Combining parametric and nonparametric models to estimate treatment effects in observational studies

Daniel Daly-Grafstein | Paul Gustafson

Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada

Correspondence
Daniel Daly-Grafstein, Department of Statistics, University of British Columbia, 2207 Main Mall, Vancouver, BC V6T 1Z4, Canada. Email: daly-grafstein@stat.ubc.ca

Funding information
Natural Sciences and Engineering Research Council of Canada, Grant/Award Number: RGPIN-2019-03957

Abstract
Performing causal inference in observational studies requires we assume confounding variables are correctly adjusted for. In settings with few discrete-valued confounders, standard models can be employed. However, as the number of confounders increases these models become less feasible as there are fewer observations available for each unique combination of confounding variables. In this paper, we propose a new model for estimating treatment effects in observational studies that incorporates both parametric and nonparametric outcome models. By conceptually splitting the data, we can combine these models while maintaining a conjugate framework, allowing us to avoid the use of Markov chain Monte Carlo (MCMC) methods. Approximations using the central limit theorem and random sampling allow our method to be scaled to high-dimensional confounders. Through simulation studies we show our method can be competitive with benchmark models while maintaining efficient computation, and illustrate the method on a large epidemiological health survey.

KEYWORDS
Bayesian methods, causal inference, g-computation, nonparametric

1 INTRODUCTION

Estimating treatment effects in observational studies can be difficult when there are high-dimensional confounding variables present or in settings with longitudinal data where treatment effects vary over time. If we are interested in causal inference, we must assume our potential outcomes are conditionally independent of the treatment assignment given the available confounding variables. This assumption is equivalent to assuming that no unmeasured confounding variables affect the treatment assignment, and thus is it often necessary to include many covariates when estimating potential outcomes in hope of satisfying this assumption.

In settings with complex high-dimensional or longitudinal data, Bayesian approaches can be used to stabilize causal estimates in the face of sparse data. When working with longitudinal data, g-methods are particularly useful as they lead to consistent effect estimation by avoiding treatment-confounder feedback in the presence of time-varying confounding variables (Mansournia et al., 2017). There has been recent work showing that Bayesian inference based on the g-formula can improve causal effect estimates in small samples (Gustafson, 2015; Keil et al., 2017).

Bayesian versions of propensity-adjusted causal estimates have also been proposed in these settings (Saarela et al., 2015). However, in our case we choose to assume a priori the exposure assignment mechanism is independent of the marginal probabilities of the confounders and the outcomes. Thus, we do not include propensity scores in our model to avoid the risk of model feedback (Zigler et al., 2013; Saarela et al., 2016). One option for the outcome distribution is to use modern nonlinear regression
techniques, with Bayesian Additive Regression Trees (BART) being a popular choice (Hill, 2011; Hahn et al., 2020). A challenge with these methods is that if we are interested in estimating treatment effects from longitudinal studies then the number of models we need to fit grows linearly with the number of timepoints. It may not be feasible to use nonlinear methods which rely on Markov chain Monte Carlo (MCMC) sampling, either due to computational concerns, or practical issues with checking MCMC diagnostics for the large number of potential models needed.

To avoid MCMC estimation, we propose a model which conceptually splits the data, processing a portion parametrically and a portion nonparametrically, while maintaining a conjugate structure. This approach is closely related to the work by Huang et al. (2020) who defined “catalytic priors,” which effectively supplement observed data with a small amount of synthetic data generated from a simpler model. These priors shrink estimates from a flexible model toward a simpler model using synthetic data to stabilize posterior distributions. Our approach differs slightly in that we are conceptually splitting the data, rather than generating synthetic data, to supplement the observed data.

The remainder of this paper is organized as follows. In Section 2, we define our problem setting and notation. We define a fully saturated Bayes model and our new method—which we call a partially saturated model (PSM)—in Section 3, as well as giving approximations to our method that allow for scaling to many confounders. In Section 4, we measure its performance under a variety of simulation settings, as well as evaluate the model on data used in the 2019 Atlantic Causal Inference Conference (ACIC) prediction competition. In Section 5, we apply our method to longitudinal data from a large-scale health survey, assessing the causal effect of reducing smoking intensity on mortality for previously heavy-smokers. We conclude in Section 6 with a discussion of the results.

