Nonparametric Bayes Analysis of the Sharp and Fuzzy Regression Discontinuity Designs

Siddhartha Chib*  
Edward Greenberg†  
March 2014, September 2016

Abstract

We develop a Bayesian analysis of the sharp and fuzzy RD designs in which the unknown functions of the forcing variable are modeled by penalized natural cubic splines, and the error is distributed as student-t. Several novel ideas are employed. First, in estimating the functions of the forcing variable, we include a knot at the threshold, which is not in general an observed value of the forcing variable, to allow for curvature in the estimated functions from the breakpoint to the nearest values on either side of the breakpoint. Second, we cluster knots close to the threshold with the aim of controlling the approximation bias. Third, we introduce a new second-difference prior on the spline coefficients that can deal with unequally spaced knots. The number of knots and other features of the model are compared through marginal likelihoods, which are easily computed by the method of Chib (1995). Fourth, we develop an analysis of the fuzzy design based on a new model that utilizes the principal stratification framework, adapted to the RD design. Posterior computations for both designs are straightforward and are implemented in two R-packages that may be downloaded. The excellent performance of the proposed Bayes ATE and (complier) ATE estimates is documented in simulation experiments.

Keywords: Bayesian inference; Causal inference; Marginal likelihood; MCMC.

*Olin Business School, Washington University in St. Louis, Campus Box 1133, 1 Bookings Drive, St. Louis, MO 63130. e-mail: chib@wustl.edu.
†Department of Economics, Washington University in St. Louis, Campus Box 1133, 1 Bookings Drive, St. Louis, MO 63130. e-mail: edg@wustl.edu.
1 Introduction

Our goal in this study is to develop Bayesian inferential procedures for the sharp and fuzzy regression discontinuity (RD) designs (Thistlethwaite and Campbell (1960), Campbell (1969)). The RD design constitutes a quasi-experimental design, in the sense that it shares some of the features of an experimental design and, when available, provides a context in which causal conclusions can be drawn from observational data. Although a great deal of work has now emerged on the frequentist analysis of these designs, for example, Hahn, Todd, and Van der Klaauw (2001), Imbens and Lemieux (2008), Lee and Lemieux (2010), Frandsen, Froelich, and Melly (2012) and Calonico, Cattaneo, and Titiunik (2014), along with scores of applications in various fields, Bayesian work is somewhat scarce. Our goal is to show that the Bayesian viewpoint is useful and leads to a practical and competitive alternative to existing approaches.

Consider, first, the sharp RD design. The distinguishing feature of such a design is that the (binary) treatment variable $x \in \{0, 1\}$ is given by the deterministic rule

$$x = I[z \geq \tau]$$

where $I[.]$ is the indicator function, $z$ is an exogenous forcing variable and $\tau$ is a known discontinuity point. Thus, $x = 0$ when $z$ is to the left of $\tau$, and $x = 1$ when $z$ is to the right of $\tau$. Let $y_0$ and $y_1$, both assumed to be in $\mathbb{R}$, denote the potential outcomes with the observed outcome given by $y = (1-x)y_0 + xy_1$. Because $x$ is a deterministic function of $z$, there can be no unobserved confounders provided that the assignment process itself is not manipulable. Now suppose, for convenience, that there are no measured confounders apart from $z$. Also suppose that the distribution of $z$ around $\tau$ is smooth (ie., has no jumps or discontinuities). Then, the relevant treatment effect,

$$\text{ATE} : E(y_1|z = \tau) - E(y_0|z = \tau)$$

called the RD average treatment effect (ATE), is non-parametrically identified (Hahn et al., 2001). Moreover, Hahn et al. (2001) discuss an approach for estimating this effect by a local polynomial regression restricted to a window around $\tau$. More recently, Imbens and Kalyanaraman (2012) and Calonico et al. (2014) have developed methods for determining this window using asymptotic arguments, conditional on the order of the local-polynomial used to construct the point-estimator, and the order of the local-polynomial used to construct the bias-correction. In our approach, instead, we model and estimate the potential outcome distributions of $y_0$ and $y_1$ over the entire support of $z$. More pre-
cisely, we suppose that $y_0 = g_0(z) + \sigma \varepsilon_0$, for $z < \tau$, and $y_1 = g_1(z) + \sigma \varepsilon_1$, for $z \geq \tau$, where the $g_j$ functions are smooth with well defined left and right limits at $\tau$ and the noise is standard student-t with $\nu > 2$ degrees of freedom.

We suppose that each $g_j(z)$ function is in the span of the natural cubic splines. We use the basis from Chib and Greenberg (2010), and a new prior on the basis coefficients, to do our prior-posterior analysis on these functions. In our chosen basis, the basis coefficients are the function heights at the chosen knots. Therefore, by placing a knot at $\tau$ in the basis expansions of both $g_0$ and $g_1$, we can learn about the RD ATE, now equal to $g_1(\tau) - g_0(\tau)$, from the posterior distribution of the difference in two basis coefficients. This fact greatly simplifies the computations of the treatment effect and has both practical and theoretical implications. If desired, our approach can be implemented with a distribution more general than the student-t, say by utilizing the Dirichlet process prior as, for example, in Chib and Greenberg (2010). We avoid this complication for two reasons. First, because the student-t distribution with small degrees of freedom offers protection from outlying observations and, to some degree, is robust to mis-specification in the same way as semiparametric formulations with less overhead. Second, the student-t assumption allows us to focus on the development of the Bayesian approach that is salient for the specifics of the RD setting. We can thus more clearly describe how we learn about $g_j(\tau)$, even though $g_0(\tau)$ cannot be observed and $g_1(\tau)$ is unlikely to be observed in most small sample situations.

In the fuzzy design, the assignment rule $I[z \geq \tau]$ and the treatment $x$ differ, but the probability that $x = 1$ has a discontinuity at $\tau$. As a result, some subjects to the left of $\tau$, though more likely to be non-treated, can receive the treatment, and some subjects to the right of the break-point, though more likely to be treated, can be non-treated. This divergence between $x$ and $I[z \geq \tau]$ is explained by an unobserved confounder (a variable that affects $x$ and also affects the potential outcomes $y_j$). In the literature to date, it has been (implicitly) assumed that the confounder is a continuous random-variable. For instance, if one supposes that individuals select into the treatment on the basis of the unobserved (to the statistician) potential gain from one vs the other treatment, then the confounder is $y_1 - y_0$, a continuous quantity if the $y_j$’s are continuous. The presence of $y_1 - y_0$ in the $x$ model has the effect of making $x$ correlated with the potential outcomes. Hahn et al. (2001) show that if one assumes (Assumption 3 in their paper) that the distribution of $y_1 - y_0$ and $x$ are both free of $z$, at least in a window around $\tau$, then a windowed version of the instrumental variable (IV) estimator is a consistent estimator of the complier average treatment effect (CATE) at $\tau$, which is the ATE for
subjects that are compliant with the assignment. This condition therefore requires that \( y_0 = g(z) + \varepsilon_0 \) and \( y_1 = g(z) + \varepsilon_1 \) (so that the gain is free of \( z \)) and a \( x \) model such as

\[
x = I[\alpha_0 + \alpha_1 I[z \geq \tau] + \varepsilon_1 - \varepsilon_0 + \varepsilon]\]

where the errors are (say) mutually independent. In fact, this is the model discussed in Frandsen et al. (2012) to which we will return later in our simulation experiments. Given such a treatment model, it is possible to define a complier as one whose \( x \) is 0 when \( I[z \geq \tau] = 0 \) and whose \( x \) is 1 when \( I[z \geq \tau] = 1 \). Similarly, a never-taker is one who has a value of \( x = 0 \) regardless of the value of \( I[z \geq \tau] \), and an always-taker is one with a value of \( x = 1 \) regardless of the value of \( I[z \geq \tau] \).

Notice that these definitions are in relation to the \( x \) model but that these types play no role in the windowed IV estimator. An unnoticed implication of the preceding condition is that the potential outcome distributions of the different types just differ by a mean shift.

Our analysis of the fuzzy design is based on a new model of the confounder. Specifically, we suppose that the confounder is a discrete random-variable \( s \), for subject type, taking the values \( \{c, n, a\} \) for complier, never-taker and always-taker, respectively. Depending on the problem, the nomenclature of these types may be different as, for example, in Chib and Jacobi (2016), where our approach has been applied. Significantly, this notion of subject-type is not defined in relation to a \( x \) model, as in the preceding continuous-confounder discussion, but is an exogenous characteristic of each subject. The discrete distribution of \( s \) is unknown and must be estimated, but the distribution of \( x \) given \( z, I[z \geq \tau] \) and \( s \) is deterministic and plays no further role in the analysis. In this setting, four potential outcomes emerge, \( y_0 \) and \( y_1 \) for the compliers, and \( y_{0n} \) and \( y_{1a} \) for never-takers and always-takers, respectively. Conditioned on \( z \) and \( s = c \), we suppose that the potential outcomes \( y_0 \) and \( y_1 \) are generated as in the sharp model. For the new types, we suppose that conditioned on \( z \) and \( s = n \),

\[ y_{0n} = g_{0n}(z) + \sigma_{0n}\varepsilon_{0n} \]

and conditioned on \( z \) and \( s = a \),

\[ y_{1a} = g_{1a}(z) + \sigma_{1a}\varepsilon_{1a} \]

both over the entire support of \( z \). For identifiability we require that the function \( g_{0n} \) and \( g_{1a} \) are continuous at \( \tau \).
The object of interest in this context is the RD average treatment effect for compliers which is given by
\[ E[y_1 | z = \tau, s = c] - E[y_0 | z = \tau, s = c] = g_1(\tau) - g_0(\tau) \]
which we show below is identified under weak assumptions. Our modeling of the fuzzy RD design is inspired by the principal stratification framework of Frangakis and Rubin (2002) but with features that take account of the specifics of the fuzzy RD setting. An important point to note is that the sharp RD design falls out of this specification if every subject is a complier, which is very useful property. In this sense, the discrete confounder formulation of the fuzzy RD design is a more natural generalization of the sharp design than the continuous confounder formulation.

In practice, one does not know if the confounder is continuous or discrete. We therefore consider experiments in which we generate data from the continuous confounder fuzzy RD model and estimate our discrete confounder model and compare the results with those of the frequentist approaches. We then reverse the data generating process, generating data from the discrete confounder model, and conduct the same comparison. From these experiments, which we detail below, we see that our approach performs well under both circumstances. The frequentist approaches, on the other hand, do less well when the confounder is discrete, especially in small samples.

In Section 2, we begin by considering the sharp RD design (which is also the model for compliers in the fuzzy RD case) and introduce the key ideas related to the estimation of the model. These key ideas relate to the soft-windowing procedure for locating knots in the cubic spline basis expansions, and the prior distribution on the basis coefficients. In Section 4 we extend the analysis to the fuzzy RDD model. Sections 5 contains simulation studies on the performance of our methods, and Section 6 contains an application of the sharp design to real data. Section 7 contains our conclusions. Details related to the basis expansions used in this article are contained in Appendix A.

2 Sharp RDD

Let \( x = j \), for \( j = 0, 1 \), denote the two levels of the treatment intake and suppose that \( x = I[z > \tau] \) is the process determining the intake. Thus, \( x = 0 \) when \( z \leq \tau \) and \( x = 1 \) when \( z > \tau \). This intake process is subject to the following assumption.

**Assumption 1** The intake process \( x = I[z > \tau] \) is not manipulable.
Let $y_j$ denote the potential outcome when $x = j$ and let the observed outcome be $y = (1 - x)y_0 + xy_1$. Note that one could conceivably define a $y_1$ for the non-treated, or a $y_0$ for the treated (even though neither is identified) but this would be unnecessary for purposes of inference following Chib (2007).

