Bayesian Causal Inference Under Conditional Ignorability*

Siddhartha Chib†

July 2017

Abstract

In this paper we describe a Bayesian approach for finding the causal effect with observational data under the assumption that the binary treatment variable is conditionally ignorable. In our approach, the potential outcome distributions are modeled directly through spline-based (basis function) regression techniques and the relevant potential outcome distributions are estimated separately from the data on the control and treated subjects. An important facet of the approach is that the average treatment effect (ATE) is calculated from a predictive perspective (post estimation) in which the missing outcomes of the control subjects are predicted from the model of the treated subjects while the missing outcomes of the treated subjects are predicted from the model of the control subjects. We show that this strategy works, even with covariate imbalance, if the knots in the basis expansions are chosen in a specific way from the combined covariate values of both the control and treated subjects. We illustrate the performance of our approach against frequentist matching-type estimators using both simulated and real data.

Key words: Average treatment effect; cubic spline; Markov chain Monte Carlo; marginal likelihood; observational data; overlap problem; semiparametric Bayesian inference.

1 Introduction

In the context of observational (non-experimental) data, suppose that $x \in \{0, 1\}$ is a binary treatment variable and let $\mathbf{z}$ denote a k-dimensional vector of observed pre-treatment covariates or confounder (control) variables. Suppose that the treatment intake mechanism is described by the probability model $\Pr(x = 1|\mathbf{z}) = e(\mathbf{z})$ and that this probability (called the propensity score) satisfies the overlap condition $0 < e(\mathbf{z}) < 1$, for all $\mathbf{z}$. Also let y_0 and y_1 denote the potential outcomes, and suppose that the treatment is conditionally ignorable, i.e., independent of the potential outcomes given the confounders. Then, the ATE, given by the difference $\mathbb{E}(y_1) - \mathbb{E}(y_0)$,

*Thanks to Dr. Sandor Kovacs of the Washington University School of Medicine for explaining the right heart catherization procedure, and to participants at seminars at Yale University (April 2010) and University of Melbourne (2014). This paper is dedicated to the memory of Edward Greenberg, friend and collaborator, whose explorations and development of the Bayesian viewpoint over several decades have left a rich legacy.

†Olin Business School, Washington University in St. Louis, St. Louis MO 63130; chib@wustl.edu

where the expectations are with respect to the marginal distribution of the potential outcomes, is identified. In this paper, we are interested in developing a Bayesian approach for estimating the ATE under the overlap and conditional ignorability assumptions.

The ATE is commonly found by frequentist matching methods, such as the method of propensity score matching (Rosenbaum and Rubin, 1983). In this method, the propensity score is estimated by a flexible logit or probit model, and then two individuals with the same propensity score, one treated and one control, are matched. The difference in outcomes of such matched subjects is the average treatment effect (ATE) conditioned on the propensity score. Averaging these differences across matched subjects leads to an estimate of the ATE.

It is not possible to develop a Bayesian approach that strictly parallels the frequentist propensity score matching method. This is because propensity score matching is an algorithm that cannot be described in likelihood terms. A more fundamental issue is that, under conditional ignorability, the treatment is independent of the outcomes and thus plays no role in inferences about the potential outcome distributions. Nonetheless, attempts at formulating causal inferences based on Bayesian versions of propensity scores are described in, for example, Hoshino (2008), An (2010), Kaplan and Chen (2012) and Zigler et al. (2013). In this paper we pursue an alternative approach from the Bayesian side which is to model the potential outcome distributions directly and to estimate the y_0 distribution from the control subjects and the y_1 distribution from the treated subjects. In this modeling we use spline-based (basis function) regression techniques to non-parametrically model the distributions of y_0 and y_1 given the confounders. We do not need to estimate the unidentified joint distribution of (y_0, y_1) for each subject, as this joint distribution is not required, following Chib (2007). We then estimate the ATE by predicting y_1 for the control subjects from the model of y_1 estimated from the treated subjects, and by predicting y_0 for the treated subjects from the model of y_0 estimated from the control subjects. We show that this strategy works (even when the distribution of the confounders is quite different for the control and treated subjects - the problem of covariate

imbalance) if the knots in the basis expansions are chosen in a specific way from the combined covariate values of both the control and treated, even while only the data on the control subjects is used to estimate the y_0 model and only the data on the treated subjects is used to estimate the y_1 model. When there is no overlap in the covariate distributions across the treatment and control subjects, our approach would fail, as would those based on matching methods, but as long as the overlap condition holds, our approach for selecting knots leads to accurate estimates of the ATE, as we show below. Our approach produces the posterior distribution of the ATE, marginalized over parameter and model uncertainties.

Our approach assumes that the set of covariates $\mathbf{z}$ that produce conditional ignorability of the treatment are known in advance. We do allow the set of available confounders to exceed those in $\mathbf{z}$. In that case, we judge the relevance of those additional confounders by comparing the marginal likelihoods of the models with and without those additional confounders. We calculate these marginal likelihoods by the method of Chib (1995).

Non-parametric modeling of the potential outcomes has also been considered by Hill (2011) but from a Bayesian CART perspective. McCandless et al. (2009) considers a quite different Bayesian approach for outcome modeling by letting the outcomes depend on the propensity score. This requires the estimation of both the propensity score and outcome models and leads to a complex estimation procedure. Joint modeling of outcome and treatment models with a particular focus on the question of confounder choice is discussed in Wang et al. (2012) while Saarela et al. (2016) provide an approach in which both models are estimated with the aim of achieving robustness to confounder misspecification. Our approach in this paper is in some sense complementary to these approaches because it explores the Bayesian analysis under the assumption that conditional ignorability holds for the given set of confounders. An important difference between Hill (2011) and our work is that we stress the issue of covariate imbalance and propose a knot selection procedure to address it, but Hill does not discuss how the CART approach would perform with significant covariate imbalance, as in one of the problems we

consider.

The rest of the paper is organized as follows. In Section 2 we present the approach for outcome modeling along with our method for selecting knots for the cubic spline basis matrices. The estimation of the models from the control and treated subject data is also described in this section followed by our approach for calculating the posterior distribution of the ATE in Section 3. The application of the methodology is first illustrated in Section 4 with an example that has considerable covariate imbalance and then with real data in Section 5. Section 6 contains our conclusions. Appendix A explains the construction of the basis matrix, and Appendix B presents details of our prior distribution.

2 Approach: outcome modeling and estimation

Let $p_0(y|\mathbf{z})$ and $p_1(y|\mathbf{z})$ denote the conditional distributions of y_0 and y_1 given the confounders. These do not depend on x because of the conditional ignorability assumption. We model these distributions in a semi-parametric way by combining a parametric student-t distribution for $p_j(.)$ with additive non-parametric modeling of the covariate affects. In addition, suppose that the vector of confounders is split into two components, $\mathbf{z} = (\mathbf{v}, w_1, ..., w_q)$, where $\mathbf{v} : k_v \times 1$ are categorical predictors including the intercept, and $\{w_r\}$ are continuous predictors with non-linear effects on the outcome. We suppose that the outcome distribution in the $x = 0$ state is

$$p_0(y_0|\mathbf{z}) = t_{\nu_0}\left(y_0|\mathbf{v}'\boldsymbol{\beta}_{00} + g_{01}(w_1) + \cdots + g_{0q}(w_q), \sigma_0^2\right) \tag{2.1}$$

and in the $x = 1$ state is

$$p_1(y_1|\mathbf{z}) = t_{\nu_1}\left(y_1|\mathbf{v}'\boldsymbol{\beta}_{10} + g_{11}(w_1) + \cdots + g_{1q}(w_q), \sigma_1^2\right) \tag{2.2}$$

where, for $j = 0, 1$, t_{ν_j} is the student-t density with $\nu_j > 2$ degrees of freedom, $g_{jr}(\cdot)$ is an unknown smooth function of w_r for $r \leq q$, and σ_j^2 is the dispersion. The following remarks are in order. The preceding specify the marginal distributions of the potential outcomes. The unidentified joint distribution of (y_0, y_1) is not needed, following Chib (2007), because the

missing counterfactuals can be simply integrated out. Second, this modeling of the marginal distributions is saturated in the sense that the mean, dispersion and degrees of freedom are allowed to differ. Finally, the student-t assumption is important in practice. It provides substantially improved models, especially when the mean function, as above, is modeled non-parametrically. Further generality can be achieved, if desired, by putting a non-parametric prior (such as the Dirichlet process) on these distributions.

