Estimation of Regression Discontinuity and Kink Designs with Multiple Running Variables

Alden Cheng*

Abstract

In regression discontinuity designs with multiple running variables (MRD), units are assigned different treatments based on whether their values on several observed running variables exceed known thresholds. In such designs, applied work commonly analyzes each running variable separately, estimating a single-dimensional RD design in the first running variable after limiting the sample to the set of individuals qualifying on the second threshold, and vice versa. In this paper, I propose a new estimator for MRD designs using thin plate splines that improves upon the applied practice in two ways. First, the estimator can be used to estimate the conditional average treatment effect at every point on the boundary separating treated and untreated units, and second, it provides efficiency gains by using the entire sample. I also develop analogous estimators for multidimensional regression kink (MRK) and multidimensional regression discontinuity/kink (MRDK) designs. I establish theoretical properties for these estimators, before presenting simulation results showing that they perform well in finite samples. Finally, I demonstrate the performance of my MRD estimator with two empirical applications: Londoño-Vélez, Rodríguez, and Sánchez (2020) on the effect of financial aid on college enrollment, and Keele and Titiunik (2015) on the effect of political ads on election turnout. Open-source software is available for implementing the proposed methods.

Keywords: Regression discontinuity, regression kink, multidimensional regression discontinuity/kink, education policy, voting behavior
Jel Classification: C1, C13, C21, C25.

1 Introduction

The regression discontinuity (RD) design, first introduced by Thistlethwaite and Campbell (1960), has enjoyed a revival in popularity over the past two decades. As Lee and Lemieux

[*] Department of Economics, Massachusetts Institute of Technology, 50 Memorial Drive, Building E52, Cambridge, MA 02142 (email: alden15@nber.org). This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. I thank Alberto Abadie, Josh Angrist, David Autor, David Card, Dave Donaldson, Ismaël Castillo, Matias Cattaneo, EunYi Chung, Jonathan Cohen, Amy Finkelstein, Florian Gunsilius, Jetson Leder-Luis, Juliana Londoño-Vélez, Richard Nickl, Parag Pathak, Rahul Singh, and all participants of MIT's labor lunch for their comments and suggestions. Titles of previous versions of this paper include "Regression Discontinuity Designs with Multiple Assignment Variables", and "Regression Discontinuity Designs with Multiple Assignment Variables". The proposed methods can be implemented using the "mrdd" package on R, which is available through `https://github.com/alden1505/rd_multiple_running_var_code`. All mistakes are my own.

(2010) document, the RD design has been used in a wide range of policy evaluations, including in areas such as education, labor market programs, health, and crime. In RD designs, treatment status is determined by whether observed running variables (also known as assignment variables) exceed known thresholds. Under the assumption that the location of observations near this threshold is as-good-as-random, the treatment effect is identified as the difference in mean outcomes for observations infinitesimally close to either side of the threshold.

In recent years, the proliferation of richer data sets has led to an increase in the number of papers using RD designs with multiple running variables, which I call multidimensional RD (MRD) designs.[1] For example, Londoño-Vélez, Rodríguez, and Sánchez (2020) studies the effect of eligibility for a financial aid program in Colombia using an MRD design with students' test scores and family wealth as the running variables,[2] while Dell (2010) studies the long-run impacts of a forced mining labor system in Peru and Bolivia leveraging a geographical discontinuity in conscription rates (with longitude and latitude being the running variables).

However, while estimation of single-dimensional RD designs has been extensively studied, there is much less work studying estimation for MRD. Given the lack of guidance, most empirical papers reduce MRDs into single-dimensional RD problems, either analyzing each running variable separately or combining multiple running variables into one. For instance, Londoño-Vélez, Rodríguez, and Sánchez (2020) focus on the set of students with low enough family income and estimate a single-dimensional RD with test score as the running variable and vice versa, whereas in geographical RD problems researchers often use distance to the boundary as the running variable.[3] While these approaches typically produce valid treatment effect estimates, they do not fully exploit the richness of the data.

Hence, in this paper I propose a non-parametric MRD estimator which improves upon the common applied practice in two ways: first, the estimator can be used to estimate the conditional average treatment effect (CATE) at every point on the boundary separating treated and untreated units (which I will henceforth refer to as the treatment frontier, and denote by $\mathbb{F}$), and second, the estimator provides efficiency gains (relative to the practice of analyzing each running variable separately) by using the entire sample. At a high-level, my approach involves estimating two conditional expectation functions (CEFs) with respect to the running variables non-parametrically by fitting two thin plate splines $\hat{g}_1(x)$ and $\hat{g}_0(x)$ over the treated and untreated regions respectively. The vertical difference between the two splines $\hat{\tau}(x) =$

[1] MRDs fall under two general categories – cases with dichotomous treatments (the two treatment conditions being either treatment or control), and those with multiple treatment arms (i.e. more than two mutually exclusive treatment conditions). Throughout this paper, I will focus my discourse on the case with dichotomous treatment, but the analysis also extends straightforwardly to the case with multiple treatment arms.

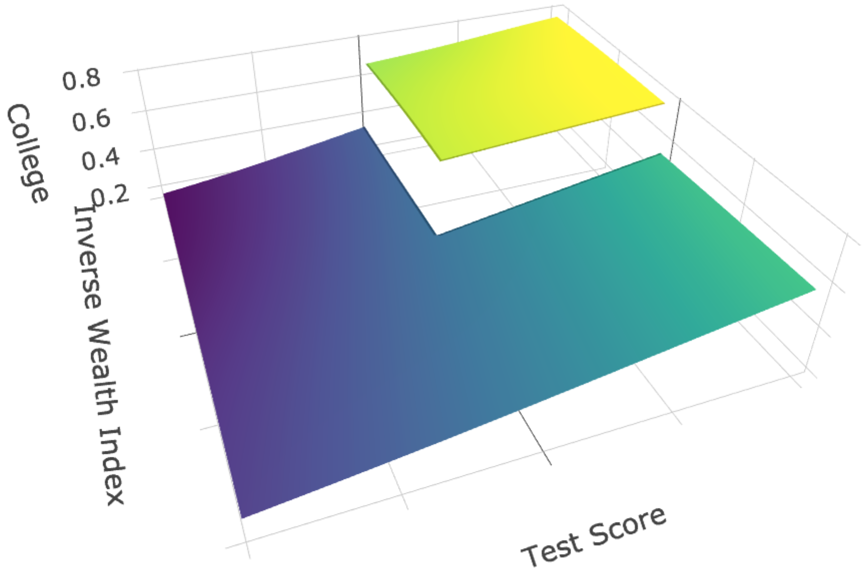
[2] Kane (2003) uses a very similar empirical strategy to study the effect of the Cal Grant program.

[3] Dell (2010) provides a rare exception by using longitude and latitude as two separate running variables. However, due to data limitations, Dell estimates the MRD using global cubic polynomial fits, an approach that is not recommended in the single-dimensional case (Gelman and Imbens 2019; Cattaneo and Titiunik 2022), and is likely to perform even worse in multiple dimensions.

$\hat{g}_1(x) - \hat{g}_0(x)$ at every point $x \in \mathbb{F}$ is then taken as the estimate of the CATE, $\tau(x)$, for individuals with values of the running variables equal to x. If desired, we can also recover the average CATE over $\mathbb{F}$ (or any subset of $\mathbb{F}$) by integrating with respect to the distribution of the running variables over $\mathbb{F}$.

Figure 1 shows a practical application of my MRD estimator using data from Londoño-Vélez, Rodríguez, and Sánchez (2020). In particular, the surfaces show the estimated probability of college enrollment as a function of the two running variables (test scores and an inverse wealth index), estimated separately for students with values of the running variables which make them eligible or ineligible for the program respectively. The vertical gap between the two surfaces corresponds to my MRD estimate of the CATE, and we observe that the effect on college enrollment seems to be decreasing in test scores. Section 4 provides more details and results for this empirical application, and discusses the economic significance of the treatment effect heterogeneity in this setting.

Figure 1: Probability of College Enrollment as a Function of Test Score and Family Wealth



Notes: The figure shows estimates of the probability of college enrollment as a function of test scores and an inverse wealth index using data from Londoño-Vélez, Rodríguez, and Sánchez (2020). Probability of college enrollment was estimated using thin plate regression splines separately for students with test scores and family wealth meeting the eligibility criteria for financial aid, and for students with test scores and family wealth that did not meet the eligibility criteria.

The popularity of the RD design also led to the development of a closely related method — the regression kink (RK) design (Card, Lee, Pei, and Weber 2015) — which is based on changes in the derivative of a continuous treatment variable at a threshold, (e.g., when the marginal tax rate increases as earnings exceed certain levels). While there has been less discussion of RK designs with multiple running variables (MRK designs), the MRD estimator discussed above can be easily extended to the MRK setting,[4] as well as to multidimensional settings where there is a discontinuity in one running variable and a kink in the other (MRDK).

Figures 2 and 3 show examples of potential MRDK and MRK designs studying the effect of UI benefits on job-finding probability, motivated by Louisiana's UI benefit schedule as described in Landais (2015). In this setting, weekly UI benefits $\mathcal{W}$ is an increasing function of prior earnings E up to a time-specific threshold $\bar{E}^t$ and is constant for higher levels of earnings, thus inducing the first kink in the figures (at $E = \bar{E}^t$). The threshold for the cap $\bar{E}^t$ is constant during the period $t < 0$, but changes after that.

