Bayesian Matching for Causal Inference

Siddhartha Chib* Edward Greenberg†

April 2010

Abstract

In this paper we provide Bayesian matching methods for finding the causal effect of a binary intake variable $x \in \{0, 1\}$ on an outcome of interest y. One technique we introduce is a Bayesian variant of the classic Rosenbaum and Rubin (1983, 1984) propensity score matching method. We show how it is possible to find the posterior distribution of the Bayesian matched sample average treatment effect (SATE) under weak semi-parametric assumptions regarding the model of the propensity score. We also consider another version of matching, which we call Bayesian self-matching, where the outcome distributions are modeled directly in a semi-parametric way and, for each subject, the expectation of the missing counterfactual (or the counterfactual itself) is obtained from the relevant outcome density. We also show how our Bayesian matching procedures generalize to randomized experiments with a compliance problem. Under the principal stratification environment of Frangakis and Rubin (2002), we show how to find the posterior distribution of the sample complier average treatment effect (SCATE) by matching on compliance probabilities and by Bayesian self-matching. We compare the Bayesian matching methods against each other and against frequentist matching estimators and causal estimates with both simulated and real data.

Key words: Bayesian propensity score matching; Bayesian self-matching; compliance problem; cubic spline; Markov chain Monte Carlo; marginal likelihood; randomized experiments; observational data; overlap problem; sample average treatment effect; sample complier average treatment effect; semiparametric Bayesian infernce.

1 Introduction

This paper is a contribution to the literature on matching methods that are used in the context of observational data to find the causal effect of a binary intake variable $x \in \{0, 1\}$ on an outcome of interest y, given confounders $\mathbf{W}$. The literature on matching methods is voluminous, and the many issues surrounding these methods, as well as the issues surrounding causal inference with observational data, are summarized in, for example, Rosenbaum (2002, 2009), Rubin (2005,

*Olin Business School, Washington University in St. Louis, St. Louis MO 63130; chib@wustl.edu
†Department of Economics, Washington University in St. Louis, St. Louis MO 63130; edg@artsci.wustl.edu

2007), Morgan and Winship (2007), and Pearl (2009). In contrast to most of this literature, we approach the matching problem from a Bayesian perspective. We show that the Bayesian perspective to matching is illuminating and opens up new possibilities for causal inference by matching methods.

The Bayesian matching techniques we assemble in this paper are simple to implement and do not rely on strong modeling assumptions. One technique we introduce is a Bayesian variant of the classic Rosenbaum and Rubin (1983, 1984) propensity score matching method. We show how it is possible to find the posterior distribution of the Bayesian matched sample average treatment effect (SATE) under weak semi-parametric assumptions regarding the model of the propensity score. We focus on the SATE because marginalizing out the covariates other than through the empirical distribution of the covariates is a difficult problem unless the true distribution of the covariates is known or can be estimated accurately, as when the sample size is truly large.

In developing the Bayesian variant, we exploit the fact that propensity score matching is an algorithmically defined function that maps the observed data into the SATE through steps in which the propensity scores are calculated, matched subjects are found, and the SATE is calculated as the sample average of the differences between outcomes of matched subjects. From a Bayesian perspective, the posterior distribution of this function can be found by simulation from the posterior distribution of the propensity scores. For each drawing of the propensity scores from the posterior distribution we can calculate the value of SATE by matching subjects on those scores. We can repeat this process for every new drawing of the propensity scores from the posterior distribution, each time forming new sets of matched subjects. The sample of SATE values generated in this way is a sample from the posterior distribution of the SATE.

We also consider another version of matching, which we call Bayesian self-matching, where the outcome distributions are modeled directly in a semi-parametric way. These outcome distributions are then used either to find the expectation of the missing counterfactual for each subject or to form predictive densities of the missing counterfactual, as in Chib and Hamilton (2002) and Chib (2007). Although this procedure may seem similar to what is termed the regression approach in the frequentist literature (Imbens, 2004), our implementation is sufficiently different that it may be classified as a new approach to the problem.

The self-matching approach we devise takes advantage of recent developments in non-parametric regression methods. We mitigate the effect of covariate imbalance by selecting the knots from the covariate values for both sets of subjects, although the fitting of the outcome

distributions is based only on the subgroups of subjects in the control and treated categories. Furthermore, from the Bayesian perspective, we have a natural scheme for generating predictive densities of the counterfactuals that also account fully for parameter and function uncertainty. These uncertainties are difficult to incorporate into the prediction of the counterfactuals from the frequentist regression approach. We show in the examples that Bayesian self-matching, in either its expectation or predictive forms, produces posterior distributions of the SATE that cover the true value even when there is considerable imbalance in the covariate distributions across the groups.

In the remainder of the paper we show how Bayesian matching can be generalized to randomized experiments with a compliance problem. In such experiments, the lack of compliance with the assignment raises the possibility of confounding on unobserved factors, even though assignment to each of the treatment arms is randomized. We develop our methods in the Frangakis and Rubin (2002) principal stratification environment where the unobserved confounder represents discrete subject types and, conditioned on the assignment and the unobserved type confounder, observed covariates are independent of the intake. In this context we take advantage of the fact that the unobserved type confounder is sampled in the fitting. These form the basis for finding the posterior distribution of the sample complier average treatment effect (SCATE), defined below, which is the relevant causal effect in this problem.

The rest of the paper is organized as follows. In Section 2, we specify the basic setting under which matching is conducted. In Section 3, we provide our Bayesian version of propensity score matching in both parametric and semi-parametric forms. In Section 4, we describe Bayesian self-matching. In Section 5, we consider the matching techniques for randomized experiments with a compliance problem. In Section 6, we apply our methods to simulated data examples and in Section 7 to the data from the right heart catheterization (RHC) study. Section 8 has our concluding remarks. Two appendices contain details related to the cubic spline basis that we employ in our semiparametric analysis and the prior distribution on the spline coefficients.

2 Setting

Until the start of Section 5 below, we consider the following situation in the context of observational data: x is an observed intake variable with $x = 0$ indicating intake of the control treatment and $x = 1$ the intake of the new treatment; (y_0, y_1) is the potential outcome pair corresponding

to $x = 0$ and $x = 1$, respectively. The observed outcome is

$$y = y_0(1 - x) + y_1 x,$$

and $\mathbf{w}$ is a q-dimensional vector of observed pre-treatment covariates that are potential confounders. The data are observations of independently distributed subjects $(i = 1, 2, \ldots, n)$, organized so that the first n_0 observations are those for $x_i = 0$ and the next $n_1 = n - n_0$ are for $x_i = 1$

$$x_i = 0,\ y_{0i},\ y^*_{1i},\ y_i = y_{0i},\ \mathbf{w}_{0i},\ i = 1, 2, \ldots, n_0 \tag{2.1}$$

$$x_i = 1,\ y^*_{0i},\ y_{1i},\ y_i = y_{1i},\ \mathbf{w}_{1i},\ i = n_0 + 1, \ldots, n, \tag{2.2}$$

where a star indicates missing data: the potential outcome y_1 is missing when $x_i = 0$, and the potential outcome y_0 is missing when $x_i = 1$. In vector notation, $\mathbf{x}_0$ is a n_0 vector with each component equal to zero, $\mathbf{y}_0 = (y_{01}, \ldots, y_{0n_0})$ and $\mathbf{y}^*_{1c} = (y^*_{11}, \ldots, y^*_{1n_0})$ are, respectively, the observed outcomes and corresponding counterfactual y_1 that are missing for the control intake group, $\mathbf{x}_1$ is a n_1 vector with each component equal to one, $\mathbf{y}_1 = (y_{1n_0+1}, \ldots, y_{1n})$ and $\mathbf{y}^*_{0t} = (y^*_{0n_0+1}, \ldots, y^*_{0n})$ are, respectively, the observed outcomes and the corresponding counterfactual y_0 that are missing for the treatment intake group, and

$$\mathbf{W}_0 = \begin{pmatrix} \mathbf{w}'_{01} \\ \vdots \\ \mathbf{w}'_{0n_0} \end{pmatrix},\ \mathbf{W}_1 = \begin{pmatrix} \mathbf{w}'_{1n_0+1} \\ \vdots \\ \mathbf{w}'_{1n} \end{pmatrix}$$

are the $n_0 \times q$ and $n_1 \times q$ associated matrices of confounders, split by intake status. In the sequel it will also be necessary to work with all n observations on a given confounder, in which case we write the stacked confounders in terms of the n-dimensional columns as

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_0 \\ \mathbf{W}_1 \end{pmatrix} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_q)\,. \tag{2.3}$$

For later reference, these definitions are summarized in Table 1.

2.1 Assumptions

In the analysis, we make two standard assumptions: (Rosenbaum and Rubin, 1983):

- conditional independence condition (CIC): x is independent of (y_0, y_1) given $\mathbf{w}$, or

$$x \perp\!\!\!\perp (y_0, y_1)|\mathbf{w}$$

x	y_0	y_1	y	w
$\mathbf{x}_0$	$\mathbf{y}_0$	$\mathbf{y}^*_{1c}$	$\mathbf{y}_0$	$\mathbf{W}_0$
$\mathbf{x}_1$	$\mathbf{y}^*_{0t}$	$\mathbf{y}_1$	$\mathbf{y}_1$	$\mathbf{W}_1$

Table 1: Vectorized data structure: $\mathbf{x}_0$ is a n_0-vector with each component equal to 0, $\mathbf{y}_0$ is a n_0-vector consisting of y_{0i}; $\mathbf{y}^*_{1c}$ is a n_0-vector consisting of the missing y^*_{1i}, etc. Definitions of the quantities in this table are immediate from equations (2.1) and (2.2).

- overlap condition: the probability of intake (the propensity score) $e(\mathbf{w}) = p_x(x = 1|\mathbf{w})$ satisfies the condition

$$0 < e(\mathbf{w}) < 1$$

for all $\mathbf{w}$.

Along with the independence of observations across subjects, this representation implies that the joint distribution of the control and treatment intake observations can be factored as

$$p(\mathbf{y}_0, \mathbf{x}_0, \mathbf{W}_0) = \prod_{i=1}^{n_0} p_w(\mathbf{w}_{0i})(1 - e(\mathbf{w}_{0i}))p_0(y_i|\mathbf{w}_{0i})$$
$$p(\mathbf{y}_1, \mathbf{x}_1, \mathbf{W}_1) = \prod_{i=n_0+1}^{n} p_w(\mathbf{w}_{1i})e(\mathbf{w}_{1i})p_1(y_i|\mathbf{w}_{1i}),$$

where $p_w(\cdot)$ is the marginal distribution of $\mathbf{w}$, $e(\mathbf{w})$ is the propensity score, and $p_j(y_j|\mathbf{w})$ is the conditional distribution of y_j. The (non-identified) joint distribution $p(y_0, y_1|\mathbf{w})$ is not needed in the analysis as pointed out in Chib (2007).

