Econometric Theory, **39**, 2023, 481–533.
doi:10.1017/S0266466622000019

NONPARAMETRIC BAYES ANALYSIS OF THE SHARP AND FUZZY REGRESSION DISCONTINUITY DESIGNS

Siddhartha Chib
Washington University in St. Louis

Edward Greenberg*
Washington University in St. Louis

Anna Simoni
CREST, CNRS, ENSAE, École Polytechnique

We develop a nonparametric Bayesian analysis of regression discontinuity (RD) designs, allowing for covariates, in which we model and estimate the unknown functions of the forcing variable by basis expansion methods. In a departure from current methods, we use the entire data on the forcing variable, but we emphasize the data near the threshold by placing some knots at and near the threshold, a technique we refer to as soft-windowing. To handle the nonequally spaced knots that emerge from soft-windowing, we construct a prior on the spline coefficients, from a second-order Ornstein–Uhlenbeck process, which is hyperparameter light, and satisfies the Kullback–Leibler support property. In the fuzzy RD design, we explain the divergence between the treatment implied by the forcing variable, and the actual intake, by a discrete confounder variable, taking three values, complier, never-taker, and always-taker, and a model with four potential outcomes. Choice of the soft-window, and the number of knots, is determined by marginal likelihoods, computed by the method of Chib [*Journal of the American Statistical Association*, 1995, 90, 1313–1321] as a by-product of the Markov chain Monte Carlo (MCMC)-based estimation. Importantly, in each case, we allow for covariates, incorporated nonparametrically by additive natural cubic splines. The potential outcome error distributions are modeled as student-*t*, with an extension to Dirichlet process mixtures. We derive the large sample posterior consistency, and posterior contraction rate, of the RD average treatment effect (ATE) (in the sharp case) and RD ATE for compliers (in the fuzzy case), as the number of basis parameters increases with sample size. The excellent performance of

The authors gratefully thank the Editor Peter Phillips, the Co-Editor Iván Fernández-Val and three anonymous referees for constructive comments. Anna Simoni gratefully acknowledges financial support from ANR grant (ANR-21-CE26-0003-01). Address correspondence to Siddhartha Chib, Olin Business School, Washington University in St. Louis, Campus Box 1133, 1 Bookings Drive, St. Louis, MO 63130, USA; e-mail: chib@wustl.edu.

*Passed away in September 2016, co-author in remembrance of contributions to this research.

the methods is documented in simulation experiments, and in an application to educational attainment of women from Meyersson [*Econometrica*, 2014, 82, 229–269].

1. INTRODUCTION

For causal inference with observational data, regression discontinuity (RD) and fuzzy RD designs (Thistlethwaite and Campbell, 1960; Campbell, 1969) have been an active area of research for many years, for example, Hahn, Todd, and Van der Klaauw (2001), Imbens and Lemieux (2008), Lee and Lemieux (2010), Frandsen, Froelich, and Melly (2012), Calonico, Cattaneo, and Titiunik (2014) and Cattaneo, Titiunik, and Vazquez-Bare (2017).

Our aim in this article is to develop a nonparametric Bayesian perspective on RD designs. One organizing theme is that we estimate the unknown functions of the forcing variable by basis expansion methods, which is relatively unexplored in this area. In a departure from most current methods, we use the entire data on the forcing variable, but we emphasize the data near the threshold by a technique that we call *soft-windowing*. The key advantage of this approach is that we can use marginal likelihoods to compare different versions of the models, say different soft-window characteristics crossed with different distributional assumptions or covariates. Such comparisons are infeasible in methods based on hard-windowing (the dominant approach in the frequentist literature) since different hard-windowing specifications produce different datasets.

To handle the nonequally spaced knots that emerge from soft-windowing, we introduce a second-difference prior on the spline coefficients that acts as a suitable regularizer, even when the number of knots is large. For the fuzzy RD design, we explain the divergence between the treatment implied by the forcing variable, and the actual intake, by a new model that adds a fresh perspective to the existing literature on these designs. Another central concern is the derivation of the theoretical large sample properties of the posterior distributions, and the rates of contraction.

The distinguishing feature of the sharp design is that the intake (the treatment) $x \in \{0,1\}$ is determined by a forcing variable z by the deterministic rule $x = I[z \geq \tau]$, where $I[.]$ is the indicator function, and τ is a known discontinuity point. As usual, there are two potential outcomes, y_0 and y_1 (say both in $\mathbb{R}$), with the observed outcome given by $y = (1-x)y_0 + xy_1$. We suppose that $y_j = g_j(z) + h(w) + \sigma_j \varepsilon_j$, where $g_j(\cdot)$ are smooth unknown functions of z that are each continuous at τ, $w \in \mathbb{R}^{k_w}$ are covariates, $h(\cdot)$ is an unknown smooth function additive in each component of w, and the errors are student-t with $\nu > 2$ degrees of freedom with separate dispersion parameters σ_j. The student-t assumption is a reasonable baseline starting point, especially when the sample size is not large, but, in Section 6, we also consider a nonparametric formulation by putting a Dirichlet process prior on this distribution.

We rely on basis expansion techniques with cubic splines to nonparametrically estimate g_0 from data on $z < \tau$, and g_1 from data on $z \geq \tau$. Instead of splines,

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

a Gaussian process as, for example, in Branson et al. (2019), could be used. Estimation in much of the frequentist literature, on the other hand, is based on variants of local polynomial methods (or kernel methods) to data limited to a window around the threshold, for example, Hahn, Todd, and Van der Klaauw (2001), Imbens and Kalyanaraman (2012) and Calonico, Cattaneo, and Titiunik (2014), or on variants of local randomization methods, see for example, Cattaneo, Frandsen, and Titiunik (2015) and Cattaneo, Titiunik, and Vazquez-Bare (2017).

Although we use the entire data on the forcing variable to estimate g_j, we emphasize the data near τ by soft-windowing, placing some knots at and near the threshold. We show that the choice of the soft-window, and the number of knots, affect the marginal likelihood (the integral of the sampling density over the prior) and, thus, these design parameters can be adjusted/optimized by the resulting marginal likelihoods, which we can compute by the method of Chib (1995) as a by-product of the Markov chain Monte Carlo (MCMC)-based estimation. In this way, the data determine the soft-window around the threshold that is most relevant for inferences about the RD effect.

In the fuzzy RD design, we explain the divergence between the assignment rule, $I[z \geq \tau]$, and the treatment, x, (the hallmark of such designs) by an *unobserved* confounder that denotes subject type, taking the values $\{c, n, a\}$ for complier, never-taker, and always-taker, respectively. See Chib and Jacobi (2016) for a concrete illustration of this approach. There are now four potential outcomes, y_0 and y_1 for the compliers, and y_{0n} and y_{1a} for never-takers and always-takers, respectively. Conditioned on (z, w) and $s = c$, the potential outcomes y_0 and y_1 are generated as in the sharp model. For the new types, conditioned on (z, w) we have $y_{0n} = g_{0n}(z) + h_n(w) + \sigma_{0n}\varepsilon_{0n}$, and $y_{1a} = g_{1a}(z) + h_a(w) + \sigma_{1a}\varepsilon_{1a}$, where the functions g_{0n} and g_{1a} are continuous at τ. These functions can be identified because never-takers and always-takers exist on both sides of τ. The functions $h_n(\cdot)$ and $h_a(\cdot)$ are unknown smooth functions of the control variables w. The $(4 + 3 \times k_w)$ nonparametric functions in this model are also estimated by basis expansion techniques. The object of interest is the RD average treatment effect for compliers, $\lim_{z \downarrow \tau^+} \mathbf{E}[y_1 | z, w, s = c] - \lim_{z \uparrow \tau^-} \mathbf{E}[y_0 | z, w, s = c]$, which we show is identified under weak assumptions.

In both models, we derive the large-sample rates of contraction of the average treatment effect (ATE) and complier average treatment effect (CATE) posterior distributions. These results (for non-Gaussian error distributions) are new to the Bayesian nonparametric literature. Branson et al. (2019) also present a posterior consistency result, but their result is only for the sharp design and under Gaussian errors with known variances. Our derivations exploit a representation of the ATE as a linear functional on the space of square integrable functions with respect to the empirical distribution of the forcing variable, along with new techniques for bounding functions that arise under our error assumptions. Interestingly, the posterior contraction rate we derive is the same, up to a logarithmic factor, as the one in Calonico, Cattaneo, and Titiunik (2014).

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

The methods proposed in this article complement and broaden the frequentist approach in several directions. Soft-windowing avoids the need to specify, or select, a hard-window. As a result, we are seamlessly able to compare models with different soft-windows by marginal likelihoods. Such comparisons of different hard-window specifications are infeasible in the frequentist case because different hard-windows produce different datasets. In addition, the posterior distribution of the ATE supplies the entire summary of the effect, conditioned on the data. This can be useful because in finite samples, the posterior distribution is not necessarily symmetric, or even unimodal. In the frequentist case, interval estimates are based on large-sample theory and are always symmetric around the point estimate. Our approach is also useful from a purely frequentist perspective. We document that the Bayesian posterior mean and interval estimates have excellent sampling properties in finite samples, competitive with the root mean square error (RMSE) optimal, and coverage-optimal, frequentist estimators. In addition, we prove that our procedures in the sharp and fuzzy cases are asymptotically valid from a frequentist point of view. Finally, these procedures are easy to implement by software produced by us. Thus, it is possible now to calculate the Bayesian effects, along with the frequentist effects, with ease that rivals that of the existing approaches.

The remainder of the article is organized as follows. In Section 2, we consider the sharp RD design and introduce the key ideas, while in Section 4, we consider the fuzzy RD design. Large sample analysis is provided in Sections 2 and 4. Simulation studies are given in Sections 3 and 5. In Section 3, we also provide a real data application of our method to an application from Meyersson (2014). Extensions of the model to nonparametric error distributions are in Section 6, and conclusions in Section 7. Details related to the basis expansions and the main proofs of the theorems are given in Appendixes A–C, and in the Online Supplementary Appendix.

2. SHARP RDD

We make the following assumptions.

Assumption 1 (Conditional expectations)**.** For $j = 0,1$, there exists two functions g_j and h that depends only z and w, respectively, such that:

$$\mathbf{E}[y_j|z,w] = g_j(z) + h(w).$$

Assumption 2 (Smoothness)**.**

1. For $j = 0,1$ and $\delta > 0$, the function $z \mapsto g_j(z)$ is δ times continuously differentiable on an interval that contains τ.
2. The function $h(w)$ is continuous.

Assumption 3 (Distributions)**.**

1. The forcing variable z and the covariates w have a continuous Lebesgue joint density.

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

2. The potential outcomes are generated as: for $j = 0, 1$,

$$y_j = \mathbf{E}[y_j|z,w] + \sigma_j \varepsilon_j,$$

where ε_j is distributed as standard student-t with $\nu > 2$ degrees of freedom and $\sigma_j > 0$.

Some remarks. Apart from the presence of the covariates w, Assumptions 1 and 2.1 correspond to Hahn et al. (2001, Assumptions (A1), (A2)), Imbens and Lemieux (2008, Assumption 2.1) and Calonico et al. (2014, Assumption 1(b)). Assumption 3.1 is also required in Calonico et al. (2014, Assumption 1(a)). The local conditional independence assumption corresponds to the first assumption in Hahn et al. (2001, Theorem 2) except for the presence of w. It is automatically satisfied in the sharp model since x is determined given z. We use this fact in showing identification. The distributional Assumption 3 relaxes the Gaussianity assumption in Branson et al. (2019). Finally, these assumptions rule out the possibility that individuals are manipulating z around τ strategically, that is, the possibility that z is related to ε_j, for example, see Lee (2008) and McCrary (2008).

The object of interest is the RD ATE, that is, the average treatment effect at τ, defined as

$$\lim_{z\downarrow\tau^+} \mathbf{E}[y_1|z,w] - \lim_{z\uparrow\tau^-} \mathbf{E}[y_0|z,w] = \lim_{z\downarrow\tau^+} g_1(z) - \lim_{z\uparrow\tau^-} g_0(z). \tag{2.1}$$

By continuity of $g_j(\cdot)$, this is equal to $g_1(\tau) - g_0(\tau)$.

Under Assumptions 1 and 2, the RD ATE is identified. Fix a w in the support of $w|(z$ around $\tau)$. Then, from the observed data on the right side of τ,

$$\begin{aligned}\lim_{z\downarrow\tau^+} \mathbf{E}[y|z,w] &= \lim_{z\downarrow\tau^+} \mathbf{E}[x|z,w] \lim_{z\downarrow\tau^+} \mathbf{E}[y_1|z,w]\\ &= \lim_{z\downarrow\tau^+} g_1(z) + h(w)\end{aligned}$$

since the first term on the right-hand side in the first line is 1. Similarly, from the observed data on the left side of τ,

$$\begin{aligned}\lim_{z\uparrow\tau^-} \mathbf{E}[y|z,w] &= \lim_{z\uparrow\tau^-} \mathbf{E}[(1-x)|z,w] \lim_{z\uparrow\tau^-} \mathbf{E}[y_0|z,w]\\ &= \lim_{z\uparrow\tau^-} g_0(z) + h(w).\end{aligned}$$

The RD ATE is the difference in these two observed data limits.

2.1. Sample Data

The available data are n independent observations (y_i, x_i, z_i), $i \leq n$, where y_i is equal to y_{0i} when $x_i = 0$ and y_{1i} when $x_i = 1$ and satisfies the above assumptions. For simplicity, until Section 6, suppose that w is absent to minimize the notational burden. Let n_0 (resp. n_1) denote the number of observations to the left (resp. right) of τ, with $n = n_0 + n_1$. Assemble the vector of observations on (y, z) to the left of τ as

$$\boldsymbol{y}_0 \triangleq (y_1, \ldots, y_{n_0})' \quad (n_0 \times 1), \quad \boldsymbol{z}_0 \triangleq (z_1, \ldots, z_{n_0})' \quad (n_0 \times 1),$$

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

and those to the right of τ as

$$\boldsymbol{y}_1 \triangleq (y_{n_0+1},\dots,y_n)' \quad (n_1 \times 1), \quad \boldsymbol{z}_1 \triangleq (z_{n_0+1},\dots,z_n)' \quad (n_1 \times 1).$$

For later reference, define $z_{j,\,\mathsf{min}} \triangleq \mathsf{min}(\boldsymbol{z}_j)$, $z_{j,\,\mathsf{max}} \triangleq \mathsf{max}(\boldsymbol{z}_j)$, for $j=0,1$, and the pth quantile of $\boldsymbol{z}_j$ by $z_{j,p}$. Moreover, we note that the likelihood function of this model is

$$p(\boldsymbol{y}|\boldsymbol{z},g_0,g_1,\sigma_0^2,\sigma_1^2) = \prod_{i=1}^{n_0} t_\nu(y_i|g_0(z_i),\sigma_0^2)\prod_{i=1}^{n_1} t_\nu(y_{n_0+i}|g_1(z_{n_0+i}),\sigma_1^2),$$

where $\boldsymbol{y} \triangleq (\boldsymbol{y}_0',\boldsymbol{y}_1')'$, $\boldsymbol{z} \triangleq (\boldsymbol{z}_0',\boldsymbol{z}_1')'$ and t_ν is the student-t density function.

2.2. Soft Windowing and Basis Expansions

We apply natural cubic spline basis expansion techniques, and the basis functions in Chib and Greenberg (2010), to estimate $g_0(z)$ and $g_1(z)$. In this basis, the basis coefficients are the function heights at the chosen knots. We take advantage of this property by expanding $g_0(z)$ and $g_1(z)$ with knots at τ. This reduces the RD ATE to the difference of two basis coefficients. As a side-benefit, by locating knots at τ, the estimates of the g functions over $(z_{0,\,\mathsf{max}},\tau)$ and $(\tau,z_{1,\,\mathsf{min}})$ are not necessarily linear.

We now explain the placement of the remaining knots. A key element of the approach is *soft-windowing*. We start by partitioning $[z_{0,\,\mathsf{min}},\tau]$ and $[\tau,z_{1,\,\mathsf{max}}]$ into intervals that are proximate and far from τ. We determine these four intervals from the quantiles z_{0,p_0} and z_{1,p_1}, for specific values of $\boldsymbol{p} \triangleq (p_0,p_1)$, for example, $\boldsymbol{p} = (0.9,0.1)$. We allocate knots to each of the four intervals under the constraint that there is at least one observation between each successive pair of knots. We let $\boldsymbol{m}_{z,\tau} = (m_{z,0,\tau},m_{z,1,\tau})$ denote the maximum number of knots in the intervals proximate to τ, and $\boldsymbol{m}_z = (m_{z,0},m_{z,1})$ to denote the maximum number of knots in the intervals further away from τ. Note that the no empty interval constraint means that the actual number of knots can be smaller than the maximum numbers.

Our algorithm for placing knots under the constraint of no-empty intervals may be characterized as "propose-check-accept-extend." Consider 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 nonempty interval. Otherwise, propose a knot at $\tau-2\Delta_\tau$, check for a nonempty 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 used on the right of τ, with the first knot at τ and the last at $z_{1,\,\mathsf{max}}$. Let $\{z_{0,\,\mathsf{min}},\kappa_{0,2},\dots,\kappa_{0,m_0-1},\tau\}$ denote the m_0 knots to the left of τ determined by this procedure, and let $\{\tau,\kappa_{1,2},\dots,\kappa_{1,m_1-1},z_{1,\,\mathsf{max}}\}$ denote the m_1 knots to the right of τ. A particular allocation of knots is shown in Figure 1, where $m_0 = 10$ and $m_1 = 7$.

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

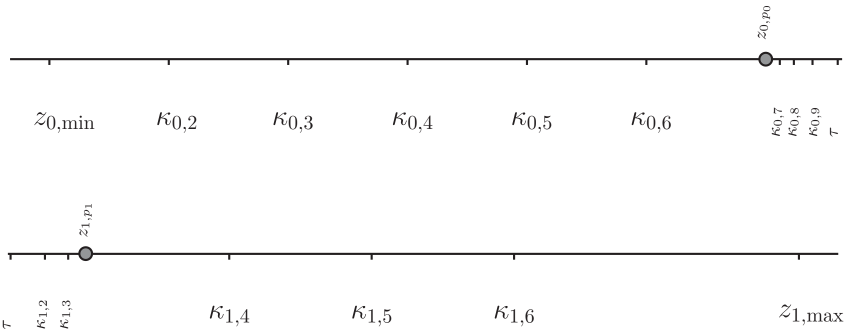


FIGURE 1. Example of knot locations in the basis expansions of g_0 (top panel) and g_1 (bottom panel), determined by $\boldsymbol{m}_z = (6,5)$, $\boldsymbol{m}_{z,\tau} = (5,5)$; these specify the maximum number of knots. Note that the no empty interval constraint means that the actual number of knots can be smaller than the maximum numbers. 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 τ.

When using this algorithm, note that

$$m_0 \leq m_{z,0} + m_{z,0,\tau} \quad \text{and} \quad m_1 \leq m_{z,1,\tau} + m_{z,1},$$

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

Now, by basis expansions, the function ordinates,

$$g_0(\boldsymbol{z}_0) \triangleq \big(g_0(z_1), \ldots, g_0(z_{n_0})\big) \quad \text{and} \quad g_1(\boldsymbol{z}_1) \triangleq \big(g_1(z_{n_0+1}), \ldots, g_1(z_n)\big)$$

can be approximated as

$$g_0(\boldsymbol{z}_0) \approx g_{m_0}(\boldsymbol{z}_0) \triangleq \boldsymbol{B}_0\boldsymbol{\alpha} \quad \text{and} \quad g_1(\boldsymbol{z}_1) \approx g_{m_1}(\boldsymbol{z}_1) \triangleq \boldsymbol{B}_1\boldsymbol{\beta}, \tag{2.2}$$

respectively, where $\boldsymbol{B}_j : n_j \times m_j$ are the basis matrices evaluated at $\boldsymbol{z}_j$, and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the basis coefficients. As shown in Section 2.5, posterior consistency requires that m_j increase with sample size at the rate $(n_j/\log(n_j))^{\nu}$, where ν is a constant dependent on the smoothness of the function g_j, $j = 0, 1$.

Since, in our basis,

$$\underset{(m_0\times 1)}{\boldsymbol{\alpha}} = \begin{pmatrix} g_0(z_{0,\,\min}) \\ g_0(\kappa_{0,2}) \\ \vdots \\ g_0(\kappa_{0,m_0-1}) \\ g_0(\tau) \end{pmatrix}, \quad \underset{(m_1\times 1)}{\boldsymbol{\beta}} = \begin{pmatrix} g_1(\tau) \\ g_1(\kappa_{1,2}) \\ \vdots \\ g_1(\kappa_{1,m_1-1}) \\ g_1(z_{1,\,\max}) \end{pmatrix}, \tag{2.3}$$

the RD ATE is the first component of $\boldsymbol{\beta}$ minus the last component of $\boldsymbol{\alpha}$:

$$\text{ATE} = \boldsymbol{\beta}_{[1]} - \boldsymbol{\alpha}_{[m_0]}. \tag{2.4}$$

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

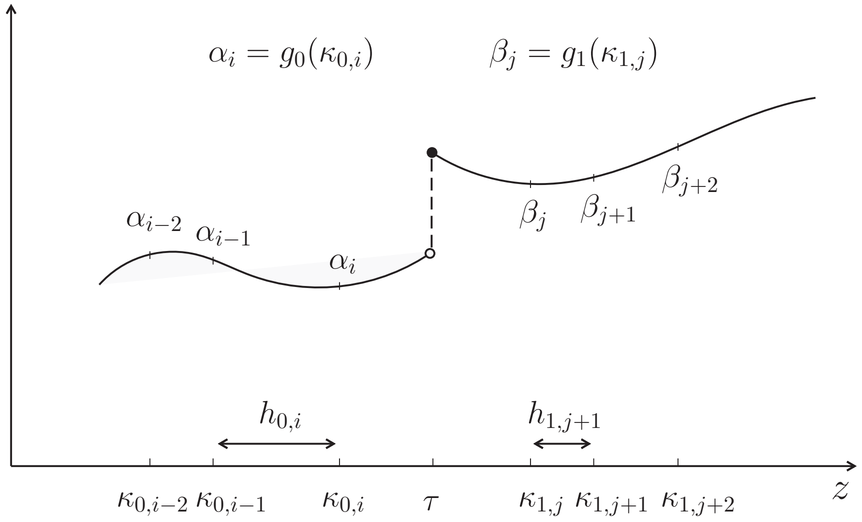


FIGURE 2. Prior formulation: three successive knots on either side of τ and the corresponding function ordinates. The latter are the basis coefficients in the natural cubic spline basis expansions of g_0 and g_1. The prior on these coefficients is defined through a second-order O–U process. The process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

2.3. Prior Distribution

To handle the nonequally spaced knots that emerge from soft-windowing, we introduce a new second-difference prior on the spline coefficients that acts as a suitable regularizer even when the number of knots is large. The prior in Lang and Brezger (2004) and Brezger and Lang (2006) assumes that the knots are equally spaced and that the first two knots have an improper prior, which precludes model comparisons by marginal likelihoods.

We develop our prior of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ from a second-order Ornstein–Uhlenbeck (O–U) process, which, in continuous time, for a diffusion $\{\varphi_t\}$, is given by 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 Euler-discretize this process, letting dt equal the spacing between successive knots, $a = 1$, $b = 0$, and $s = 1/\sqrt{\lambda}$, where λ is a penalty parameter.

Prior of $\boldsymbol{\alpha}$: 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

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

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

and define the spacings between knots by $h_{0,i} \triangleq \kappa_{0,i} - \kappa_{0,i-1}$, as shown in Figure 2. We suppose now that, a priori, $(\alpha_3, \alpha_4, \ldots, \alpha_{m_0})$, conditioned on (α_1, α_2), follow the process

$$\Delta^2\alpha_i = -(\alpha_{i-1} - \alpha_{i-2})h_{0,i} + u_{0i}, \tag{2.5}$$

$$u_{0i}|\lambda_0 \sim \mathcal{N}\left(0, \lambda_0^{-1}h_{0,i}\right), \tag{2.6}$$

where $(\alpha_{i-1} - \alpha_{i-2})h_{0,i}$ introduces mean reversion and λ_0 is an unknown precision (smoothness) parameter.

