Nonparametric Bayes Analysis of the Sharp and Fuzzy Regression Discontinuity Designs

Siddhartha Chib *
Edward Greenberg †

March 17, 2014

Abstract

We develop a Bayesian analysis of the sharp and fuzzy RD designs in which the unknown functions of the forcing variable and the other covariates are modeled by penalized natural cubic splines, and the errors are distributed as either Gaussian or 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 many unequally spaced knots. The number of knots and other features of the model are compared through marginal likelihoods and Bayes factors. Fourth, we develop an analysis of the fuzzy RD design based on a new model that utilizes the principal stratification framework, adapted to the RD design. In this model, the sharp RD model holds for compliers, while the outcome models for never-takers and always-takers are assumed to satisfy the usual exclusion rule with respect to the forcing variable. In each design, posterior computations are straightforward and are implemented in two R-packages that may be downloaded. Calculations with simulated data show that the frequentist RMSE of the Bayes ATE estimate (in the sharp model) and the Bayes ATE for compliers (in the fuzzy model) are smaller than that of the frequentist local ATE estimate in all examples, in several cases by a significant factor.

*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@artsci.wustl.edu.

1 Introduction

The sharp and fuzzy regression discontinuity (RD) designs are simple but powerful quasi-experimental designs for conducting causal inferences with observational data. Interest in these designs has grown substantially in recent years following the realization that many cause-effect problems in such fields as economics, finance, political science and other fields can be viewed in terms of one or the other RD design (Imbens and Lemieux, 2008, Lee and Lemieux, 2010).

At their core, these designs rely on the fact that treatment assignment in some problems is determined by an exogenously specified rule. The rule is characterized by a forcing variable z and a known break-point τ. In the sharp RD design, subjects with a value of z less than or equal to τ receive the control treatment $x = 0$, while those with z value greater than the break-point are the treated and receive the treatment $x = 1$. Thus, in this case, treatment assignment is a deterministic function of the forcing variable. In contrast, in the fuzzy RD design, treatment assignment is a non-deterministic function of the forcing variable but with the property that the treatment probability changes discontinuously at the break-point. As a result, some subjects to the left of the break-point, 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. In effect, in the fuzzy RD design, the forcing variable along with the assignment rule is an instrument.

Such rules can be exploited to find the relevant RD treatment effects. Let y be the outcome of interest, and suppose for concreteness that it is continuous. If it can be assumed that z is not manipulable and that the distributions of z and other observed confounders w is smooth around τ, then, in the sharp design, the RD average treatment effect is given by the size of the discontinuity in the relation between y and z at τ, and it can be consistently estimated (Hahn et al., 2001). In the fuzzy RD design, these and additional assumptions imply the consistency and identification of the LATE estimator, a version of the instrumental variable treatment effect (Hahn et al., 2001).

In this paper, we provide the first Bayesian analysis of these designs. We employ non-parametric modeling and use flexible and novel priors for the non-parametric functions. Let

y_j denote the potential outcome when $x = j$, for $j = 0, 1$. The observed outcome is $y = (1 - x)y_0 + xy_1$. Let w denote a single continuous confounder (for simplicity), and let v denote q categorical and linear confounders not including an intercept. Then, in the sharp RD design, the model we analyze assumes that the potential outcomes y_j conditioned on (z, w, v) are generated as

$$y_j = g_j(z) + v'\gamma + h(w) + \varepsilon_j, \tag{1.1}$$

where $g_j(\cdot)$ are smooth unknown functions of z, one for each value of value of x, $h(\cdot)$ is a smooth function that depends on w but not x, and ε_j is the random noise that is independent of z. We assume that the distribution of the noise is Gaussian or Student-t with degrees of freedom not less than two. We refrain from modeling the noise distribution non-parametrically to focus on the main features of our approach, but it is straightforward to extend our analysis to a fully non-parametric model. Given this model and a random sample of data on n subjects, we show how to estimate the posterior distribution of the RD ATE

$$\mathsf{E}[y_1|z = \tau, w, v] - \mathsf{E}[y_0|z = \tau, w, v] = g_1(\tau) - g_0(\tau). \tag{1.2}$$

In our approach, we do not limit the data to a window around the threshold, as is done in the local linear regression method of Imbens and Kalyanaraman (2012). Our modeling of the unknown functions is by penalized splines. These are a compromise between regression splines, which have fewer knots than data points and no penalty, and smoothing splines, which have one knot for each data point and a penalty for roughness of the fit (Ruppert et al., 2003). We employ splines because they are flexible and easy to work with, and they have known asymptotic properties Zhou et al. (1998), Claeskens et al. (2009).

Our choice of the number of knots is guided by a result in Claeskens et al. (2009) that fewer knots provide better asymptotic rates and lower mean-squared errors than a large number of knots. We allow the number of knots to grow moderately with the sample size, according to the rates in Claeskens et al. (2009), but the distribution of knots, for any given sample size, is tailored to the specifics of the RD model. A key idea in our approach is to have a knot at τ in the basis expansion of each $g_j(\cdot)$ function. This allows the estimated functions to have curvature from τ to the nearest z value on either side of τ. We also recognize the special importance of the region near the threshold by dividing the support of the g_0 function into two

regions, one close to τ and one further away. Analogously, we divide the support of g_1 into two regions. We then cluster some knots in the regions that are proximate to τ, subject to the requirement that there is at least one observation between each pair of knots. We also impose this requirement on knots placed in the regions farther away from τ.

The cubic spline basis we use is described in Chib and Greenberg (2010), where its value for Bayesian function smoothing is highlighted. An attractive property of this basis is that its coefficients are interpretable as function ordinates at the knots, which is useful in constructing priors. In this paper, we develop a new prior on the basis coefficients. We suppose that the function values, conditioned on the first two ordinates and precision (penalty) parameters λ_j, arise from a discrete second-order Ornstein–Uhlenbeck (O-U) process. This formulation is essentially a generalization of the second-difference penalty in Eilers and Marx (1996) to the case of unequally spaced knots. In the next step, the two initial ordinates of each g_j function are modeled by Zellner's g-prior, in the first use of this prior in this context. The final element of this prior is a flexible distribution on the penalty parameters to promote data-driven smoothness. It is worth noting that this prior requires only a small number of inputs and can be implemented in a default way.

Our analysis of the fuzzy RD design is based on a new model that utilizes the principal stratification framework of Frangakis and Rubin (2002), adapted to the RD design. In this model, the mismatch between the treatment implied by the forcing variable, and the treatment actually observed, is explained by a discrete confounder variable s that represents one of three subject types (or strata): compliers, never-takers and always-takers, that are distributed a priori as

$$\Pr(s = k) = q_k,\ k \in \{c, n, a\}\,, \tag{1.3}$$

with $q_c + q_n + q_a = 1$. For subjects of the type $s = c$, the compliers, the sharp RD model holds precisely, that is, as z passes the break-point τ, the treatment state changes from 0 to 1 with probability one:

$$\Pr(x = 0|z \leq \tau, s = c) = 1 \text{ and } \Pr(x = 1|z > \tau, s = c) = 1 \tag{1.4}$$

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,

$$\mathsf{Pr}(x = 0|z, s = n) = 1 \tag{1.5}$$

and for subjects with $s = a$, the always-takers, the probability that $x = 1$ is one regardless of the value of the forcing variable

$$\mathsf{Pr}(x = 1|z, s = a) = 1 \tag{1.6}$$

For compliers, therefore, we can assume that the sharp RD model holds. On the other hand, for never-takers and always-takers, the outcome models are assumed to satisfy the usual exclusion rule with respect to z. In detail, by subject type, our assumption is that outcomes are generated as

$$\begin{aligned} s = c &: y_j = g_j(z) + v'\gamma + h(w) + \varepsilon_j, \\ s = n &: y_{0n} = [1, v']\gamma_n + h_n(w) + \varepsilon_{0n}, \\ s = a &: y_{1a} = [1, v']\gamma_a + h_a(w) + \varepsilon_{1a}, \end{aligned} \tag{1.7}$$

where ε_{0n} is either Gaussian or Student-t with dispersion σ_n^2, and ε_{0a} is either Gaussian or Student-t with dispersion σ_a^2. We include intercepts in the n and a models because v is assumed to be free of an intercept. Under these assumptions, the fuzzy RD ATE is again the difference $g_1(\tau) - g_0(\tau)$ which we can interpret as the ATE for compliers (ATEC) at τ:

$$\begin{aligned} \text{ATEC} &= \mathsf{E}[y_1|z = \tau, w, v, s = c] - \mathsf{E}[y_0|z = \tau, w, v, s = c] \\ &= g_1(\tau) - g_0(\tau). \end{aligned} \tag{1.8}$$

We note that although the principal stratification model has been used in many different situations, this is the first use of the model to describe the fuzzy RD problem. It is interesting to see how this model straightforwardly generalizes the sharp RD model. Fitting of this model takes advantage of the usual Bayesian techniques for mixtures and involves the augmentation and sampling of the latent variables s_i $(i \leq n)$. Label switching does not occur in this model, even with a non-informative prior, because of three reasons: one, the exclusion restriction ensures that the distributional component for compliers is distinct from the other two components in the mixture; second, information from the observations with $z > \tau$ and $x = 0$ supply

a revision of the prior distribution of the parameters of the n model without contamination by compliers; and, third, information from the $z \leq \tau$ and $x = 1$ observations provide an update of the a model, again without contamination by compliers.

Posterior computations for each model, which are described below, have been coded in two R-packages and are available for download on request. We use these packages to examine the performance of our methods in both Bayesian and frequentist terms. Comparisons of results with simulated data show that the frequentist root mean squared error (RMSE) of our ATE estimates in each case are smaller than of the frequentist local ATE estimate, in several cases by a significant factor. The frequentist coverage of the Bayesian interval estimate of the ATE is often closer to the nominal value than that of the corresponding frequentist interval ATE estimate. These superior frequentist properties of the Bayes point and interval estimators may be due to the principled way in which we take advantage of the full Bayesian apparatus to address the particular challenges that arise in the RD problem.

The rest of the article is organized as follows. In Section 2, we specify the data generation process and the cubic spline formulation. Section 3 contains the prior specification, and Section 4 presents the posterior distributions and summarizes an MCMC algorithm to sample the posterior distributions. Sections 5 and 6 contains examples of simulated data for the sharp and fuzzy RD designs, respectively, and Section 7 contains our conclusions. Details of the basis functions are contained in Appendices A and B.

2 Modeling $g_j(z)$, basis expansions and likelihoods

We begin by describing our modeling of the unknown $g_0(z)$, $g_1(z)$, $h(w)$, $h_n(w)$ and $h_a(w)$ 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. The cubic splines are expressed as affine functions of basis functions. The specific basis we use is explained in Chib and Greenberg (2010) and described in Appendix A. We favor this basis, as opposed to others, for example, the B-spline basis, because of the interpretability of its basis coefficients. The modeling is

further characterized by an approach for placing the knots that has been tailored to the RD problem, and a new regularizing prior on the basis coefficients.