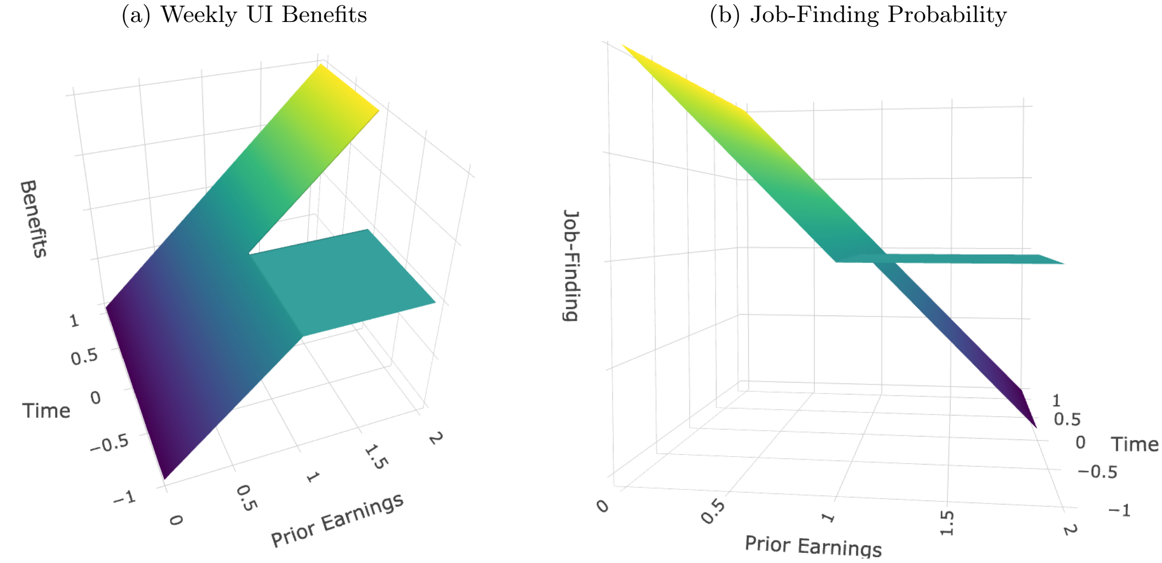
Figure 2 considers the scenario where the threshold is abruptly increased at time $t = 0$ from $\bar{E}^0$ to a much higher threshold $\bar{E}^1$. This leads to a discrete jump in weekly UI benefits for individuals with earnings above the previous cap who apply at time of the cap raise. This results in the discontinuity at $t = 0$ and $E \geq \bar{E}^0$ seen in Figure 2a, giving rise to an MRDK design.[5] Figure 3 considers instead a case where the threshold is gradually raised after time $t = 0$. This results in a second kink in Figure 3a at $t = 0$ and $E \geq \bar{E}^0$, thus leading to an MRK design.

If there is a causal relationship between UI benefits and job-finding probability, these discontinuities/kinks in the benefit schedule will result in similar discontinuities/kinks in job-finding probability. The MRDK and MRK estimands are then given by the ratio of the discontinuity/kink for job-finding shown in Figures 2b and 3b (which assume a negative causal relationship) to the discontinuity/kink for benefits (seen in Figures 2a and 3a). A more detailed description of the UI benefits schedule that give rise to these MRDK and MRK designs is given in Section 2.3.

[4] Essentially, the only additional step to estimating an MRK design compared to the procedure for a fuzzy MRD design is to take the derivatives of the estimated surfaces.

[5] In reality, this may give rise to incentives for individuals to delay the start date of their claims until after t_0. On the other hand, if eligibility for the higher cap is based on individuals' age at time t_0 (rather than being introduced for all individuals at t_0), then we can use age (at t_0) as the running variable instead of time, which may ameliorate concerns over manipulation of UI start dates.

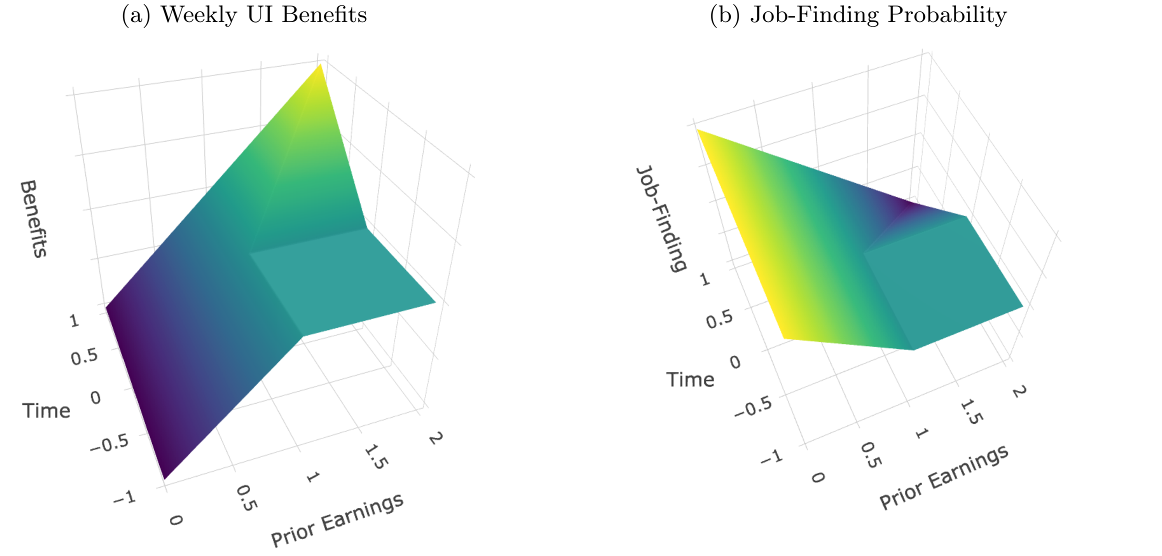
Figure 2: Example of an MRDK Design


(a) Weekly UI Benefits

(b) Job-Finding Probability

Notes: Panels A and B show the conditional expectation functions of weekly UI benefits and job-finding probability respectively as functions of prior earnings and time.

Figure 3: Example of an MRK Design


(a) Weekly UI Benefits

(b) Job-Finding Probability

Notes: Panels A and B show the conditional expectation functions of weekly UI benefits and job-finding probability respectively as functions of prior earnings and time.

While there are no papers (to the best of my knowledge) that propose an estimator for MRK designs and study its theoretical properties, there are a few papers studying MRD estimation. Closest in spirit to this paper is Zajonc (2012), who proposes a non-parametric MRD estimator based on local linear regressions. While local linear regressions (or more generally, local polynomial regressions) are popular for single-dimensional RD estimation due to their attractive boundary properties, they are more difficult to implement in multiple dimensions, and I instead consider MRD estimation using thin plate splines due to its simplicity.

In particular, to control the flexibility of thin plate splines, one only has to choose a scalar penalty parameter λ, and only a single estimation procedure is required to estimate the CATE along the entire treatment frontier $\mathbb{F}$. The ease of thin plate spline estimation allows me to implement a simple undersmoothing procedure to obtain asymptotically valid confidence intervals (CIs) for my estimators,[6] and to construct simultaneous confidence bands for the CATE function. By contrast, for multidimensional local linear regressions, the choice of a continuum of bandwidths for each $x \in \mathbb{F}$ is required,[7] and the estimation procedure needs to be repeated separately for each $x \in \mathbb{F}$ in order to estimate the CATE at that point.[8]

An alternative approach to MRD estimation which does not involve estimating CEFs of the potential outcome functions is given in Imbens and Wager (2018). Instead, they use convex numerical optimization to obtain a finite-sample-minimax linear estimator of the treatment effect subject to bounds on the second derivative of the CEF locally (for RD designs with either a single or multiple running variables). This is similar in spirit to methods in Armstrong and Kolesár (2018), and Kolesár and Rothe (2018), and requires the user to assume a reasonable bound for the curvature of an unknown function.

This paper fits more broadly into a vast literature on RD designs, excellent reviews of which can be found in Lee and Lemieux (2010), Cattaneo and Titiunik (2022), and Cattaneo, Idrobo, and Titiunik (2023). While there are fewer papers on MRD designs, several studies consider related settings: for example, Keele and Titiunik (2015), and Cattaneo, Titiunik, Vazquez-Bare, and Keele (2016) consider RD designs with a single running variable but multiple cutoffs,[9] while van Dijcke and Gunsilius (2023) study methods for MRD settings where the treatment frontier $\mathbb{F}$ is unknown. In addition, Abdulkadiroglu, Angrist, Narita, and Pathak

[6] Calonico, Cattaneo, and Titiunik (2014) derive asymptotically valid CIs for single-dimensional RD through bias correction (while accounting for the variance of the bias estimate). However, multidimensional extensions of these formulae may be complex and challenging to implement without very large data sets. There are also ad hoc methods for undersmoothing such as dividing the MSE-optimal bandwidth by half (Hall 2012) or using the minimum of the continuum of MSE-optimal bandwidths estimated along the treatment frontier (Zajonc 2012), but these methods are not theoretically justified.

[7] Methods for estimating a continuum of MSE-optimal bandwidths for multidimensional local linear regressions include Zajonc (2012), who essentially extends Imbens and Kalyanaraman's (2012) method for choosing MSE-optimal bandwidths for single-dimensional RDs to multiple dimensions, as well as a cross-validation method proposed by Papay, Willett and Murnane (2011).

[8] Perhaps reflecting the difficulty of estimating a continuum of optimal bandwidths in the multidimensional setting, in one of the few empirical papers to estimate MRD using local linear regressions, Snider and Williams (2015) choose the bandwidth in an ad hoc manner.

[9] Empirical settings where such methods may be used include Angrist and Lavy (1999), Angrist, Lavy, Leder-Luis, and Shany (2019), and Finkelstein, Hendren, and Shepard (2019).

(2022) consider a market design setting with many running variables and many cutoffs, and develop a method to estimate an average of the local average treatment effects across all cutoffs by combining RD with a local propensity score approach.

The rest of this paper proceeds as follows. Section 2 presents theoretical results on estimation and inference for my estimators, and section 3 evaluates their finite-sample performance through a simulation study. Section 4 shows my MRD estimator in action based on two empirical applications, and section 5 concludes.