We make the following assumption about the potential outcomes.

**Assumption 2** The potential outcomes $y_j$ conditioned on $z$ are generated as

\[
\begin{align*}
y_0 &= g_0(z) + \sigma \varepsilon_0, \quad z < \tau \\
y_1 &= g_1(z) + \sigma \varepsilon_1, \quad z \geq \tau
\end{align*}
\]  

(2.1)

where the functions $g_j(\cdot)$, $j = 0, 1$, are smooth unknown functions of $z$ with well defined left and right limits at $\tau$

\[
\lim_{z \uparrow \tau} g_0(z) = g_0(\tau) \quad \text{and} \quad \lim_{z \downarrow \tau} g_1(z) = g_1(\tau)
\]

and the random noise $\varepsilon_j$, $j = 0, 1$, is independent of $z$ (because of Assumption 1) and is distributed as $\mathcal{N}(0, 1)$, the standard Gaussian distribution.

Under these assumptions, the goal is to do prior-posterior inference on the RD ATE

\[
E[y_1|z = \tau] - E[y_0|z = \tau] = g_1(\tau) - g_0(\tau).
\]  

(2.2)

2.1 Sample data

Suppose now that the available data are $n$ independent observations on $(y, x, z)$. These are indicated by $(y_i, x_i, z_i)$, $i \leq n$, where $y_i$ is $y_{0i}$ when $x_i = 0$ and $y_{1i}$ when $x_i = 1$. Denote the number of observations to the left of $\tau$ by $n_0$ and the number of observations to the right of $\tau$ by $n_1$, with $n = n_0 + n_1$. For convenience, rearrange the data so that the first $n_0$ observations correspond to those for $x_i = 0$ and the next $n_1$ to those for $x_i = 1$. Let the vector of observations on $(y, z)$ to the left of $\tau$ be assembled as

\[
y_0 \triangleq (y_1, \ldots, y_{n_0}) \quad (n_0 \times 1), \quad z_0 \triangleq (z_1, \ldots, z_{n_0}) \quad (n_0 \times 1),
\]

and those to the right of $\tau$ as

\[
y_1 \triangleq (y_{n_0+1}, \ldots, y_n) \quad (n_1 \times 1), \quad z_1 \triangleq (z_{n_0+1}, \ldots, z_n) \quad (n_1 \times 1)
\]

For later reference, define $z_{j,\min} \triangleq \min(z_j)$, $z_{j,\max} \triangleq \max(z_j)$, $(j = 0, 1)$, and the $p$th quantile of $z_j$ by $z_{j,p}$.
2.2 Soft windowing and basis expansions

We begin by describing our modeling of the unknown $g_0(z)$ and $g_1(z)$ functions. Our approach is based on penalized natural cubic splines. A natural cubic spline is a smooth curve constructed from sections of cubic polynomials joined together at knot points under the constraints that the function has continuous second derivatives at the knot points and that the second derivatives are zero at the end knots. As in known, any cubic spline can be expressed as a linear combination of basis functions, where the weights, called the basis coefficients, depend on the specific basis.

Consider now the basis given in Chib and Greenberg (2010), also summarized in Appendix A. We adopt this basis over (say) the B-spline basis, because under this basis, the basis coefficients are the function heights at the chosen knots. We take advantage of this rather remarkable property by placing the last knot in the $g_0(z)$ basis expansion not at $z_{0,\text{max}}$ (the normal choice) but at $\tau$, a point necessarily to the right of $z_{0,\text{max}}$. Analogously, in the case of the $g_1(z)$ expansion, we place the first knot not at $z_{1,\text{min}}$ but at $\tau$, a value in most cases to the left of $z_{1,\text{min}}$. We do this because, then, the RD ATE reduces to the difference of two basis coefficients, which greatly simplifies the computation of the treatment effect. Another motivation for placing a knot at $\tau$ in this way is to allow the $g$ functions to have curvature over the intervals $(z_{0,\text{max}}, \tau)$ and $(\tau, z_{1,\text{min}})$. Otherwise, by properties of the natural cubic spline, those estimated functions are simply linear over those intervals.

The question now is how best to place the other knots to extract information from the data in the vicinity of $\tau$. Our idea is to cluster some knots in the regions around $\tau$ and then sprinkle the other knots in the regions further away. We refer to this approach as soft-windowing. It is implemented as follows. Partition the closed intervals $[z_{0,\text{min}}, \tau]$ and $[\tau, z_{1,\text{max}}]$ into intervals that are proximate and far from $\tau$. Let these four intervals be determined by the quantiles $z_{0,p_0}$ and $z_{1,p_1}$, for specific values of $p = (p_0, p_1)$, for example, $p = (0.9, 0.1)$. A particular distribution of knots is shown in Figure 1. Knots are now allocated to each of the four segments with the provision that there is at least one observation between each successive pair of knots. In placing these knots, we place $m_{z,\tau} = (m_{z,0,\tau}, m_{z,1,\tau})$ knots in the intervals proximate to $\tau$, and $m_z = (m_{z,0}, m_{z,1})$ knots in the intervals that are further away from $\tau$.

Setting up an algorithm that places the desired number of knots under the constraint of no-empty intervals can be a bit tricky, especially when the data is sparse. One algorithm, which may be characterized as ‘propose-check-accept-extend,’ is simple to implement and ensures that the number of knots produced is close to, but not necessarily equal to, the desired numbers. It proceeds in the following
Figure 1: Example of knot locations in the basis expansions of $g_0$ (top panel) and $g_1$ (bottom panel), determined by $m_z = (6, 5), m_{z, \tau} = (5, 5)$. Note that the no empty interval constraint meant that the number of knots is smaller than what is implied by these choices. The circled points are the $p_0$ and $p_1$ quantiles of $z_0$ and $z_1$, respectively. Both $g_0$ and $g_1$ have a knot at $\tau$.

way: For the two intervals to the left of $\tau$, place a knot at $\tau$ and let $\Delta_\tau = (\tau - z_0,p_0)/(m_{z,0,\tau} - 1)$ be the initial spacing for the remaining knots in the interval proximate to $\tau$. Propose the next knot at $\tau - \Delta_\tau$, and accept it as a knot if it produces a non-empty interval. Otherwise, propose a knot at $\tau - 2\Delta_\tau$, check for a non-empty interval, accept or extend the interval, and continue in this way until either $z_0,p_0$ is reached or exceeded. Then calculate the spacing $\Delta_0 = (z_0,p_0 - z_{0,\text{min}})/m_{z,0}$ and proceed from the last accepted knot in the same way as before, making sure that $z_{0,\text{min}}$ is a knot at the end of this stage. The same propose-check-accept-extend approach is applied to the right of $\tau$ after placing the first knot at $\tau$ and ending with a knot at $z_{1,\text{max}}$. Let $\{z_{0,\text{min}}, \kappa_{0,2}, \ldots, \kappa_{0,m_0-1}, \tau\}$ denote the $m_0$ knots to the left of $\tau$ determined by this procedure, and let $\{\tau, \kappa_{1,2}, \ldots, \kappa_{1,m_1-1}, z_{1,\text{max}}\}$ denote the $m_1$ knots to the right of $\tau$. An example is shown in Figure 1, where $m_0 = 10$ and $m_1 = 7$. When using this algorithm, note that

$$m_0 \leq m_{z,0} + m_{z,0,\tau}$$

and

$$m_1 \leq m_{z,1,\tau} + m_{z,1}$$

and that, in general, the knots are not equally-spaced.

In practice, one can choose the values of $p_j, m_{z,j},$ and $m_{z,j,\tau}$, by examining $z_0$ and $z_1$, and placing more knots where there is a greater concentration of observations. These choices can then be adjusted on the basis of the marginal likelihoods of various models, as discussed below. As the sample size
increases, one can increase the default number of knots roughly as $cn_j^\nu$, for some constant $c$ and $\nu \geq 1/5$, following the rate derived in Claeskens, Krivobokova, and Opsomer (2009).

Given the knots, the function ordinates,

$$g_0(z_0) \triangleq (g_0(z_1), \cdots, g_0(z_{n_0}))$$

and

$$g_1(z_1) \triangleq (g_1(z_{n_0+1}), \cdots, g_1(z_n))$$

can be expanded in terms of the natural cubic spline basis functions in the appendix as

$$g_0(z_0) = B_0 \alpha$$
$$g_1(z_1) = B_1 \beta,$$

respectively, where $B_j : n_j \times m_j$ are the basis matrices, and $\alpha$ and $\beta$ are the basis coefficients. Under our basis, these are explicitly the function ordinates at the knots,

$$\alpha_{(m_0 \times 1)} = \begin{pmatrix} g_0(z_{0,\text{min}}) \\ g_0(\kappa_{0,2}) \\ \vdots \\ g_0(\kappa_{0,m_0-1}) \\ g_0(\tau) \end{pmatrix}, \quad \beta_{(m_1 \times 1)} = \begin{pmatrix} g_1(\tau) \\ g_1(\kappa_{1,2}) \\ \vdots \\ g_1(\kappa_{1,m_1-1}) \\ g_1(\tau_{\text{max}}) \end{pmatrix},$$

which implies that the ATE is simply the first component of $\beta$ minus the last component of $\alpha$:

$$\text{ATE} = \beta_{[1]} - \alpha_{[m_0]}.$$  \hfill (2.5)

2.3 Likelihood

Given the preceding basis expansions, the likelihood function of the sharp RD model is given by

$$p(y|\theta, \sigma^2) = \prod_{i=1}^{n_0} t_\nu(y_i|B_0, \alpha, \sigma^2) \prod_{i=1}^{n_1} t_\nu(y_{n_0+i}|B_1, \beta, \sigma^2)$$

where $t_\nu$ is the student-t density function and $B_{j,i}$ is the $i$th row of $B_j$.

Note that by employing the usual scale mixture of normals representation of the student-t distribution, the sharp RD model for all $n$ observations is also expressible as

$$\begin{pmatrix} y_0 \\ y_1 \end{pmatrix}_{(n_0+n_1 \times 1)} = \begin{pmatrix} B_0 & 0 \\ 0 & B_1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}_{(m_0+m_1 \times 1)} + \begin{pmatrix} \varepsilon_0 \\ \varepsilon_1 \end{pmatrix}_{(n_0+n_1 \times 1)},$$

\hfill (2.6)
or as \( y = X\theta + \varepsilon \), where \( \theta \equiv (\alpha, \beta) \) is the regression parameter of length \( k = (m_0 + m_1) \), \( \varepsilon \equiv (\varepsilon_0, \varepsilon_1) \) is \( N_n(0, \sigma^2 \Xi^{-1}) \),

\[
\Xi = \text{diag}(\xi_1, \ldots, \xi_n)
\]

and

\[
\xi_i \sim \text{Gamma} \left( \frac{\nu}{2}, \frac{\nu}{2} \right), \quad i \leq n
\]

### 2.4 Prior Distribution

In spline estimation with no regularizing penalty, there is a trade-off between the model fit and the smoothness of the function estimates. As the model fit is improved by adding knots, the function estimates tend to become less smooth. In non-Bayesian penalized spline estimation, the smoothness of the function is controlled by adding an \( l_2 \)-based roughness penalty to the negative log-likelihood, or least squares, objective function. A commonly chosen penalty is the integrated squared second-order derivative of the spline function. In the case of the B-spline basis, the second derivatives are in terms of weighted second-order differences of the basis coefficients, where the weights depend on the spacing between knots. As a result, translating this penalty into a prior on the basis coefficients, as done for example, in Lang and Brezger (2004) and Brezger and Lang (2006), leads to a second order auto-regressive process with a unit root on the basis coefficients.