2 | PROBLEM SETTING AND NOTATION

Throughout the paper, we will focus on a scenario with a single binary outcome $Y$, single binary treatment $X$, and a set of $p$ binary confounders $C$, although our method can easily be applied to confounders with more than two categories. We are interested in estimating either the average treatment affect (ATE) $\Delta$ or average effect on the treated (ATT) $\Delta_t$. These measures correspond to the difference in expectation of the outcome $Y$ between hypothetical worlds where the treatment is set to either $X = 1$ or $X = 0$ for the entire population, or for the treated population, respectively. These measures can be defined using the potential outcomes framework as

$$\Delta = E\{Y^{(1)}\} - E\{Y^{(0)}\},$$

$$\Delta_t = E\{Y^{(1)}|X = 1\} - E\{Y^{(0)}|X = 1\},$$

where $Y^{(1)}$ denotes the outcome had the treatment been set to $X = 1$ and $Y^{(0)}$ denotes the outcome had the treatment been set to $X = 0$ (Hernán & Robins, 2020). Assuming all confounders are measured in our set $C$, such that $\{Y^{(1)}, Y^{(0)}\} \perp X|C$, we can reexpress $\Delta$ and $\Delta_t$ as

$$\Delta = E[E(Y|X = 1, C) - E(Y|X = 0, C)],$$

$$\Delta_t = E[\{E(Y|X = 1, C) - E(Y|X = 0, C)\}|X = 1].$$

For our setting with a binary outcome, treatment, and confounders, $\Delta$ in (3) can be computed using the $g$-formula by estimating $f(c)$ and $f(y|x, c)$, where $f(c) = Pr(C = c)$ and $f(y|x, c) = Pr(Y = y|X = x, C = c)$. Our goal in this paper is to construct prior distributions for $f(c)$ and $f(y|x, c)$ in a way that allows for posterior distributions to be represented in closed form, while still incorporating smoothness over $c$ to stabilize the estimates of $\Delta$ and $\Delta_t$. The following section illustrates our approach, with prior distributions for $f(y|x, c)$ that are related to the catalytic priors given in Huang et al. (2020).

3 | BAYESIAN SATURATED AND PSMs

3.1 | Saturated Bayes model

Consider the scenario described in Section 2. For posterior inference on $\Delta$ or $\Delta_t$, we need to estimate distributions of $(Y|X, C)$ and $C$. One option for the former is to use a Bayesian saturated binary regression model (BSAT) that estimates probabilities $\theta_{c,c} = Pr(Y = 1|X = x, C = c)$ for each unique combination of $C$ (Gustafson, 2015). Given a single treatment covariate $X$ and $p$ confounders $C$, $\theta$ is a vector of length $2^{p+1}$, thus requiring estimation of $2^{p+1}$ separate probabilities. We take conjugate beta$(\phi, \phi)$ priors independently for each element of $\theta$, resulting in a posterior independent beta distributions. Let $\gamma_c = P(C = c)$ and $\gamma_c = P(C = c|X = 1)$. We assume $\gamma_c$ (or $\gamma_c$) is independent of $\theta$, taking a conjugate Dir$(\epsilon, ..., \epsilon)$ prior for $\gamma$ (or $\gamma$). As a result, the joint posterior of $(\theta, \gamma)$ can be represented numerically using direct Monte Carlo (MC) sampling.

The BSAT model performs well in small settings with few confounders $C$ where model assumptions are not
necessary. However, for observational studies with more than a few confounders the model quickly becomes insufficient as many of the confounder combinations will have no sample data. For cells with no data the posterior \( \theta_{x,c} \) will equal the prior, in our case \( \text{beta}(\phi, \phi) \), resulting in a bias toward 0 when estimating \( \Delta \) or \( \Delta_t \). For example, Figure 1 shows a BSAT model with i.i.d. \( \text{beta}(\phi, \phi) \) priors in settings with various numbers of confounders and a fixed sample size \( n = 200 \). As the number of confounders increases for a fixed data size, the posterior distribution of \( \Delta_t \) becomes more strongly biased toward zero as a larger proportion of posterior distributions of \( \theta \) resemble the \( \text{beta}(\phi, \phi) \) prior.

### 3.2 Partially saturated model

Here, we introduce our PSM that augments the BSAT model with a prior for the outcome model \( f(Y|X,C) \) computed using a portion of the data. We again define a BSAT model with independent \( \text{beta}(\phi, \phi) \) priors for each element of \( \theta \) and a \( \text{Dir}(\epsilon, \ldots, \epsilon) \) prior for \( \gamma \). However, we augment the uninformative \( \text{beta}(\phi, \phi) \) on \( \theta_{x,c} \) by fitting a parametric model \( g(Y|X,C) \) to the data. The BSAT and parametric model \( g \) are analogous to the “working” and “simplified” models described in Huang et al. (2020). The parametric model \( g \) produces estimates for each of the \( 2^{p+1} \) elements of \( \theta \), which we denote as \( \hat{\theta}_g \). These smoothed estimates are incorporated into our beta prior by conversion into a number of pseudo-successes and pseudo-failures spread evenly for each covariate combination. The independent beta priors for each \( \theta_{x,c} \) become

\[
\theta_{x,c} \sim \text{beta} \left\{ \phi + \frac{bn}{2^{p+1}} \hat{\theta}_{x,c,g} + (1-b) n_{x,c} - y_{x,c}, \phi + \frac{bn}{2^{p+1}} (1 - \hat{\theta}_{x,c,g}) + (1-b) (n_{x,c} - y_{x,c}) \right\},
\]

where \( b \) is a tuning parameter that represents how smooth we believe \( \theta \) is relative to the smoothness implied by \( g \). While \( g \) is fit to the entire data set of size \( n \), \( b \) can be thought of as determining the proportion of the data which is processed parametrically, with \( b = 0 \) returning us to the BSAT prior and \( b = 1 \) giving us a model completely determined by \( g \) and the hyperparameter \( \phi \). Pseudo data points are distributed evenly between all \( 2^{p+1} \) cells, with a total of \( bn \) pseudo data points allocated. For each cell with \( X = x \) and covariates \( C = c \), we have a total of \( bn/2^{p+1} \) pseudo data points. These pseudo data points are split into pseudo successes and failures, where the proportion of each is determined by the parametric estimate \( \hat{\theta}_{x,c,g} \), resulting in the augmented prior given in (5).