2 Theory

In this section, I start by discussing identification and estimation for sharp and fuzzy MRD designs, before moving on to MRDK and MRK designs. In addition, I briefly discuss issues related to practical uses, interpretation, and implementation for these methods.

2.1 Sharp MRD Designs

In a canonical MRD design, there is a vector of running variables $X_i \in \mathbb{R}^d$ (also known as assignment variables) which determines the treatment that individual i is *assigned* to, $Z_i \in \{0, 1\}$. The treatment that i actually *takes up* is denoted by $W_i \in \{0, 1\}$, and in this subsection we will focus on the case where all individuals comply with the treatment they are assigned to (i.e., $W_i = Z_i$), which is also known as a sharp MRD design. The scenario where $W_i \neq Z_i$ for some individuals is known as a fuzzy MRD design, and will be discussed in the next subsection.

Writing treatment assignment as a function of the running variables, $Z_i = Z(X_i)$, we can denote regions in the running variable space corresponding to different treatment assignments by $\Omega_1 \equiv \{x \in \mathbb{R}^d | Z(x) = 1\}$, and $\Omega_0 \equiv \{x \in \mathbb{R}^d | Z(x) = 0\}$. In addition, I will call the boundary separating Ω_1 and Ω_0 the *treatment frontier*, and denote it by $\mathbb{F}$. In many cases of interest, treatment is assigned based on whether each of the d running variables exceed their respective thresholds (assumed to be zero here without loss of generality):[10]

$$Z_i = \prod_{k=1}^{d} \mathbb{I}\left[X_{ki} \geq 0\right], \tag{1}$$

[10] While I have described treatment assignment as determined by an "AND" condition here, this is without loss of generality in the sense that cases where the treatment is assigned by an "OR" condition (and/or one or both of the running variables have to fall below the threshold) can be transformed into the formulation above by appropriately redefining the treatment (and/or switching the sign(s) of the running variable(s)). There are certain cases where treatment assignment cannot be easily transformed into this formulation such as geographical RD settings, but the estimators proposed in this paper can typically still be applied by estimating two or more thin plate splines, and taking their difference at the boundary.

in which case, we have

$$\Omega_1 = \{x \in \mathbb{R}^d | \prod_{k=1}^{d} \mathbb{I}\left[x_k \geq 0\right] = 1\}, \quad \Omega_0 = \{x \in \mathbb{R}^d | \prod_{k=1}^{d} \mathbb{I}\left[x_k \geq 0\right] = 0\},$$

$$\mathbb{F} = \left\{ x \in \mathbb{R}^d | x_k = 0 \text{ for some } k, \text{ and } x_j \geq 0 \; \forall j \neq k \right\}.$$

Throughout this paper, I also assume that X_i has a strictly positive density in a neighborhood of every point $x \in \mathbb{F}$.

Denote the potential outcome for individual i in a world where she receives treatment w by $Y_i(w)$, and let $B^z_\epsilon(x) \equiv B_\epsilon(x) \cap \Omega_z$ where $B_\epsilon(x)$ is the open ball of radius ϵ centered at x. The key identifying assumption for sharp MRD designs is as follows.

Assumption 1. *(Continuity of Mean Potential Outcomes)* For all $x \in \mathbb{F}$ and $w \in \{0, 1\}$:

$$\lim_{\epsilon \to 0} \mathbb{E}\left[Y_i(w) | X_i = x', x' \in B^1_\epsilon(x)\right] = \lim_{\epsilon' \to 0} \mathbb{E}\left[Y_i(w) | X_i = x', x' \in B^0_{\epsilon'}(x)\right]. \tag{2}$$

Under Assumption 1, the CATE $\tau(x)$ at each $x \in \mathbb{F}$ is identified and is given by the difference between two limits of the conditional expectation function (CEF): the limit along a sequence in the treated region minus the limit along a sequence in the untreated region:

$$\begin{aligned}
\tau(x) &\equiv \mathbb{E}\left[Y_i(1) - Y_i(0) | X_i = x\right] \\
&= \lim_{\epsilon \to 0} \mathbb{E}\left[Y_i(1) | X_i = x', x' \in B^1_\epsilon(x)\right] - \lim_{\epsilon' \to 0} \mathbb{E}\left[Y_i(0) | X_i = x', x' \in B^0_{\epsilon'}(x)\right] \\
&= \lim_{\epsilon \to 0} \mathbb{E}\left[Y_i | X_i = x', x' \in B^1_\epsilon(x)\right] - \lim_{\epsilon' \to 0} \mathbb{E}\left[Y_i | X_i = x', x' \in B^0_{\epsilon'}(x)\right] \qquad (3)
\end{aligned}$$

Equation (3) suggests a natural way to estimate the treatment effect: one can simply estimate the CEFs $g_1(x) \equiv \mathbb{E}[Y_i(1)|X_i = x]$ using $X_i \in \Omega_1$ and $g_0(x) \equiv \mathbb{E}[Y_i(0)|X_i = x]$ using $X_i \in \Omega_0$, and take the difference $\hat{\tau}(x) = \hat{g}_1(x) - \hat{g}_0(x)$ as the CATE estimate at any $x \in \mathbb{F}$. If desired, one can also recover the average treatment effect over any subset of the $\mathbb{F}$ by integrating with respect to the distribution of X_i. Note that this estimation approach also extends straightforwardly to the case with more than two treatment arms, where individuals in (the more than two) different regions of the running variable space are assigned to different treatments.[11]

This estimation approach offers two advantages over the common empirical practice of analyzing each running variable separately. First, it allows us to estimate heterogeneous treatment effects $\tau(x)$ for $x \in \mathbb{F}$, and in many cases this can give rise to economically meaningful insights.[12] Second, this approach yields more precise estimates by using all of the data si-

[11] In this case, one can estimate the CEF over each region of the running variable space, and take the difference between the estimated CEFs at the boundaries separating the different regions to obtain an estimate of the relative effects between two different treatments.

[12] Estimating a single-dimensional RD separately for each running variable will instead only give us estimates

multaneously. By contrast, when reducing the MRD to a single-dimensional RD, for each estimation we are discarding a substantial fraction of observations.[13]

My choice of estimator for the CEFs $g_1(x)$ and $g_0(x)$ is motivated by several goals. First, the treatment effect estimates should not depend heavily on points far away from the treatment frontier $\mathbb{F}$ (Gelman and Imbens 2019; Cattaneo and Titiunik 2022). Second, the estimator should be regularized in order to avoid overfitting the data. Third, the estimator should be computationally tractable: ideally, implementation should not require estimation of a large number of nuisance parameters.

With these objectives in mind, I estimate the CEFs using a type of non-parametric estimator known as thin plate splines (Duchon 1977), given that its flexibility is regularized by a scalar penalty term λ, an MSE-optimal choice of which can be easily computed (Golub, Heath, and Wahba 1979). In the following, I give a brief definition of thin plate splines; see Appendix section A for more details.

For notational purposes, I use superscript/subscript $z \in \{0, 1\}$ to denote parameters and quantities for observations that lie in the region of the running variable space Ω_z in which observations are assigned treatment $Z_i = z$: for example, n_1 and n_0 denote the number of observations in Ω_1 and Ω_0 respectively. Assume that Ω_z is an open bounded subset of $\mathbb{R}^d$, and consider the Sobolev space of functions:

$$H^m(\Omega_z) = \left\{ u \in \mathscr{D}'(\Omega_z) : \int_{\Omega} \sum_{|\alpha| \leq m} |D^\alpha u|^2 < \infty \right\},$$

where $\mathscr{D}'(\Omega_z)$ is the space of Schwartz distributions, and:

$$D^\alpha u \equiv \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} ... \partial x_d^{\alpha_d}} u,$$

where $\alpha = (\alpha_1, ..., \alpha_d)'$ is a multi-index, $|\alpha| \equiv |\alpha_1| + ... + |\alpha_d|$, and $2m > d$.[14] Suppose that $g_z(x)$ are functions in $H^m(\Omega_z)$ for $z \in \{0, 1\}$, and let us write:

$$y_i^z = g_z(x_i^z) + \epsilon_i^z,$$

for $i = 1, ..., n_z$. Assume also that ϵ_i^z are i.i.d. random variables that are conditionally independent of x_i^z with mean zero and variance ν_z^2, and that $\epsilon_i^1 \perp \epsilon_{i'}^0$ for all i and i'. The thin

of $\int_{x \in \mathbb{F}_k} \tau(x) dF_{\mathbb{F}_k}(x)$, where $\mathbb{F}_k \equiv \{x \in \mathbb{R}^d | x_k = 0, x_j \geq 0 \; \forall j \neq k\}$.

[13] Specifically, assuming that observations are distributed uniformly over the running variable space and that the running variable space is a hypercube, for each estimation we will be discarding $2^{-(d-1)}$ of the data.

[14] To be more precise, elements in $u_z \in H^m(\Omega_z)$ are equivalence classes of functions, $[u_z]$, since one can always redefine a function in the space up to a set of measure zero without affecting the norm.

plate spline of order m and penalty parameter $\lambda_z > 0$ is defined by :

$$\hat{g}_z = argmin_{u \in \Omega_z} \sum_{i=1}^{n_z} (y_i^z - u(x_i^z))^2 + \lambda_z J_{md}^z(u),$$

where the penalty $J_{md}^z(u)$ is given by:

$$J_{md}^z(u) \equiv \int_{\Omega_z} \sum_{|\alpha| \leq m} |D^\alpha u|^2 dx.$$

Next, I introduce some regularity conditions required for my theoretical results.[15]

Definition. (Adams and Fournier 2001) Ω_z satisfies the *uniform cone condition* if there exists a locally finite open cover $\{U_j^z\}$ of the boundary of Ω_z and a corresponding sequence $\{C_j^z\}$ of finite cones, each congruent to some fixed finite cone C_z, such that:

1. There exists $M_z < \infty$ such that every U_j^z has diameter less than M_z.
2. $\Omega_\delta^z \subset \cup_{j=1}^\infty U_j^z$ for some $\delta > 0$ where Ω_δ^z is the set of points in Ω^r with distance less than δ from the boundary of Ω_z.
3. $Q_j^z \equiv \cup_{x^z \in \Omega_z \cap U_j^z}(x^z + C_j) \subset \Omega^z$ for every j.
4. For some finite $\bar{R}_z$, every collection of $\bar{R}_z + 1$ of the sets Q_j^z has empty intersection.