2.2 Sample average treatment effect

The goal of the inferential analysis is to find the causal effect of x on y measured, for example, in terms of the sample average treatment effect (SATE),

$$\mathsf{SATE} = \frac{n_0}{n}\mathsf{SATE}_c + \frac{n_1}{n}\mathsf{SATE}_t,$$

where

$$\mathsf{SATE}_c = n_0^{-1} \sum_{i=1}^{n_0} \big(\mathbb{E}(y^*_{1i}|\mathbf{w}_{0i}) - \mathbb{E}(y_{0i}|\mathbf{w}_{0i})\big) \tag{2.4}$$

is the sample average treatment effect for the control intake and

$$\mathsf{SATE}_t = n_1^{-1} \sum_{i=n_0+1}^{n} \big(\mathbb{E}(y_{1i}|\mathbf{w}_{1i}) - \mathbb{E}(y^*_{0i}|\mathbf{w}_{1i})\big) \tag{2.5}$$

is the sample average treatment effect for the treatment intake. Clearly, the SATE is not identified in general, because it involves expectations of the missing $\mathbf{y}^*_{1c} = \{y^*_{1i}\}_{i=1}^{n_0}$ and $\mathbf{y}^*_{0t} = \{y^*_{0i}\}_{i=n_0+1}^{n}$. Under the assumptions of the previous section, however, the SATE is identified. To see this, it suffices to consider a particular subject in the summand of SATE_c and note that

$$\begin{aligned}\mathbb{E}(y^*_{1i}|\mathbf{w}_{0i}) - \mathbb{E}(y_{0i}|\mathbf{w}_{0i}) &= \mathbb{E}(y^*_{1i}|\mathbf{w}_{0i}, x_i = 0) - \mathbb{E}(y_{0i}|\mathbf{w}_{0i}, x_i = 0) \\ &= \mathbb{E}(y^*_{1i}|\mathbf{w}_{0i}, x_i = 1) - \mathbb{E}(y_i|\mathbf{w}_{0i}, x_i = 0),\end{aligned}$$

where in the first line we are able to insert x_i since x is independent of the potential outcomes, given $\mathbf{w}$. Here $\mathbb{E}(y^*_{1i}|\mathbf{w}_{0i}, x_i = 0)$ is the expectation of the missing y_1 for the controls. In the first term of the second line, we replace $x_i = 0$ with $x_i = 1$ because of the CIC, and in the second term we replace y_{0i} by y_i since y_{0i} is observed for these subjects. Since the $\mathbf{w}_i$ are drawn from the marginal distribution and the overlap condition holds, for large n, for every $\mathbf{w}_{0i}$ for the control intake subjects, it will be possible to find a subject $i_t \in \{n_0 + 1, \ldots, n\}$ from the treatment intake group whose $\mathbf{w}_{1i_t}$ is matched to $\mathbf{w}_{0i}$. Thus, the first term of the second line will be estimable. The second term is estimable directly from the data.

3 Bayesian propensity score matching

Matching methods build on the idea, which stems from the CIC, that each term in SATE_c,

$$\mathbb{E}(y^*_{1i}|\mathbf{w}_{0i}, x_i = 1) - \mathbb{E}(y_{0i}|\mathbf{w}_{0i}, x_i = 0),$$

can be estimated by the difference in outcomes of subjects, one from the control intake group and one from the treatment intake group, that are matched on $\mathbf{w}_{0i}$. Similarly, each term in SATE_t,

$$\mathbb{E}(y_{1i}|\mathbf{w}_{1i}, x_i = 1) - \mathbb{E}(y^*_{0i}|\mathbf{w}_{1i}, x_i = 0),$$

can be estimated by the difference in outcomes of subjects that are matched on $\mathbf{w}_{1i}$. Averaging these differences in outcomes across such matched pairs yields a consistent estimator of the SATE. Matching in this way is feasible if the covariate dimension is moderate and the overlap problem is not severe. When the covariate dimension is large, matching may be done on the propensity score rather than on the $\mathbf{w}_{0i}$ and $\mathbf{w}_{1i}$. This propensity score matching is justified by the fact that, if CIC holds, then

$$x \perp\!\!\!\perp (y_0, y_1)|e(\mathbf{w}).$$

In practice, the propensity score is modeled parametrically through a logit or probit link, and the covariates are assumed to enter the link linearly. The model parameters are estimated by maximum likelihood, and the resulting estimated propensity scores are used to implement the matching strategy. See, for example, Rosenbaum (2009) for the many ways in which matches can be found from the estimated propensity scores.

Our Bayesian version of propensity score matching is a simple but novel variant of this method. The idea stems from the recognition that matching is an algorithmically defined function that maps $(\mathbf{y}, \mathbf{x}, \mathbf{W})$ into the SATE through steps in which the propensity scores are calculated, matched subjects are found, and the SATE is calculated as the sample average of the differences between outcomes of matched subjects. Thus,

$$\mathsf{SATE} = M(\mathbf{y}, \mathbf{x}, \mathbf{W}, \boldsymbol{\gamma}),$$

where $M(\cdot)$ is the matching procedure and $\boldsymbol{\gamma}$ is a possibly high dimensional parameter vector that is used in the modeling of the propensity scores. From the Bayesian perspective, with a prior distribution on $\boldsymbol{\gamma}$ and a sample of drawings $\left\{\boldsymbol{\gamma}^{(1)}, \ldots, \boldsymbol{\gamma}^{(G)}\right\}$ from the posterior distribution of $\boldsymbol{\gamma}$ conditioned on the intake data $(\mathbf{x}, \mathbf{W})$, the posterior distribution of this function can be obtained from

$$\left\{M(\mathbf{y}, \mathbf{x}, \mathbf{W}, \boldsymbol{\gamma}^{(g)})\right\}_{g=1}^{G}. \tag{3.1}$$

This is the basic idea of Bayesian matching, which, to our knowledge, has not previously been proposed in the literature.

This idea can be used to find the prior distribution of SATE, conditioned only on $(\mathbf{y}, \mathbf{W})$, by first drawing $\boldsymbol{\gamma}$ from the prior, then drawing $\mathbf{x}$ from the intake model given $(\mathbf{W}, \boldsymbol{\gamma})$, and finally calculating the function M for each drawing of $(\mathbf{x}, \boldsymbol{\gamma})$. The resulting sample is from the prior distribution of SATE. This approach can also be used to find the prior and posterior distributions of SATE that incorporate model uncertainty, as we discuss in the examples.

3.1 Parametric framework

To illustrate this method, consider for simplicity a parametric setup that is common in frequentist discussions. Let

$$e(\mathbf{w}, \boldsymbol{\gamma}) = \Phi(\gamma_0 + \mathbf{w}'\boldsymbol{\gamma}_1),$$

where Φ is the cdf of the standard normal distribution, let

$$e(\mathbf{W}, \boldsymbol{\gamma}) = \Phi(\gamma_0 \mathbf{i}_n + \mathbf{w}_1\gamma_{11} + \cdots + \mathbf{w}_q\gamma_{1q}) \tag{3.2}$$

$$= \Phi(\gamma_0 \mathbf{i}_n + \mathbf{W}\boldsymbol{\gamma}_1)$$

be the n-vector of propensity scores, where the Φ function is applied point-wise to each component of the vector $\gamma_0 \mathbf{i}_n + \mathbf{W}\boldsymbol{\gamma}_1$, and let $\pi(\boldsymbol{\gamma})$ be the prior on $\boldsymbol{\gamma} = (\gamma_0, \boldsymbol{\gamma}_1)$. The posterior distribution of $\boldsymbol{\gamma}$ given $\mathbf{x} = (\mathbf{x}_0, \mathbf{x}_1)$ is then

$$\pi(\boldsymbol{\gamma}|\mathbf{x}) \propto \pi(\boldsymbol{\gamma}) \prod_{i=1}^{n_0} (1 - e(\mathbf{w}_{0i})) \prod_{i=n_0+1}^{n} e(\mathbf{w}_{1i}).$$

To find the posterior distribution of SATE, we proceed as follows. First, the posterior distribution of $(\boldsymbol{\gamma}|\mathbf{x})$ may be sampled by MCMC methods (see Chib, 2001, for a summary of these methods) to produce a correlated sample of draws $\boldsymbol{\gamma}^{(1)}, \ldots, \boldsymbol{\gamma}^{(G)}$. Second, for each drawing, evaluate the matching function $M(\mathbf{y}, \mathbf{x}, \mathbf{W}, \boldsymbol{\gamma}^{(g)})$ by calculating the vector of propensity scores

$$e(\mathbf{W}, \boldsymbol{\gamma}^{(g)}) \ = \Phi(\gamma_0^{(g)} \mathbf{i}_n + \mathbf{W}\boldsymbol{\gamma}_1^{(g)}).$$

Next, use these propensity scores to match each control intake subject with a treatment intake subject that is closest to it in terms of the propensity score and to match each treatment intake subject with a control intake subject that is closest to it in terms of the propensity score. Finally, from each of these n matched pairs, calculate the difference in outcomes and average those differences across subjects. By the theory of Monte Carlo sampling, this is a drawing of SATE from its posterior distribution. Repeated across MCMC iterations, this procedure produces a sequence of SATE values

$$\left\{ \mathsf{SATE}^{(1)}, \ldots, \mathsf{SATE}^{G)} \right\}$$

that are a sample from the posterior distribution of SATE. We can summarize these sampled values to find the mean and various quantiles of the posterior distribution of SATE. These are the Bayesian propensity score matched summaries of the SATE.

3.2 Semiparametric framework

A more general version of Bayesian matching is possible if the effect of the covariates is modeled nonparametrically. We illustrate our approach by modeling the functions as cubic splines, although other methods of semi-parametric modeling could be used (Denison et al., 2002, Dunson et al., 2007). It is also possible to let the link function be non-parametric, say along the lines of Basu and Chib (2003), but we do not consider that extension in this paper, because neglected covariate non-linearity is usually of greater concern than a misspecification of the link

function. Accordingly, to generalize the model of the propensity scores in (3.2), we assume, with a redefined $\boldsymbol{\gamma}$, that

$$e(\mathbf{W}, \boldsymbol{\gamma}) = F(\mathbf{L}\boldsymbol{\gamma}_0 + g_1(\mathbf{w}_1) + \cdots + g_q(\mathbf{w}_q)),$$

where F is a known distribution function applied point wise, $\mathbf{L}$ is a $n \times k_0$ matrix of categorical variables, and for $r = 1, \ldots, q$, $g_r(\mathbf{w}_r) = (g_r(w_{r1}), \ldots, g_r(w_{rn}))$, also applied point-wise, are unknown functions. We obtain the posterior distribution of SATE in this semiparametric framework by placing a prior on $\boldsymbol{\gamma}_0$ and the unknown functions, finding the posterior distribution of these quantities, sampling the posterior distribution by MCMC methods, calculating $e(\mathbf{W}, \boldsymbol{\gamma})$ for each drawing, matching as in the parametric case described above, and repeating the matching procedure for each MCMC drawing. This provides the posterior distribution of SATE in a semiparametric framework.

We operationalize this framework by approximating $g_r(\mathbf{w}_r) = (g_r(w_{r1}), \ldots, g_r(w_{rn}))$ by a natural cubic spline with the basis that appears in Chib and Greenberg (2010). With $M_r \ll n$ knots, potentially different for each function, each $g_r(\mathbf{w}_r)$ can be expressed as

$$g_r(\mathbf{w}_r) = \mathbf{B}_r\boldsymbol{\gamma}_r,$$

where $\mathbf{B}_r$ is a $n \times (M_r - 1)$ matrix and $\boldsymbol{\gamma}_r : (M_r - 1)$ are the cubic spline parameters with the convenient property of being related to the function ordinates at each of the knots, starting with the second. (See Appendix A for details.) In terms of this representation, the propensity score vector is given by

$$e(\mathbf{W}, \boldsymbol{\gamma}) = F(\mathbf{L}\boldsymbol{\gamma}_0 + \mathbf{B}_1\boldsymbol{\gamma}_r + \cdots + \mathbf{B}_q\boldsymbol{\gamma}_r) = F(\mathbf{B}_x\boldsymbol{\gamma}) \ ,$$

where $\mathbf{B}_x = (\mathbf{L}, \mathbf{B}_1, \ldots, \mathbf{B}_q)$ is a $n \times k_x$ matrix with $k_x = k_0 + \sum_{r=1}^{q}(M_r - 1)$ and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_q)$ is a k_x vector. We index the spline matrix by x to distinguish it from the spline matrices that appear below in our discussion of the potential outcome distributions.