2.1 Data

The available data are n independent observations on (y, x, z, w, v). These are indicated by $(y_i, x_i, z_i, w_i, v_i)$, $i \leq n$, where y_i is y_{0i} when $x_i = 0$ and y_{1i} when $x_i = 1$. In the sharp RD case, denote the number of observations to the left of τ by n_0 and the number of observations to the right of τ 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 τ be assembled as

$$y_0 = (y_1, \ldots, y_{n_0}) \quad (n_0 \times 1), \qquad z_0 = (z_1, \ldots, z_{n_0}) \quad (n_0 \times 1),$$

and those to the right of τ as

$$y_1 = (y_{n_0+1}, \ldots, y_n) \quad (n_1 \times 1), \qquad z_1 = (z_{n_0+1}, \ldots, z_n) \quad (n_1 \times 1),$$

and similarly for the observations on (w, v). For later reference, define

$$z_{j,\mathsf{min}} = \mathsf{min}(z_j), \qquad z_{j,\mathsf{max}} = \mathsf{max}(z_j), \ (j = 0, 1),$$

and the pth quantile of z_j by $z_{j,p}$.

In the fuzzy RD design, these data structures are modified appropriately. In particular, since observations on either side of τ can be controls or treated, the data is arranged in sequence in four cells, defined by $I_{00} = \{i : z_i \leq \tau, x_i = 0\}$, $I_{10} = \{i : z_i > \tau, x_i = 0\}$, $I_{01} = \{i : z_i \leq \tau, x_i = 1\}$ and $I_{11} = \{i : z_i > \tau, x_i = 1\}$, as

	$x = 0$	$x = 1$
$z \leq \tau$	y_{00}, z_{00}, w_{00}	y_{01}, z_{01}, w_{01}
$z > \tau$	y_{10}, z_{10}, w_{10}	y_{11}, z_{11}, w_{11}

(2.1)

The number of observations in these cells is denoted by n_{lj} $(l, j = 0, 1)$. The outcome data in cell I_{00} is indicated by y_{00} and consists of those outcomes for which $z_i \leq \tau$ and $x_i = 0$ and the outcome data in cell I_{11} is indicated by y_{11} and consists of outcomes for which $z_i > \tau$ and $x_i = 1$. Similarly, the w data in the I_{ij} cells is indicated by w_{ij}, a vector of size $n_{ij} \times 1$.

2.2 Knots and basis matrices: sharp RD

Normally, it is enough to place knots equally-spaced through the range of the data or at particular quantiles of the data. For estimating the g_0 and g_1 functions, however, it is advantageous to place the knots more strategically. Because interest is centered on the difference of the two g functions at τ, we propose a procedure that clusters some knots in the regions around τ. Partition the intervals $[z_{0,\mathsf{min}}, \tau]$ and $[\tau, z_{1,\mathsf{max}}]$ into intervals that are proximate and far from τ. Let these four intervals be determined by the quantiles

$$z_{0,p_0} \text{ and } z_{1,p_1}$$

for specific values of p_0 and p_1, for example, $(p_0, p_1) = (0.9, 0.1)$. A particular distribution of knots is shown in Figure 1.

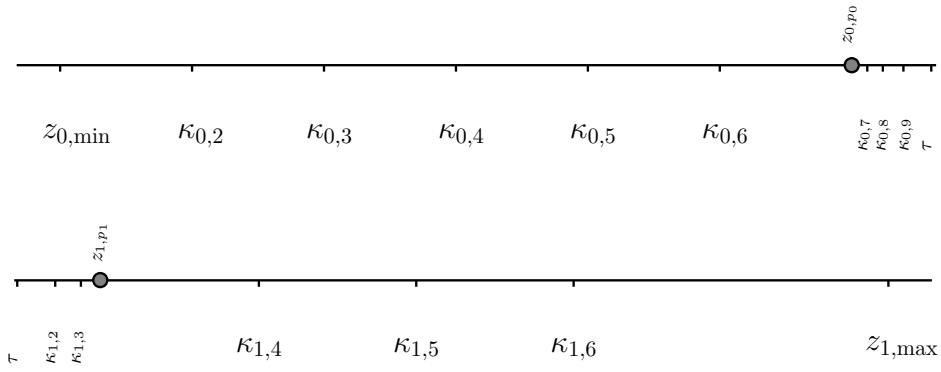

Figure 1: Example of knot locations in the basis expansions of g_0 (top panel) and g_1 (bottom panel), determined by $m_z = (6, 5)$, $m_{z,\tau} = (5, 5)$. Note that the no empty interval constraint meant that the number of knots is smaller than what is implied by these choices. The circled points are the p_0 an p_1 quantiles of z_0 and z_1, respectively. Both g_0 and g_1 have a knot at τ.

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 first place a knot at τ for each g_j function. Even though τ is not an observed value of z in general, this key idea allows the estimated g functions to have curvature from the breakpoint to the nearest z value on either side of it. Otherwise, the functions are linear over those intervals. We then place $m_{z,0,\tau}$ and $m_{z,1,\tau}$ knots in the intervals proximate to τ. Finally, we place $m_{z,0}$ and $m_{z,1}$ knots in the intervals that are further away from τ. It is convenient to define $m_z = (m_{z,0}, m_{z_1})$ and $m_{z,\tau} = (m_{z,0,\tau}, m_{z,1,\tau})$.

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 way: For the two intervals to the left of τ, place a knot at τ 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 τ. 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 $\Delta_0 = (z_{0,p_0} - z_{0,\mathsf{min}})/m_{z,0}$ and proceed from the last accepted knot in the same way as before, making sure that $z_{0,\mathsf{min}}$ is a knot at the end of this stage. The same propose-check-accept-extend approach is applied to the right of τ after placing the first knot at τ and ending with a knot at $z_{1,\mathsf{max}}$. Let

$$\{z_{0,\mathsf{min}}, \kappa_{0,2}, \ldots, \kappa_{0,m_0-1}, \tau\} \tag{2.2}$$

denote the m_0 knots to the left of τ determined by this procedure, and let

$$\{\tau, \kappa_{1,2}, \ldots, \kappa_{1,m_1-1}, z_{1,\mathsf{max}}\} \tag{2.3}$$

denote the m_1 knots to the right of τ. 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 specifying the number of knots, m_j can be selected to be cn_j^ν, for some constant c and $\nu \geq \frac{1}{5}$, following the rate derived in Claeskens et al. (2009). To choose the values of p_j, $m_{z,j}$, and $m_{z,j,\tau}$, the observed z_0 and z_1 can be examined, 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.

Given the knots and the basis functions, the unknown function ordinates g_0 and g_1 at z_0 and z_1

$$g_0(z_0) = \begin{pmatrix} g_0(z_1) \\ g_0(z_2) \\ \vdots \\ g_0(z_0) \end{pmatrix} \text{ and } \quad g_1(z_1) = \begin{pmatrix} g_1(z_{n_0+1}) \\ g_1(z_{n_0+2}) \\ \vdots \\ g_1(z_n) \end{pmatrix}, \tag{2.4}$$

respectively, can now be expressed in terms of the knots and the basis functions in the appendix as

$$g_0(z_0) = B_0\alpha$$

and

$$g_1(z_1) = B_1\beta,$$

where $B_j : n_j \times m_j$ are the basis matrices and α, and β are the basis coefficients. Since the basis coefficients, as noted above, are the function values at the knots, the components of α and β are explicitly,

$$\underset{(m_0\times 1)}{\alpha} = \begin{pmatrix} g_0(z_{0,\mathsf{min}}) \\ g_0(\kappa_{0,2}) \\ \vdots \\ g_0(\kappa_{0,m_0-1}) \\ g_0(\tau) \end{pmatrix}, \quad \underset{(m_1\times 1)}{\beta} = \begin{pmatrix} g_1(\tau) \\ g_1(\kappa_{1,2}) \\ \vdots \\ g_1(\kappa_{1,m_1-1}) \\ g_1(z_{1,\mathsf{max}}) \end{pmatrix}, \tag{2.5}$$

which implies that the ATE under our parameterization is the first component of β minus the last component of α:

$$\mathsf{ATE} = \beta_{[1]} - \alpha_{[m_0]}. \tag{2.6}$$

Knot placement for the function $h(w)$ needs little comment. One can provisionally allocate m_w equally-spaced knots on the interval $[w_{\mathsf{min}}, w_{\mathsf{max}}]$, again under the constraint that there are no empty intervals. The value m_w can be chosen to be $c_w n^\nu$, for some constant c_w and $\nu \geq \frac{1}{5}$. The function ordinates

$$h(w) = (h(w_1), h(w_2), \cdots, h(w_n))'$$

can then be expanded as

$$h(w) = B_w\delta$$

where the basis matrix B_w has one column less than the number of knots because of an identifiability condition (see Chib and Greenberg, 2010).

2.3 Knots and basis matrices: fuzzy RD

In dealing with the fuzzy RD design, the knots are selected as above, except that the data used is the one that is relevant for that function. For instance, in dealing with the g_0 function, one takes the data z_{00}, padded with τ at the right, to locate the knots, so that

$$g_0(z_{00}) = (g_0(z_1), g_0(z_2), \cdots, g_0(z_{n_{00}}))'$$

can be expressed as

$$g_0(z_{00}) = B_{00}\alpha,$$

where we use the notation B_{00} to emphasize that this basis matrix is constructed from the z data in the I_{00} cell. Similarly, from the z_{11} data, padded with τ at the left, we express the g_1 function values

$$g_1(z_{11}) = (g_1(z_{n_0+n_{01}+1}), g_1(z_{n_0+n_{01}+2}), \cdots, g_1(z_n))'$$

as

$$g_1(z_{11}) = B_{11}\beta.$$

Notice that we abuse notation in denoting the basis coefficients by the same symbols as in the sharp RDD model. This is done to emphasize the correspondence with the sharp model. The ATEC is again

$$\mathsf{ATEC} = \beta_{[1]} - \alpha_{[m_0]}. \tag{2.7}$$

For the h function in the model for compliers, the knots are computed from the data (w_{00}, w_{11}), and

$$h(w) = (h(w_1), \cdots, h(w_{n_{00}}), h(w_{n_0+n_{01}+1}), \cdots, h(w_n))'$$

has the expansion

$$h(w) = B_{00,11}\delta\ ,$$

where we employ the somewhat cumbersome (but informative) subscripting to clarify that this basis matrix is based on the w data in the I_{00} and I_{11} cells. In the same manner, the basis expansion of the h_n function makes use of the data (w_{00}, w_{10}) so that