Condition F. Suppose that there exists CDFs F^z on Ω_z defined such that:

$$\lim_{n_z \to \infty} \sup_{x \in \Omega_z} |F^z(x) - F_{n_z}^z(x)| = 0,$$

where $F_{n_z}^z(t)$ is the distribution that assigns mass n_z^{-1} to all points in Ω_z, and that the limiting distribution F^z has density $f^z \in C^\infty(\bar{\Omega})$ with respect to the Lebesgue measure in Ω_z such that for all $x^z \in \Omega_z$, we have:

$$0 < \alpha_1^z \leq f^z(x^z) \leq \alpha_2^z.$$

Definition. We say that a sequence $\{T_{n_z}^z\}$ is *quasi-uniform* if there exists a constant $B_z > 0$ such that for each n_z, we have $h_{max}^z(T_{n_z}^z)/h_{min}^z(T_{n_z}^z) \leq B_z$, where:

$$h_{max}^z(T_{n_z}^z) \equiv \sup_{x \in \Omega_z} \inf_{i=1,\ldots,n_r} |x - x_i^z|,$$

$$h_{min}^z(T_{n_z}^z) \equiv \min_{i \neq j} |x_i^z - x_j^z|.$$

[15] An alternative set of regularity conditions can in principle be derived using functional Bahadur representation, by generalizing the analysis in Shang and Cheng (2013) for single-dimensional smoothing splines to multivariate thin plate splines.

Heuristically, the uniform cone condition places some restrictions on the smoothness of the boundaries Ω_z, while condition F ensures that x_i^z are sufficiently spread out over Ω_z, in the sense that the quasi-uniform condition holds with high probability. Also, throughout this paper we consider asymptotics where $n_1/n_0 = O(1)$.

The following theorem tells us that if the penalty parameter goes to zero at an appropriate rate, the estimate $\hat{\tau}(x)$ is consistent and asymptotically Gaussian.

Theorem 1. *Let Ω_z be open bounded subsets of $\mathbb{R}^d$ satisfying the uniform cone condition and having a Lipschitz boundary, and suppose that condition F and Assumption 1 are satisfied. In addition, assume that the penalty parameters λ_z are chosen such that $\lambda_z \to 0$ and $\lambda_z^{-1} = o(n_z^{2m/d})$. Then, for each $x \in \mathbb{F}$, as $n_0 \to \infty$, and $n_1 \to \infty$,*

1. *$\hat{\tau}(x) \to^p \tau(x)$.*
2. *$\sqrt{n\lambda^{d/2m}}\left(\hat{\tau}(x) - \tau(x)\right) \to^d N\left(b(x), \sigma^2(x)\right)$ for some constants $b(x)$ and $\sigma^2(x)$.*

All proofs are shown in Appendix section E.

In practice, we still need a finite-sample method for choosing λ_z. A common method of choosing the penalty parameter is generalized cross-validation (GCV), which is a modification of leave-one-out cross-validation (LOOCV) with two advantages: first, GCV is less computationally expensive than LOOCV, and second, it is invariant to rotation of the outcome vector and basis matrix. This procedure minimizes the GCV score, which is defined by:

$$GCV(\lambda_z) = \frac{n_z||y_z - A(\lambda_z)||^2}{[n_z - tr(A(\lambda_z))]^2},$$

where $A(\lambda_z)$ denotes the influence matrix for the model fit using λ_z (see Appendix section A for a formula for the influence matrix). The following proposition shows that the optimal rate of convergence is achieved if λ_z is chosen via GCV.

Proposition 2. *Let Ω_z be open bounded subsets of $\mathbb{R}^d$ satisfying the uniform cone condition and having a Lipschitz boundary, and suppose that condition F and Assumption 1 are satisfied. For each $x \in \mathbb{F}$, the optimal rate of convergence of $\hat{\tau}(x)$ for $\tau(x)$ is given by $O(n_z^{-2m/(2m+d)})$, and this rate is achieved if we set $\lambda_z = O(n_z^{-2m/(2m+d)})$, or choose λ_z using GCV.*

Turning next to inference, I start by showing that it is straightforward to compute Bayesian standard errors for $\hat{\tau}(x)$, before showing that they also have a frequentist interpretation. Note that the thin plate spline estimates $\hat{g}_z$ can be written in the form:

$$\hat{g}_z(x) = \sum_{k=1}^{K_z} \hat{\beta}_{z,k} s_{z,k}(x),$$

where $s_{z,k}(x)$, is the kth basis function for the thin plate spline estimate, and $\hat{\beta}_{z,k}$ are the estimates of the coefficients on these K_z basis functions, where typically, $K_z = O(n_z)$. With the appropriate choice of prior (Wahba 1990; Wood 2006) one obtains a Gaussian posterior distribution on $\hat{\beta}_{z,k}$, with covariance matrix which I denote by $\hat{\Sigma}_z$. Since the treatment effect estimate $\hat{\tau}(x)$ is linear in $\hat{\beta}_z$, this implies that the posterior distribution of $\hat{\tau}(x)$ is also Gaussian. It is natural then to consider Bayesian confidence sets of the form:

$$C(\hat{\tau}(x), \hat{se}(\hat{\tau}(x)), 1-\alpha) \equiv \left[\hat{\tau}(x) - q_{1-\alpha/2} \cdot \hat{se}\left(\hat{\tau}(x)\right), \hat{\tau}(x) + q_{1-\alpha/2} \cdot \hat{se}\left(\hat{\tau}(x)\right)\right], \tag{4}$$

where $q_{1-\alpha/2}$ denotes $(1-\alpha/2)$th quantile of a standard Gaussian distribution, and $\hat{se}$ is the estimate of the standard error computed using the posterior distribution of $(\hat{\beta}_0', \hat{\beta}_1')'$.

However, there are still a couple of concerns about these confidence sets. First, one may worry about the dependency on our choice of priors, i.e., that there may not be a frequentist interpretation of these inference results.[16,17] Second, even if we have correct standard errors, we need to ensure that the asymptotic distribution of the CATE estimate is centered at the truth (i.e., that $b = 0$ in Theorem 1), and typically the MSE-optimal choice of λ_z (e.g., as described in Proposition 2) does not guarantee this. Nonetheless, the next theorem shows that with appropriate undersmoothing, the Bayesian confidence set has a frequentist interpretation.

Theorem 3. *Let Ω_z be open bounded subsets of $\mathbb{R}^d$ satisfying the uniform cone condition and having a Lipschitz boundary, In addition, suppose that condition F and Assumption 1 are satisfied, and that the penalty parameters λ_z are chosen such that $\lambda_z = o(n_z^{-2m/(2m+d)})$ and $\lambda_z^{-1} = o(n_z^{2m/d})$. Then, denoting the $(1-\alpha)$th quantile of a standard Gaussian distribution by $q_{1-\alpha/2}$, the Bayesian confidence set in equation (4) has coverage rate $1-\alpha$ asymptotically:*

$$Pr\left(\tau(x) \in C(\hat{\tau}(x), \hat{se}(\hat{\tau}(x)), 1-\alpha)\right) \to 1-\alpha,$$

as $n_0 \to \infty$, $n_1 \to \infty$.

In order to apply the previous theorem, we need a way to implement undersmoothing.[18] The next proposition shows that one can obtain a penalty parameter that satisfies the rate condition by using the GCV choice of λ_z from a thin plate spline of higher penalty order. The intuition is similar to the result in local polynomial regressions, where the asymptotic

[16] The Bernstein-von Mises theorem as described in van der Vaart (2000) does not directly apply here, given that the results typically do not hold for infinite-dimensional statistical models such as non-parametric regression without further restrictions (Freedman 1999).

[17] Nychka (1988), and Marra and Wood (2011) show that in the case of thin plate splines, under regularity conditions, the Bayesian confidence sets described above have the frequentist property that they have close to nominal "across-the-function" (ACF) coverage. However, this does not tell us whether the Bayesian confidence set has nominal coverage for the CATE at any given point $x \in \mathbb{F}$.

[18] A commonly used method in practice is to simply divide the penalty parameter by two (Hall 1992). While the resulting sequence of λ_z does not satisfy the asymptotic rate condition, one might view this as a reasonable finite-sample approximation.

bias of the local quadratic regression estimate vanishes if one chooses the bandwidth using the MSE-optimal bandwidth for local linear regression (Fan and Gibjels 1992; Calonico, Cattaneo, and Titiunik 2014).[19,20]

Proposition 4. *Let Ω_z be open bounded subsets of $\mathbb{R}^d$ satisfying the uniform cone condition and having a Lipschitz boundary. In addition, suppose that $g_z \in H^{m+1}(\Omega_z)$ for $z = 0, 1$, and denote the penalty parameter chosen using GCV when the penalty order of the thin plate spline is m by $\hat{\lambda}^{GCV}_{m,z}$, and similarly, denote the estimated CEFs under penalty order m with penalty parameter λ_z by $\hat{g}_{m,z}(x; \lambda_z)$. Then, we have:*

$$Pr\left(\tau(x) \in C(\hat{\tau}_m(x), \hat{se}(\hat{\tau}_m(x)), 1-\alpha)\right) \to 1-\alpha,$$

where $\hat{\tau}_m(x) \equiv \hat{g}_{m,1}(x; \hat{\lambda}^{GCV}_{m+1,1}) - \hat{g}_{m,0}(x; \hat{\lambda}^{GCV}_{m+1,0})$.