Given the \( b \) portion of the data used to create the above priors, we take the yet unused \( (1 - b) \) portion of the data to update these parameters in a conjugate fashion. Specifically, the parameters in (5) are updated by \( (1 - b) \) times the number of successes and failures in each covariate combination in the data. The resulting posterior distribution is

\[
(\theta_{x,c}|Y,X,C) \sim \text{beta} \left\{ \phi + \frac{bn}{2^{p+1}} \hat{\theta}_{x,c,g} + (1-b) y_{x,c}, \phi + \frac{bn}{2^{p+1}} (1 - \hat{\theta}_{x,c,g}) + (1-b) (n_{x,c} - y_{x,c}) \right\},
\]

for each \( x \) and \( c \).
where \( y_{x,c} \) denotes the total number of successes with \( X = x \) and \( C = c \) and \( n_{x,c} \) denotes the corresponding total number of observations. This setup has the effect of borrowing strength in estimating elements of \( \hat{\theta}_{x,c} \) from the parametric model \( g \) where there is not a lot of data, while allowing the BSAT model to take over when estimating \( \hat{\theta}_{x,c} \) parameters where a lot of data is present.

Given we are interested in the ATT as defined in (4), under our assumption of no unmeasured confounding variables the posterior distribution of \( \Delta_t \) is estimated as

\[
\Delta_t = \sum_c \hat{\gamma}_c (\hat{\theta}_{1,c} - \hat{\theta}_{0,c}),
\]

where the constituent posterior distributions of \( \hat{\theta}_{1,c} \) and \( \hat{\theta}_{0,c} \) are independent. The conjugate structure of the PSM allows us to calculate the first two moments of (7) in closed form (Supporting Information S1).

While the PSM has a similar spirit to empirical Bayes, the same data are not used in both the prior and the likelihood. For example, given a sample size of \( n \) we can show that the total number of data points used as part of the data-driven prior and the likelihood equals the sample size. For each \( \hat{\theta}_{x,c} \) if we combine the pseudo data points from the partially parametric prior and the data from the likelihood scaled by \((1 - b)\) (while ignoring the beta \((\phi, \phi)\) hyperprior) and then sum over all \( \hat{\theta}_{x,c} \) we have

\[
\sum_{x,c} \left\{ \frac{bn}{2^{p+1}} \hat{\theta}_{x,c} + (1 - b)y_{x,c} + \frac{bn}{2^{p+1}}(1 - \hat{\theta}_{x,c}) + (1 - b)(n_{x,c} - y_{x,c}) \right\}
\]

\[
= \sum_{x,c} \left( \frac{bn}{2^{p+1}} \right) + (1 - b)n
\]

\[
= \frac{2^{p+1}bn}{2^{p+1}} + (1 - b)n = n.
\]

We can see that combining the pseudo successes and failures with the likelihood scaled by \((1 - b)\) gives us a total of \( n \) effective data points used to estimate the \( 2^{p+1} \hat{\theta}_{x,c} \) parameters. Scaling the prior and likelihood by the tuning parameter \( b \) and \((1 - b)\) ensures we are not overly confident in our posterior uncertainty by using the same data in both the prior and likelihood.

### 3.3 Approximating the PSM

While the conjugate structure of the PSM makes direct sampling of the posterior trivial, computational issues can still arise when the number of confounders becomes large. When estimating \( \Delta_t \) using the PSM we require sampling of \( 2^{p+1} \) independent beta distributions as given in (6). The number of beta distributions needed grows exponentially with \( p \), and likely starts to become infeasible above \( p = 20 \), with at least \( 2^{21} = 2,097,152 \) beta distributions to sample from and sum over. To allow for scaling to high-dimensional settings, we provide approximations to the PSM by either approximating part of the linear combination of beta distributions that make up the estimator \( \Delta_t \) in (7), or by sampling a subset of this linear combination. We discuss each of these approximation methods in turn.