$$h_n(w) = (h_n(w_1), \cdots, h_n(w_{n_{00}}), h_n(w_{n_{00}+1}), \cdots, h_n(w_{n_0}))'$$

is expressed as

$$h_n(w) = B_{00,10}\delta_n.$$

Finally, for the h_a function in the fuzzy model, we base the knots on (w_{01}, w_{11}) and express

$$h_a(w) = \begin{pmatrix} h_a(w_{n_0+1}) & h_a(w_{n_0+2}) & \cdots & h_a(w_n) \end{pmatrix}'$$

as

$$h_a(w) = B_{00,10}\delta_\alpha.$$

2.4 Likelihood: sharp RD

With the basis expansions in hand, we can express the sharp RD model $y_j = g_j(z) + v'\gamma + h(w) + \varepsilon_j$ in a form convenient for computing by collecting all n observations as:

$$\begin{pmatrix} y_{0\,(n_0\times 1)} \\ y_{1\,(n_1\times 1)} \end{pmatrix} = \begin{pmatrix} B_0 & 0 & B_{w,1:n_0} & V_{1:n_0} \\ 0 & B_1 & B_{w,n_{0+1}:n} & V_{n_{0+1}:n} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{pmatrix} + \begin{pmatrix} \varepsilon_{0\,(n_0\times 1)} \\ \varepsilon_{1\,(n_1\times 1)} \end{pmatrix}, \tag{2.8}$$

where the notation $B_{w,1:n_0}$ is shorthand for the first n_0 rows of B_w, and the notation $B_{w_{0+1}:n}$ for its last n_1 rows, and V is the matrix of observations on v, partitioned conformably. In abbreviated form, this model can be written as

$$y = X\theta + \varepsilon,$$

where $\varepsilon = (\varepsilon_0, \varepsilon_1)$ are the independently distributed errors, and

$$\theta = (\alpha, \beta, \gamma, \delta)$$

is the regression parameter of length $k = (m_0 + m_1 + m_w - 1 + q)$. Therefore, the Gaussian likelihood is

$$y|\theta, \sigma^2 \sim \mathcal{N}(X\theta, \sigma^2 I_n), \tag{2.9}$$

where I_n is the identity matrix of order n, and the Student-t likelihood is

$$y|\theta, \sigma^2, \nu \sim t_\nu(X\theta, \sigma^2 I_n) \tag{2.10}$$

for given degrees of freedom $\nu \geq 2$.

2.5 Likelihood: fuzzy RD

The likelihood function in the fuzzy RD model is that of a mixture model. Consider for simplicity the Gaussian case. Let $B_{00,i}$ denote the ith row of B_{00}, with similar notation for the other basis matrices. Then, the likelihood contribution of the ith observation by cell are

$$\begin{aligned}
L_{00,i} &= q_c N(y_i|B_{00,i}\alpha + v_i\gamma + B_{00,11,i}\delta, \sigma^2) + q_n N(y_i|[1, v_i]\gamma_n + B_{00,10,i}\delta_n, \sigma_n^2),\\
L_{10,i} &= q_n N(y_i|[1, v_i]\gamma_n + B_{00,10,i}\delta_n, \sigma_n^2),\\
L_{01,i} &= q_a N(y_i|[1, v_i]\gamma_a + B_{01,11,i}\delta_a, \sigma_a^2),\\
L_{11,i} &= q_c N(y_i|B_{11,i}\beta + v_i\gamma + B_{00,11,i}\delta, \sigma^2) + q_a N(y_i|[1, v_i]\gamma_a + B_{01,11,i}\delta_a, \sigma_a^2), \qquad (2.11)
\end{aligned}$$

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

$$L = \prod_{i\in I_{00}} L_{00,i} \times \prod_{i\in I_{10}} L_{10,i} \times \prod_{i\in I_{01}} L_{01,i} \times \prod_{i\in I_{11}} L_{11,i}. \qquad (2.12)$$

3 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, which, because the function is linear in the basis parameters, is a quadratic form in the basis parameters.

Our approach to specifying this penalty is through a novel prior distribution. Consider the sharp RD model for specificity. The parameter of interest in the mean function is $\theta = (\alpha, \beta, \gamma, \delta)$. The prior we construct is conditioned on four positive penalty parameters, namely $\lambda = (\lambda_0, \lambda_1, \lambda_w, \lambda_a)$, and the error variance σ^2. It has the form

$$\pi(\theta|\lambda_0, \lambda_1, \lambda_\gamma, \lambda_\delta, \sigma^2) = \pi_0\left(\alpha|\lambda_0, \sigma^2\right)\pi_1\left(\beta|\lambda_1, \sigma^2\right)\pi_\gamma\left(\gamma|\lambda_\gamma, \sigma^2\right)\pi_\delta\left(\delta|\lambda_\delta, \sigma^2\right)$$

where each of the distributions on the right-hand side is Gaussian, and except for the distribution π_γ, is of a form that has not been used before. The other three distributions, π_0, π_1 and

π_δ, which are the distributions of the basis parameters in the expansions $B_0\alpha$, $B_1\beta$ and $B_w\delta$, respectively, are constructed by supposing that the basis parameters (the function ordinates at the knots) are each realizations of mutually independent discrete, second-order Ornstein–Uhlenbeck (O-U) processes. Our approach naturally handles unequally-spaced knots and produces a proper distribution, unlike the prior in Eilers and Marx (1996).

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. We propose to use this process in its Euler discretized form as a prior for the ordinates of the non-parametric functions at their respective knots. One simple possibility is to let $a = 1$, $\mu = 0$ and $s = \sigma/\sqrt{\lambda}$, where λ is equal to λ_j for the g_j functions or λ_δ for the h function. Furthermore, in discrete time, dt is the spacing between successive knots. This discrete second order O-U process is the basis of the prior on the unknown function ordinates.

3.1 Prior of α

Consider the situation shown in Figure 2 for values of g_0 computed at three successive knots,

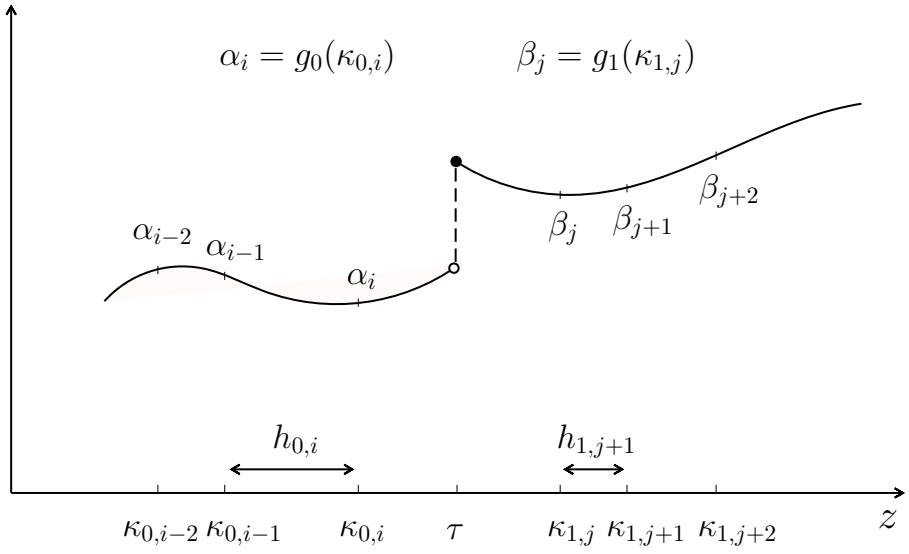

Figure 2: Function ordinates and prior formulation.

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

$$\Delta^2\alpha_i = (\alpha_i - \alpha_{i-1}) - (\alpha_{i-1} - \alpha_{i-2}) \ , \ i > 2$$

and define the spacings between knots by

$$h_{0,i} = \kappa_{0,i} - \kappa_{0,i-1}$$

as shown in Figure 2. Then, in the Gaussian error model, our prior assumption on $(\alpha_3, \alpha_4, \ldots, \alpha_{m_0})$ conditioned on (α_1, α_2) is that

$$\Delta^2\alpha_i = -(\alpha_{i-1} - \alpha_{i-2})h_{0,i} + u_{0i}, \tag{3.1}$$

$$u_{0i}|\sigma^2, \lambda_0 \sim \mathcal{N}\left(0, \frac{\sigma^2}{\lambda_0}h_{0,i}\right), \tag{3.2}$$

where $(\alpha_{i-1} - \alpha_{i-2})h_{0,i}$ introduces mean reversion and λ_0 is an unknown penalty parameter. Note that the assumed dependence between the spline coefficients and σ^2 is uncommon, but requiring conditional conjugacy with respect to σ^2 is also quite reasonable. In the Student-t error model, on the other hand, conditional prior conjugacy with respect to σ^2 cannot be achieved, and it is more reasonable to suppose that the spline coefficients and σ^2 are a priori independent.

Next consider the starting ordinates, (α_1, α_2). Instead of an improper prior as in Lang and Brezger (2004) and Brezger and Lang (2006), we specify a proper g-type prior that smoothly avoids any elicitation difficulties. We let

$$T^{-1}_{\alpha,1:2} = (B_0'B_0)_{1:2}$$

denote the first two rows and columns of $B_0'B_0$ and then suppose that

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} g_0(z_{0,\min}) \\ g_0(\kappa_{0,2}) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \alpha_{1,0} \\ \alpha_{2,0} \end{pmatrix}, \frac{\sigma^2}{\lambda_0}T_{\alpha,1:2}\right)$$

where $\alpha_{1,0}$ and $\alpha_{2,0}$ are 2 free hyperparameters that we roughly set to equal the prior means of g_0, if such information is available, otherwise, we set them to zero. In our experience, posterior inferences are quite robust to the choice of these values.

By straightforward calculations it can be shown that these assumptions imply that

$$\pi_0(\alpha|\sigma^2, \lambda_0) = \mathcal{N}\left(\alpha|D_\alpha^{-1}\alpha_0, \frac{\sigma^2}{\lambda_0}D_\alpha^{-1}T_\alpha D_\alpha^{-1'}\right) \tag{3.3}$$

where

$$\alpha_0 = (\alpha_{1,0}, \alpha_{2,0}, 0, \ldots, 0)' : m_0 \times 1,$$

D_α is a tri-diagonal matrix (given in Appendix B) that depend entirely on the spacings, and

$$T_\alpha = \mathsf{blockdiag}(T_{\alpha,1:2}, I_{m_0-2}) : m_0 \times 1$$

Thus, the penalty matrix of the g_0 function is $\lambda_0 D_\alpha T_\alpha^{-1} D_\alpha'$.