In addition to pointwise confidence intervals, we may also be interested in obtaining simultaneous confidence bands for the entire CATE function $\{\tau(x)\}_{x\in\mathbb{F}}$, which allow us to test a number hypotheses about the CATE function, including:

- $H_0 : \tau(x) = 0$ for all $x \in \mathbb{F}$.
- $H_0 : \tau(x) = \bar{\tau}$ for all $x \in \mathbb{F}$, for some constant $\bar{\tau}$.
- $H_0 : \tau(x)$ is a linear function of x, for $x \in \mathbb{F}$.

In order to obtain simultaneous confidence bands, we can simulate the maximum of a Gaussian process using the following procedure.

1. Consider a fine grid for the running variables $\{x_g\}_{g\in\mathcal{G}}$ over the treatment frontier $\mathbb{F}$, and let $\vec{x}_g$ be the $|\mathcal{G}| \times d$ matrix with gth row equal to x_g.

[19] Intuitively, for larger penalty order m, the MSE-optimal penalty term goes to zero at a faster rate. So, if we use the MSE-optimal penalty choice $\hat{\lambda}^{GCV}_{m+1,z}$ for thin plate splines of a higher order $(m+1)$ than the order of the thin plate spline that we are ultimately fitting for the CEFs (m), the asymptotic bias in the estimator vanishes. The reason that $\hat{\lambda}^{GCV}_{m+1,z}$ goes to zero at a faster rate than $\hat{\lambda}^{GCV}_{m,z}$ is that for a given λ_z, $\hat{g}_{m+1,z}(x; \lambda_z)$ will be smoother than $\hat{g}_{m,z}(x; \lambda_z)$ since derivatives of order $m+1$ are also being penalized in the former. Hence, in order to "correct" for the fact that $\hat{g}_{m+1,z}(x; \lambda_z)$ is smoother, the MSE-optimal λ_z for $\hat{g}_{m+1,z}(x; \lambda_z)$ should be smaller. Another way to understand this is to note that the MSE-optimal λ_z is chosen so that the squared bias and variance terms in the MSE tend to zero at the same rate (otherwise, the MSE will tend to zero at the slower of the two rates). As m increases, the variance term increases at a slower rate as a function of λ_z as λ_z tends to zero, whereas the bias term remains linear in λ_z. Hence, in order to balance the rate at which the bias and variance tend to zero as m increases, λ_z needs to tend to zero at a faster rate.

[20] A less-used procedure for choosing the smoothing parameter is via generalized maximum likelihood (GML), which can be motivated from a Bayesian framework. Wahba (1985) shows that in the case of single-dimensional smoothing splines, under regularity conditions, λ_z chosen via GML tend to zero at a faster rate than the MSE-optimal rate. If the same holds true for multivariate thin plate splines, then we can implement undersmoothing by choosing λ_z via GML.

2. We can compute the covariance matrix of $\{\hat{\tau}(x_g)\}_{g\in\mathcal{G}}$ using:

$$\hat{V}\left(\{\hat{\tau}(x_g)\}_{g\in\mathcal{G}}\right) = s_1(\mathbf{x}_g)\hat{\Sigma}_1 s_1(\mathbf{x}_g)' + s_0(\mathbf{x}_g)\hat{\Sigma}_0 s_0(\mathbf{x}_g)',$$

   where $s_z(\mathbf{x}_g)$ is the $|\mathcal{G}| \times K_z$ matrix where the (g,k)th element is equal to $s_{z,k}(x_g)$. The standard error estimate of $\hat{\tau}(x_g)$ is given by $\hat{se}(\hat{\tau}(x_g)) = \sqrt{\hat{V}(\{\hat{\tau}(x_g)\}_{g\in\mathcal{G}})_{gg}}$.

3. For $b = 1, ..., B$:

   (a) Take a draw $\{\beta^{*(b)}_{(z),k}\}_{k=1}^{K_z}$ from the posterior distribution of $\{\hat{\beta}^{(b)}_{(z),k}\}_{k=1}^{K_z}$, and compute the treatment effect estimate $\tau^{*(b)}(x_g)$ based on these simulated parameters at each point x_g, $g \in \mathcal{G}$:

$$\tau^{*(b)}(x_g) = \sum_{k=1}^{K_z} \beta^{*(b)}_{z,k} s_{z,k}(x_g).$$

   (b) Denote the standardized difference between the simulated and estimated treatment effects by:

$$t^{*(b)}(x_g) = \frac{\tau^{*(b)}(x_g) - \hat{\tau}(x_g)}{\hat{se}(\hat{\tau}(x_g))},$$

   for each x_g, $g \in \mathcal{G}$.

4. We can then take the critical value $\bar{c}$ to be the $(1-\alpha)$th percentile of $\{\max_{g\in\mathcal{G}} |t^{*(b)}(x_g)|\}_{b=1}^{B}$.

See Appendix section B for details on how to test various null hypotheses about the CATE using simultaneous confidence bands.

In certain settings, we may want to account for heteroscedasticity or cluster our standard errors. To do so, we can formulate the estimation of $\tau(x)$ as a ridge regression problem. Specifically, we will write:

$$y_i = \mathbf{X}_i(x)'\beta(x) + u_i,$$

where:

$$\mathbf{X}_i(x) \equiv \begin{pmatrix} 1 \\ W_i \\ Z_i \cdot s_{1,2}(x_i - x) \\ \vdots \\ Z_i \cdot s_{1,K_1}(x_i - x) \\ (1 - Z_i) \cdot s_{0,2}(x_i - x) \\ \vdots \\ (1 - Z_i) \cdot s_{0,K_0}(x_i - x) \end{pmatrix} \in \mathbb{R}^{K_1+K_0} \tag{5}$$

where I omit the constant terms $s_{1,1}(\cdot) = s_{2,1}(\cdot) = 1$ from the basis functions.

Let $\mathbf{X}(x)$ be the design matrix, and observe that typically, $K_1 = n_1 + M$, $K_0 = n_0 + M$, and $M = \begin{pmatrix} m+d-1 \\ d \end{pmatrix}$, so $\mathbf{X}(x)'\mathbf{X}(x)$ does not have full rank. Hence, we consider the minimization problem:

$$\min_b \sum_{i=1}^{n} (y_i - \mathbf{X}(x)_i' b)^2 + (\lambda^{1/2} * b)' \mathbf{M}_x(\lambda)(\lambda^{1/2} * b),$$

where $\lambda^{1/2} \equiv (0, \sqrt{\lambda_1}\iota_{n_1}', \sqrt{\lambda_0}\iota_{n_0}')'$, the symbol $*$ denotes element-wise multiplication, and ι_{n_z} is a vector of ones that is of length n_z.[21] This is equivalent to fitting two thin plate splines over Ω_1 and Ω_0 respectively based on translated coordinates, with x now being the origin. Moreover, the second element of the solution vector is equal to the thin plate spline estimate of $\hat{\tau}(x)$, considering that the spline basis functions (other than the constant) all vanish at zero (under suitably chosen basis functions).

The solution to the minimization problem is given by:

$$\hat{\beta}(x) = (\mathbf{X}(x)'\mathbf{X}(x) + \mathbf{M}_x(\lambda))^{-1}\mathbf{X}(x)'\mathbf{Y}.$$

From this, we obtain the familiar "sandwich" formula for the conditional variance estimate:

$$\hat{Var}(\hat{\beta}(x)|\mathbf{X}(x)) = (\mathbf{X}(x)'\mathbf{X}(x) + \mathbf{M}_x(\lambda))^{-1}\mathbf{X}(x)'\hat{\Omega}(x)\mathbf{X}(x)(\mathbf{X}(x)'\mathbf{X}(x) + \mathbf{M}_x(\lambda))^{-1},$$

where $\hat{\Omega}(x)$ is an estimate of the covariance matrix for the residuals. In the case of heteroscedasticity, we can estimate $\hat{\Omega}(x)$ using the diagonal matrix with the squared residuals on the diagonal, or alternatively in the case of clustering, we may compute the Liang-Zeger standard errors. We can then take $\sqrt{\hat{Var}(\hat{\beta}(x)|\mathbf{X}(x))_{2,2}}$ as our estimate of the standard error of $\hat{\tau}(x)$.

Finally, we may also conduct inference using nonparametric bootstrap, which allows us to obtain pointwise confidence intervals as well as simultaneous confidence bands. Moreover, the bootstrap procedure can be modified to accommodate different assumptions about the covariance structure for the error terms (e.g., cluster bootstrap for cluster-robust standard errors).[22]

[21] The precise definition of $\mathbf{M}_x(\lambda)$ is given in Appendix section C.

[22] Here, I give an informal argument for why bootstrap may be theoretically justified in this setting. Invoking the interpretation of the thin plate spline as a Gaussian process, we use Theorem 2.4 from Giné and Zinn (1990) on the bootstrapping of general empirical measures. We observe that the main conditions of the theorem are satisfied: for part (a) we have $\int \sup_{u \in H_{m+1}(\Omega_z)} |u - Pu|^2 dP < \infty$, and for part (b), we use the result from Marcus (1985) which shows that the unit ball of a Sobolev space with $2m > d$ is a Donsker class for any P. To complete the argument, one must verify or assume that $H_{m+1}(\Omega_z)$ satisfies certain measurability conditions with respect to P,å spelled out in detail in Giné and Zinn (1984).

2.2 Fuzzy MRD Designs

Next, I study fuzzy MRD designs, where the jump in treatment probability at the treatment frontier is less than one. I denote the potential treatment of an individual i if she is assigned treatment z by $W_i(z)$. I also make several additional assumptions for the fuzzy MRD design.