In this paper we also employ a prior that is a second-order autoregressive process but because the knots in our formulation are typically not equally based we formulate this process by discretizing a second-order second-order Ornstein–Uhlenbeck (O-U) process. In continuous time, the second-order O-U process for a diffusion \( \{\varphi_t\} \) can be defined through the stochastic differential equation

\[
d^2 \varphi_t = -a(d\varphi_t - b)dt + s dW_t,
\]

where \( a > 0 \), and \( \{W_t\} \) is the standard Wiener process. Our idea is to use the Euler-discretized form of this process as a prior for the basis coefficients, letting \( dt \) be the spacing between successive knots. In this construction, we also let \( a = 1, b = 0 \) and \( s = 1/\sqrt{\lambda} \), where \( \lambda \) is the penalty parameter.

**Prior of \( \alpha \):** Consider the situation shown in Figure 2 for values of \( g_0 \) computed at three successive knots, represented by \( \alpha_i = g_0(\kappa_{0,i}) \), \( \alpha_{i-1} = g_0(\kappa_{0,i-1}) \) and \( \alpha_{i-2} = g_0(\kappa_{0,i-2}) \). Let

\[
\Delta^2 \alpha_i \equiv (\alpha_i - \alpha_{i-1}) - (\alpha_{i-1} - \alpha_{i-2}), \quad i > 2
\]
\[ \alpha_i = g_0(\kappa_{0,i}) \quad \beta_j = g_1(\kappa_{1,j}) \]

Figure 2: Prior formulation: three successive knots on either side of \( \tau \) and the corresponding function ordinates. The latter are the basis coefficients in the cubic spline basis expansions of \( g_0 \) and \( g_1 \). The prior on these coefficients is defined through a second-order O-U process. The process moves from left to right on the \( \alpha_i \) \((i > 2)\), and from right to left on the \( \beta_j \) \((j < m_1 - 1)\).

and define the spacings between knots by \( h_{0,i} = \kappa_{0,i} - \kappa_{0,i-1} \), as shown in Figure 2. We suppose now that, a priori, \((\alpha_3, \alpha_4, \ldots, \alpha_{m_0})\), conditioned on \((\alpha_1, \alpha_2)\), follow the process

\[
\Delta^2 \alpha_i = - (\alpha_{i-1} - \alpha_{i-2}) h_{0,i} + u_{0,i},
\]

\[ u_{0,i} | \lambda_0 \sim \mathcal{N} \left( 0, \lambda_0^{-1} h_{0,i} \right), \]

where \((\alpha_{i-1} - \alpha_{i-2}) h_{0,i}\) introduces mean reversion and \( \lambda_0 \) is an unknown precision (smoothness) parameter.

To complete this process, we specify a distribution on \((\alpha_1, \alpha_2)\). Our prior on these parameters is proper, unlike [Lang and Brezger (2004)] and [Brezger and Lang (2006)], in order to allow computation of marginal likelihoods for comparing models. Our choice of this distribution is motivated by Zellner’s \( g \)-prior and, to the best of our knowledge, has not been used in a similar way before. Let \( T_{\alpha,1:2}^{-1} \triangleq (B_0' B_0)^{-1}_{1:2} \) denote the first two rows and columns of \( B_0' B_0 \). We then let

\[
\begin{pmatrix}
\alpha_1 \\
\alpha_2
\end{pmatrix} = \begin{pmatrix}
g_0(z_{0,\min}) \\
g_0(\alpha_{0,2})
\end{pmatrix} \sim \mathcal{N}_2 \left( \begin{pmatrix}
\alpha_{1,0} \\
\alpha_{2,0}
\end{pmatrix}, \lambda_0^{-1} T_{\alpha,1:2} \right)
\]

where \( \alpha_{1,0} \) and \( \alpha_{2,0} \) (the expected levels of \( g_0 \) at the first two knots) are the only two free hyperparameters.
By straightforward calculations we can show that the joint prior distribution is given by

$$\alpha|\lambda_0 \sim \mathcal{N}_{m_0} \left( D_\alpha^{-1} \alpha_0, \lambda_0^{-1} D_\alpha^{-1} T_\alpha D_\alpha^{-1}' \right)$$  \hfill (2.9)

where \( \alpha_0 = (\alpha_1, 0, 0, \ldots, 0)' : m_0 \times 1 \), \( D_\alpha \) is a tri-diagonal matrix (given in Appendix B) that depends entirely on the spacings, and \( T_\alpha = \text{blockdiag}(T_{\alpha,1:2}, I_{m_0-2}) : m_0 \times 1 \). Note that, under this prior, the diagonal elements of \( D_\alpha^{-1} T_\alpha D_\alpha^{-1}' \) increase as one moves down the diagonal, which implies that \( \text{Var}(g_0(z_{0,\min})) < \text{Var}(g_0(\tau)) \). Also note that this prior is fully specified by the two hyperparameters, \( \alpha_{1,0} \) and \( \alpha_{2,0} \), which is convenient.

Prior of \( \beta \): Our prior of \( \beta \) is similar except for one key difference. Rather than assuming that the O-U process moves from the smallest knot (namely \( \tau \)) to the largest, we orient the process from right to left. By following this approach, the prior on the key parameter \( \beta_1 = g_1(\tau) \) is determined by the O-U process, in the same way that the prior on \( \alpha_{m_0} = g_0(\tau) \) is determined by the \( \alpha \) O-U process. This refinement helps to control the magnitude of the shrinkage-bias for \( \alpha_{m_0} \) and \( \beta_1 \).

Consider the three successive knots of \( g_1 \), shown in Figure 2 and the corresponding function values \( \beta_j = g_1(\kappa_1,j), \beta_{j+1} = g_1(\kappa_1,j+1) \) and \( \beta_{j+2} = g_1(\kappa_1,j+2) \). Conditioned on the right end-points \((\beta_{m_1-1}, \beta_{m_1})\), let

$$\Delta^2 \beta_j \triangleq (\beta_{j} - \beta_{j+1}) - (\beta_{j+1} - \beta_{j+2}), \ j < m_1 - 1$$

denote a sequence of second differences. Then, under the prior, our assumption is that

$$\Delta^2 \beta_j = -(\beta_{j+1} - \beta_{j+2})h_{1,j+1} + u_{ji}, \hfill (2.10)$$

$$u_{ji}|\lambda_1 \sim \mathcal{N} \left( 0, \lambda_1^{-1} h_{1,j+1} \right), \hfill (2.11)$$

where \( h_{1,j+1} = \kappa_{1,j+1} - \kappa_{1,j} \) is the spacing between knots and \( \lambda_1 \) is an unknown precision parameter. Note that we let this prior process have its own smoothness parameter.

As above, we complete the prior modeling by placing a \( g \)-type prior distribution on \((\beta_{m_1-1}, \beta_{m_1})\). Let \( T_{\beta,m_1-1:m_1}^{-1} \triangleq (B_1'B_1)_{m_1-1:m_1} \) denote the last two rows and columns of \( B_1'B_1 \). Then, our assumption is that

$$\begin{pmatrix} \beta_{m_1-1} \\ \beta_{m_1} \end{pmatrix} = \begin{pmatrix} g_1(\kappa_{1,m_1-1}) \\ g_1(\kappa_{1,m_1}) \end{pmatrix} \sim \mathcal{N}_2 \left( \begin{pmatrix} \beta_{m_1,0} \\ \beta_{m_1-1,0} \end{pmatrix}, \lambda_1^{-1} T_{\beta,m_1-1:m_1} \right),$$

which implies that

$$\beta|\lambda_1 \sim \mathcal{N}_{m_1} \left( D_\beta^{-1} \beta_0, \lambda_1^{-1} D_\beta^{-1} T_\beta D_\beta^{-1}' \right), \hfill (2.12)$$
where $\beta_0 = (0, \ldots, 0, \beta_{m_1-1,0}, \beta_{m_1,0})' : m_1 \times 1$, $D_\beta$ is the tri-diagonal matrix in Appendix B, and $T_\beta = \text{blockdiag}(I_{m_1-2}, T_{\beta, m_1-1, m_1}) : m_1 \times 1$.

Prior of $\lambda$ and $\sigma^2$: We complete our prior with a Gamma prior distribution on $\lambda_j$ ($j = 0, 1$). Just as in the frequentist interpretation of the penalized smoothing spline, for fixed $n$, $\lambda_j \rightarrow 0$ implies an unpenalized regression spline, and $\lambda_j \rightarrow \infty$ implies that the second differences are forced to zero, leading to piece-wise linearity. Keeping the latter facts in mind, and depending on the situation, we generally consider two methods of choosing the hyperparameters of the prior distribution. The first is to specify prior values of $E(\lambda_j)$ and $\text{sd}(\lambda_j)$ and match a Gamma distribution to these choices. For example, as a default we could let $E(\lambda_j) = 1$ and then let $\text{sd}(\lambda_j) = 5$, where the latter would be allowed to increase with the sample size. The second is to choose $E(\lambda_j)$ to make the smallest diagonal element of the variance matrix equal to one, that is, choose $E(\lambda_j)$ so that

$$\min \left\{ \text{diag} \left( \frac{1}{E(\lambda_j)} D_j^{-1} T_j D_j^{-1}' \right) \right\} = 1,$$

and let $\text{sd}(\lambda_j)$ be a multiple of this prior mean. Given the prior mean and standard deviation, we can then find independent matching Gamma distributions, denoted (say) as

$$\lambda_j \sim \text{Ga} \left( \frac{a_j0}{2}, \frac{b_j0}{2} \right), \ (j = 0, 1).$$

We note that if one is interested in the unpenalized regression spline model, one could let the prior mean of $\lambda_j$ be small and the prior standard deviation be even smaller.

The prior on $\sigma^2$ is of the usual form. Independent of $\lambda = (\lambda_0, \lambda_1)$, we suppose that

$$\sigma^2 \sim \text{IG} \left( \frac{\nu_00}{2}, \frac{\delta_00}{2} \right)$$

an inverse-gamma distribution, where $\nu_00$ and $\delta_00$ are chosen to reflect the researcher’s views about the mean and standard deviation of $\sigma^2$.

2.5 Posterior distributions and MCMC sampling

The sharp RD model under the preceding assumptions has the form

$$y|\theta, \sigma^2, \{\xi_i\} \sim \mathcal{N}_n(X\theta, \sigma^2\Xi),$$

$$\theta|\lambda \sim \mathcal{N}_k \left( \theta_0, \sigma^2 A_0 \right), \ \lambda_j \sim \text{Ga} \left( \frac{a_j0}{2}, \frac{b_j0}{2} \right), \ (j = 0, 1),$$

$$\sigma^2 \sim \text{IG} \left( \frac{\nu_00}{2}, \frac{\delta_00}{2} \right)$$
where $\theta_0 \triangleq \left( D_{\alpha}^{-1}\alpha_0, D_{\beta}^{-1}\beta_0 \right)'$ and

$$A_0 \triangleq \text{blockdiag} \left( \frac{1}{\lambda_0} D_{\alpha}^{-1} T_{\alpha} D_{\alpha}^{-1}', \frac{1}{\lambda_1} D_{\beta}^{-1} T_{\beta} D_{\beta}^{-1}' \right)$$

The posterior distribution of the parameters of this model can be sampled by the following MCMC algorithm, which is iterated $n_0 + m$ times, where $n_0$ is the number of burn-in iterations and $m$ is the number of iterations retained:

- Given $(y, \sigma^2, \{\xi_i\}, \{\lambda_j\})$, sample $\theta$ from $N_k(\hat{\theta}, A)$, where $\hat{\theta} = A_0^{-1}\theta_0 + \sigma^{-2}X'\Xi^{-1}y$ and $A = (A_0^{-1} + \sigma^{-2}X'\Xi^{-1}X)^{-1}$

- Given $(y, \theta, \{\xi_i\}, \{\lambda_j\})$, sample $\sigma^2$ from $\text{IG} \left( \nu_0 + n, \delta_0 + (y - X\theta)'\Xi^{-1}(y - X\theta) \right)$

- Given $(y, \theta, \sigma^2)$, sample $\{\xi_i\}$ from $\text{Ga} \left( \nu + 1 \over 2, \left( y_i - X_i\theta \right)^2 / \sigma^2 \right)$

- Given $\theta$, sample $\lambda$ from

$$\lambda_0|\alpha \sim \text{Ga} \left( a_0 + m_0 \over 2, b_0 + (D_{\alpha}\alpha - \alpha_0)'T_{\alpha}^{-1}(D_{\alpha}\alpha - \alpha_0) \over 2 \right)$$

$$\lambda_1|\beta \sim \text{Ga} \left( a_1 + m_1 \over 2, b_1 + (D_{\beta}\beta - \beta_0)'T_{\beta}^{-1}(D_{\beta}\beta - \beta_0) \over 2 \right)$$

- After the burn-in iterations, extract the last element of $\alpha$ and the first element of $\beta$ to obtain drawings of the ATE from its posterior distribution.