2.1 Sample data

Suppose we have sample data $(x_i, y_i, \mathbf{v}_i', \mathbf{w}_i')$ on n independently distributed subjects $(i = 1, \ldots, n)$, where $y_i = x_i y_{1i} + (1 - x_i) y_{0i}$, organized so that the first n_0 observations are those for the controls $(x_i = 0)$ and the next $n_1 = n - n_0$ are for the treated $(x_i = 1)$:

$$x_i = 0,\ y_{0i},\ y_{1i}^*,\ y_i = y_{0i},\ \mathbf{v}_i', \mathbf{w}_i',\ i = 1, \ldots, n_0, \tag{2.3}$$

$$x_i = 1,\ y_{0i}^*,\ y_{1i},\ y_i = y_{1i},\ \mathbf{v}_i', \mathbf{w}_i',\ i = n_0 + 1, \ldots, n, \tag{2.4}$$

where a star indicates the missing counterfactual outcome. In vector notation, in the control group, the observed outcome data are

$$\mathbf{y}_0 = (y_{01}, \ldots, y_{0n_0}) : n_0 \times 1$$

and the missing counterfactual outcomes are

$$\mathbf{y}_{1c}^* = (y_{11}^*, \ldots, y_{1n_0}^*)$$

to be read as "y_1 for the controls." Similarly, in the treated group, the observed outcome data are

$$\mathbf{y}_1 = (y_{1n_0+1}, \ldots, y_{1n}) : n_1 \times 1$$

and the missing counterfactual outcomes are the "y_0 for the treated"

$$\mathbf{y}_{0t}^* = (y_{0n_0+1}^*, \ldots, y_{0n}^*)$$

The associated matrix of linear confounders, split by intake status, are indicated by

$$\mathbf{V}_0 = (\mathbf{v}_1, ..., \mathbf{v}_{n_0})' : n_0 \times k_v \text{ and } \mathbf{V}_1 = (\mathbf{v}_{n_0+1}, ..., \mathbf{v}_n)' : n_1 \times k_v$$

and those of the non-linear confounders by

$$\mathbf{W}_0 = (\mathbf{w}_1, ..., \mathbf{w}_{n_0})' : n_0 \times q \text{ and } \mathbf{W}_1 = (\mathbf{w}_{n_0+1}, ..., \mathbf{w}_n)' : n_1 \times q$$

In the sequel it will also be necessary to work with all n observations on the confounders in $\mathbf{w}$, in which case we write

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_0 \\ \mathbf{W}_1 \end{pmatrix} = (\mathbf{w}_{.1}, \mathbf{w}_{.2}, \ldots, \mathbf{w}_{.q}) \tag{2.5}$$

where $\mathbf{w}_{.r}$ $(r \leq q)$ denotes the rth column of $\mathbf{W}$.

2.2 Knots and basis matrices

We now discuss the estimation of the $p_0(y_0|\mathbf{z})$ model given the data on the control subjects and of the $p_1(y_1|\mathbf{z})$ given the data on the treated subjects. Then, the ATE is calculated (post-estimation) from the predictions of $\mathbf{y}_{1c}^*$ for the control subjects using the model $p_1(y_1|\mathbf{z})$ estimated on the treated subjects, and from the predictions of $\mathbf{y}_{0t}^*$ for the treated subjects using the model $p_0(y_0|\mathbf{z})$ estimated on the control subjects.

To make this procedure concrete, we suppose that each $g_{jr}(w_r)$ function is in the span of the natural cubic splines. We use the basis from Chib and Greenberg (2010) and a prior on the basis coefficients from Chib and Greenberg (2014) to do our prior-posterior analysis on these functions. Since the ATE is calculated by a predictive approach, it is important to recognize the potential problems that can arise from covariate imbalance across the two groups of subjects. If the problem of covariate imbalance is extreme, no method can be expected to work adequately. In the case of matching, the problem of extreme covariate imbalance would manifest itself in the form of fewer matches. In our method, the problem would be revealed by a more dispersed posterior distribution of the ATE.

The question now is how to develop a predictive approach that would work in other less problematic cases of covariate imbalance. The key idea here is to estimate the $p_0(y_0|\mathbf{z})$ and

$p_1(y_1|\mathbf{z})$ models in such a way that accurate predictions of the counterfactuals are possible. Our study of this problem reveals the key role played by the choice of knots. The default strategy of equally spaced knots often turns to be unsatisfactory, even with moderate levels of covariate imbalance across the two groups of subjects. In that case, equally spaced knots based on (say) the control observations would miss some of the covariate values in the treated group, which would lead to intervals in the control and treatment groups with no observations, and instability of the spline basis matrices and inaccurate extrapolation of the spline estimated from the control observations to those covariate values in the treated group.

Instead, we propose another strategy that overcomes the preceding problem. As motivation, consider for example the function $g_{0r}(w_r)$ in the control model. For estimating this function, the outcome data available to us is just that from the n_0 control subjects. But for purposes of making the basis functions we have data on w_r not just from the control subjects but also from the treated subjects. One can take advantage of this additional data on w_r for placing knots and constructing the basis matrix. Having made the basis matrix from all n observations, one simply drops the last n_1 rows while estimating the control model. The remaining n_1 basis matrix rows are used when predicting y^*_{0i} for the treated. Similarly, we use all the n observations on w_r to make the basis matrix for the function $g_{1r}(w_r)$ but then we remove the first n_0 rows of that basis matrix while estimating the outcome model of the treated subjects. Those n_0 rows are used in the prediction of y^*_{1i} for the controls.

We now explain how this strategy is implemented. For a given covariate w_r, suppose we need to locate m_{jr} knots. The notation m_{jr} indicates that the number of knots for a given covariate w_r could vary by control and treated subjects. Consider now the entire data on w_r, namely $(w_{1r}, \ldots, w_{nr})$. Let the first knot τ_{r1} be located at $\mathsf{min}(w_{1r}, \ldots, w_{nr})$ and the last knot τ_{rm_r} at $\mathsf{max}(w_{1r}, \ldots, w_{nr})$. To find the remaining $m_{jr} - 2$ knots, define

$$\mathsf{maxmin}_r = \mathsf{max}(\mathsf{min}(w_{1r}, \ldots, w_{n_0 r}), \mathsf{min}(w_{n_0+1r}, \ldots, w_{nr})),$$

$$\mathsf{minmax}_r = \mathsf{min}(\mathsf{max}(w_{1r}, \ldots, w_{n_0 r}), \mathsf{max}(w_{n_0+1r}, \ldots, w_{nr}))$$

Now let $a_{jr} = (\mathsf{maxmin}_r, a_{r2}, \ldots, a_{r,m_{jr}-1}, \mathsf{minmax})$ denote m_{jr} evenly spaced values between maxmin_r and minmax_r. Then, our set of m_{jr} knots is given by the collection

$$\tau_r = (\tau_{r1}, a_{r2}, \ldots, a_{r,m_{jr}-1}, \tau_{rm_r}).$$

It can be checked that with this (novel) approach, even with covariate imbalance, use of these knots ensures that the smallest and largest knots include the required interval for both the control and treated observations and that there are no empty intervals between knots. If there is no covariate imbalance or only minor imbalance, these knots will be very nearly evenly spaced between the smallest and largest values of w_r.

Given the knots, we express each of the $g_{jr}(\mathbf{w}_{.r})$ functions at the n covariate values $\mathbf{w}_{.r}$ by a natural cubic spline. Applying the spline transformation given in Appendix A we have

$$g_{0r}(\mathbf{w}_{.r}) = \mathbf{B}_{0r}\boldsymbol{\beta}_{0r} \ , \ r = 1, 2, ..., q$$

and

$$g_{1r}(\mathbf{w}_{.r}) = \mathbf{B}_{1r}\boldsymbol{\beta}_{1r} \ , \ r = 1, 2, ..., q$$

where $\mathbf{B}_{jr} : n \times (M_{jr} - 1)$ is the basis matrix and $\boldsymbol{\beta}_{jr} : (M_{jr} - 1) \times 1$ are the spline coefficients.

It is helpful to assemble these basis matrices in the following way. The basis matrices from expanding the $g_{0r}(\mathbf{w}_{.r})$ functions can be assembled as

$$\mathbf{B}_c = (\mathbf{V}, \mathbf{B}_{01}, \ldots, \mathbf{B}_{0q})$$

starting with the matrix of linear covariates $\mathbf{V} = (\mathbf{V}_0', \mathbf{V}_1')'$. Only the first n_0 rows of $\mathbf{B}_c$ are used in the estimation of the control model. Thus, it is further useful to partition $\mathbf{B}_c$ at row n_0 as

$$\mathbf{B}_c = \begin{pmatrix} \mathbf{B}_0 \\ \mathbf{B}_{0t} \end{pmatrix}$$

where the matrix $\mathbf{B}_{0t}$ is used for the prediction of the missing $\mathbf{y}_{0t}^*$ for the treated subjects in the sample.

Similarly, the basis matrices from expanding the $g_{1r}(\mathbf{w}_{\cdot r})$ functions can be assembled as

$$\mathbf{B}_t = (\mathbf{L}, \mathbf{B}_{11}, \ldots, \mathbf{B}_{1q})$$

starting again with the matrix of linear covariates $\mathbf{V}$. Because only the last n_1 rows of $\mathbf{B}_t$ are to be used in the estimation of the treatment model it is further useful to partition $\mathbf{B}_t$ at row n_0 as

$$\mathbf{B}_t = \begin{pmatrix} \mathbf{B}_{1c} \\ \mathbf{B}_1 \end{pmatrix}$$

where the matrix $\mathbf{B}_{1c}$ is used for the prediction of the missing $\mathbf{y}_{1c}^*$ for the control subjects in the sample.

2.3 Estimation of the control subject model

The model of the observed data on the control subjects after the basis function expansions can now be expressed as

$$\mathbf{y}_0 = \mathbf{B}_0\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}_0, \tag{2.6}$$

where $\mathbf{B}_0$ are the first n_0 rows of $\mathbf{B}_c$,

$$\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{00}', \boldsymbol{\beta}_{01}', ..., \boldsymbol{\beta}_{0r}')' : k_0 \times 1$$

is of size $k_0 = k_v + \sum_{r=1}^{q}(m_{0r} - 1)$, and $\boldsymbol{\varepsilon}_0$ is a n_0 vector of independently distributed student-t random variables with ν degrees of freedom and dispersion σ_0^2.