***Assumption* 2.** *(Continuity of Mean Potential Treatments)* For all $x \in \mathbb{F}$ and $z \in \{0,1\}$:

$$\lim_{\epsilon \to 0} \mathbb{E}\left[W_i(z)|X_i = x', x' \in B^1_\epsilon(x)\right] = \lim_{\epsilon' \to 0} \mathbb{E}\left[W_i(z)|X_i = x', x' \in B^0_{\epsilon'}(x)\right]. \tag{6}$$

***Assumption* 3.** *(First Stage)* There exists a positive constant $\delta > 0$ such that:

$$\lim_{\epsilon \to 0} \mathbb{E}\left[W_i|X_i = x', x' \in B^1_\epsilon(x)\right] - \lim_{\epsilon' \to 0} \mathbb{E}\left[W_i|X_i = x', x' \in B^0_{\epsilon'}(x)\right] \geq \delta, \tag{7}$$

for all $x \in \mathbb{F}$.

***Assumption* 4.** *(Monotonicity)* $W_i(1) \geq W_i(0)$ almost surely.

Under Assumptions 1–4, the conditional local average treatment effect (CLATE), $\tau^{FMRD}(x)$ for $x \in \mathbb{F}$ is identified (Imbens and Angrist 1994; Hahn, Todd, and Van der Klaauw 2001), and is given by:

$$\begin{aligned}\tau^{FMRD}(x) &\equiv \mathbb{E}\left[Y_i(1) - Y_i(0)|X_i = x, W_i(1) > W_i(0)\right] \\ &= \frac{\lim_{\epsilon \to 0} \mathbb{E}\left[Y_i|X_i = x', x' \in B^1_\epsilon(x)\right] - \lim_{\epsilon' \to 0} \mathbb{E}\left[Y_i|X_i = x', x' \in B^0_{\epsilon'}(x)\right]}{\lim_{\epsilon'' \to 0} \mathbb{E}\left[W_i|X_i = x', x' \in B^1_{\epsilon''}(x)\right] - \lim_{\epsilon''' \to 0} \mathbb{E}\left[W_i|X_i = x', x' \in B^0_{\epsilon'''}(x)\right]}.\end{aligned} \tag{8}$$

The numerator of this Wald ratio can be estimated in the same manner as for sharp MRD, and the only difference for the denominator is that W_i replaces Y_i as the left-hand side variable.

Let us denote $h_z(x) \equiv \mathbb{E}[W_i(z)|X_i = x]$, and write the difference between two functions u_z as $\Delta u \equiv u_1 - u_0$. The following proposition describes properties of the thin plate spline estimator of the CLATE in a fuzzy MRD design.

Proposition 5. *Let Ω_z be open bounded subsets of $\mathbb{R}^d$ satisfying the uniform cone condition and having a Lipschitz boundary, and suppose that condition F and Assumptions 1–4 are satisfied. Also, assume that the penalty parameters λ_z and λ^h_z for the thin plate spline estimates $\hat{g}_z$ and $\hat{h}_z$ of g_z and h_z respectively are chosen such that $\lambda_z = o(1)$, $\lambda^h_z = o(1)$, $\lambda_z^{-1} = O(n_z^{2m/d})$, $(\lambda^h_z)^{-1} = O(n_z^{2m/d})$. Then, for all $x \in \mathbb{F}$,*

$$\hat{\tau}^{FMRD}(x) \equiv \Delta\hat{g}(x)/\Delta\hat{h}(x) \to^p \tau^{FMRD}(x),$$

as $n_0 \to \infty$ and $n_1 \to \infty$.

One can compute standard errors for $\hat{\tau}^{FMRD}(x)$ using nonparametric bootstrap. Alternatively, we can may obtain standard error estimates by formulating $\hat{\tau}^{FMRD}(x)$ as the solution to a 2SLS ridge problem. I give a brief overview here, and discuss details in Appendix section C. For any $x \in \mathbb{F}$, consider the same setup as for the ridge formulation for sharp MRD:

$$y_i = \mathbf{X}(x)_i'\beta(x) + u_i,$$

except that here, we instrument $\mathbf{X}(x)_i$ with $\mathbf{Z}(x)_i$, defined as:

$$\mathbf{Z}(x)_i \equiv \begin{pmatrix} 1 \\ Z_i \\ Z_i \cdot s_{1,1}(x_i - x) \\ \vdots \\ Z_i \cdot s_{1,K_1}(x_i - x) \\ (1 - Z_i) \cdot s_{0,1}(x_i - x) \\ \vdots \\ (1 - Z_i) \cdot s_{0,K_0}(x_i - x) \end{pmatrix} \in \mathbb{R}^{K_1 + K_0}. \tag{9}$$

Defining $\mathbf{Z}(x)$ as the "design matrix" but with $\mathbf{Z}(x)_i$ as the regressor, and suppressing the dependence of terms on x for notational simplicity, we can write the solution to this 2SLS ridge problem as $\hat{\beta}_{2SLS} = \mathbf{QY}$, where:

$$\mathbf{Q} \equiv (\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z} + \mathbf{M}(\lambda_h))^{-1}\mathbf{Z}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z} + \mathbf{M}(\lambda_h))^{-1}\mathbf{Z}'\mathbf{X} + \mathbf{M}(\lambda))^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z} + \mathbf{M}(\lambda_h))^{-1}\mathbf{Z}',$$

and λ and λ_h are vectors representing the penalty terms for the thin plate splines in the numerator and denominator respectively. An estimate of the conditional variance of $\hat{\beta}_{2SLS}$ is given by:

$$\hat{Var}(\hat{\beta}_{2SLS}|\mathbf{Z}, \mathbf{X}) = \mathbf{Q}\hat{\Omega}\mathbf{Q}'.$$

where we can use different estimators of $\hat{\Omega}$ depending on whether we assume heteroscedasticity or if we want clustered standard errors. Considering that $\hat{\tau}^{FMRD}(x)$ is given by the second element of $\hat{\beta}_{2SLS}$, we can take $\sqrt{\hat{Var}(\hat{\beta}_{2SLS}|\mathbf{Z}, \mathbf{X})_{2,2}}$ as our estimate of the standard error of $\hat{\tau}^{FMRD}(x)$.

2.3 MRK and MRDK Designs

Our MRD estimation approach can easily be extended to MRDK and MRK designs. Before discussing identification and estimation, I provide a motivating example of these designs based on the unemployment insurance (UI) system in Louisiana (Landais 2015). Here, we are interested in estimating the causal effect of weekly UI benefits $\mathcal{W}_i$ on employment outcomes Y_i. The benefit amount $\mathcal{W}_i$ is a linear function of prior earnings E_i up to a time-specific threshold $\bar{E}^t$ and is constant for higher levels of earnings. This threshold is constant during the period

$t < t_0$, but changes after t_0.

First, consider the case where the threshold increases discretely at time t_0 from $\bar{E}^0$ to a much higher threshold $\bar{E}^1$. We can write the formula for benefits as:

$$\mathcal{W}_i = \mathcal{W}(X_i) = \mathcal{W}(E_i, t_i) = \begin{cases} \alpha E_i & E_i \leq \bar{E}^0,\ t_i < 0, \\ \alpha \bar{E}^0 & E_i > \bar{E}^0,\ t_i < 0, \\ \alpha E_i & E_i \leq \bar{E}^1,\ t_i \geq 0, \\ \alpha \bar{E}^1 & E_i > \bar{E}^1,\ t_i \geq 0. \end{cases}$$

This is illustrated graphically in Figure 4a, where the solid and dashed lines represent the benefit formula prior to and after $t = 0$ respectively, as well as in Figure 2a, which plots benefits as a function of prior earnings and time. Figure 2a shows that individuals applying for UI before $t = 0$ face a kink in the benefit amount as prior earnings cross the threshold $\bar{E}^0$, whereas individuals with prior earnings falling between $\bar{E}^0$ and $\bar{E}^1$ face a discontinuity in benefits amount depending on whether they apply before or after $t = 0$.[23] Hence, we have an MRDK design, with a discontinuity on one dimension, and a kink on the other.

Next, as an example of an MRK design, suppose that instead of a large discrete increase in the cap at $t = 0$, the cap is increased gradually after that, as shown in Figure 4b. This UI benefit formula can be written as:

$$\mathcal{W}_i = \mathcal{W}(X_i) = \mathcal{W}(E_i, t_i) = \begin{cases} \alpha E_i & E_i \leq \bar{E}^0,\ t_i < t_0, \\ \alpha \bar{E}^0 & E_i > \bar{E}^0,\ t_i < t_0, \\ \alpha E_i & E_i \leq \bar{E}^{t_i},\ t_i \geq t_0, \\ \alpha \bar{E}^{t_i} & E_i > \bar{E}^{t_i},\ t_i \geq t_0, \end{cases} \qquad \bar{E}^{t_i} = \bar{E}^0 + \gamma t_i,$$

and is also shown graphically in Figure 3a, which plots UI benefits as a function of the two running variables. We observe that similar to the MRDK example, there is a kink in the benefit amount as prior earnings cross the threshold $\bar{E}^0$ for individuals applying for UI before $t = 0$. However, individuals with prior earnings falling between $\bar{E}^0$ and $\bar{E}^1$ now face a kink instead of a discontinuity at $t = 0$. In addition, there is another a kink for individuals applying for UI after $t = 0$ when their earnings cross threshold $\bar{E}^{t_i}$.