3.2 Prior of β

Our prior $\pi_1(\beta|\lambda_1, \sigma^2)$ on β is constructed in a way analogous to $\pi_0(\alpha|\sigma^2, \lambda_0)$ but with the key difference that the O-U process is oriented for knots going from right to left. Consider again Figure 2 and now consider the three successive values of g_1, ordered from right to left, and represented by $\beta_j = g_1(\kappa_{1,j})$, $\beta_{j+1} = g_1(\kappa_{1,j+1})$ and $\beta_{j+2} = g_1(\kappa_{1,j+2})$. Our proposal is to imagine the specific second differences

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

with spacings between knots given by

$$h_{1,j+1} = \kappa_{1,j+1} - \kappa_{1,j},$$

conditioned on the right end-points $(\beta_{m_1-1}, \beta_{m_1})$, as following the process

$$\Delta^2\beta_j = -(\beta_{j+1} - \beta_{j+2})h_{1,j+1} + u_{ji}, \tag{3.4}$$

$$u_{ji}|\sigma^2, \lambda_1 \sim \mathcal{N}\left(0, \frac{\sigma^2}{\lambda_1} h_{1,j+1}\right), \tag{3.5}$$

where λ_1 is another unknown penalty parameter.

We orient the β process in this way to circumvent the direct specification of a distribution on $\beta_1 = g_1(\tau)$, which can be both difficult and consequential. In our approach, the prior on $g_1(\tau)$ is determined by the O-U process, in the same way that the prior on the other key parameter $\alpha_{m_0} = g_0(\tau)$ is determined by the α O-U process. Our experiments have indicated that this formulation helps to reduce the extent of the shrinkage-bias for α_{m_0} and β_1.

The distribution of the initial values $(\beta_{m_1-1}, \beta_{m_1})$ is once again a type of g-prior. Let

$$T^{-1}_{\beta,m_1-1:m_1} = (B_1' B_1)_{m_1-1:m_1}$$

denote the last two rows and columns of $B_1' B_1$, then our assumption is that

$$\begin{pmatrix} \beta_{m_1-1} \\ \beta_{m_1} \end{pmatrix} = \begin{pmatrix} g_1(z_{1,m_1-1}) \\ g_0(\kappa_{1,m_1}) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \beta_{m_1-1,0} \\ \beta_{m_1,0} \end{pmatrix}, \frac{\sigma^2}{\lambda_1} T_{\beta,m_1-1:m_1} \right).$$

These two assumptions immediately imply that

$$\pi_1(\beta|\sigma^2, \lambda_1) = \mathcal{N}\left(D_\beta^{-1}\beta_0, \frac{\sigma^2}{\lambda_1} D_\beta^{-1} T_\beta D_\beta^{-1'} \right), \tag{3.6}$$

where

$$\beta_0 = (0, \ldots 0, \beta_{m_1-1,0}, \beta_{m_1,0})' : m_1 \times 1,$$

D_β is the tri-diagonal matrix in Appendix B, and

$$T_\beta = \mathsf{blockdiag}(I_{m_1-2}, T_{\beta,m_1-1:m_1}) : m_1 \times 1.$$

Thus, the penalty matrix for the g_1 function is $\lambda_1 D_\beta T_\beta^{-1} D_\beta'$.

3.3 Prior of γ and δ

For the linear parameters γ our proposed prior is a semi-conjugate g-prior with a prior mean of γ_0 and precision matrix equal to $\sigma^{-2}\lambda_v V'V$. Specifically,

$$\pi_\gamma(\gamma|\lambda_\gamma, \sigma^2) = \mathcal{N}\left(\gamma|\gamma_0, \frac{\sigma^2}{\lambda_v}(V'V)^{-1} \right)$$

where we generally set γ_0 to equal zero.

Finally, for the basis parameters δ, the prior construction proceeds entirely analogously to that of α. Omitting details,

$$\pi_\delta\left(\delta|\lambda_\delta, \sigma^2\right) = \mathcal{N}\left(\delta|D_w^{-1}\delta_0, \frac{\sigma^2}{\lambda_\delta} D_w^{-1} T_w D_w^{-1'} \right)$$

where the matrices D_w and T_w are those given in the appendix.

We remark that in the fuzzy model, the preceding prior is the prior on the parameters of the complier model, where the penalty matrices are computed from the data in the cells I_{00} and I_{11}. The prior on the parameters of the n and a models is constructed analogously to that of (γ, δ) in the complier model. Because the details are clear, they are omitted.

3.4 Prior of σ^2 and λ

The prior on σ^2 is of the usual form. Independent of the precision parameters, we suppose that

$$\sigma^2 \sim \mathsf{IG}\left(\frac{\nu_0}{2}, \frac{\delta_0}{2}\right),$$

where ν_0 and δ_0 are chosen to reflect the researcher's views about the mean and standard deviation of σ^2. We have a similar prior on σ_n^2 and σ_a^2 in the fuzzy model. In the sharp model, this leaves us with the parameters $\lambda = (\lambda_0, \lambda_1, \lambda_\gamma, \lambda_\delta)$ and with the parameters $\lambda = (\lambda_0, \lambda_1, \lambda_\gamma, \lambda_\delta, \lambda_{\gamma_n}, \lambda_{\delta_n}, \lambda_{\gamma_a}, \lambda_{\delta_a})$ in the fuzzy model. Specifying a general prior assumption about these parameters is not easy. In the frequentist interpretation of the penalized smoothing spline, for fixed n, $\lambda_j \to 0$ implies an unpenalized regression spline, and $\lambda_j \to \infty$ implies piece-wise linearity. Also, in that interpretation, the size of λ increases with n. For our setting, Claeskens et al. (2009) derive the optimal rate $\lambda = O(n^{1/5})$.

In our Bayesian formulation, λ helps determine the prior variances of θ, which may be regarded as a measure of the strength of our belief in the specification of the prior mean. The prior variances of the elements of α, for example, are equal to the diagonal elements of $(\sigma^2/\lambda_0)D_\alpha^{-1}T_\alpha D_\alpha^{-1'}$. These variances depend on σ^2, λ_0 and the $h_{0,k}$, which vary from problem to problem. We consider two methods of choosing the hyperparameters of the distribution of λ_0. One is simply to specify prior values of $\mathsf{E}(\lambda_0)$ and $\mathsf{sd}(\lambda_0)$ and match a Gamma distribution to these choices. The second idea is to choose $\mathsf{E}(\lambda_0)$ to make the smallest diagonal element of the variance matrix equal to one, that is, choose $\mathrm{E}(\lambda_0)$ so that

$$\min\left\{\mathsf{diag}\left(\frac{\mathsf{E}(\sigma^2)}{\mathsf{E}(\lambda_0)}D_\alpha^{-1}T_\alpha D_\alpha^{-1'}\right)\right\} = 1,$$

and let $\mathsf{sd}(\lambda_0)$ be a multiple of the prior mean. For specificity, in the sharp model, the same approach can be taken for λ_1, λ_v and λ_w. Given the prior mean and standard deviation, we can find the matching Gamma distributions, which are denoted as

$$\lambda_j \sim \mathsf{Gamma}\left(\frac{a_{j0}}{2}, \frac{b_{j0}}{2}\right), \quad j = 0, 1, \gamma, \delta$$

and assumed to be distributed independently. An identical approach can be employed in the fuzzy model. Note that we obtain a result comparable to an unpenalized regression spline model by letting the prior mean of λ_j $(j = 0, 1, \gamma, \delta)$ be small and the prior standard deviation

be even smaller. In the Bayesian interpretation, these settings have the effect of enforcing a small precision on the prior distribution of $(\alpha, \beta, \gamma, \delta)$.

3.5 Prior of q

In the fuzzy RD model, a prior distribution on the 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 Posterior Distributions and MCMC Sampling

We begin with the sharp RD design and then show how to modify it for the fuzzy case.

4.1 Sharp RD design

4.1.1 Gaussian error

Consider first the Gaussian error model which has the form

$$\begin{aligned} y|\theta, \sigma^2 &\sim \mathcal{N}_n(X\theta, \sigma^2 I_n), \\ \theta|\sigma^2, \{\lambda_j\} &\sim \mathcal{N}_k\left(\theta_0, \sigma^2 A_0\right), \\ \sigma^2 &\sim \mathsf{IG}\left(\frac{\nu_{00}}{2}, \frac{\delta_{00}}{2}\right), \\ \lambda_j &\sim \mathsf{Gamma}\left(\frac{a_{j0}}{2}, \frac{b_{j0}}{2}\right), \ (j = 0, 1, \gamma, \delta) \end{aligned}$$

where

$$\theta_0 = \left(D_\alpha^{-1}\alpha_0, D_\beta^{-1}\beta_0, \gamma_0, D_w^{-1}\delta_0\right)' ,$$

and

$$A_0 = \mathsf{blockdiag}\left(\frac{1}{\lambda_0}D_\alpha^{-1}T_\alpha D_\alpha^{-1'}, \frac{1}{\lambda_1}D_\beta^{-1}T_\beta D_\beta^{-1'}, \frac{1}{\lambda_\gamma}(V'V)^{-1}, \frac{1}{\lambda_\delta}D_w^{-1}T_w D_w^{-1'}\right)$$

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 $(y, \{\lambda_j\})$, sample (θ, σ^2) from the updated normal-inverse gamma distribution:

$$\begin{aligned}\pi(\theta, \sigma^2|y, \{\lambda_j\}) &= \pi(\theta|\sigma^2, y, \{\lambda_j\})\ \pi(\sigma^2|y, \{\lambda_j\}) \\ &\sim \mathcal{N}_k(\theta_1, \sigma^2 A_1)\ \mathsf{IG}(\nu_1/2, \delta_1/2),\end{aligned}$$

  where

$$\begin{aligned}\theta_1 &= A_1(X'y + A_0^{-1}\theta_0), \\ A_1 &= (X'X + A_0^{-1})^{-1}, \\ \nu_1 &= \nu_{00} + n, \\ \delta_1 &= \delta_{00} + y'y + \theta_0' A_0^{-1}\theta_0 - \theta_1' A_1^{-1}\theta_1\end{aligned}$$

- Given (α, σ^2), sample λ_0 from

  an updated Gamma distribution, independent of y:

$$\pi(\lambda_0|\alpha, \sigma^2) \sim \mathsf{Gamma}(a_{01}/2, b_{01}/2),$$

  where

$$\begin{aligned}a_{01} &= a_{00} + m_0, \\ b_{01} &= b_{00} + \frac{(D_\alpha \alpha - \alpha_0)' T_\alpha^{-1} (D_\alpha \alpha - \alpha_0)}{\sigma^2}.\end{aligned}$$

- Given (β, σ^2), sample λ_1 from an updated Gamma distribution, independent of y:

$$\pi(\lambda_1|\beta, \sigma^2) \sim \mathsf{Gamma}(a_{11}/2, b_{11}/2),$$

  where

$$\begin{aligned}a_{11} &= a_{10} + m_1, \\ b_{11} &= b_{10} + \frac{(D_\beta \beta - \beta_0)' T_\beta^{-1} (D_\beta \beta - \beta_0)}{\sigma^2}.\end{aligned}$$

- Given (γ, σ^2), sample λ_γ from an updated Gamma distribution, independent of y:

$$\pi(\lambda_\gamma|\gamma, \sigma^2) \sim \mathsf{Gamma}(a_{\gamma 1}/2, b_{\gamma 1}/2),$$

where

$$a_{\gamma 1} = a_{\gamma 0} + q,$$
$$b_{\gamma 1} = b_{\gamma 0} + \frac{(\gamma - \gamma_0)'(V'V)^{-1}(\gamma - \gamma_0)}{\sigma^2}.$$

- Given (δ, σ^2), sample λ_δ from an updated Gamma distribution, independent of y:

$$\pi(\lambda_\delta|\gamma, \sigma^2) \sim \mathsf{Gamma}(a_{\delta 1}/2, b_{\delta 1}/2),$$

where

$$a_{\delta 1} = a_{\delta 0} + m_w,$$
$$b_{\delta 1} = b_{\delta 0} + \frac{(D_w\delta - \delta_0)'T_w^{-1}(D_w\delta - \delta_0)}{\sigma^2}.$$

- After the burn-in iterations, extract the last element of α and the first element of β to obtain drawings of the ATE from its posterior distribution.