#### 3.3.1 Approximating missing data cells with the central limit theorem (CLT)

Our goal is to approximate the posterior of \( \Delta_t \) as given in (7) when the number of confounders \( p \) makes it infeasible to sample all parameters. We can produce an approximation that bounds the number of beta distributions needed to sample from at the sample size \( n \) by treating all beta distributions from data rich cells as in Section 3.2, and using a single distribution to approximate the linear combination of betas for cells where no data are present. For a given sample size of \( n \), let \( M_1 \) represent the set of all covariate combinations \( c \) present in the data for either \( X = 1 \) or \( X = 0 \), and let \( M_0 \) be the set of \( c \) missing in the data. For covariate combinations with no data present, the posteriors in (6) will equal the priors given in (5). The terms in the sum of random variables \( \hat{\gamma}_c(\hat{\theta}_{1,c} - \hat{\theta}_{0,c}), c \in M_0 \) that we wish to approximate are not independent, with dependence induced by the correlation among components of the Dirichlet posterior \( \hat{\gamma}_c \). However, if the dependence between these random variables is small enough, and goes to zero as the number of confounders becomes large, then the CLT can be applied (Peligrad, 1986). See Supporting Information S2 for more details. Figure 2 compares the posterior distribution of \( \Delta_t \) computed using the full PSM and the approximation using the CLT for missing data cells. The approximation using the CLT results in a posterior distribution for \( \Delta_t \) similar to the full PSM.

#### 3.3.2 Randomly sampling missing data cells

If \( p \) is large enough eventually even summing over the cells to calculate the mean and variance as described previously becomes infeasible. In these cases, we propose an approximation where we take a random sample of those covariate combinations which are not manifested in the sample. Here, we take a random sample of \( c ' s \) in \( M_0 \), then calculate \( \Delta_t \) treating this random sample along with all cells \( c \in M_1 \) as if it were all possible combinations of \( c \). Specifically, we compute the posterior by taking an MC sample of all \( c \in M_1 \) as well as \( c \in R \) where \( R \subset M_0 \) is the set of randomly sampled covariate combinations with missing data. Under the full PSM model, we have a Dirichlet
distribution with \(2^p\) elements, whereas with our random sample approximation we have \(|M_1| + |R|\) elements. We know the total mass for missing cells in our Dirichlet prior is \(\varepsilon|M_0|\). Keeping \(\varepsilon\) the same as in the full PSM for \(\{c \in M_1\}\), we can upweight \(\varepsilon\) in our missing sampled cells to give equivalent probabilities with \(\varepsilon' = \varepsilon \frac{|M_0|}{R}, \forall c \in R\).

In addition, in the standard PSM we essentially have \(bn\) data points as part of the prior. We split these prior data points in each cell evenly, with \(\frac{bn}{2^p+1}\) data points in each cell. For example, in the cell with \(X = 1, C = c\) we have \(\frac{bn}{2^p+1} \hat{\delta}_{1,c,g}\) pseudo-successes and \(\frac{bn}{2^p+1}(1 - \hat{\delta}_{1,c,g})\) pseudo-failures as part of the PSM outcome model prior. However, in the case where we are only taking a random sample of missing cells, we still have \(\frac{|M_1|bn}{2^p+1}\) prior data points in cells with data, but now only have \(\frac{Rbn}{2^p+1}\) in missing cells instead of \(\frac{|M_0|bn}{2^p+1}\). We can scale this by \(\frac{|M_0|}{R}\) such that we have \(\frac{|M_1|bn}{2^p+1} \frac{Rbn}{2^p+1} = bn\).

To maintain the same expected value for each missing data cell the hyperparameter \(\phi\) also needs to be included in this upscaling. The posterior beta distribution for each randomly sampled missing data cell becomes

\[
(\theta_{x,c} | Y,X,C) \sim \text{beta}\left(\frac{|M_0|}{R} \left(\phi + \frac{bn}{2^p+1} \hat{\delta}_{x,c,g}\right), \frac{|M_0|}{R} \left(1 - \hat{\delta}_{x,c,g}\right)\right),
\]

(9)

We find the posterior of \(\Delta_t\) computed using the full PSM and this approximation taking a random sample of missing cells result in similar distributions (Figure 2).

4 | EMPIRICAL EVALUATIONS

In this section, we compare the PSM to various alternative models under a variety of simulation scenarios. Alternative models include the BSAT model, fully parametric model \(g\), and a BART estimator implemented using the \(\text{1bart}\) function from the \(R\) package \(\text{BART}\) with the default hyperparameter settings and 100 trees. The BSAT model and fully parametric model can be thought of as special cases of the PSM model with \(b = 0\) and \(b = 1\), respectively.
In all simulation cases to evaluate the models, we compare the root-mean-square error (RMSE) of the posterior estimator of Δ, using various sample sizes and number of confounders. In each case, we set $\phi = 100/2^{p+1}$ and $\epsilon = 100/2^p$ to maintain a constant effective prior strength independent of the number of confounders. We set $b$ to be the proportion of covariation combinations $c$ not present in the data, with a minimum of $b = 0.1$ and maximum of $b = 0.9$. See Supporting Information S3 for details on tuning and hyperparameter specification. In Section 4.1, we compare the RMSEs of the models using data generated with a purposely misspecified model $g$. In Section 4.2, we present results from dichotomized simulated data using data generating processes (DGPs) from the 2019 ACIC data challenge. In Section 4.3, we provide an extension to the PSM for continuous confounders and rerun results using data from the ACIC that has not been dichotomized.

