldsymbol{D}^{obs}}\Big(\Delta(\boldsymbol{D}^{(m)}, \boldsymbol{\Psi})\Big)\Big] + E^2_{\boldsymbol{D}^{(m)}}\Big[E_{\boldsymbol{\Psi}|\boldsymbol{D}^{obs}}\Big(\Delta(\boldsymbol{D}^{(m)}, \boldsymbol{\Psi})\Big)\Big] \\
&\approx E_{\boldsymbol{D}}\Big[E^2_{\boldsymbol{\Psi}|\boldsymbol{D}}\Big(\Delta(\boldsymbol{D}, \boldsymbol{\Psi})\Big)\Big] - E^2_{\boldsymbol{D}}\Big[E_{\boldsymbol{\Psi}|\boldsymbol{D}}\Big(\Delta(\boldsymbol{D}, \boldsymbol{\Psi})\Big)\Big] \\
&\quad - E_{\boldsymbol{D}}\Big[E^2_{\boldsymbol{\Psi}|\boldsymbol{D}^{obs}}\Big(\Delta(\boldsymbol{D}, \boldsymbol{\Psi})\Big)\Big] + E^2_{\boldsymbol{D}}\Big[E_{\boldsymbol{\Psi}|\boldsymbol{D}^{obs}}\Big(\Delta(\boldsymbol{D}, \boldsymbol{\Psi})\Big)\Big] \\
&\approx E_{\boldsymbol{D}}\Big[E^2_{\boldsymbol{\Psi}|\boldsymbol{D}}\Big(\Delta(\boldsymbol{D}, \boldsymbol{\Psi})\Big)\Big] - E_{\boldsymbol{D}}\Big[E^2_{\boldsymbol{\Psi}|\boldsymbol{D}^{obs}}\Big(\Delta(\boldsymbol{D}, \boldsymbol{\Psi})\Big)\Big].
\end{aligned}$$

The first approximation stems from the fact that we replaced $E_{\boldsymbol{D}^{(m)}}$ with $E_{\boldsymbol{D}}$. This is a reasonable approximation as the distribution of $\boldsymbol{D}^{(m)}$ is the empirical distribution of a sample from the distribution of $\boldsymbol{D}$. The second approximation stems from the fact that the average of the posterior mean using the observed data posterior, $E_{\boldsymbol{D}}E_{\boldsymbol{\Psi}|\boldsymbol{D}^{obs}}$ should closely resemble $E_{\boldsymbol{D}}E_{\boldsymbol{\Psi}|\boldsymbol{D}}$ as the posterior mean from one sample is a reasonable estimate of the posterior mean averaged over repeated samples.

We now look at the correction term, the term added to the initial variance. We show this term closely resembles the difference above, thereby producing a variance estimate that is of the correct magnitude.

$$\begin{aligned}
\text{Correction} &= E_{\boldsymbol{D}^{(m)}}\Big[\text{Var}_{\boldsymbol{\Psi}|\boldsymbol{D}^{obs}}\Big(\Delta(\boldsymbol{D}^{(m)}, \boldsymbol{\Psi})\Big)\Big] \\
&\approx E_{\boldsymbol{D}}\Big[\text{Var}_{\boldsymbol{\Psi}|\boldsymbol{D}^{obs}}\Big(\Delta(\boldsymbol{D}, \boldsymbol{\Psi})\Big)\Big] \\
&= E_{\boldsymbol{D}}\Big[E_{\boldsymbol{\Psi}|\boldsymbol{D}^{obs}}\Big(\Delta(\boldsymbol{D}, \boldsymbol{\Psi})^2\Big)\Big] - E_{\boldsymbol{D}}\Big[E^2_{\boldsymbol{\Psi}|\boldsymbol{D}^{obs}}\Big(\Delta(\boldsymbol{D}, \boldsymbol{\Psi})\Big)\Big]
\end{aligned}$$

Again, the approximation stems from the fact that we replace $E_{\boldsymbol{D}^{(m)}}$ with $E_{\boldsymbol{D}}$. Note that the second term in the correction is exactly the same as the second term in the difference.

The question is whether the first term in the variance correction $E_{\boldsymbol{D}}\left[E_{\boldsymbol{\Psi}|\boldsymbol{D}^{obs}}\left(\Delta(\boldsymbol{D},\boldsymbol{\Psi})^2\right)\right]$ is of a similar magnitude to the first term in the difference $E_{\boldsymbol{D}}\left[E^2_{\boldsymbol{\Psi}|\boldsymbol{D}}\left(\Delta(\boldsymbol{D},\boldsymbol{\Psi})\right)\right]$. We see that the first term in the correction involves the second moment of Δ with respect to $\boldsymbol{\Psi}|\boldsymbol{D}^{obs}$, while the first term in the difference involves the first moment with respect to $\boldsymbol{\Psi}|\boldsymbol{D}$ squared. It is difficult to compare these two terms with any certainty as they refer to moments with respect to different distributions. Due to Jensen's equality, we generally expect the second moment to be larger than the first moment squared. While this cannot be shown to be true in general, because these two quantities refer to different posterior distributions, this provides some evidence that our correction will lead to a conservative estimate of the variance. In Appendix H we evaluate how this correction does across all simulation scenarios looked at, and we see that it always leads to estimates of the variance that are close to the sampling distribution variance as desired, and generally are slightly conservative.

D Using the bootstrap for competing approaches

Here, we will assess whether bootstrapping the competing approaches can provide improved inference over the asymptotic intervals used in the manuscript. Bootstrapping is justified and has been used for both TMLE (Schnitzer *et al.* , 2015) and double machine learning (Knaus, 2018). For the other estimators, it is not clear whether the standard nonparametric bootstrap would provide valid inference, however, here we will perform inference using the nonparametric bootstrap for all competing approaches to evaluate whether it provides better finite sample performance. This is to address the question of whether our approach is only doing better in finite samples because we are resampling, while the other approaches are not. The results of the two approaches to inference can be seen in Figure A.1.

In the left panel, we see the results seen in the manuscript that show approaches based on asymptotic confidence intervals are anti-conservative and obtain confidence interval coverages well below the nominal level. In the right figure, we see that the bootstrap intervals present the opposite problem. With the exception of the residual de-biasing approach that achieves low coverages due to large amounts of bias in the estimator, most estimators have confidence interval coverages at 100% or 99% in the case of the lasso-DR approach. These intervals are far too wide with average estimated standard errors well above two times the standard error of the estimators. This indicates that these intervals are extremely conservative and lead to confidence intervals that are far too wide. This is due to the erratic nature of the bootstrapped estimators in small samples when variable selection is involved in the estimation procedure.

To assess whether these approaches to inference correct themselves as the sample size increases, we tried the same test, but with $n = 400$ and $p = 50$. The results can be seen in Figure A.2. We see that when we are in a scenario with a larger sample size and more favorable p/n ratio that both methods to inference provide interval coverages very close to the nominal level for all approaches considered.