In our application, we assume that F is the cdf of the standard t distribution with ν degrees of freedom, where ν is unknown. The Student-t link, an important extension of the popular probit and logit links, is discussed in Albert and Chib (1993). In conjunction with the latent variable approach introduced in that paper, it leads to a posterior distribution that can sampled easily by MCMC methods. Write $\mathbf{B}_x$ in terms of its rows as

$$\mathbf{B}_x = \begin{pmatrix} \mathbf{r}_1' \\ \vdots \\ \mathbf{r}_n' \end{pmatrix}$$

and let x_i^* be a latent random-variable such that $x_i = I[x_i^* > 0]$. Then, independently across i, we assume that

$$x_i^* | \boldsymbol{\gamma}, \lambda_i \sim \mathcal{N}\left(\mathbf{r}_i' \boldsymbol{\gamma}, \lambda_i^{-1}\right) \ , \ \lambda_i | \nu \sim \mathcal{G}\left(\frac{\nu}{2}, \frac{\nu}{2}\right).$$

Marginalized over λ_i, the link function in the propensity score model is Student-t with ν degrees of freedom. This semi-parametric Bayesian propensity score model can be fit by the method of Albert and Chib (1993). The posterior distribution of interest conditioned on $(\mathbf{x}, \nu)$ is

$$\pi(\mathbf{x}^*, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\sigma}_e^2, \boldsymbol{\sigma}_d^2 | \mathbf{x}, \nu) \propto \pi(\boldsymbol{\sigma}_e^2)\pi(\boldsymbol{\sigma}_d^2)\pi(\boldsymbol{\gamma}|\boldsymbol{\sigma}_e^2, \boldsymbol{\sigma}_d^2) \prod_{i=1}^{n_0} \mathcal{G}\left(\nu_i | \frac{\nu}{2}, \frac{\nu}{2}\right) \mathcal{N}\left(x_i^* | \mathbf{r}_i' \boldsymbol{\gamma}, \lambda_i^{-1}\right) I[x_i^* < 0]$$
$$\times \prod_{i=n_0+1}^{n} \mathcal{G}\left(\nu_i | \frac{\nu}{2}, \frac{\nu}{2}\right) \mathcal{N}\left(x_i^* | \mathbf{r}_i' \boldsymbol{\gamma}, \lambda_i^{-1}\right) I[x_i^* > 0],$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$, $\boldsymbol{\sigma}_e^2$ and $\boldsymbol{\sigma}_d^2$ are the $2q$-vectors of variances that appear in the prior distribution of the spline coefficients in $\boldsymbol{\gamma}$, and $\pi(\boldsymbol{\sigma}_e^2)\pi(\boldsymbol{\sigma}_d^2)\pi(\boldsymbol{\gamma}|\boldsymbol{\sigma}_e^2, \boldsymbol{\sigma}_d^2)$ is the prior distribution on the parameters (see appendix B for details). We can sample this distribution by the MCMC methods of Albert and Chib (1993), except for the new steps involving $\boldsymbol{\sigma}_e^2$ and $\boldsymbol{\sigma}_d^2$. These can be found in Chib and Greenberg (2010).

It is possible to incorporate model uncertainty relative to (say) the number of knots and/or ν in these calculations by fitting the above model for different number of knots or values of ν and then finding the marginal likelihood of each of those models, calculated by the method of Chib (1995). These can be used to find the posterior probabilities of the models which can then be used to model average the posterior distributions of SATE from each of the models.

4 Bayesian self-matching

An alternative, complementary strategy for calculating the SATE is to model the outcome distributions directly in a semi-parametric way. The semi-parametric modeling discussed here involves modeling the effect of the covariates through cubic splines with appropriately chosen knots and basis functions. These outcome distributions can be used to find the expectation of the missing counterfactual y_{1c}^* for each control intake subject, the expectation of the missing counterfactual y_{0t}^* for each treated subject, or to form predictive densities of these missing counterfactuals, as in Chib and Hamilton (2002) and Chib (2007). Differences between these expectations (or the predictions) and the observed outcome for each subject can then be calculated and averaged across subjects.

We show here that Bayesian self-matching in either its expectation or predictive forms is an effective approach for finding the SATE. We are able to mitigate the effect of possible covariate imbalance by determining the knots from the covariate values for both sets of subjects (see the second paragraph of Appendix A for details), although the fitting of the outcome distributions is based only on the subgroup of subjects in each category.

4.1 Outcome models

Consider first the control intake subjects. Assuming that the outcomes are real-valued, we write the outcome model for such subjects for all n observations as

$$\begin{pmatrix} \mathbf{y}_0 \\ \mathbf{y}_{0t}^* \end{pmatrix} = \mathbf{L}\boldsymbol{\beta}_{00} + g_{01}(\mathbf{w}_1) + \cdots + g_{0q}(\mathbf{w}_q) + \begin{pmatrix} \boldsymbol{\varepsilon}_0 \\ \boldsymbol{\varepsilon}_{0t} \end{pmatrix}, \tag{4.1}$$

where $\mathbf{y}_0 : n_0 \times 1$ and $\mathbf{y}_{0t}^* : n_1 \times 1$ are the quantities defined in Table 1, $\mathbf{L}$ is a $n \times k_l$ matrix of categorical covariates, the $g_{0r}(\cdot)$, $r = 1, \ldots, q$ are unknown functions, and the errors are independently distributed with a distribution F_0, which we take to be Student-t with dispersion σ_0^2 and ν_0 degrees of freedom. By modeling each of the $g_{0r}(\cdot)$ by a natural cubic spline with M_{r0} knots per function, we can write the outcome model as

$$\underbrace{\begin{pmatrix} \mathbf{y}_0 \\ \mathbf{y}_{0t}^* \end{pmatrix}}_{\mathbf{y}_c} = \underbrace{\begin{pmatrix} \mathbf{B}_0 \\ \mathbf{B}_{0t} \end{pmatrix}}_{\mathbf{B}_c} \boldsymbol{\beta}_0 + \begin{pmatrix} \boldsymbol{\varepsilon}_0 \\ \boldsymbol{\varepsilon}_{0t} \end{pmatrix}, \tag{4.2}$$

where the basis matrix $\mathbf{B}_c$ is $n \times k_0$, $k_0 = k_l + \sum_{r=1}^{q}(M_{r0} - 1)$, and given by $\mathbf{B}_c = (\mathbf{L}, \mathbf{B}_{01}, \ldots, \mathbf{B}_{0q})$, where each $\mathbf{B}_{0r} : n \times (M_{r0} - 1)$ is obtained by applying the spline transformation given in Appendix A and $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{00}, \boldsymbol{\beta}_{01}, \ldots, \boldsymbol{\beta}_{0q}) : k_0 \times 1$, with $\boldsymbol{\beta}_{0r}$ $(r = 1, \ldots, q)$ being the vector of spline coefficients corresponding to the basis matrix $\mathbf{B}_{0r}$.

If we now let $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \sigma_0^2, \boldsymbol{\sigma}_{0e}^2, \boldsymbol{\sigma}_{0d}^2)$ denote the parameters in this model distributed a priori as $\pi(\boldsymbol{\beta}_0, \sigma_0^2, \boldsymbol{\sigma}_{0e}^2, \boldsymbol{\sigma}_{0d}^2)$, where the prior distribution is specified by the approach described in Appendix B, then the posterior distribution $\pi(\boldsymbol{\beta}_0, \sigma_0^2, \boldsymbol{\sigma}_{0e}^2, \boldsymbol{\sigma}_{0d}^2, \boldsymbol{\lambda}_0 | \mathbf{y}_0, \mathbf{B}_0, \nu_0)$, after the augmentation with the Gamma distributed mixing variables $\boldsymbol{\lambda}_0 = (\lambda_1, \ldots, \lambda_{n_0})$, conditioned on the given outcomes $\mathbf{y}_0$, is proportional to

$$\pi(\boldsymbol{\beta}_0, \sigma_0^2, \boldsymbol{\sigma}_{0e}^2, \boldsymbol{\sigma}_{0d}^2) \times \mathcal{N}_{n_0}(\mathbf{y}_0 | \mathbf{B}_0\boldsymbol{\beta}_0, \sigma_0^2\boldsymbol{\Lambda}_0^{-1}) \prod_{i=1}^{n_0} \mathcal{G}\left(\lambda_i | \frac{\nu_0}{2}, \frac{\nu_0}{2}\right), \tag{4.3}$$

where $\boldsymbol{\Lambda}_0^{-1} = \mathsf{diag}\left(\lambda_1^{-1}, \ldots, \lambda_{n_0}^{-1}\right)$. This distribution is easily sampled by MCMC methods.

If the outcome is not real-valued, for instance if it is binary as in one of our examples below, then the modeling and fitting is similar to the case of the intake. For example, one way to proceed is to assume that

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

where the probability is 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 can be estimated by the method of Albert and Chib (1993) as in the discussion of the intake above.

In a manner analogous to that for the control intake subjects, we write the outcome model for the treatment intake group for all n observations in terms of a different set of unknown covariate functions as

$$\begin{pmatrix} \mathbf{y}^*_{1c} \\ \mathbf{y}_1 \end{pmatrix} = \mathbf{L}\beta_{10} + g_{11}(\mathbf{w}_1) + \cdots + g_{1q}(\mathbf{w}_q) + \begin{pmatrix} \boldsymbol{\varepsilon}_{1c} \\ \boldsymbol{\varepsilon}_1 \end{pmatrix}, \tag{4.4}$$

where $\mathbf{y}^*_{1c} : n_0 \times 1$ and $\mathbf{y}_1 : n_1 \times 1$ are defined in Table 1, and the errors are independently distributed with distribution F_1, say Student-t with dispersion σ_1^2 and ν_1 degrees of freedom. After modeling each of the functions $g_{1r}(\cdot)$ by a natural cubic spline with M_{r1} knots per function, the model has the form

$$\underbrace{\begin{pmatrix} \mathbf{y}^*_{1c} \\ \mathbf{y}_1 \end{pmatrix}}_{\mathbf{y}_t} = \underbrace{\begin{pmatrix} \mathbf{B}_{1c} \\ \mathbf{B}_1 \end{pmatrix}}_{\mathbf{B}_t}\boldsymbol{\beta}_1 + \begin{pmatrix} \boldsymbol{\varepsilon}_{1c} \\ \boldsymbol{\varepsilon}_1 \end{pmatrix}, \tag{4.5}$$

where the basis matrix $\mathbf{B}_t$ is $n \times k_1$, $k_1 = k_l + \sum_{r=1}^q (M_{r1} - 1)$, and given by $\mathbf{B}_t = (\mathbf{L}, \mathbf{B}_{11}, \ldots, \mathbf{B}_{1q})$, where each $\mathbf{B}_{1r}$ is an $n \times (M_{r1} - 1)$ matrix and $\boldsymbol{\beta}_1 = (\boldsymbol{\beta}_{10}, \boldsymbol{\beta}_{11}, \ldots, \boldsymbol{\beta}_{1q})$.

In this case, with $\boldsymbol{\lambda}_1 = (\lambda_{n_0+1}, \ldots, \lambda_n)$, the posterior distribution

$$\begin{aligned} &\pi(\boldsymbol{\beta}_1, \sigma_1^2, \boldsymbol{\sigma}_{1e}^2, \boldsymbol{\sigma}_{1d}^2, \boldsymbol{\lambda}_1 | \mathbf{y}_1, \mathbf{B}_1, \nu_1) \\ &\qquad \propto \pi(\boldsymbol{\beta}_1, \sigma_1^2, \boldsymbol{\sigma}_{1e}^2, \boldsymbol{\sigma}_{1d}^2) \times \mathcal{N}_{n_1}(\mathbf{y}_1 | \mathbf{B}_1\boldsymbol{\beta}_1, \sigma_1^2\boldsymbol{\Lambda}_1^{-1}) \prod_{i=n_0+1}^{n} \mathcal{G}\left(\lambda_i | \frac{\nu_1}{2}, \frac{\nu_1}{2}\right), \end{aligned} \tag{4.6}$$

where $\boldsymbol{\Lambda}_1^{-1} = \mathsf{diag}\left(\lambda_{n_0+1}^{-1}, \ldots, \lambda_n^{-1}\right)$. Just as for the controls, this distribution is sampled easily by MCMC methods.

Note that if the outcome is binary, as we discussed above, we assume that $\mathsf{Pr}(\mathbf{y}_t = 1|\boldsymbol{\theta}_1) = T_{\nu_1}(\mathbf{B}_t\boldsymbol{\beta}_1)$ and estimate the model as in the case of the intake.