Let $\boldsymbol{\xi}_0 = (\xi_{01}, \ldots, \xi_{0n_0})$ where

$$\xi_{0i} \sim \mathcal{G}\left(\xi_{0i}|\frac{\nu_0}{2}, \frac{\nu_0}{2}\right),\ i \leq n_0$$

denote the Gamma distributed mixing variables in the hierarchical representation of the student-error error distribution

$$\varepsilon_{0i}|\xi_{0i} \sim \mathcal{N}(0, \xi_{0i}^{-1}\sigma_0^2)$$

Also let the prior take the form

$$\pi(\boldsymbol{\beta}_0|\boldsymbol{\lambda}_0)\pi(\sigma_0^2)\pi(\boldsymbol{\lambda}_0)$$

where $\boldsymbol{\lambda}_0 : (r+1) \times 1$ is a vector of smoothness parameters, as detailed in Appendix B. Then, the posterior distribution of $(\boldsymbol{\beta}_0, \sigma_0^2)$, augmented with $(\boldsymbol{\xi}_0, \boldsymbol{\lambda}_0)$, and conditioned on the given outcomes $\mathbf{y}_0$, the covariates in $\mathbf{B}_0$ and the degrees of freedom ν_0 of the student-t error distribution is

$$\pi(\boldsymbol{\beta}_0, \sigma_0^2, \boldsymbol{\xi}_0, \boldsymbol{\lambda}_0 | \mathbf{y}_0, \mathbf{B}_0, \nu_0) \propto \pi(\boldsymbol{\beta}_0, \sigma_0^2, \boldsymbol{\lambda}_0) \times \mathcal{N}_{n_0}(\mathbf{y}_0 | \mathbf{B}_0 \boldsymbol{\beta}_0, \sigma_0^2 \boldsymbol{\Xi}_0^{-1}) \prod_{i=1}^{n_0} \mathcal{G}\left(\xi_i | \frac{\nu_0}{2}, \frac{\nu_0}{2}\right), \quad (2.7)$$

where

$$\boldsymbol{\Xi}_0^{-1} = \mathsf{diag}\left(\xi_{01}^{-1}, \ldots, \xi_{0n_0}^{-1}\right).$$

This distribution can be sampled easily by MCMC methods.

2.4 Estimation of the treated subject model

In parallel with the preceding discussion, the model of the observed data on the treated subjects after the basis function expansions can be written as

$$\mathbf{y}_1 = \mathbf{B}_1 \boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1 \quad (2.8)$$

where $\mathbf{B}_1$ are the last n_1 rows of $\mathbf{B}_t$

$$\boldsymbol{\beta}_1 = (\boldsymbol{\beta}_{10}', \boldsymbol{\beta}_{11}', ..., \boldsymbol{\beta}_{1r}')' : k_1 \times 1$$

is of size $k_1 = k_v + \sum_{r=1}^{q} (m_{1r} - 1)$, and $\boldsymbol{\varepsilon}_1$ is a n_1 vector of independently distributed student-t random variables with ν degrees of freedom and dispersion σ_1^2.

Then, after the augmentation $\boldsymbol{\xi}_1 = (\xi_{11}, \ldots, \xi_{1n_1})$ where

$$\xi_{1i} \sim \mathcal{G}\left(\xi_{1i} | \frac{\nu_1}{2}, \frac{\nu_1}{2}\right), \; i > n_0$$

and the prior of the form

$$\pi(\boldsymbol{\beta}_1 | \boldsymbol{\lambda}_1) \pi(\sigma_1^2) \pi(\boldsymbol{\lambda}_1)$$

the posterior distribution of interest, conditioned on the given outcomes $\mathbf{y}_1$, the covariates in $\mathbf{B}_1$ and the degrees of freedom ν_1 of the student-t error distribution, is

$$\pi(\boldsymbol{\beta}_1, \sigma_1^2, \boldsymbol{\xi}_1, \boldsymbol{\lambda}_1 | \mathbf{y}_1, \mathbf{B}_1, \nu_1) \propto \pi(\boldsymbol{\beta}_1, \sigma_1^2, \boldsymbol{\lambda}_1) \times \mathcal{N}_{n_1}(\mathbf{y}_1 | \mathbf{B}_1 \boldsymbol{\beta}_1, \sigma_1^2 \boldsymbol{\Xi}_1^{-1}) \prod_{i=n_0+1}^{n} \mathcal{G}\left(\xi_i | \frac{\nu_1}{2}, \frac{\nu_1}{2}\right), \quad (2.9)$$

where

$$\mathbf{\Xi}_1^{-1} = \mathsf{diag}\left(\xi_{1,n_0+1}^{-1}, \ldots, \xi_{1n}^{-1}\right).$$

Sampling by MCMC methods is straightforward.

2.5 Binary outcomes

If the outcome is binary, as in one of the examples below, we assume that

$$\mathsf{Pr}(\mathbf{y}_0 = 1|\boldsymbol{\beta}_0) = T_{\nu_0}(\mathbf{B}_0\boldsymbol{\beta}_0),$$

and

$$\mathsf{Pr}(\mathbf{y}_1 = 1|\boldsymbol{\beta}_1) = T_{\nu_1}(\mathbf{B}_1\boldsymbol{\beta}_1)$$

where the probabilities are computed point-wise and $T_\nu(\cdot)$ is the cdf of the standard t distribution with ν degrees of freedom applied point-wise to its vector argument. This model is analyzed in exactly the same way as the continuous models above following the latent variable augmentation method of Albert and Chib (1993).

3 Posterior distribution of the ATE

We can now turn to finding the posterior distribution of the ATE. Under the modeling of the outcome distributions of the control and treated subjects, it follows that the ATE can be expressed as

$$\mathsf{ATE}(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1) = \mathsf{mean}\left(\begin{array}{c} \mathbf{B}_{1c}\boldsymbol{\beta}_1 - \mathbf{B}_0\boldsymbol{\beta}_0 \\ \mathbf{B}_1\boldsymbol{\beta}_1 - \mathbf{B}_{0t}\boldsymbol{\beta}_0 \end{array}\right), \tag{3.1}$$

where $\mathsf{mean}\,(\cdot)$ denotes the average of the components. Three key remarks are in order.

- First, $\mathbf{B}_{1c}\boldsymbol{\beta}_1 - \mathbf{B}_0\boldsymbol{\beta}_0$ is the conditional ATE of the control subjects and $\mathbf{B}_{1c}\boldsymbol{\beta}_1$ is the forecast of the missing $\mathbf{y}_{1c}^*$ for the control observations given the data on the treated subjects $\mathbf{y}_1$, the covariates in the basis matrix $\mathbf{B}_t$ and the parameters $\boldsymbol{\beta}_1$. Specifically,

$$\mathbb{E}(\mathbf{y}_{1c}^*|\mathbf{y}_1, \mathbf{B}_t, \boldsymbol{\theta}_1) = \mathbf{B}_{1c}\boldsymbol{\beta}_1.$$

- Second, $\mathbf{B}_1\boldsymbol{\beta}_1 - \mathbf{B}_{0t}\boldsymbol{\beta}_0$ is conditional ATE of the treated subjects and $\mathbf{B}_{0t}\boldsymbol{\beta}_0$ is the forecast of the missing $\mathbf{y}^*_{0t}$ for the treated subjects given the data on the control subjects $\mathbf{y}_0$, the covariates in the basis matrix $\mathbf{B}_c$ and the parameters $\boldsymbol{\beta}_0$. Specifically,

$$\mathbb{E}(\mathbf{y}^*_{0t}|\mathbf{y}_0, \mathbf{B}_c, \boldsymbol{\theta}_0) = \mathbf{B}_{0t}\boldsymbol{\beta}_0$$

- Third, each of the four quantities in the expression of the ATE is a function of the parameters. Hence, the posterior distribution of the ATE can be computed from the MCMC output of the parameters, as follows. Let the MCMC output on the regression parameters from the simulation of (2.7) and (2.9) be $\left\{\boldsymbol{\beta}_0^{(1)}, \ldots, \boldsymbol{\beta}_0^{(G)}\right\}$ and $\left\{\boldsymbol{\beta}_1^{(1)}, \ldots, \boldsymbol{\beta}_1^{(G)}\right\}$. Then, the sequence of values

$$\mathsf{ATE}^{(g)} = \mathsf{mean}\begin{pmatrix} \mathbf{B}_{1c}\boldsymbol{\beta}_1^{(g)} - \mathbf{B}_0\boldsymbol{\beta}_0^{(g)} \\ \mathbf{B}_1\boldsymbol{\beta}_1^{(g)} - \mathbf{B}_{0t}\boldsymbol{\beta}_0^{(g)} \end{pmatrix}, \; g = 1, \ldots, G. \tag{3.2}$$

  is a sample from the posterior distribution of the ATE.

If the outcome is binary, we can proceed as above, now letting

$$\mathsf{ATE}(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1) = \mathsf{mean}\begin{pmatrix} T_{\nu_1}(\mathbf{B}_{1c}\boldsymbol{\beta}_1) - T_{\nu_0}(\mathbf{B}_0\boldsymbol{\beta}_0) \\ T_{\nu_1}(\mathbf{B}_1\boldsymbol{\beta}_1) - T_{\nu_0}(\mathbf{B}_{0t}\boldsymbol{\beta}_0) \end{pmatrix} \tag{3.3}$$

where T_{ν_0} and T_{ν_1} are the cdf's of the standard student-t distribution with ν_0 and ν_1 degrees of freedom, respectively.

4 Simulated data example

We illustrate the proposed method by generating a data set on intake and outcomes that involves three binary and two continuous confounders with highly nonlinear effects, under the assumption that conditional ignorability holds. The simulation design ensures that the covariate imbalance problem is non-trivial. We then describe a small search to find the best model according to the Bayes factor criterion and use that model to implement our method. Sample sizes of 500, 1,000, 2,000, and 4,000 subjects are generated.