E Additional simulation scenarios

Here we will run a number of additional simulation scenarios to assess the performance of the proposed approach. The first three scenarios will be in the binary treatment setting, while the fourth will look at a continuous exposure response curve. The first two scenarios will be from sparse data generating models that have different functional forms from the simulations in the paper, while the third scenario looks at a non sparse setting to see how the method performs when the assumption of sparsity does not hold. The fourth simulation differs from the continuous treatment simulation of the paper in that it has linear relationships between the covariates and treatment / outcome instead of nonlinear ones.

Scenario 1

Here we will run an identical simulation to the linear simulation scenario from the manuscript with $n = 100$ and $p = 500$, except now we will generate data from the following treatment and outcome models:

$$\mu_i = T_i + 0.45X_{1i} + 0.7X_{2i} - 0.6X_{3i} + 1.3X_{4i} - 0.5X_{5i}$$
$$p_i = \Phi(0.35X_{1i} + 0.2X_{2i} - 0.3X_{3i} - 0.4X_{5i})$$

The results of this simulation study can be seen in Figure A.3. We see that the proposed approach does very well in terms of variance, as only the TMLE estimator is lower. All approaches with the

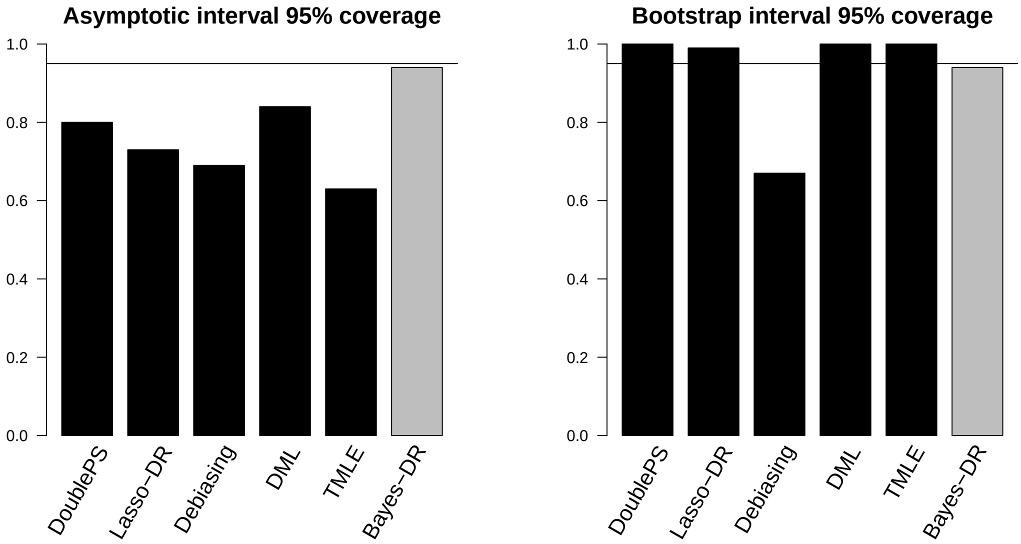

Figure A.1: The left panel shows the results of the linear simulation study from the manuscript using asymptotic standard errors for the competing approaches. The right panel shows the same simulation study with bootstrap based confidence intervals. In both figures, the Bayes-DR approach uses the approach to inference seen in Section 4 of the manuscript.

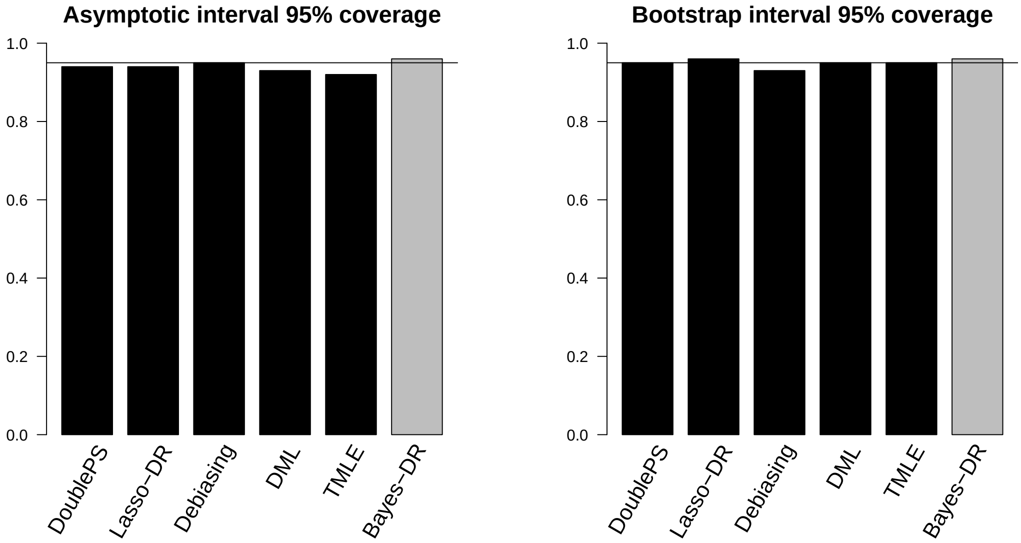

Figure A.2: Results from the linear simulation study with $n = 400$ and $p = 50$. The left panel shows the results using asymptotic standard errors for the competing approaches. The right panel shows the same simulation study with bootstrap based confidence intervals. In both figures, the Bayes-DR approach uses the approach to inference seen in Section 4 of the manuscript.

exception of TMLE had non-negligible bias so none of them obtain the nominal coverage rate. Our approach, however, appears to be accounting for more of the finite sample uncertainty in the estimation of the causal effect, which leads to a much higher coverage rate. Our approach has a ratio of the true to Monte Carlo standard errors that is slightly above 1, while the remaining approaches with the exception of the debiasing approach are far below 1. While not shown here, we again tried to use bootstrapping approaches to fix the problems with undercoverage, and we found erratic confidence intervals from the bootstrap that were far too wide and led to 100% coverages for the competing approaches.

Scenario 2

Now, similarly to an experiment run by Belloni *et al.* (2014), we will generate data such that $\mu_i = T_i + \boldsymbol{X_i\beta}$, and $p_i = \Phi(\boldsymbol{X_i\alpha})$, where $\boldsymbol{\beta} = \boldsymbol{\alpha} = (1, 1/4, 1/9, \dots, 1/p^2)$. This situation is not strictly sparse as none of the coefficients are exactly zero, though it is approximately sparse in the sense that a small number of confounders can remove essentially all of the confounding bias. We will set $n = 100$ and $p = 300$.

The results of this simulation study can be seen in Figure A.4. Most approaches have small bias with the exception of TMLE and debiasing. The Lasso-DR and DML procedures have a very high variance in this setting while the other approaches have similar variances to each other. The Double PS, Lasso-DR, and DML procedures all obtain reasonable levels of coverage only slightly below the nominal level, which is caused by the fact that the ratio of their estimated to true standard errors is below 1. Our approach is slightly conservative in this setting as we obtain coverages slightly above 95% with a ratio of true to estimated standard errors that is around 1.2.