Marginal likelihoods play an important role in our approach. We compute these by the procedure of [Chib (1995)]. We use marginal likelihoods to compare models that differ in the value of the soft-windowing parameter $p$, and in the number of knots in the four regions implied by a given $p$.

3 Example: sharp design

3.1 Design

[Imbens and Kalyanaraman (2012), henceforth IK, and Calonico et al. (2014), henceforth CCT, are two recent papers that have developed windowed frequentist estimators for the sharp design. In order
to study the performance of their estimators, each paper considered simulated data from the model
\[ y_j = g_j(z) + \varepsilon_j, \]
where
\[ g_0(z) = 0.48 + 1.27z + 7.18z^2 + 20.21z^3 + 21.54z^4 + 7.33z^5, \]
\[ g_1(z) = 0.52 + 0.84z - 3.00z^2 + 7.99z^3 - 9.01z^4 + 3.56z^5, \]
and \( z \sim 2 \times \text{Beta}(2, 4) - 1 \)
\[ \varepsilon_0 \sim 0.1295 \times \mathcal{N}(0, 1) \]
and \( \varepsilon_1 \sim 0.1295 \times \mathcal{N}(0, 1) \). The true value of the ATE at the break-point \( \tau = 0 \) is 0.04.

A noteworthy feature of this design is that there are relatively fewer observations for \( z \geq 0 \) than for \( z < 0 \) which makes the estimation of the (small-sized) ATE quite challenging, especially in small samples. To make the problem even more interesting, we suppose that the error distribution is student-t. In particular, we suppose that \( \varepsilon_0 \sim 0.1295 \times t_\nu(0, 1) \) and \( \varepsilon_1 \sim 0.1295 \times t_\nu(0, 1) \), where \( \nu = 3 \).

### 3.2 Conditional analysis

Consider now two sample sizes, \( n = 500 \) and \( n = 4000 \). Let us generate one particular data set for each sample size (results from a repeated sampling experiment are in the next section) and consider the usual Bayesian analysis conditional on these data. Suppose that the prior mean and standard deviation of \( \sigma^2 \) are 0.3 and 1.0, respectively, and that the prior mean and standard deviation of \( \lambda \) are \((1, 1)\) and \((5, 5)\), respectively. This leaves us with having to fix the soft-window quantiles \( p = (p_0, p_1) \), the knots \( m_z \) far from \( \tau \), and \( m_{z, \tau} \) the knots close to \( \tau \). Because of the sparseness of the distribution of \( z \) (especially to the right of \( \tau \)), one should consider a wider soft-window setting. The soft-window settings can be tightened as the sample size increases. For illustration, the settings of \( p \) for each sample size are in Table I. Also we could be unsure of the value of \( \nu \), the degrees of freedom of the student-t error distribution. Let us suppose we consider three possible values for this parameter, 3, 3.5 and 4. Finally, again because of sparsity of the data, we use few knots on each side of \( \tau \) but allow for an extra knot on the left side for the larger sample size.

The results in Table I show that the choice of the soft-window width matters and leads to different marginal likelihoods. For \( n = 500 \), the wider soft-window is supported while for \( n = 4000 \) there is more support for the narrower soft-window. The second point to note is that the marginal likelihood can be used to determine the degrees of freedom. For both sample sizes we can see that the marginal
Table 1: Sharp design: Simulated data with error distributed as $0.1295 \times t_3(0, 1)$. This table shows that the soft-window quantiles influence the marginal likelihood (computed by the method of Chib (1995)) and that the marginal likelihood worsens as the degrees of freedom used in the estimation moves further away from the true value.

Table: Sharp design: Simulated data with error distributed as $0.1295 \times t_3(0, 1)$. This table shows that the soft-window quantiles influence the marginal likelihood (computed by the method of Chib (1995)) and that the marginal likelihood worsens as the degrees of freedom used in the estimation moves further away from the true value.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$p$</th>
<th>$m_z$</th>
<th>$m_{z, \tau}$</th>
<th>log marg lik</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>$t_3$</td>
<td>(.7, .3)</td>
<td>(3, 3)</td>
<td>(3, 2)</td>
</tr>
<tr>
<td></td>
<td>$t_3$</td>
<td>(.9, .1)</td>
<td>(3, 3)</td>
<td>(3, 2)</td>
</tr>
<tr>
<td></td>
<td>$t_{3.5}$</td>
<td>(.7, .3)</td>
<td>(3, 3)</td>
<td>(3, 2)</td>
</tr>
<tr>
<td></td>
<td>$t_4$</td>
<td>(.7, .3)</td>
<td>(3, 3)</td>
<td>(3, 2)</td>
</tr>
<tr>
<td>4000</td>
<td>$t_3$</td>
<td>(.7, .3)</td>
<td>(4, 3)</td>
<td>(3, 2)</td>
</tr>
<tr>
<td></td>
<td>$t_3$</td>
<td>(.9, .1)</td>
<td>(4, 3)</td>
<td>(3, 2)</td>
</tr>
<tr>
<td></td>
<td>$t_{3.5}$</td>
<td>(.9, .1)</td>
<td>(4, 3)</td>
<td>(3, 2)</td>
</tr>
<tr>
<td></td>
<td>$t_4$</td>
<td>(.9, .1)</td>
<td>(4, 3)</td>
<td>(3, 2)</td>
</tr>
</tbody>
</table>

3.2.1 Function estimates

The Bayes estimates of the $g_0$ and $g_1$ functions are given in Figure 3. In this figure, the true value of the functions are the dotted lines, the estimates are the solid lines, the 95% point-wise credibility intervals of the functions are the shaded bands, and the distribution of the $z$ values is notched on the horizontal axis. This figure shows that when $n = 500$, the function $g_1$ is not well estimated (because of the sparseness of the data) but that function estimates improve for the larger sample size. The right panel are the corresponding zoom plots, zoomed to the interval given by the soft-window quantiles.

3.2.2 Role of $\lambda_j$

Our fitting thus far is based on a small number of knots, as supported by the marginal likelihood criterion. In such a case, the regularizing effect of the prior is essentially minimal, and the posterior distribution of $\lambda_0$ and $\lambda_1$ concentrates on small values. What would happen instead if, for a particular sample size, we were to increase the number of knots? For instance, in the case of $n = 4000$, we can try fitting the $g_0$ function with 75 knots far from $\tau$ and 10 knots close to $\tau$ by letting $m_z = (75, 3)$ and $m_{z, \tau} = (10, 2)$. We leave the $g_1$ knots as is because of the paucity of data on the right of $\tau$. In this case, $\lambda_0$ should play a regularizing role, increasing in value to ensure a degree of smoothness in the function estimates. This is precisely what happens. With a prior on $\lambda_0$ that has a mean of 1 (as
Figure 3: Sharp design: Simulated data with error distributed as $0.1295 \times t_3(0, 1)$. This shows the function estimates and credibility bands for two different sample sizes. The right panel are the corresponding zoom plots, zoomed to the interval given by the soft-window quantiles.

before) with a standard deviation of 50 (instead of 5 as before), the posterior mean of $\lambda_0$ is 31.02 with posterior sd equal to 11.20. The resulting estimates of the two functions are given in the left panel of Figure 4 (the estimate of the $g_1$ which is unchanged is reproduced for completeness). One can notice that the $g_0$ estimate is less smooth than in Figure 3. But consider what happens if $\lambda_0$ is prevented from taking large values. This can be done with a prior mean of 1 and prior sd of .01. Then as shown in the right panel of Figure 4, the estimate of the $g_0$ function is even less smooth, confirming the key role that $\lambda_j$ plays in promoting smoothness. Of course, such a large number of knots are not supported by the marginal likelihood criterion. The log marginal likelihood of the model that generates the estimates in
the left panel of Figure 4 is 993.25.

(a) $\lambda_0$ allowed to freely adjust  
(b) $\lambda_0$ restricted to stay close to 1

Figure 4: Sharp design ($n = 4000$) to show how $\lambda_j$ helps to promote smoothness when there are a large number of knots: Simulated data with error distributed as $0.1295 \times t_3(0, 1)$ and 83 knots to the left of $\tau$. The estimate of the $g_0$ function (left panel) is less smooth than in Figure 3, but the posterior distribution of $\lambda_0$ concentrates on larger values which ensures a modicum of smoothness, whereas in the right panel with $\lambda_0$ restricted to near 1 with a tight prior sd, the function estimate is considerably less smooth.

3.3 Sampling investigation

It is also interesting to examine the sampling performance of the Bayes estimates of the ATE. We benchmark our against the windowed frequentist estimators of IK and CCT, both as implemented the R package rdrobust. The IK and CCT estimators rely on several estimation parameters, such as the parameter $p$ which specifies the order of the local-polynomial used to construct the point-estimator, $q$ the order of the local-polynomial used to construct the bias-correction, and one of three kernel functions. Our experiments utilize the default choices of these parameters as given in the rdrobust package.

We consider $R = 1000$ simulated data sets. The simulation is geared to examining the sampling performance along two dimensions - the sampling root mean square error (RMSE) of the estimators of the ATE and CATE, and the coverage of the 95% interval estimators. The IK estimator uses the MSE-optimal bandwidth and should be expected to produce the minimal RMSE. The CCT estimator is coverage-optimal, and by construction uses a bandwidth that is smaller than MSE-optimal. Nonetheless, we consider the RMSE and coverage of both frequentist estimators. We do the same for the Bayesian estimates, even though the Bayes estimates are developed from a conditional perspective.
The Bayes RMSE and sampling coverage are calculated from the posterior mean and the posterior SD of the ATE and the CATE, and from the .025 and .975 quantiles of the posterior distribution, across samples.

For the Bayesian results, the soft-window parameter $p$, and the knot values $m_z$ and $m_{z,\tau}$ are based on marginal likelihoods, calculated by the method of Chib (1995). Once determined, these values are used for every repeated sample. Alternatively, we could re-determine these parameters for every sample, which would improve the performance of our procedure. We have found, however, that the final sampling results are not improved greatly by this effort, largely because the selected soft window and knot values do not change much (if at all) across the repeated samples.

Finally, as in the preceding section, the prior mean and standard deviation of $\sigma^2$ are 0.3 and 1.0, respectively, and the prior mean and standard deviation of $\lambda$ are (1, 1) and (5, 5), respectively. No tuning was used to arrive at this prior to demonstrate that the performance of our approach is not dependent on a tuned prior.

We consider the same data generating process (DGP) as in the preceding section though our experiments consider the Gaussian error case with $\sigma = .1295$ and the student-t error case with 3 degrees of freedom and dispersion parameter equal to .1295. The true value of the ATE at the break-point $\tau = 0$ is 0.04.