4.1 Design

The three categorical confounders are binary and are generated for each subject as Bernoulli $\mathcal{B}(p)$ random variables with success probability p, as

$$v_1 \sim \mathcal{B}(.2)\,, \quad v_2 \sim \mathcal{B}(.4)\,, \quad v_3 \sim \mathcal{B}(.5).$$

The two continuous confounders are uniform random variables,

$$w_1 \sim \mathcal{U}(0,1)\,, \quad w_2 \sim \mathcal{U}(0,1).$$

Intake is generated for each subject according to the model

$$\mathsf{Pr}(x = 1|\mathbf{v},\mathbf{w}) = T_5\left(-.5 + .2v_1 - .3v_2 + .5v_3 + g_1(w_1) + g_2(w_2)\right),$$

where

$$g_1(w_1) = -50(w_1 - .5)^4\,, \quad g_2(w) = \frac{\sin\left(\pi w_2/2\right)}{\left(1 + w_2^2 \mathsf{sign}\left(w_2 + 1\right)\right)},$$

and the potential outcomes are generated as

$$y_0 = 1 + .1v_1 + .7v_2 - .2v_3 + g_{01}(w_1) + g_{02}(w_2) + \varepsilon_0,$$

$$y_1 = 1 - .1v_1 + .5v_2 - .6v_3 + g_{11}(w_1) + g_{12}(w_2) + \varepsilon_1,$$

where

$$g_{01}(w_1) = w_1 + w_1^5\,, \quad g_{02}(w_2) = \sin\left(2\pi\left(1 - w_2\right)^2\right)\,,$$

$$g_{11}(w_1) = 5w_1 + 8w_1^4\,, \quad g_{12}(w_2) = \frac{1}{w_2 + .1} + 8\exp\left(-400\left(w_2 - .5\right)^2\right);$$

ε_0 is distributed as Student-t with $\nu_0 = 7$ degrees of freedom, and ε_1 as Student-t with $\nu_1 = 5$ degrees of freedom. There are 336, 639, 1,277, and 2,582 control subjects, respectively, in the samples of 500, 1,000, 2,000, and 4,000 observations.

One aim of this design is to incorporate a complex dependence of the confounders on the intake. This dependence is shown in Figure 1, which plots the propensity score for each of the subjects in the $n = 500$ sample against the values of w_1 and w_2 for that subject. control

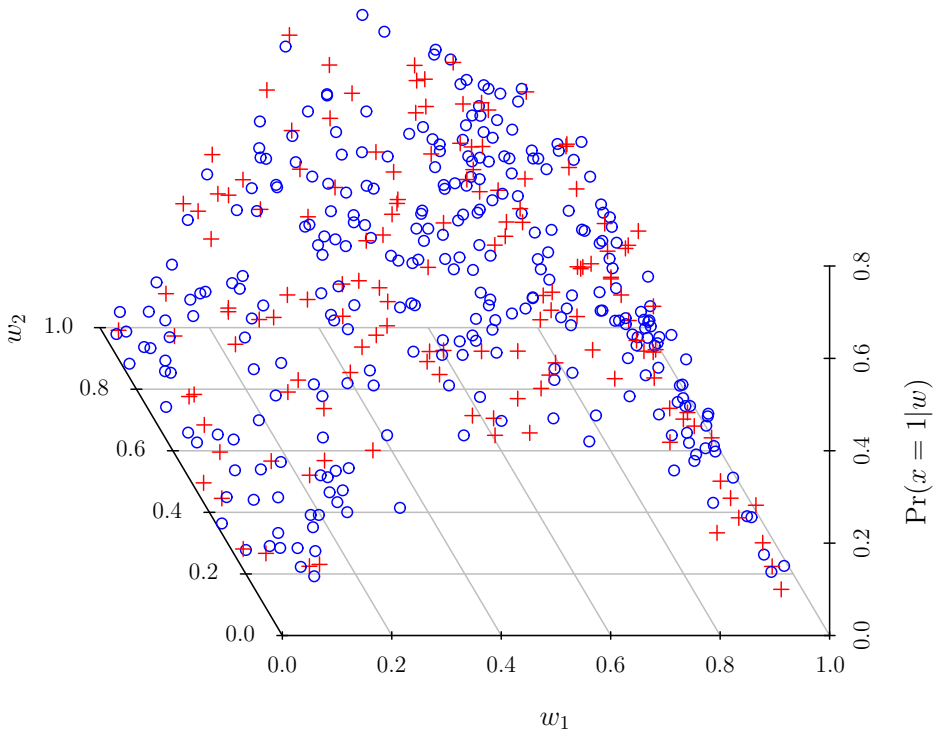

Figure 1: Plot of the true propensity score with simulated data ($n = 500$) at the generated values of the confounders. The control observations are marked in circles and the treated observations with pluses.

subjects are indicated by circles and treated subjects by pluses. This 3-D scatterplot shows an arch-like structure of the propensity scores. Small and large values of w_1 have small values of the propensity score and values of w_1 in the mid-range of the $(0, 1)$ interval generate larger values of the propensity scores. As a result, there are fewer treated observations at each end of the w_1 interval.

A second aim of this design is to produce a non-trivial overlap problem. Again focusing on the $n = 500$ sample, this problem can be seen from the contour plots of the (w_1, w_2) distribution by intake group that are given in Figure 2. The left plot in the figure, which has the distribution of $(w_1, w_2)|x = 0$, shows that the regions of high density (indicated by the higher numbers on the contour lines) are separated from one another. The distribution of $(w_1, w_2)|x = 1$ in the right side of the plot is quite different from the first distribution, with clear regions of limited overlap.

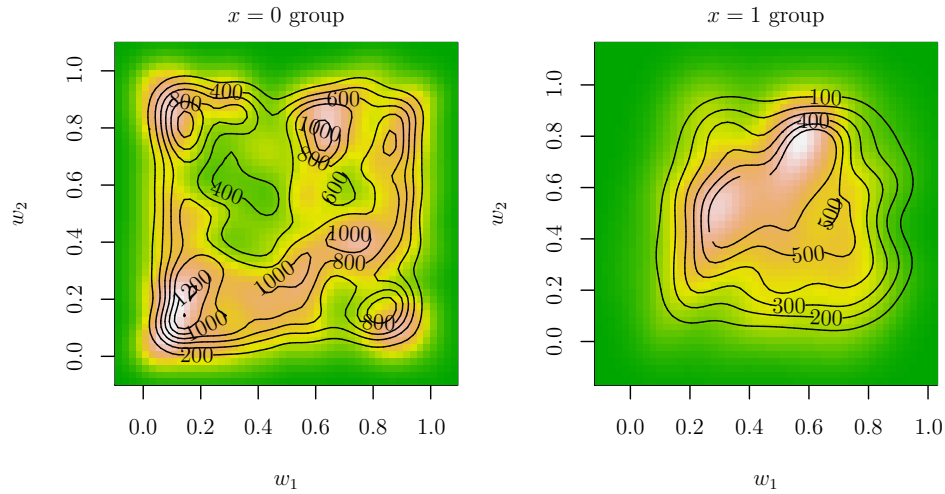

Figure 2: Contour plot of the distribution of (w_1, w_2) by intake group in the simulated data ($n = 500$); this shows the severe overlap problem.

4.2 Model fitting

The prior distribution in (B.4) and (B.5) requires specifying the prior on two initial values for each of the g_0 and g_1 functions, the priors on the variances, and the priors on the smoothness parameters λ_v, λ_0 and λ_1. We set these priors to be same across the potential outcome models and across models with different degrees of freedom by assuming that all regression coefficients are centered at zero, that both variances have a gamma prior distribution with mean 1 and standard deviation 5, and that all of the λs have means of 1 and standard deviations of 10.

Our results are based on 10,000 MCMC draws following a burn-in of 1,000 MCMC cycles. Although we do not report the results on the mixing of the MCMC chains, the inefficiency factors for each of the parameters in each model are mostly less than 2 or 3, indicating that the sampling procedures are highly efficient. We assume, incorrectly, 5 degrees of freedom for the Student-t distributions of ε_0 and ε_1.

We undertake a small model search as part of our fitting of the outcome models by considering models that have different number of knots in the spline formulation and 5 degrees of freedom in the Student-t distributions. Examination of the marginal likelihoods, computed by the method of Chib (1995) and shown in Table 1, reveals that 6 knots are sufficient for g_{11}, and

that 15 knots are necessary for g_{12}. More knots are needed in the latter equation to capture the sharp rise and fall that occurs for values of w_2 between 0.4 and 0.6. In Figure 3 we show

Knots		Sample Size			
w_1	w_2	500	1,000	2,000	4,000
		Controls			
5	5	-563.139	-1118.375	-2181.854	-4319.271
6	**6**	-562.851	-1115.194	-2181.844	-4317.586
5	10	-569.577	-1118.786	-2189.281	-4319.283
6	10	-570.442	-1119.891	-2192.645	-4322.671
10	10	-572.468	-1121.318	-2199.103	-4330.002
		Treated			
6	6	-390.496	-791.885	-1487.727	-3058.220
5	10	-375.601	-753.338	-1382.359	-2821.052
6	**15**	-366.579	-672.804	-1270.652	-2460.075
10	10	-385.074	-763.282	-1395.639	-2838.971
15	15	-377.211	-687.960	-1288.669	-2489.898
18	18	-378.880	-698.198	-1304.038	-2539.988

Table 1: Marginal likelihoods for the nonlinear equations, various knot combinations, simulated data. Knots in boldface yield greatest values of marginal likelihood.

the true functions and estimated functions for a sample of size 500.

4.3 Posterior distribution of ATE

Estimates of ATE by frequentist matching are obtained from the R package *Matching*. We report results for propensity score matching based on propensity scores from a logit link and linear covariate effects. The results are reported in Table 2.

As a simple criterion for accuracy, we determine whether the estimate $\pm$ two standard deviations includes the true value. According to this criterion, three of the four intervals based on propensity score matching cover the true value, and all four of the intervals based on our Bayesian method cover the true value. Note also that the Bayesian approach yields smaller standard deviations for all sample sizes.