Scenario 3

Next, we will look at scenario that comes from Athey *et al.* (2018), where the propensity score is dense. First we define 20 clusters, $\{\mathbf{c_1}, \dots, \mathbf{c_{20}}\}$ where $\mathbf{c_k} \sim \mathcal{N}(0, I_{pxp})$. Second, we draw $\mathbf{C}_i$ uniformly at random from one of the 20. Third, we draw the covariates from a multivariate normal distribution centered at $\mathbf{C}_i$ with the identity matrix as the covariance. Fourth, we set $T_i = 1$ with probability 0.25 for the first 10 clusters, and $T_i = 1$ with probability 0.75 for the remaining clusters. Finally, we generate data from the outcome model defined as $Y_i = 10T_i + \boldsymbol{X\beta} + \epsilon_i$, where $\boldsymbol{\beta} \propto (1, \frac{1}{2}, \dots, \frac{1}{p})$ and is normalized such that $||\boldsymbol{\beta}||_2^2 = 18$. Here we will again set $n = 100$ and $p = 300$. Intuitively, this is a simulation scenario in which the outcome model is approximately sparse, though the treatment model is dense as all of the covariates are associated with the treatment.

The results of this simulation study can be seen in Figure A.5. The Double PS approach works the best in this setting in terms of bias and variance. This is probably due to the fact that it uses an outcome model in the end to perform inference, and therefore is less affected by the dense propensity model. Our approach has a higher variance due to the dense propensity score model, but importantly still obtains 95% interval coverage as we seem to be accurately estimating the uncertainty in our estimator.

Scenario 4

Here we will set $n = 200$ and $p = 200$. We will generate data from the following models:

$$\begin{aligned} Y_i|T_i, \boldsymbol{X}_i &\sim \mathcal{N}(\mu_i^y, 1) \\ T_i|\boldsymbol{X}_i &\sim \mathcal{N}(\mu_i^t, 1) \\ \boldsymbol{X}_i &\sim N(\mathbf{0}_p, \boldsymbol{I}_n), \end{aligned}$$

where

$$\begin{aligned} \mu_i^y &= 5 - 0.1T_i + 0.05T_i^3 + 0.5X_{1i} + 0.5X_{2i} - 0.3X_{5i} \\ \mu_i^t &= 0.4X_{1i} + 0.6X_{2i} - 0.5X_{4i}, \end{aligned}$$

The results can be seen in Figure A.6. In this scenario, the confounding structure is linear and therefore all of the approaches are able to achieve small amounts of bias. Unsurprisingly, the model assuming linearity does the best of all the approaches, while the performance gets slightly worse in terms of MSE as the amount of nonlinearity increases. Importantly though, the DR-Bayes estimator is still able to achieve 95% interval coverages across the range of the exposure, and the ratio of estimated to true standard errors is quite close to 1. This indicates that our strategy for variance estimation is performing well and leading to valid inference.

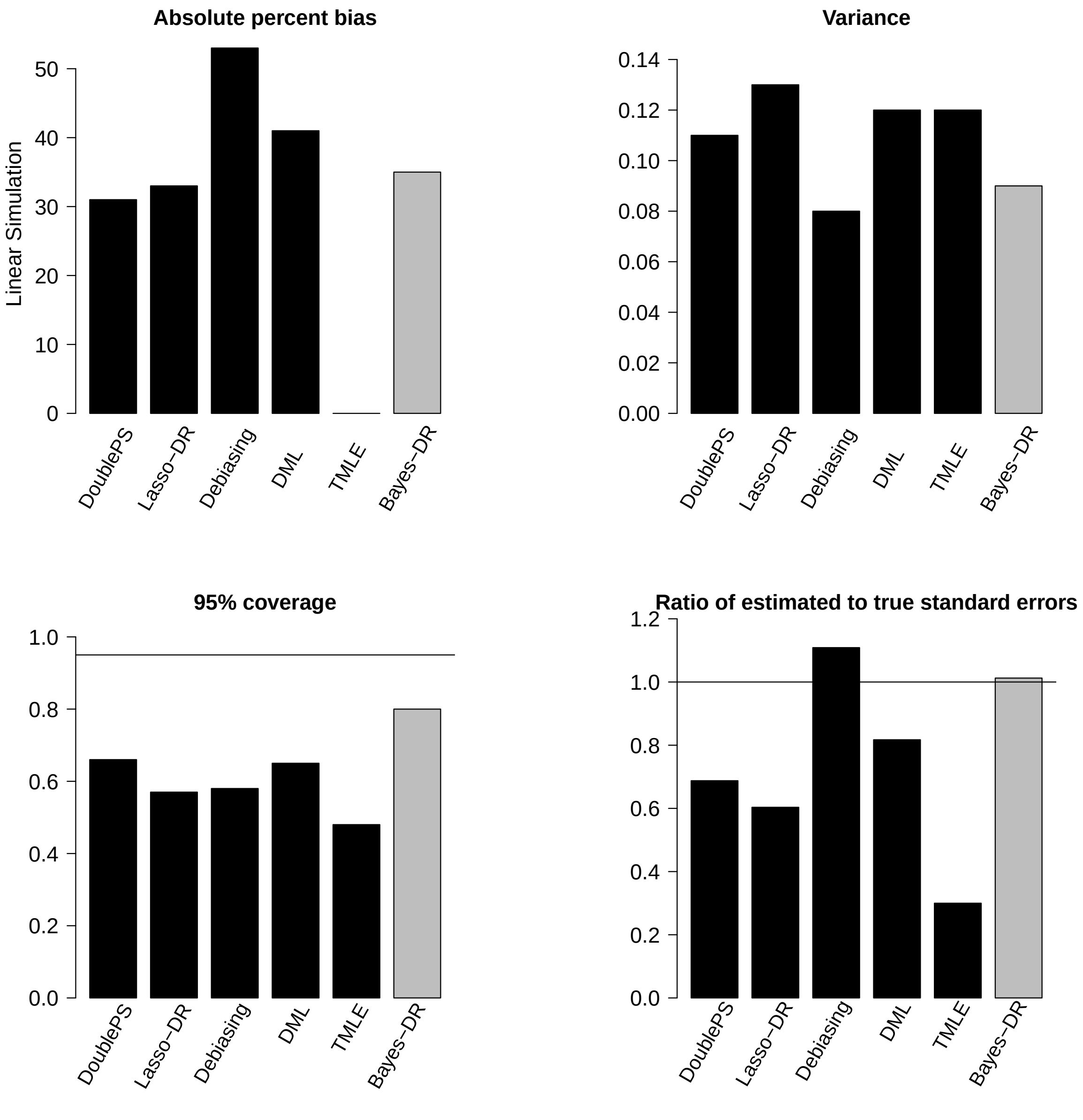

Figure A.3: Results from the first additional simulation scenario. The top left panel shows absolute bias, the top right panel shows the variance, the bottom left panel shows 95% interval coverages, while the bottom right panel is the ratio of estimated to Monte Carlo standard errors.