Table 2 gives a summary of the sampling results from these experiments. The results show that in

<table>
<thead>
<tr>
<th>$n$</th>
<th>Gaussian error</th>
<th>Student-t error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Coverage</td>
<td>RMSE</td>
</tr>
<tr>
<td>$n = 500$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bayes</td>
<td>0.042</td>
<td>0.958</td>
</tr>
<tr>
<td>IK (1,2,T)</td>
<td>0.060</td>
<td>0.904</td>
</tr>
<tr>
<td>CCT (1,2,T)</td>
<td>0.054</td>
<td>0.916</td>
</tr>
<tr>
<td>$n = 4000$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bayes</td>
<td>0.040</td>
<td>0.954</td>
</tr>
<tr>
<td>IK (1,2,T)</td>
<td>0.057</td>
<td>0.801</td>
</tr>
<tr>
<td>CCT (1,2,T)</td>
<td>0.051</td>
<td>0.884</td>
</tr>
</tbody>
</table>

Table 2: Simulated data: Sharp RD designs, true value of the ATE is 0.04. Summary of results from 1000 repeated samples from Gaussian and student-t DGP’s and two different sample sizes.

the Gaussian DGP, for both sample sizes, the Bayes estimates have the smaller bias and better coverage, and either the smallest RMSE or close to the smallest RMSE. The findings are similar for the student-t DGP.
4 Fuzzy RD design

We now formalize our Bayesian approach for the fuzzy RD design. This approach is inspired by the principal stratification framework of Frangakis and Rubin (2002). The main idea is to explain the mismatch between the assignment process $I[z > \tau]$ and the treatment intake $x$ by an unobserved discrete confounder variable $s$ that represents one of three subject types (or strata): compliers, never-takers, and always-takers. We make the following assumptions.

4.1 Assumptions

Assumption 3 The unobserved confounder $s$ is a discrete random variable that represents subject type.

A subject can be of three types, a complier, never-taker or always-taker, who acts as follows:

$$x = I[z \geq \tau] \text{ if } s = c; x = 0 \text{ if } s = n, \text{ and } x = 1 \text{ if } s = a.$$  

Our next assumption is about the distribution of these types.

Assumption 4 Subject types are distributed smoothly around $\tau$ with unknown distribution $Pr(s = k) = q_k$, where $q_c + q_n + q_a = 1$.

The model for the type probability in Assumption 4 encapsulates the assumption that the distribution of type around $\tau$ is independent of $z$. As mentioned in Section 1, in the continuous confounder model, types are not modeled explicitly, but the implied distribution of type can be derived from the assumed distribution of $x$. The assumption that the distribution of $x$ is free of $z$ around $\tau$ (as mentioned above) implies that the type probabilities are free of $z$, consistent with our latter assumption.

Note that for subjects of the type $s = c$, the compliers, assignment and intake agree; that is, as $z$ passes the break-point $\tau$, the treatment state changes from 0 to 1 with probability one:

$$Pr(x = 0|z \leq \tau, s = c) = 1 \text{ and } Pr(x = 1|z > \tau, s = c) = 1 \quad (4.1)$$

On the other hand, for subjects of the type $s = n$, the never-takers, the probability that $x = 0$ is one regardless of the value of the forcing variable, $Pr(x = 0|z, s = n) = 1$, and for subjects with $s = a$, the always-takers, the probability that $x = 1$ is one regardless of the value of the forcing variable, $Pr(x = 1|z, s = a) = 1$. It follows that, for compliers, the sharp design holds.
In this set-up there are four potential outcomes: $y_0$ and $y_1$ for compliers, and $y_{0n}$ and $y_{1a}$ for never-takers and always-takers, respectively. We make the following assumption about these potential outcomes.

**Assumption 5** Conditioned on $z$, the potential outcomes $y_0$ and $y_1$ (for compliers) satisfy Assumption 1, while $y_{0n}$ (the outcome for $s = n$) satisfies

$$y_{0n} = g_{0n}(z) + \sigma_{0n}\varepsilon_{0n}$$

and $y_{1a}$ (the outcome for $s = a$) satisfies

$$y_{1a} = g_{1a}(z) + \sigma_{1a}\varepsilon_{1a}$$

both over the entire support of $z$, where the function $g_{0n}$ and $g_{1a}$ are continuous at $\tau$ and the noise terms are independently distributed as standard student-t with $\nu$ degrees of freedom.

### 4.2 Sample data

Suppose that the data consist of $n$ independent copies of $(y, x, z)$. Because observations on either side of $\tau$ can be controls or treated, it is helpful to place the data into four cells, cross-classified by $I[z \geq \tau] = l, l = 0, 1$ and $x = j, j = 0, 1$. We can indicate each of these cells by $(lj)$. The observations in each of these cells are indicated in vector notation and displayed in Table 1. The indices of the observations in each cell are denoted by $I_{00} = \{i : z_i \leq \tau, x_i = 0\}$, $I_{10} = \{i : z_i \leq \tau, x_i = 1\}$, $I_{01} = \{i : z_i > \tau, x_i = 0\}$ and $I_{11} = \{i : z_i > \tau, x_i = 1\}$. We denote the number of observations in these cells by $n_{lj}$ ($l, j = 0, 1$). We also denote the union of data down the columns of this table by a single subscript, as before, since the columns indicate the treatment state. Thus, for example, $z_0 = (z_{00}, z_{10})$ and $z_1 = (z_{01}, z_{11})$.

### 4.3 Possible types cross-classified by $z$ and $x$

In the manner of the preceding data table, one can now display the possible subject types, as shown in Table 2. Specifically, an individual in cell $(00)$ can be either a complier or never-taker; a person in cell
Table 4: Possible subject types on either side of $\tau$ by treatment state.

<table>
<thead>
<tr>
<th>$x = 0$</th>
<th>$x = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$z &lt; 0$</td>
<td>$c, n$</td>
</tr>
<tr>
<td>$z \geq 0$</td>
<td>$n$</td>
</tr>
</tbody>
</table>

This division of types by cell is key to understanding the subsequent inferential procedure for this model. It also makes clear that the fuzzy RD model with our discrete confounder is a mixture model. This is readily seen once the outcome distributions are averaged over the unknown subject types.

### 4.4 Identification of the RD CATE

Under our assumptions, the RD CATE, the ATE for compliers at $\tau$:

$$CATE = E[y_1 | z = \tau, s = c] - E[y_0 | z = \tau, s = c] = g_1(\tau) - g_0(\tau).$$

is identified. This requires showing that our fuzzy RD model is not subject to label switching (a problem otherwise inherent in mixture models). The following result speaks to this issue. We let $\mathcal{N}$ denote the density of the Gaussian distribution.

**Theorem 1** Suppose Assumptions 1-5 hold. Also suppose that $g_0(\tau) \neq g_{0n}(\tau)$ and $g_1(\tau) \neq g_{1a}(\tau)$.

Then, the mixture likelihood of the fuzzy RD model, for given independently distributed data $(y_i, x_i, z_i), i \leq n$,

$$\prod_{i \in I_{10}} \left\{ q_{ct}(y_i | g_0(z_i), \sigma^2) + q_{nt}(y_i | g_{0n}(z_i), \sigma_{0n}^2) \right\} \times \prod_{i \in I_{10}} q_{nt}(y_i | g_{0n}(z_i), \sigma_{0n}^2) \times \prod_{i \in I_{01}} q_{at}(y_i | g_{1a}(z_i), \sigma_{1a}^2) \times \prod_{i \in I_{01}} \left\{ q_{ct}(y_i | g_1(z_i), \sigma^2) + q_{at}(y_i | g_{1a}(z_i), \sigma_{1a}^2) \right\}$$

is not subject to label-switching provided the cells $I_{10}$ and $I_{01}$ are non-empty.

The proof of this result is simply based on the contribution of the data in the (10) and (01) cells. Under Assumption 3, these cells will be non-empty, at least for a sufficiently large sample. The second
and the third set of terms in the likelihood involve the parameters \( g_{0n}(\tau^+) \) and \( g_{1a}(\tau^+) \), whereas the first set of terms involve the parameters \( g_0(\tau^-) \) and \( g_{0n}(\tau^-) \), and the last set of terms involve the parameters \( g_1(\tau^+) \) and \( g_{1a}(\tau^+) \), where \( \tau^- \) and \( \tau^+ \) are points just to the left and right of \( \tau \) in cells (00) and (11), respectively. Under the supposition of the theorem, \( g_0(\tau) \neq g_{0n}(\tau) \) and \( g_1(\tau) \neq g_{1a}(\tau) \) which means that under the continuity Assumption 5, \( g_0(\tau^-) \neq g_{0n}(\tau^-) \) and \( g_1(\tau^+) \neq g_{1a}(\tau^+) \). Thus, the component distributions in the (00) and (11) cells cannot be permuted without violating the continuity assumption, or the probability distribution of the data in the (10) and (01) cells. This means that one can do inference on types in this model or, equivalently, revise our prior beliefs about the distribution of types in these four cells, and thereby estimate the component models, and hence, learn about the RD CATE.

4.5 Basis expansions

We construct the basis matrices in the same way as in the sharp model but with data taken from the appropriate cells in Table 1. For instance, for the \( g_0 \) function, we use the data \( z_{00} \), padded with \( \tau \) at the right, to locate the desired number of knots according to the soft-windowing method. These knots are given by \( \{z_{00,\min}, \kappa_{0,2}, \ldots, \kappa_{0,m_0-1}, \tau\} \). For the \( g_1 \) function, the knots are calculated from the data \( z_{11} \), padded with \( \tau \) at the left. These knots are given by \( \{\tau, \kappa_{1,2}, \ldots, \kappa_{1,m_1-1}, z_{11,\max}\} \). Then, the function ordinates

\[
g_j(z_{jj}) \triangleq \left( g_j(z_{jj,1}), \ldots, g_j(z_{jj,n_{jj}}) \right), \quad j = 0, 1
\]

are expressed using the natural cubic spline basis functions in the appendix as

\[
\begin{align*}
g_0(z_{00}) &= B_{00} \alpha, \\
g_1(z_{11}) &= B_{11} \beta,
\end{align*}
\]

respectively, where \( B_{jj} : n_{jj} \times m_j \) are the basis matrices, and \( \alpha \) and \( \beta \) are the basis coefficients. The notation \( B_{jj} \) emphasizes the fact that these matrices are based on data in the \((jj)\) cell. Under our basis, the basis coefficients are the function ordinates at the knots,

\[
\alpha_{(m_0 \times 1)} = \begin{pmatrix} g_0(z_{00,\min}) \\ g_0(\kappa_{0,2}) \\ \vdots \\ g_0(\kappa_{0,m_0-1}) \\ g_0(\tau) \end{pmatrix}, \quad \beta_{(m_1 \times 1)} = \begin{pmatrix} g_1(\tau) \\ g_1(\kappa_{1,2}) \\ \vdots \\ g_1(\kappa_{1,m_1-1}) \\ g_1(z_{11,\max}) \end{pmatrix},
\]

(4.4)
which implies that the CATE is simply the first component of $\beta$ minus the last component of $\alpha$:

$$\text{CATE} = \beta[1] - \alpha[m_a].$$  \hfill (4.5)

Now consider the functions $g_{0n}$ and $g_{1a}$. The support of these functions is given by the $z$ values in each treatment state (in other words from both sides of $\tau$), $z_j \triangleq (z'_{0j}, z'_{1j})', (n_j \times 1)$, where $n_j = n_{0j} + n_{1j}$. Our way for placing knots for these functions is as follows. Some knots are based on the data $z_{0j}$ and some are based on $(\tau, z_{1j})$, making sure that $\tau$ is one of the knots and that every pair of knots has at least one observation in between. Requiring a knot ensures that the assumption of continuity of the functions $g_{0n}$ and $g_{1a}$ at $\tau$ is satisfied. Finding such a placement of knots is relatively straightforward. Suppose then that $m_n$ knots are intended for $g_{0n}$, and $m_a$ knots for $g_{1a}$. We then use our basis functions in the appendix to expand the function ordinates

$$g_{0n}(z_0) \triangleq (g_{0n}(z_{001}), \cdots, g_{0n}(z_{10n}))$$

as

$$g_{0n}(z_0) = \begin{pmatrix} B_{00,n} \\ B_{10,n} \end{pmatrix}_{(n_0 \times m_n)} \alpha_n \triangleq B_{0,n} \alpha_n$$ \hfill (4.6)

and

$$g_{1a}(z_1) \triangleq (g_{1a}(z_{011}), \cdots, g_{1a}(z_{111}))$$

as

$$g_{1a}(z_1) = \begin{pmatrix} B_{01,a} \\ B_{11,a} \end{pmatrix}_{(n_1 \times m_a)} \beta_a \triangleq B_{1,a} \beta_a,$$ \hfill (4.7)

where $\alpha_n : m_n \times 1$ and $\beta_a : m_a \times 1$ are the basis coefficients.