Finally, even though the data come from a complicated design, Figure 4 shows that the posterior distribution of the ATE centers quickly on the true value and becomes more concentrated as the sample size increases.

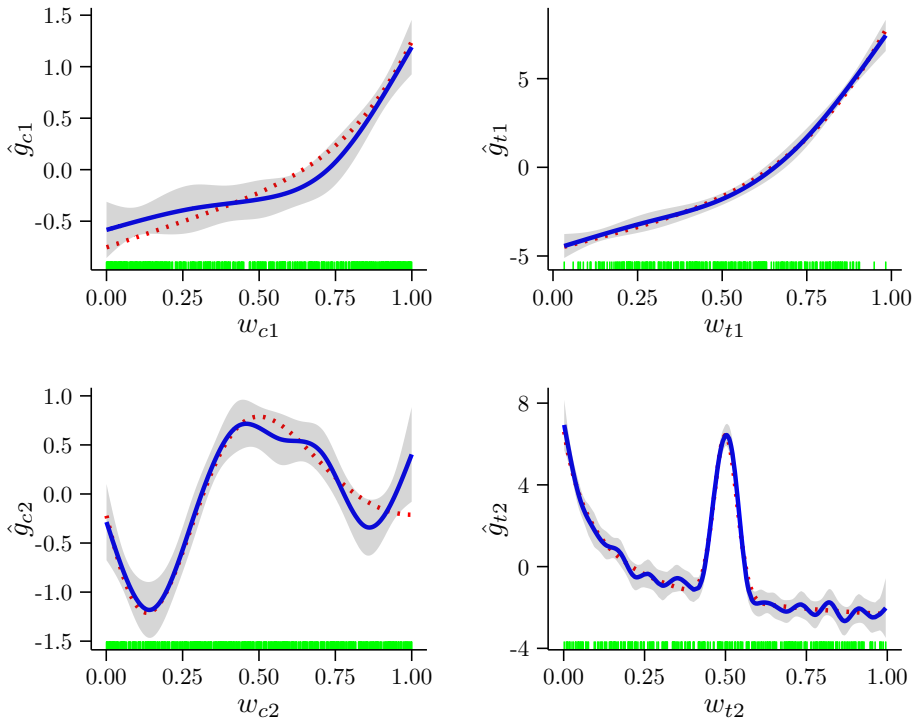

Figure 3: True (dotted lines) and estimated functions (solid lines) for simulated data and sample size 500.

	Sample size			
	500	1000	2000	4000
True value	6.072	6.145	6.031	6.037
Propensity Score Matching	5.726	5.732	5.728	5.544
	(0.332)	(0.242)	(0.168)	(0.116)
Bayesian ATE	5.799	6.308	6.119	6.023
	(0.206)	(0.111)	(0.075)	(0.053)

Table 2: True and estimated values of ATE (standard deviations in parentheses) by frequentist propensity score matching and by the Bayes approach in the text.

5 Real data examples

This section contains the application of our method to two real data sets. The first considers the effect on academic achievement of receiving AFDC payments, and the second examines the effectiveness of a medical procedure on 30-day survival rates.

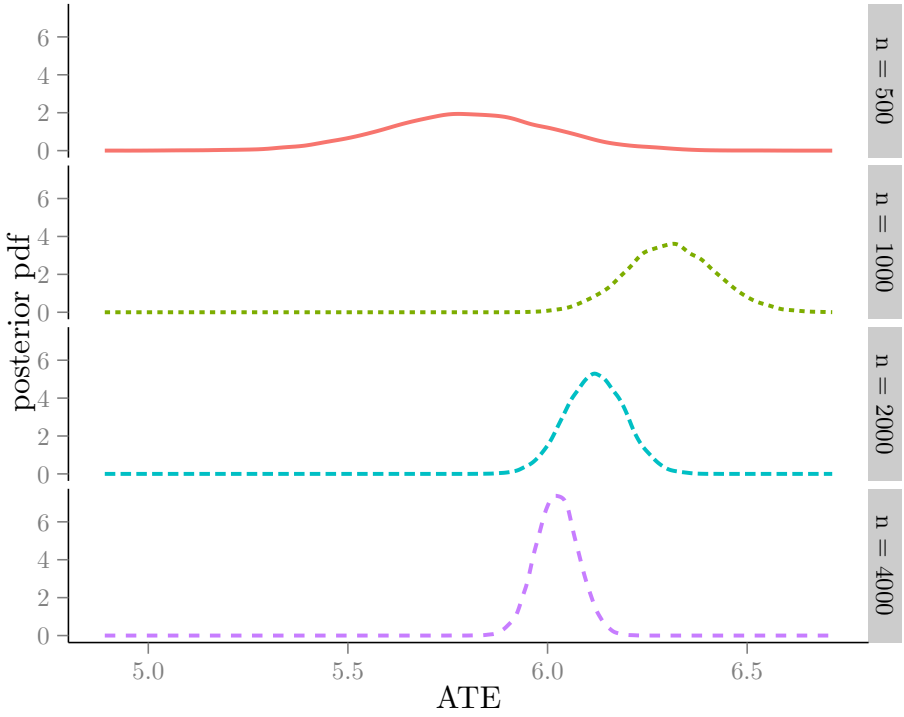

Figure 4: Simulated data: Posterior distributions of the ATE by sample size.

5.1 Academic achievement data

5.1.1 Background

This example considers a data set taken from the 1997 Child Development Supplement to the Panel Study of Income Dynamics. We use the sample analyzed by Guo and Fraser (2015, Section 5.8.2), which includes only female caregivers. The object of the study is to estimate the effect of childhood welfare dependency on academic achievement. The continuous outcome variable y is measured by the child's score on the "letter-word identification" section of the Woodcock-Johnson Revised Tests of Achievement. The treatment variable x equals one if the child received AFDC benefits at any time from birth to 1997 (the survey year) and equals zero if the child never received benefits during that period. The linear covariate in $\mathbf{v}$ are an intercept and two binary covariates: *race* is one for African-American children and zero for other, and *male* is one if the child is male and zero if female. The nonlinear confounders in $\mathbf{w}$

are *mratio*97, the ratio of family income to the poverty line in in 1997; *pcged*97, the caregiver's years of schooling; *pcg_adc*, the number of years in which the caregiver received AFDC in her childhood; and *age*97, the child's age in 1997. The sample size n is 1,003, composed of $n_0 = 729$ controls and $n_1 = 274$ treated subjects. The ATE is expected to be negative, reflecting the hypothesis that welfare dependency has an adverse effect on academic achievement.

Guo and Fraser (2015) examine these data with propensity score methods. They apply a large number of matching methods and carefully show how alternative methods affect the results. In our empirical study, we compare our Bayesian results with the matching algorithm included in the R package *Matching*.

Our outcome models are specified through Student-t links with 5 degrees of freedom. The effects of the continuous confounders are modeled by cubic splines with six knots for each confounder. This number was determined by examination of the marginal likelihoods for 5, 6, and 7 knots for the controls and 5 or 6 knots for the treated; we did not try 7 knots for the treated, because of the relatively small number of observations in that group. Since the scores are standardized with a mean of 100 and a standard deviation of 15, we set the prior expected value of the intercept to 100, the prior expected value of the dispersion parameter σ_0^2 to 200, and the prior variance of σ_0^2 to 50. Two observations were dropped from the sample because their values for *mratio96* were far larger than the other values of this variable.

5.1.2 Function estimates

Function estimates for the four continuous variables are graphed in Figure 5. The sample of observations on controls displays considerably more curvature than that of the treated, but, as noted above, the Bayes factor criterion favors 6 knots for both sets of observations. We conclude from this result that it is desirable to allow for nonlinearities in the outcome functions.

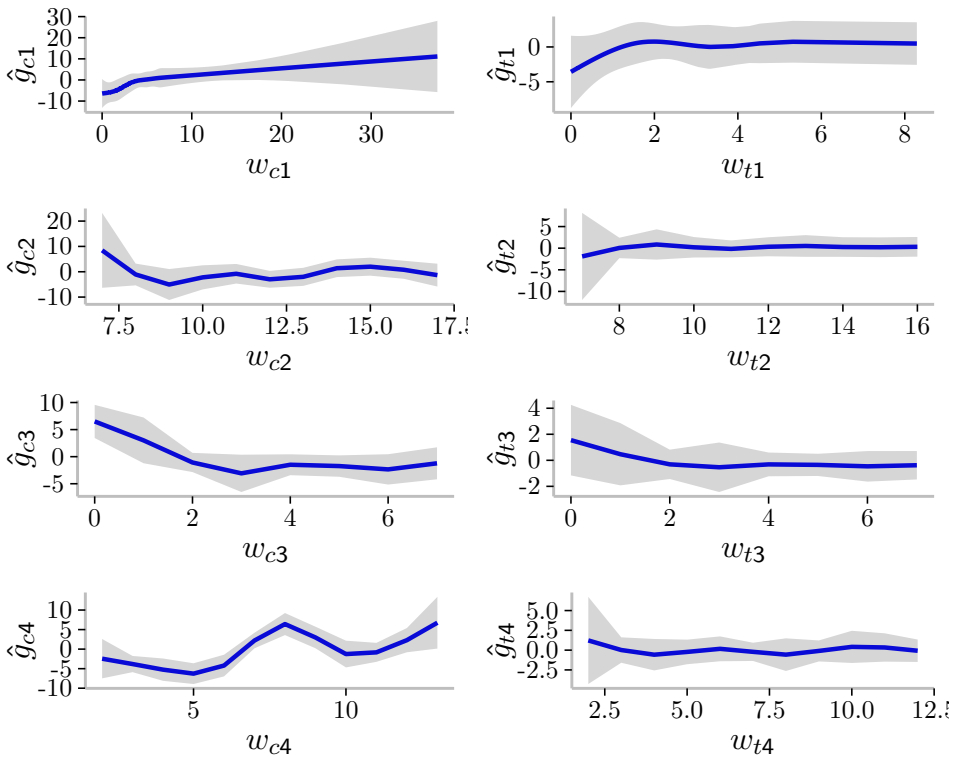

Figure 5: Academic achievement data: Estimated confounder functions in the model of the control subjects (left panel) and estimated confounder functions in the model of the treated subjects (right panel).