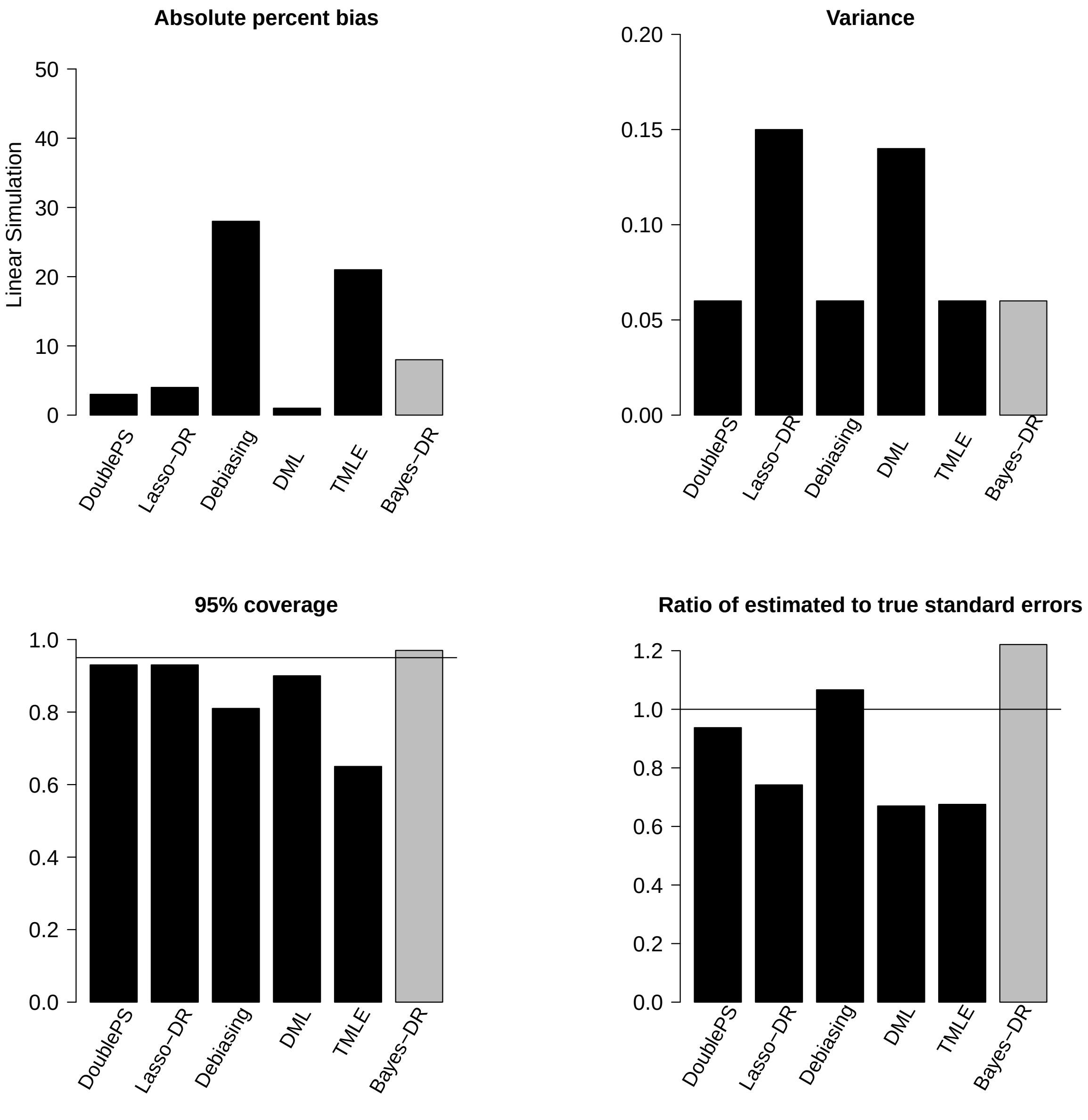

Figure A.4: Results from the second additional simulation scenario. The top left panel shows absolute bias, the top right panel shows the variance, the bottom left panel shows 95% interval coverages, while the bottom right panel is the ratio of estimated to Monte Carlo standard errors.