### 4.6 Likelihood function

The likelihood function of $\theta \triangleq (\alpha, \beta, \alpha_n, \beta_a)$ and $\sigma^2 = (\sigma^2, \sigma_n^2, \sigma_a^2)$ follows straightforwardly from Theorem 1. Let $B_{00,i}$ denote the $i$th row of $B_{00}$, with similar notation for the other basis matrices. Then, the likelihood contribution of the $i$th observation by cell is

$$L_{00,i} = q_{0i} t_\nu(y_i | B_{00,i} \alpha, \sigma^2) + q_{n_i} t_\nu(y_i | B_{00,n,i} \alpha_n, \sigma_n^2), \ i \in I_{00}$$

$$L_{10,i} = q_{n_i} t_\nu(y_i | B_{00,n,i} \alpha_n, \sigma_n^2), \ i \in I_{10}$$

24
\[
L_{01,i} = q_a t_r(y_i | B_{01,a,i} \beta_a, \sigma_a^2), \ i \in I_{01}
\]
\[
L_{11,i} = q_c t_r(y_i | B_{11,i} \beta, \sigma^2) + q_a t_r(y_i | B_{11,a,i} \beta_a, \sigma_a^2), \ i \in I_{11}
\]

and the likelihood function is the product of these contributions over all the observations:

\[
L = \prod_{i \in I_{00}} L_{00,i} \times \prod_{i \in I_{10}} L_{10,i} \times \prod_{i \in I_{01}} L_{01,i} \times \prod_{i \in I_{11}} L_{11,i}.
\]

### 4.7 Prior

Except for an increase in dimension of the parameter space, the prior on the parameters is specified in the manner of the sharp RD model. The prior on \((\alpha, \beta)\) is exactly the same as in (2.9) and (2.12) except that the matrices \(D_\alpha\) and \(D_\beta\), which have the form given in the Appendix, are built up from the data in the cells (00) and (11) respectively. The prior on the parameters \(\alpha_n\) and \(\beta_a\) of the \(n\) and \(a\) models, is given by

\[
\alpha_n | \lambda_n \sim N_{m_n} \left( D_n^{-1} \beta_{0n}, \lambda_n^{-1} D_n^{-1} T_n D_n^{-1} \right)
\]
\[
\beta_a | \lambda_a \sim N_{m_a} \left( D_a^{-1} \beta_{0a}, \lambda_a^{-1} D_a^{-1} T_a D_a^{-1} \right)
\]

where the matrices \(D_n\) and \(D_a\) are constructed analogously to \(D_\alpha\). The penalty parameters, now given by \((\lambda_0, \lambda_1, \lambda_n, \lambda_a)\), is each specified a prior distribution as in the sharp model. In the examples below, for simplicity, we specify a Gamma distribution for each \(\lambda\) with a prior mean of 1 and prior standard deviation of 10. Finally, a prior distribution on the type probabilities \(q = (q_c, q_n, q_a)\) is needed. Following the usual custom, this prior is taken to be Dirichlet with parameters \((n_{0c}, n_{0n}, n_{0a})\). We normally set these hyperparameters to reflect the belief that half the sample consists of compliers and that the remaining half is equally divided between never-takers and always-takers.

### 4.8 MCMC sampling

Estimation of the fuzzy RD model relies on the usual augmentation of the mixture indicators, here the type variables \(s_i\) \((i \leq n)\), and the augmentation of the Gamma random variables in the scale mixture of normals representation of the student-t distribution. Conditioned on the parameters, these type variables have to be sampled only in the (00) and (11) cells because the subjects in the cells (10) and (01) are necessarily of the types \(n\) and \(a\), respectively. From the likelihood contributions given above, for observations in the set \(I_{00}\), conditioned on the data and \((\theta, \sigma^2)\),

\[
Pr(s_i = c | y_i, \theta, \sigma^2) \propto q_c t_r(y_i | B_{00,c} \alpha, \sigma^2)
\]
\[
\Pr(s_i = n|y_i, \theta, \sigma^2) \propto q_n t_\nu(y_i | B_{00,n}, \alpha_n, \sigma_n^2)
\]
and for observations in the set \(I_{11}\)

\[
\Pr(s_i = c|y_i, \theta, \sigma^2) \propto q_c t_\nu(y_i | B_{11,i}, \beta, \sigma^2)
\]
\[
\Pr(s_i = a|y_i, \theta, \sigma^2) \propto q_a t_\nu(y_i | B_{11,a}, \beta_a, \sigma_a^2)
\]

Suppose that in a particular MCMC iteration, in the cell \((00)\), the sampling of \(\{s_i\}\) with these probabilities produces \(n_{00}^c\) compliers and \(n_{00}^n = n_{00} - n_{00}^c\) never-takers. Similarly, suppose that in the cell \((11)\), the sampling produces \(n_{11}^c\) compliers and \(n_{11}^n = n_{11} - n_{11}^c\) always-takers. Given the sampled types, the probabilities \(q = (q_c, q_n, q_a)\) are sampled from an updated Dirichlet distribution with parameters

\[
(n_{0c} + n_{00}^c + n_{11}^c, n_{0n} + n_{00}^n + n_{10}, n_{0a} + n_{01} + n_{11}^a).
\]

Conditioned on the sampled types, the posterior distribution decomposes into three independent distributions, one for each type. These distributions can be calculated as follows. For the observations classified as \(s = c\), using the basis matrices in (4.3) we can write

\[
\begin{pmatrix}
\hat{y}_{00} \\
\hat{y}_{11}
\end{pmatrix}
= \begin{pmatrix}
B_{00}^c & 0 \\
0 & B_{11}^c
\end{pmatrix}
\begin{pmatrix}
\alpha \\
\beta
\end{pmatrix}
+ \begin{pmatrix}
\epsilon_{00}^c \\
\epsilon_{11}^c
\end{pmatrix},
\]

(4.10)

where the \(c\) superscript indicates the sub-vectors and sub-matrices consisting of the rows (observations) classified as compliers in the indicated cells, and each component of the error conditioned on \(\{\xi^c_i\}\) is distributed as \(N(0, \sigma^2/\xi^c_i)\). This model of the data is analogous to (2.6) in the sharp RD model. Therefore, the posterior distribution of the parameters \((\alpha, \beta), (\lambda_0, \lambda_1), \sigma^2\) and \(\{\xi^c_i\}\), conditional on \(s_i\) \((i \leq n)\), can be sampled according to one step of the sharp MCMC algorithm.

Similarly, given the \(n_{00}^n\) observations sampled as never-takers in the cell \((00)\), using the basis matrices in (4.6) we can write

\[
\begin{pmatrix}
\hat{y}_{00}^n \\
\hat{y}_{10}
\end{pmatrix}
= \begin{pmatrix}
B_{00,n}^0 & 0 \\
0 & B_{10,n}
\end{pmatrix}
\begin{pmatrix}
\alpha_n \\
\beta_n
\end{pmatrix}
+ \begin{pmatrix}
\epsilon_{00,n}^0 \\
\epsilon_{10,n}
\end{pmatrix},
\]

(4.11)

where \(\hat{y}_{00}^n\) and \(B_{00,n}^0\) consist of the rows of \(y_{00}\) and \(B_{00,n}\) in cell \((00)\) that are classified as never-takers, and each component of the error, conditioned on \(\{\xi^n_i\}\) is distributed as \(N(0, \sigma^n/\xi^n_i)\). Again, the model of these data is analogous to that of sharp RD model and, therefore, the conditional posterior.
distribution of the parameters $\alpha_n$, $\lambda_n$, $\sigma^2_n$ and $\{\xi^n_i\}$ can be sampled using one step of the sharp MCMC algorithm.

Last, given the $n_{11}^a$ observations classified as always-takers in the cell (11), using the basis matrices in (4.7) we can write

$$
\begin{pmatrix}
  y_{01} \\
  y_{11}^a
\end{pmatrix}
\begin{pmatrix}
  n_{01} \times 1 \\
  n_{11}^a \times 1
\end{pmatrix}
= 
\begin{pmatrix}
  B_{01,a} \\
  B_{11,a}^a
\end{pmatrix}
\begin{pmatrix}
  \beta_a \\
  \varepsilon_{11,a}
\end{pmatrix}
+ 
\begin{pmatrix}
  \varepsilon_{01,a} \\
  \varepsilon_{11,a}^a
\end{pmatrix}
\begin{pmatrix}
  n_{01} \times 1 \\
  n_{11}^a \times 1
\end{pmatrix}
$$

(4.12)

where $y_{11}^a$ and $B_{11,a}^a$ consist of the rows of $y_{11}$ and $B_{11,a}$ in cell (11) that are classified as always-takers, and each component of the error, conditioned on $\{\xi^a_i\}$ is distributed as $\mathcal{N}(0, \sigma^2_a/\xi^a)$. This shows that conditioned on $\{s_i\}$, the parameters $\beta_a$, $\lambda_a$, $\sigma^2_a$ and $\{\xi^a_i\}$ can be sampled as in the sharp model.

These simulation steps, which constitute one iteration of the MCMC algorithm in the fuzzy RD model, are repeated, and beyond the burn-in phase, the last element of $\alpha$ and the first element of $\beta$ are extracted to create drawings of the CATE from its posterior distribution.

Finally, for any version of the fuzzy RD model – defined by differing number of knots and differing soft-window widths– the marginal likelihood can be calculated by the method of [Chib (1995)]. The details are straightforward and hence omitted.

5 Example: Fuzzy design

5.1 Discrete confounder

To illustrate our ideas, we consider simulated data from a relatively simple design that satisfies the conditions of Theorem 1. It has four unrestricted mean functions and is defined by the following specification:

$$
y_{0c} = -8z + z^2 + z^3 + \sin(2z) + \varepsilon_0 \\
y_{1c} = 4z + z^2 + z^3 + 5\sin(z) + 1 + \varepsilon_1 \\
y_{0n} = 0 + 10z + \varepsilon_n \\
y_{1a} = 3 - 20z + \varepsilon_{1a}
$$

where $z \sim 0.1295 \times \mathcal{N}(0,1)$, $\tau = 0$, $\varepsilon_0$ and $\varepsilon_1$ are standard normal, and $\varepsilon_{0n}$ and $\varepsilon_{1a}$ are .5 and 2 times standard normals, respectively, and the types $s \in \{c, n, a\}$ are generated from a discrete distribution with probabilities $q = (0.5, 0.25, 0.25)$. In this design, the true value of the CATE is 1.0. We also consider the same design with student-t errors.
by letting $\varepsilon_0$ and $\varepsilon_1$ be distributed as standard student-t, and $\varepsilon_{0n}$ and $\varepsilon_{1a}$ distributed as .5 and 2 times standard student-t, respectively. Our data consists of one sample each of sizes $n = 500$ and $n = 4000$ (which are used to conduct the conditional Bayes analysis) and 1000 repeated samples which are used to evaluate the sampling properties of the CATE estimate.