5.1.3 Distribution of the ATE

Table 3 and Figure 6 present summary statistics and a graph of the estimated ATE distribution. Our approach and the propensity score matching method find negative values for the mean ATE, and the interval estimates from both methods indicate that the ATE is less than zero.

	Mean	sd	Median	0.025	0.975
Propensity Score Matching	-5.499	1.999		-9.496	-1.502
Bayesian ATE	-6.437	1.415	-6.449	-9.200	-3.669

Table 3: Academic achievement data: Summary of the posterior distribution of the ATE.

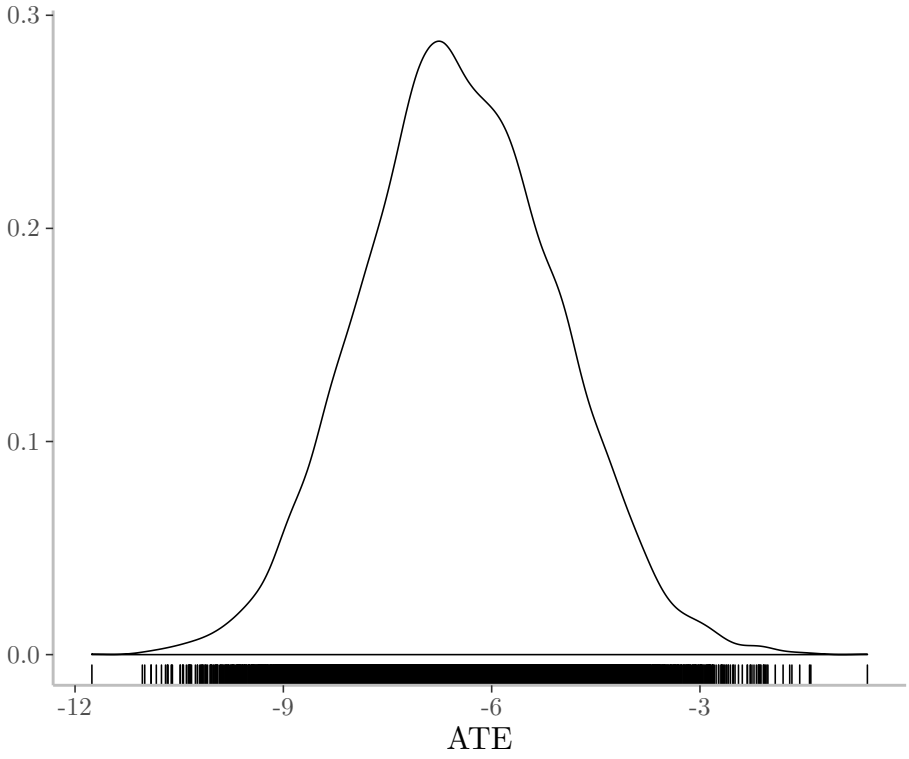

Figure 6: Academic achievement data: Posterior distribution of the ATE.

5.2 RHC data

5.2.1 Background

We next apply our method to a binary response problem which deals with the effect of a diagnostic tool called right heart catheterization (RHC) on life expectancy. The data were collected as part of the SUPPORT study, a major research effort to study physician decision making and outcomes of seriously ill, hospitalized adult patients at five medical centers. We aim to do inferences on the ATE of RHC on life expectancy in the presence of 40 linear and 16 nonlinear confounders.

In our analysis, we define the intake x to be 1 if the patient is exposed to the RHC procedure and 0 otherwise. The outcome y is 1 if the patient dies within 30 days and 0 if the patient survives beyond 30 days. Thus, a positive value of ATE implies that exposure to the intake increases the probability of dying within 30 days.

For both the controls and treated, the confounders in $\mathbf{v}$ consist of 40 categorical variables that represent primary and secondary diseases, comorbidities, whether the patient has cancer and whether it is metastatic, sex, race, income groups, insurance status, admission diagnosis, and whether the patient chose to be resuscitated. There are 16 continuous confounders that constitute $\mathbf{w}$ and these measure a variety of physical measurements and other information about the patient taken at the time of admission into the hospital. The effects of each confounder in $\mathbf{w}$ is modeled by a cubic spline. After dropping some observations because of missing and extreme observations, our final sample contains 3,515 control and 2,163 treated subjects.

The probability of the binary outcome for both the control and treated subjects is modeled by a Student-t link with 5 degrees of freedom. In addition, five knots are used in the cubic spline basis expansions. This was determined by estimating models with different number of knots and comparing the marginal likelihoods (computed by the method of Chib (1995)). We found that the marginal likelihood dropped of considerably when more than 5 knots were used. Thus, in each final model, there are 127 regression and basis function parameters, and 17 unknown λ smoothness parameters.

5.2.2 Function estimates

Posterior estimates of selected functions in y_0 and y_1 are displayed in Figure 7. The figure shows considerable nonlinearities in the effect of the continuous variables in the outcome functions; the effect of *das2d3pc* is an example. Differences in the effects of the covariates on the outcome functions suggest that estimating the treatment effect by a simple shift in the function is not appropriate; for example, at low values of *wblc1*, the probability of death increases for the controls, but decreases for the treated, and the level of *sod1* has no effect on the treated but a highly nonlinear effect on the controls. As other examples, note that *temp1* and *sod1* have nonlinear effects on the controls but no effect on the treated.

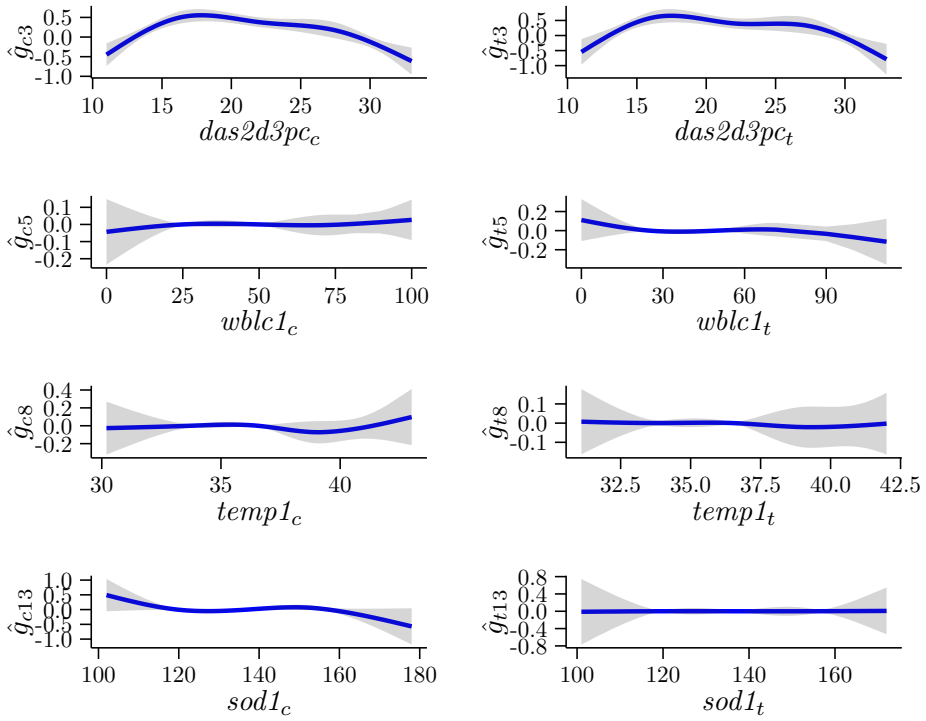

Figure 7: RHC data: Cubic spline estimates of selected functions in y_0 (first column) and y_1 (second column) models, Student-t link with 5 degrees of freedom, 5 knots for each function.

5.2.3 Distribution of the ATE

A summary of the posterior distribution of the ATE appears in Table 4. The posterior mean of the ATE is 0.043 in contrast with the propensity score based ATE of 0.039 (obtained from the *Matching* R-package.

	Mean	sd	Median	0.025	0.975
Propensity Score Matching	0.054	0.021		0.012	0.096
Bayesian ATE	0.043	0.012	0.044	0.019	0.067

Table 4: RHC data: Summary of the posterior distribution of the ATE.

The posterior distribution of the ATE is given in Figure 8. With virtually no mass in the negative region, this distribution supports the findings in previous research that the RHC procedure was not helpful in prolonging life.