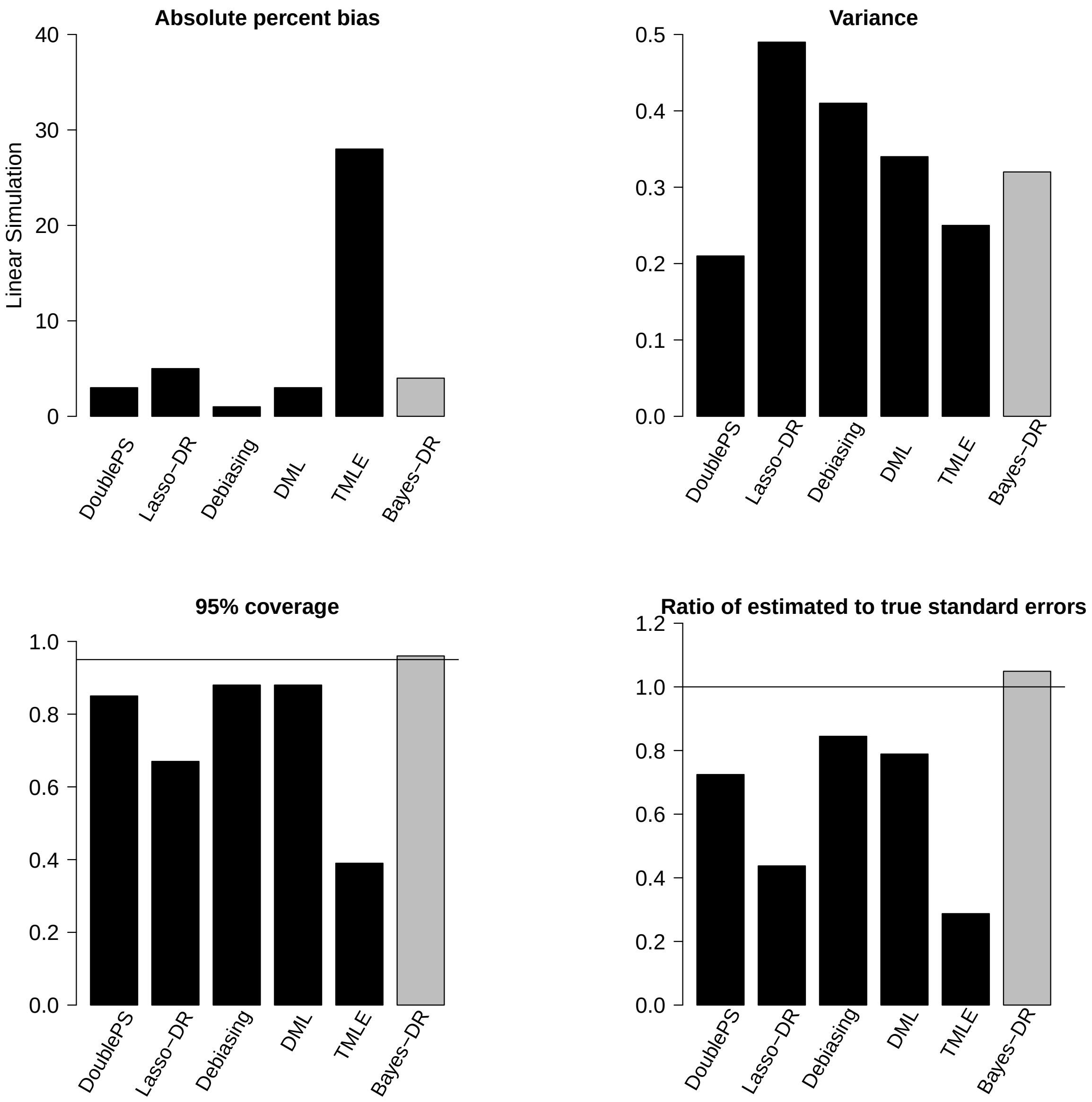

Figure A.5: Results from the third additional simulation scenario. The top left panel shows absolute bias, the top right panel shows the variance, the bottom left panel shows 95% interval coverages, while the bottom right panel is the ratio of estimated to Monte Carlo standard errors.

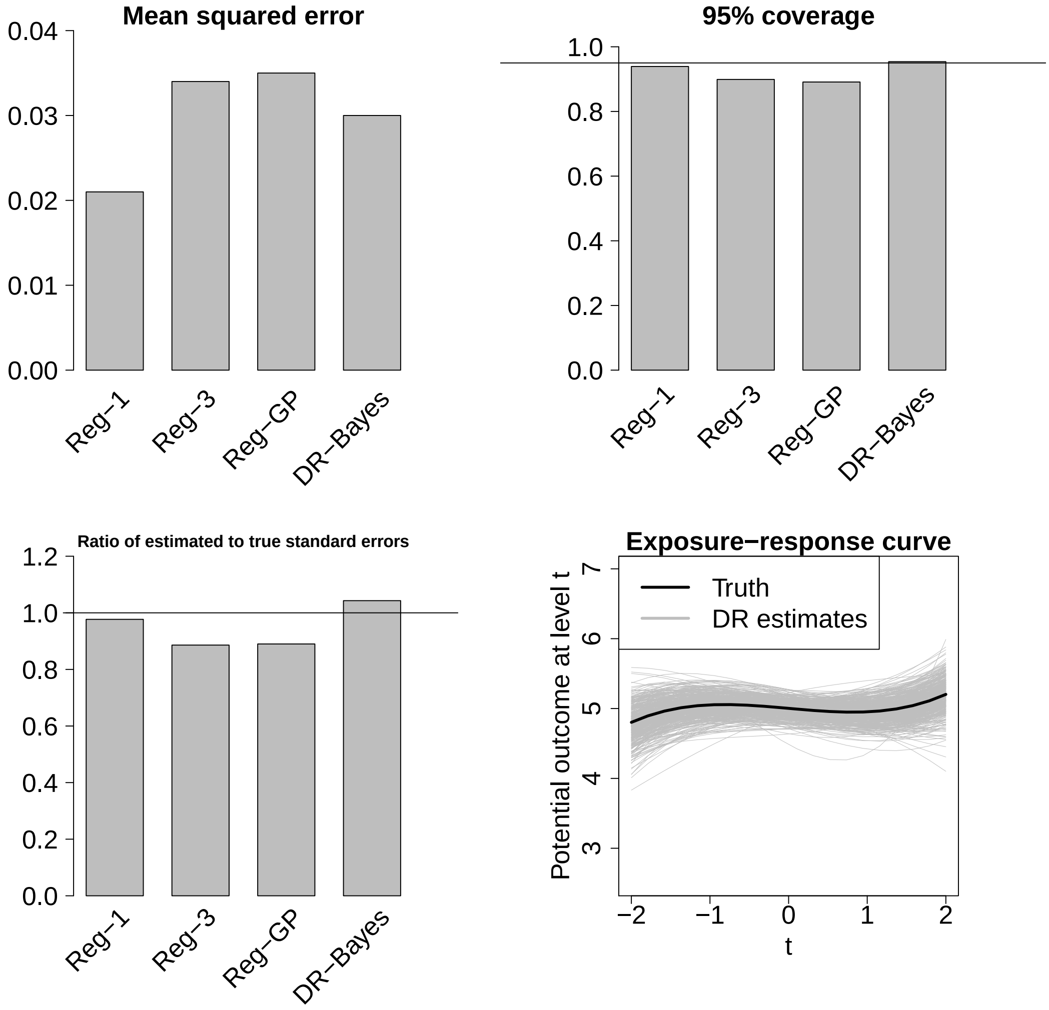

Figure A.6: Results from the fourth additional simulation scenario. The top left panel shows mean squared error, the top right panel shows the 95% interval coverages, the bottom left panel is the ratio of estimated to Monte Carlo standard errors, and the bottom right panel shows the true exposure-response curve and the simulated estimates of it.

F Simulation without additivity of models

Here we show that additivity of the treatment and outcome models is not required for our approach as long as there exists a Bayesian modeling approach that incorporates interactions among the variables considered. To highlight this issue, we will use the SoftBart approach of Linero & Yang (2018), which is an extension of the commonly used Bayesian Additive Regression Trees (BART, Chipman *et al.* (2010)) to allow for smoother functions and high-dimensional covariate spaces. As this is a tree-based approach, it naturally allows for interactions among the high-dimensional covariate space, thereby alleviating the additivity assumptions used in the simulations of the main manuscript. We will use SoftBart to estimate both the treatment and outcome models, and we will compare with linear models that assume additivity and impose sparsity via spike-and-slab priors. Both models will be embedded within our estimation and inferential strategy to combine posterior samples with resampling.

We will set $n = 150$ and $p = 200$, and generate covariates from independent, standard normal distributions. We will generate data from the following models:

$$\begin{aligned}\mu_i &= T_i + 0.8X_{1i} + 0.5X_{2i} + 0.8X_{1i}X_{2i} + 0.6X_{3i}X_{4i} \\ p_i &= \Phi(0.5X_{1i} - 0.4X_{2i} + 0.3X_{1i}X_{2i} + 0.4X_{3i}).\end{aligned}$$

The additive model will be misspecified for both the treatment and outcome models as there are nonzero interaction terms in each model, while the SoftBart approach allows for such interactions between the covariates. The results can be found in Table 1. We see that when the additive model is used for the treatment and outcome models that there is a substantial bias of 29%. When SoftBart is used for the two models, the bias drops to only 6% and the MSE is cut by more than half. In both cases, we end up with reasonable ratios for the estimated to true standard errors, and the SoftBart approach, due to it's lack of bias, achieves close to nominal coverage rates.