To complete this process, we specify a distribution of (α_1, α_2). Let $\boldsymbol{T}_{\alpha,1:2}^{-1} \triangleq \left(\boldsymbol{B}_0'\boldsymbol{B}_0\right)_{1:2}$ denote the first two rows and columns of $\boldsymbol{B}_0'\boldsymbol{B}_0$. We then let

$$\begin{pmatrix}\alpha_1 \\ \alpha_2\end{pmatrix} = \begin{pmatrix} g_0(z_{0,\,\mathsf{min}}) \\ g_0(\kappa_{0,2}) \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix}\alpha_{1,0} \\ \alpha_{2,0}\end{pmatrix}, \lambda_0^{-1}\boldsymbol{T}_{\alpha,1:2}\right),$$

where $\alpha_{1,0}$ and $\alpha_{2,0}$ (the prior expected levels of g_0 at the first two knots) are the only two free hyperparameters.

By straightforward calculations, the implied joint prior distribution is

$$\boldsymbol{\alpha}|\lambda_0 \sim \mathcal{N}_{m_0}\left(\boldsymbol{D}_\alpha^{-1}\boldsymbol{\alpha}_0, \lambda_0^{-1}\boldsymbol{D}_\alpha^{-1}\boldsymbol{T}_\alpha\boldsymbol{D}_\alpha^{-1'}\right), \tag{2.7}$$

where $\boldsymbol{\alpha}_0 \triangleq (\alpha_{1,0}, \alpha_{2,0}, 0, \ldots, 0)'$ is $m_0 \times 1$, $\boldsymbol{D}_\alpha$ is a tri-diagonal matrix (given in Appendix B) that depends entirely on the spacings, and $\boldsymbol{T}_\alpha \triangleq \text{blockdiag}(\boldsymbol{T}_{\alpha,1:2}, \boldsymbol{I}_{m_0-2})$ is $m_0 \times m_0$. Note that, under this prior, the diagonal elements of $\boldsymbol{D}_\alpha^{-1}\boldsymbol{T}_\alpha\boldsymbol{D}_\alpha^{-1'}$ increase as one moves down the diagonal, which implies that $\text{Var}\big(g_0(z_{0,\,\mathsf{min}})\big) < \text{Var}\,(g_0(\tau))$. Also note that this prior is fully specified by the two hyperparameters, $\alpha_{1,0}$ and $\alpha_{2,0}$, which is convenient.

Prior of $\boldsymbol{\beta}$: The prior of $\boldsymbol{\beta}$ is similar except that we orient the process from right to left. We do this in order that the prior of $\beta_1 = g_1(\tau)$ is determined by the O–U process. 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, we suppose that

$$\Delta^2\beta_j = -(\beta_{j+1} - \beta_{j+2})h_{1,j+1} + u_{1j}, \tag{2.8}$$

$$u_{1j}|\lambda_1 \sim \mathcal{N}\left(0, \lambda_1^{-1}h_{1,j+1}\right), \tag{2.9}$$

where $h_{1,j+1} = \kappa_{1,j+1} - \kappa_{1,j}$ is the spacing between knots and λ_1 is an unknown precision parameter, specific to $\boldsymbol{\beta}$.

For $(\beta_{m_1-1}, \beta_{m_1})$, let $\boldsymbol{T}_{\beta,m_1-1:m_1}^{-1} \triangleq \left(\boldsymbol{B}_1'\boldsymbol{B}_1\right)_{m_1-1:m_1}$ denote the last two rows and columns of $\boldsymbol{B}_1'\boldsymbol{B}_1$. Then, assume that

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

which implies that

$$\boldsymbol{\beta}|\lambda_1 \sim \mathcal{N}_{m_1}\left(\boldsymbol{D}_\beta^{-1}\boldsymbol{\beta}_0, \lambda_1^{-1}\boldsymbol{D}_\beta^{-1}\boldsymbol{T}_\beta\boldsymbol{D}_\beta^{-1'}\right), \qquad \textbf{(2.10)}$$

where $\boldsymbol{\beta}_0 \triangleq (0,\ldots 0, \beta_{m_1-1,0}, \beta_{m_1,0})'$ is $m_1 \times 1$, $\boldsymbol{D}_\beta$ is the tri-diagonal matrix in Appendix B, and $\boldsymbol{T}_\beta \triangleq \text{blockdiag}(\boldsymbol{I}_{m_1-2}, \boldsymbol{T}_{\beta,m_1-1:m_1})$ is $m_1 \times m_1$.

The preceding prior on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ induces a prior on the approximating functions $g_{m_j}, j = 0,1$, which, together with a degenerate prior on zero for the remaining part $g_j - g_{m_j}$, gives a prior for the whole function $g_j, j = 0,1$. This is a sieve type prior as we let the number of components m_j increase with n.

Prior of $\boldsymbol{\lambda}$ and $\boldsymbol{\sigma^2}$: We complete our prior with a Gamma prior distribution on $\lambda_j, j = 0,1$. It is important to keep in mind that for any value of n, $\lambda_j \to 0$ implies an unpenalized regression spline, and $\lambda_j \to \infty$ implies that the second differences are forced to zero, leading to piece-wise linearity. Therefore, depending on the situation, we fix the Gamma hyperparameters in two ways. The first, the default, we specify prior values of $\mathbf{E}(\lambda_j)$ and $\text{sd}(\lambda_j)$ and match a Gamma distribution to these choices. For example, we let $\mathbf{E}(\lambda_j) = 1$ and then let $\text{sd}(\lambda_j) = 5$, where the latter typically increases with the sample size. The second is to choose $\mathbf{E}(\lambda_j)$ to make the smallest diagonal element of the variance matrix equal to one, that is, choose $\mathbf{E}(\lambda_j)$ so that

$$\min\left\{\text{diag}\left(\frac{1}{\mathbf{E}(\lambda_j)}\boldsymbol{D}_j^{-1}\boldsymbol{T}_j\boldsymbol{D}_j^{-1'}\right)\right\} = 1,$$

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

$$\lambda_j \sim \text{Ga}\left(\frac{a_{j0}}{2}, \frac{b_{j0}}{2}\right), \; (j = 0,1). \qquad \textbf{(2.11)}$$

Note that if we let the prior mean of λ_j be small (relative to the prior SD), then that puts more weight on the unpenalized regression spline. On the other hand, a large value for a_{j0} puts more weight on the penalty, as required when the number of knots increase with sample size. We prove below that for posterior consistency, $a_{j0} = C(n_j/\log(n_j))^\nu$, for positive constants C and ν that depend on the smoothness of the function g_j, for $j = 0,1$.

The prior on $\boldsymbol{\sigma^2} \triangleq (\sigma_0^2, \sigma_1^2)$ is of the usual form. Independent of $\boldsymbol{\lambda} \triangleq (\lambda_0, \lambda_1)$, we suppose that

$$\sigma_j^2 \sim \text{IG}\left(\frac{\nu_{00}}{2}, \frac{\delta_{00}}{2}\right), \; (j = 0,1), \qquad \textbf{(2.12)}$$

an inverse-gamma distribution, where ν_{00} and δ_{00} are chosen to reflect the researcher's views about the mean and standard deviation of σ_j^2.

Remark. As shown later, this prior satisfies an important property. Suppose that one admits the existence of a true value (g_j^*, σ_{j*}^2) of (g_j, σ_j^2) for $j = 0,1$. Then, one is interested in knowing whether these true values are in the Kullback–Leibler (KL) support of the prior. More precisely, the true data distribution characterized by (g_j^*, σ_{j*}^2) belongs to the KL support of the prior distribution if the prior assigns positive probability to any KL neighborhood of the true distribution. Our prior satisfies the KL support property if the functions g_0 and g_1 are well approximated by g_{m_0} and g_{m_1}, respectively. In particular, the KL support property is satisfied even for a shrinking KL neighborhood if the hyperparameter a_{j0} of the Gamma prior for λ_j is set equal to $a_{j0} = C(n_j/\log(n_j))^\nu$, for positive constants C and ν that depend on the smoothness of the function g_j, for $j = 0, 1$. We use this property to derive large-sample asymptotic results using KL neighborhoods that shrink at a rate $\epsilon_n \to 0$ indexed by the sample size n that is decreasing in n.

2.4. Posterior Distributions and MCMC Sampling

For observations on either side of τ, the following Bayesian linear models hold (after we integrate with respect to the degenerate prior on zero for $g_j - g_{m_j}$):

$$\underset{(n_j\times 1)}{\boldsymbol{y}_j} = \boldsymbol{B}_j\boldsymbol{\theta}_j + \underset{(n_j\times 1)}{\boldsymbol{\varepsilon}_j}, \boldsymbol{\varepsilon}_j \sim \mathcal{N}_{nj}(0, \boldsymbol{\Xi}_j), \tag{2.13}$$

$$\boldsymbol{\theta}_j|\lambda_j \sim \mathcal{N}_k\left(\boldsymbol{\theta}_{j0}, \boldsymbol{A}_{j0}\right), \ \lambda_j \sim \mathrm{Ga}\left(\frac{a_{j0}}{2}, \frac{b_{j0}}{2}\right), \tag{2.14}$$

$$\sigma_j^2 \sim \mathrm{IG}\left(\frac{\nu_{00}}{2}, \frac{\delta_{00}}{2}\right), \tag{2.15}$$

where $\boldsymbol{\theta}_j$ is $\boldsymbol{\alpha}$ for $j = 0$ and $\boldsymbol{\beta}$ for $j = 1$, and by the Gamma scale-mixture of normals representation of the student-t distribution,

$$\boldsymbol{\Xi}_0^{-1} \triangleq \mathrm{diag}\left(\frac{\xi_1}{\sigma_0^2}, \ldots, \frac{\xi_{n_0}}{\sigma_0^2}\right), \ \boldsymbol{\Xi}_1^{-1} \triangleq \mathrm{diag}\left(\frac{\xi_{(n_0+1)}}{\sigma_1^2}, \ldots, \frac{\xi_n}{\sigma_1^2}\right),$$

where

$$\xi_i \sim \mathrm{Ga}\left(\frac{\nu}{2}, \frac{\nu}{2}\right), \ 1 \le i \le n,$$

and

$$\boldsymbol{\theta}_{00} \triangleq \boldsymbol{D}_\alpha^{-1}\boldsymbol{\alpha}_0\,, \ \boldsymbol{\theta}_{10} \triangleq \boldsymbol{D}_\beta^{-1}\boldsymbol{\beta}_0,$$

$$\boldsymbol{A}_{00} \triangleq \frac{1}{\lambda_0}\boldsymbol{D}_\alpha^{-1}\boldsymbol{T}_\alpha\boldsymbol{D}_\alpha^{-1'}, \ \boldsymbol{A}_{00} \triangleq \frac{1}{\lambda_1}\boldsymbol{D}_\beta^{-1}\boldsymbol{T}_\beta\boldsymbol{D}_\beta^{-1'}.$$

The joint posterior distribution of the parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, (λ_0, λ_1), $\boldsymbol{\sigma}^2 \triangleq (\sigma_0^2, \sigma_1^2)$, and $\{\xi_i\}$ can be sampled easily by MCMC methods (Chib and Greenberg, 1996). The MCMC steps are iterated $N_0 + M$ times, where N_0 is the number of burn-in iterations and M is the number of iterations retained:

- Given $(\boldsymbol{y}_j, \sigma_j^2, \{\xi_i\}, \{\lambda_j\})$, sample $\boldsymbol{\theta}_j$ from $\mathcal{N}_k(\hat{\boldsymbol{\theta}}_j, \boldsymbol{A}_j)$, where $\hat{\boldsymbol{\theta}}_j = \boldsymbol{A}_j(\boldsymbol{A}_{j0}^{-1}\boldsymbol{\theta}_{j0} + \boldsymbol{B}_j'\boldsymbol{\Xi}_j^{-1}\boldsymbol{y}_j)$ and $\boldsymbol{A}_j = (\boldsymbol{A}_{j0}^{-1} + \boldsymbol{B}_j'\boldsymbol{\Xi}_j^{-1}\boldsymbol{B}_j)^{-1}$.

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

- Given $(\boldsymbol{y}_j, \boldsymbol{\theta}_j, \{\xi_i\}, \{\lambda_j\})$, sample σ_j^2 from $\mathrm{IG}\left(\frac{\nu_{00}+n_j}{2}, \frac{\delta_{00}+(\boldsymbol{y}_j-\boldsymbol{B}_j\boldsymbol{\theta}_j)'\Xi_0^{-1}(\boldsymbol{y}_j-\boldsymbol{B}_j\boldsymbol{\theta}_j)}{2}\right)$.
- Given $(\boldsymbol{y}_j, \boldsymbol{\theta}_j, \sigma_j^2)$, sample $\{\xi_i\}$ from

$$\mathrm{Ga}\left(\frac{\nu+1}{2}, \frac{\left(y_{ji}-B_{ji}\boldsymbol{\theta}_j\right)^2/\sigma_j^2}{2}\right).$$

- Given $\boldsymbol{\theta}$, sample $\boldsymbol{\lambda}$ from

$$\lambda_0|\boldsymbol{\alpha} \sim \mathrm{Ga}\left(\frac{a_{00}+m_0}{2}, \frac{b_{00}+(\boldsymbol{D}_\alpha\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)'\boldsymbol{T}_\alpha^{-1}(\boldsymbol{D}_\alpha\boldsymbol{\alpha}-\boldsymbol{\alpha}_0)}{2}\right),$$

$$\lambda_1|\boldsymbol{\beta} \sim \mathrm{Ga}\left(\frac{a_{10}+m_1}{2}, \frac{b_{10}+(\boldsymbol{D}_\beta\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{T}_\beta^{-1}(\boldsymbol{D}_\beta\boldsymbol{\beta}-\boldsymbol{\beta}_0)}{2}\right).$$

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

We also use the output of this MCMC simulation to calculate the marginal likelihood by the method of Chib (1995). Marginal likelihoods are used to optimize $\boldsymbol{p}$ and the number of knots.

2.5. Large Sample Analysis

Assume for simplicity that there are no control variables w. In this section, we conduct large sample analysis from a frequentist point of view, that is, we assume a true value of the parameters that generates the data. We use these notations and definitions: $g_0^*(\cdot)$, $g_1^*(\cdot)$, σ_{0*}^2, and σ_{1*}^2 denote the true values of $g_0(\cdot)$, $g_1(\cdot)$, σ_0^2, and σ_1^2, respectively. The true data distribution, conditional on $z^{(n)}$ is denoted by $P_*^{(n)}(y^{(n)}|z^{(n)})$, or $P_*^{(n)}$ for short, where $y^{(n)} \triangleq \{y_i, i=1,\ldots,n\}$ and $z^{(n)} \triangleq \{z_i, i=1,\ldots,n\}$ (and similarly for n replaced by n_0 and n_1). Denote by $\|\cdot\|_2$ the norm in $L_2(P_z)$ where P_z is the distribution of z, that is, $\forall f \in L_2(P_z)$, $\|f\|_2 \triangleq \left(\int |f(z)|^2 dP_z(z)\right)^{1/2}$ and by $\|\cdot\|_n$ the norm in $L_2(P_n)$ where P_n is the empirical distribution of z, that is, $\forall f \in L_2(P_n)$,

$$\|f\|_n^2 \triangleq \frac{1}{n}\sum_{i=1}^{n}|f(z_i)|^2.$$

Moreover, for a vector $\boldsymbol{v}$ we denote $\|\boldsymbol{v}\|_\infty \triangleq \max_i |v_i|$ and for $a,b \in \mathbb{R}$, $\mathcal{C}^\delta[a,b] \triangleq \{h:[a,b] \to \mathbb{R};\ h \text{ is } \delta \text{ times continuously differentiable}\}$. For a probability measure P, the notation $Ph = \int h dP$.

For $j=0,1$ denote by $\mathcal{C}_{m_j}$ the set of piecewise cubic functions and by $\mathcal{S}_{m_j} \subset \mathcal{C}_{m_j}$ the subspace of natural cubic splines on the set of m_j knots defined in Section 2.2. Therefore, $g_{m_0}(z) \triangleq \boldsymbol{B}_0(z)'\boldsymbol{\alpha} \in \mathcal{S}_{m_0} \subset \mathcal{C}_{m_0}$ (resp. $g_{m_1}(z) \triangleq \boldsymbol{B}_1(z)'\boldsymbol{\beta} \in \mathcal{S}_{m_1} \subset \mathcal{C}_{m_1}$) is the unique natural cubic spline interpolating g_0 (resp. g_1) at the knots

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

$\{z_{0,\,\mathsf{min}}, \kappa_{0,2}, \ldots, \kappa_{0,m_0-1}, \tau\}$ (resp. $\{\tau, \kappa_{1,2}, \ldots, \kappa_{1,m_1-1}, z_{1,\,\mathsf{max}}\}$), where $\boldsymbol{B}_j(z)$ is a m_j-vector defined in Appendix A. We denote by $\boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}^*$ the coefficients corresponding to the interpolations of g_0^* and g_1^*, respectively, which in turn are denoted by $g_{m_0}^*(z)$ and $g_{m_1}^*(z)$. Finally, π denotes both the prior on $(g_0, g_1, \sigma_0^2, \sigma_1^2, \lambda_0, \lambda_1)'$ that is specified in Section 2.3 and the corresponding posterior.

2.5.1. *Posterior Contraction Rate.* The goal of the asymptotic analysis is to establish the consistency of the ATE posterior distribution and to find the corresponding contraction rate. To derive these results, we first have to find the contraction rate for the posterior distribution of (g_j, σ_j^2) for $j = 0, 1$, which, given our model assumptions, is a new, interesting result in its own right. Our result shows that there exist two sequences $\epsilon_{n_0}, \epsilon_{n_1} \to 0$, such that for all $M_n, \widetilde{M}_n \to \infty$,

$$P_*^{(n)}\pi\left(\|g_j - g_j^*\|_{n_j} \geq M_n\epsilon_{n_j},\ \left|1 - \frac{\sigma_j}{\sigma_{j*}}\right| \geq \frac{\widetilde{M}_n}{\sqrt{n_j}} \text{ for } j = 0, 1 \,\middle|\, y^{(n)}, z^{(n)} \right) \to 0.$$

Since the posterior of $(g_0, g_1, \sigma_0^2, \sigma_1^2)$ is equal to the product of the posteriors of (g_0, σ_0^2) and of (g_1, σ_1^2), it is enough to show the separate convergence to zero of the following probability for $j = 0, 1$:

$$P_*^{(n_j)}\pi\left(\underbrace{\|g_j - g_j^*\|_{n_j} \geq M_n\epsilon_{n_j},\ |1 - \sigma_j^2/\sigma_{j*}^2| \geq \widetilde{M}_n/\sqrt{n_j}}_{\triangleq B_*^c(\epsilon_{n_j}, n^{-1/2})} \,\middle|\, y^{(n_j)}, z^{(n_j)} \right) \to 0$$

for all $M_n, \widetilde{M}_n \to \infty$. This means that the posterior distribution of (g_j, σ_j^2) concentrates on balls of radius of the order $(\epsilon_{n_j}, n_j^{-1/2})$ around the true (g_j^*, σ_{j*}^2). The sequence $(\epsilon_{n_j}, n_j^{-1/2})$ is a posterior contraction rate and, by definition, every sequence that tends to zero at a slower rate is also a contraction rate.

We make the next assumption to establish that (g_j^*, σ_{j*}^2) lies in the KL support of the prior distribution and that $\pi(\mathcal{G}_j \setminus \mathcal{C}_{n,j})$ is at most of the order $e^{-n_j\epsilon_{n_j}^2}$, where $\mathcal{G}_j$ and $\mathcal{C}_{n,j}$ are introduced below, as sketched in Section 2.5.2.

Assumption 4. (i) For every $m_0, m_1 \in \mathbb{N}_+$, there exist finite $w_0 \in \mathbb{R}^{m_0}$ and $w_1 \in \mathbb{R}^{m_1}$ such that $\boldsymbol{\alpha}^* - \boldsymbol{D}_\alpha^{-1}\boldsymbol{\alpha}_0 = \boldsymbol{D}_\alpha^{-1}\boldsymbol{T}_\alpha^{1/2}w_0$ and $\boldsymbol{\beta}^* - \boldsymbol{D}_\beta^{-1}\boldsymbol{\beta}_0 = \boldsymbol{D}_\beta^{-1}\boldsymbol{T}_\beta^{1/2}w_1$.
(ii) For every $m_0, m_1 \in \mathbb{N}_+$, the maximum eigenvalues of $\boldsymbol{D}_\alpha^{-1}\boldsymbol{T}_\alpha\boldsymbol{D}_\alpha^{-1'}$ and $\boldsymbol{D}_\beta^{-1}\boldsymbol{T}_\beta\boldsymbol{D}_\beta^{-1'}$ are bounded away from zero and infinity.

Assumption 4 (i) is quite standard in settings with Gaussian process priors and is a condition about the smoothness of the true function. It requires that the parameters $\boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}^*$ associated with the true functions belong to the ellipsoid associated with the prior covariance operator. Assumption 4 (ii) is weak and is satisfied for instance if the prior covariance matrix is a discretization of a bounded covariance operator.

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

We then have the following theorem. The notation $\asymp$ denotes equality up to a fixed constant.

THEOREM 2.1. *Assume that for some* $-\infty < a < b < \infty$ *there exists a* $\delta > 0$ *and a constant* $C_2 > 0$ *such that:* $g_0^* \in \mathcal{G}_0 \triangleq \mathcal{C}^\delta[a,\tau)$ *and* $g_1^* \in \mathcal{G}_1 \triangleq \mathcal{C}^\delta[\tau,b]$, *and* $\|g_j^* - g_{m_j}^*\|_\infty \leq C_2 m_j^{-\delta}$ *for* $j = 0,1$. *Moreover,* $0 < \underline{\sigma}_j^2 < \sigma_{j*}^2 < \overline{\sigma}_j^2 < \infty$, *for* $j = 0,1$ *and two constants* $\underline{\sigma}_j^2, \overline{\sigma}_j^2$. *Let* π *denote the prior on* $(g_0, g_1, \sigma^2, \{\lambda_j, j = 0,1\})'$ *that is specified in Section 2.3 with* $m_j = m_j^* \asymp \left(\frac{n_j}{\log(n_j)}\right)^{1/(2\delta+1)}$, $\nu_{00} > 2$, $\delta_{00} > 2$, *and* $a_{j0} = 2\eta m_j^*$ *for* $j = 0,1$, *and* $\eta > \min\{3, \delta_{00}\}(2\delta+1)$. *Moreover, suppose that Assumptions 1–4 hold. Then, for all* $M_n, \widetilde{M}_n \to \infty$ *the posterior of* $(g_0, g_1, \sigma_0^2, \sigma_1^2)$, *denoted by* $\pi(\cdot|y^{(n)}, z^{(n)})$ *satisfies the following:*

$$P_*^{(n)} \pi \Bigg((g_j, \sigma_j^2) \in \mathcal{G}_j \times \mathbb{R}_+ \textit{ for } j = 0,1; \|g_j - g_j^*\|_{n_j} \geq M_n \epsilon_{n_j}, \left|1 - \frac{\sigma_j}{\sigma_{j*}}\right| \geq \frac{\widetilde{M}_n}{\sqrt{n_j}} \Bigg| y^{(n)}, z^{(n)} \Bigg) \to 0, \tag{2.16}$$

where $\epsilon_{n_j} \asymp \left(\frac{\log(n_j)}{n_j}\right)^{\delta/(2\delta+1)}$.

We remark that if $n_j \asymp n$ for $j = 0,1$ then the two posterior distributions contract at the same rate $\left(\frac{\log(n)}{n}\right)^{\delta/(2\delta+1)}$. The smoothness assumption in this theorem is standard in the literature (see e.g., Calonico et al., 2014, Assumption 1; Imbens and Kalyanaraman, 2012, Assumption 3.3). The rate ϵ_{n_j} for the nonparametric parameter corresponds to the minimax rate, up to a logarithmic factor. The logarithmic factor in the convergence rate is typical in Bayesian nonparametric problems, see for example, (Ghosal, 2017). The constant M_n can be any diverging sequence and so it does not affect the rate of contraction. For the parametric component σ_j the usual parametric rate is obtained since the sequence $\widetilde{M}_n$ is allowed to grow indefinitely. Therefore, the joint posterior of (g_j, σ_j^2) contracts at the rate $n_j^{-1/2}$ along the σ_j^2-direction, but at a slower nonparametric rate ϵ_{n_j} along the g_j direction. We point out that our result (2.16) has been established in terms of the norm $\|\cdot\|_n$ but it holds also for the norm $\|\cdot\|_2$ as long as we add the assumption that the sets $\mathcal{C}^\delta[a,\tau)$ and $\mathcal{C}^\delta[\tau,b]$ are Glivenko–Cantelli with respect to P_n. In fact, in this case P_n converges to P_z uniformly over $g_0 \in \mathcal{C}^\delta[a,\tau)$ and $g_1 \in \mathcal{C}^\delta[\tau,b]$.