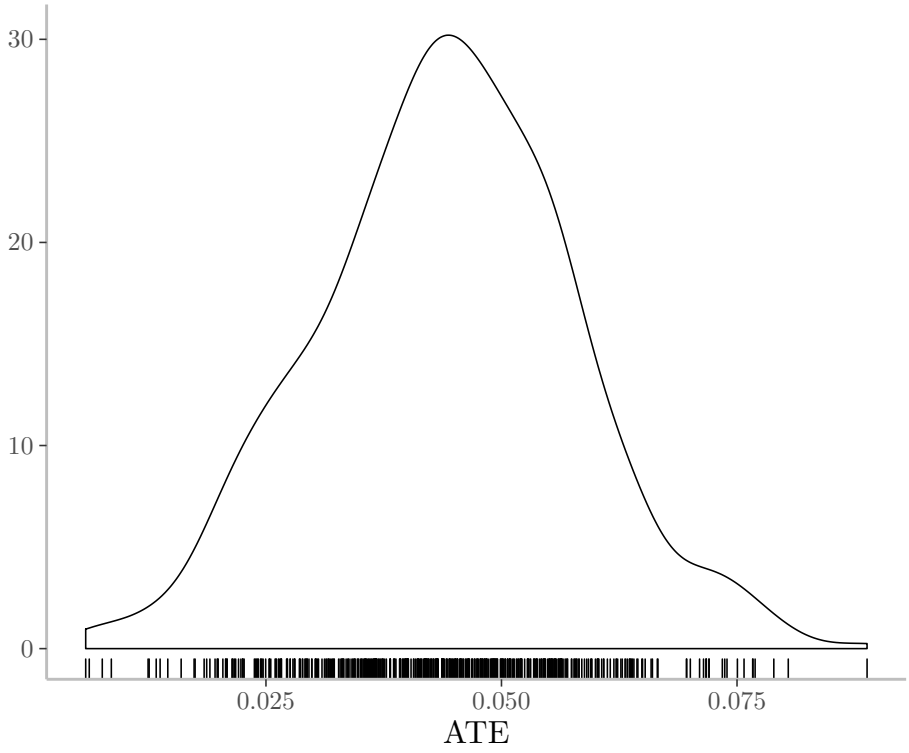

Figure 8: RHC data: Posterior distribution of the ATE.

6 Conclusions

In this paper, we have introduced a Bayesian approach for learning about the ATE of a binary treatment, when the treatment satisfies the overlap and conditional ignorability assumptions, by modeling and estimating the outcome distributions of the control and treated subjects without involvement of the propensity scores. The outcome models include flexible, non-parametric covariate functional forms. An important facet of the approach is that the ATE is calculated by predicting the missing outcomes of the control subjects from the estimated model of the treated subjects and by predicting the missing outcomes of the treated subjects from the estimated model of the control subjects. No missing counterfactuals are used in the estimation of the model parameters, however, following Chib (2007). In order to ensure that the ATE can be efficiently calculated in this way, even in the presence of overt covariate imbalance, the knots in the cubic spline basis expansions are selected in a novel manner. The approach is

illustrated with both continuous and binary outcome data. The model fitting shows evidence of nonlinear confounder effects, supporting the modeling approach advanced in this paper, and the posterior ATE results in both the simulated and real data cases show that the method has promise for use in practice.

APPENDIX

A Construction of the cubic spline basis matrix

Consider a single function $f(\mathbf{w})$ with unknown function ordinates $(f(w_1), \ldots, f(w_n))$ at each of the sample values of the covariate $\mathbf{w} = (w_1, \ldots, w_n)$. In the text, we approximate these function values by a natural cubic spline. The basis we use is described fully in Chib and Greenberg (2010).

We now show how to calculate the basis expansion

$$\begin{pmatrix} f(w_1) \\ f(w_2) \\ \vdots \\ f(w_n) \end{pmatrix} = \mathbf{B}_w \boldsymbol{\beta}_w,$$

where $\mathbf{B}_w$ is a $n \times (m-1)$ basis matrix and $\boldsymbol{\beta}_w$ is a $(m-1)$ vector of cubic spline basis parameters. We index these quantities by w because they depend on the input vector $\mathbf{w}$. The basis functions we use for our cubic spline are the functions Φ_s and Ψ_s, $s = 1, \ldots, m$, have compact support and are given by

$$\Phi_s(a) = \begin{cases} 0, & a < \tau_{s-1}, \\ -(2/h_s^3)(a-\tau_{s-1})^2(a-\tau_s-0.5h_s), & \tau_{s-1} \le a < \tau_s, \\ (2/h_{s+1}^3)(a-\tau_{s+1})^2(a-\tau_s+0.5h_{s+1}), & \tau_s \le a < \tau_{s+1}, \\ 0, & a \ge \tau_{s+1}, \end{cases}$$

$$\Psi_s(a) = \begin{cases} 0, & a < \tau_{s-1}, \\ (1/h_s^2)(a-\tau_{s-1})^2(a-\tau_s), & \tau_{s-1} \le a < \tau_s, \\ (1/h_{s+1}^2)(w-\tau_{s+1})^2(a-\tau_s), & \tau_s \le a < \tau_{s+1}, \\ 0, & a \ge \tau_{s+1}, \end{cases}$$

where $h_s = \tau_s - \tau_{s-1}$ is the spacing between the $(s-1)$st and sth knots. (The basis for the first and last knots is defined differently; see Chib and Greenberg (2010).) Next, evaluate the

basis functions for each element of $\mathbf{w}$ and each knot, and arrange them in the $n \times m$ matrices Φ and Ψ as

$$\mathbf{\Phi} = \begin{pmatrix} \Phi_1(w_1) & \cdots & \Phi_m(w_1) \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \Phi_1(w_n) & \cdots & \Phi_m(w_n) \end{pmatrix}, \ \mathbf{\Psi} = \begin{pmatrix} \Psi_1(w_1) & \cdots & \Psi_m(w_1) \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \Psi_1(w_n) & \cdots & \Psi_m(w_n) \end{pmatrix}.$$

Now let $\omega_s = h_s/(h_s+h_{s+1})$, $\mu_s = 1-\omega_s$, and define the $(m\times m)$ tri-diagonal matrix $\mathbf{A}$ with 2 on the principal diagonal, $(\omega_2, \omega_3, \ldots, \omega_{m-1}, 1)$ on the first sub-diagonal, and $(1, \mu_2, \mu_3, \ldots, \mu_{m-1})$ on the first super-diagonal. Also define the $(m\times m)$ matrix $\mathbf{C}$ equal to 3 times a tri-diagonal matrix that has $(-\frac{1}{h_2}, \frac{\omega_2}{h_2}-\frac{\mu_2}{h_3}, \ldots, \frac{\omega_{m-1}}{h_{m-1}}-\frac{\mu_{m-1}}{h_m}, \frac{1}{h_m})$ on the principal diagonal, $(-\frac{\omega_2}{h_2}, -\frac{\omega_3}{h_3}, \ldots, -\frac{\omega_{m-1}}{h_{m-1}}, -\frac{1}{h_m})$ on the first sub-diagonal, and $(\frac{1}{h_2}, \frac{\mu_2}{h_3}, \ldots, \frac{\mu_{m-1}}{h_m})$ on the first super-diagonal. Let

$$\mathbf{B}^\dagger = \mathbf{\Phi} + \mathbf{\Psi}\mathbf{A}^{-1}\mathbf{C} \equiv (\mathbf{b}_1, \ldots, \mathbf{b}_m),$$

where $\mathbf{b}_s \in \Re^n$ is the sth column of $\mathbf{B}^\dagger$. For identification purposes, we restrict the m coefficients of $\mathbf{B}^\dagger$, $\boldsymbol{\beta}^\dagger$, by requiring $\sum \beta_k^\dagger = 0$ and use the restriction to eliminate

$$\beta_1^\dagger = -(\beta_2^\dagger + \cdots + \beta_m^\dagger).$$

The cubic spline basis matrix for a given covariate is then given by

$$\mathbf{B}_w = (\mathbf{b}_2 - \mathbf{b}_1, \ldots, \mathbf{b}_m - \mathbf{b}_1)$$

with coefficient vector $\boldsymbol{\beta}_w = (\beta_2, \ldots, \beta_m)'$. A nice property of this basis is that each component of the cubic spline parameters $\boldsymbol{\beta}_w$ is the value of the unknown function at the corresponding knot i.e.,

$$\boldsymbol{\beta}_w = \begin{pmatrix} f(\tau_2) \\ \vdots \\ f(\tau_m) \end{pmatrix}.$$

This property of the spline coefficients is particularly helpful because it aids in the formulation of the prior distribution.

B Prior distribution

Now consider the basis coefficients of a continuous confounder w_r in the control and treated models given by $\boldsymbol{\beta}_{0r} : (m_{0r} - 1) \times 1$ and $\boldsymbol{\beta}_{1r} : (m_{1r} - 1) \times 1$. A priori we suppose that each

follows a discrete time, second-order Ornstein–Uhlenbeck (O-U) process, where

$$\boldsymbol{\beta}_{0r} = (\beta_{0r2}, \beta_{0r3}, \beta_{0r4} \ldots, \beta_{0rm_{0}r}) : (m_{0r} - 1) \times 1,$$
$$\boldsymbol{\beta}_{1r} = (\beta_{1r2}, \beta_{1r3}, \beta_{1r4} \ldots, \beta_{1rm_{1r}}) : (m_{1r} - 1) \times 1$$

are the function values at the respective knots. In defining these O-U processes, we condition on $(\beta_{0r2}, \beta_{0r3})$ and $(\beta_{1r2}, \beta_{1r3})$. Effectively, such an assumption generalizes the second-difference penalty in Eilers and Marx (1996) to the situation of unequally spaced knots.

The next part of the construction is the prior distribution on the initial ordinates. Instead of an improper prior as in Lang and Brezger (2004) and Brezger and Lang (2006), this distribution is assumed to be proper, which is necessary for utilizing marginal likelihoods for comparing models. The hierarchical prior model is completed by specifying a flexible Gamma prior on the smoothness parameters with the aim of achieving data-driven smoothness.