	Coverage	SE ratio	MSE	absolute bias
SoftBart	0.924	0.96	0.08	6
Additive	0.852	1.17	0.17	29

Table 1: Results from simulation with interactions among the covariates. SE ratio represents the ratio of the average estimated standard errors to the Monte Carlo standard errors of the estimator.

G Inference when models are misspecified

Here we will evaluate the extent to which our approach to inference will work if one or both of the models is misspecified. Clearly if both models are misspecified then the approach will likely not attain 95% coverage, but we can still evaluate the ratio of estimated to true standard errors in this setting. We will restrict to $p = 10$ in this setting as our question of interest is based more in model misspecification, which we don't want to conflate with the high-dimensional aspect of the procedure. Our data generating models are either linear or squared functions of the covariates, though we assume linear models for both the treatment and outcome in all cases. The ratio of the estimated to Monte Carlo standard errors can be seen in Figure A.7. We can see that in any of the four scenarios considered, our approach to inference either appropriately captures the uncertainty or is slightly conservative.

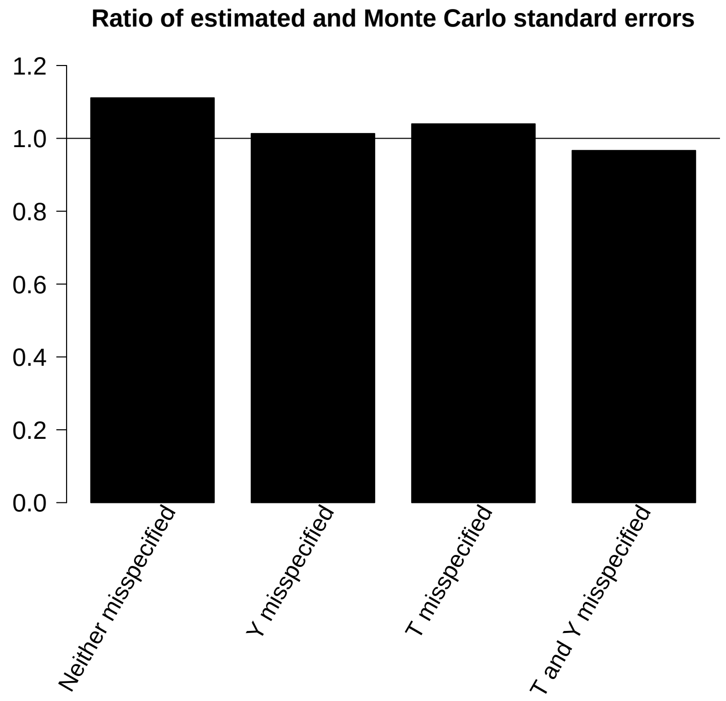

Figure A.7: Ratio of the estimated to Monte Carlo standard errors when different models are misspecified.

H Assessing impact of variance correction

In Section 3 of the manuscript we detailed our approach to variance estimation which entailed an approximation to the variance of interest, plus a correction that attempts to eliminate anti-conservative variances induced by error in the initial variance estimate. Here, we summarize the impact of this correction on estimation across all scenarios looked at in the paper. Figure A.8 shows the ratio of the estimated to true standard errors when only the approximate variance is used, as well as the ratio when the correction is applied on top of the initial variance estimate. We see that in some scenarios, such as the misspecification scenarios that were low-dimensional, the approximate variance is quite good and is close to achieving a ratio of 1. In other scenarios, such as the nonlinear simulation from the paper or the third appendix simulation, the approximation is nowhere near the correct variance, achieving a ratio below 0.4. Importantly though, the correction brings these variances up to a ratio close to, or above, 1 in nearly all scenarios. When the approximation is already doing quite well such as the misspecification scenarios, the correction only adds a small amount as the initial variance is already adequate. In scenarios where the approximation is poor, the variance correction dramatically increases the variance estimate to the correct levels.

I Implementation of existing approaches

First we will detail how we implement the competing approaches for the linear simulation study. For all estimators, linear models are assumed, in contrast to our approach that attempts to find the degree of nonlinearity required. The residual de-biasing approach is implemented using the `balanceHD` R package, which is available at `github.com/swager/balanceHD`. This R package estimates the treatment effect and provides confidence intervals, with which we perform inference. The TMLE approach is implemented using the `tmle` R package (Gruber & van der Laan, 2012). In the super learner library for TMLE, we only included SL.glmnet, as the true model is contained within this high-dimensional linear model. This package also gives both estimates and confidence intervals, which is how we performed inference. For the double machine learning approach, we used sample splitting with $K = 5$ splits and linear models based on the `glmnet` package (Friedman *et al.* , 2010) for both the treatment and outcome model. We tried fitting the treatment and outcome model with the output from lasso models, but also attempted fitting both the treatment and outcome models using post selection estimates of both models. For the post selection estimates, we ran an initial variable selection step

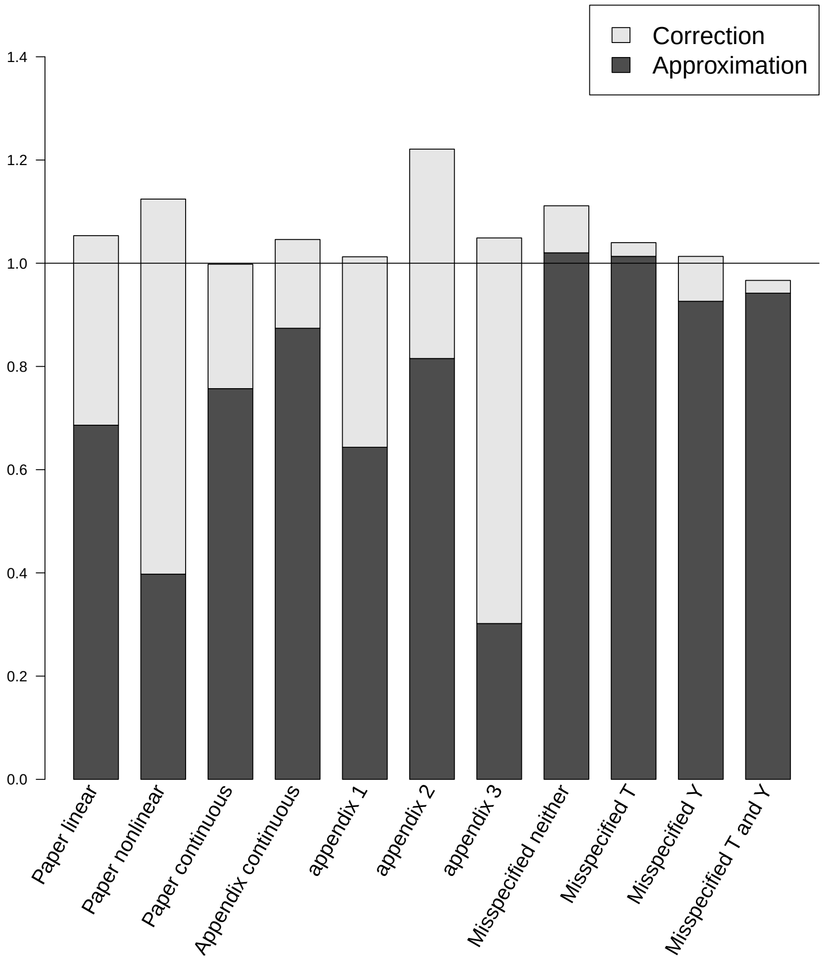

Figure A.8: Ratio of the estimated to Monte Carlo standard errors when only the initial variance is used and when the correction is added.