4.2 Posterior distribution of self-matched SATE

From (4.2) and (4.5) it follows that $\mathbb{E}(\mathbf{y}_{0t}^*|\mathbf{y}_0, \mathbf{B}_c, \boldsymbol{\theta}_0) = \mathbf{B}_{0t}\boldsymbol{\beta}_0$ and $\mathbb{E}(\mathbf{y}_{1c}^*|\mathbf{y}_1, \mathbf{B}_t, \boldsymbol{\theta}_1) = \mathbf{B}_{1c}\boldsymbol{\beta}_1$, which implies that the SATE can be expressed as

$$\mathsf{SATE}(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1) = \mathsf{mean}\begin{pmatrix} \mathbf{B}_{1c}\boldsymbol{\beta}_1 - \mathbf{y}_0 \\ \mathbf{y}_1 - \mathbf{B}_{0t}\boldsymbol{\beta}_0 \end{pmatrix}, \tag{4.7}$$

where $\mathsf{mean}\,(\cdot)$ denotes the average of the components. We refer to this as the expectation self-matched SATE. To obtain a sample of draws from the posterior distribution of the Bayes self-matched expectation SATE, we can utilize the sample of draws from the MCMC simulations of (4.3) and (4.6), $\left\{\boldsymbol{\theta}_0^{(1)}, \ldots, \boldsymbol{\theta}_0^{(G)}\right\}$ and $\left\{\boldsymbol{\theta}_1^{(1)}, \ldots, \boldsymbol{\theta}_1^{(G)}\right\}$ and calculate

$$\mathsf{SATE}^{(g)} = \mathsf{mean}\begin{pmatrix} \mathbf{B}_{1c}\boldsymbol{\beta}_1^{(g)} - \mathbf{y}_0 \\ \mathbf{y}_1 - \mathbf{B}_{0t}\boldsymbol{\beta}_0^{(g)} \end{pmatrix}, \; g = 1, 2, \ldots, G. \tag{4.8}$$

Alternatively, following Chib and Hamilton (2002) and Chib (2007), we can define the self-matched SATE directly in terms of the missing counterfactuals to obtain the self-matched predictive SATE (PSATE)

$$\mathsf{PSATE}(\mathbf{y}_{1c}^*, \mathbf{y}_{0t}^*) = \mathsf{mean}\begin{pmatrix} \mathbf{y}_{1c}^* - \mathbf{y}_0 \\ \mathbf{y}_1 - \mathbf{y}_{0t}^* \end{pmatrix}. \tag{4.9}$$

We can sample the posterior distribution of the predictive SATE by sampling the missing counterfactuals by the method of composition. Given $\boldsymbol{\theta}_0^{(g)}$, we sample

$$\mathbf{y}_{0t}^{*(g)} \sim t_{n_1}(\mathbf{B}_{0t}\boldsymbol{\beta}_0^{(g)}, \sigma_0^{2(g)}),$$

where t_{n_1} denotes the independent Student-t distribution (denoted by the scalar dispersion parameter), and given $\boldsymbol{\theta}_1^{(g)}$, we sample

$$\mathbf{y}_{1c}^{*(g)} \sim t_{n_0}(\mathbf{B}_{1c}\boldsymbol{\beta}_1^{(g)}, \sigma_1^{2(g)})$$

Then

$$\mathsf{PSATE}^{(g)} = \mathsf{mean}\begin{pmatrix} \mathbf{y}_{1c}^{*(g)} - \mathbf{y}_0 \\ \mathbf{y}_1 - \mathbf{y}_{0t}^{*(g)} \end{pmatrix}, \; g = 1, 2, \ldots, G. \tag{4.10}$$

is a sample of draws from the posterior distribution of PSATE.

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

$$\mathsf{SATE}(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1) = \mathsf{mean}\begin{pmatrix} T_{\nu_1}(\mathbf{B}_{1c}\boldsymbol{\beta}_1) - \mathbf{y}_0 \\ \mathbf{y}_1 - T_{\nu_0}(\mathbf{B}_{0t}\boldsymbol{\beta}_0) \end{pmatrix} \tag{4.11}$$

under the (default) assumption that the outcomes are modeled through a Student-t link function, with the PSATE defined as above.

5 Bayes matching in randomized experiments

We now explain how the ideas developed above can be applied in a setting in which the CIC condition fails. Suppose that subjects are randomized into control and treatment arms, and let $z = 0$ denote assignment into the control arm and $z = 1$ assignment into the treatment arm. If the intake x is equal to z for each subject, we have the classic situation of a randomized experiment in which the intake is independent of the potential outcomes (y_0, y_1) and, importantly, independent of observed $(\mathbf{w})$ and unobserved (s) confounders. The CIC and overlap conditions obviously hold, and the SATE is identified and estimable without the need for matching or other sophisticated technique.

5.1 Compliance problem

In some settings, however, the intake x is not the same as the assignment z. This is called the compliance problem. In that case, there may be concern that the lack of compliance is due to confounders, both observed and unobserved. A number of studies, for example, Sommer and Zeger (1991), Imbens and Rubin (1997) and Frangakis and Rubin (2002), have emerged that deal with this case, exploiting the fact that assignment z can be treated as an instrument. We show how our methods can apply to this problem. We follow the notation and description of the model set-up given by Chib and Jacobi (2008).

Suppose that s is a discrete (unobserved) random variable that represents subject types such that intake is determined completely by z and s. In particular, assume that s can take 3 levels (with a 4th possible type, that of defiers, ruled out by assumption),

$$\begin{array}{ll} s = c, & c \ \text{ for complier} \\ s = n, & n \ \text{ for never taker} \\ s = a, & a \ \text{ for always taker}, \end{array}$$

where these attributes determine the intake in each arm of z. For example,

$$\begin{aligned} &\mathsf{Pr}(x = 0 | z = 0, s = c) = 1 \\ &\mathsf{Pr}(x = 0 | z = 0, s = n) = 1 \\ &\mathsf{Pr}(x = 1 | z = 1, s = a) = 1. \end{aligned}$$

Under this structure,

$$x \perp\!\!\!\perp (y_0, y_1, w) | z, s,$$

which means that there is no overlap problem, conditioned on (z, s). Naturally, because of the randomization of z, $s \perp\!\!\!\perp z$.

Now suppose that n subjects are randomly exposed to z. Then, categorized by assignment and intake, the available data and possible subject types can be arranged as in Table 2. As

	$x = 0$	$x = 1$
$z = 0$	$\mathbf{y}_{00}$, $\mathbf{W}_{00}$ (compliers, never-takers)	$\mathbf{y}_{01}$, $\mathbf{W}_{01}$ (always-takers)
$z = 1$	$\mathbf{y}_{10}$, $\mathbf{W}_{10}$ (never-takers)	$\mathbf{y}_{11}$, $\mathbf{W}_{11}$ (compliers, always-takers)

Table 2: Vectorized data structure and possible subject types by assignment and intake: $\mathbf{y}_{lk}$ and $\mathbf{W}_{lk}$ are of dimension $n_{lk} \times 1$ and $n_{lk} \times q$, respectively, $(l = 0, 1)$ and $(k = 0, 1)$ with $n_{00} + n_{01} + n_{10} + n_{11} = n$. The (0,0) cell can only have compliers and never-takers, the (0,1) cell only always-takers, the (1,0) cell only always-takers, and the (1,1) cell only compliers and always-takers.

before, the counterfactual outcomes are not observed. We denote the counterfactuals of interest by $\mathbf{y}^*_{1,00} : n_{00} \times 1$, which are the missing y_1 outcomes for the subjects in the (0,0) cell, and by $\mathbf{y}^*_{0,11} : n_{11} \times 1$, the missing y_0 outcomes for the subjects in the $(1, 1)$ cell. For convenience in indexing the sample subjects, we arrange the complete observed data as

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_{00} \\ \mathbf{y}_{11} \\ \mathbf{y}_{10} \\ \mathbf{y}_{01} \end{pmatrix}, \ \mathbf{W} = \begin{pmatrix} \mathbf{W}_{00} \\ \mathbf{W}_{11} \\ \mathbf{W}_{10} \\ \mathbf{W}_{01} \end{pmatrix}.$$

5.2 Sample Complier Average Treatment Effect

A causal effect of interest in this context is the complier average treatment effect

$$\mathsf{CATE} = \mathbb{E}(y_1|\mathbf{w}, s = c) - \mathbb{E}(y_0|\mathbf{w}, s = c).$$

which is the treatment effect for compliers. We focus on another quantity, the sample complier average treatment effect (SCATE), which is an average of CATE over the subjects that are compliers. Let $I_{00} = \{i : i \leq n_{00}\}$ be the indices of the subjects in the $(0, 0)$ cell and $I_{11} = \{i : n_0 + 1 \leq i \leq n_{00} + n_{11}\}$ be the indices of the subjects in the $(1, 1)$ cell. Also let $s_i = c$ indicate that the ith subject is an complier and define $C_{jj} = \{i : i \in I_{jj} \text{ and } s_i = c\}$ as the indices of the subjects in each diagonal cell that are compliers. The SCATE can then be defined as

$$\mathsf{SCATE} = \frac{|C_{00}|}{|C_{00}| + |C_{11}|} \frac{1}{|C_{00}|} \left(\sum_{i \in C_{00}} \left(\mathbb{E}(y^*_{1,00i}|\mathbf{w}_{00i}, s_i = c) - \mathbb{E}(y_i|\mathbf{w}_{00i}, s_i = c) \right) \right)$$

$$+\frac{|C_{11}|}{|C_{00}|+|C_{11}|}\frac{1}{|C_{11}|}\left(\sum_{i\in C_{11}}\left(\mathbb{E}(y_i|\mathbf{w}_{11i},s_i=c)-\mathbb{E}(y^*_{0,11i}|\mathbf{w}_{11i},s_i=c)\right)\right), \quad (5.1)$$

where $|C_{jj}|$ is the cardinality of the set C_{jj}.

5.3 Bayesian modeling and posterior distribution

The posterior distribution of SCATE can be found either by Bayesian matching on compliance probabilities or by Bayesian self-matching. For simplicity and to focus on the main point, we discuss the ideas in a parametric Gaussian framework, which can be extended in the semi-parametric direction by modeling covariate effects by splines and letting the outcome distributions be Student-t, as in the discussion above.

Assume that the outcome is continuous. Let $\mathbf{s}_{lj}$ be an n_{lj} vector that indicates the type of each subject in cell (l,j). Let $\boldsymbol{\theta}$ denote all the parameters in the model. Then, following Chib and Jacobi (2008), the Bayesian model of outcomes and types in each of the four cells is defined in terms of four Gaussian distributions as follows:

$$p(\mathbf{y}_{00},\mathbf{s}_{00}|\mathbf{W}_{00},\boldsymbol{\theta})=\prod_{i\in I_{00}}\mathcal{N}(y_{00i}|\mathbf{w}'_{00i}\boldsymbol{\beta}_{0c},\sigma^2_{0c})I\left[s_i=c\right]q_c+\mathcal{N}(y_{00i}|\mathbf{w}'_{00i}\boldsymbol{\beta}_{n},\sigma^2_{n})I\left[s_i=n\right]q_n$$

$$p(\mathbf{y}_{11},\mathbf{s}_{11}|\mathbf{W}_{11},\boldsymbol{\theta})=\prod_{i\in I_{11}}\mathcal{N}(y_{11i}|\mathbf{w}'_{11i}\boldsymbol{\beta}_{1c},\sigma^2_{1c})I\left[s_i=c\right]q_c+\mathcal{N}(y_{11i}|\mathbf{w}'_{11i}\boldsymbol{\beta}_{a},\sigma^2_{a})I\left[s_i=a\right]q_a$$

$$p(\mathbf{y}_{10},\mathbf{s}_{10}|\mathbf{W}_{10},\boldsymbol{\theta})=\prod_{i\in I_{10}}\mathcal{N}(y_{10i}|\mathbf{w}'_{10i}\boldsymbol{\beta}_{n},\sigma^2_{n})I\left[s_i=n\right]q_n$$

$$p(\mathbf{y}_{01},\mathbf{s}_{01}|\mathbf{W}_{01},\boldsymbol{\theta})=\prod_{i\in I_{01}}\mathcal{N}(y_{01i}|\mathbf{w}'_{01i}\boldsymbol{\beta}_{a},\sigma^2_{a})I\left[s_i=a\right]q_a,$$

where $q_k=\mathsf{Pr}(s_i=k)$, $k\in\{c,n,a\}$, are the probabilities of type k that sum to one.