Let

$$\Delta^2 \beta_{0ri} = (\beta_{0ri} - \beta_{0ri-1}) - (\beta_{0ri-1} - \beta_{0ri-2}), \quad i > 2,$$

and define the spacings between knots by

$$h_{0ri} = \tau_{0ri} - \tau_{0ri-1}.$$

Then, our prior assumption on $\beta_{04:m_0} = (\beta_{0r4}, \beta_{0r5}, \ldots, \beta_{3rm_{0}r})$ conditioned on $(\beta_{0r2}, \beta_{0r3})$ is that

$$\Delta^2 \beta_{0ri} = -(\beta_{0ri-1} - \beta_{0ri-2}) h_{0ri} + u_{0ri},$$
$$u_{0ri} | \lambda_{0r} \sim \mathcal{N}\left(0, \frac{1}{\lambda_{0r}} h_{0ri}\right),$$

where $(\beta_{0ri-1} - \beta_{0ri-2}) h_{0ri}$ introduces mean reversion and λ_{0r} is an unknown smoothness parameter. Under an analogous set of assumptions for β_{1r} we have

$$\Delta^2 \beta_{1ri} = -(\beta_{1ri-1} - \beta_{1ri-2}) h_{1ri} + u_{1ri},$$
$$u_{1ri} | \lambda_{1r} \sim \mathcal{N}\left(0, \frac{1}{\lambda_{1r}} h_{1ri}\right),$$

where λ_{1r} is another unknown smoothness parameter.

Next consider the starting ordinates, $(\beta_{0r2}, \beta_{0r3})$ and $(\beta_{1r2}, \beta_{1r3})$. Let

$$\mathbf{T}^{-1}_{0r1:2} = (\mathbf{B}'_{0r} \mathbf{B}_{0r})_{1:2}$$

denote the first two rows and columns of $\mathbf{B}_{0r}'\mathbf{B}_{0r}$ and

$$\mathbf{T}_{1r1:2} = \left(\mathbf{B}_{1r}'\mathbf{B}_{1r}\right)_{1:2}$$

denote the first two rows and columns of $\mathbf{B}_{1r}'\mathbf{B}_{1r}$. Then, our prior assumption about these quantities is that

$$\begin{pmatrix} \beta_{0r2} \\ \beta_{0r3} \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} \beta_{0r2,0} \\ \beta_{0r3,0} \end{pmatrix}, \frac{1}{\lambda_{0r}}\mathbf{T}_{0r1:2}\right)$$

and

$$\begin{pmatrix} \beta_{1r2} \\ \beta_{1r3} \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} \beta_{1r2,0} \\ \beta_{1r3,0} \end{pmatrix}, \frac{1}{\lambda_{1r}}\mathbf{T}_{1r1:2}\right).$$

It is worth observing that this prior is based on only four free hyper-parameters for each covariate, $\beta_{0r2,0}$ and $\beta_{0r3,0}$ for the $\boldsymbol{\beta}_{0r}$-prior, and $\beta_{1r2,0}$ and $\beta_{1r3,0}$ for the $\boldsymbol{\beta}_{1r}$-prior, apart from the smoothness parameters. Our experience shows that inferences are not sensitive to the choice of these hyperparameters. A rule of thumb is to set the hyper-parameters to equal roughly the prior means of g_{r2} and g_{r3}, if such information is available. Otherwise, these values may be set to equal zero.

Straightforward calculations show that the foregoing prior assumptions can be conveniently rewritten as

$$\boldsymbol{\beta}_{0r}|\lambda_{0r} \sim \mathcal{N}_{m_{0r}-1}\left(\mathbf{D}_{0r}^{-1}\boldsymbol{\beta}_{0r,0}, \frac{1}{\lambda_{0r}}\mathbf{D}_{0r}^{-1}\mathbf{T}_{0r}\mathbf{D}_{0r}^{-1'}\right)$$

and

$$\boldsymbol{\beta}_{1r}|\lambda_{1r} \sim \mathcal{N}_{m_{1r}-1}\left(\mathbf{D}_{1r}^{-1}\boldsymbol{\beta}_{1r,0}, \frac{1}{\lambda_{1r}}\mathbf{D}_{1r}^{-1}\mathbf{T}_{1r}\mathbf{D}_{1r}^{-1'}\right),$$

where $\mathbf{D}_{0r}$ and $\mathbf{D}_{1r}$ are tri-diagonal matrices of the form

$$\boldsymbol{D}_{jr} = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & \dots & 0 & 0 \\ \frac{(1-h_{jr,3})}{\sqrt{h_{jr,3}}} & \frac{(h_{jr,3}-2)}{\sqrt{h_{jr,3}}} & \frac{1}{\sqrt{h_{jr,3}}} & 0 & 0 & 0 & \dots & 0 \\ 0 & \frac{(1-h_{jr,4})}{\sqrt{h_{jr,4}}} & \frac{(h_{jr,4}-2)}{\sqrt{h_{jr,4}}} & \frac{1}{\sqrt{h_{jr,4}}} & 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \frac{(1-h_{jr.m_{jr}-1})}{\sqrt{h_{jr,m_{jr}-1}}} & \frac{(h_{jr,m_{jr}-2})}{\sqrt{h_{jr,m_{jr}-1}}} & \frac{1}{\sqrt{h_{jr,m_r-1}}} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{(1-h_{jr,m_{jr}})}{\sqrt{h_{jr,m_{jr}}}} & \frac{(h_{jr,m_{jr}-2})}{\sqrt{h_{jr,m_{jr}}}} & \frac{1}{\sqrt{h_{j,m_{jr}}}} \end{pmatrix}.$$

$$\boldsymbol{\beta}_{0r,0} = (\beta_{0r2,0}, \beta_{0r3,0}, 0, \dots, 0)' : (m_{0r}-1) \times 1,$$

$$\boldsymbol{\beta}_{1r,0} = (\beta_{1r2,0}, \beta_{1r3,0}, 0, \dots, 0)' : (m_{1r}-1) \times 1,$$

$$\mathbf{T}_{0r} = \mathsf{blockdiag}(\mathbf{T}_{0r,1:2}, \mathbf{I}_{m_0-2}) : (m_{0r} - 1) \times 1,$$

and

$$\mathbf{T}_{1r} = \mathsf{blockdiag}(\mathbf{T}_{1r,1:2}, \mathbf{I}_{m_1-2}) : (m_{1r} - 1) \times 1.$$

Thus, the penalty matrices of the g_{0r} and g_{1r} functions are $\lambda_{0r}\mathbf{D}'_{0r}\mathbf{T}_{0r}^{-1}\mathbf{D}_{0r}$ and $\lambda_{1r}\mathbf{D}'_{1r}\mathbf{T}_{1r}^{-1}\mathbf{D}'_{1r}$, respectively.

The prior of these coefficients is completed by supposing that each smoothness parameter λ_j is distributed as Gamma with a prior mean of 1 and prior standard deviation of 10. Following Claeskens et al. (2009), we also suppose that the number of knots increases with the sample size as does the size of each size λ_j. We thus suppose that the prior mean of λ_j is adjusted upwards with n and the number of knots. Finally, the prior on the coefficients of the linear covariates is joint normal and on the error dispersion σ_j^2 is inverse-gamma.

References

Albert, J. H. and Chib, S. (1993), "Bayesian analysis of binary and polychotomous response data," *Journal of the American Statistical Association*, 88, 669–679.

An, W. (2010), "Bayesian Propensity Score Estimators: Incorporating Uncertainties In Propensity Scores Into Causal Inference," *Sociological Methodology, Vol 40*, 40, 151–189.

Brezger, A. and Lang, S. (2006), "Generalized structured additive regression based on Bayesian P-splines," *Computational Statistics & Data Analysis*, 50, 967–99.

Chib, S. (1995), "Marginal likelihood from the Gibbs output," *Journal of the American Statistical Association*, 90, 1313–1321.

— (2007), "Analysis of treatment response data without the joint distribution of potential outcomes," *Journal of Econometrics*, 140, 401–412.

Chib, S. and Greenberg, E. (2010), "Additive cubic spline regression with Dirichlet process mixture errors," *Journal of Econometrics*, 156, 322–336.

Claeskens, G., Krivobokova, T., and Opsomer, J. D. (2009), "Asymptotic properties of penalized spline estimators," *Biometrika*, 96, 529–544.

Eilers, P. H. C. and Marx, B. D. (1996), "Flexible Smoothing with B-Splines and Penalties (with discussion)," *Statistical Science*, 11, 89–121.

Guo, S. and Fraser, M. W. (2015), *Propensity Score Analysis: Statistical Methods and Applications*, Advanced Quantitative Techniques in the Social Sciences, Thousand Oaks, CA: Sage, 2nd ed.

Hill, J. L. (2011), "Bayesian Nonparametric Modeling for Causal Inference," *Journal of Computational and Graphical Statistics*, 20, 217–240.

Hoshino, T. (2008), "A Bayesian propensity score adjustment for latent variable modeling and MCMC algorithm," *Computational Statistics & Data Analysis*, 52, 1413–1429.

Kaplan, D. and Chen, J. S. (2012), "A Two-Step Bayesian Approach for Propensity Score Analysis: Simulations and Case Study," *Psychometrika*, 77, 581–609.

Lang, S. and Brezger, A. (2004), "Bayesian P-Splines," *Journal of Computational and Graphical Statistics*, 13, 183–212.

McCandless, L. C., Gustafson, P., and Austin, P. C. (2009), "Bayesian propensity score analysis for observational data," *Statistics in Medicine*, 28, 94–112.

Rosenbaum, P. R. and Rubin, D. B. (1983), "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70, 41–55.

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

Wang, C., Parmigiani, G., and Dominici, F. (2012), "Bayesian Effect Estimation Accounting for Adjustment Uncertainty," *Biometrics*, 68, 661–671.

Zigler, C. M., Watts, K., Yeh, R. W., Wang, Y., Coull, B. A., and Dominici, F. (2013), "Model Feedback in Bayesian Propensity Score Estimation," *Biometrics*, 69, 263–273.