### 4.1 Simulation studies

We are interested in evaluating the PSM in situations with heterogeneous treatment effects, strong confounding variables, and a parametric model $g$ that is a simplification of the true DGP, which we feel are common features in the applied settings of interest described in Section 1. Consider the problem setting described in Section 2. To recreate this applied setting we simulate data using the following DGP:

$$
\begin{align*}
\logit\{E(Y \mid X = 0, C)\} &= \beta_0 \\
+ \beta_1^T(C - \mu_1) + (\beta_2 + \lambda_1\mathbf{1}_p)^T(C - \mu_1),
\end{align*}
$$

(10)

$$
\begin{align*}
\logit\{E(Y \mid X = 1, C)\} &= \beta_0 + \lambda_0 \\
+ \beta_1^T(C - \mu_1) + (\beta_2 - \lambda_1\mathbf{1}_p)^T(C - \mu_1),
\end{align*}
$$

(11)

where $\mu_1 = E(C \mid X = 1)$. Here $\lambda_0$ controls the effect of $X$, $\beta_1$ controls the main effects of $C$, and the $(\beta_2 + \lambda_1\mathbf{1}_p)^T(C - \mu_1)$ and $(\beta_2 - \lambda_1\mathbf{1}_p)^T(C - \mu_1)$ terms induce two-way interactions between $X$ and $C$. To induce confounding effects between $X$ and $C$, we simulate equicorrelated binary variables $C$ via a thresholded multivariate normal distribution with all pairwise correlations between the latent normals set to $\rho_c$ (Leish et al., 1998). In all cases, we set the marginal probabilities of $\mu_1 = 0.5$. After simulating $C$, we generate probabilities for the treatment $X$ by $\logit\{E(X \mid C)\} = \alpha^T C$.

By including the $\lambda_0$ and $\lambda_1$ terms in (9) and (10) we can specify a DGP with a given $\Delta$ value. In addition, if we choose $g(\cdot)$ to be a logistic regression with only main effects terms for $X$ and $C$ (rather than both main effects and two-way interactions), then the $\lambda_1$ parameter can be tuned to give a specific bias level for $g(\cdot)$. For some prespecified $p$, $\Delta_1, \beta_0, \beta_1, \beta_2, \rho_c$, and $\omega$ we search for $\lambda_0$ and $\lambda_1$ values that result in a specific $\Delta$ and bias in $g(\cdot)$, which we will refer to as the main effects bias (MEB). Table 1 gives examples of how these parameters are generated. We consider six different DGPs, using 4, 8, and 12 confounders $C$, MEB set to 0.1 or $-0.1$, and $\Delta_1 = 0.3$. For each, we simulate data using sample sizes ranging from $n = 100$ to $n = 3000$ and compare the RMSE of the joint posterior of $(\theta, \phi)$ using the BSAT, PSM, $g(\cdot)$, and BART estimators. We set $g(\cdot)$ to be a logistic regression with main effects as $\logit\{g(Y \mid X, C)\} = \alpha_0 + \alpha_1X + \phi C$.

Figure 3 shows the RMSE over various sample sizes for each DGP. We can see that as the number of confounders increases, the PSM starts to outperform the BSAT model due to the larger proportion of covariate combinations not present in the data. Because we have set the tuning parameter $b$ to be the proportion of cells without data, as the sample size increases the proportion of data $b$ used as part of the parametric model prior decreases and the PSM converges toward the BSAT model. As the number of confounders increases, the proportion of covariate combinations missing in the sample remains high over all sample sizes and the PSM follows the purely parametric estimator more closely.

There are significant differences in the performance of the PSM depending on whether the MEB is set to 0.1 or $-0.1$. In this scenario, we have specified a true $\Delta_1$ value of 0.3, while utilizing beta($\phi, \phi$), $\phi = 100/2^{p+1}$ priors for our BSAT model. At low sample sizes, the priors bias the BSAT model toward 0, creating a negative bias for the estimator. When the MEB is set greater than 0, the positive bias of the parametric model cancels some of the negative bias of the BSAT model when combined in the PSM estimator, resulting in good performance across all sample sizes. When the MEB is less than 0, both the biases of the parametric and BSAT models are in the same direction and this canceling does not occur. However, we still expect the PSM to be an improvement in many of these cases. Without data, the posterior of the BSAT model for cells without data...
defaults to the $\text{beta}(100/2^{p+1}, 100/2^{p+1})$ prior, thus information shared from the parametric model should improve estimation of these $\theta_{x,c}$ parameters even in these scenarios.