This model can be easily summarized by Bayesian MCMC methods after including $\{\mathbf{s}_{lj}\}$ in the sampling. One places a prior distribution $\pi(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$, for example, a product of independent normal distributions on $(\boldsymbol{\beta}_{0c},\boldsymbol{\beta}_{1c},\boldsymbol{\beta}_{n},\boldsymbol{\beta}_{a})$, a product of inverse-gamma distributions on $(\sigma^2_{0c},\sigma^2_{1c},\sigma^2_{n},\sigma^2_{a})$, and a Dirichlet distribution on (q_c,q_n,q_a), and then samples the posterior distribution of the parameters and types

$$\pi(\boldsymbol{\theta},\{\mathbf{s}_{lj}\}|\mathbf{y},\mathbf{W})\propto\pi(\boldsymbol{\theta})\prod_{l=0,1}\prod_{j=0,1}p(\mathbf{y}_{lj},\mathbf{s}_{lj}|\mathbf{W}_{lj},\boldsymbol{\theta}) \quad (5.2)$$

by standard MCMC methods. Because the mixture distributions are present only in the diagonal cells, the sampling of this posterior does not suffer from the label switching problem that arises in general finite mixture models.

5.4 Posterior of SCATE by Bayesian matching

One can do Bayesian matching in the course of the MCMC sampling of the posterior distribution in (5.2) by matching on compliance probabilities in the following way: In each MCMC iteration and for each subject in the diagonal cells, the conditional posterior probabilities

$$\mathsf{Pr}(s_i = c|\mathbf{y}, \mathbf{W}, \theta) = \frac{q_c \mathcal{N}(y_{00i}|\mathbf{w}'_{00i}\boldsymbol{\beta}_{0c}, \sigma^2_{0c})}{q_c \mathcal{N}(y_{00i}|\mathbf{w}'_{00i}\boldsymbol{\beta}_{0c}, \sigma^2_{0c}) + q_n \mathcal{N}(y_{00i}|\mathbf{w}'_{00i}\boldsymbol{\beta}_{n}, \sigma^2_{n})} \ , \ i \in I_{00},$$

and

$$\mathsf{Pr}(s_i = c|\mathbf{y}, \mathbf{W}, \theta) = \frac{q_c \mathcal{N}(y_{11i}|\mathbf{w}'_{11i}\boldsymbol{\beta}_{1c}, \sigma^2_{1c})}{q_c \mathcal{N}(y_{11i}|\mathbf{w}'_{11i}\boldsymbol{\beta}_{1c}, \sigma^2_{1c}) + q_a \mathcal{N}(y_{11i}|\mathbf{w}'_{i}\boldsymbol{\beta}_{a}, \sigma^2_{a})} \ , \ i \in I_{11},$$

that are needed to sample subject types can also be used to match subjects. In particular, in every iteration, each subject in the $(0, 0)$ cell can be matched on compliance probabilities with a subject in the $(1, 1)$ cell, and then the difference in outcomes

$$d_{00i} = (y_{11i^*} - y_{00i}) \ , \ i = 1, 2, \ldots, n_{00},$$

can be computed, where $i^* \in I_{11}$ is the subject matched to subject $i \in I_{00}$. Similarly, in every MCMC iteration, each subject in the $(1, 1)$ cell can be matched on the above compliance probabilities with a subject in the $(0, 0)$ cell, leading to the difference in outcomes

$$d_{11i} = (y_{11i} - y_{00i^*}) \ , \ i = n_{00} + 1, \ldots, n_{00} + n_{11}.$$

The mean value of these $n_{00} + n_{11}$ differences is the Bayes drawing of the compliance matched SCATE at each MCMC iteration. The sequence of these SCATE values across MCMC iterations can be used to find the posterior Bayes estimate of SCATE and various quantiles of the posterior distribution.

5.5 Posterior of SCATE by self-matching

The posterior distribution of SCATE by self-matching can be found unhindered by any concern over the overlap problem. The basic idea is that the terms in (5.1) can be calculated directly for the subjects that are currently classified as compliers because the type s_i is sampled in each MCMC iteration. For instance, for subjects in the $(0, 0)$ cell that are compliers, we can express the expectation of the missing counterfactual as

$$\mathbb{E}(y^*_{1,00i}|\mathbf{w}_{00i}, s_i = c) = \mathbb{E}(y^*_{1,00i}|\mathbf{w}_{00i}, s_i = c, z_i = 0),$$

because z is independent of the potential outcomes given type. But, we know that intake is independent of the covariates $\mathbf{w}$ conditioned on type and assignment. The latter expectation can therefore be calculated as $\mathbf{w}'_{00i}\boldsymbol{\beta}_{1c}$ from the model of y_1 in the $(1,1)$ cell for compliers without any concern about the overlap problem. The same argument can be used to deal with $\mathbb{E}(y^*_{0,11i}|\mathbf{w}_{11i}, s_i = c)$. Thus, given $(\boldsymbol{\beta}_{0c}, \boldsymbol{\beta}_{1c}, \mathbf{s}_{00}, \mathbf{s}_{11})$, after calculating $C_{jj} = \{i : i \in I_{jj} \text{ and } s_i = c\}$, $j = 0, 1$, we can compute SCATE as

$$\mathsf{SCATE}\left(\boldsymbol{\beta}_{0c}, \boldsymbol{\beta}_{1c}, \mathbf{s}_{00}, \mathbf{s}_{11}\right) = \frac{1}{|C_{00}| + |C_{11}|}\left(\sum_{i \in C_{00}}(\mathbf{w}'_{00i}\boldsymbol{\beta}_{1c} - y_{00i}) + \sum_{i \in C_{11}}(y_{11i} - \mathbf{w}'_{11i}\beta_{0c})\right). \tag{5.3}$$

This function is evaluated in every MCMC iteration, given the current drawing of the parameters and types. The sequence of these SCATE values across MCMC iterations is a sample from the posterior distribution of the self-matched SCATE.

6 Simulated data example

6.1 CIC data example

We begin our illustration of the proposed methods 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 CIC holds. The simulation design ensures that the overlap problem is non-trivial. We then describe a small model search through marginal likelihoods and Bayes factors to find the best model under that criterion and use that model to implement our matching strategies. We consider sample sizes of 500, 1,000, 2,000, and 4,000 subjects.

6.1.1 Design

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

$$l_1 \sim \mathcal{B}(.2)\ ,\ l_2 \sim \mathcal{B}(.4)\ ,\ l_3 \sim \mathcal{B}(.5)\ ,$$

and the two continuous confounders as uniform random-variables

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

Intake is generated for each subject according to the model

$$\mathsf{Pr}(x = 1|\mathbf{l}, \mathbf{w}) = T_5\left(-.5 + .2l_1 - .3l_2 + .5l_3 + g_1(w_1) + g_2(w_2)\right),$$

where

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

and the potential outcomes according to

$$y_0 = 1 + .1l_1 + .7l_2 - .2l_3 + g_{01}(w_1) + g_{02}(w_2) + \varepsilon_0$$
$$y_1 = 1 - .1l_1 + .5l_2 - .6l_3 + g_{11}(w_1) + g_{12}(w_2) + \varepsilon_1,$$

where

$$g_{01}(w_1) = w_1 + w_1^5 \ , \ g_{02}(w_2) = \sin\left(2\pi (1 - w_2)^2\right) \ ,$$
$$g_{11}(w_1) = 5w_1 + 8w_1^4 \ , \ g_{12}(w_2) = \frac{1}{w_2 + .1} + 8\exp\left(-400 (w_2 - .5)^2\right);$$

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 intake 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 intake

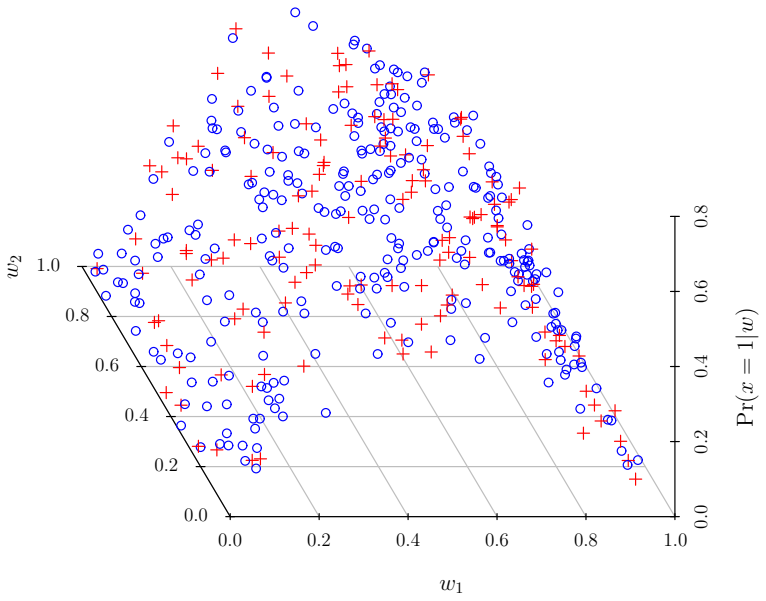

Figure 1: Plot of the true propensity score with simulated data ($n = 500$) at the generated values of the confounders. The control intake observations are marked in circles and the treatment intake observations with pluses.

subjects are indicated by circles and treatment intake 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, which generates fewer treatment intake observations at each end of the w_1 interval. A consequence of this feature is that when a control subject with a small value of w_1, for example, is matched with a treatment intake person on the propensity score, it is quite likely that that subject would be from the right end of the w_1 interval and, therefore, well matched on the propensity score but not well matched on w_1. This problem tends to become less important as the sample size increases because there will be more treatment intake subjects for small values of w_1 that can form matches. Because of this feature, as we show below, the posterior distribution of the SATE from Bayes propensity score matching is bimodal when $n = 500$, indicating the presence of two different centers of the SATE depending on whether or not propensity score matching results in a close match with w_1.

Another aim of this design is to produce a non-trivial overlap problem. Again focusing on the $n = 500$ sample, this problem can be observed from the contour plots of the (w_1, w_2) distribution, by intake group shown in Figure 2. The left plot in the figure, which has the

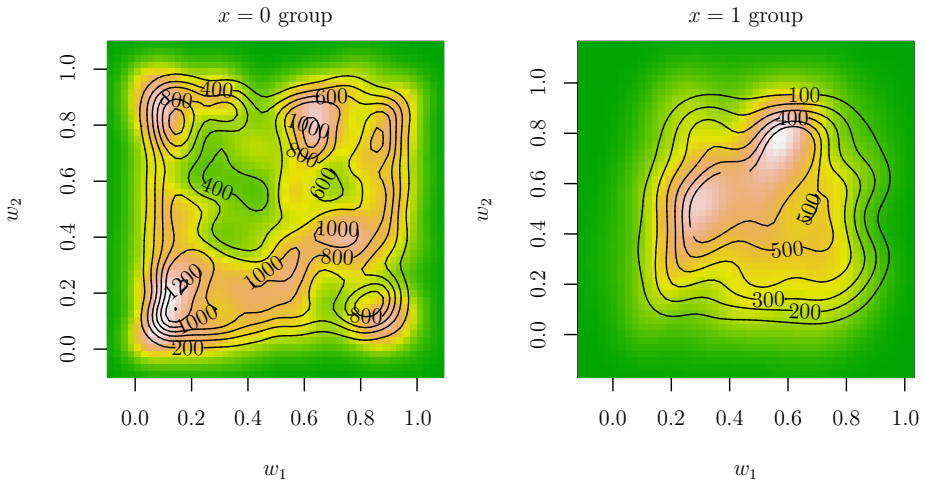

Figure 2: Contour plot of the distribution of (w_1, w_2) by intake group in the simulated data ($n = 500$); this shows the severe overlap problem.