The description so far assumes that the benefit amount $\mathcal{W}_i$ is a deterministic function of X_i, which corresponds to a sharp design. However, it is often the case that the benefits formula may depend on variables unavailable in the data in addition to the running variables (e.g., marital status and number of dependents), so the simplified formula we are forced to rely on may be incorrect for some individuals. In this case, we have a fuzzy MRDK/MRK

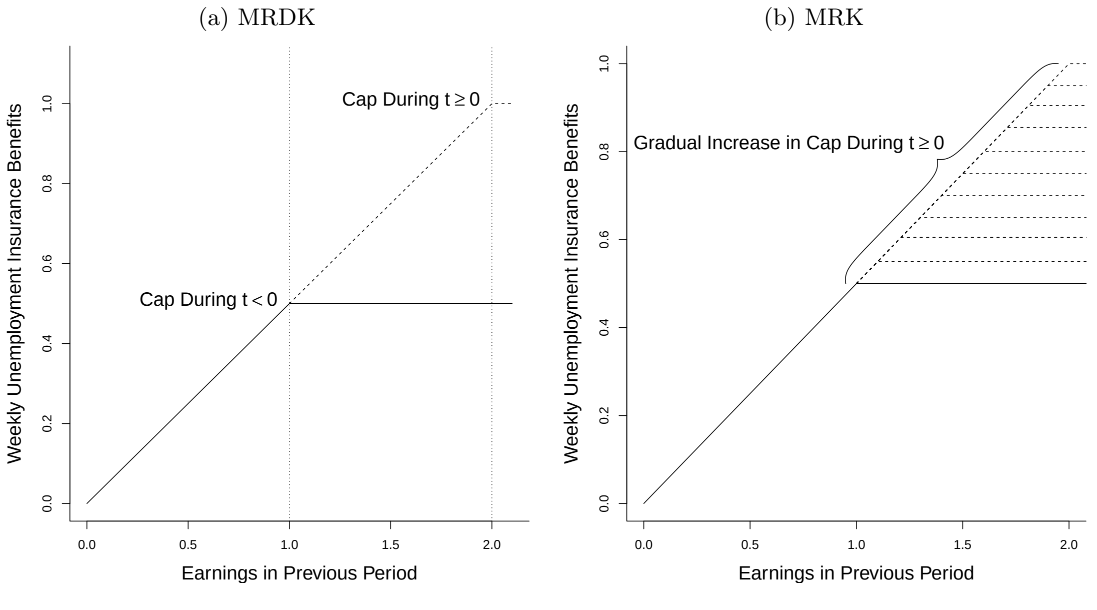
[23] There is also another kink for individuals applying for UI after $t = 0$ as their earnings cross the threshold $\bar{E}^1$, but this kink is outside the range of Figure 2a.

design, where $\mathcal{W}_i | X_i = x$ is a random variable.

Under assumptions about the smoothness of the potential outcome functions, if there is a causal relationship between the treatment variable (UI benefits) and the outcome variable (employment outcome), we might expect to see kinks/discontinuities in the CEF of the outcome variable at the same points in the running variable space. I illustrate this in Figures 2b and 3b by plotting the (CEFs of) job-finding probability as a function of the running variables, assuming that there is a negative causal relationship between UI benefits and employment outcomes. The MRDK/MRK estimand is given by the ratio of the kink/discontinuity of the outcome variable to the kink/discontinuity of the treatment variable at the same point in running variable space.

In order to estimate the MRDK/MRK CATE, we need to fit separate thin plate splines for different sets of individuals, similar to MRD estimation. However, instead of splitting the sample based on assigned treatment (as in the MRD case), in the MRDK/MRK examples here, we split the sample based on whether individuals are subject to the cap. Regions of the running variable space corresponding to individuals with benefit amounts that are either subject to or not subject to the cap are shown in Appendix Figure A.1 for the examples here.

Figure 4: UI Benefit Schedules Leading to MRDK and MRK Designs


(a) MRDK

(b) MRK

Notes: Panels A and B show the unemployment insurance (UI) weekly benefits described in the MRDK and MRK examples respectively in the main text, as functions of prior earnings. The benefit formula in the period before $t = 0$ is shown using solid lines, whereas the benefit formula occuring after $t = 0$ is shown using dashed lines.

Next, I discuss assumptions required for identification of MRK designs. I focus on MRK instead of MRDK given that the latter is simply a combination of MRD and MRK designs and does not require any assumptions above and beyond what is required for MRK. Considering that derivatives of a multivariate function taken from different directions will generally differ, I denote the directional derivative with respect to a unit vector v by D_v. In addition, I make the following assumptions for the MRK design.

***Assumption* 5.** *(Continuity of Mean Potential Outcome Derivatives)* For all $x \in \mathbb{F}$ and all w in the interior of the support of $\mathcal{W}_i$, $\mathbb{E}\left[Y_i(w)|X_i = x\right]$ has continuous partial derivatives with respect to $x_1, ..., x_d$.

***Assumption* 6.** *(Continuity of Mean Potential Treatment Derivatives)* For all $x \in \mathbb{F}$ and $z \in \{0, 1\}$, $\mathbb{E}\left[\mathcal{W}_i(z)|X_i = x\right]$ has continuous partial derivatives with respect to $x_1, ..., x_d$.

***Assumption* 7.** *(First Stage of MRK Design)* For all $x \in \mathbb{F}$, there exists a unit vector $v \in \mathbb{R}^d$ such that $x + \epsilon \cdot v \notin \mathbb{F}$ for sufficiently small $\epsilon > 0$, and some $\delta > 0$ such that:

$$\lim_{\epsilon \to 0} D_v \mathbb{E}\left[\mathcal{W}_i|X_i = x + \epsilon \cdot v, X_i \in B^1_\epsilon(x)\right] - \lim_{\epsilon' \to 0} D_v \mathbb{E}\left[\mathcal{W}_i|X_i = x + \epsilon' \cdot v, X_i \in B^0_{\epsilon'}(x)\right] \geq \delta.$$

***Assumption* 8.** *(Monotonicity for Fuzzy MRK)* For all $x \in \mathbb{F}$, and for any unit vector $v \in \mathbb{R}^d$ satisfying $x + \epsilon \cdot v \notin \mathbb{F}$ for sufficiently small $\epsilon > 0$, we have:

$$\lim_{\epsilon \to 0} D_v \mathcal{W}_i|(X_i = x + \epsilon \cdot v, X_i \in B^1_\epsilon(x)) \geq \lim_{\epsilon' \to 0} D_v \mathcal{W}_i|(X_i = x + \epsilon' \cdot v, X_i \in B^0_{\epsilon'}(x)),$$

almost surely.

Note that given the partial derivatives with respect to $x_1, ..., x_d$, assuming they are continuous, we can compute the derivative with respect to any unit vector v using the dot product. Hence, Assumption 5 also implies that:

$$\lim_{\epsilon \to 0} \mathbb{E}\left[D_v Y_i(w)|X_i = x', x' \in B^1_\epsilon(x)\right] = \lim_{\epsilon' \to 0} \mathbb{E}\left[D_v Y_i(w)|X_i = x', x' \in B^0_{\epsilon'}(x)\right],$$

which parallels Assumption 1 in the MRD case.

Under Assumptions 5–7 in the case of sharp MRK (and Assumptions 5–8 for fuzzy MRK), the MRK estimand $\partial \mathbb{E}\left[Y_i(w)|X = x\right] / \partial w$, denoted by $\tau^{MRK}(x)$ (respectively, $\tau^{FMRK}(x)$ for fuzzy MRK), is given by:

$$\frac{\lim_{\epsilon \to 0} D_v \mathbb{E}\left[Y_i|X_i = x + \epsilon \cdot v, X_i \in B^1_\epsilon(x)\right] - \lim_{\epsilon' \to 0} D_v \mathbb{E}\left[Y_i|X_i = x + \epsilon' \cdot v, X_i \in B^0_{\epsilon'}(x)\right]}{\lim_{\epsilon \to 0} D_v \mathbb{E}\left[\mathcal{W}_i|X_i = x + \epsilon \cdot v, X_i \in B^1_\epsilon(x)\right] - \lim_{\epsilon' \to 0} D_v \mathbb{E}\left[\mathcal{W}_i|X_i = x + \epsilon' \cdot v, X_i \in B^0_{\epsilon'}(x)\right]},$$

for any $v \in \mathbb{R}^d$ satisfying $x + \epsilon \cdot v \notin \mathbb{F}$ for sufficiently small $\epsilon > 0$. In the case of a sharp single-dimensional regression kink design, the RK estimand is termed the treatment-on-the-treated (TOT) by Florens, Heckman, Meghir, and Vytlacil (2008) or the local average response (LAR) by Altonji and Matzkin (2005). Since the treatment effect is conditional on individuals having running variables equal to x in the MRK case, I call this the conditional TOT (CTOT) or conditional LAR (CLAR). For a fuzzy MRK design, the estimand $\tau^{FMRK}(x)$ likely still represents a weighted average of the marginal effects of $\mathcal{W}_i$ on Y_i if one extends arguments from Card, Lee, Pei, and Weber (2015) to multiple dimensions.

Next, I describe some theoretical results for my MRK estimator, which is given by $\hat{\tau}^{MRK}(x) = \Delta D_v \hat{g}(x))/\Delta D_v h(x)$ and $\hat{\tau}^{FMRK}(x) = \Delta D_v \hat{g}(x))/\Delta D_v \hat{h}(x)$, where $h_z(x) \equiv \mathbb{E}[\mathcal{W}_i | X_i = x, x \in \Omega_z]$, and $\hat{h}_z(x)$ is the corresponding thin plate spline estimate.[24] The following theorem establishes the main properties of the the MRK and FMRK estimators.

Theorem 6. *Let Ω_z be open bounded subsets of $\mathbb{R}^d$ satisfying the uniform cone condition and having a Lipschitz boundary, and suppose that condition F and Assumptions 5–7 are satisfied. Moreover, assume that the penalty parameters are chosen such that $\lambda_{MRK,z} = o(1)$ and $\lambda^{-1}_{MRK,z} = o\left(n_z^{2m/(2+d)}\right)$. Then,*

1. $\hat{\tau}^{MRK}(x) \to^p \tau^{MRK}(x)$ *for all $x \in \mathbb{F}$* as $n_0 \to \infty$, and $n_1 \to \infty$.