### 4.2 ACIC 2019 data challenge

The ACIC data challenge is an annual competition where participants are invited to develop methods for estimating treatment effects from synthetic data sets. The 2019 competition had participants submit estimates for the ATE of a binary treatment on a binary or continuous outcome, as well as a 95% confidence interval. Data sets were split into low-dimensional and high-dimensional, with 3200 separate data sets in each track. Within each track, 100 data sets were generated from 32 DGPs. Methods were evaluated in terms of RMSE of the ATE estimates as well as confidence interval coverage.

Of these 32 DGPs, we chose four which involved binary treatments and responses. These DGPs were generated in two pairs of two, the first based on credit card default data from Yeh and Lien (2009), and the second created using email spam data from Blake et al. (1998). Because our methods are best suited for handling discrete covariates, we dichotomized each continuous covariate by thresholding at its median value and recalculated true ATEs based on the publicly available simulation code. For each DGP, we generate 100 data sets of size $n = 500$. We compare the bias and RMSE of the BSAT, fully parametric main-effects logistic regression, standard BART, and PSM approximated using random sampling. For all methods, we calculate the ATE using a linear combination weighted with a Dirichlet distribution as described in Section 3.1. We set $b = 0.9$ and $\phi = 100/2^{p+1}, \epsilon = 100/2^p$ with $p = 22$. Table 2 shows the results of each method averaged over 100 simulated data sets for each DGP. While the BART and parametric models fit using MCMC methods perform best overall, relative to the BSAT model the PSM stabilizes inference and reduces RMSE, performing competitively with the BART and parametric models in some cases.

### 4.3 Extension for continuous data

Here, we provide a method for using the PSM on data with continuous confounders without resorting to full discretization. The primary benefit of the PSM is the fast computation afforded by the conjugate structure of the discrete response surface. While we cannot modify this portion, we can fit the parametric portion of the model $g$ using continuous data. Once fit, we can then discretize the data and proceed with the PSM procedure as usual. To generate pseudo-successes and pseudo-failures from

---

**FIGURE 3** RMSE of the BSAT, PSM, parametric, and BART models estimating $\Delta_t$ with 4, 8, and 12 confounding variables $C$. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.
TABLE 2 Results for the simulated ACIC data. Bias and RMSE are averaged over 100 data sets for each of the four DGPs. DGPs 1 and 2 are created using the credit card data set of Yeh and Lien (2009) while 3 and 4 are created using the spam data set of Blake et al. (1998). True ATEs for the “binary” (“continuous”) columns are calculated from DGPs on dichotomized (nondichotomized) data from the ACIC. The PSM-cont-2 method corresponds to a PSM where \( g \) is calculated on the continuous data, then continuous columns dichotomized, PSM-cont-5 continuous columns are split into five categories.

| DGP | Method      | Binary Bias | RMSE | Continuous Bias | RMSE |
|-----|-------------|-------------|------|-----------------|------|
| 1   | BSAT        | 0.200       | 0.201| -0.043          | 0.045|
| 1   | BART        | 0.025       | 0.045| 0.105           | 0.115|
| 1   | Parametric  | -0.002      | 0.032| 0.104           | 0.113|
| 1   | PSM         | 0.059       | 0.066| 0.101           | 0.115|
| 1   | PSM-cont-2  |             |      | 0.116           | 0.127|
| 1   | PSM-cont-5  |             |      | 0.063           | 0.078|
| 2   | BSAT        | -0.142      | 0.144| -0.088          | 0.090|
| 2   | BART        | -0.013      | 0.036| 0.052           | 0.070|
| 2   | Parametric  | -0.003      | 0.033| 0.072           | 0.087|
| 2   | PSM         | -0.044      | 0.053| 0.057           | 0.074|
| 2   | PSM-cont-2  |             |      | 0.093           | 0.106|
| 2   | PSM-cont-5  |             |      | 0.005           | 0.048|
| 3   | BSAT        | -0.038      | 0.047| -0.110          | 0.112|
| 3   | BART        | -0.011      | 0.051| 0.015           | 0.050|
| 3   | Parametric  | -0.010      | 0.050| 0.008           | 0.047|
| 3   | PSM         | -0.017      | 0.045| 0.016           | 0.046|
| 3   | PSM-cont-2  |             |      | 0.026           | 0.055|
| 3   | PSM-cont-5  |             |      | -0.005          | 0.044|
| 4   | BSAT        | -0.051      | 0.058| -0.054          | 0.056|
| 4   | BART        | -0.005      | 0.038| -0.002          | 0.032|
| 4   | Parametric  | -0.004      | 0.036| 0.006           | 0.033|
| 4   | PSM         | -0.017      | 0.039| 0.001           | 0.031|
| 4   | PSM-cont-2  |             |      | 0.030           | 0.051|
| 4   | PSM-cont-5  |             |      | -0.018          | 0.038|

We fit \( g \) on continuous data, we index each discretized bin at a midpoint. For example, if we dichotomize a continuous confounder into two categories at the median, we index \( g \) at the midpoint of these two bins—the first and third quartiles of the dichotomized confounder.