In setting up our prior, we assume that a priori, $(\sigma^2, \sigma^2_n, \sigma^2_a)$ have a mean equal to $(2, 2, 2)$ and standard deviation equal to $(10, 10, 10)$, and that the four smoothness parameters $(\lambda_0, \lambda_1, \lambda_n, \lambda_a)$ have a mean equal to $(1, 1, 1, 1)$ and standard deviation equal to $(5, 5, 5, 5)$. A priori we also assume that $q$ is Dirichlet with hyperparameters equal to $(2, 2, 2)$. In addition, on the basis of marginal likelihood comparisons, for $n = 500$, the soft-windowing parameter $p$ is $(0.5, 0.5)$, and for $n = 4000$ it is $(0.6, 0.4)$. For both sample sizes, the knots for the four functions are set by the values $m_z = (3, 3)$, $m_{z,n} = 5$ and $m_{z,a} = 5$. Finally, for $n = 500$, $m_{z,\tau} = (2, 2)$, and for $n = 4000$, $m_{z,\tau} = (3, 3)$.

5.1.1 Conditional analysis

Suppose we were to apply our procedure to each sample size for data generated from $t_3$ distributed errors. The results given in Table 5 show that even for the smaller sample size, which is a particularly challenging case, the 95% posterior credibility intervals include the true values and that the marginal posterior distributions concentrate around the true values with the sample size.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True Value</th>
<th>Prior</th>
<th>Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>mean</td>
<td>std dev</td>
</tr>
<tr>
<td>$n = 500$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>1</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>$\sigma^2_n$</td>
<td>0.25</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>$\sigma^2_a$</td>
<td>4</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>$q_c$</td>
<td>0.5</td>
<td>0.33</td>
<td>0.178</td>
</tr>
<tr>
<td>$q_n$</td>
<td>0.25</td>
<td>0.33</td>
<td>0.178</td>
</tr>
<tr>
<td>$q_a$</td>
<td>0.25</td>
<td>0.33</td>
<td>0.178</td>
</tr>
<tr>
<td>$n = 4000$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>1</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>$\sigma^2_n$</td>
<td>0.25</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>$\sigma^2_a$</td>
<td>4</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>$q_c$</td>
<td>0.5</td>
<td>0.33</td>
<td>0.178</td>
</tr>
<tr>
<td>$q_n$</td>
<td>0.25</td>
<td>0.33</td>
<td>0.178</td>
</tr>
<tr>
<td>$q_a$</td>
<td>0.25</td>
<td>0.33</td>
<td>0.178</td>
</tr>
</tbody>
</table>

Table 5: Fuzzy RD with discrete confounder and student $t_3$ errors: summary results for the $t_3$ dispersion parameters $(\sigma^2, \sigma^2_n, \sigma^2_a)$ and probabilities of types. Inefficiency factors in the last column. For each parameter, the 95% posterior credibility intervals include the true values and the marginal posterior distributions concentrate around the true values with the sample size.
procedure to infer the mixture type probabilities is particularly striking.

Also interesting is to consider the inferences about the four smoothness parameters \((\lambda_0, \lambda_1, \lambda_n, \lambda_a)\). The results, given in Table 6, show that the marginal posterior distribution of \(\lambda_0\) is relatively more dispersed and includes some mass on larger values. In addition, since the \(g_n\) and \(g_a\) functions in this DGP are linear, enforcing smoothness on the second differences of the basis coefficients through the prior is unnecessary with few knots and, correctly, the marginal posterior distributions of the corresponding smoothness parameters are concentrated on values close to zero. Note that the posterior distributions of \(\lambda_0\) and \(\lambda_1\) also concentrate on small values with sample size because the number (and proportion) of knots used in the estimation of the \(g_0\) and \(g_1\) functions remain small. The Bayes estimates of the four functions are given in Figure 5. In this figure, the true value of the functions are the dotted lines, the estimates are the solid lines, the 95% point-wise credibility intervals of the functions are the shaded bands, and the distribution of the z values is notched on the horizontal axis. As can be seen from this figure, the functions are well estimated. Finally, for this sample size, the posterior mean of the CATE is 1.007, with 95% credibility interval equal to (0.542, 1.463). The true value of the CATE as mentioned above is 1.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Prior</th>
<th>Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>std dev</td>
</tr>
<tr>
<td>(\lambda_0)</td>
<td>1 5</td>
<td>3.877 0.171</td>
</tr>
<tr>
<td>(\lambda_1)</td>
<td>1 5</td>
<td>1.275 0.084</td>
</tr>
<tr>
<td>(\lambda_n)</td>
<td>1 5</td>
<td>0.086 0.009</td>
</tr>
<tr>
<td>(\lambda_a)</td>
<td>1 5</td>
<td>0.039 0.006</td>
</tr>
</tbody>
</table>

Table 6: Fuzzy RD with discrete confounder and student t_3 errors: summary results for \((\lambda_0, \lambda_1, \lambda_n, \lambda_a)\). The marginal posterior distributions of \(\lambda_n\) and \(\lambda_a\) are concentrated on small values because the underlying functions \(g_n\) and \(g_a\) in this DGP are linear and there are only few knots involved in the fitting. Because the number of knots in the fitting of \(g_0\) and \(g_1\) also remains small with sample size, the enforcement of the smoothness condition is less important, and the marginal posterior distributions of \(\lambda_0\) and \(\lambda_1\) show the move to small values.
Figure 5: Fuzzy RD with discrete confounder and student $t_3$ errors: This shows the function estimates and credibility bands for $n = 4000$. Note that, as required by our conditions, the functions $g_n$ and $g_0$ are both continuous at $\tau$. This is achieved by having a knot at $\tau$. See text for further details.

### 5.1.2 Sampling experiments

Next, we examine the frequentist properties of the Bayes estimates of the CATE and compare its sampling performance with that of the IK and CCT estimators. Our sampling results are based on 1000 replications and are given in Table 7. Because the confounder is discrete, the frequentist estimators, which implicitly assume that the confounder is continuous, have larger bias and RMSE than the Bayes estimator, though the performance of the frequentist estimators improves with the sample size. Importantly, though, even in the $n = 4000$ case, the RMSE of the frequentist estimators is 4-5 times larger than that of the Bayesian estimator.
Table 7: Summary of results from 1000 repeated samples: fuzzy RD design, discrete confounder with Guasian and t3 errors and two sample sizes, true value of the CATE is 1.0.

### 5.2 Continuous confounder

We now consider a fuzzy RD design in which the confounder is continuous, to assess the degree to which the performance of our approach is affected by a key assumption that is counter to our modeling. Rather interestingly, even though this design does not conform to our modeling, our approach performs satisfactorily. Once again, we compare the results from our approach with those from the fuzzy estimators of Imbens and Kalyanaraman (2012) and Calonico et al. (2014).

The data for our first set of experiments is generated from a design that appears in Frandsen et al. (2012). Potential outcomes in this design are generated as $y_j = g(z) + \varepsilon_j$, where $z \sim \mathcal{N}(0, 1)$ and $\varepsilon_j \sim \mathcal{N}(0, 1)$, and the treatment $x$ is generated as $x = I[y_1 - y_0 + \alpha_0 + \alpha_1 I(z \geq 0) + \varepsilon_x \geq 0]$, where $\varepsilon_x \sim \mathcal{N}(0, 1)$. The continuous confounder in this model is $y_1 - y_0 = \varepsilon_1 - \varepsilon_0$. As mentioned in the Introduction, in a model such as this, which satisfies the conditions of Theorem 3 in Hahn et al. (2001), the mean value of the outcomes for a complier, never-taker, or always-taker, are intercept shifts of the function $g(z)$. In particular,

$$
\mathbb{E}[y_j|z, s = c] = g(z) + \mathbb{E}[\varepsilon_j|-(\alpha_0 + \alpha_1) < \varepsilon_1 - \varepsilon_0 + \varepsilon_x \leq -\alpha_0]
$$

$$
\mathbb{E}[y_0|z, s = n] = g(z) + \mathbb{E}[\varepsilon_0|\varepsilon_1 - \varepsilon_0 + \varepsilon_x \leq -\alpha_0] \quad \text{and} \quad \mathbb{E}[y_1|z, s = a] = g(z) + \mathbb{E}[\varepsilon_1|\varepsilon_1 < \varepsilon_1 - \varepsilon_0 + \varepsilon_x].
$$

Thus, each of these expectations equals the function of $z$ plus a constant value.

We let $g(z) = z^2$ and consider two cases of this design, the first defined with the values $\alpha_0 = -0.70$ and $\alpha_1 = 1.75$, which generates about 40% compliers and a CATE of $-0.105$, and a second design in which $\alpha_0 = -0.70$ and $\alpha_1 = 2.50$, which generates about 55% compliers and a CATE of $-0.307$.

In the fitting of our Bayesian discrete confounder model, we model the four mean functions by an unknown function of $z$ with intercept shifts for each of the types. This model does not require a soft
window specification and has only two λ’s. Not surprisingly, this simpler model is heavily preferred on the marginal likelihood criterion relative to the general model with fully unrestricted mean functions. The prior means of all variance parameters are assumed to be 2.0 with a prior standard deviation of 5, and the Dirichlet prior of \( \mathbf{q} \) is defined by the hyperparameters \((n_0c, n_{00}, n_{0a}) = (2, 2, 2)\). As above, we determine the number of knots by comparing the marginal likelihood of models with varying numbers of knots and arrive at a specification with 6 knots for \( n = 500 \) and 7 knots for \( n = 4000 \) to model the function \( g(z) \).

Results are reported in Table 8. Unexpectedly, the two frequentist estimators, which are designed for this continuous confounder case, do not perform well in estimating the CATE when the sample size is small. Interestingly, the Bayesian estimate of the CATE, despite coming from a mis-specified model, has a sampling RMSE that is smaller than even the MSE-optimal frequentist estimator, for both sample sizes, and for both levels of compliance. The Bayesian coverage is also close to the nominal value and comparable to those of the coverage-optimal CCT estimators, results that are especially significant given that, in this setting, the confounder is continuous, counter to our assumption.

<table>
<thead>
<tr>
<th></th>
<th>( \alpha = 1.75 )</th>
<th>( \alpha = 2.50 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( n = 500 )</td>
<td>( n = 4000 )</td>
</tr>
<tr>
<td>( n )</td>
<td>Mean</td>
<td>Coverage</td>
</tr>
<tr>
<td>Bayes</td>
<td>-0.074</td>
<td>0.976</td>
</tr>
<tr>
<td>IK (1,2,T)</td>
<td>-0.225</td>
<td>0.986</td>
</tr>
<tr>
<td>CCT (1,2,T)</td>
<td>-0.680</td>
<td>0.983</td>
</tr>
</tbody>
</table>

Table 8: Summary of results from 1000 repeated samples.: fuzzy RD design with Gaussian errors, continuous confounder. True value of the CATE is \(-0.105\) for \( \alpha = 1.75 \), and \(-0.307\) for \( \alpha = 2.50 \).

### 6 Example: Senate data

For an application of the sharp RD design to real data, we consider the example discussed in Calonico et al. (2014) and Cattaneo, Frandsen, and Titiunik (2015). Data for this example are contained in the R package `rdrobust`. This example is concerned with the size of a party-level incumbency effect in U. S. Senate elections. Specifically, the question is what effect a particular party winning a seat in a Senate election has on the vote share for the same party in the election for the same seat six years later. In
previous work, Lee (2008) considered this question in elections to the U. S. House of Representatives and found an incumbency effect of approximately 8 percentage points.