using glmnet, and then fit unpenalized models for both treatment and outcome. We found that the post selection estimators worked best for the double machine learning approach, and therefore we present them throughout the paper. We performed inference using the asymptotic standard errors derived in Chernozhukov *et al.* (2016). For the double post selection approach, we fit lasso models for both the treatment and outcome using glmnet with the tuning parameter chosen via cross validation. We then take the union of the selected variables from the two lasso regressions, and re-fit an unpenalized outcome regression model using the union of the selected covariates as confounders. Standard errors are calculated using the asymptotic standard errors derived in Belloni *et al.* (2014). We used a similar approach to implementing the lasso-DR approach of Farrell (2015). We fit lasso models using glmnet and cross validation and identified important covariates for both the treatment and outcome models. Then, given these respective set of covariates for each model, we fit unpenalized estimates of the outcome and treatment model. These are then used to estimate the doubly robust estimator described in our manuscript. Inference is again done using the asymptotic standard errors derived in Farrell (2015).

For the nonlinear simulation section, we restricted attention to TMLE and double machine learning as these were the most readily available to include nonlinear terms into the model. For both approaches, we ran an initial variable selection step for both the treatment and outcome model, and then identified important covariates as those that are in either the treatment or outcome model. For TMLE, we take these covariates and use the TMLE package with a super learner that includes generalized additive models. Whenever the model only required linearity, we only included GLM into the super learner to avoid using the overly flexible models when they are not necessary. For double machine learning, we take the reduced set of covariates and fit either linear models (when the truth is linear), or nonlinear additive models using spline representations of the covariates with 3 degrees of freedom. Otherwise, the implementation of the double machine learning approach is the same as for the linear case.

J Illustration of how asymptotics suffer in high-dimensions

In this section we will compare the finite sample variance of our proposed estimator and the one proposed by Farrell (2015). We will simulate data from sparse, linear models for both the treatment and outcome. We will apply our doubly robust estimator with Bayesian linear models and sparsity inducing priors as described in Section 4 of the manuscript. To build a doubly robust estimator, Farrell (2015) fit lasso (or group lasso) regressions (Tibshirani, 1996; Yuan & Lin, 2006) on both a treatment and outcome model to identify covariates that are associated with the treatment and outcome respectively. Then, they re-fit non-penalized estimators of the treatment and outcome models using only the covariates identified by the original lasso regressions. From these two regressions they can calculate the doubly robust estimator defined in Equation 1 where p_{ti} and m_{ti} are estimated using the non-penalized regression models. The authors derived some important theoretical results that demonstrate that their proposed double robust estimator is consistent and asymptotically normal. Our goal of this brief illustration is to elucidate why utilizing Bayesian methods to account for parameter uncertainty, which do not rely on asymptotics, can provide a more accurate assessment of the finite sample uncertainty, especially in high-dimensional scenarios. Here we focus on the estimator from Farrell (2015) as it uses the exact same doubly robust estimator, with the main difference coming in how inference is performed. As seen in the simulation study of the main manuscript, these ideas extend to other estimators rooted in asymptotics.

For each of the two doubly robust estimators, we will plot two lines. First, we will show the sampling distribution of the estimator as taken by the empirical distribution of the estimators across a large number of simulated datasets. Next, we will plot a normal density centered at the mean of the estimates across all datasets with a standard deviation that is the average estimated standard error across all datasets. If the estimated standard errors are correct, then this average standard error should be the same as the standard deviation of the sampling distribution and the two curves should

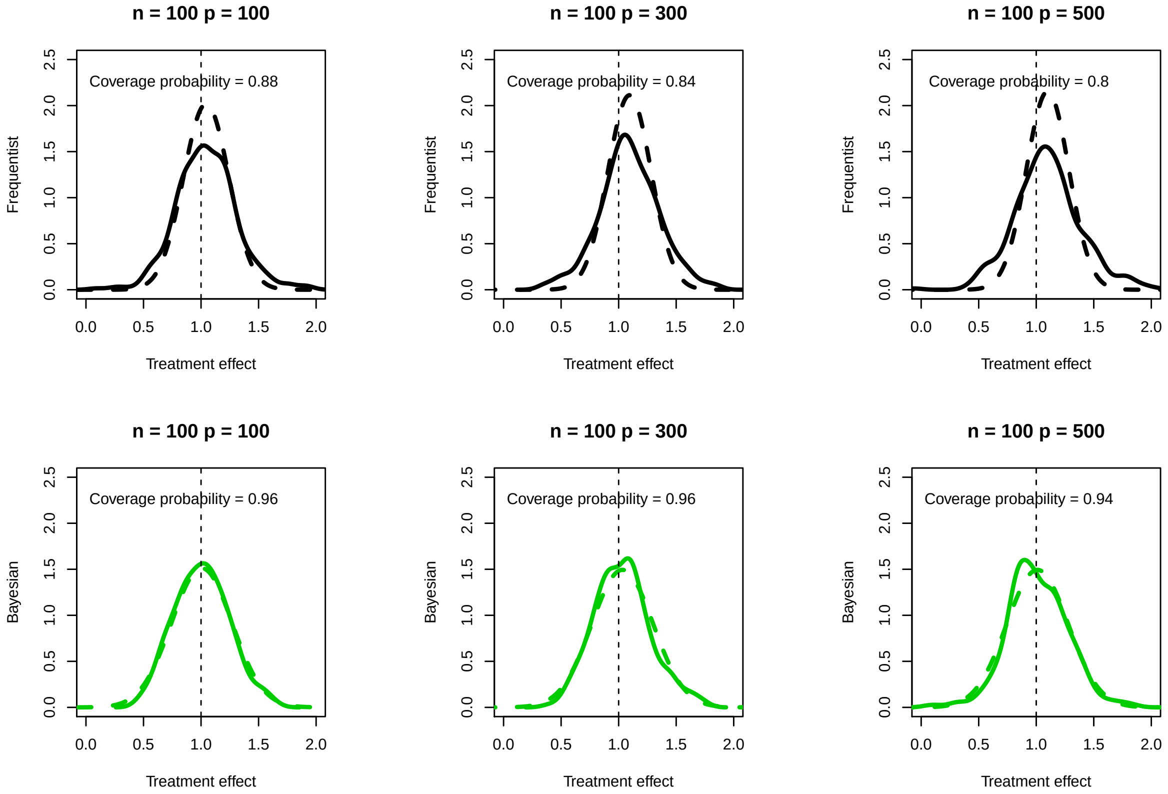

Figure A.9: Comparison of empirical and assumed distributions for the doubly robust estimator of Farrell (2015) and our approach from Section 3. The solid lines are the empirical sampling distributions, while the dashed lines are normal distributions with standard deviation equal to the average estimated standard deviation across the simulations