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, with clear regions of limited overlap with the first distribution.

6.1.2 Model fitting

The prior distribution in (B.4) and (B.5) requires specifying the prior on the coefficients of the intercept and the categorical variables and the priors on $\sigma_e^2 = (\sigma_{e1}^2, \sigma_{e2}^2)$ and $\sigma_d^2 = (\sigma_{d1}^2, \sigma_{d2}^2)$. (Although these parameters are model specific, this feature is suppressed in the notation.) We set this prior to be same across the intake and potential outcome models and across models with different degrees of freedom by assuming that the linear coefficients are centered at zero with a prior variance of 10, that the distribution of each element of σ_e^2 has a prior mean of 20 and standard deviation of 10, and that each element of σ_d^2 has a prior mean of 0.5 and standard deviation of 1. The prior on the spline coefficients is determined within the algorithm as explained in Appendix B.

Our results are based on 20,000 MCMC draws following a burn-in of 2,500 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 small and mostly less than 2 or 3, indicating that the sampling procedures are highly efficient.

We incorporate a small model search as part of our fitting of the intake and outcome models by considering models of the intake and the outcomes 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 shown in Table 3 for the intake model equation reveal that the combination of (15, 15) knots is slightly preferred over (10, 10) to approximate g_1 and g_2, that 10 knots are sufficient for $g_{1,1}$, 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 appears for values of w_2 between 0.4 and 0.6. We consider only the combination of (10, 10) knots for the y_0 equation. The results that follow are based on the models for x and y_1 that yield the largest ML value. In Figure 3 we show the true

Knots		Sample Size			
w_1	w_2	500	1,000	2,000	4,000
		x Equation			
10	10	-129.600	-260.286	-481.430	-978.455
15	15	-128.990	-259.719	-480.328	-978.052
		y_1 Equation			
10	10	-145.390	-289.119	-554.895	-1005.262
10	15	-126.551	-230.655	-437.373	-751.991

Table 3: Marginal likelihoods (log base 10) for various knot combinations, simulated data

functions and estimated functions for sample sizes of 500 and 4,000. The functions are well

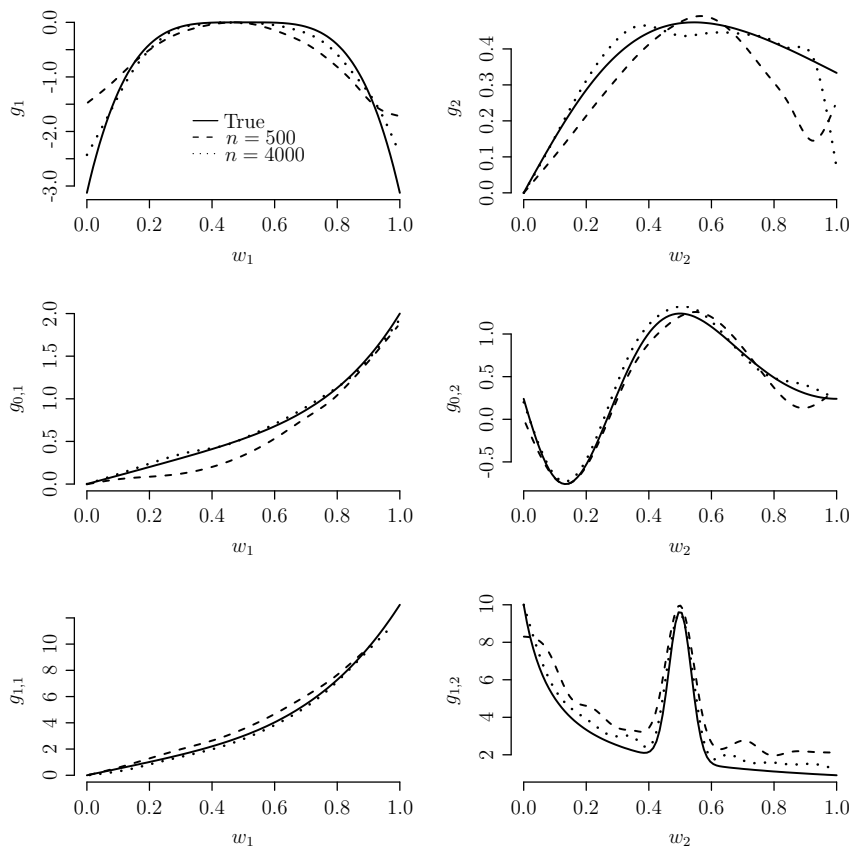

Figure 3: True and estimated functions in the intake (top row) and outcome models (second and third rows) with simulated data and sample sizes 500 and 4000. The intercepts of the estimates are shifted to coincide roughly with the maximum value of the true function.

estimated for all models and both sample sizes, and, as expected, the fit is better for the larger sample size.

6.1.3 Posterior distribution of SATE

We first consider the estimation of SATE by versions of frequentist matching. The estimates are obtained from the GenMatch package in R. We report results for (a) propensity score matching based on propensity scores from a logit link and linear covariate effects (b) propensity score matching based on propensity scores from a logit link and covariates effects estimated by the generalized additive model function in R and (c) matching based on distance between covariates across subjects measured in terms of the Mahanalobis measure. The latter method method is less commonly used because it is does not generalize easily to a large number of confounders or to cases with many categorical confounders. Nonetheless, it can be applied here. The results

are reported in the top-half of Table 4.

As a simple criterion for accuracy, we determine whether the estimate $\pm$ two standard deviations includes the true value. According to this criterion, none of intervals based on method (a) cover the true value. Intervals derived from method (b) contain the true value for one of the sample sizes, and those based on method (c) contain the true value for three of the four sample sizes.

	Sample size			
	500	1000	2000	4000
True value	5.913	5.876	6.049	5.969
Frequentist matching				
Logit (linear effects)	5.174	5.404	5.457	5.431
	(0.271)	(0.209)	(0.139)	(0.104)
Logit (non-linear effects)	5.360	5.616	5.324	5.361
	(0.276)	(0.204)	(0.141)	(0.102)
Mahalanobis	5.399	6.067	5.364	5.706
	(0.325)	(0.285)	(0.193)	(0.144)
Bayesian matching				
Propensity score	5.545	6.194	5.810	5.876
	(0.559)	(0.264)	(0.270)	(0.241)
Expectation SATE	5.763	5.899	5.992	5.957
	(0.091)	(0.053)	(0.043)	(0.024)
Predictive SATE	5.753	5.892	5.991	5.960
	(0.122)	(0.075)	(0.058)	(0.035)

Table 4: True and estimated values of SATE (standard deviations in parentheses) by three different frequentist matching methods (as explained in the text) and by Bayes propensity score matching, Bayes expectation SATE and Bayes predictive SATE

Summary results for the Bayesian calculations are contained in the second section of Table 4. As can be seen, in this example, the Bayesian matched estimates are more accurate than the frequentist estimates. For all four sample sizes and all three methods, the Bayesian estimated value $\pm$ 2 standard deviations includes the true value. As can be seen from Figure 4, the posterior distribution of the Bayes propensity score matched SATE is bimodal when $n = 500$, picking up the two possible centers of the SATE that depend on whether or not the matched propensity scores have a similar value of w_1. For the larger sample sizes, with more treated subjects across the support of the w_1 distribution, this bi-modality disappears. It may be reiterated that in the self-matching case it is essential to place the knots as described in Appendix

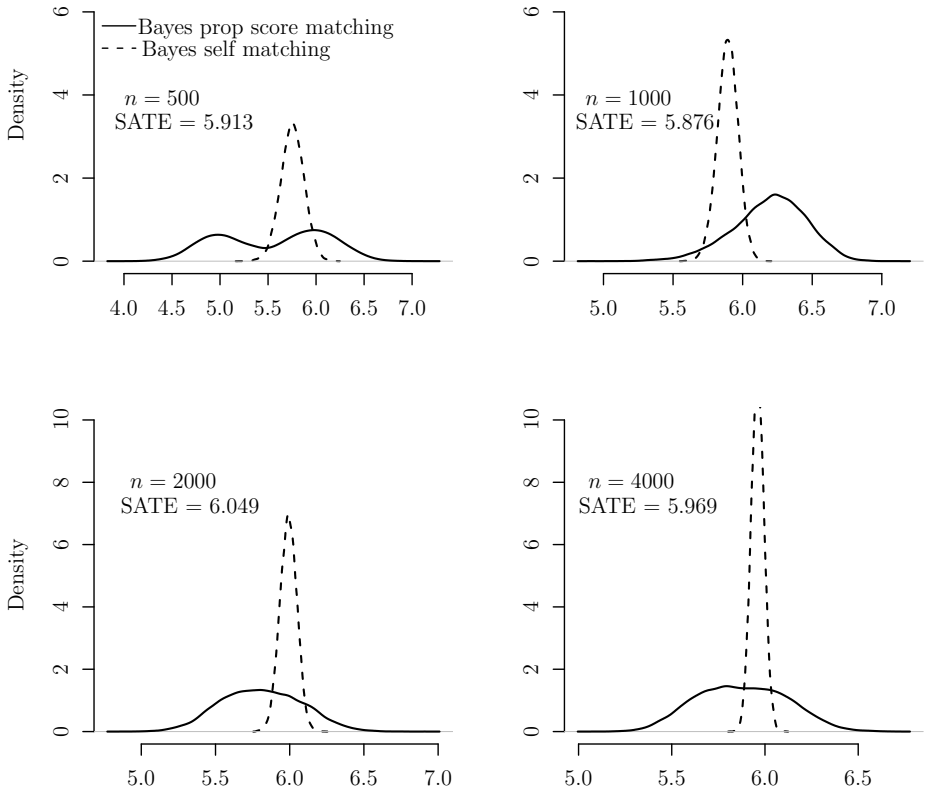

Figure 4: Posterior distribution of Bayes propensity score matched SATE and posterior distribution of PSATE with simulated data by sample size.

A. Otherwise, because of the overlap problem the functions estimates do not extrapolate well over the regions with limited overlap. However, under our approach, even though the data come from a complicated design, the posterior distribution of the PSATE centers quickly on the true value and becomes more concentrated around that value as the sample size increases.

6.2 Simulated Compliance Data

To illustrate our method in the setting of a randomized trial with a compliance problem, we consider the small $(n = 10)$ data set without covariates that is given in Table 5. In the generation process, the expected value of y_0 for the generated type in column 3, is assumed to be $(66, 63, 59, 57, 54, 53, 51, 51, 42, 39)$. The corresponding y_0 for the subjects with these types is equal to this expectation plus a standard normal error, and the corresponding y_1 is this expectation plus 10 plus 2 times a standard normal error. The true value of CATE is therefore 10.

It is easy to check that the naive estimate of the causal effect for these data, $\mathbb{E}(y|x = 1) -$

Subject	z	s	x	y_0	y_1	y
1	0	c	0	56.038	69.226*	56.038
2	0	c	0	53.637	67.706*	53.637
3	0	c	0	50.137	63.110*	50.137
4	0	n	0	42.057	48.959*	42.057
5	0	a	1	63.617*	74.942	74.942
6	1	n	0	38.663	47.597*	38.663
7	1	c	1	59.549*	68.786	68.786
8	1	c	1	54.528*	62.767	62.767
9	1	c	1	51.003*	58.228	58.228
10	1	a	1	67.593*	75.383	75.383

Table 5: Simulated data from a randomized trial with compliance modeled by subject types. Starred values are not observed. The true CATE is 10.