The incumbency effect problem can be cast into the sharp RD framework by setting \( z_{it} = 2(DS_{it} - 50) \), where \( DS_{it} \) is the share of the vote won by the candidate of the Democratic Party in state \( i \) in year \( t \), and letting \( x_{it} = I[z_{it} > 0] \), in which case \( x_{it} = 1 \) if the Democratic candidate wins, and 0 otherwise. The outcome variable \( y_{i,t+6} \) is the percentage of votes won by the Democratic Party candidate in the election for the same seat six years later.

On the basis of the marginal likelihood ranking, we settle on the Bayes model defined by \( p = (0.8, 0.2) \), knot values \( m_z = (3, 3) \) and \( m_{z,\tau} = (2, 2) \), prior mean and variance of \( \sigma^2 \) equal to 75 and 150, respectively, and \( \alpha_0 \) and \( \beta_0 \) equal to the zero vectors.

Table 9 displays the Bayes results and the IK and CCT estimates for different procedural parameters. All three estimates indicate a clear party-level incumbency effect, with estimates ranging from 7.4 to 8.9 points. The Bayes 95% interval estimate is 5.229 to 12.540, while the frequentist interval estimates are broadly similar. These results are comparable to Lee’s findings for the House incumbency effect. We note that the analysis of Cattaneo et al. (2015), with data on additional confounder variables that are not available to us, produces a party-incumbency effect of 9.32 and a 95% confidence interval of [4.60, 14.78].

For a different summary of the Bayesian results, consider Figure 5 where the top panel has the Bayes point and credibility estimates of the \( g_0 \) and \( g_1 \) functions, and the bottom panel has the posterior distribution of the ATE. Interestingly, there is a slight dip in the slope of \( g_0 \) near \( \tau \) that also appears in the graphical output of the \texttt{rdrobust} package (not shown). This dip can be explained by the data: The mean value of \( y \) in the interval between -2.494 (the largest knot less than \( \tau \) of the \( g_0 \) function) is 44.068. Comparing values of the \( y \) mean near that knot, we find that it is 46.483 for \( z \in [-4, -2.494] \) and 44.225 for \( z \in [-6, -4] \). Thus, the \( y \) values that determine the \( g_0 \) function have a tendency to fall

<table>
<thead>
<tr>
<th>Method</th>
<th>Estimate</th>
<th>sd</th>
<th>lower</th>
<th>upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayes</td>
<td>8.921</td>
<td>1.882</td>
<td>5.229</td>
<td>12.540</td>
</tr>
<tr>
<td>IK (1,2,T)</td>
<td>7.425</td>
<td>1.495</td>
<td>4.494</td>
<td>10.356</td>
</tr>
<tr>
<td>CCT (1,2,T)</td>
<td>7.527</td>
<td>1.764</td>
<td>4.070</td>
<td>10.983</td>
</tr>
<tr>
<td>CCT (2,3,T)</td>
<td>8.006</td>
<td>2.031</td>
<td>4.026</td>
<td>11.987</td>
</tr>
</tbody>
</table>

Table 9: Senate data: Summary of the Bayes and IK and CCT estimates, where for IK and CCT, the values of \( p, q \) and the kernel (triangular or uniform), are indicated in brackets. Estimation settings for Bayes results are based on marginal likelihoods and are given in the text.
as they pass the last knot before $\tau$. Although we cannot explain why this occurs, it makes clear that the estimated function is accurately reflecting the information contained in the data.

7 Conclusions

In this paper, we have introduced several novel ideas in the analysis of the sharp and fuzzy RD designs. First, we specify a new, flexible and novel second-difference prior on the spline coefficients that is capable of handling the situation of many unequally spaced knots. The information required of the investigator – essentially a rough idea of the first two ordinates at the extreme points on both sides of $\tau$ – should be known to an investigator with knowledge of the specifics of the application. Second, we include a knot at the threshold, which is not in general an observed value of $z$, to allow for curvature in the estimated function from the breakpoint to the nearest $z$ value on either side of the breakpoint. Third, our procedure allows for the clustering of knots close to the threshold with the aim of controlling the approximation bias. The number of knots and other inputs into the model can be compared
through marginal likelihoods and Bayes factors. Fourth, we introduce a probability model for the fuzzy RD design, inspired by the principal stratification framework, in which the unobserved confounder is modeled as a discrete random variable, in a departure from the assumption made in the literature to date. The models and estimation approaches are easily implemented through available R packages, and comparisons show that the Bayesian RD CATE estimates perform satisfactory even when the confounder is continuous, counter to our assumption, while the frequentist estimators do less well when the confounder is discrete, especially in small samples.

The framework we have provided can be extended in different directions. For instance, we can consider binary and categorical outcomes in both designs by taking recourse to the latent variable modeling of Albert and Chib [1993]. It is also possible to extend our framework to multivariate outcomes and multiple thresholds. Finally, one can further robustify the modeling assumptions by modeling the distribution of the potential outcomes by the approach of say Kundu and Dunson [2014]. These extensions are ongoing and will be reported elsewhere.

A Appendix: Basis functions

In this appendix, we let $g(\cdot)$ denote any function that is to be represented by a cubic spline and let $z \in \mathbb{R}$ denote its argument. Let $\kappa_j (j = 1, \ldots, m)$ denote the knots, and $h_j = \kappa_j - \kappa_{j-1}$ the spacing between the $(j-1)$st and $j$th knots. The basis functions are the collections of cubic splines $\{\Phi_j(z)\}_{j=1}^m$ and $\{\Psi_j(z)\}_{j=1}^m$, where for $2 \leq j \leq m - 1$,

$$
\Phi_j(z) = \begin{cases} 
0, & z \geq \kappa_j+1 \\
-(2/h_j^3)(z - \kappa_{j-1})^2(z - \kappa_j - 0.5h_j), & \kappa_j-1 \leq z < \kappa_j \\
(2/h_{j+1}^3)(z - \kappa_j+1)^2(z - \kappa_{j} + 0.5h_{j+1}), & \kappa_j \leq z < \kappa_{j+1} \\
0, & z < \kappa_{j-1}, 
\end{cases}
$$

(A.1)

$$
\Psi_j(z) = \begin{cases} 
0, & z \geq \kappa_{j+1} \\
(1/h_j^2)(z - \kappa_{j-1})^2(z - \kappa_j), & \kappa_{j-1} \leq z < \kappa_j \\
(1/h_{j+1}^2)(z - \kappa_{j+1})^2(z - \kappa_{j}), & \kappa_j \leq z < \kappa_{j+1} \\
0, & z < \kappa_{j-1}, 
\end{cases}
$$

(A.2)

and for $j = 1$, $\Phi_j$ and $\Psi_j$ are defined by the last two lines of equations (A.1) and (A.2), respectively, and for $j = m$, $\Phi_j$ and $\Psi_j$ are defined by the first two lines. In these two cases, the strong inequality at the upper limit is replaced by a weak inequality.
The representation of $g(z)$ as a natural cubic spline is given by

$$g(z) = \sum_{j=1}^{m} \Phi_j(z) f_j + \sum_{j=1}^{m} \Psi_j(z) s_j,$$

(A.3)

where $f = (f_1, \ldots, f_m)'$ and $s = (s_1, \ldots, s_m)'$ are the coefficients of this cubic spline. Conveniently, $f_j = g(\kappa_j)$ is the function value at the $j$th knot, and $s_j = g'(\kappa_j)$ is the slope at the $j$th knot.

The fact that $g(z)$ is a natural cubic spline implies that $g''(\kappa_1) = 0 = g''(\kappa_m)$ and that the second derivatives are continuous at the knot points. These conditions place restrictions on the $s_j$. If we define $\omega_j = h_j / (h_j + h_{j+1})$, and $\mu_j = 1 - \omega_j$ for $j = 2, \ldots, m$, then Lancaster and Šalkauskas (1986, Sec. 4.2) show that the ordinates and slopes are related by the relations

$$C f = A s, \quad s = A^{-1} C f,$$

where

$$A = \begin{pmatrix}
2 & 1 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 \\
\omega_2 & 2 & \mu_2 & 0 & 0 & \ldots & 0 & 0 & 0 \\
0 & \omega_3 & 2 & \mu_3 & 0 & \ldots & 0 & 0 & 0 \\
\vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & 0 & \ldots & \omega_{m-1} & 2 & \mu_{m-1} \\
0 & 0 & 0 & 0 & 0 & \ldots & 0 & 1 & 2 \\
\end{pmatrix},$$

and

$$C = 3 \begin{pmatrix}
-\frac{1}{h_2} & -\frac{\omega_2}{h_2} & \frac{\mu_2}{h_2} & \frac{\mu_2}{h_2} & \frac{\mu_2}{h_2} & \ldots & 0 & 0 & 0 \\
-\frac{\omega_3}{h_2} & \frac{\omega_2}{h_2} & -\frac{\mu_2}{h_2} & \mu_2 & \frac{\mu_2}{h_2} & \ldots & 0 & 0 & 0 \\
0 & -\frac{\omega_3}{h_3} & -\frac{\omega_3}{h_4} & \frac{\mu_2}{h_4} & \frac{\mu_2}{h_4} & \ldots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \ldots & -\frac{\omega_{m-1}}{h_{m-1}} & \omega_{m-1} & \omega_{m-1} & 0 \\
0 & 0 & 0 & 0 & \ldots & 0 & -\frac{\mu_{m-1}}{h_{m-1}} & \mu_{m-1} & \mu_{m-1} \\
\end{pmatrix}.$$  

For any observation of $z$, $z_i$, it follows that $g(z_i)$ in (A.3) can be re-expressed as

$$g(z_i) = \Phi(z_i)' f + \Psi(z_i)' A^{-1} C f$$

or as $g(z_i) = c_i' f$, where $\Phi(z_i)' = (\Phi_1(z_i), \ldots, \Phi_m(z_i))$, $\Psi(z_i)' = (\Psi_1(z_i), \ldots, \Psi_m(z_i))$ and $c_i = \Phi(z_i)' + \Psi(z_i)' A^{-1} C$, which implies the following representation for the $n \times m$ basis matrix: $B = (c_1, \ldots, c_n)' = [b_1, \ldots, b_m]$. This is the form of the basis matrices $B_0$ and $B_1$. 

36
B Appendix: $D_\alpha$, $D_\beta$

Suppose that $m$ is the dimension of $\alpha$ and $\beta$. Then, the matrices $D_\alpha$ and $D_\beta$ in equations (2.9) and (2.12) take the following forms:

$$D_\alpha = \begin{pmatrix}
1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\
0 & 1 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\frac{1}{h_{0,3}} & \sqrt{h_{0,3}} & \frac{1}{h_{0,3}} & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & \frac{1}{h_{0,4}} & \sqrt{h_{0,4}} & h_{0,4} & \cdots & \cdots & \cdots & \cdots \\
\vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & 0 & 0 & 0 & 0 & \cdots & \cdots & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{h_{0,m-1}} & \sqrt{h_{0,m-1}} \\
\end{pmatrix}$$

and

$$D_\beta = \begin{pmatrix}
\frac{1}{h_{1,2}} & \frac{1}{h_{1,2}} & (1-h_{1,2}) & 0 & \cdots & 0 & 0 & 0 \\
0 & 0 & \frac{1}{h_{1,3}} & \sqrt{h_{1,3}} & \cdots & \cdots & \cdots & \cdots \\
0 & 0 & \frac{1}{h_{1,4}} & \sqrt{h_{1,4}} & \cdots & \cdots & \cdots & \cdots \\
\vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\vdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & 0 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{h_{1,m-1}} & \sqrt{h_{1,m-1}} \\
\end{pmatrix}$$

respectively.

References