2.5.2. *Main Idea of the Proof of Theorem 2.1.* The general strategy to establish (2.16) in a semiparametric setting is the test and metric entropy approach. In our framework this means that, given an event $\mathcal{A}_j$ and its complement $\mathcal{A}_j^c$, the following upper bound holds: for $j = 0,1$

$$\pi\left(B^c_*(\epsilon_{n_j},n^{-1/2})\middle|y^{(n_j)},z^{(n_j)}\right)\leq\phi_{n_j}+(1-\phi_{n_j})\pi\left(B^c_*(\epsilon_{n_j},n^{-1/2})\middle|y^{(n_j)},z^{(n_j)}\right)$$
$$\leq\phi_{n_j}+I[\mathcal{A}_j^c]+(1-\phi_{n_j})\pi\left(B^c_*(\epsilon_{n_j},n^{-1/2})\middle|y^{(n_j)},z^{(n_j)}\right)I[\mathcal{A}_j],\qquad\textbf{(2.17)}$$

where ϕ_{n_j} is a testing function for the null hypothesis $H_0:(g_j,\sigma_j^2)=(g_j^*,\sigma_{j*}^2)$ against the complement of a neighborhood $\mathcal{U}$ of (g_j^*,σ_{j*}^2) denoted by $\mathcal{U}^c$, where the posterior probability of $\mathcal{U}$ converges to one. The event $\mathcal{A}_j$, which has probability converging to 1, is used to lower bound the denominator of the posterior distribution. This lower bound also depends on the KL property of $P_*^{(n_j)}$. That is, the true data distribution $P_*^{(n_j)}$ characterized by (g_j^*,σ_{j*}^2) has to belong to the KL support of the prior distribution: namely, for some constant $C>0$ the prior has to assign a probability not smaller than $e^{-Cn_j\epsilon_{n_j}^2}$ to any KL neighborhood of $P_*^{(n_j)}$ (see definition D.1 in the Online Supplementary Appendix). This requires lower bounding the prior of the KL neighborhood of $P_*^{(n_j)}$. While with a normal sampling distribution, and a known variance σ_j^2, it is easy to show that the KL neighborhood contains a set that admits an immediate characterization in terms of $\|g_j-g_j^*\|_{n_j}^2$, showing this for a student distribution and an unknown σ_j^2 is considerably more involved. We rely on upper bounds for the logarithmic function and a Taylor expansion which is valid over a set with prior probability of order at least equal to $e^{-n_j\epsilon_{n_j}^2}$.

The test ϕ_{n_j} formalizes the idea that (g_j^*,σ_{j*}^2) and $\mathcal{U}^c$ can be separated. It has to be a uniformly exponentially consistent test, that is, a test for which there exist constants $C>0$, $\varepsilon>0$ such that $P_*^{(n_j)}\phi_{n_j}\leq e^{-n_j\varepsilon}$ and $\inf_{(g_j,\sigma_j^2)\in\mathcal{U}^c}P^{(n_j)}_{g_j,\sigma_j^2}\phi_{n_j}\geq 1-Ce^{-n_j\varepsilon}$ where $P^{(n_j)}_{g_j,\sigma_j^2}\triangleq P^{(n_j)}_{g_j,\sigma_j^2}(y^{(n_j)}|z^{(n_j)},g_j,\sigma_j^2)$. Such tests exist for nonparametric regression models with normal errors and known variance. We develop such tests for our context. In the proof of Theorem 2.1, we exploit the fact that the student distribution can be written as a mixture of normal distributions and, conditional on the latent mixing variable, the true model is normal with unknown variance. Then, conditional on the latent mixing structure, we construct uniformly exponentially consistent tests for our set-up with an ε of the order $\epsilon_{n_j}^2$.

One way to guarantee the existence of such tests is to use the fact that the models in $\mathcal{U}^c$ are not too complex, where complexity is measured in terms of metric entropy and prior mass. More precisely, if $\mathcal{G}_j\times\mathbb{R}_+$ denotes the parameter space for (g_j,σ_j^2), for ε to be of the order ϵ_n^2 one needs to find a sequence of measurable sets $\mathcal{C}_{n,j}\subset\mathcal{G}_j$ such that: (i) as $n_j\to\infty$, $\mathcal{C}_{n,j}$ becomes big enough to approximate $\mathcal{G}_j$, (ii) the prior $\pi(\mathcal{G}_j\setminus\mathcal{C}_{n,j})$ is at most of the order $e^{-n_j\epsilon_{n_j}^2}$, and (iii) $\mathcal{C}_{n,j}$ can be written as the countable union of sets whose metric entropy is at most of the order $n_j\epsilon_{n_j}^2$.

The proof of Theorem 2.1 is constructed around these arguments, developed precisely in Appendix C.2 and in the Supplementary Appendix.

2.5.3. *Posterior Consistency of the ATE.* We now have the following result on the consistency of the ATE posterior distribution and its contraction rate.

THEOREM 2.2. *Let the assumptions of Theorem 2.1 hold. Then, for all* $M_{1,n} \to \infty$

$$P_*^{(n)}\pi\left(|\mathrm{ATE}-\mathrm{ATE}^*| \geq M_{1,n}\epsilon_n,\ |1-\sigma_j/\sigma_{j*}| \geq \widetilde{M}_n/\sqrt{n_j},\ \textit{for}\ j=0,1\,\middle|\,y^{(n)},z^{(n)}\right) \to 0,$$

where

$$\epsilon_n \asymp \max\left\{\left(\frac{\log(n_0)}{n_0}\right)^{\delta/(2\delta+1)}, \left(\frac{\log(n_1)}{n_1}\right)^{\delta/(2\delta+1)}\right\}.$$

Note that if $n_0 \asymp n_1$, then the contraction rate ϵ_n is the same as the rate in Calonico, Cattaneo, and Titiunik (2014) for $\delta = 3$, up to a logarithmic factor, which is usual in Bayesian nonparametric estimation. To establish this result, we have used the characterization of the ATE as $\boldsymbol{\beta}_{[1]} - \boldsymbol{\alpha}_{[m_0]}$ and then expressed $\boldsymbol{\beta}_{[1]}$ and $\boldsymbol{\alpha}_{[m_0]}$ as two linear functionals on the $L_2(P_{n_j})$ space for $j = 1,0$, respectively. This strategy is an alternative to the use of point evaluation functionals to express the ATE, as in Branson et al. (2019), which is unbounded in the $L_2(P_z)$ space because its representer does not belong to $L_2(P_z)$.

It is also possible to consider the ATE consistency under the assumption of fixed variances, as in Branson et al. (2019). Unless $\sigma_j^2 = \sigma_{j*}^2$, however, which Branson et al. (2019) apparently assume, the conditional model is misspecified and the posterior concentrates around the least-favorable model, see Bickel and Kleijn (2012). Therefore, to establish the ATE consistency (without assuming that the variances are equal to the true values), one has to consider the contraction of the conditional posterior of g_j, and of the ATE, conditioned on a sequence of variances, $\sigma_{jn}^2 \triangleq \sigma_{jn}^2(h_n) \triangleq \sigma_{j*}^2 + h_n/\sqrt{n_j}$ for all bounded, stochastic sequences h_n. By slightly modifying our proofs, we can show that the contraction rate of these conditional posteriors is the same as the rate given in Theorems 2.1 and 2.2. That is, under the assumptions of Theorems 2.1 and 2.2: for all $M_n, \widetilde{M}_n \to \infty$ and every bounded, stochastic h_n:

$$P_*^{(n)}\pi\left(g_j \in \mathcal{G}_j;\ \|g_j - g_j^*\|_{n_j} \geq M_n\epsilon_{n_j}\ \text{for}\ j=0,1\,\middle|\right.$$
$$\left.\sigma_j^2 = \sigma_{j*}^2 + n_j^{-1/2}h_n\ \text{for}\ j=0,1; y^{(n)}, z^{(n)}\right) \to 0$$

and for all $M_{1,n} \to \infty$

$$P_*^{(n)}\pi\left(|\mathrm{ATE}-\mathrm{ATE}^*| \geq M_{1,n}\epsilon_n\,\middle|\,\sigma_j^2 = \sigma_{j*}^2 + n_j^{-1/2}h_n\ \text{for}\ j=0,1; y^{(n)}, z^{(n)}\right) \to 0.$$

3. EXAMPLES: SHARP DESIGN

3.1. Design

We simulate data from the following data generating process: $y_j = g_j(z) + \sigma\varepsilon_j$ for $j = 0,1$, where $\varepsilon_0 \sim t_\nu(0,1)$ and $\varepsilon_1 \sim t_\nu(0,1)$ with $\nu = 3$, and the z variable and the

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

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 log marginal likelihood (computed by the method of Chib, 1995) and that the log marginal likelihood worsens as the degrees of freedom used in the estimation moves further away from the true value. The log marginal likelihood of the best model, for each sample size, is highlighted in bold.

	$\boldsymbol{p}$	$\boldsymbol{m}_z$	$\boldsymbol{m}_{z,\tau}$	log marg lik
		$n = 500$		
t_3	$(0.7, 0.3)$	$(3,3)$	$(3,2)$	**55.61**
t_3	$(0.9, 0.1)$	$(3,3)$	$(3,2)$	52.48
t_4	$(0.7, 0.3)$	$(3,3)$	$(3,2)$	55.20
t_5	$(0.7, 0.3)$	$(3,3)$	$(3,2)$	53.12
t_{100}	$(0.7, 0.3)$	$(3,3)$	$(3,2)$	8.99
		$n = 4,000$		
t_3	$(0.7, 0.3)$	$(4,3)$	$(3,2)$	1,026.47
t_3	$(0.9, 0.1)$	$(4,3)$	$(3,2)$	**1,027.71**
t_4	$(0.9, 0.1)$	$(4,3)$	$(3,2)$	1,016.16
t_5	$(0.9, 0.1)$	$(4,3)$	$(3,2)$	992.84
t_{100}	$(0.9, 0.1)$	$(4,3)$	$(3,2)$	568.95

parameters are chosen as in Imbens and Kalyanaraman (2012) (IK) and Calonico, Cattaneo, and Titiunik (2014) (CCT), that is,

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

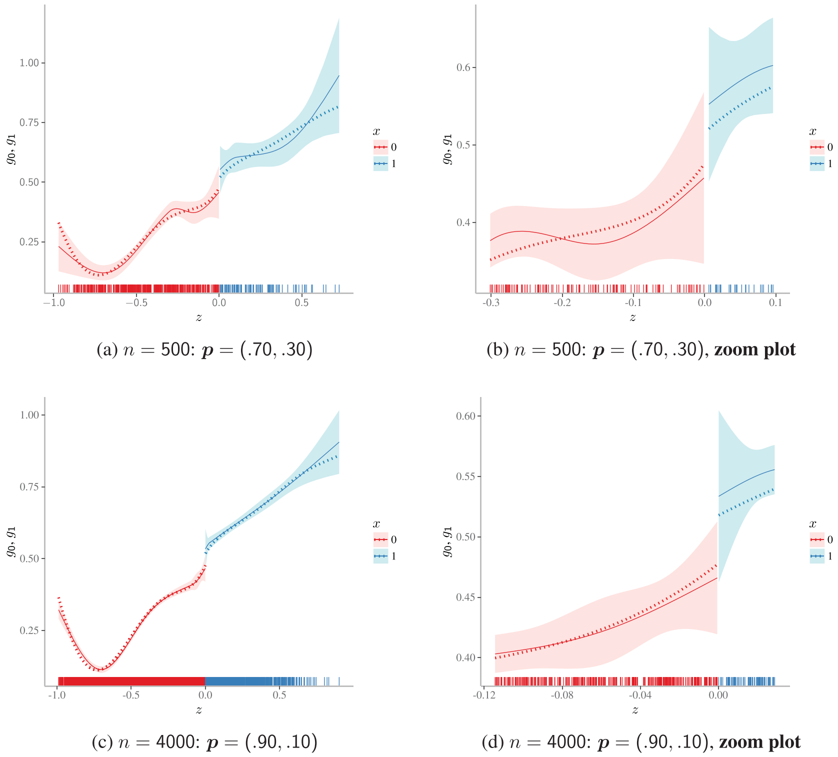
$\sigma = 0.1295$. The true value of the RD ATE at the break-point $\tau = 0$ is 0.04.

In this design, there are relatively fewer treated observations than controls, which makes the estimation of the ATE challenging in small samples. We consider two sample sizes, $n = 500$ and $n = 4,000$.

We estimate the model as follows:

1. The prior mean and SD of σ^2 are 0.3 and 1.0, respectively.
2. The prior mean and SD of $\boldsymbol{\lambda}$ are $(1,1)$ and $(5,5)$, our default choices.
3. The soft-window quantiles $\boldsymbol{p} \triangleq (p_0, p_1)$ are based on the distribution of z. We optimize these values according to the value of the marginal likelihoods, as shown in Table 1.
4. The number of knots is also chosen based on marginal likelihoods. In general, as n increases, we shrink the soft-window width and increase the number of knots.
5. Finally, the degrees of freedom parameter is set on the grids 3, 4, 5, and 100.

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press



(a) $n = 500$: $\boldsymbol{p} = (.70, .30)$

(b) $n = 500$: $\boldsymbol{p} = (.70, .30)$, **zoom plot**

(c) $n = 4000$: $\boldsymbol{p} = (.90, .10)$

(d) $n = 4000$: $\boldsymbol{p} = (.90, .10)$, **zoom plot**

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 for the best models selected in Table 1. The right panel are the corresponding zoom plots, zoomed to the interval given by the soft-window quantiles.

The results in Table 1 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 $n = 4,000$ there is more support for a narrower soft-window.

3.1.1. *Function Estimates.* The Bayes estimates of g_0 and g_1 are given in Figure 3. 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 $n = 500$, g_1 is not well estimated (because of the sparseness of the data) but the function estimate improves with n. The right panel displays the same function estimates, but restricted to the soft-window intervals.

3.2. Sampling Experiments

We now consider the sampling performance of the Bayes RD ATE estimate along two dimensions—the sampling root mean square error (RMSE) and the coverage

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

Table 2. Simulated Data: Sharp RD Designs, True Value of the ATE is 0.04. Summary of Results from 1,000 Repeated Samples of Student-t data generating process (DGP) and Two Different Sample Sizes.

Mean	Coverage	RMSE
	$n = 500$	
0.043	0.961	0.073
	$n = 4,000$	
0.040	0.945	0.045

of the 95% posterior credibility interval. In these experiments, $\boldsymbol{p}$, $\boldsymbol{m}_z$ and $\boldsymbol{m}_{z,\tau}$ are determined once. We have found, however, that the final sampling results are not improved greatly by optimizing these choices for every new sample.

Finally, as in the preceding section, the prior mean and standard deviation of σ^2 are 0.3 and 1.0, respectively, and the prior mean and standard deviation of $\boldsymbol{\lambda}$ 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.

The results show that, as the sample size increases, the Bayes estimate gets closer to the true value, the coverage approaches the nominal value, and the RMSE declines.

3.3. Example: Meyersson (2014)

For an application of the sharp RD design to real data, we consider the study of Meyersson (2014), used as an example by Cattaneo, Idrobo, and Titiunik (2020) to discuss various frequentist procedures. The study (based on data from Turkey) deals with the effect of an Islamic mayor on the percentage of women aged 15–20 in 2000 who had completed high school in 2000 (this is the outcome variable y). The running variable z is the vote margin obtained by the Islamic party in the 1994 mayoral election over its strongest secular party opponent (the cut-point τ is zero). The treatment variables x is 1 if the Islamic party candidate won the mayoral election, and 0 otherwise. Four regional indicator variables, which we denote by $\boldsymbol{v}$, and three continuous variables, w_1, w_2, w_3, representing the Islamic vote percentage in 1994, the number of parties receiving votes in 1994, and the logarithm of the population in 1994, respectively, are available as controls.

Cattaneo, Idrobo, and Titiunik (2020) calculate the sharp RD effect by different local polynomial based methods without and with covariate adjustment and under various procedural choices involving the kernel and bandwidth parameters. These different methods and choices produce estimates that are similar, but it is not clear from the frequentist analysis if covariate adjustment should be done, and which procedural choices are preferred. It is also possible that some covariates are more important for some bandwidths, and not for others, but these sort of comparisons

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

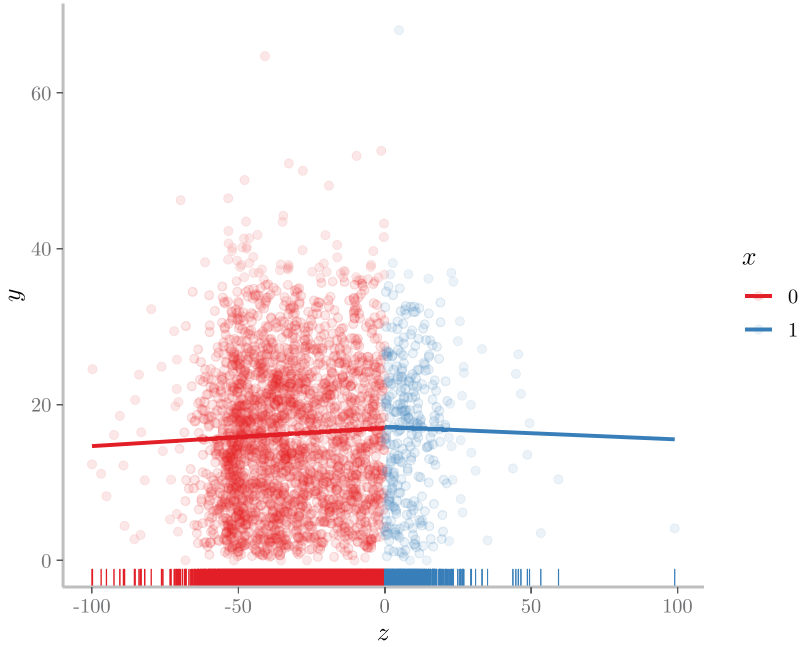


Figure 4. Meyersonn (2014) data: scatter plot of y against x with least squares fits from data on each side of τ.

are not possible by hard-windowing methods where different bandwidths produce different data sets.

As we have demonstrated above, soft-windowing plus marginal likelihoods enable a systematic comparison (and ranking) of the different modeling choices. One primary choice is about $\boldsymbol{p}$, $\boldsymbol{m}_z$, and $\boldsymbol{m}_{z,\tau}$. A scatter-plot of y against z, with least squares fits on each side of τ, given in Figure 4, shows (noisy) observations concentrated close to τ, right-skewness in the distribution of y with some rather outlying observations, and a small positive jump at τ, at least as measured by the (inappropriate) global linear fits. These observations suggest that covariate adjustment will be required to reduce the noise, that the student-t distribution with its thicker tails should outperform the Gaussian (both assumptions should be viewed as approximations given that the support of y is limited in these data), and that a few knots should be adequate to model g_j.

Hence, we consider Gaussian (approximated by $\nu = 100$) and student-t ($\nu = 5$) models, with and without covariate adjustment. A range of soft-window parameters are scanned, starting with $\boldsymbol{p} = (0.7, 0.4)$. In these models, $\boldsymbol{m}_z = (2,2)$, and $\boldsymbol{m}_{z,\tau} = (3,2)$, as other larger number of knots have substantially lower marginal likelihoods, regardless of $\boldsymbol{p}$. In each of these models, the prior on σ_j^2, has a mean

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

and sd of 70 and 30, respectively, and the prior of each element of $\boldsymbol{\lambda}$ has a mean of 1 and sd of 5.

Consider first the case of models without covariate adjustment. The best model (selected by the marginal likelihood) has $\boldsymbol{p} = (0.4, 0.3)$ and a log-marginal likelihood of **−9669.249**. It shows a preference for a wider soft-window to the left of τ, and a narrower soft-window to the right of τ. From this best model, the posterior mean of the ATE is 3.213, with a 95% posterior credibility interval of (−0.398, 6.864).

A similar analysis can be conducted with covariates. One can consider all possible adjustments, by taking different combinations of the four geographical indicators and three continuous covariates. As an illustration, however, and to draw a contrast with the preceding no-covariate adjusted analysis, we only consider models in which all seven covariates are included, that is, where

$$y_j = g_j(z) + \boldsymbol{v}'\boldsymbol{\delta}_v + \sum_{l=1}^{3} h_l(w_j) + \sigma_j \varepsilon_j.$$

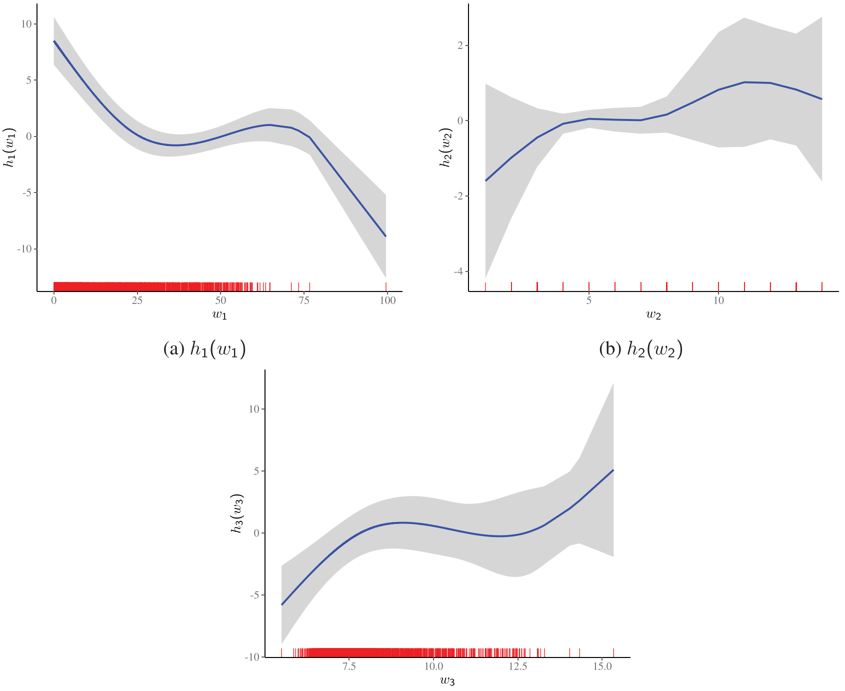
Thus, the three continuous covariates are entered additively and nonparametrically. Each is estimated as natural cubic splines with five knots. Again, we consider the same soft-window parameters and Gaussian and student-t distributions as in the no-covariate model scan. The best model now is the student-t model with $\nu = 5$ and $\boldsymbol{p} = (0.60, 0.50)$ indicating support for different soft-window dimensions. The substantially larger log-marginal likelihood of **−9330.464** shows decisive support for covariate-adjustment. Within the covariate-adjusted models, the support for the student-t over the Gaussian error distribution is also decisive. The nonparametric estimates of the three covariate functions, given in Figure 5, show the importance of entering (w_1, w_2, w_3) nonparametrically.

From this best model, the posterior mean of the ATE is 3.254, with a 95% credibility interval of (0.430, 6.049), with a positive lower-limit, and a narrower interval than before. Cattaneo et al. (2020, pg 73) report similar estimates: an ATE estimate of 3.108 and a 95% confidence interval of (0.194, 6.132). A graphical summary of the posterior means of the g_0 and g_1 functions and of the ATE, is given in Figure 6.

4. FUZZY RD DESIGN

We now formalize our Bayesian approach for the fuzzy RD design, inspired by the principal stratification framework of Frangakis and Rubin (2002) and Chib and Jacobi (2008). Our essential idea is to capture the mismatch between the assignment process $I[z \geq \tau]$ 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.

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press



(a) $h_1(w_1)$

(b) $h_2(w_2)$

(c) $h_3(w_3)$

Figure 5. Meyersson (2014) data: Posterior mean and 95% credibility bands of nonparametric covariate functions.

4.1. Assumptions

Assumption 5 (Local conditional independence)**.** x is independent of y_0, y_1 conditional on z around τ and w, that is,

$$x \perp\!\!\!\perp (y_0, y_1) | z \text{ around } \tau, w.$$

Note that this assumption is trivially satisfied in the sharp RD design. Also note that this assumption does not restrict the dependence between x and (y_0, y_1) for values of z far from τ.

Assumption 6. The unobserved confounder s is an unobserved 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 on the treatment intake x:

$$\begin{aligned} x = I[z \geq \tau] &\quad \text{if} \quad s = c, \\ x = 0 &\quad \text{if} \quad s = n, \\ x = 1 &\quad \text{if} \quad s = a. \end{aligned}$$

Our next assumption is about the distribution of these types.