$\mathbb{E}(y|x=0)$, is 19.915 and the so-called intention-to-treat effect, given by $\mathbb{E}(y|z=1)-\mathbb{E}(y|z=0)$, is 5.403, which are both incorrect. The ratio of these effects, the instrumental variable estimator of the CATE, is 9.006, which is closer to the correct answer.

For our Bayesian analysis, we assume the form of $p_j(y|w,s=k)$, place a prior on the unknown parameters, sample the resulting posterior by MCMC methods, and calculate the two Bayesian matching effects discussed above. In particular, we assume that

$$\begin{aligned}
p_0(y|w,s=c) &= \mathcal{N}(y|\beta_{0c},\sigma^2_{0c}) \\
p_0(y|w,s=n) &= \mathcal{N}(y|\beta_{n},\sigma^2_{n}) \\
p_1(y|w,s=c) &= \mathcal{N}(y|\beta_{1c},\sigma^2_{1c}) \\
p_1(y|w,s=a) &= \mathcal{N}(y|\beta_{a},\sigma^2_{a}),
\end{aligned}$$

and for the prior we assume that

$$\beta_k \sim \mathcal{N}(55, 15^2)\ ,\ \sigma^2_k \sim \mathcal{IG}(\text{mean}=5, \text{sd}=20)\ ,\ q \sim \mathcal{D}(2,1,1).$$

From 20,000 MCMC iterations, following a burn-in of 2,500, our Bayes estimate of SCATE by matching on compliance probabilities is 9.114 and that from self-matching is 9.102. The quantiles of the posterior distribution of SCATE from each of these matching methods is given in Table 6. Although each posterior is centered around the same value, the distribution of the self-matching estimator is more dispersed. Even for such a small sample, when one might have concerns about an adverse impact from the prior, Bayes matching works well.

		Quantiles				
	Mean	.05	.25	.50	.75	.95
Bayes compliance matching	9.114	7.119	8.457	9.041	9.843	11.354
Bayes self-matching	9.102	5.628	7.525	9.122	10.477	12.788

Table 6: Posterior quantiles of the Bayes matched SCATE by Bayes compliance matching and Bayes self-matching

7 Real data example: RHC data

7.1 Background

We now apply our methods to a real data set in which the outcome is binary, several categorical and continuous confounders are available, and the CIC condition is assumed to hold. The data are concerned with the effect of a procedure called right heart catheterization, a diagnostic tool, on life expectancy. The data set we use was 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. The RHC aspect of the data was subjected to a propensity score analysis by Connors et al. (1996). The data can be found at lib.stat.cmu.edu/S/Harrell/data/descriptions/rhc.html.

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

We utilize 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. These variables comprise the matrix $\mathbf{L}_0$ as defined above. We also include 16 continuous variables whose effects are each modeled by a cubic spline. These variables measure a variety of physical measurements taken at the time of admission into the hospital as well as additional information about the patient taken at that time. After dropping some variables and patients because of data problems, our final sample contains 3,515 controls and 2,163 treated.

Our intake and outcome models are each specified through Student-*t* links with 5 degrees of freedom. The effects of the continuous confounders are modeled by cubic splines with different numbers of knots. The appropriate number of knots is then determined through the computation of marginal likelihoods and Bayes factors, which are computed by the method of Chib

(1995), from the output of the MCMC simulations. In our model formulation, all the available variables are included in the intake and y_0 models but 5 continuous variables are omitted from the y_1 specification because these were not important from the marginal likelihood-Bayes factor perspective.

7.2 Marginal likelihoods

Our first set of results, on the marginal likelihoods of the various models, appear in Table 7. These results are based on a prior distribution of $\boldsymbol{\beta}_0$ that is $\mathcal{N}_{47}(\mathbf{0}, 10 \times \mathbf{I}_{47})$, an inverted gamma prior for each σ^2_{ej} with a mean of 1 and standard deviation of 10, and an inverted gamma prior for each each σ^2_{dj} with a mean of 1 and standard deviation of 2. For the x and y_0 functions,

	Number of knots						
	5	6	7	10	12	K_1	K_2
x	-1447.342	-1457.089	-1464.326				
y_0	-929.368	-931.898	-934.400				
y_1	-681.505			-664.669	-650.513	-649.845	-654.625

Table 7: Log (base ten) marginal likelihoods for selected intake and outcome models, RHC data, categorized by number of knots - common to each function in the model - each in models with Student-t link and $\nu = 5$ degrees of freedom. Models with K_1 and K_2 knots have varying number of knots for each function as detailed in the text.

the marginal likelihoods in the table show that 5 knots are sufficient for each of the 16 cubic splines in the intake model and for each of the 16 cubic splines in the y_0 model. For the y_1 model, however, it appears that each of the 11 functions is better modeled by a different number of knots. For instance, the choice $K_1 = (5, 7, 5, 7, 7, 7, 7, 7, 7, 7, 5)$, which means that 5 knots were used for the first function and 7 for the second function, and so on, is preferred to one with the choice of $K_2 = (5, 9, 5, 9, 9, 9, 9, 9, 9, 9, 5)$ knots.

7.3 Function estimates

We display our posterior estimates of the functions in the intake model in Figure 5. To conserve space, we omit the estimates of the functions in the outcome models. The figures show considerable nonlinearities in the effect of the continuous variables on the intake and outcome functions. Figures not presented also show some differences in the effects of the covariates on the outcome variables. 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.

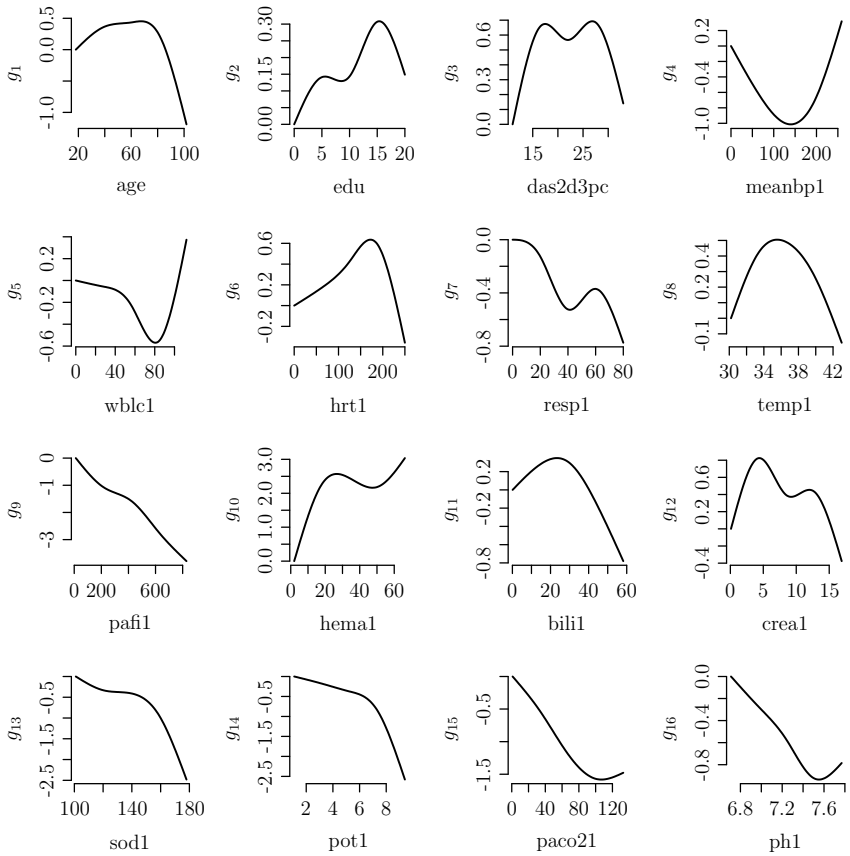

Figure 5: Cubic spline estimates of functions in intake model, RHC data, Student-t link (5 degrees of freedom), 5 knots for each function

7.4 Distributions of SATE

Summaries of the posterior distribution of SATE appear in Table 8, where we present the mean and selected quantiles of the posterior distributions from both forms of Bayesian matching across the various fitted models. The mean SATE values vary from 0.051 to 0.075. In contrast, a propensity score matching approach from the R-package GenMatch yields an estimated SATE of 0.04.

The posterior distributions of SATE from the models with the highest marginal likelihood are given in Figure 6. With little or no mass in the negative region, and with the posterior distribution from self-matching even further away from zero, these distributions support and reinforce the claim that the RHC procedure was not helpful.

Knots	Mean	sd	.025	.975
Bayes propensity score matching				
5	0.054	0.016	0.024	0.088
6	0.051	0.017	0.020	0.089
7	0.054	0.017	0.022	0.090
Bayes self-matching				
5	0.060	0.010	0.041	0.079
10	0.060	0.010	0.042	0.079
12	0.061	0.009	0.042	0.079
K_1	0.074	0.009	0.057	0.093
K_2	0.075	0.009	0.056	0.093

Table 8: Summary of the posterior distributions of SATE (mean, standard deviation, .025 and .975 quantiles), RHC data, from Bayes propensity score and Bayes self-matching methods, across the various intake and potential outcome models defined by the number of knots (indicated in the rows) used in the fitting of the functions.

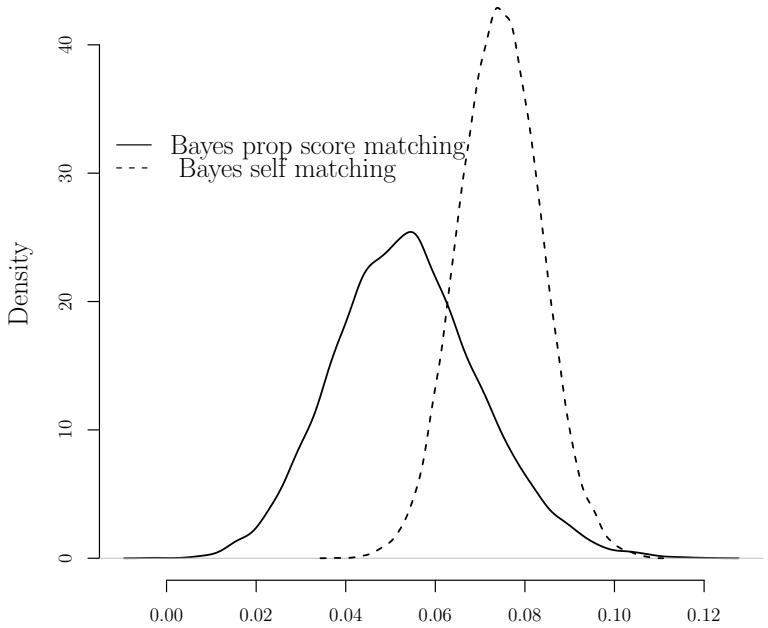

Figure 6: Posterior distributions of SATE in the RHC data from each of the Bayesian matching methods and the best fitting (in terms of the marginal likelihoods) intake and outcome models.

8 Conclusion

In this paper we provide a discussion of matching methods from a Bayesian perspective. The methods we supply are easy to implement and do not rely on strong modeling assumptions. One method we discuss is a Bayesian variant of the widely used propensity score matching

method. We show how it possible to find the posterior distribution of the SATE under weak semiparametric assumptions regarding the model of the propensity score. We also develop a second version of matching, Bayesian self-matching, and show how it can be used to provide another Bayesian view of the SATE. Although our methods have been illustrated in the setting of cubic splines, they can be easily modified or extended to another semi-parametric setting.

We also explain how the Bayesian matching ideas can be applied to randomized trials with a compliance problem. In this case, we show that how can match on the compliance probabilities that are calculated as part of the MCMC sampling procedure. Alternatively, we one can find the posterior distribution of SCATE by self-matching, without any concerns about the overlap problem.

We believe that the Bayesian perspective on matching methods developed here is illuminating and opens up new possibilities for causal inference by matching methods. Due to the simplicity and effectiveness of these methods, it is possible that causal inference by Bayesian matching methods will be of much general interest.