We repeat the above study using the ACIC DGPs, except this time we use the true ATEs generated using continuous confounders where applicable. We refit the BSAT and PSM models as previously by first dichotomizing the confounders, but we also include versions of the PSM where \( g \) is fit on the nondichotomized data (Table 2). We include a PSM with dichotomized confounders (PSM-cont-2), and confounders split into five categories (PSM-cont-5) after fitting \( g \) on the continuous data. For PSM-cont-5, as well as the BART and parametric models, we use the sample ATE as an approximation for the population ATE. We can see from Table 2 that this extension to the PSM allows it to maintain a reasonable level of performance even when working with continuous data.

5 | EXAMPLE: NHANES I EPIDEMIOLOGIC FOLLOW-UP STUDY (NHEFS)

We illustrate the PSM on data from the NHEFS, part of the National Health and Nutrition Examination Survey (NHANES I). This study consists of a broad set of clinical, nutritional, and behavioral data, with a cohort aged between 25 and 74 initially interviewed in 1971–1975, and then again in a series of follow-up studies in 1982–1984, 1986, 1987, and 1992. We analyze a subset of these data, taking individuals aged 25–50 interviewed in 1971–1975, and followed-up in 1982–1984. Our goal is to estimate the effect of reducing smoking on the probability of mortality. We
take $X_{1i}$ to be a binary exposure denoting whether individuals smoked 20 or more cigarettes a day in 1971–1975, $X_{2i}$ an binary variable indicating whether patients smoked at least 10 cigarettes a day in 1982–1984, and the outcome $Y_i$ to be vital status (dead/alive) in the year 1992. Our treatment effect of interest is the effect of smokers reducing their smoking at follow-up, that is, $\Delta = E(Y | X_1 = 1, X_2 = 0, C_1, C_2) - E(Y | X_1 = 0, C_1, C_2)$. We control for 12 categorical confounders $C_1$ and 10 categorical confounders $C_2$ at each timepoint, including age, weight, sex, education level, and income, and comorbidities such as diabetes, high blood pressure, heart disease, asthma, and level of alcohol consumption. We discretize all continuous confounders. Due to the potential of treatment-confounder feedback, we estimate the treatment effect as

$$\Delta = \sum_{C_1} \sum_{C_2} (E(Y | X_1 = 1, X_2 = 1, C_1, C_2) - E(Y | X_1 = 1, X_2 = 0, C_1, C_2)) \cdot P(C_1, C_2, X_1 = 1) \cdot P(C_1, C_2).$$

We treat $P(C_1)$ and $E(Y | X_1, X_2, C_1, C_2)$ as described in Section 3 using the PSM, with separate beta distributions sampled over all combinations of $(X_1, X_2, C_1, C_2)$. For $P(C_2 | X_1, C_1)$, we use the same PSM machinery, fitting separate Dirichlet distributions over $C_2$ for each combination of $(X_1, C_1)$. We first fit a main effects logistic regression for each element of $C_2$ as the parametric portion $g_{C_2}$, scaling the outcome probabilities such that the sum of possible combinations of $C_2$ equals 1. We then take the total number of data points for a given combination of $(C_1, X_1)$, and spread a portion $b_{C_2}$ over all cells of the Dirichlet prior, taking the remaining $(1 - b_{C_2})$ portion of the data to update the prior in conjugate fashion. We take a random sample of missing combinations of $C_1$ and $C_2$ as in Section 3.3.2 to make sampling and summing (11) feasible. Fitting the PSM we estimate $\Delta$ to be 0.024, with a 90% credible interval of (0.012, 0.035), suggesting that decreasing smoking from over one pack a day to less than half a pack per day reduces the relative risk of mortality by 8.8% (4.4%, 12.5%).

6 | DISCUSSION

In this paper, we proposed a new estimator for treatment effects based on a nonparametric model augmented with a data-driven parametric prior, with suitable adjustment to avoid double-use of the same data. Using a Bayesian g-computation framework, our PSM takes a portion of the data to inform prior distributions for the treatment model via parametric modeling, allowing us to incorporate smoothness while maintaining a Bayesian framework. This setup allows us to avoid the use of MCMC computation for our posterior distribution by defining conjugate priors, giving the potential for this model to be used for causal inference in longitudinal settings with high-dimensional confounders and many timepoints. A limitation of this approach is that it is difficult to work with continuous data. While we provide one option to partially remedy this in Section 4.3, this approach still requires we discretize continuous confounders and index the parametric model $g$ at specific discrete points. The PSM is likely to be most useful when working with discrete treatments, outcomes, and confounders, or with a few continuous confounders than can be discretized without large amounts of information loss.