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

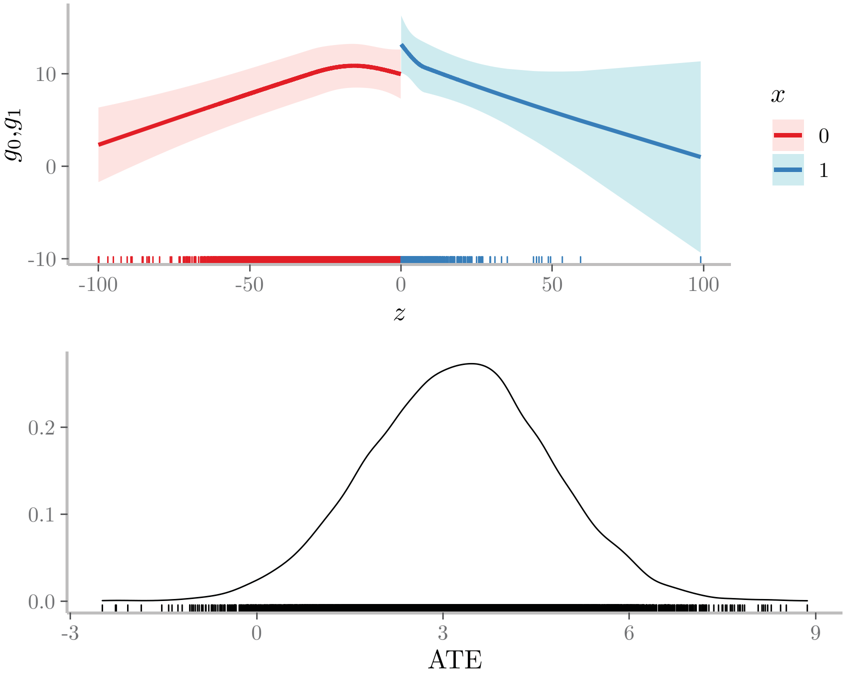


Figure 6. Meyersson (2014) data (sharp RD covariate-adjusted): posterior means of the g_0 and g_1 functions (along with the 95% point-wise credibility bands) and the posterior distribution of the ATE from $m = 10,000$ MCMC draws from the best fitting model in which there are four geographical indicators and three continuous covariates that are entered nonparameterically and the error distribution is student-t with 5 degrees of freedom. The posterior mean of the ATE is 3.254, with a 95% credibility interval of (0.430, 6.049).

Assumption 7. Subject types are distributed around τ with unknown distribution $\mathsf{Pr}(s = k) = q_k > 0$, where $q_c + q_n + q_a = 1$.

The model for the subject type probability in Assumption 6 encapsulates the assumption that the distribution of types around τ is independent of z. For simplicity, we assume that these type probabilities are also free of w.

Note from the first row of Assumption 6 that for subjects of type $s = c$, the compliers, assignment, and intake agree, that is, as z passes the break-point τ, the treatment state changes from 0 to 1 with probability one:

$$\mathsf{Pr}(x = 0|z < \tau, w, s = c) = 1 \text{ and } \mathsf{Pr}(x = 1|z \geq \tau, w, s = c) = 1. \tag{4.1}$$

It follows that, for compliers, the sharp design holds. On the other hand, for subjects of the type $s = n$, the never-takers, $\mathsf{Pr}(x = 0|z, w, s = n) = 1$, and for subjects with $s = a$, the always-takers, $\mathsf{Pr}(x = 1|z, w, s = a) = 1$, both regardless of the value of (z, w).

Table 3. Sample Data in the Fuzzy RD Case.

	$x = 0$	$x = 1$
$z < \tau$	$\mathbf{y}_{00}, \mathbf{z}_{00}$	$\mathbf{y}_{01}, \mathbf{z}_{01}$
$z \geq \tau$	$\mathbf{y}_{10}, \mathbf{z}_{10}$	$\mathbf{y}_{11}, \mathbf{z}_{11}$

Four potential outcomes emerge: 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.

Assumption 8. Conditioned on (z, w) and s, Assumptions 1–4 hold for $s = c$. In addition, the following conditional expectations hold

$$E[y_{0n}|z, w, s = n] = g_{0n}(z) + h_n(w),$$

and

$$E[y_{1a}|z, w, s = a] = g_{1a}(z) + h_a(w)$$

both over the entire support of z, where the function g_{0n} and g_{1a} are δ-times continuously differentiable on an interval that contains τ with $\delta > 0$, and $h_n(w)$ and $h_a(w)$ are continuous in w. Furthermore, these two potential outcome are independently distributed as standard student-t with $\nu > 2$ degrees of freedom.

4.2. Sample Data

Suppose that the data consist of n independent copies of (y, x, z), where for simplicity, we assume that w is absent from the model. Because observations on either side of τ can be controls or treated, we place the sample data into four cells, cross-classified by $I[z \geq \tau] = l$, $l = 0, 1$ and $x = j$, $j = 0, 1$. We indicate each of these cells by (lj). The observations in each of these cells are indicated in vector notation and displayed in Table 3.

Indices of the observations in each cell are denoted by $I_{00} = \{i : z_i < \tau, x_i = 0\}$, $I_{10} = \{i : z_i \geq \tau, x_i = 0\}$, $I_{01} = \{i : z_i < \tau, x_i = 1\}$ and $I_{11} = \{i : z_i \geq \tau, x_i = 1\}$, and 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, $\mathbf{z}_0 = (\mathbf{z}_{00}, \mathbf{z}_{10})$ and $\mathbf{z}_1 = (\mathbf{z}_{01}, \mathbf{z}_{11})$.

4.3. Possible Types Cross-Classified by *z* and *x*

In the manner of the preceding data table, the possible subject types are shown in Table 4. Specifically, an individual in cell (00) can be either a complier or never-taker, a person in cell (10) is of type never-taker, a subject in cell (01) is of type always-taker, while a person in cell (11) can be either a complier or an always-taker.

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

TABLE 4. Possible Subject Types on Either Side of τ by Treatment State.

	$x=0$	$x=1$
$z<\tau$	c,n	a
$z\geq\tau$	n	c,a

This division of types by cell is key to understanding the subsequent inferential procedure. It also clarifies that this model is a mixture model. For instance, consider the case where $z_i<\tau$ (the first row of this table). Then,

$$\begin{aligned} p(y_i,x_i=j|z_i<\tau,\theta) &= \sum_{k\in\{c,n,a\}} p(y_i,x_i=j|z_i<\tau,\theta,s_i=k)\,\mathrm{Pr}(s_i=k|\theta) \\ &= \sum_{k\in\{c,n,a\}} p(y_i|x_i=j,z_i<\tau,\theta,s_i=k)\,\mathrm{Pr}(x_i=j|z_i<\tau,\theta,s_i=k)\,\mathrm{Pr}(s_i=k|\theta), \end{aligned} \tag{4.2}$$

which for $j=0$, cell (00), reduces to $t_\nu(y_i|g_0(z_i),\sigma_0^2)q_c+t_\nu(y_i|g_{0n}(z_i),\sigma_{0n}^2)q_n$ and for $j=1$, cell (01), reduces to $t_\nu(y_i|g_{1a}(z_i),\sigma_{1a}^2)q_a$ (since the middle term is either one or zero).

4.4. Identification of the RD CATE

Under our assumptions, the RD CATE, the ATE for compliers at τ:

$$\begin{aligned} \text{CATE} &= \lim_{z\downarrow\tau^+}\mathbf{E}[y_1|z,w,s=c]-\lim_{z\uparrow\tau^-}\mathbf{E}[y_0|z,w,s=c] \\ &= g_1(\tau)-g_0(\tau). \end{aligned} \tag{4.3}$$

is identified. The proof involves showing that there is a function of the data that gives the CATE.

THEOREM 4.1. *Suppose Assumptions 1–3 and 5–8 hold. Also suppose that $g_0(\tau)\neq g_{0n}(\tau)$ and $g_1(\tau)\neq g_{1a}(\tau)$. Then, the CATE is identified.*

First consider $\lim_{z\downarrow\tau^+}\mathbf{E}[y|z,w]$. By extending the argument along the second row of Table 4,

$$\begin{aligned} \lim_{z\downarrow\tau^+}\mathbf{E}[y|z,w] = &\lim_{z\downarrow\tau^+}\mathbf{E}[y|z,w,s=c,x=1]\,\mathrm{Pr}(x=1|z,w,s=c)\,\mathrm{Pr}(s=c) \\ &+\lim_{z\downarrow\tau^+}\mathbf{E}[y|z,w,s=n,x=0]\,\mathrm{Pr}(x=0|z,w,s=n)\,\mathrm{Pr}(s=n) \\ &+\lim_{z\downarrow\tau^+}\mathbf{E}[y|z,w,s=a,x=1]\,\mathrm{Pr}(x=1|z,w,s=a)\,\mathrm{Pr}(s=a), \end{aligned}$$

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

which is

$$\begin{aligned}\lim_{z\downarrow\tau^+} \mathbf{E}[y|z,w] = {} & \lim_{z\downarrow\tau^+} \mathbf{E}[y_{1c}|z,w,s=c]\,\mathrm{Pr}(s=c)\\ & + \lim_{z\downarrow\tau^+} \mathbf{E}[y_{0n}|z,w,s=n]\,\mathrm{Pr}(s=n)\\ & + \lim_{z\downarrow\tau^+} \mathbf{E}[y_{1a}|z,w,s=a]\,\mathrm{Pr}(s=a),\end{aligned}$$

since to the right of τ each of the conditional probabilities of x is equal to one. Similarly, $\lim_{z\uparrow\tau^-} \mathbf{E}[y|z,w]$ is equal to

$$\begin{aligned}\lim_{z\uparrow\tau^-} \mathbf{E}[y|z,w] = {} & \lim_{z\uparrow\tau^-} \mathbf{E}[y_{0c}|z,w,s=c]\,\mathrm{Pr}(s=c)\\ & + \lim_{z\uparrow\tau^-} \mathbf{E}[y_{0n}|z,w,s=n]\,\mathrm{Pr}(s=n)\\ & + \lim_{z\uparrow\tau^-} \mathbf{E}[y_{1a}|z,w,s=a]\,\mathrm{Pr}(s=a).\end{aligned}$$

By the assumed continuity, the second and third terms in each of these expressions are equal. Therefore,

$$\begin{aligned}&\lim_{z\downarrow\tau^+} \mathbf{E}[y|z,w] - \lim_{z\uparrow\tau^-} \mathbf{E}[y|z,w]\\ &= \left(\lim_{z\downarrow\tau^+} \mathbf{E}[y_{1c}|z,w,s=c] - \lim_{z\uparrow\tau^-} \mathbf{E}[y_{0c}|z,w,s=c]\right)\mathrm{Pr}(s=c).\end{aligned}$$

Now consider

$$\lim_{z\downarrow\tau^+} \mathbf{E}[x|z,w] = \lim_{z\downarrow\tau^+} \mathrm{Pr}(x=1|z,w),$$

which by extending the argument along the second row of Table 4 is

$$\begin{aligned}\lim_{z\downarrow\tau^+} \mathrm{Pr}(x=1|z,w) = {} & \lim_{z\downarrow\tau^+} \mathrm{Pr}(x=1|z,w,s=c)\,\mathrm{Pr}(s=c)\\ & + \lim_{z\downarrow\tau^+} \mathrm{Pr}(x=1|z,w,s=a)\,\mathrm{Pr}(s=a)\\ & + \lim_{z\downarrow\tau^+} \mathrm{Pr}(x=1|z,w,s=n)\,\mathrm{Pr}(s=n),\end{aligned}$$

where the first and second conditional probabilities are one and the third is zero. Thus,

$$\lim_{z\downarrow\tau^+} \mathrm{Pr}(x=1|z,w) = \mathrm{Pr}(s=c) + \mathrm{Pr}(s=a).$$

By similar calculations,

$$\begin{aligned}\lim_{z\uparrow\tau^-} \mathbf{E}[x|z,w] &= \lim_{z\uparrow\tau^-} \mathrm{Pr}[x=1|z,w]\\ &= \mathrm{Pr}(s=a).\end{aligned}$$

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

Hence,

$$\frac{\mathsf{lim}_{z\downarrow\tau^+}\ \mathbf{E}[y|z,w]-\mathsf{lim}_{z\uparrow\tau^-}\ \mathbf{E}[y|z,w]}{\mathsf{lim}_{z\downarrow\tau^+}\ \mathbf{E}(x|z,w)-\mathsf{lim}_{z\uparrow\tau^-}\ \mathbf{E}[x|z,w]}=g_1(\tau)-g_0(\tau),$$

since all the functions involving w cancel, as well $\mathsf{Pr}(s=c)$ in the numerator and denominator. Thus, the CATE is identified.

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 3. For instance, for the g_0 function, we use the data z_{00}, padded with τ at the right, to locate the desired number of knots according to the soft-windowing method. These knots are given by $\left\{z_{00,\,\mathsf{min}},\kappa_{0,2},\ldots,\kappa_{0,m_0-1},\tau\right\}$. For the g_1 function, the knots are calculated from the data z_{11}, padded with τ at the left. These knots are given by $\left\{\tau,\kappa_{1,2},\ldots,\kappa_{1,m_1-1},z_{11,\,\mathsf{max}}\right\}$. Then, the function ordinates

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

are approximated using the natural cubic spline basis functions given in the Appendix as

$$\begin{aligned}&g_0(z_{00})\approx g_{m_0}(z_{00})\triangleq\boldsymbol{B}_{00}\boldsymbol{\alpha},\\&g_1(z_{11})\approx g_{m_1}(z_{11})\triangleq\boldsymbol{B}_{11}\boldsymbol{\beta},\end{aligned}\tag{4.4}$$

respectively, where $\boldsymbol{B}_{jj}:n_{jj}\times m_j$ are the basis matrices evaluated at z_{jj}, and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are the basis coefficients. The notation $\boldsymbol{B}_{jj}$ 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,

$$\underset{(m_0\times 1)}{\boldsymbol{\alpha}}=\begin{pmatrix}g_0(z_{00,\,\mathsf{min}})\\g_0(\kappa_{0,2})\\\vdots\\g_0(\kappa_{0,m_0-1})\\g_0(\tau)\end{pmatrix},\quad\underset{(m_1\times 1)}{\boldsymbol{\beta}}=\begin{pmatrix}g_1(\tau)\\g_1(\kappa_{1,2})\\\vdots\\g_1(\kappa_{1,m_1-1})\\g_1(z_{11,\,\mathsf{max}})\end{pmatrix},\tag{4.5}$$

which implies that the CATE is simply the first component of $\boldsymbol{\beta}$ minus the last component of $\boldsymbol{\alpha}$:

$$\text{CATE}=\boldsymbol{\beta}_{[1]}-\boldsymbol{\alpha}_{[m_0]}.\tag{4.6}$$

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 τ), $z_j\triangleq(z_{0j}',z_{1j}')'$, $(n_j\times 1)$, where $n_j=n_{0j}+n_{1j}$, for $j=0,1$. 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 (τ,z_{1j}), making sure that τ is one of the knots and that every pair of knots has at least one observation in between. Locating knots in this way is relatively

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

straightforward. Note that by placing a knot at τ, we ensure that the functions g_{0n} and g_{1a} at τ are continuous, which is required by our assumptions.

Suppose then that m_n and m_a knots are used in the basis expansions of the g_{0n} and g_{1a}, respectively. The function ordinates at these knots

$$g_{0n}(z_0) \triangleq \left(g_{0n}(z_{00,1}),\ldots,g_{0c}(z_{10,n_{10}})\right),$$

and

$$g_{1a}(z_1) \triangleq \left(g_{1a}(z_{01,1}),\ldots,g_{1a}(z_{11,n_{11}})\right)$$

can then be approximated by the basis functions given in the Appendix as

$$g_{0n}(z_0) \approx g_{m_n}(z_0) \triangleq \begin{pmatrix} \underset{(n_{00}\times m_n)}{\boldsymbol{B}_{00,n}} \\ \underset{(n_{10}\times m_n)}{\boldsymbol{B}_{10,n}} \end{pmatrix} \boldsymbol{\alpha}_n \triangleq \boldsymbol{B}_{0,n}\boldsymbol{\alpha}_n, \tag{4.7}$$

and

$$g_{1a}(z_1) \approx g_{m_a}(z_1) \triangleq \begin{pmatrix} \underset{(n_{01}\times m_a)}{\boldsymbol{B}_{01,a}} \\ \underset{(n_{11}\times m_a)}{\boldsymbol{B}_{11,a}} \end{pmatrix} \boldsymbol{\beta}_a \triangleq \boldsymbol{B}_{1,a}\boldsymbol{\beta}_a, \tag{4.8}$$

respectively, where $\boldsymbol{\alpha}_n : m_n \times 1$ and $\boldsymbol{\beta}_a : m_a \times 1$ are the basis coefficients.

As for g_0 and g_1, for every z_j, $j = 0,1$, the goodness of the approximation of g_{0n} and g_{1a} at z_j depends on the smoothness of g_{0n} and g_{1a}, respectively, and the approximation bias decreases as m_n and m_a increase. We show in Section 4.9 that, for posterior consistency, m_j must increase at the rate of $(n/\log n)^{\nu}$, for some constant ν dependent on the smoothness of the function $g_j, j = 0,1,n,a$.

4.6. Likelihood Function

The basis function approximations of g_0, g_1, g_{0n}, and g_{1a} given in Section 4.5 suggest to construct a prior that puts all its mass on the approximations g_{m_j}, for $j = 0,1,n,a$, as we construct in the next section. This corresponds to specify a Dirac probability at zero for the remaining part of g_j, that is, $g_j - g_{m_j}$. The marginal likelihood function of $\boldsymbol{\theta} \triangleq (\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\alpha}_n,\boldsymbol{\beta}_a)$ and $\boldsymbol{\sigma}^2 = \left(\sigma_0^2,\sigma_1^2,\sigma_n^2,\sigma_a^2\right)$ is obtained by integrating with respect to this Dirac probability and follows straightforwardly from Theorem 1. Let $\boldsymbol{B}_{00,i}$ denote the ith row of $\boldsymbol{B}_{00}$, with similar notation for the other basis matrices. Then, the likelihood contribution of the ith observation by cell is

$$\begin{aligned}
L_{00,i} &= q_c t_\nu(y_i|\boldsymbol{B}_{00,i}\boldsymbol{\alpha},\sigma_0^2) + q_n t_\nu(y_i|\boldsymbol{B}_{00,n,i}\boldsymbol{\alpha}_n,\sigma_n^2),\ i \in I_{00},\\
L_{10,i} &= q_n t_\nu(y_i|\boldsymbol{B}_{00,n,i}\boldsymbol{\alpha}_n,\sigma_n^2),\ i \in I_{10},\\
L_{01,i} &= q_a t_\nu(y_i|\boldsymbol{B}_{01,a,i}\boldsymbol{\beta}_a,\sigma_a^2),\ i \in I_{01},\\
L_{11,i} &= q_c t_\nu(y_i|\boldsymbol{B}_{11,i}\boldsymbol{\beta},\sigma_1^2) + q_a t_\nu(y_i|\boldsymbol{B}_{11,a,i}\boldsymbol{\beta}_a,\sigma_a^2),\ i \in I_{11},
\end{aligned} \tag{4.9}$$

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

$$L = \prod_{i \in I_{00}} L_{00,i} \times \prod_{i \in I_{10}} L_{10,i} \times \prod_{i \in I_{01}} L_{01,i} \times \prod_{i \in I_{11}} L_{11,i}. \tag{4.10}$$

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 $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is exactly the same as in (2.7) and (2.10) except that the matrices $\boldsymbol{D}_\alpha$ and $\boldsymbol{D}_\beta$, 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 $\boldsymbol{\alpha}_n$ and $\boldsymbol{\beta}_a$ of the n and a models, is given by

$$\boldsymbol{\alpha}_n | \lambda_n \sim \mathcal{N}_{m_n}\left(\boldsymbol{D}_n^{-1}\boldsymbol{\alpha}_{0n}, \lambda_n^{-1}\boldsymbol{D}_n^{-1}\boldsymbol{T}_n\boldsymbol{D}_n^{-1\prime}\right),$$
$$\boldsymbol{\beta}_a | \lambda_a \sim \mathcal{N}_{m_a}\left(\boldsymbol{D}_a^{-1}\boldsymbol{\beta}_{0a}, \lambda_a^{-1}\boldsymbol{D}_a^{-1}\boldsymbol{T}_a\boldsymbol{D}_a^{-1\prime}\right),$$

where the matrices $\boldsymbol{D}_n$ and $\boldsymbol{D}_a$ are constructed analogously to $\boldsymbol{D}_\alpha$ and $\boldsymbol{D}_\beta$. The penalty parameters, now given by $(\lambda_0, \lambda_1, \lambda_n, \lambda_a)$, have a prior distribution as in the sharp model: $\lambda_j \sim \mathrm{Ga}\left(\frac{a_{j0}}{2}, \frac{b_{j0}}{2}\right)$, $j = 0, 1, n, a$. As in the sharp case, to obtain frequentist asymptotic results we require that $a_{j0} = C(n/\log n)^{\nu}$, for positive constants C and ν that depend on the smoothness of the function g_j, for $j = 0, 1, 0n, 1a$. Similarly as in the sharp case, the prior on the variance parameters is independent of λ and inverse Gamma: $\sigma_j^2 \sim \mathrm{IG}\left(\frac{\nu_{00}}{2}, \frac{\delta_{00}}{2}\right)$, for $j = 0, 1$, and for $j = n, a$: $\sigma_j^2 \sim \mathrm{IG}\left(\frac{\nu_{0j}}{2}, \frac{\delta_{0j}}{2}\right)$. In the examples below, for simplicity, we specify a Gamma distribution for each λ with a prior mean of 1 and prior SD of 10. Finally, we suppose that the prior of $\boldsymbol{q} = (q_c, q_n, q_a)$ is Dirichlet with parameters (n_{0c}, n_{0n}, n_{0a}), where we set the hyperparameters to imply that half the sample consists of compliers and the remaining half is equally divided between never-takers and always-takers.