4.1.2 Student-t error

The Student-t error model differs from the Gaussian model for only two distributions

$$y|\theta, \sigma^2 \sim t_\nu(X\theta, \sigma^2 I_n)$$
$$\theta|\{\lambda_j\} \sim \mathcal{N}_k\left(\theta_0, A_0\right),$$

since the prior distributions of σ^2 and $\{\lambda_j\}$ are the same. The posterior distribution of the parameters in this model are easily sampled by using the well known representation of the Student-t distribution as a gamma scale mixture of normals:

$$\varepsilon_i|\sigma^2, \xi_i \sim \mathcal{N}(0, \sigma^2/\xi_i),$$
$$\xi_i \sim \mathsf{Gamma}(\frac{\nu}{2}, \frac{\nu}{2}), i \leq n.$$

Then, by augmenting the posterior distribution of θ by the $\{\xi_i\}$, MCMC sampling proceeds in much the same way as above, except that θ is sampled conditioned on both σ^2 and $\{\xi_i\}$, σ^2 is sampled conditioned on θ and $\{\xi_i\}$, and a layer is added in which the $\{\xi_i\}$ are sampled conditioned on θ and σ^2.

4.2 Fuzzy RD design

Estimation of the fuzzy RD model relies on the usual augmentation of the mixture indicators, here the type variables s_i $(i \leq n)$. Conditioned on the parameters, these type variables have to be sampled only in the cells I_{00} and I_{11} (because the subjects in cells I_{10} and I_{10} are necessarily of types n and a, respectively). From the likelihood contributions given above, for observations in cell I_{00}

$$\begin{aligned}
\mathsf{Pr}(s_i = c|y_i, \theta) &= \frac{q_c N(y_i|B_{00,i}\alpha + v_i\gamma + B_{00,11,i}\delta, \sigma^2)}{q_c N(y_i|B_{00,i}\alpha + v_i\gamma + B_{00,11,i}\delta, \sigma^2) + q_n N(y_i|[1, v_i]\gamma_n + B_{00,10,i}\delta_n, \sigma_n^2)} \\
\mathsf{Pr}(s_i = n|y_i, \theta) &= 1 - \mathsf{Pr}(s_i = c|y_i, \theta)
\end{aligned}$$

and for observations in cell I_{11}

$$\begin{aligned}
\mathsf{Pr}(s_i = c|y_i, \theta) &= \frac{q_c N(y_i|B_{11,i}\beta + v_i\gamma + B_{00,11,i}\delta, \sigma^2)}{q_c N(y_i|B_{11,i}\beta + v_i\gamma + B_{00,11,i}\delta, \sigma^2) + q_a N(y_i|[1, v_i]\gamma_a + B_{01,11,i}\delta_a, \sigma_a^2)} \\
\mathsf{Pr}(s_i = a|y_i, \theta) &= 1 - \mathsf{Pr}(s_i = c|y_i, \theta)
\end{aligned}$$

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

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

Again, conditioned on the sampled types, the model decomposes into three separate models, one for each type. Then, we can write

$$\begin{pmatrix} y_{00}^c \\ {\scriptstyle (n_{00}^c \times 1)} \\ y_{11}^c \\ {\scriptstyle (n_{11}^c \times 1)} \end{pmatrix} = \begin{pmatrix} B_{00}^c & 0 & B_{00,11}^c & V_{00}^c \\ {\scriptstyle (n_{00}^c \times m_0)} & & {\scriptstyle (n_{00}^c \times m_w - 1)} & \\ 0 & B_{11}^c & B_{00,11}^c & V_{11}^c \\ & {\scriptstyle (n_{11}^c \times m_1)} & {\scriptstyle (n_{11}^c \times m_w - 1)} & \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{pmatrix} + \begin{pmatrix} \varepsilon_{00}^c \\ {\scriptstyle (n_{00}^c \times 1)} \\ \varepsilon_{11}^c \\ {\scriptstyle (n_{11}^c \times 1)} \end{pmatrix}, \tag{4.1}$$

where the c superscript indicates the sub-vectors and sub-matrices consisting of the rows (observations) sampled as compliers in the indicated cells. This is analogous to (2.8) in the sharp RD model. Therefore, the parameters $(\alpha, \beta, \gamma, \delta)$, $(\lambda_0, \lambda_1, \lambda_\gamma, \lambda_\delta)$ and σ^2 can be sampled according to one step of the sharp RD MCMC algorithm.

Similarly, given the n^n_{00} observations sampled as never-takers in the cell I_{00}, we can write

$$\begin{pmatrix} y^n_{00} \\ (n^n_{00}\times 1) \\ y_{10} \\ (n_{10}\times 1) \end{pmatrix} = \begin{pmatrix} V^n_{00} & B^n_{00,10} \\ & (n^n_{00}\times m^n_w - 1) \\ V_{10} & B_{00,11} \\ & (n_{10}\times m_w - 1) \end{pmatrix} \begin{pmatrix} \gamma_n \\ \delta_n \end{pmatrix} + \begin{pmatrix} \varepsilon^n_{00} \\ (n^n_{00}\times 1) \\ \varepsilon^n_{11} \\ (n_{10}\times 1) \end{pmatrix}, \tag{4.2}$$

where V_{00} and V_{10} contain an intercept, and $B^n_{00,10}$ consists of the rows of $B_{00,11}$ in cell I_{00} that are classified as never-takers. This model again is similar in structure to the sharp RD model and, therefore, its parameters (γ_n, δ_n), $(\lambda_{\gamma_n}, \lambda_{\delta_n})$ and σ^2_n can be sampled using one step of the sharp RD MCMC algorithm.

Next, given n^a_{11} observations classified as always-takers in the cell I_{11}, we have

$$\begin{pmatrix} y_{01} \\ (n_{01}\times 1) \\ y^a_{11} \\ (n^a_{11}\times 1) \end{pmatrix} = \begin{pmatrix} V_{01} & B_{01,11} \\ & (n_{01}\times m^a_w - 1) \\ V^a_{11} & B^a_{01,11} \\ & (n^a_{11}\times m^a_w - 1) \end{pmatrix} \begin{pmatrix} \gamma_\alpha \\ \delta_a \end{pmatrix} + \begin{pmatrix} \varepsilon^a_{01} \\ (n_{01}\times 1) \\ \varepsilon^a_{11} \\ (n^a_{11}\times 1) \end{pmatrix} \tag{4.3}$$

where V_{01} and V^a_{11} contain an intercept, and $B^a_{01,11}$ consists of the rows of $B_{01,11}$ in cell I_{11} that are classified as always-takers. Its parameters (γ_a, δ_a), $(\lambda_{\gamma_a}, \lambda_{\delta_a})$ and σ^2_a can likewise be sampled as above.

These steps, which constitute one iteration of the MCMC sampling in the fuzzy RD model, are repeated, and sampled values beyond the burn-in are retained for analysis, just as in the sharp MCMC algorithm above.

4.3 Marginal likelihood computation

For both these models, the procedure of Chib (1995) can be used to calculate the marginal likelihood. We use the marginal likelihood to compare models that differ in the value of p and in the number of knots in the four regions implied by a given p. We also use marginal likelihoods to compare the Gaussian and Student-t assumptions, and the Student-t model with different degrees of freedom.

5 Simulated Data: Sharp RD Design

5.1 Gaussian errors

This section examines the performance of our method in both Bayesian and frequentist terms using simulated data with a Gaussian error. The first goal is to investigate the posterior esti-

mates of the ATE as a function of sample size. The second goal is to examine the sampling properties of the posterior ATE estimates, again as a function of sample size. The performance is benchmarked against those of the frequentist estimator of Imbens and Kalyanaraman (2012) as implemented in the R package *rdd*. As will be seen, these side-by-comparisons show that the frequentist RMSE of the Bayes ATE estimate is smaller than of the frequentist ATE estimate, in several cases by a significant factor. The coverage of the Bayes ATE estimate is also better than that of the frequentist ATE estimate. All results are calculated and easily reproduced from an R package that is available from us on request.

5.1.1 Data generating process

The data generating process (DGP) is similar to that of Rau (2011), but the g_0 and g_1 used here have derivatives up to second order throughout their support. We also include control variables W and V in the DGP. In detail,

$$\begin{aligned} g_0(z) &= z + z^2 + z^3 + \sin(30z), \\ g_1(z) &= z + z^2 + z^3 + 5\sin(10z) + 1, \\ h(W) &= \frac{\sin(\pi W/2)}{1 + W^2(\mathrm{sign}(W) + 1)}, \\ V &\sim U(0,1),\ \delta = 1, \\ \varepsilon &\sim \mathcal{N}(0,1), \\ \tau &= 0. \end{aligned}$$

The true value of the ATE is one. A particular data set of $n = 500$ observations drawn from this design is shown in Figure 3. It can be seen from this figure that the range of the outcome data is quite large relative to the size of the ATE and that the g_0 and g_1 functions have quite different cycles and amplitudes. The outcome data is also noisy.

In the analysis that follows, this design is called Design 1. It is used to generate 10,000 data sets for each of the sample sizes $n = 500, 1000, 2000$ and 4000.