We illustrated the proposed model through a number of simulation studies as well as on modified data from the 2019 ACIC competition. Our simulation protocol described in Section 4.1 allows for exact specifications of treatment effects and parametric model bias and flexibility in specifying both the strength of confounding and complexity of the relationship between the outcome and covariates. However, as with any simulation study our presented results only encompass a small fraction of possible scenarios where this model can be used. For example, in all cases in both our simulations and in the ACIC data the true treatment effects were moderately large. In addition, we exclusively used beta priors on the outcome distribution that biased estimates toward a null treatment effect. This resulted in posteriors that, for three of four binary DGPs, were systematically biased toward 0. The effect is most notable when looking at the coverage of nominal 95% credible intervals of our posteriors when modeling data from the ACIC. Coverage for the BART models were between 0.87 and 0.93, for the purely parametric method between 0.86 and 0.94, the BSAT model between 0.80 and 0.93, and for the PSM between 0.54 and 0.97 for the four DGPs examined in Section 4.2. It may be the case that the hyperpriors in the PSM based on the hyperparameters $\epsilon$ and $\phi$ are contributing to the low coverage in some cases (Supporting Information S4).

Although we have focused primarily on applications in causal inference settings, this method is simply a Bayesian regression model that can be used for any regression problem involving discrete variables. It may be useful whenever a simpler Bayesian approach is desired that does not require implementation of potentially complex MCMC methods, while still allowing some of the flexibility of nonparametric estimation as well as the incorporation of prior information. The PSM also has a benefit of giving treatment effect estimates (or approximations) for all possible covariate combinations $C$. By explicitly modeling over all combinations of $C$, we may be better able to capture the uncertainty present in the entire population. This may be useful any time marginal effects are of interest,
rather than effects conditional on the observed covariates in the sample.

**ACKNOWLEDGEMENTS**

Daly-Grafstein and Gustafson were supported by NSERC grant RGPIN-2019-03957.

**DATA AVAILABILITY STATEMENT**

The data that support the findings in this paper are publicly available from the National Health and Nutrition Examination Survey Epidemiologic Follow-up Study at https://wwwn.cdc.gov/nchs/nhanes/nhefs.

**ORCID**

Daniel Daly-Grafstein https://orcid.org/0000-0002-2331-5991

Paul Gustafson https://orcid.org/0000-0002-2375-5006

**REFERENCES**

Blake, C., Keogh, E. & Merz, C.J. (1998) UCI Repository of machine learning databases. Department of Information and Computer Science, University of California, Irvine.

Gustafson, P. (2015) Discussion of “On Bayesian estimation of marginal structural models”. *Biometrics*, 71, 291–293.

Hahn, P.R, Murray, J.S. & Carvalho, C.M. (2020) Bayesian regression tree models for causal inference: regularization, confounding and heterogeneous effects. *International Society for Bayesian Analysis*, 15, 965–1056.

Hernán, M.A. & Robins, J.M. (2020) Causal inference: What if. Boca Raton: Chapman & Hall/CRC.

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

Huang, D., Stein, N., Rubin, D.B. & Kour, S.C. (2020) Catalytic prior distributions with application to generalized linear models. *Proceedings of the National Academy of Sciences*, 117, 12004–12010.

Keil, A.P., Daza, E.J., Engel, S.M., Buckley, J.P. & Edwards, J.K. (2017) A Bayesian approach to the g-formula. *Statistical Methods in Medical Research*, 27, 3183–3204.

Leisch, F., Weingessel, A. & Hornik, K. (1998) On the generation of correlated artificial binary data. Working Paper Series, SFB Adaptive Information Systems and Modelling in Economics and Management Science, Vienna University of Economics.

Mansournia, M., Etminan, M., Danaei, G., Kaufman, J. & Collins, G. (2017) Handling time varying confounding in observational research. *BMJ: British Medical Journal*, 359, j4587.

[dataset] National Center for Health Statistics: Plan and Operation of the Health and Nutrition Examination Survey United States 1971–1973. Vital and Health Statistics Series 1, No. 10a. Washington DC: Govt Printing Office, 1973. Available from: https://wwwn.cdc.gov/nchs/nhanes/nhefs/

Peligrad, M. (1986) Recent advances in the central limit theorem and its weak invariance principle for mixing sequences of random variables (a survey). In *Dependency in Probability and Statistics: A Survey of Recent Results*, edited by E. Eberlein and M.S. Taqqu (Springer 1986), 193–223.

Saarela, O., Belzile, L.R. & Stephens, D.A. (2016) A Bayesian view of doubly robust causal inference. *Biometrika*, 103, 667–681.

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

Yeh, I.C. & Lien, C.H. (2009) The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36, 2473–2480.

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

**SUPPORTING INFORMATION**

Web appendices referenced in Sections 3, 4, and 6 are available with this paper at the Biometrics website on Wiley Online Library. Additionally, code to run the simulations and analyses in Sections 3.1, 3.3, 4.1, 4.2, 4.3 and 5 is available. The original code for the ACIC 2019 Data Challenge is available at https://sites.google.com/view/acic2019datachallenge/data-challenge.

Data SI

**How to cite this article:** Daly-Grafstein, D. & Gustafson, P. (2023) Combining parametric and nonparametric models to estimate treatment effects in observational studies. *Biometrics*, 79, 1986–1995. https://doi.org/10.1111/biom.13776