4.8. MCMC Sampling

We estimate the model by sampling the posterior distribution of the parameters, the type variables $s_i (i \leq n)$, and the Gamma mixing variables, by MCMC methods. Note that conditioned on the parameters, the type variables have to be sampled only in the (00) and (11) cells. Specifically, for observations in the set I_{00}, conditioned on the data and $\left(\boldsymbol{\theta}, \boldsymbol{\sigma}^2\right)$, the type variables are sampled from the conditional distributions

$$\Pr(s_i = c | y_i, \boldsymbol{\theta}, \boldsymbol{\sigma}^2) \propto q_c t_\nu(y_i | \boldsymbol{B}_{00,i}\boldsymbol{\alpha}, \sigma_0^2),$$
$$\Pr(s_i = n | y_i, \boldsymbol{\theta}, \boldsymbol{\sigma}^2) \propto q_n t_\nu(y_i | \boldsymbol{B}_{00,n,i}\boldsymbol{\alpha}_n, \sigma_n^2),$$

and for observations in the set I_{11} from

$$\Pr(s_i = c|y_i, \boldsymbol{\theta}, \boldsymbol{\sigma}^2) \propto q_c t_\nu(y_i|\boldsymbol{B}_{11,i}\boldsymbol{\beta}, \sigma_1^2),$$
$$\Pr(s_i = a|y_i, \boldsymbol{\theta}, \boldsymbol{\sigma}^2) \propto q_a t_\nu(y_i|\boldsymbol{B}_{11,a,i}\boldsymbol{\beta}_a, \sigma_a^2).$$

Suppose that in a particular MCMC iteration, in the cell (00), the sampling of $\{s_i\}$ 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 $n_{11}^a = n_{11} - n_{11}^c$ always-takers. Then, in the next MCMC step, we sample $\boldsymbol{q} = (q_c, q_n, q_a)$ 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 posterior distribution decomposes into three independent distributions, one for each type. These distributions take the following form. For the observations classified as $s = c$, using the basis matrices in (4.4), we can write

$$\underset{(n_{00}^c \times 1)}{\boldsymbol{y}_{00}^c} = \underset{(n_{00}^c \times m_0)}{\boldsymbol{B}_{00}^c \boldsymbol{\alpha}} + \underset{(n_{00}^c \times 1)}{\boldsymbol{\varepsilon}_{00}^c},$$

and

$$\underset{(n_{11}^c \times 1)}{\boldsymbol{y}_{11}^c} = \underset{(n_{11}^c \times m_1)}{\boldsymbol{B}_{11}^c \boldsymbol{\beta}} + \underset{(n_{11}^c \times 1)}{\boldsymbol{\varepsilon}_{11}^c},$$

where the c superscript indicates the subvectors and submatrices consisting of the rows (observations) classified as compliers in the indicated cells, and each component of the error conditioned on $\{\xi_i^c\}$ is distributed as $\mathcal{N}(0, \sigma_j^2/\xi_i^c)$. These models are analogous to (2.13) in the sharp RD model. Therefore, the posterior distribution of the parameters $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, (λ_0, λ_1), $\sigma^2 = (\sigma_0^2, \sigma_1^2)$ and $\{\xi_i^c\}$, 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 (4.7), we can write

$$\begin{pmatrix} \underset{(n_{00}^n \times 1)}{\boldsymbol{y}_{00}^n} \\ \underset{(n_{10} \times 1)}{\boldsymbol{y}_{10}} \end{pmatrix} = \begin{pmatrix} \underset{(n_{00}^n \times m_n)}{\boldsymbol{B}_{00,n}^n} \\ \underset{(n_{10} \times m_n)}{\boldsymbol{B}_{10,n}} \end{pmatrix} \boldsymbol{\alpha}_n + \begin{pmatrix} \underset{(n_{00}^n \times 1)}{\boldsymbol{\varepsilon}_{00,n}^n} \\ \underset{(n_{10} \times 1)}{\boldsymbol{\varepsilon}_{10,n}} \end{pmatrix}, \quad \textbf{(4.11)}$$

where $\boldsymbol{y}_{00}^n$ and $\boldsymbol{B}_{00,n}^n$ consist of the rows of $\boldsymbol{y}_{00}$ and $\boldsymbol{B}_{00,n}$ in cell (00) that are classified as never-takers, and each component of the error, conditioned on $\{\xi_i^n\}$ is distributed as $\mathcal{N}(0, \sigma_n^2/\xi_i^n)$. Again, the model of these data is analogous to that of the sharp RD model and, therefore, the conditional posterior distribution of the

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

parameters $\boldsymbol{\alpha}_n$, λ_n, σ_n^2 and $\left\{\xi_i^n\right\}$ 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 (4.8), we can write

$$\begin{pmatrix} \underset{(n_{01}\times 1)}{\boldsymbol{y}_{01}} \\ \underset{(n_{11}^a\times 1)}{\boldsymbol{y}_{11}^a} \end{pmatrix} = \begin{pmatrix} \underset{(n_{01}\times m_a)}{\boldsymbol{B}_{01,a}} \\ \underset{(n_{11}^a\times m_a)}{\boldsymbol{B}_{11,a}^a} \end{pmatrix} \boldsymbol{\beta}_a + \begin{pmatrix} \underset{(n_{01}\times 1)}{\boldsymbol{\varepsilon}_{01,a}} \\ \underset{(n_{11}^a\times 1)}{\boldsymbol{\varepsilon}_{11,a}^a} \end{pmatrix}, \tag{4.12}$$

where $\boldsymbol{y}_{11}^a$ and $\boldsymbol{B}_{11,a}^a$ consist of the rows of $\boldsymbol{y}_{11}$ and $\boldsymbol{B}_{11,a}$ in cell (11) that are classified as always-takers, and each component of the error, conditioned on $\left\{\xi_i^a\right\}$ is distributed as $\mathcal{N}(0,\sigma_a^2/\xi_i^a)$. This shows that conditioned on $\{s_i\}$, the parameters $\boldsymbol{\beta}_a$, λ_a, σ_a^2 and $\left\{\xi_i^a\right\}$ 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 $\boldsymbol{\alpha}$ and the first element of $\boldsymbol{\beta}$ 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—we calculate the marginal likelihood by the method of Chib (1995). The details are straightforward and hence omitted.

4.9. Large Sample Analysis

As in Section 2.5, we now derive the large-sample properties of our Bayesian procedure for the fuzzy RD design without control variables w. We denote $\boldsymbol{g} \triangleq (g_0, g_1, g_{0n}, g_{1a})$, $\boldsymbol{\sigma^2} \triangleq (\sigma_0^2, \sigma_1^2, \sigma_{0n}^2, \sigma_{1a}^2)$, $\boldsymbol{q} \triangleq (q_c, q_n, q_a)$, and by $\boldsymbol{g}^*$, $\boldsymbol{\sigma_*^2}$ and $\boldsymbol{q}^*$ their true values. The true data distribution, conditional on $z^{(n)}$, of the sample is denoted by $P_*^{(n)}(y^{(n)}, x^{(n)}|z^{(n)})$, or $P_*^{(n)}$ for short, where $y^{(n)} \triangleq \{y_i, i=1,\ldots,n\}$, $x^{(n)} \triangleq \{x_i, i=1,\ldots,n\}$, and $z^{(n)} \triangleq \{z_i, i=1,\ldots,n\}$.

The quantities $\|\cdot\|_n$ and $\mathcal{C}^\delta[a,b]$ are defined as in Section 2.5 while for every g_j and $j = 0, 1, 0n, 1a$, $\|g_j\|_{\sup} \triangleq \max_{i\in I_j}|g_j(z_i)|$ with $I_0 \triangleq I_{00}$, $I_1 \triangleq I_{11}$, $I_{0n} \triangleq I_{10}$, and $I_{1a} \triangleq I_{01}$. For $j = 0, 1, n, a$ denote by $\mathcal{C}_{m_j}$ the set of piecewise cubic functions and by $\mathcal{S}_{m_j} \subset \mathcal{C}_{m_j}$ the subspace of natural cubic splines on the set of m_j knots defined in Section 4.5. Therefore, $g_{m_0}(z) \triangleq \boldsymbol{B}_{00}(z)'\boldsymbol{\alpha} \in \mathcal{S}_{m_0} \subset \mathcal{C}_{m_0}$ is the unique natural cubic spline interpolating g_0 at the knots $\{z_{00,\,\mathsf{min}}, \kappa_{0,2}, \ldots, \kappa_{0,m_0-1}, \tau\}$. Similarly, define $g_{m_1}(z) \triangleq \boldsymbol{B}_{11}(z)'\boldsymbol{\beta}$, $g_{m_n}(z) \triangleq \boldsymbol{B}_{0,n}(z)'\boldsymbol{\alpha}_n$, and $g_{m_a}(z) \triangleq \boldsymbol{B}_{1,a}(z)'\boldsymbol{\beta}_a$ for given m_1, m_n, and m_a and corresponding knots described in Section 4.5, where $\boldsymbol{B}_{j\ell}$ and $\boldsymbol{B}_{j,\ell}$ are m_ℓ-vectors defined in Appendix A. We denote by $\boldsymbol{\alpha}^*$, $\boldsymbol{\beta}^*$, $\boldsymbol{\alpha}_n^*$, and $\boldsymbol{\beta}_a^*$ the coefficients corresponding to the interpolations of g_0^*, g_1^*, g_{0n}^*, and g_{1a}^*, respectively. Finally, π denotes both the prior on $(\boldsymbol{g}, \boldsymbol{\sigma^2}, \boldsymbol{q}, \lambda_0, \lambda_1, \lambda_n, \lambda_a)'$ that is specified in Section 4.7 and the corresponding posterior, and for a probability measure P, the notation Ph will abbreviate $\int h dP$.

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

To derive our consistency result, we need the following assumption, which is similar to Assumption 4, and the same remarks apply.

Assumption 9. (i) For $j = 0, 1, n, a$ and for every $m_j \in \mathbb{N}_+$ there exists a finite $w_j \in \mathbb{R}^{m_j}$ such that $\boldsymbol{\alpha}^* - \boldsymbol{D}_\alpha^{-1}\boldsymbol{\alpha}_0 = \boldsymbol{D}_\alpha^{-1}\boldsymbol{T}_\alpha^{1/2} w_0$, $\boldsymbol{\alpha}_{\boldsymbol{n}}^{\boldsymbol{*}} - \boldsymbol{D}_n^{-1}\boldsymbol{\alpha}_{\boldsymbol{0n}} = \boldsymbol{D}_n^{-1}\boldsymbol{T}_n^{1/2} w_n$, $\boldsymbol{\beta}^* - \boldsymbol{D}_\beta^{-1}\boldsymbol{\beta}_0 = \boldsymbol{D}_\beta^{-1}\boldsymbol{T}_\beta^{1/2} w_1$, and $\boldsymbol{\beta}_{\boldsymbol{a}}^{\boldsymbol{*}} - \boldsymbol{D}_a^{-1}\boldsymbol{\beta}_{0a} = \boldsymbol{D}_a^{-1}\boldsymbol{T}_a^{1/2} w_a$.
(ii) For $j = 0, 1, n, a$, for $k = \alpha, \beta, n, a$, and for every $m_j \in \mathbb{N}_+$ the maximum eigenvalue of $\boldsymbol{D}_k^{-1}\boldsymbol{T}_k\boldsymbol{D}_k^{-1'}$ is bounded away from zero and infinity.

We are now ready to establish the consistency and contraction rate of the posterior distribution of the RD CATE.

THEOREM 4.2. *Assume that for some* $-\infty < a < b < \infty$ *there exists a* $\delta > 0$ *and a constant* $C_2 > 0$ *such that:* $g_0^* \in \mathcal{G}_0 \triangleq \mathcal{C}^\delta[a, \tau)$, $g_1^* \in \mathcal{G}_1 \triangleq \mathcal{C}^\delta[\tau, b]$ *and* $g_{0n}^*, g_{1a}^* \in \mathcal{C}^\delta[a, b]$, *and* $\|g_j^* - g_{m_k}^*\|_\infty \le C_2 m_k^{-\delta}$ *for* $j = 0, 1, 0n, 1a$ *and* $k = 0, 1, n, a$. *Moreover,* $0 < \underline{\sigma}_j^2 < \sigma_{j*}^2 < \overline{\sigma}_j^2 < \infty$, *for* $j = 0, 1, 0n, 1a$ *and two constants* $\underline{\sigma}_j^2, \overline{\sigma}_j^2$. *Let* π *be the prior on* $(\boldsymbol{g}, \boldsymbol{\sigma}^2, \boldsymbol{q}, \{\lambda_j, j = 0, 1, n, a\})'$ *specified in Section 4.7 with* $m_j = m_j^* \asymp \left(\frac{n}{\log n}\right)^{1/(2\delta+1)}$, $\nu_{00} > 2$, $\delta_{00} > 5$, *and* $a_{j0} = 2\eta m_j^*$ *for* $j = 0, 1, n, a$, *and* $\eta > \min\{7, \delta_{00}\}(2\delta + 1)/4$. *Moreover, suppose that Assumptions 1–3 and 5–9 hold and let* $n_{lj} \asymp n$ *for* $l, j = 0, 1$. *Then, for all* $M_{1,n}, \widetilde{M}_n \to \infty$ *we have that*

$$P_*^{(n)}\pi\left(|\mathrm{CATE} - \mathrm{CATE}^*| \ge M_{1,n}\epsilon_n, |1 - \sigma_j/\sigma_{j*}| \right.$$
$$\left. \ge \widetilde{M}_n/\sqrt{n} \; for \; j = 0, 1, 0n, 1a | y^{(n)}, x^{(n)}, z^{(n)}\right) \to 0,$$

where $\epsilon_n \asymp \left(\frac{\log n}{n}\right)^{\delta/(2\delta+1)}$.

The smoothness assumption in this theorem is standard in the literature (see e.g., Calonico et al., 2014, Assumption 3). To prove this result, we write the RD CATE as a linear functional of $\boldsymbol{g}$ with respect to the empirical measure, as in the proof of Theorem 2.2. Other than this connection, however, the proof of this theorem is substantially different because the fuzzy model is a mixture over two types of latent variables: the scaling variable ξ in the normal model (which is also present in the sharp case), and the unobserved confounder s. Among other things, under this double mixture, the construction of an uniformly exponentially consistent test is more difficult. Moreover, determining whether the true distribution $P_*^{(n)}$, characterized by $(\boldsymbol{g}^*, \boldsymbol{\sigma}_{\boldsymbol{*}}^{\boldsymbol{2}}, \boldsymbol{q}^{\boldsymbol{*}})$, is in the KL support of the prior is more complex. This is due to the fact that the KL divergence, and the KL second moment, involve the logarithms of the mixture over the unobserved confounder s. We give the essentials of the proof in Appendix C.3, but because the proof is long and involved, we provide its complete version in the Online Supplementary Appendix F.

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

5. EXAMPLE: FUZZY DESIGN

To illustrate our ideas, we consider simulated data from a relatively simple design that satisfies the conditions of Theorem 4.1. The four unknown functions are defined as:

$$y_{0c} = -8z + z^2 + z^3 + \sin(2z) + \varepsilon_0,$$
$$y_{1c} = 4z + z^2 + z^3 + 5\sin(z) + 1 + \varepsilon_1,$$

$y_{0n} = 0 + 10z + \varepsilon_{0n}$ and $y_{1a} = 3 - 20z + \varepsilon_{1a}$, where $z \sim 0.1295 \times \mathcal{N}(0,1)$, $\tau = 0$, $\varepsilon_0 \sim t_3(0,1)$ and $\varepsilon_1 \sim t_3(0,1)$, and ε_{0n} and ε_{1a} are $.5t_3(0,1)$ and $2t_3(0,1)$, respectively, and the types $s \in \{c,n,a\}$ are generated from a discrete distribution with probabilities $\boldsymbol{q} = (0.5, 0.25, 0.25)$. In this design, the true value of the CATE is 1.0. We also consider the same design with student-t errors by letting ε_0 and ε_1 be distributed as standard student-t, and ε_{0n} and ε_{1a} distributed as 0.5 and 2 times standard student-t, respectively. Our simulated data consists of a sample of size n and we consider $n = 500$ and $n = 4,000$.

For the prior distribution, we assume that $\sigma_0^2 = \sigma_1^2 = \sigma^2$ and a priori, $(\sigma^2, \sigma_n^2, \sigma_a^2)$ 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 SD equal to $(5,5,5,5)$. A priori, we also assume that $\boldsymbol{q}$ 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 $\boldsymbol{p}$ is $(0.5, 0.5)$, and for $n = 4,000$, it is $(0.6, 0.4)$. For both sample sizes, the knots for the four functions are set by the values $\boldsymbol{m}_z = (3,3)$, $m_{z,n} = 5$, and $m_{z,a} = 5$. Finally, for $n = 500$, $\boldsymbol{m}_{z,\tau} = (2,2)$, and for $n = 4,000$, $\boldsymbol{m}_{z,\tau} = (3,3)$.

We apply our procedure to each sample size for data generated from t_3 distributed errors. The results from the estimation, 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.

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 λ_0 is relatively more dispersed and includes some mass on larger values. In addition, since the g_{0n} and g_{1a} 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 λ_0 and λ_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 g_0, g_1, g_{0n}, and g_{1a} are given in Figure 7. 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

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

TABLE 5. Fuzzy RD with Data Simulated with Student t_3 errors: Summary Results for the Dispersion Parameters $(\sigma^2, \sigma_n^2, \sigma_a^2)$ 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.

Parameter	True value	Prior		Posterior			
		Mean	SD	Mean	Lower	Upper	Ineff
				$n = 500$			
σ^2	1	2	5	0.998	0.746	1.313	4.402
σ_n^2	0.25	2	5	0.290	0.202	0.408	2.466
σ_a^2	4	2	5	3.371	2.246	4.942	4.641
q_c	0.5	0.33	0.178	0.513	0.460	0.566	2.374
q_n	0.25	0.33	0.178	0.216	0.180	0.252	1.000
q_a	0.25	0.33	0.178	0.271	0.223	0.321	2.730
				$n = 4{,}000$			
σ^2	1	2	5	0.973	0.880	1.076	4.707
σ_n^2	0.25	2	5	0.279	0.246	0.314	2.981
σ_a^2	4	2	5	4.101	3.547	4.727	4.302
q_c	0.5	0.33	0.178	0.510	0.491	0.529	2.312
q_n	0.25	0.33	0.178	0.246	0.232	0.260	1.035
q_a	0.25	0.33	0.178	0.244	0.228	0.261	2.557

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.

5.0.1. Sampling Experiments

We conclude this discussion with some evidence about the sampling properties of the Bayes CATE estimate based on 1,000 replications of the design used above. The results, which are given in Table 7, parallel those in the sharp design.

6. EXTENSION

We briefly mention that the framework we have developed is straightforwardly extended to a nonparametric distribution of the errors. For example, in the sharp-design, we could let

$$y_{0i} = g_0(z_i) + h(w_i) + \sigma_0 \varepsilon_{0i}\,, \; z < \tau,$$
$$y_{1i} = g_1(z_i) + h(w_i) + \sigma_1 \varepsilon_{1i}\,, \; z \geq \tau,$$

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 λ_n and λ_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 λ_0 and λ_1 show the move to small values.

Parameter	Prior		Posterior			
	Mean	SD	Mean	Lower	Upper	Ineff
			$n = 500$			
λ_0	1	5	3.877	0.171	15.066	2.259
λ_1	1	5	1.275	0.084	4.416	2.007
λ_n	1	5	0.086	0.009	0.297	1.254
λ_a	1	5	0.039	0.006	0.107	1.239
			$n = 4,000$			
λ_0	1	5	1.236	0.082	6.391	6.444
λ_1	1	5	0.060	0.008	0.182	2.051
λ_n	1	5	0.327	0.035	1.103	1.254
λ_a	1	5	0.050	0.008	0.130	1.034

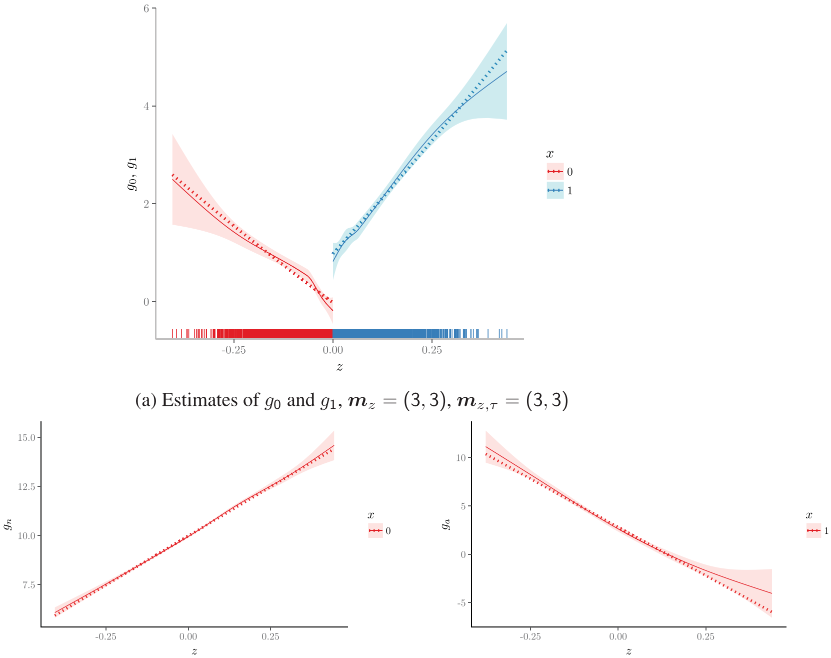
where now the error distributions are unknown and modeled (say) by a Dirichlet process mixture (DPM) prior. For instance, following Chib and Greenberg (2010), we could suppose that

$$\begin{aligned}\varepsilon_{0i}|\xi_{0i} &\sim N(0, \xi_{0i}^{-1}),\\ \xi_{0i}|F &\overset{iid}{\sim} F,\\ F|\alpha_0, F_0 &\sim DP(\alpha_0 F_0),\\ F_0 &= Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}\right),\end{aligned}$$

where DP is the Ferguson (1973) and Antoniak (1974) Dirichlet process with mass parameter α_0 and base distribution F_0. Thus, under the base measure, the distribution of the error is student-t with ν degrees of freedom, as before, though the distribution of the error is now nonparametric. We could model the distribution of ε_1 in the same way by letting

$$\begin{aligned}\varepsilon_{1i}|\xi_{1i} &\sim N(0, \xi_{1i}^{-1}),\\ \xi_{1i}|G &\overset{iid}{\sim} G,\\ G|\alpha_0, G_0 &\sim DP(\alpha_0 G_0),\\ G_0 &= Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}\right).\end{aligned}$$

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press



(a) Estimates of g_0 and g_1, $\boldsymbol{m}_z = (3,3)$, $\boldsymbol{m}_{z,\tau} = (3,3)$

(b) Estimate of g_n, $m_{z,n} = 5$

(c) Estimate of g_a, $m_{z,a} = 5$

FIGURE 7. Fuzzy RD with discrete confounder and student t_3 errors: This shows the function estimates and credibility bands for $n = 4,000$. Note that, as required by our conditions, the functions g_{0n} and g_{1a} are both continuous at τ. This is achieved by having a knot at τ. See text for further details.

TABLE 7. Summary of Results from 1,000 Repeated Samples: Fuzzy RD Design with t_3 Errors and Two Sample Sizes, True Value of the CATE is 1.0.

Mean	Coverage	RMSE
	$n = 500$	
0.976	0.970	0.358
	$n = 4,000$	
1.014	0.970	0.222

The advantage of using this specific construction of the DPM prior is that the MCMC updates are all tractable (see Chib and Greenberg, 2010) and the marginal likelihood of the model is also tractable, being calculable by the method of Basu and Chib (2003). We can complete the model using a basis expansion approach for the $h(w)$ function, making sure that a knot is placed at τ to ensure continuity of the function at τ. Thus, in this model, both the mean functions and error distributions are nonparametric.

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

6.1. Example

As an illustrative example, consider the design of Section 3.1 where as before

$$g_0(z) = 0.48 + 1.27z + 7.18z^2 + 20.21z^3 + 21.54z^4 + 7.33z^5,$$
$$g_1(z) = 0.52 + 0.84z - 3.00z^2 + 7.99z^3 - 9.01z^4 + 3.56z^5,$$

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

but with the model modified to include the effect of the control covariate w through the function

$$h(w) = \frac{\sin(\pi w/2)}{1 + w^2(sign(w)+1)},$$

$$w \sim \mathrm{Uniform}(-\pi, \pi),$$

and with the error distribution defined by $\sigma_0 = 0.1295$, $\sigma_1 = 0.20$ and ε_0 and ε_1 following a *multi-modal* mixture of three-component normal distributions with component means equal to $(-2.5, 0, 2.5)$, component SDs equal to $(1,1,1)$, and component weights equal to $(1/3, 1/3, 1/3)$. The true value of the RD ATE at the break-point $\tau = 0$ is 0.04. We draw a sample of $n = 500$ and $n = 4{,}000$ from this design.

This design is clearly considerably more complex than the one considered in Section 3.1. Estimation of the RD ATE is based on adding two steps to our MCMC algorithm, one in which the ξ_{0i} and ξ_{1i} are sampled, and a second step in which the DPM mass parameters are sampled, both as described in Chib and Greenberg (2010). The prior is, once again, not tuned to the specifics of this model. For $n =$ 500, we use a soft-window specification defined by $\boldsymbol{p} = (0.6, 0.4)$, with number of knots in the basis expansion of g_0 and g_1 given by $\boldsymbol{m}_z = (3,3)$, $\boldsymbol{m}_{z,\tau} = (3,2)$, the number of knots in the basis expansion of $h(w)$ equal to 8, prior mean of the two initial ordinates of each of the three unknown functions equal to $(0,0)$, the prior mean of the (now) three λ's equal to $(1,1,1)$ with prior standard deviations equal to $(15,15,15)$, and the prior mean and prior standard deviations of σ_0 and σ_1 equal to 0.3 and 5, respectively. The prior of the two DP mass parameters α_0 and α_1 is defined by a prior mean equal to 5 and prior standard deviation of 1. We fix the value of ν at 3. The prior for the larger sample size is exactly the same. The only change is in the design of the window width parameters and the number of knots in the basis expansions, which take the values $\boldsymbol{p} = (0.9, 0.1)$, $\boldsymbol{m}_z = (4,3)$, $\boldsymbol{m}_{z,\tau} = (3,2)$, and 20 knots in the case of $h(w)$. These values are determined by marginal likelihoods, calculated by the method of Basu and Chib (2003).

The results on the RD-ATE from our fully nonparametric estimation are given in Table 8. The posterior mean for the smaller sample size is 0.079 with posterior SD equal to 0.168, and for the larger sample the posterior mean is 0.062 with posterior SD of 0.078. In contrast, the frequentist estimator of the RD ATE is 0.2946 with standard error equal to 0.1302, and for the larger sample size the estimate is 0.104

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

Table 8. Sharp Design: Posterior Summaries for Two Different Sample Sizes from the Dirichlet Process Mixture Estimation on Simulated Data with Nonparametric Control, Error Distributed as Mixture of Normals and Different SDs.