look similar. Figure A.9 shows the results for $n = 100$ and $p \in \{100, 300, 500\}$. The top row shows the results for the estimator based on asymptotic confidence intervals and the dashed line has much smaller tails than the solid line, indicating that the asymptotic distribution used for inference is not properly accounting for the uncertainty in the estimator. This phenomenon gets worse as p grows larger, and we see that the coverage probabilities decrease from 88% to 80%. Our approach to the same estimator, however, maintains the correct coverage probabilities for any dimension of the data, and the dashed and solid lines are very similar, showing that the uncertainty in the estimator is fully accounted for.

References

Antonelli, Joseph, Cefalu, Matthew, Palmer, Nathan, & Agniel, Denis. 2016. Doubly robust matching estimators for high dimensional confounding adjustment. *Biometrics*.

Antonelli, Joseph, Parmigiani, Giovanni, & Dominici, Francesca. 2017. High dimensional confounding adjustment using continuous spike and slab priors. *arXiv preprint arXiv:1704.07532*.

Athey, Susan, Imbens, Guido W, & Wager, Stefan. 2018. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **80**(4), 597–623.

Bang, Heejung, & Robins, James M. 2005. Doubly robust estimation in missing data and causal inference models. *Biometrics*, **61**(4), 962–973.

Belloni, Alexandre, Chernozhukov, Victor, & Hansen, Christian. 2014. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, **81**(2), 608–650.

Castillo, Ismaël, Schmidt-Hieber, Johannes, Van der Vaart, Aad, *et al.* . 2015. Bayesian linear regression with sparse priors. *The Annals of Statistics*, **43**(5), 1986–2018.

Chernozhukov, Victor, Chetverikov, Denis, Demirer, Mert, Duflo, Esther, Hansen, Christian, & Newey, Whitney K. 2016. *Double machine learning for treatment and causal parameters*. Tech. rept. cemmap working paper, Centre for Microdata Methods and Practice.

Chipman, Hugh A, George, Edward I, McCulloch, Robert E, *et al.* . 2010. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, **4**(1), 266–298.

Ertefaie, Ashkan, Asgharian, Masoud, & Stephens, David A. 2018. Variable selection in causal inference using a simultaneous penalization method. *Journal of Causal Inference*, **6**(1).

Farrell, Max H. 2015. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, **189**(1), 1–23.

Freedman, David. 1999. Wald Lecture: On the Bernstein-von Mises theorem with infinite-dimensional parameters. *The Annals of Statistics*, **27**(4), 1119–1141.

Friedman, Jerome, Hastie, Trevor, & Tibshirani, Robert. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, **33**(1), 1–22.

Gill, Richard D, & Robins, James M. 2001. Causal inference for complex longitudinal data: the continuous case. *Annals of Statistics*, 1785–1811.

Gruber, Susan, & van der Laan, Mark J. 2012. tmle: An R Package for Targeted Maximum Likelihood Estimation. *Journal of Statistical Software*, **51**(13), 1–35.

Hahn, P Richard, Carvalho, Carlos M, Puelz, David, He, Jingyu, *et al.* . 2018. Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Analysis*, **13**(1), 163–182.

Hill, Jennifer L. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, **20**(1), 217–240.

Hirano, Keisuke, & Imbens, Guido W. 2004. The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, **226164**, 73–84.

Kennedy, Edward H, Ma, Zongming, McHugh, Matthew D, & Small, Dylan S. 2017. Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **79**(4), 1229–1245.

Knaus, Michael C. 2018. A Double Machine Learning Approach to Estimate the Effects of Musical Practice on Student's Skills. *arXiv preprint arXiv:1805.10300*.

Linero, Antonio R, & Yang, Yun. 2018. Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **80**(5), 1087–1110.

Little, Roderick J, & Rubin, Donald B. 2000. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual review of public health*, **21**(1), 121–145.

Mitra, Robin, & Dunson, David. 2010. Two-level stochastic search variable selection in GLMs with missing predictors. *The international journal of biostatistics*, **6**(1).

Patel, Chirag J, & Ioannidis, John PA. 2014. Studying the elusive environment in large scale. *Jama*, **311**(21), 2173–2174.

Patel, Chirag J, Cullen, Mark R, Ioannidis, John PA, & Butte, Atul J. 2012. Systematic evaluation of environmental factors: persistent pollutants and nutrients correlated with serum lipid levels. *International journal of epidemiology*, **41**(3), 828–843.

Reich, Brian J, Storlie, Curtis B, & Bondell, Howard D. 2009. Variable selection in Bayesian smoothing spline ANOVA models: Application to deterministic computer codes. *Technometrics*, **51**(2), 110–120.

Robins, James M, & Ritov, Ya'acov. 1997. Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in medicine*, **16**(3), 285–319.

Rosenbaum, Paul R, & Rubin, Donald B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**(1), 41–55.

Saarela, Olli, Stephens, David A, Moodie, Erica EM, & Klein, Marina B. 2015. On Bayesian estimation of marginal structural models. *Biometrics*, **71**(2), 279–288.

Saarela, Olli, Belzile, Léo R, & Stephens, David A. 2016. A Bayesian view of doubly robust causal inference. *Biometrika*, **103**(3), 667–681.

Schnitzer, Mireille E, Lok, Judith J, & Bosch, Ronald J. 2015. Double robust and efficient estimation of a prognostic model for events in the presence of dependent censoring. *Biostatistics*, **17**(1), 165–177.

Shortreed, Susan M, & Ertefaie, Ashkan. 2017. Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*.

Tibshirani, Robert. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Van der Laan, Mark J, & Rose, Sherri. 2011. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media.

Van Der Laan, Mark J, & Rubin, Daniel. 2006. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, **2**(1).

Watanabe, Sumio. 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, **11**(Dec), 3571–3594.

Wild, Christopher Paul. 2005. Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiology Biomarkers & Prevention*, **14**(8), 1847–1850.

Wilson, Ander, Zigler, Corwin, Patel, Chirag, & Dominici, Francesca. 2018. Model-averaged confounder adjustment for estimating multivariate exposure effects with linear regression. *Biometrics*.

Yang, Yun, Tokdar, Surya T, *et al.* . 2015. Minimax-optimal nonparametric regression in high dimensions. *The Annals of Statistics*, **43**(2), 652–674.

Yoo, William Weimin, Ghosal, Subhashis, *et al.* . 2016. Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. *The Annals of Statistics*, **44**(3), 1069–1102.

Yuan, Ming, & Lin, Yi. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**(1), 49–67.