APPENDIX

A Construction of the cubic spline basis matrix

Let $f(\mathbf{w}) = (f(w_1), \ldots, f(w_n))$ denote the unknown function values 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 comes from Chib and Greenberg (2010) though the identification scheme on the spline coefficients presented here is simpler. The choice of knots $\tau = (\tau_1, \ldots, \tau_M)$ is also different to deal with the particular facets of the problem in this paper.

If the function is in the intake model, we let $\tau_1 = \mathsf{min}(\mathbf{w})$ and $\tau_M = \mathsf{max}(\mathbf{w})$ and let the remaining knots be equally spaced. If the function is in one of the outcome models, then equally spaced knots are unsatisfactory if the covariate has an overlap problem across the two groups of subjects. Then, equally spaced knots based on the control observations would miss some of the covariate values in the treated group. In that case, extrapolation of the spline estimated from the control observations to those covariate values in the treated group will be inaccurate. This problem cannot be avoided by calculating the equally spaced knots from the entire n covariate values because then there can be intervals in the control and treatment groups with no observations. This would make the basis matrices $\mathbf{B}_0$ and $\mathbf{B}_1$ unstable. Therefore, to mitigate the effect of the overlap problem when predicting $y^*_{1i}|\mathbf{w}_{0i}$ for the controls and $y^*_{0i}|\mathbf{w}_{1i}$ for the treated, we suggest the following procedure. We let the first knot τ_1 be $\mathsf{min}(w_1, \ldots, w_n)$,

and the last knot τ_M be $\max(w_1, \ldots, w_n)$, where the min and max are based on all the available observations on $\mathbf{w}$. We then define

$$\begin{aligned}\text{maxmin} &= \max(\min(\mathbf{w}_0), \min(\mathbf{w}_1)) \\ \text{minmax} &= \min(\max(\mathbf{w}_0), \max(\mathbf{w}_1)),\end{aligned}$$

and let $a = (\text{maxmin}, a_2, \ldots, a_{M-1}, \text{minmax})$ be M evenly spaced values between maxmin and minmax. The knots are then given by

$$\tau = (\tau_1, a_2, \ldots, a_{M-1}, \tau_M).$$

If there is an overlap problem, 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 overlap problem or only a minor overlap problem, these knots will be very nearly evenly spaced between the smallest and largest values of $\mathbf{w}$.

We now show how to express

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

where $\mathbf{B}_w$ is a $n \times (M-1)$ matrix and $\boldsymbol{\beta}_w$ is a $(M-1)$ vector of cubic spline 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_m and Ψ_m, $m = 1, 2, \ldots, M$, that have compact support and are given by

$$\Phi_m(a) = \begin{cases} 0, & a < \tau_{m-1}, \\ -(2/h_m^3)(a - \tau_{m-1})^2(a - \tau_m - 0.5h_m), & \tau_{m-1} \le a < \tau_m, \\ (2/h_{m+1}^3)(a - \tau_{m+1})^2(a - \tau_m + 0.5h_{m+1}), & \tau_m \le a < \tau_{m+1}, \\ 0, & a \ge \tau_{m+1}, \end{cases} \tag{A.1}$$

$$\Psi_m(a) = \begin{cases} 0, & a < \tau_{m-1}, \\ (1/h_{mr}^2)(a - \tau_{m-1})^2(a - \tau_m), & \tau_{m-1} \le a < \tau_m, \\ (1/h_{m+1,r}^2)(w - \tau_{m+1})^2(a - \tau_m), & \tau_m \le a < \tau_{m+1}, \\ 0, & a \ge \tau_{m+1}, \end{cases} \tag{A.2}$$

where $h_m = \tau_m - \tau_{m-1}$ is the spacing between the $(m-1)$st and mth knots. 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 $\omega_m = h_m/(h_m + h_{m+1})$, $\mu_m = 1 - \omega_m$, 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 $\mathbf{b}_m \in \Re^n$ is the mth column of $\mathbf{B}^\dagger$. Then, after dropping the first column for identification purposes, the cubic spline basis matrix is given by

$$\mathbf{B}_w = (\mathbf{b}_2, \ldots, \mathbf{b}_M).$$

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 minus the value at the first knot, i.e.,

$$\boldsymbol{\beta}_w = \begin{pmatrix} f(\tau_2) - f(\tau_1) \\ \vdots \\ f(\tau_M) - f(\tau_1) \end{pmatrix}.$$

This property of the spline coefficients is particularly helpful in the Bayesian context because it can be used in the formulation of the prior distribution. Also notice that the matrices that have to be inverted above (and in the prior distribution to follow) are conveniently banded.

B Prior distribution

Our prior on the spline coefficients incorporates an assumption of smoothness. This is done by requiring that differences in the ordinates at the end knots, and the differences in slopes between adjacent knots at the interior knots, have a prior mean of 0. In particular, for the end knots we assume that

$$\begin{aligned} \frac{f(\tau_2) - f(\tau_1)}{h_2} &\sim \mathcal{N}(0, \sigma_e^2) \\ \frac{f(\tau_M) - f(\tau_{M-1})}{h_M} &\sim \mathcal{N}(0, \sigma_e^2), \end{aligned} \tag{B.1}$$

and for the interior knots that

$$\frac{f(\tau_{m+1}) - f(\tau_m)}{h_{m+1}} - \frac{f(\tau_m) - f(\tau_{m-1})}{h_m} \sim \mathcal{N}(0, \sigma_d^2), \quad m = 2, \ldots M-2. \tag{B.2}$$

The parameters σ_e^2 and σ_d^2 are smoothness parameters in the sense that small variances smooth the function because the differences in coefficients are presumed to be small, while large variances have the opposite effect.

The foregoing assumptions imply that for a given covariate $\mathbf{w}$,

$$\boldsymbol{\Delta}_w \boldsymbol{\beta}_w | \sigma_{ew}^2, \sigma_{dw}^2 \sim \mathcal{N}_{M-1}(0, \mathbf{T}_w),$$

or that

$$\boldsymbol{\beta}_w | \sigma_{ew}^2, \sigma_{dw}^2 \sim \mathcal{N}_{M-1}(0, \boldsymbol{\Delta}_w^{-1} \mathbf{T}_w \boldsymbol{\Delta}_w^{-1\prime}), \tag{B.3}$$

where

$$\boldsymbol{\Delta}_w = \begin{pmatrix} \frac{1}{h_2} & 0 & \cdots & \cdots & \cdots & 0 \\ -\left(\frac{1}{h_2}+\frac{1}{h_3}\right) & \frac{1}{h_3} & 0 & \cdots & \cdots & 0 \\ \frac{1}{h_3} & -\left(\frac{1}{h_3}+\frac{1}{h_4}\right) & \frac{1}{h_4} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \frac{1}{h_{M-2}} & -\left(\frac{1}{h_{M-2}}+\frac{1}{h_{M-1}}\right) & \frac{1}{h_{M-1}} & 0 \\ 0 & \cdots & 0 & 0 & -\frac{1}{h_M} & \frac{1}{h_M} \end{pmatrix},$$

is a banded lower-triangular matrix, and $\mathbf{T}_w = \mathsf{diag}(\sigma_{ew}^2, \sigma_{dw}^2 \mathbf{I}_{M-3}, \sigma_{ew}^2)$.

If there are k_0 categorical covariates with coefficients $\boldsymbol{\beta}_0$ and prior distribution $\mathcal{N}_{k_0}(\mathbf{b}_{0l}, \mathbf{V}_{0l})$, and q continuous covariates $\mathbf{w}_1, \ldots, \mathbf{w}_q$, the spline coefficients $(\boldsymbol{\beta}_{w_1}, \ldots, \boldsymbol{\beta}_{w_q}) \equiv (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q)$ conditioned on $\boldsymbol{\sigma}_e^2 = (\sigma_{e1}^2, \ldots, \sigma_{eq}^2)$ and $\boldsymbol{\sigma}_d^2 = (\sigma_{d1}^2, \ldots, \sigma_{dq}^2)$ are assumed to be independently distributed according to the proceeding prior, so that $\boldsymbol{\beta}_0 = \left(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q\right)$ has the prior

$$\boldsymbol{\beta} | \boldsymbol{\sigma}_e^2, \boldsymbol{\sigma}_d^2 \sim \mathcal{N}_k(\mathbf{b}_0, \mathbf{V}_0), \tag{B.4}$$

where $k = k_0 + \sum_r (M_r - 1)$,

$$\mathbf{b}_0 = \begin{pmatrix} \mathbf{b}_{0l} \\ \mathbf{0} \end{pmatrix} \text{ and } \mathbf{V}_0 = \mathsf{diag}(\mathbf{V}_{0l}, \boldsymbol{\Delta}_1^{-1} \mathbf{T}_1 \boldsymbol{\Delta}_1^{-1\prime}, \ldots, \boldsymbol{\Delta}_q^{-1} \mathbf{T}_q \boldsymbol{\Delta}_q^{-1\prime}).$$

The prior model is completed by assuming that independently

$$(\sigma_e^2, \sigma_d^2) \sim \prod_{j=1}^{q} \mathsf{inverse\ gamma}\left(\frac{\alpha_{ej0}}{2}, \frac{\delta_{ej0}}{2}\right) \mathsf{inverse\ gamma}\left(\frac{\alpha_{dj0}}{2}, \frac{\delta_{dj0}}{2}\right) \tag{B.5}$$

for given values of the hyperparameters $\{\alpha_{ej0}, \delta_{ej0}, \alpha_{dj0} \text{ and } \delta_{dj0}\}_{j=1}^{q}$.

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.

Basu, S. and Chib, S. (2003), "Marginal likelihood and Bayes Factors for Dirichlet process mixture models," *Journal of the American Statistical Association*, 98, 224–235.

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

— (2001), "Markov chain Monte Carlo methods: Computation and inference," in *Handbook of Econometrics*, eds. Heckman, J. J. and Leamer, E., Amsterdam: North-Holland, vol. 5, pp. 3569–3649.

— (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.

Chib, S. and Hamilton, B. H. (2002), "Semiparametric Bayes analysis of longitudinal data treatment models," *Journal of Econometrics*, 110, 67–89.

Chib, S. and Jacobi, L. (2008), "Analysis of treatment response data from eligibility designs," *Journal of Econometrics*, 144, 465–478.

Denison, D. G. T., Holmes, C. C., Mallick, B. K., and Smith, A. F. M. (2002), *Bayesian Methods for Nonlinear Classification and Regression (Wiley Series in Probability and Statistics)*, Wiley, 1st ed.

Dunson, D. B., Pillai, N., and Park, J. H. (2007), "Bayesian density regression," *Journal of The Royal Statistical Society Series B-Statistical Methodology*, 69, 163–183.

Frangakis, C. E. and Rubin, D. B. (2002), "Principal stratification in causal inference," *Biometrics*, 58, 21–29.

Imbens, G. W. (2004), "Nonparametric estimation of average treatment effects under exogeneity: A review," *Review of Economics and Statistics*, 86, 4–29.

Imbens, G. W. and Rubin, D. B. (1997), "Bayesian inference for causal effects in randomized experiments with noncompliance," *The Annals of Statistics*, 25, 305–327.

Morgan, S. L. and Winship, C. (2007), *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, New York, NY: Cambridge University Press, 1st ed.

Pearl, J. (2009), *Causality: Models, Reasoning and Inference*, New York, NY: Cambridge University Press, 2nd ed.

Rosenbaum, P. R. (2002), *Observational Studies*, New York, NY: Springer, 2nd ed.

— (2009), *Design of Observational Studies (Springer Series in Statistics)*, New York, NY: Springer, 1st ed.

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.

— (1984), "Reducing bias in observational studies using subclassification on the propensity score," *Journal of The American Statistical Association*, 79, 516–524.

Rubin, D. B. (2005), "Causal inference using potential outcomes: Design, modeling, decisions," *Journal of the American Statistical Association*, 100, 322–331.

— (2007), "The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials," *Statistics in Medicine*, 26, 20–36.

Sommer, A. and Zeger, S. (1991), "On estimating efficacy from clinical trials," *Statistics in Medicine*, 10, 45–52.