5.1.2 Knots and prior distribution

We define the two regions proximate to $\tau = 0$ through the quantiles $p = (.8, .2)$. In estimating the model when $n = 500$, the number of knots is selected from a small marginal likelihood

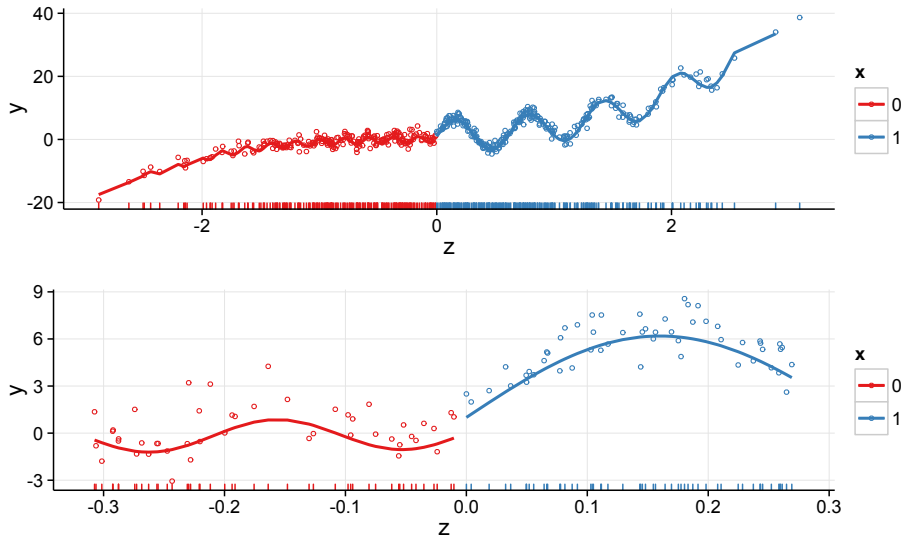

Figure 3: Simulated data - Design 1, $n = 500$: True functions and sample data. Bottom panel is a zoom plot of the plot in the top panel.

comparison of alternative models. This yields the values $m_z = (30, 20)$, $m_{z,\tau} = (5, 5)$ and $m_w = 5$. For the remaining sample sizes, the number of knots is increased moderately from these values: when $n = 1000$, $m_z = (32, 22)$, $m_{z,\tau} = (8, 8)$, and $m_w = 10$; when $n = 2000$, $m_z = (35, 30)$, $m_{z,\tau} = (9, 9)$, and $m_w = 15$; and when $n = 4000$, $m_z = (40, 35)$, $m_{z,\tau} = (10, 10)$, and $m_w = 18$.

The prior means of the first two ordinates of g_0, the last two of g_1 and the first two of h are assumed to be

$$\alpha_{1,0} = \alpha_{2,0} = -15, \quad \beta_{m_1-1,0} = \beta_{m_1,0} = 30, \quad \gamma_{1,0} = \gamma_{2,0} = 0,$$

respectively. The prior mean of δ is assumed to be 0. The hyperparameters ν_0 and δ_0 are chosen so that $\mathsf{E}(\sigma^2) = 2$ and $\mathsf{sd}(\sigma^2) = 20$. Finally, the hyperparameters a_{j0} and b_{j0} for each of the four λ's are chosen to make $\mathsf{E}(\lambda_j) = 1$ and $\mathsf{sd}(\lambda_j) = 10$. This same prior distribution is used for all the data sets generated in this section.

5.1.3 Conditional analysis

Consider first a conditional analysis of this model for varying sample sizes. All results are based on a burn-in of 1,000 iterations and a retained sample of 10,000 iterations.

Figure 4 contains the posterior mean of the g_j functions along with the 95% point-wise credibility bands. The true values of the functions are indicated by squares, and the outcome observations are indicated by open circles. For better clarity and to see how the functions are estimated near τ, a zoom plot is provided in the bottom panel of Figure 4. These results show that the method closely tracks the functions even with a relatively small sample size.

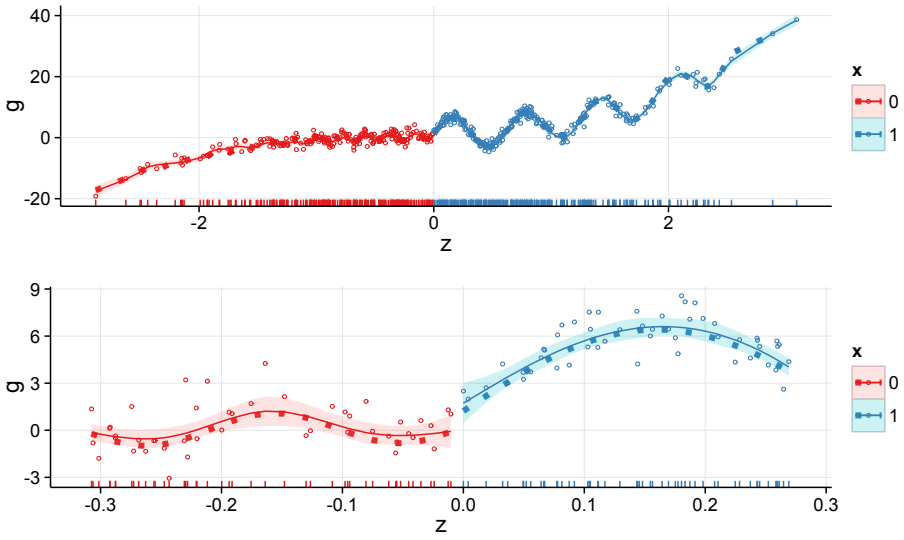

Figure 4: Simulated data - design 1, $n = 500$: Posterior mean of the g_j functions along with the 95% point-wise credibility bands. True function values indicated by squares. The outcome data is indicated by open circles. Zoom plot in the bottom panel.

The posterior results for the ATE by sample size are given in Table 1 and Figure 5. As

	Bayes		Frequentist	
n	Mean	sd	Estimate	se
500	1.69	0.90	3.03	0.70
1000	0.75	0.61	2.12	0.65
2000	1.67	0.50	2.48	0.57
4000	1.34	0.46	1.38	0.39

Table 1: Simulated data - Design 1: Summary of the ATE estimates, conditional on a given sample, by sample size. The true value of the ATE is one.

expected, the posterior standard deviation of the ATE declines with the sample size. A comparison with the estimates from the Imbens and Kalyanaraman (2012) method shows that, at

least for this data set, the Bayes posterior mean is closer to the true value than the frequentist estimate for each sample size.

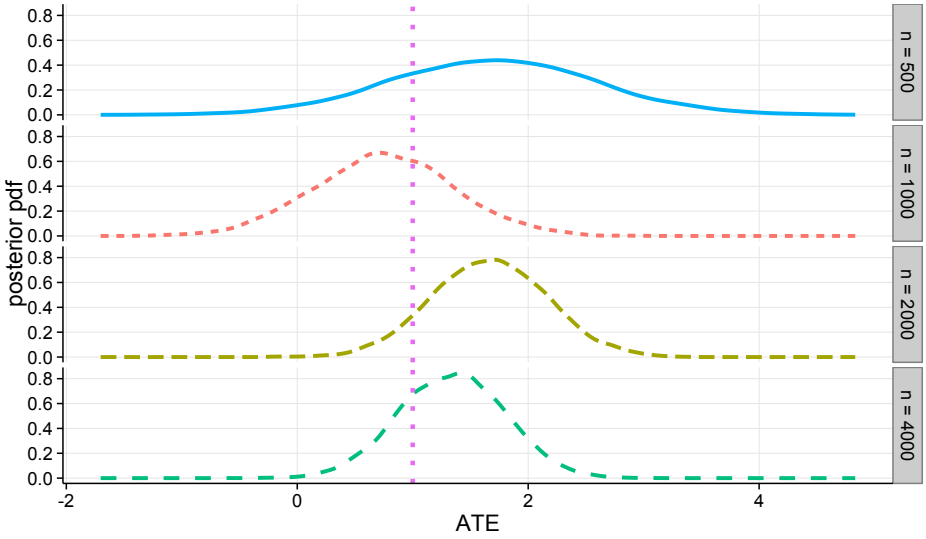

Figure 5: Simulated data - Design 1: Posterior pdf of the ATE conditional on a given sample, by sample size. True value of one is indicated by a vertical line.

5.1.4 Sampling performance

Further information about the Bayesian method can be generated by evaluating the frequentist properties of the Bayesian point and interval estimators. By repeating the estimation 10,000 times for each sample size, we obtain Monte Carlo estimates of the frequentist bias, the RMSE, and the coverage behavior of the Bayesian credibility interval. These properties can be contrasted with those of the corresponding frequentist estimators. The sampling results reported in Table 2 are striking. They show that the sampling behavior of the Bayesian point and interval estimates is superior to that of the frequentist estimators. The sampling distribution of the posterior mean shows faster convergence to the true value and substantially smaller RMSEs, especially for the first three sample sizes. The coverage properties of the Bayesian interval estimator is also close to that of the nominal value for every sample size.

	Bayes			Frequentist		
	ATE	coverage	RMSE	ATE	coverage	RMSE
$n = 500$						
mean	1.4847	0.9328	0.9499	2.791	0.427	2.1211
q.025	0.9438	1.0000		2.053	0.0000	
q.975	2.0104	1.0000		3.413	1.0000	
$n = 1000$						
mean	1.1371	0.9675	0.7279	2.3589	0.4122	1.6006
q.025	0.6754	1.0000		1.8223	0.0000	
q.975	1.6131	1.0000		2.8303	1.0000	
$n = 2000$						
mean	1.0599	0.9618	0.5560	1.7589	0.6043	0.9874
q.025	0.6939	1.0000		1.3477	0.0000	
q.975	1.4243	1.0000		2.1700	1.0000	
$n = 4000$						
mean	1.0215	0.9534	0.4305	1.2642	0.8627	0.5307
q.025	0.7318	1.0000		0.9602	1.0000	
q.975	1.3145	1.0000		1.5764	1.0000	

Table 2: Simulated data - Design 1: Summary of the ATE sampling distributions from 10,000 repeated samples. The true value of the ATE is one.

5.2 Student-t error

In this section, we consider the performance of the Bayes method using a second design, this one based on Imbens and Kalyanaraman (2012), that has features that are not present in the preceding example. In particular, under this design, the data around τ is sparse, and only about 10% of the observations are treated. We introduce another dimension of complexity into the model by assuming that the errors are distributed as Student-t with 2 degrees of freedom. This allows us to examine the impact, if any, of a thick-tailed error distribution on inferences about the ATE. There are no controls in the design other than z.

5.2.1 Data generating process

The design, which we call Design 2, is defined by

$$
\begin{aligned}
g_0(z) &= 0.48 + 1.27z + 7.18z^2 + 20.21z^3 + 21.54z^4 + 7.33z^5, \\
g_1(z) &= 0.52 + 0.84z - 3.00z^2 + 7.99z^3 - 9.01z^4 + 3.56z^5, \\
z &\sim 2 \times \mathsf{Beta}(2, 4) - 1,
\end{aligned}
$$

$$\varepsilon \sim t_2(0,1),$$

$$\tau = 0.$$

The true ATE for this model is 0.04. In Figure 6, we plot a particular sample of 500 observations drawn from this design. The sparseness around τ and the small number of treated observations are apparent in the plot.