Mean	SD	Median	Lower	Upper
		$\underline{n = 500}$: log marglik $= -225.129$		
0.079	0.168	0.075	−0.240	0.410
		$\underline{n = 4,000}$: log marglik $= 1277.468$		
0.062	0.078	0.062	−0.090	0.213

Table 9. Sharp Design (True Value of RD ATE is 0.04) Repeated Sampling Results from 1,000 Samples Drawn from the Design with Error Distributed as Mixture of Normals, a Control Covariate w and Different SDs on Each Side of τ (See Text for Further Details). RMSE of the Bayesian posterior mean under the Dirichlet process mixture assumption for the error, and frequentist coverage of the Bayesian 95% credibility interval. The IK and CCT estimates are computed from the R package rdrobust with the call outr = rdrobust(y, z, covs = w, all = TRUE), where outr is a list object. The element coef of this list contains the IK and CCT estimates as the first and third elements, respectively.

	Mean	Coverage	RMSE
		$\underline{n = 500}$	
Bayes	0.061	0.947	0.083
IK	0.053	1	0.252
CCT	0.046	0.915	0.294
		$\underline{n = 2,000}$	
Bayes	0.018	0.926	0.066
IK	0.068	0.933	0.123
CCT	0.063	0.942	0.140

with standard error of 0.048. Despite the complexity of the problem, the Bayesian results from the nonparametric control and error model are satisfactory.

As a final exercise, we consider 1,000 simulated data sets generated from the preceding design. The simulation is geared to examining the sampling performance of our Bayesian approach along two dimensions—the sampling RMSE of the posterior mean, and the coverage of the 95% interval estimator. For comparison, we also calculate the RMSE and coverage of two frequentist estimators: the Imbens and Kalyanaraman (2012) (IK) estimator, which uses the MSE-optimal bandwidth and should be expected to produce the minimal RMSE; and the Calonico, Cattaneo, and Titiunik (2014) (CCT) estimator, which is coverage-optimal and, by construction, uses a bandwidth that is smaller than the MSE-optimal one. The results are in Table 9.

The table shows that the sampling performance of our Bayesian point and interval estimators is excellent. Specifically, the sampling RMSE of the Bayesian

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

posterior mean is the lowest of the three estimates, for both sample sizes. The sampling coverage of the Bayesian posterior credibility interval is closer to the nominal value for $n = 500$, but below that of the coverage-optimal frequentist estimate for $n = 2,000$, at least in this example.

7. CONCLUSIONS

In summary, we have provided a Bayesian framework for inference in the sharp and fuzzy RD designs that is based on several novel ideas. First, this framework uses all the data in the sample, along with local nonparametric methods, to model and estimate the distributions of the outcomes on either side of the threshold. We use basis expansion techniques for the unknown functions of the forcing variable with extra knots sprinkled in the vicinity of the threshold, and one knot (for each function) exactly at the threshold. The local nature of our cubic spline procedure, along with the strategic placement of knots, ensures that the data around the threshold are automatically more important in inferences about the jump size. We develop a new, flexible 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 to specify this prior—essentially a rough idea of the first two ordinates at the extreme points on both sides of τ—should be known to an investigator with knowledge of the specifics of the application. We show that this prior satisfies the KL property and is key in our derivation of the large-sample behavior of the posterior distribution of the RD effects. Our probability model for the fuzzy RD design, inspired by the principal stratification framework, is new. In this model, the unobserved confounder is a discrete random variable, not continuous as in the literature to date. One interesting aspect of this formulation is that the sharp design is a special case, and the estimation approach for the fuzzy RD design also similarly reduces to that of the sharp case when all individuals are compliers. Finally, for both designs, we establish the large-sample properties of our procedures, and the consistency and posterior contraction rates for the causal effects.

Given the importance of RD designs in practice, this broadening of the analytical framework for inference in such designs is likely to prove useful. Just as researchers can apply different frequentist estimators to these designs to see how the RD effects vary across estimators, it is now possible for researchers to also calculate the effects from our Bayesian perspective with ease that rivals that of the existing approaches. Comparing and contrasting the estimates (whether identical or different in any particular instance) from the various approaches should deepen understanding of the effects in practice.

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 $h_j = \kappa_j - \kappa_{j-1}$

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

the spacing between the $(j-1)$st and jth knots. The basis functions are the collections of cubic splines $\left\{\Phi_j(z)\right\}_{j=1}^{m}$ and $\{\Psi_j(z)\}_{j=1}^{m}$, where for $2 \le j \le m-1$,

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

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

and for $j = 1$, Φ_j and Ψ_j are defined by the last two lines of equations (A.1) and (A.2), respectively, and for $j = m$, Φ_j and Ψ_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

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

where $\boldsymbol{f} = (f_1, \ldots, f_m)'$ and $\boldsymbol{s} = (s_1, \ldots, s_m)'$ 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 $\boldsymbol{Cf} = \boldsymbol{As}$, $\boldsymbol{s} = \boldsymbol{A}^{-1}\boldsymbol{Cf}$, where

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

and

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

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

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

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

or as $g(z_i) = \boldsymbol{c}_i'\boldsymbol{f}$, where $\boldsymbol{\Phi}(z_i)' = (\Phi_1(z_i), \ldots, \Phi_m(z_i))$, $\boldsymbol{\Psi}(z_i)' = (\Psi_1(z_i), \ldots, \Psi_m(z_i))$, and $\boldsymbol{c}_i' = \boldsymbol{\Phi}(z_i)' + \boldsymbol{\Psi}(z_i)'\boldsymbol{A}^{-1}\boldsymbol{C}$, which implies the following representation for the $n \times m$ basis matrix: $\boldsymbol{B} = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n)' = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m]$. This is the form of the basis matrices $\boldsymbol{B}_0$, $\boldsymbol{B}_1$, $\boldsymbol{B}_{00}$, $\boldsymbol{B}_{01}$, $\boldsymbol{B}_{0n}$, and $\boldsymbol{B}_{1a}$. Further, we denote $\boldsymbol{B}_j(z) \triangleq (\boldsymbol{\Phi}(z)' + \boldsymbol{\Psi}(z)'\boldsymbol{A}^{-1}\boldsymbol{C})'$ the m_j-vector used in Sections 2.5 and 4.9.

B. APPENDIX: $\boldsymbol{D}_\alpha, \boldsymbol{D}_\beta$

Suppose that m is the dimension of α and β. Then, the matrices $\boldsymbol{D}_\alpha$ and $\boldsymbol{D}_\beta$ in equations (2.7) and (2.10) take the following forms:

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

and

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

respectively.

C. PROOFS

C.1. Notation

The notation $\lesssim$ (resp. $\asymp$) will be used to denote inequality (resp. equality) up to a fixed constant. For a vector v, denote by $\|v\|_\infty \triangleq \max_i |v_i|$. The indicator function of an event $\mathcal{A}$ is denoted by both $I[\mathcal{A}]$ and $\mathbb{1}_{\mathcal{A}}$.

The true data distribution, conditional on z, of the sample is denoted by $P_*^{(n)}$ in both designs and is as specified in Assumptions 1 and 5–7, respectively. The true data distribution, conditional on z_i, for one observation is denoted by P_*^i in both designs for

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

short. The Lebesgue density of $P_*^{(n)}$ (resp. P_*^i) is denoted by $p_*^{(n)}$ (resp. $p_*(y_i|z_i)$ in the sharp RD design and $p_*(y_i,x_i|z_i)$ in the fuzzy RD design). In the proof, we exploit the fact that we can write the student-t distribution as a mixture of Normal distributions: in the sharp RD design $p_*(y_i|z_i) = t_\nu(y_i|g_j^*(z_i),\sigma_{j*}^2) = \int \mathcal{N}(y_i|z_i;g_j^*,\sigma_{j*}^2/\xi_i)\mathcal{G}a\left(\xi_i;\frac{\nu}{2},\frac{\nu}{2}\right)d\xi_i$ for $j=0,1$ (and similarly in the fuzzy RD design). The expectation taken with respect to $P_*^{(n)}$ is denoted by $\mathbf{E}_*$. For a probability measure P, the notation Ph will abbreviate $\int hdP$. Define the KL divergence between two probability measures P and Q as $K(P,Q) \triangleq P\log(p/q)$, where p and q are the Lebesgue densities of P and Q respectively, and define the discrepancy measure

$$V(P,Q) \triangleq P\left|\log(p/q) - K(P,Q)\right|^2.$$

For a set $\mathcal{F}$ and a metric d on it let $N(\epsilon,\mathcal{F},d)$ be the ϵ-covering number which is defined as the minimum number of balls of radius ϵ needed to cover $\mathcal{F}$ see for example, van der Vaart (2000).

Notation specific to the Sharp RD design. Let $\mathcal{G}_0 \triangleq \mathcal{C}^\delta[a,\tau)$, $\mathcal{G}_1 \triangleq \mathcal{C}^\delta[\tau,b]$ and

$$B_*^c(\epsilon_{n_j},n^{-1/2}) \triangleq \{(g_j,\sigma_j^2) \in \mathcal{G}_j \times \mathbb{R}_+; \|g_j - g_j^*\|_{n_j} \geq M_n\epsilon_{n_j}, |1-\sigma_j/\sigma_{j*}| \geq \widetilde{M}_n/\sqrt{n_j}\}$$

for $j=0,1$ and any sequences $M_n, \widetilde{M}_n \to \infty$. For $j=0,1$, let $P_{g_j,\sigma_j^2}^{(n_j)}$ (resp. $P_{g_j,\sigma_j^2}^i$) denote the conditional distribution of the sample $y^{(n_j)}$ given $(z^{(n_j)},g_j,\sigma_j^2)$ (resp. of y_i given (z_i,g_j,σ_j^2)), and $p_{g_j,\sigma_j^2}^{(n_j)}$ (resp. $p_{g_j,\sigma_j^2}(y_i|z_i)$) denote the Lebesgue density of $P_{g_j,\sigma_j^2}^{(n_j)}$ (resp. of $P_{g_j,\sigma_j^2}^i$). Denote the subsamples $I_0 \triangleq \{i; z_i < \tau\}$ and $I_1 \triangleq \{i; z_i \geq \tau\}$. For $j=0,1$, the KL neighborhood of (g_j^*,σ_{j*}^2) is defined as, $\forall \epsilon_{n_j} > 0$

$$B_n^{KL}((g_j^*,\sigma_{j*}^2),\epsilon_{n_j}) \triangleq \left\{(g_j,\sigma_j^2) \in \mathcal{G}_j \times \mathbb{R}_+; K\left(P_*^{(n_j)},P_{g_j,\sigma_j^2}^{(n_j)}\right) \leq n_j\epsilon_{n_j}^2, V\left(P_*^{(n_j)},P_{g_j,\sigma_j^2}^{(n_j)}\right) \leq 2n_j\epsilon_{n_j}^2\right\}. \tag{C.1}$$

Finally, in the proofs we use the following sequence of measurable sets for $M_k = 4\gamma_k n_k\epsilon_n^2 \asymp n_k\epsilon_n^2 \asymp n\epsilon_n^2$, where $\gamma_k \triangleq \max_{1\leq j\leq m_k^*}|(\boldsymbol{D}_{\boldsymbol{\alpha_k}}^{-1}\boldsymbol{T}_{\boldsymbol{\alpha_k}}\boldsymbol{D}_{\boldsymbol{\alpha_k}}^{-1'})_{jj}|$, $\boldsymbol{\alpha}_k \triangleq \boldsymbol{\alpha}I[k=0] + \boldsymbol{\beta}I[k=1]$ and $m_k^* \asymp \left(\frac{n_k}{\log(n_k)}\right)^{1/(2\delta+1)}$, $k=0,1$:

$$\mathcal{C}_{n,0} \triangleq \bigcup_{m_0=1}^{m_0^*}\left\{g_{m_0}(z) \in \mathcal{S}_{m_0}; \boldsymbol{\alpha} \in [-M_0,M_0]^{m_0}\right\},$$
$$\mathcal{C}_{n,1} \triangleq \bigcup_{m_1=1}^{m_1^*}\left\{g_{m_1}(z) \in \mathcal{S}_{m_1}; \boldsymbol{\beta} \in [-M_1,M_1]^{m_1}\right\}, \tag{C.2}$$

where $g_{m_0}(z) \triangleq \boldsymbol{B}_0(z)'\boldsymbol{\alpha}$ and $g_{m_1}(z) \triangleq \boldsymbol{B}_1(z)'\boldsymbol{\beta}$.

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

Notation specific to the Fuzzy RD design. For the Fuzzy RD design we in addition use the following notation. Let $\mathcal{G}_{0n} = \mathcal{G}_{1a} \triangleq \mathcal{C}^{\delta}[a,b]$, and

$$\begin{aligned}\mathcal{F} \triangleq \Big\{ \Big(& g_{0,0n}(s,z) I[s \in \{c,n\}] + g_{1a}(z) I[s = a] \Big) I[z < \tau] + \Big(g_{1,1a}(s,z) I[s \in \{c,a\}] \\ & + g_{0n}(z) I[s = n] \Big) I[z \geq \tau], \quad \text{with } g_{0,0n}(s,z) \triangleq g_0(z) I[s = c] + g_{0n}(z) I[s = n], \\ & g_{1,1a}(s,z) \triangleq g_1(z) I[s = c] + g_{1a}(z) I[s = a],\, g_j \in \mathcal{G}_j, j = 0,1,0n,1a \Big\}.\end{aligned}$$

We use the notation $\bar{n}_{10} \triangleq n_{00} + n_{10}$, $\bar{n}_{01} \triangleq n_{00} + n_{10} + n_{01}$. For the sample size of the cells I_{10} and I_{01} we use both the notations $n_{0n} \triangleq n_{10}$ and $n_{1a} \triangleq n_{01}$; moreover, $I_{0n} \triangleq I_{10}$ and $I_{1a} \triangleq I_{01}$. For a given sequence $\boldsymbol{s} \triangleq \{\boldsymbol{s_0}, \boldsymbol{s_1}\}$ where $\boldsymbol{s_k} \triangleq \{s_i\}_{i \in I_{kk}}$ with values in $\{c,n\}$ if $k = 0$ and in $\{c,a\}$ if $k = 1$ denote:

$$\begin{aligned} \text{for } f \in \mathcal{F}, \quad & \boldsymbol{f_s} \triangleq (\boldsymbol{f}'_{\boldsymbol{s_0}}, f(n, z_{n_{00}+1}), \ldots, f(n, z_{\bar{n}_{10}}), f(a, z_{\bar{n}_{10}+1}), \ldots, f(n, z_{\bar{n}_{01}}), \boldsymbol{f}'_{\boldsymbol{s_1}})', \\ & \boldsymbol{f_{s_0}} \triangleq (f(s_1, z_1), \ldots, f(s_{n_{00}}, z_{n_{00}}))', \qquad \boldsymbol{f_{s_1}} \triangleq (f(s_{\bar{n}_{01}+1}, z_{\bar{n}_{01}+1}), \ldots, f(s_n, z_n))', \\ & \boldsymbol{\sigma^2_{0s_0}} \triangleq (\sigma^2_{0s_1}, \ldots, \sigma^2_{0s_{n_{00}}})', \qquad \boldsymbol{\sigma^2_{1s_1}} \triangleq (\sigma^2_{1s_{\bar{n}_{01}+1}}, \ldots, \sigma^2_{1s_n})', \\ & \boldsymbol{g_{0s_0}} \triangleq (g_{0s_1}(z_1), \ldots, g_{0s_{n_{00}}}(z_{n_{00}}))', \qquad \boldsymbol{g_{1s_1}} \triangleq (g_{1s_{\bar{n}_{01}+1}}(z_{\bar{n}_{01}+1}), \ldots, g_{1s_n}(z_n))'. \end{aligned} \tag{C.3}$$

Moreover, define $B_*(\epsilon_n, n^{-1/2}) \triangleq \{(f, \sigma_j^2) \in \mathcal{F} \times \mathbb{R}_+ \text{ for } j = 0,1,0n,1a; |1 - \sigma_j/\sigma_{j*}| \geq \widetilde{M}_n/\sqrt{n}$, and $\forall \boldsymbol{s}, \|\boldsymbol{f_s} - \boldsymbol{f^*_s}\|_n \geq M_n \epsilon_n\}$ for any sequences $M_n, \widetilde{M}_n \to \infty$.

Prior specification for both the Sharp and the Fuzzy RD designs. The prior specification for $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\alpha}_n, \boldsymbol{\beta}_a, \boldsymbol{\lambda}, \sigma_0^2, \sigma_1^2, \sigma_n^2, \sigma_a^2, \boldsymbol{q})$, where $\boldsymbol{\lambda} \triangleq (\lambda_0, \lambda_1, \lambda_n, \lambda_a)$ and $\boldsymbol{q} \triangleq (q_c, q_n, q_a)$ is an independent prior as follows:

$$\begin{aligned} & \boldsymbol{\alpha}|\lambda_0 \sim \mathcal{N}_{m_0}(\boldsymbol{D}_\alpha^{-1}\boldsymbol{\alpha}_0, \lambda_0^{-1}\boldsymbol{D}_\alpha^{-1}\boldsymbol{T}_\alpha\boldsymbol{D}_\alpha^{-1'}), \qquad \boldsymbol{\beta}|\lambda_1 \sim \mathcal{N}_{m_1}(\boldsymbol{D}_\beta^{-1}\boldsymbol{\beta}_0, \lambda_1^{-1}\boldsymbol{D}_\beta^{-1}\boldsymbol{T}_\beta\boldsymbol{D}_\beta^{-1'}), \\ & \boldsymbol{\alpha}_n|\lambda_n \sim \mathcal{N}_{m_n}(\boldsymbol{D}_n^{-1}\boldsymbol{\alpha}_{0n}, \lambda_n^{-1}\boldsymbol{D}_n^{-1}\boldsymbol{T}_n\boldsymbol{D}_n^{-1'}), \qquad \boldsymbol{\beta}_a|\lambda_a \sim \mathcal{N}_{m_a}(\boldsymbol{D}_a^{-1}\boldsymbol{\beta}_{0a}, \lambda_a^{-1}\boldsymbol{D}_a^{-1}\boldsymbol{T}_a\boldsymbol{D}_a^{-1'}), \\ & \lambda_j \sim \mathcal{G}a\left(\frac{a_{j0}}{2}, \frac{b_{j0}}{2}\right), \qquad j = 0,1,n,a, \\ & \sigma_j^2 \sim \mathcal{IG}\left(\frac{\nu_{00}}{2}, \frac{\delta_{00}}{2}\right), \text{ for } j = 0,1, \quad \sigma_n^2 \sim \mathcal{IG}\left(\frac{\nu_{0n}}{2}, \frac{\delta_{0n}}{2}\right), \quad \sigma_a^2 \sim \mathcal{IG}\left(\frac{\nu_{0a}}{2}, \frac{\delta_{0a}}{2}\right), \\ & \boldsymbol{q} \sim Dir(n_{0c}, n_{0n}, n_{0a}), \end{aligned} \tag{C.4}$$

where $\nu_{00} > 2$.

C.2. Proofs for Section 2

C.2.1. *Proof of Theorem 2.1.* For every n, denote $X^{(n)} \triangleq (y^{(n)}, z^{(n)})$ and let $I_0 \triangleq \{i; z_i < \tau\} = \{i = 1, \ldots n_0\}$ and $I_1 \triangleq \{i; z_i \geq \tau\} = \{i = n_0 + 1, \ldots n_1\}$. Moreover, for $k = 0,1$ denote by D_{n_k} the denominator of the posterior distribution of (g_k, σ_k^2). In the following, we use the notation $\boldsymbol{\alpha}_k \triangleq \boldsymbol{\alpha} I[k = 0] + \boldsymbol{\beta} I[k = 1]$ for $k = 0,1$ (and similarly $\boldsymbol{\alpha}_{k0}$ for the prior mean). Therefore, for $k = 0, 1$, the marginal posterior of (g_k, σ_k^2) is

$$\pi((g_k, \sigma_k^2)|X^{(n_k)}) = \frac{1}{D_{n_k}} \int \prod_{\{i \in I_k\}} t_\nu(y_i | \boldsymbol{B}_k(z_i)'\boldsymbol{\alpha}_k + g_k^\perp, \sigma_k^2) \mathcal{N}_{m_k}$$
$$\times (\boldsymbol{D}_{\boldsymbol{\alpha}_k}^{-1} \boldsymbol{\alpha}_{k0}, \lambda_k^{-1} \boldsymbol{D}_{\boldsymbol{\alpha}_k}^{-1} \boldsymbol{T}_{\boldsymbol{\alpha}_k} \boldsymbol{D}_{\boldsymbol{\alpha}_k}^{-1'}) \otimes \delta_0(dg_k^\perp) d\pi(\lambda_k) \pi(\sigma_k^2),$$

where δ_0 denotes the Dirac mass at zero and $g_k^\perp \triangleq g_k^\perp(z) \triangleq g_k(z) - \boldsymbol{B}_k(z)'\boldsymbol{\alpha}_k$. For $k = 0, 1$, for some $C > 0$ and $\epsilon_{n_k} > 0$ denote by $\mathcal{A}_k^c$ the following event:

$$\mathcal{A}_k^c \triangleq \left\{ \int \frac{\prod_{i \in I_k} p_{g_k, \sigma_k^2}(y_i|z_i)}{p_*^{(n_k)}} d\overline{\pi}(\boldsymbol{\alpha}_k, g_k^\perp, \lambda_k, \sigma_k^2) \leq e^{-(1+C) n_k \epsilon_{n_k}^2} \right\},$$

where $d\overline{\pi}(\boldsymbol{\alpha}_k, g_k^\perp, \lambda_k, \sigma_k^2)$ denotes the prior supported on $B_n^{KL}((g_j^*, \sigma_{j*}^2), \epsilon_{n_j})$. By Ghosal and van der Vaart (2007), Lemma 10), $P_*^{(n_k)}[\mathcal{A}_k^c] \leq \frac{1}{C^2 n_k \epsilon_{n_k}^2}$ for every $C > 0$.

To prove the result of the theorem, we use the test approach which relies on the uniformly exponential consistent test ϕ_{n_k} constructed in Lemma C.3. By this lemma, for $k = 0, 1$ there exists a test ϕ_{n_k} that satisfies:

$$Q_{g_k^*, \sigma_{k*}^2}^{(n_k)} \phi_{n_k} \leq e^{n_k \epsilon_{n_k}^2} (1 - e^{-KM^2 n_k \epsilon_{n_k}^2})^{-1} e^{-KM^2 n_k \epsilon_{n_k}^2} \qquad \text{and}$$
$$Q_{g_k, \sigma_k^2}^{(n_k)} (1 - \phi_{n_k}) \leq e^{-KM^2 n_k \epsilon_{n_k}^2 j^2}, \qquad \forall (g_k, \sigma_k^2) \in \mathcal{C}_{n,k} \times \left[\frac{1}{2n_k}, e^{3 n_k \epsilon_{n_k}^2} \right]$$
$$\text{such that } |1 - \sigma_k / \sigma_{k*}| > j \varepsilon_\sigma \text{ and } \| \Xi_k^{1/2} (g_k - g_k^*) \|_{n_k} > M \epsilon_{n_k} j, \forall j \in \mathbb{N} \qquad \textbf{(C.5)}$$

for some $\varepsilon_\sigma > 0$, where $Q_{g_k^*, \sigma_{k*}^2}^{(n_k)} \triangleq \prod_{\{i \in I_k\}} \mathcal{N}(y_i | z_i; g_k^*, \sigma_{k*}^2 / \xi_i)$ which, conditional on $\{\xi_i\}_{i=1}^n$, is the true model, $Q_{g_k, \sigma_k^2}^{(n_k)} \triangleq \prod_{\{i \in I_k\}} \mathcal{N}(y_i | z_i; g_k, \sigma_k^2 / \xi_i)$ and $\Xi_0 \triangleq$ diagonal $(\xi_1, \ldots, \xi_{n_0})$, $\Xi_1 \triangleq$ diagonal$(\xi_{n_0+1}, \ldots, \xi_{n_1})$.