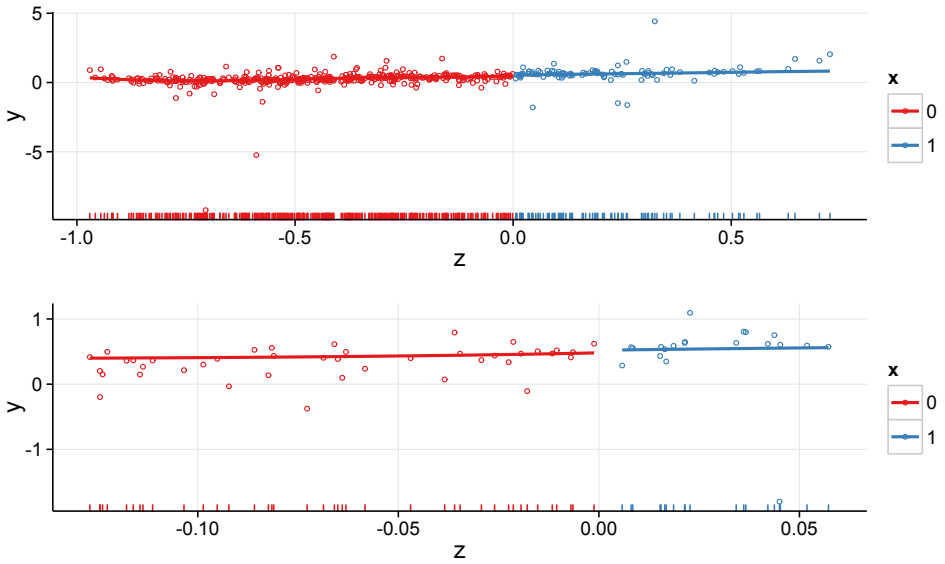

Figure 6: Simulated data - design 2, $n = 500$: True functions and sample data. Bottom panel is data in the top panel plot restricted to the quantiles $p = (.9, .2)$.

5.2.2 Conditional analysis

As before, the analysis and results below are obtained from our R package and are therefore easily reproduced. Our computational strategy for determining the size of the proximate intervals around τ and the number of knots is to perform a marginal likelihood comparison of a limited number of models around the starting specifications $p = (.8, .2)$, $m_z = (4, 4)$, $m_{z,\tau} = (2, 2)$ with Gaussian and Student-t errors. The starting specification is our standard default. We then consider nearby models by adding knots in the various regions, as shown in Table 3. We do not consider models far from the default, because those models are not competitive in the marginal likelihood comparison, which is reasonable in the case of estimating fifth-degree polynomials. The degrees of freedom of the t-distribution are allowed to take the values 4 and 8. The value 2 is intentionally omitted to avoid considering the true model. In

Dist.	m_z	$m_{z,\tau}$	p	log ML
$n = 500$				
t_4	**4 3**	**2 2**	**0.9 0.2**	**-78.674**
t_4	4 4	2 2	0.8 0.2	-78.702
t_8	4 3	2 2	0.9 0.2	-129.022
Gaussian	4 3	2 2	0.9 0.2	-478.017
$n = 1000$				
t_4	4 3	2 2	0.9 0.1	-7.150
t_4	5 3	3 2	0.9 0.1	-9.571
t_4	5 3	2 2	0.9 0.1	-8.552
t_4	4 4	2 2	0.8 0.2	-7.497
t_8	**4 3**	**2 2**	**0.9 0.1**	**-79.060**
t_8	5 3	3 2	0.9 0.1	-81.382
Gaussian	4 3	2 2	0.9 0.1	-672.978
$n = 2000$				
t_4	**4 3**	**2 2**	**0.9 0.1**	**23.780**
t_4	5 3	3 2	0.9 0.1	20.609
t_4	5 3	2 2	0.9 0.1	21.744
t_4	4 4	2 2	0.8 0.2	21.599
t_8	4 3	2 2	0.9 0.1	-116.907
t_8	5 3	3 2	0.9 0.1	-120.266
Gaussian	4 3	2 2	0.9 0.1	-1116.926
$n = 4000$				
t_4	4 4	2 2	0.9 0.1	121.110
t_4	5 4	3 2	0.9 0.1	117.953
t_4	5 3	2 2	0.9 0.1	120.183
t_4	**4 4**	**2 2**	**0.8 0.2**	**121.492**
t_8	4 4	2 2	0.9 0.1	-152.629
t_8	5 4	3 2	0.9 0.1	-155.468
Gaussian	4 4	2 2	0.9 0.1	-1856.054

Table 3: Simulated data - Design 2: Marginal likelihoods for selected models.

the fitting, we suppose that the four hyperparameters in the prior of θ equal 0, that the prior mean of σ^2 equals 0.3 and its prior standard deviation equals 2, and that the prior means and standard deviations of the λ_j equal 1. This prior is relatively benign.

Results of the marginal likelihood comparison are contained in Table 3. The model specifications that yield the largest value of the marginal likelihoods are in bold. It is noteworthy that the marginal likelihood is informative about the length of the proximate region around τ and that it strongly prefers the Student-t model with 4 degrees of freedom to the model with

8. Furthermore, the support for the Student-t assumption over the Gaussian is overwhelming.

5.2.3 Comparison

Consider now the models that are picked in Table 3. It is worthwhile to compare the Bayesian and frequentist estimates of the ATE (the latter computed by the Imbens and Kalyanaraman (2012) method), by sample size. These results, given in Table 4, show that for these data the

	Bayes		Frequentist	
n	Mean	sd	Estimate	sd
500	0.029	0.104	0.172	0.114
1000	0.034	0.097	0.001	0.077
2000	0.178	0.063	0.164	0.083
4000	0.060	0.033	0.005	.053

Table 4: Simulated data - Design 2: Bayes and Frequentist ATE estimates, conditional on a given sample, by sample size. The true value of the ATE is 0.04.

Bayes posterior mean is closer to the true value, except in the $n = 2000$ sample. In addition, the posterior standard deviations are smaller than the frequentist standard errors for every sample size.

It remains now to discuss the results from the sampling investigation of the Bayesian and frequentist point and interval estimates of the ATE. The RMSE and coverage estimates are again based on 10,000 data sets for each sample size. The results, reported in Table 5, demonstrate that, even in this rather complex setting, the sampling distribution of the Bayes posterior mean is less biased and has substantially smaller frequentist RMSE than the frequentist estimator, for all sample sizes. The frequentist coverage of the Bayesian and frequentist ATE interval estimates is similar in this design though the frequentist coverage tends to be smaller than the nominal value.

6 Simulated Data: Fuzzy RD Design

This section is devoted to a study of our approach for the fuzzy RD design. The intent, as above, is to document both the conditional and sampling behavior of the Bayesian estimate of the RD ATE (specifically, the RD ATE for compliers) in comparison with the corresponding

	Bayes			Frequentist		
	ATE	coverage	RMSE	ATE	coverage	RMSE
$n = 500$						
mean	0.042	0.968	0.096	0.043	0.926	0.570
q.025	-0.021	1.000		-0.051	1.000	
q.975	0.106	1.000		0.154	1.000	
$n = 1000$						
mean	0.039	0.967	0.086	0.049	0.934	0.238
q.025	-0.018	1.000		-0.031	1.000	
q.975	0.095	1.000		0.132	1.000	
$n = 2000$						
mean	0.041	0.966	0.061	0.046	0.942	0.148
q.025	0.001	1.000		-0.019	1.000	
q.975	0.082	1.000		0.111	1.000	
$n = 4000$						
mean	0.040	0.964	0.042	0.043	0.937	0.112
q.025	0.011	1.000		0.960	1.000	
q.975	0.068	1.000		1.576	1.000	

Table 5: Simulated data - design 2: Summary of the Bayesian and frequentist ATE sampling distributions from 10,000 samples. The true value of the ATE is 0.04.

frequentist estimator of Imbens and Kalyanaraman (2012) as implemented in the R package *rdd*. The latter estimator is a version of the standard IV estimator, adapted to the specifics of the fuzzy RD design. Surprisingly, the performance of the freqentist fuzzy RD estimator has not been examined before in any simulation experiment.

6.1 Data generating process

Our data is generated by simulating s for each observation with probabilities of type in (1.3) given by

$$q = (.5, .25, .25)$$

and supposing, as in the previous designs, that

$$z \sim 2\text{Beta}(2, 4) - 1$$

and $\tau = 0$. Given the type and z for any observation, x is generated according to the assignment model in (1.4)-(1.6). We then generate the continuous and nonlinear confounders

as

$$w \sim N(0,1), \quad v \sim U(0,1),$$

and then given the sampled type s and treatment x, we generate the outcome from the appropriate model in (1.7) where we suppose that the model for compliers is the same as the one above. Specifically, we suppose that

$$\begin{aligned}
g_0(z) &= 0.48 + 1.27z + 7.18z^2 + 20.21z^3 + 21.54z^4 + 7.33z^5, \\
g_1(z) &= 0.52 + 0.84z - 3.00z^2 + 7.99z^3 - 9.01z^4 + 3.56z^5, \\
h(w) &= \frac{\sin(\pi w/2)}{1 + w^2(\text{sign}(w) + 1)}, \\
h_n(w) &= \frac{\sin(\pi w/2)}{2 + w^2(\text{sign}(w) + 1)}, \\
h_a(w) &= \frac{\sin(\pi w/2)}{3 + w^2(\text{sign}(w) + 1)}, \\
\gamma &= 1, \quad \gamma_n = 3, \quad \gamma_a = 4, \\
\sigma^2 &= 0.1295^2, \quad \sigma_n^2 = \sigma_a^2 = 0.1^2
\end{aligned}$$

The true value of ATEC is 0.04. In the experiments, we use this DGP to create data sets with sample sizes of $n = 500$, 1000, 2000 and 4000.

6.2 Knots and prior distribution

As before, the analysis is conducted under our default preferences: the two regions proximate to $\tau = 0$ are defined by the quantile values $p = (0.9, 0.2)$ when $n = 500$ and $p = (0.9, 0.1)$ for the other values of n. The number of knots in the expansion of the g_0 and g_1 functions are defined as $m_z = (4, 3)$ and $m_{z,\tau} = (2, 2)$, for $n = 500, 1000, 2000$, and by $m_z = (4, 4)$ and $m_{z,\tau} = (3, 3)$, for $n = 4000$. Each of the three h functions, namely h, h_n and h_a, has a basis expansion with and $m_w = 5$ knots, for each sample size.

The prior distribution is likewise specified in a default way. The prior means of the first two ordinates of g_0, the last two of g_1, and the first two of h, h_n, and h_a, are assumed to be zero. The prior means of γ, γ_n and γ_a are also assumed to be 0. The prior means of σ^2, σ_n^2, and σ_a^2 are each assumed to be 0.3 with a prior standard deviation of 3.0. The prior mean of each of $(\lambda_0, \lambda_1, \lambda_\gamma, \lambda_\delta)$ in the complier model, each of $(\lambda_{\gamma_n}, \lambda_{\delta_n})$ in the never-takers

model, and each of $(\lambda_{\gamma_a}, \lambda_{\delta_a})$ in the always-takers model, is assumed to be one with a prior standard deviation of 1. Finally, the Dirichlet prior of q is defined by the hyperparameters $(n_{0c}, n_{0n}, n_{0a}) = (5, 2, 2)$. This same prior is used for each sample size.

6.3 Conditional analysis

For brevity, we focus attention only on the conditional results as they pertain to the ATEC, for each of the four sample sizes. These are given in Figure 7 and are based on output from our MCMC sampling on a burn-in of 1,000 iterations and a retained sample of 10,000 iterations. Once again we observe the tendency of the posterior distribution to concentrate on the true value as the sample size increases.

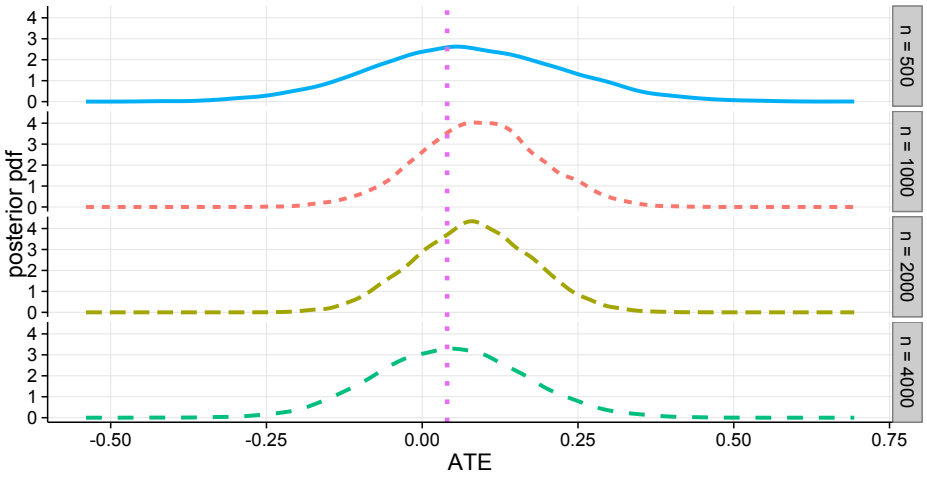

Figure 7: Simulated data - Fuzzy design: Posterior pdf of the ATEC conditional on a given sample, by sample size. True value of 0.04 is indicated by a vertical line.

Next, in Table 6, we compare the Bayesian posterior mean and posterior standard deviation of the ATEC, by sample size, against the corresponding frequentist estimates. We observe that the Bayes estimates are closer to the true ATEC except in the first case.

6.4 Sampling performance

We conclude by providing the frequentist properties of the Bayesian and frequentist point and interval estimates of the ATEC. The RMSE and coverage estimates are based on 10,000 data sets for each sample size. The numbers of knots and prior parameters are those specified

	Bayes		Frequentist	
n	Mean	sd	Estimate	sd
500	0.0724	0.1553	0.0457	0.5364
1000	0.0917	0.0992	-0.0670	0.4534
2000	0.0807	0.0942	-0.2271	0.2359
4000	0.0449	0.1198	0.3383	0.1823

Table 6: Simulated data - fuzzy design: Bayes and frequentist ATEC estimates, conditional on a given sample, by sample size. The true value of the ATEC is 0.04.

in Section 6.2. It is evident from the results in Table 7 that, for each sample size, the sam-

	Bayes			Frequentist		
	ATEC	coverage	RMSE	ATEC	coverage	RMSE
$n = 500$						
mean	0.033	0.968	0.157	0.057	0.958	0.648
q.025	-0.052	1.000		-0.289	1.000	
q.975	0.123	1.000		0.438	1.000	
$n = 1000$						
mean	0.030	0.955	0.143	0.071	0.952	0.399
q.025	-0.057	1.000		-0.172	1.000	
q.975	0.116	1.000		0.325	1.000	
$n = 2000$						
mean	0.036	0.945	0.109	0.072	0.944	0.277
q.025	-0.030	1.000		-0.103	1.000	
q.975	0.105	1.000		0.255	1.000	
$n = 4000$						
mean	0.036	0.940	0.079	0.075	0.942	0.192
q.025	-0.011	1.000		-0.044	1.000	
q.975	0.087	1.000		0.196	1.000	

Table 7: Simulated data - Fuzzy design: Summary of the Bayesian and frequentist ATE sampling distributions from 10,000 samples. The true value of the ATEC is 0.04.

pling distribution of the Bayes ATEC estimate is less biased and has smaller RMSE than the frequentist ATEC estimate, paralleling the results seen in the sharp RD cases.

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 second-difference prior on the spline coefficients that

is capable of handling the situation of many unequally spaced knots. 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 features of the model can be compared through marginal likelihoods and Bayes factors. Our methods are also easily implemented through available R packages, and examples show that the Bayesian RD ATE and RD ATEC estimates have superior frequentist properties than the corresponding frequentist estimates.

Extensions of the method are possible. For instance, the method of Albert and Chib (1993) can be easily embedded in the approach to fit RD models with binary and categorical outcomes. The approach can also be extended to multivariate outcomes and multiple thresholds. 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 denote its argument. For any point $z \in R$ and the set of knots κ_j, $j = 1, \ldots, m$, the basis functions are the collections of cubic splines $\{\Phi_j(z)\}_{j=1}^m$ and $\{\Psi_j(z)\}_{j=1}^m$, where

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

$$\Psi_j(z) = \begin{cases} 0, & z < \kappa_{j-1}, \\ (1/h_j^2)(z - \kappa_{j-1})^2(z - \kappa_j), & \kappa_{j-1} \le z < \kappa_j, \\ (1/h_{j+1}^2)(z - \kappa_{j+1})^2(z - \kappa_j), & \kappa_j \le z < \kappa_{j+1}, \\ 0, & z \ge \kappa_{j+1}, \end{cases} \tag{A.2}$$

and $h_j = \kappa_j - \kappa_{j-1}$ is the spacing between the $(j-1)$st and jth knots. Note that Φ_1 and Ψ_1 are defined by the last two lines of equations (A.1) and (A.2), respectively, and that Φ_m and Ψ_m are defined by only the first two lines. In both cases the strong inequality at the upper limit should be replaced by a weak inequality.

The representation of $g(z)$ as a natural cubic spline is given by

$$g(z) = \sum_{j=1}^{m} \left(\Phi_j(z) f_j + \Psi_j(z) s_j \right), \tag{A.3}$$

where

$$f = (f_1, \ldots, f_m)' \quad \text{and} \quad s = (s_1, \ldots, s_m)'$$

are the coefficients of this cubic spline. Conveniently, $f_j = g(\kappa_j)$ is the function value at the jth knot, and $s_j = g'(\kappa_j)$ is the slope at the jth 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 $Cf = As$, or

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

where

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

and

$$C = 3 \begin{pmatrix} -\frac{1}{h_2} & \frac{1}{h_2} & 0 & 0 & \ldots & 0 & 0 & 0 \\ -\frac{\omega_2}{h_2} & \frac{\omega_2}{h_2} - \frac{\mu_2}{h_3} & \frac{\mu_2}{h_3} & 0 & \ldots & 0 & 0 & 0 \\ 0 & -\frac{\omega_3}{h_3} & \frac{\omega_3}{h_3} - \frac{\mu_3}{h_4} & \frac{\mu_3}{h_4} & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ldots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \ldots & -\frac{\omega_{m-1}}{h_{m-1}} & \frac{\omega_{m-1}}{h_{m-1}} - \frac{\mu_{m-1}}{h_m} & \frac{\mu_{m-1}}{h_m} \\ 0 & 0 & 0 & 0 & \ldots & 0 & -\frac{1}{h_m} & \frac{1}{h_m} \end{pmatrix}.$$

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

$$\begin{aligned} g(z_i) &= \sum_{j=1}^{m} \left(\Phi_j(z_i) f_j + \Psi_j(z_i) s_j \right) \\ &= \left(\Phi(z_i)' + \Psi(z_i)' A^{-1} C \right) f \\ &= b_i' f, \end{aligned}$$

where

$$\Phi(z_i)' = \left(\Phi_1(z_i), \ldots, \Phi_m(z_i) \right),$$

$$\Psi(z_i)' = (\Psi_1(z_i), \ldots, \Psi_m(z_i)),$$

and

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

which implies the following representation for the $n \times m$ basis matrix:

$$B(z) = (b_1, \ldots, b_m).$$

B Appendix: D_α and D_β

Suppose that m is the dimension of α and β. Then, the matrices D_α and D_β in equations (3.3) and (3.6) take the following forms:

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

and

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

respectively.

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.

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

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

Chib, S. and Greenberg, E. (2010), "Additive Cubic Spline Regression with Dirichlet Process Mixture Errors," *Journal of Econometrics*, 156, 322–336.

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

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

Frangakis, C. E. and Rubin, D. B. (2002), "Principal Stratification in Causal Inference," *Biometrics*, 58, 21–29.

Hahn, J. Y., Todd, P., and Van der Klaauw, W. (2001), "Identification and estimation of treatment effects with a regression-discontinuity design," *Econometrica*, 69, 201–209.

Imbens, G. and Kalyanaraman, K. (2012), "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," *Review of Economic Studies*, 79, 933–959.

Imbens, G. W. and Lemieux, T. (2008), "Regression discontinuity designs: A guide to practice," *Journal of Econometrics*, 142, Banff Int Res Stn.

Lancaster, P. and Šalkauskas, K. (1986), *Curve and Surface Fitting: An Introduction*, San Diego: Academic Press.

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

Lee, D. S. and Lemieux, T. (2010), "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, 48, 281–355.

Rau, T. (2011), "Bayesian inference in the regression discontinuity model," 1Pontificia Universidad Católica de Chile.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semparametric Regression*, Cambridge Series in Statistical and Probabilistic Mathematics, New York: Cambridge University Press.

Zhou, S., Shen, X., and Wolfe, D. (1998), "Local Asymptotics for Regression Splines and Confidence Regions," *Annals of Statistics*, 26, 1760–1782.