We now use the tests ϕ_{n_0} and ϕ_{n_1} to upper bound the posterior of $B_*^c(\epsilon_{n_k}, n^{-1/2})$. For this, we use the fact that, since $t_\nu(y_i | g_j^*(z_i), \sigma_{j*}^2) = \int \mathcal{N}(y_i | z_i; g_j^*, \sigma_{j*}^2 / \xi_i) \mathcal{G}a\left(\xi_i; \frac{\nu}{2}, \frac{\nu}{2}\right) d\xi_i$ for $j = 0, 1$, the true Lebesgue density can be written as a mixture of Normal distributions:

$$p_*^{(n)} = \int Q_{g_0^*, \sigma_{0*}^2}^{(n_0)} Q_{g_1^*, \sigma_{1*}^2}^{(n_1)} \prod_{i=1}^n \mathcal{G}a\left(\xi_i; \frac{\nu}{2}, \frac{\nu}{2}\right) d\xi_i.$$

For $k = 0, 1$:

$$\mathbf{E}_*[\pi(B_*^c(\epsilon_{n_k}, n^{-1/2})|X^{(n_k)})] \leq \mathbf{E}_*[\overbrace{\pi(B_*^c(\epsilon_{n_k}, n^{-1/2})|X^{(n_k)})}^{\leq 1} \phi_{n_k}]$$
$$+ \mathbf{E}_*[\overbrace{\pi(B_*^c(\epsilon_{n_k}, n^{-1/2})|X^{(n_k)})}^{\leq 1} \mathbb{1}_{\mathcal{A}_k^c}] + \mathbf{E}_*[\pi(B_*^c(\epsilon_{n_k}, n^{-1/2})|X^{(n_k)})(1 - \phi_{n_k}) \mathbb{1}_{\mathcal{A}_k}]$$

$$
\begin{aligned}
&\leq \int \frac{e^{-(KM^2-1)n_k\epsilon_{n_k}^2}}{(1-e^{-Kn_kM^2\epsilon_{n_k}^2})} \prod_{i\in I_k} \mathcal{G}a\left(\xi_i;\frac{\nu}{2},\frac{\nu}{2}\right)d\xi_i + \frac{1}{C^2 n_k\epsilon_{n_k}^2} \\
&\quad + \mathbf{E}_*\left[\frac{\int\int_{B_*^c(\epsilon_{n_k},n^{-1/2})} \prod_{\{i\in I_k\}} \frac{p_{g_k,\sigma_k^2}(y_i|z_i)}{p_*(y_i|z_i)} d\pi(\boldsymbol{\alpha}_k, g_k^{\perp}, \lambda_k, \sigma_k^2)}{D_{n_k}/p_*^{(n_k)}}(1-\phi_{n_k})\mathbb{1}_{\mathcal{A}_k}\right],
\end{aligned}
$$

where to get the second inequality, we have used the first line in (C.5) with $M=\xi_{\mathsf{min}}^{1/2}M_n/J$ for $J\in\mathbb{N}$ and $\xi_{\mathsf{min}} \triangleq \mathsf{min}_i(\Xi_k)_{i,i}$. If M is sufficiently large to ensure that $KM^2-1 > KM^2/2$, $\frac{e^{-(KM^2-1)n_k\epsilon_{n_k}^2}}{(1-e^{-Kn_kM^2\epsilon_{n_k}^2})} \leq e^{-KM^2n\epsilon_{n_k}^2/2}$ and using the definition of $\mathcal{A}_k$, we obtain

$$
\begin{aligned}
&\mathbf{E}_*[\pi(B_*^c(\epsilon_{n_k},n^{-1/2})|X^{(n_k)})] \\
&\quad \leq \int e^{-K\xi_{\mathsf{min}}M_n^2n_k\epsilon_{n_k}^2/(2J^2)} \prod_{i\in I_k} \mathcal{G}a\left(\xi_i;\frac{\nu}{2},\frac{\nu}{2}\right)d\xi_i + \frac{1}{C^2n_k\epsilon_{n_k}^2} \\
&\quad + \mathbf{E}_*\left[\int\int_{B_*^c(\epsilon_{n_k},n^{-1/2})} \prod_{\{i\in I_k\}} \frac{p_{g_k,\sigma_k^2}(y_i|z_i)}{p_*(y_i|z_i)} d\pi(\boldsymbol{\alpha}, g_k^{\perp}, \lambda_k, \sigma_k^2)(1-\phi_{n_k})\right] \\
&\quad \times \frac{e^{(1+C)n_k\epsilon_{n_k}^2}}{\pi(B_n^{KL}((g_j^*,\sigma_{j*}^2),\epsilon_{n_j}))}.
\end{aligned}
$$

Next, remark that

$$
\begin{aligned}
&\int e^{-K\xi_{\mathsf{min}}M_n^2n_k\epsilon_{n_k}^2/(2J^2)} \prod_{i\in I_k} \mathcal{G}a\left(\xi_i;\frac{\nu}{2},\frac{\nu}{2}\right)d\xi_i \\
&\qquad \leq \int e^{-K\xi_{\mathsf{min}}M_n^2n_k\epsilon_{n_k}^2/(2J^2)} n_k \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}(\lambda_{\mathsf{min}})^{\nu/2-1}e^{-\xi_{\mathsf{min}}(\nu/2)}d\xi_{\mathsf{min}} \\
&\qquad \leq 2e^{-Kc_qM_n^2n_k\epsilon_{n_k}^2/(2J^2)} n_k \int_{c_q}^{+\infty} \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}(\lambda_{\mathsf{min}})^{\nu/2-1}e^{-\xi_{\mathsf{min}}(\nu/2)}d\xi_{\mathsf{min}} \\
&\qquad \leq 2e^{-n_k\epsilon_{n_k}^2(Kc_qM_n^2/(2J^2)-\mathsf{log}(n_k)/(n_k\epsilon_{n_k}^2))}, \qquad \textbf{(C.6)}
\end{aligned}
$$

where, for $q\in(0,1/2)$, c_q is the q-quantile of a $\mathcal{G}a\left(\frac{\nu}{2},\frac{\nu}{2}+\frac{KM_n^2n_k\epsilon_{n_k}^2}{2J^2}\right)$. Moreover, using the result of Lemma C.1: $\pi(B_n^{KL}((g_j^*,\sigma_{j*}^2),\epsilon_{n_j})) \gtrsim e^{-n_k\epsilon_{n_k}^2}$ for $k=0,1$, and by Fubini's theorem

$$
\mathbf{E}_*[\pi(B_*^c(\epsilon_{n_k},n^{-1/2})|X^{(n_k)})] \lesssim e^{-n_k\epsilon_{n_k}^2(Kc_qM_n/(2J)-\epsilon_{n_k})} + \frac{1}{C^2n_k\epsilon_{n_k}^2} + e^{(1+C+1)n_k\epsilon_{n_k}^2} \times
$$

$$\int\int_{B_*^c(\epsilon_{n_k},n^{-1/2})}\int \underbrace{\prod_{\{i\in I_k\}}\mathcal{N}(y_i|z_i;\boldsymbol{\alpha},g_0^{\perp},\sigma_k^2/\xi_i)}_{=Q^{(n_k)}_{g_k,\sigma_k^2}}(1-\phi_{n_k})dy_i d\pi(\boldsymbol{\alpha},g_k^{\perp},\lambda_k,\sigma_k^2)$$

$$\times\prod_{i\in I_k}\mathcal{G}a\left(\xi_i;\frac{\nu}{2},\frac{\nu}{2}\right)d\xi_i. \qquad \textbf{(C.7)}$$

Set for $k=0,1$ and for $j\in\mathbb{N}$:

$$B_{k,j}\triangleq\{(g_k,\sigma_k^2)\in\mathcal{C}_{n,k}\times\left[\frac{1}{2n_k},e^{3n_k\epsilon_{n_k}^2}\right];M\epsilon_{n_k}j<\|\Xi_k^{1/2}(g_k-g_{k*})\|_n\le 2jM\epsilon_{n_k},$$

$$j\varepsilon_\sigma<|1-\sigma_k/\sigma_{k*}|<2j\varepsilon_\sigma\}$$

and $\mathcal{G}_{k,n}\triangleq\{(g_k,\sigma_k^2)\in\mathcal{C}_{n,k}\times\left[\frac{1}{2n_k},e^{3n_k\epsilon_{n_k}^2}\right];\|g_k-g_{k*}\|_n$

$$\ge M_n\epsilon_{n_k},|1-\sigma_k/\sigma_{k*}|>\widetilde{M}_n/\sqrt{n}\}$$

$$\subseteq\left\{(g_k,\sigma_k^2)\in\mathcal{C}_{n,k}\times\left[\frac{1}{2n_k},e^{3n_k\epsilon_{n_k}^2}\right];\|\Xi_k^{1/2}(g_j-g_j^*)\|_n\right.$$

$$\left.\ge\xi_{\min}^{1/2}M_n\epsilon_{n_k},\ |1-\sigma_k/\sigma_{k*}|>\widetilde{M}_n/\sqrt{n}\right\},$$

then $B_*^c(\epsilon_{n_k},n^{-1/2})\subseteq(\mathcal{G}_k\times\mathbb{R}_+)\backslash\left(\mathcal{C}_{n,k}\times\left[\frac{1}{2n_k},e^{3n_k\epsilon_{n_k}^2}\right]\right)\cup\mathcal{G}_{k,n}$ and $\mathcal{G}_{k,n}\subseteq\bigcup_{j\ge J}B_{k,j}$ for $\xi_{\min}^{1/2}M_n=JM$ and $\widetilde{M}/\sqrt{n_k}=J\varepsilon_\sigma$. Therefore, by decomposing the integral over $B_*^c(\epsilon_{n_k},n^{-1/2})$ in the sum of two integrals over the ranges $(\mathcal{G}_k\times\mathbb{R}_+)\backslash\left(\mathcal{C}_{n,k}\times\left[\frac{1}{2n_k},e^{3n_k\epsilon_{n_k}^2}\right]\right)$ and $\mathcal{G}_{k,n}$ and by upper bounding $(1-\phi_{n_k})$ by 1 over $\mathcal{G}_k\backslash\mathcal{C}_{n,k}$, for $k=0,1$, we get:

$$\int\int_{B_*^c(\epsilon_{n_k},n^{-1/2})}\int Q^{(n_k)}_{g_k,\sigma_k^2}(1-\phi_{n_k})d\pi(\boldsymbol{\alpha},g_k^{\perp},\lambda_k,\sigma_k^2)\prod_{i\in I_k}\mathcal{G}a\left(\xi_i;\frac{\nu}{2},\frac{\nu}{2}\right)d\xi_i$$

$$\le\pi(\mathcal{G}_k\backslash\mathcal{C}_{n,k})\int_{(2n_k)^{-1}}^{e^{3n_k\epsilon_{n_k}^2}}\pi(\sigma_k^2)d\sigma_k^2+\underbrace{\pi(\mathcal{G}_k)}_{\le 1}\left(\int_0^{(2n_k)^{-1}}\pi(\sigma_k^2)d\sigma_k^2+\int_{e^{3n_k\epsilon_{n_k}^2}}^{+\infty}\pi(\sigma_k^2)d\sigma_k^2\right)$$

$$+\int\int_{\mathcal{G}_{k,n}}\int Q^{(n_k)}_{g_k,\sigma_k^2}(1-\phi_{n_k})d\pi(\boldsymbol{\alpha},g_k^{\perp},\lambda_k,\sigma_k^2)\prod_{i\in I_k}\mathcal{G}a\left(\xi_i;\frac{\nu}{2},\frac{\nu}{2}\right)d\xi_i$$

$$\le e^{-n_k\epsilon_{n_k}^2\eta/(2\delta+1)}+e^{-n_k\epsilon_{n_k}^2\delta_{00}}+\frac{e^{-3n_k\epsilon_{n_k}^2\delta_{00}}}{\nu_{00}-2}$$

$$+\int\sum_{j\ge J}e^{-Kj^2M^2n_k\epsilon_{n_k}^2}\prod_{i\in I_k}\mathcal{G}a\left(\xi_i;\frac{\nu}{2},\frac{\nu}{2}\right)d\xi_i, \qquad \textbf{(C.8)}$$

where we have used Lemma E.2 to upper bound $\pi(\mathcal{G}_k \backslash \mathcal{C}_{n,k})$, the concentration inequality for sub-Gamma random variables to upper bound the integral over $\left[0, \frac{1}{2n_k}\right]$ (since $(\sigma_k^{-2} - \mathbf{E}[\sigma_k^{-2}])$ is sub-Gamma $\left(2\frac{\nu_{00}}{\delta_{00}^2}, \frac{2}{\delta_{00}}\right)$), and the second inequality in (C.5) with $JM = \xi_{\mathsf{min}}^{1/2} M_n$ to control the term $Q_{g_k,\sigma_k^2}^{(n_k)}(1-\phi_{n_k})$. Using the same argument to get (C.6), we obtain $\mathbf{E}[e^{-KJ^2M^2 n_k \epsilon_{n_k}^2}] \leq 2\,\mathsf{exp}\{-n_k\epsilon_{n_k}^2(Kc_qM_n^2 - \epsilon_{n_k})\}$. By putting this, (C.7) and (C.8) together, we get that for $k = 0, 1$:

$$\mathbf{E}_*[\pi(B_*^c(\epsilon_{n_k}, n^{-1/2})|X^{(n_k)})] \lesssim e^{-n_k\epsilon_{n_k}^2(Kc_qM_n^2/(2J^2)-\epsilon_{n_k})} + \frac{1}{C^2 n_k \epsilon_{n_k}^2} + e^{(1+C+1)n_k\epsilon_{n_k}^2}$$
$$\times \left(e^{-n_k\epsilon_{n_k}^2 \eta/(2\delta+1)} + e^{-n_k\epsilon_{n_k}^2 \delta_{00}} + e^{-3n_k\epsilon_{n_k}^2} \frac{\delta_{00}}{\nu_{00}-2} + e^{-n_k\epsilon_{n_k}^2(Kc_qM_n^2 - \epsilon_{n_k})}\right),$$

which converges to zero for every M_n, K sufficiently large and every $0 < C < \mathsf{min}\{1, \delta_{00} - 2\}$ such that $1 + C + 1 < \frac{\eta}{(2\delta+1)}$ which implies $\eta > \mathsf{min}\{3, \delta_{00}\}(2\delta + 1)$. This establishes the statement of the theorem.

C.2.2. *Proof of Theorem 2.2.* For every n, denote $X^{(n)} \triangleq (y^{(n)}, z^{(n)})$. For given $m_0, m_1 \in \mathbb{N}$, define the functional spaces $\mathcal{G}_{m_0}$ and $\mathcal{G}_{m_1}$ as

$$\mathcal{G}_{m_0} \triangleq \left\{g_{m_0}(z);\, g_{m_0}(z) = \boldsymbol{B}_0(z)'\boldsymbol{\alpha},\, z \in \{z_1, \ldots, z_{n_0}\}, \boldsymbol{\alpha} \in \mathbb{R}^{m_0}\right\} \subset L_2(P_{n_0})$$
$$\mathcal{G}_{m_1} \triangleq \left\{g_{m_1}(z);\, g_{m_1}(z) = \boldsymbol{B}_1(z)'\boldsymbol{\beta},\, z \in \{z_{n_0+1}, \ldots, z_{n_1}\}, \boldsymbol{\beta} \in \mathbb{R}^{m_1}\right\} \subset L_2(P_{n_1}),$$

where P_{n_j} has support $z^{(n_j)}$ for $j = 0, 1$. Associated to each space (and then to each measure P_{n_j}), define the linear functionals

$$L_{z^{(n_0)}} : \mathcal{G}_{m_0} \to \mathbb{R}$$
$$g \mapsto \boldsymbol{e}'_{m_0,m_0}\mathbb{B}_0^{-1}\frac{1}{n_0}\sum_{i=1}^{n_0}\boldsymbol{B}_0(z_i)g(z_i)I[z_i < \tau],$$
$$L_{z^{(n_1)}} : \mathcal{G}_{m_1} \to \mathbb{R}$$
$$g \mapsto \boldsymbol{e}'_{m_1,1}\mathbb{B}_1^{-1}\frac{1}{n_1}\sum_{i=n_0+1}^{n_1}\boldsymbol{B}_1(z_i)g(z_i)I[z_i \geq \tau],$$

where $\boldsymbol{e}_{i,j}$ denotes the $(i \times 1)$ canonical vector with all components equal to zero but the jth one, $\mathbb{B}_0 \triangleq \frac{1}{n_0}\sum_{i=1}^{n_0}\boldsymbol{B}_0(z_i)\boldsymbol{B}_0(z_i)'I[z_i < \tau]$ and $\mathbb{B}_1 \triangleq \frac{1}{n_1}\sum_{i=n_0+1}^{n_1}\boldsymbol{B}_1(z_i)\boldsymbol{B}_1(z_i)'I[z_i \geq \tau]$. Therefore, for every $g_{m_0} \in \mathcal{G}_{m_0}$, $g_{m_1} \in \mathcal{G}_{m_1}$, $\boldsymbol{\alpha}_{[m_0]} = L_{z^{(n_0)}}g_{m_0}$, and $\boldsymbol{\beta}_{[1]} = L_{z^{(n_1)}}g_{m_1}$. The linear functionals $L_{z^{(n_0)}}$ and $L_{z^{(n_1)}}$ can be written: $\forall\varphi_0 \in L_2(P_{n_0})$ and $\forall\varphi_1 \in L_2(P_{n_1})$,

$$L_{z^{(n_0)}}\varphi_0 = \langle\varphi_0, \ell_{z^{(n_0)}}\rangle_{P_{n_0}}, \qquad L_{z^{(n_1)}}\varphi_1 = \langle\varphi_1, \ell_{z^{(n_1)}}\rangle_{P_{n_1}},$$

where for $j = 0,1$, $\langle\cdot,\cdot\rangle_{P_{n_j}}$ (resp. $\|\cdot\|_{n_j}$) denotes the scalar product (resp. the induced norm) in $L_2(P_{n_j})$ and

$$\ell_{z^{(n_0)}}(s) \triangleq \boldsymbol{e}'_{m_0,m_0}\mathbb{B}_0^{-1}\boldsymbol{B}_0(z)I[z < \tau] \in L_2(P_{n_0}),$$
$$\ell_{z^{(n_1)}}(s) \triangleq \boldsymbol{e}'_{m_1,1}\mathbb{B}_1^{-1}\boldsymbol{B}_1(z)I[z \geq \tau] \in L_2(P_{n_1}).$$

Therefore, by the Riesz theorem, $\|L_{z^{(n_0)}}\|_{n_0}^2 = \|\ell_{z^{(n_0)}}\|_{n_0}^2 = e'_{m_0,m_0}\mathbb{B}_0^{-1}e_{m_0,m_0}$ and $\|L_{z^{(n_1)}}\|_{n_1}^2 = \|\ell_{z^{(n_1)}}\|_{n_1}^2 = e'_{m_1,1}\mathbb{B}_0^{-1}e_{m_1,1}$. By the definition of $\boldsymbol{B}_j(z)$, $j = 0,1$, there exists a constant $c_j > 0$ such that $\|\ell_{z^{(n_j)}}\|_{n_j} \leq c_j < \infty$ for every n_j.

Since $\left|\text{ATE} - \text{ATE}^*\right| \triangleq \left|(\boldsymbol{\beta}_{[1]} - \boldsymbol{\alpha}_{[m_0]}) - (\boldsymbol{\beta}^*_{[1]} - \boldsymbol{\alpha}^*_{[m_0]})\right| \leq |\boldsymbol{\beta}_{[1]} - \boldsymbol{\beta}^*_{[1]}| + |\boldsymbol{\alpha}_{[m_0]} - \boldsymbol{\alpha}^*_{[m_0]}|$ it holds that (by denoting with Σ the event $\{|1 - \sigma_j/\sigma_{j*}| \geq \widetilde{M}_n/\sqrt{n_j}$ for $j = 0,1\}$)

$$\begin{aligned}
&\pi\left(|\text{ATE} - \text{ATE}^*| \geq M_{1,n}\epsilon_n,\ \underbrace{|1 - \sigma_j/\sigma_{j*}| \geq \widetilde{M}_n/\sqrt{n_j} \text{ for } j = 0,1}_{\triangleq\Sigma}\,\middle|\, X^{(n)}\right)\\
&\quad \leq \pi\left(|\boldsymbol{\beta}_{[1]} - \boldsymbol{\beta}^*_{[1]}| + |\boldsymbol{\alpha}_{[m_0]} - \boldsymbol{\alpha}^*_{[m_0]}| \geq M_{1,n}\epsilon_n,\ \Sigma, |\boldsymbol{\beta}_{[1]} - \boldsymbol{\beta}^*_{[1]}| \geq \frac{M_{1,n}\epsilon_n}{2}\,\middle|\, X^{(n)}\right)\\
&\qquad + \pi\left(|\boldsymbol{\beta}_{[1]} - \boldsymbol{\beta}^*_{[1]}| + |\boldsymbol{\alpha}_{[m_0]} - \boldsymbol{\alpha}^*_{[m_0]}| \geq M_{1,n}\epsilon_n,\ \Sigma, |\boldsymbol{\beta}_{[1]} - \boldsymbol{\beta}^*_{[1]}| < \frac{M_{1,n}\epsilon_n}{2}\,\middle|\, X^{(n)}\right).
\end{aligned} \tag{C.9}$$

We analyze the two terms in the right-hand side of (C.9) separately by starting from the first one. Since for two events A and B, $\pi(A \cap B|X^{(n)}) \leq \pi(B|X^{(n)})$ and using the independence of the posterior of (σ_0^2) and $(\boldsymbol{\beta}_{[1]}, \sigma_1^2)$ we get:

$$\begin{aligned}
&\pi\left(|\boldsymbol{\beta}_{[1]} - \boldsymbol{\beta}^*_{[1]}| + |\boldsymbol{\alpha}_{[m_0]} - \boldsymbol{\alpha}^*_{[m_0]}| \geq M_{1,n}\epsilon_n,\ \Sigma, |\boldsymbol{\beta}_{[1]} - \boldsymbol{\beta}^*_{[1]}| \geq \frac{M_{1,n}\epsilon_n}{2}\,\middle|\, X^{(n)}\right)\\
&\qquad \leq \pi\left(\Sigma, |\boldsymbol{\beta}_{[1]} - \boldsymbol{\beta}^*_{[1]}| \geq \frac{M_{1,n}\epsilon_n}{2}\,\middle|\, X^{(n)}\right)\\
&\qquad \leq \pi\left(\|L_{z^{(n_1)}}\|_{n_1}\|(g_{m_1} - g^*_{m_1})\|_n \geq \frac{M_{1,n}\epsilon_{n_1}}{2},\ \left|1 - \frac{\sigma_1}{\sigma_{1*}}\right| \geq \frac{\widetilde{M}_n}{\sqrt{n_j}}\,\middle|\, X^{(n_1)}\right),
\end{aligned} \tag{C.10}$$

which converges to zero in $P_*^{(n_1)}$-probability by the result of Theorem 2.1 with $M_n = M_{1,n}/(2c_1)$ and where we have used $\|L_{z^{(n_1)}}\|_{n_1} \leq c_1 < \infty$. We now analyze the second term on the right-hand side of (C.9). Using the independence of the posterior of $(\boldsymbol{\alpha}_{[m_0]}, \sigma_0^2)$ and $(\boldsymbol{\beta}_{[1]}, \sigma_1^2)$ we get:

$$\begin{aligned}
&\pi\left(|\boldsymbol{\beta}_{[1]} - \boldsymbol{\beta}^*_{[1]}| + |\boldsymbol{\alpha}_{[m_0]} - \boldsymbol{\alpha}^*_{[m_0]}| \geq M_{1,n}\epsilon_n,\ \Sigma, |\boldsymbol{\beta}_{[1]} - \boldsymbol{\beta}^*_{[1]}| < \frac{M_{1,n}\epsilon_n}{2}\,\middle|\, X^{(n)}\right)\\
&\quad \leq \pi\left(\frac{M_{1,n}\epsilon_n}{2} + |\boldsymbol{\alpha}_{[m_0]} - \boldsymbol{\alpha}^*_{[m_0]}| \geq M_{1,n}\epsilon_n,\ \Sigma, |\boldsymbol{\beta}_{[1]} - \boldsymbol{\beta}^*_{[1]}| < \frac{M_{1,n}\epsilon_n}{2}\,\middle|\, X^{(n)}\right)
\end{aligned}$$

$$
\begin{aligned}
&\le \pi\left(|\boldsymbol{\alpha}_{[m_0]} - \boldsymbol{\alpha}^*_{[m_0]}| \ge \frac{M_{1,n}\epsilon_n}{2}, \left|1-\frac{\sigma_0}{\sigma_{0*}}\right| \ge \widetilde{M}_n/\sqrt{n_j}\,\middle|\, X^{(n_0)}\right)\\
&\qquad\qquad \underbrace{\times\, \pi\left(|\boldsymbol{\beta}_{[1]} - \boldsymbol{\beta}^*_{[1]}| < \frac{M_{1,n}\epsilon_n}{2}, \left|1-\frac{\sigma_1}{\sigma_{1*}}\right| \ge \widetilde{M}_n/\sqrt{n_j}\,\middle|\, X^{(n_1)}\right)}_{\le 1}\\
&\le \pi\left(|L_{z^{(n_0)}}(g_{m_0} - g^*_{m_0})| \ge \frac{M_{1,n}\epsilon_{n_0}}{2}, \left|1-\frac{\sigma_0}{\sigma_{0*}}\right| \ge \widetilde{M}_n/\sqrt{n_j}\,\middle|\, X^{(n_0)}\right)\\
&\le \pi\left(\|L_{z^{(n_0)}}\|_n \|(g_{m_0} - g^*_{m_0})\|_n \ge \frac{M_{1,n}\epsilon_{n_0}}{2}, \left|1-\frac{\sigma_0}{\sigma_{0*}}\right| \ge \widetilde{M}_n/\sqrt{n_j}\,\middle|\, X^{(n_0)}\right),
\end{aligned}
\tag{C.11}
$$

where we have used the independence of the posterior of $(\boldsymbol{\alpha}_{[m_0]}, \sigma_0^2)$ and $(\boldsymbol{\beta}_{[1]}, \sigma_1^2)$ to get the second inequality. By the result of Theorem 2.1 with $M_n = M_{1,n}/(2c_0)$, (C.11) converges to zero in $P_*^{(n_0)}$-probability. By putting together (C.9)–(C.11), we get the statement of the theorem.

C.3. Proofs for Section 4

C.3.1. *Proof of Theorem 4.2.* For every n, denote $X^{(n)} \triangleq (y^{(n)}, x^{(n)}, z^{(n)})$. For given $m_0, m_n \in \mathbb{N}$ and for every $s \in \{c,n\}$, let $\boldsymbol{\alpha}_s \triangleq \boldsymbol{\alpha} I[s=c] + \boldsymbol{\alpha}_n I[s=n]$, $\boldsymbol{\alpha} \in \mathbb{R}^{m_0}$, $\boldsymbol{\alpha}_n \in \mathbb{R}^{m_n}$, and $\forall z$: $\boldsymbol{B}_{0s}(z) \triangleq \boldsymbol{B}_{00}(z) I[s=c] + \boldsymbol{B}_{00,n}(z) I[s=n]$, with $\boldsymbol{B}_{00}(z) \in \mathbb{R}^{m_0}$ and $\boldsymbol{B}_{00,n}(z) \in \mathbb{R}^{m_n}$. For given $m_0, m_n, m_a, m_1 \in \mathbb{N}$, define the following functional spaces

$$
\begin{aligned}
\mathcal{G}_{m_0,m_n} &\triangleq \left\{ g_{m_0,m_n}(s,z;\boldsymbol{\alpha},\boldsymbol{\alpha}_n) = \boldsymbol{B}_{0s}(z)'\boldsymbol{\alpha}_s : \{c,n\} \times \mathbf{z_{00}} \to \mathbb{R},\ \boldsymbol{\alpha} \in \mathbb{R}^{m_0}, \boldsymbol{\alpha}_n \in \mathbb{R}^{m_n} \right\},\\
\mathcal{G}_{m_n} &\triangleq \left\{ g_{m_n}(z;\boldsymbol{\alpha}_n) = \boldsymbol{B}_{10,n}(z)'\boldsymbol{\alpha}_n : \mathbf{z_{10}} \to \mathbb{R},\ \boldsymbol{\alpha}_n \in \mathbb{R}^{m_n} \right\},\\
\mathcal{G}_{m_a} &\triangleq \left\{ g_{m_a}(z;\boldsymbol{\beta}_a) = \boldsymbol{B}_{01,a}(z)'\boldsymbol{\beta}_a : \mathbf{z_{01}} \to \mathbb{R},\ \boldsymbol{\beta}_n \in \mathbb{R}^{m_a} \right\},\\
\mathcal{G}_{m_1,m_a} &\triangleq \left\{ g_{m_1,m_a}(s,z;\boldsymbol{\beta},\boldsymbol{\beta}_a) = \boldsymbol{B}_{1s}(z)'\boldsymbol{\beta}_s : \{c,a\} \times \mathbf{z_{11}} \to \mathbb{R},\ \boldsymbol{\beta} \in \mathbb{R}^{m_1}, \boldsymbol{\beta}_a \in \mathbb{R}^{m_a} \right\},
\end{aligned}
$$

where for every $s \in \{c,a\}$, $\boldsymbol{\beta}_s \triangleq \boldsymbol{\beta} I[s=c] + \boldsymbol{\beta}_a I[s=a]$, $\boldsymbol{\beta} \in \mathbb{R}^{m_1}$, $\boldsymbol{\beta}_a \in \mathbb{R}^{m_a}$, and $\boldsymbol{B}_{1s}(z) \triangleq \boldsymbol{B}_{11}(z) I[s=c] + \boldsymbol{B}_{11,a}(z) I[s=a]$, with $\boldsymbol{B}_{11}(z) \in \mathbb{R}^{m_1}$, $\forall z$, and $\boldsymbol{B}_{11,a}(z) \in \mathbb{R}^{m_a}$, $\forall z$. Moreover, for $\boldsymbol{m} \triangleq (m_0, m_n, m_a, m_1)$ define

$$
\begin{aligned}
\mathcal{F}_{\boldsymbol{m}} \triangleq \Big\{ f(s,z) = &\Big(g_{m_0,m_n}(s,z) I[s \in \{c,n\}] + g_{m_a}(z) I[s=a] \Big) I[z<\tau]\\
&+ \Big(g_{m_1,m_a}(s,z) I[s \in \{c,a\}] + g_{m_n}(z) I[s=n] \Big) I[z \ge \tau],\ g_{m_0,m_n}(s,z)\\
&\in \mathcal{G}_{m_0,m_n}, g_{m_a}(z) \in \mathcal{G}_{m_a}, g_{m_1,m_a}(s,z) \in \mathcal{G}_{m_1,m_a}, g_{m_n} \in \mathcal{G}_{m_n} \Big\}
\end{aligned}
$$

and for every sequence $\boldsymbol{s}$, $\mathcal{F}_{\boldsymbol{m}} \subset L_2(P_{n,\boldsymbol{s}})$ where $P_{n,\boldsymbol{s}}$ has support $(s_1,z_1), \ldots, (s_{n_{00}}, z_{n_{00}})$, $(n, z_{n_{00}+1}), \ldots, (n, z_{\bar{n}_{10}})$, $(a, z_{\bar{n}_{10}+1}), \ldots, (a, z_{\bar{n}_{01}})$, $(s_{\bar{n}_{01}+1}, z_{\bar{n}_{01}+1}), \ldots, (s_n, z_n)$. Associated to every sequence $\boldsymbol{s}$, we also define two linear functionals: $L_{0\boldsymbol{s}} : \mathcal{F}_{\boldsymbol{m}} \to \mathbb{R}$ and $L_{1\boldsymbol{s}}$:

$\mathcal{F}_{\boldsymbol{m}} \to \mathbb{R}$ as

$$\begin{aligned}
L_{0\boldsymbol{s}}f &\mapsto \boldsymbol{e}'_{m_0,m_0}\mathbb{B}_{00}^{-1}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{B}_{00}(z_i)f(s_i,z_i)I[s_i=c]I[z_i<\tau]\\
&=\boldsymbol{e}'_{m_0,m_0}\frac{\sharp\{i;s_i=c,z_i<\tau\}}{n}\mathbb{B}_{00}^{-1}\frac{\sum_{i=1}^{n}\boldsymbol{B}_{00}(z_i)g_{m_0,m_n}(s_i,z_i)I[s_i=c]I[z_i<\tau]}{\sharp\{i;s_i=c,z_i<\tau\}},\\
L_{1\boldsymbol{s}}f &\mapsto \boldsymbol{e}'_{m_1,1}\mathbb{B}_{11}^{-1}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{B}_{11}(z_i)f(s_i,z_i)I[s_i=c]I[z_i\geq\tau]\\
&=\boldsymbol{e}'_{m_1,1}\frac{\sharp\{i;s_i=c,z_i\geq\tau\}}{n}\mathbb{B}_{11}^{-1}\frac{\sum_{i=1}^{n}\boldsymbol{B}_{11}(z_i)g_{m_1,m_a}(s_i,z_i)I[s_i=c]I[z_i\geq\tau]}{\sharp\{i;s_i=c,z_i\geq\tau\}},
\end{aligned}$$

where $\boldsymbol{e}_{i,j}$ denotes the $(i\times 1)$ canonical vector with all components equal to zero but the jth one, $\mathbb{B}_{00}\triangleq\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{B}_{00}(z_i)\boldsymbol{B}_{00}(z_i)'I[s_i=c]I[z_i<\tau]=P_{n,\boldsymbol{s}}\boldsymbol{B}_{00}(z_i)\boldsymbol{B}_{00}(z_i)'I[s_i=c]I[z_i<\tau]$and $\mathbb{B}_{11}\triangleq\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{B}_{11}(z_i)\boldsymbol{B}_{11}(z_i)'I[z_i=c]I[z_i\geq\tau]$. Therefore, for every $f\in\mathcal{F}_{\boldsymbol{m}}$ and every sequence $\boldsymbol{s}$, $\boldsymbol{\alpha}_{[m_0]}=L_{0\boldsymbol{s}}f$ and $\boldsymbol{\beta}_{[1]}=L_{1\boldsymbol{s}}f$. The linear functionals $L_{0\boldsymbol{s}}$ and $L_{1\boldsymbol{s}}$ can be written: $\forall\varphi\in L_2(P_{n,\boldsymbol{s}})$, $L_{0\boldsymbol{s}}\varphi=\langle\varphi,\ell_{0\boldsymbol{s}}\rangle_{P_{n,\boldsymbol{s}}}$, $L_{1\boldsymbol{s}}\varphi=\langle\varphi,\ell_{1\boldsymbol{s}}\rangle_{P_{n,\boldsymbol{s}}}$, where for every sequence $\boldsymbol{s}$, $\langle\cdot,\cdot\rangle_{P_{n,\boldsymbol{s}}}$ (resp. $\|\cdot\|_{n,\boldsymbol{s}}$) denotes the scalar product (resp. the induced norm) in $L_2(P_{n,\boldsymbol{s}})$ and

$$\begin{aligned}
\ell_{0\boldsymbol{s}}(s,z)&\triangleq\boldsymbol{e}'_{m_0,m_0}\mathbb{B}_{00}^{-1}\boldsymbol{B}_{00}(z)I[s=c]I[z<\tau],\\
\ell_{1\boldsymbol{s}}(s,z)&\triangleq\boldsymbol{e}'_{m_1,1}\mathbb{B}_{11}^{-1}\boldsymbol{B}_{11}(z)f(s,z)I[s=c]I[z\geq\tau].
\end{aligned}$$

Therefore, by the Riesz theorem, for $j=0,1$, $\|L_{j\boldsymbol{s}}\|_{n,\boldsymbol{s}}^2=\|\ell_{j\boldsymbol{s}}\|_{n,\boldsymbol{s}}^2$ and $\|\ell_{0\boldsymbol{s}}\|_{n,\boldsymbol{s}}^2=e'_{m_0,m_0}\mathbb{B}_{00}^{-1}e_{m_0,m_0}I[s=c]I[z<\tau]$, $\|\ell_{1\boldsymbol{s}}\|_{n,\boldsymbol{s}}^2=e'_{m_1,1}\mathbb{B}_{11}^{-1}e_{m_1,1}I[s=c]I[z\geq\tau]$. By the definition of $\boldsymbol{B}_{jj}(z)$, $j=0,1$, there exists a constant $c_j>0$ such that $\|\ell_{j\boldsymbol{s}}\|_{n,\boldsymbol{s}}\leq c_j<\infty$ for every n and for every $\boldsymbol{s}$.

Since $\left|\text{CATE}-\text{CATE}^*\right|\triangleq\left|(\boldsymbol{\beta}_{[1]}-\boldsymbol{\alpha}_{[m_0]})-(\boldsymbol{\beta}^*_{[1]}-\boldsymbol{\alpha}^*_{[m_0]})\right|\leq|\boldsymbol{\beta}_{[1]}-\boldsymbol{\beta}^*_{[1]}|+|\boldsymbol{\alpha}_{[m_0]}-\boldsymbol{\alpha}^*_{[m_0]}|$ it holds that (by denoting with Σ the event $\{|1-\sigma_j/\sigma_{j*}|\geq\widetilde{M}_n/\sqrt{n}$ for $j=0,1,0n,1a\}$)

$$\begin{aligned}
&\pi\left(|\text{CATE}-\text{CATE}^*|\geq M_{1,n}\epsilon_n,\ \Sigma\middle|X^{(n)}\right)\\
&\quad\leq\pi\left(|\boldsymbol{\beta}_{[1]}-\boldsymbol{\beta}^*_{[1]}|+|\boldsymbol{\alpha}_{[m_0]}-\boldsymbol{\alpha}^*_{[m_0]}|\geq M_{1,n}\epsilon_n,\ \Sigma\middle|X^{(n)}\right)\\
&\quad\leq\pi\left(\forall\boldsymbol{s},\forall f\in\mathcal{F}_{\boldsymbol{m}};\ |L_{1\boldsymbol{s}}(f-f^*)|+|L_{0\boldsymbol{s}}(f-f^*)|\geq M_{1,n}\epsilon_n,\ \Sigma\middle|X^{(n)}\right)\\
&\quad\leq\pi\left(\forall\boldsymbol{s},\forall f\in\mathcal{F}_{\boldsymbol{m}};\ \left(\|L_{1\boldsymbol{s}}\|_{n,\boldsymbol{s}}+\|L_{0\boldsymbol{s}}\|_{n,\boldsymbol{s}}\right)\|f-f^*\|_{n,\boldsymbol{s}}\geq M_{1,n}\epsilon_n,\ \Sigma\middle|X^{(n)}\right).\qquad\textbf{(C.12)}
\end{aligned}$$

Remark that $\forall f\in\mathcal{F}_{\boldsymbol{m}}$, for $f^*\in\mathcal{F}_{\boldsymbol{m}}$ and for every given $\boldsymbol{s}=(\boldsymbol{s_0}',\boldsymbol{s_1})'$ with $\boldsymbol{s_0}\triangleq(s_1,\ldots,s_{n_{00}})$, $\boldsymbol{s_1}\triangleq(s_{\bar{n}_{01}+1},\ldots,s_n)$:

$$\begin{aligned}
&\|f-f^*\|_{n,\boldsymbol{s}}^2\leq\\
&\frac{1}{n}\sum_{i\in I_{00}}\left[2(g_{0s_i}(z_i)-g^*_{0s_i}(z_i))^2+4(g_{m_0,m_n}(s_i,z_i)-g_{0s_i}(z_i))^2+4(g^*_{0s_i}(z_i)-g^*_{m_0,m_n}(s_i,z_i))^2\right]
\end{aligned}$$

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

$$+\sum_{j\in\{0n,1a\}}\frac{1}{n}\sum_{i\in I_j}\Big[2(g_j(z_i)-g_j^*(z_i))^2+4(g_{m_j}(z_i)-g_j(z_i))^2+4(g_j^*(z_i)-g^*_{m_j}(z_i))^2\Big]$$

$$+\frac{1}{n}\sum_{i\in I_{11}}\Big[2(g_{1s_i}(z_i)-g^*_{1s_i}(z_i))^2+4(g_{m_1,m_a}(s_i,z_i)-g_{1s_i}(z_i))^2$$

$$+4(g^*_{1s_i}(z_i)-g^*_{m_1,m_a}(s_i,z_i))^2\Big]$$

$$=2\|\boldsymbol{f_s}-\boldsymbol{f_s^*}\|_n^2+4\|f-\boldsymbol{f_s}\|_{n,\boldsymbol{s}}^2+4\|\boldsymbol{f_s^*}-f^*\|_{n,\boldsymbol{s}}^2, \tag{C.13}$$

where $\boldsymbol{f_s}$ (resp. $\boldsymbol{f_s^*}$) is defined in (C.3) (resp. for the true value of the parameters) and remark that $f^*\in\mathcal{F}_{\boldsymbol{m}}$. Remark that the second term in the right-hand side of (C.13) has zero prior mass and, under the assumption of the theorem $\|\boldsymbol{f_s^*}-f^*\|_{n,\boldsymbol{s}}\leq C_2(\min\{m_0,m_1,m_{0n},m_{1a}\})^{-\delta}$. By plugging this in (C.12), we get:

$$\pi\left(|\mathrm{CATE}-\mathrm{CATE}^*|\geq M_{1,n}\epsilon_n,\ \Sigma\,\Big|\,X^{(n)}\right)$$

$$\leq\pi\Bigg(\forall\boldsymbol{s},\forall f\in\mathcal{F};\ \|\boldsymbol{f_s}-\boldsymbol{f_s^*}\|_n^2$$

$$\geq\frac{1}{2}\left(\frac{M_{1,n}^2\epsilon_n^2}{(c_0+c_1)^2}-4C_2^2(\min\{m_0,m_1,m_{0n},m_{1a}\})^{-2\delta}\right),\ \Sigma|X^{(n)}\Bigg)$$

$$=\pi\left(\forall\boldsymbol{s},\forall f\in\mathcal{F};\ \|\boldsymbol{f_s}-\boldsymbol{f_s^*}\|_n^2\geq\frac{\epsilon_n^2}{2}\left(\frac{M_{1,n}^2}{(c_0+c_1)^2}-4C_2^2\right),\ \Sigma\,\Big|\,X^{(n)}\right), \tag{C.14}$$

where we have set $m_0\asymp m_1\asymp m_{0n}\asymp m_{1a}\asymp(n/\log(n))^{1/(2\delta+1)}$ to get the last line. If $M_{1,n}$ is large enough such that $\frac{M_{1,n}^2}{2(c_0+c_1)^2}-2C_2^2>0$, then (C.14) converges to zero by Theorem F.1 in the Supplementary Appendix with $M_n\leq\frac{M_{1,n}^2}{2(c_0+c_1)^2}-2C_2^2$.

C.4. Technical Lemmas

The proof of the lemmas in this section is provided in Section E of the Online Supplementary Appendix.

LEMMA C.1. *Assume the conditions of Theorem 2.1 hold. Then, for* $k=0,1$ *there exists a* $N>0$ *such that* $\forall n_k\geq N$ *and* $\forall\epsilon_{n_k}>0$*:*

$$\pi(B_n^{KL}((g_k^*,\sigma_{k*}^2),\epsilon_{n_k}))\gtrsim\exp\left\{-n_k\epsilon_{n_k}^2\right\}.$$

LEMMA C.2. *Assume the conditions of Theorem 2.1 hold. Then, for* $k=0,1$ *the sequence of measurable sets* $\mathcal{C}_{n,k}$ *defined in (C.2) satisfies*

$$\pi(\mathcal{G}_k\backslash\mathcal{C}_{n,k})\lesssim\exp\left\{-n_k\epsilon_{n_k}^2\frac{\eta}{2\delta+1}\right\}, \tag{C.15}$$

where η *is defined in Theorem 2.1.*

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press

LEMMA C.3 (Testing). *For each* $k = 0,1$, *let* $q_{k*} \triangleq (g_k^*, \sigma_{k*}^2)$ *and* $\boldsymbol{E}_{q_{k*}}$ *(resp.* $\boldsymbol{E}_{q_k}$*) denote the expectation taken with respect to the distribution* $\mathcal{N}_{n_k}(g_k^*, \sigma_{k*}^2)$ *(resp.* $\mathcal{N}_{n_k}(g_k, \sigma_k^2)$*). For each* $k = 0,1$, *there exists a test* ϕ_{n_k} *such that for some* $K, M > 0$ *and* $\forall \ell \in \mathbb{N}$*:*

$$\boldsymbol{E}_{q_{k*}}\phi_{n_k} \le \frac{e^{-(M^2K-1)n_k\epsilon_{n_k}^2}}{1-e^{-M^2Kn_k\epsilon_{n_k}^2}}, \qquad \sup_{q_k=(g_k,\sigma_k^2)\in\mathcal{A}_{n,k,\ell}} \boldsymbol{E}_{q_k}(1-\phi_{n_k}) \le e^{-M^2K\ell n_k\epsilon_{n_k}^2}, \tag{C.16}$$

where $\mathcal{A}_{n,k,\ell} \triangleq \{g_k \in \mathcal{C}_{n,k}; \|\Xi_k^{1/2}(g_k - g_k^*)\|_{n_k} > \ell M\epsilon_{n_k}\} \times \left\{\sigma_k^2 \in \left[\frac{1}{2n_k}, e^{n_k\epsilon_{n_k}^2}\right]; |1-\sigma_k/\sigma_{k*}| > \ell\varepsilon_\sigma\right\}$ *for some* $\varepsilon_\sigma > 0$, $\Xi_0 \triangleq \text{diagonal}(\xi_1, \ldots, \xi_{n_0})$, $\Xi_1 \triangleq \text{diagonal}(\xi_{n_0+1}, \ldots, \xi_{n_1})$ *and* $\{\xi_i\}_i$ *are the latent variables in the mixture representation of the student-t distribution.*

Supplementary Material

To view the supplementary material for this article, please visit: https://doi.org/10.1017/S0266466622000019.

REFERENCES

Antoniak, C.E. (1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics* 2, 1152–1174.

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

Bickel, P.J. & B.J.K. Kleijn (2012) The semiparametric Bernstein-von Mises theorem. *The Annals of Statistics* 40, 206–237.

Branson, Z., M. Rischard, L. Bornn, & L.W. Miratrix (2019) A nonparametric Bayesian methodology for regression discontinuity designs. *Journal of Statistical Planning and Inference* 202, 14–30.

Brezger, A. & S. Lang (2006) Generalized structured additive regression based on Bayesian *P*-splines. *Computational Statistics & Data Analysis* 50, 967–999.

Calonico, S., M.D. Cattaneo, & R. Titiunik (2014) Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica* 82, 2295–2326.

Campbell, D.T. (1969) Reforms as experiments. *American Psychologist* 24, 409–429.

Cattaneo, M.D., B.R. Frandsen, & R. Titiunik (2015) Randomization inference in the regression discontinuity design: An application to party advantages in the US senate. *Journal of Causal Inference* 3, 1–24.

Cattaneo, M.D., N. Idrobo, & R. Titiunik (2020) *A Practical Introduction to Regression Discontinuity Designs: Foundations*. Cambridge University Press.

Cattaneo, M.D., R. Titiunik, & G. Vazquez-Bare (2017) Comparing inference approaches for rd designs: A reexamination of the effect of head start on child mortality. *Journal of Policy Analysis and Management* 36, 643–681.

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

Chib, S. & E. Greenberg (1996) Markov chain Monte Carlo simulation methods in econometrics. *Econometric Theory* 12, 409–431.

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

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

Chib, S. & L. Jacobi (2016) Bayesian fuzzy regression discontinuity analysis and returns to compulsory schooling. *Journal of Applied Econometrics* 31, 1026–1047.

Ferguson, T.S. (1973) A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* 1, 209–230.

Frandsen, B.R., M. Froelich, & B. Melly (2012) Quantile treatment effects in the regression discontinuity design. *Journal of Econometrics* 168, 382–395.

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

Ghosal, S. & A. van der Vaart (2017) *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Ghosal, S. & A. van der Vaart (2007) Convergence rates of posterior distributions for non iid observations. *The Annals of Statistics* 35, 192–223.

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

Imbens, G. & K. Kalyanaraman (2012) Optimal bandwidth choice for the regression discontinuity estimator. *Review of Economic Studies* 79, 933–959.

Imbens, G.W. & T. Lemieux (2008) Regression discontinuity designs: A guide to practice. *Journal of Econometrics* 142, 615–635.

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

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

Lee, D.S. (2008) Randomized experiments from non-random selection in U.S. house elections. *Journal of Econometrics* 142, 675–697.

Lee, D.S. & T. Lemieux (2010) Regression discontinuity designs in economics. *Journal of Economic Literature* 48, 281–355.

McCrary, J. (2008) Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics* 142, 698–714.

Meyersson, E. (2014) Islamic rule and the empowerment of the poor and pious. *Econometrica* 82, 229–269.

Thistlethwaite, D.L. & D.T. Campbell (1960) Regression-discontinuity analysis: An alternative to the ex-post facto experiment. *Journal of Educational Psychology* 51, 309–317.

van der Vaart, A.W. (2000) *Asymptotic Statistics*. Lectures on Probability Theory. École d'Été de Probailités de St. Flour XX.

https://doi.org/10.1017/S0266466622000019 Published online by Cambridge University Press