2. *For each $x \in \mathbb{F}$, $\sqrt{n\lambda^{(2+d)/2m}}\left(\hat{\tau}^{MRK}(x) - \tau^{MRK}(x)\right) \to^d N(b_{MRK}(x), \sigma^2_{MRK}(x))$ for some constants $b_{MRK}(x)$ and $\sigma^2_{MRK}(x)$.*

3. *If we choose λ_z such that $\lambda_z = O(n_z^{-2m/(2m+d)})$, then the optimal rate of convergence for $\hat{\tau}^{MRK}(x)$ is achieved, and this rate is given by $O(n^{-2(m-1)/(2m+d)})$.*

4. *Suppose that we choose λ_z such that $\lambda_z = o(n_z^{-2m/(2m+d)})$. Then, for each $x \in \mathbb{F}$,*

$$Pr\left(\tau^{MRK}(x) \in C(\hat{\tau}^{MRK}(x), \hat{se}(\hat{\tau}^{MRK}(x)), 1-\alpha)\right) \to 1 - \alpha,$$

where:

$$\begin{aligned} &C(\hat{\tau}^{MRK}(x), \hat{se}(\hat{\tau}^{MRK}(x)), 1-\alpha) \\ &\qquad \equiv \left[\hat{\tau}^{MRK}(x) - q_{1-\alpha/2} \cdot \hat{se}\left(\hat{\tau}^{MRK}(x)\right), \hat{\tau}^{MRK}(x) + q_{1-\alpha/2} \cdot \hat{se}\left(\hat{\tau}^{MRK}(x)\right)\right], \end{aligned} \tag{10}$$

and $\hat{se}\left(\hat{\tau}^{MRK}(x)\right)$ denotes the standard error of $\hat{\tau}^{MRK}(x)$ computed using the posterior distribution of the thin plate spline estimates.

5. *In the case of fuzzy MRK, if Assumption 8 holds and the penalty parameters for $\Delta D_v \hat{h}(x)$ are chosen such that:*

$$\lambda^h_{FMRK,z} = o(1) \text{ and } \left(\lambda^h_{FMRK,z}\right)^{-1} = o\left(n_z^{2m/(2+d)}\right),$$

then for each $x \in \mathbb{F}$, $\hat{\tau}^{FMRK}(x) \to^p \tau^{FMRK}(x)$ as $n_0 \to \infty$, and $n_1 \to \infty$.

[24] Note that the formula is essentially the same for both sharp and fuzzy MRK estimators, the only difference being that there is no uncertainty in the denominator for sharp MRK.

Corollary 7. *Suppose that the assumptions of Theorem 6 hold, and that $g_z \in H^{m+1}(\Omega_z)$ for $z = 0, 1$. Denote the penalty parameter chosen using GCV when the penalty order of the thin plate spline is m by $\hat{\lambda}^{GCV}_{m,z}$, and similarly, denote the estimated CEFs under penalty order m with penalty parameter λ_z by $\hat{g}_{m,z}(x; \lambda_z)$. Then, we have:*

$$Pr\left(\tau^{MRK}(x) \in C(\hat{\tau}^{MRK}_m(x), \hat{se}(\hat{\tau}^{MRK}_m(x)), 1-\alpha)\right) \to 1-\alpha,$$

where $\hat{\tau}_m(x) \equiv D_v \hat{g}_{m,1}(x; \hat{\lambda}^{GCV}_{m+1,1}) - D_v \hat{g}_{m,0}(x; \hat{\lambda}^{GCV}_{m+1,0})$.

Similar to the MRD estimator, we may compute the standard error of $\hat{\tau}^{MRK}(x)$ and $\hat{\tau}^{FMRK}(x)$ using bootstrap. Alternatively, we can formulate FMRK estimation as a seemingly unrelated ridge regression (SURR), which then allows us to compute standard errors using the delta method (for details, see Appendix section C).

2.4 Discussion

First, in addition to the conditional average treatment effects, we may also be interested in the average effect over a subset of the treatment frontier. We can recover this by estimating the distribution of the running variables over $\mathbb{F}$, and then integrating the CATE estimates over the relevant subset of $\mathbb{F}$ with respect to this distribution. The standard errors can then be calculated using the delta method (see Appendix section G).

Second, one may be worried that if there are many more observations on one side of the treatment threshold $\mathbb{F}$ than the other, this may induce a discrepancy in the degree of regularization chosen for $\hat{g}_1(x)$ and $\hat{g}_0(x)$, potentially leading to spurious estimates of treatment effect heterogeneity. This is arguably less of a concern in the present setting if one focuses on observations relatively close to $\mathbb{F}$, given that a large imbalance in the number of observations on opposite sides of $\mathbb{F}$ suggests that precise manipulation of the running variables may be possible, which will likely invalidate the identification assumption for the MRD (or MRK) design.

Third, even in an MRD setting, the estimated kink at the threshold (i.e., $D_v \hat{\tau}(x)$) may still be of economic interest. In particular, the estimand $D_v \Delta g(x) = D_v \tau(x)$ describes whether the treatment effect is likely to increase or decrease if the running variables' thresholds were perturbed slightly. In MRD designs for program evaluation, thresholds for the running variables are typically chosen by the policymaker, so $D_v \tau(x)$ may be of policy relevance.

Fourth, for implementation purposes, it is preferable that both running variables have comparable scale. This is can easily be achieved by normalizing both variables to have unit variance for MRD estimation, and then converting the estimates back into the original units after that.

Fifth, in practice, MRK estimation requires substantially more data than MRD estimation.

This is evident from the MRK's slower convergence rate,[25] and the fact that even though the same choice of λ guarantees consistency and asymptotic validity of CIs for both sharp MRD and MRK estimators, the squared bias is $O(\lambda)$ for the former, whereas it is $O(\lambda^{(m-1)/m})$ for the latter.

Finally, a practical issue with estimating thin plate splines is its computational expense.[26] Hence, I use an approximation to thin plate splines suggested by Wood (2003) — thin plate regression splines (TPRS) — which allows the user to balance the tradeoff between computational efficiency and accuracy by choosing a parameter k_z (larger values of which correspond to greater accuracy).[27] Due to the larger bias for kink designs, a greater value of k_z is recommended for MRK and MRDK estimation (compared to MRD estimation). A more detailed description of TPRS can be found in Appendix section F.

3 Simulations

In this section, I present simulation results for the MRD, MRDK, and MRK estimators described in the previous section.

3.1 MRD Simulation

I first present simulation results for the MRD estimator, which I compare to various single-dimensional methods applied to the MRD setting, specifically, the local linear estimator with MSE-optimal bandwidths as in Imbens and Kalyanaraman (2012, henceforth IK), the bias-corrected local linear estimator in Calonico, Cattaneo, and Titiunik (2014, henceforth CCT), and the bounding approach based on an assumed bound for the second derivative as described in Kolesár and Rothe (2018), henceforth KR. In addition, I consider three versions of the MRD estimator in these simulations: an estimator using the MSE-optimal choice of penalty parameter for the thin plate regression splines (TPRS), a bias-corrected estimator using the MSE-optimal penalty parameter from higher-order TPRS, and an undersmoothed estimator using half of the MSE-optimal penalty parameter.

For my simulations, I consider variants of the following data-generating process (DGP)

[25] Specifically, the optimal convergence rates are $O\left(n^{-2(m-1)/(2m+d)}\right)$ for sharp MRK and $O\left(n^{-2m/(2m+d)}\right)$ for sharp MRD.

[26] In particular, while an efficient $O(n_z)$ algorithm exists for the single-dimensional thin plate splines (also known as smoothing splines), computational costs for thin plate splines with $d \geq 2$ are generally of order $O(n_z^3)$.

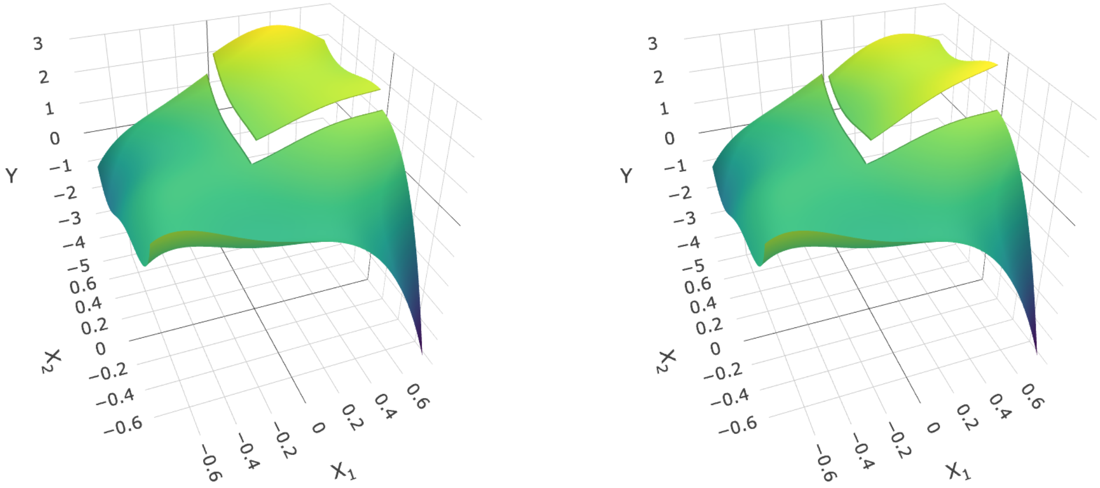
[27] At a high level, thin plate regression splines (TPRS) uses a basis matrix of rank k_z instead of $K_z \approx n_z$ in the case of thin plate splines, where the k_z basis functions for TPRS are chosen so that the minimization problem for thin plate splines is perturbed in the smallest possible way in a minimax sense (made precise in Appendix section F). This approximation reduces the computational expense from $O(n_z^3)$ to $O(k_z n_z^2)$.

based on a fifth order polynomial:

$$Y_i = \begin{cases} \sum_{p+q\leq 5} a_{p,q} X_{1i}^p X_{2i}^q + \tau(X_{1i}, X_{2i}) + \epsilon_i & X_{1i} \geq 0, X_{2i} \geq 0, \\ \sum_{p+q\leq 5} a_{p,q} X_{1i}^p X_{2i}^q + \epsilon_i & \text{otherwise.} \end{cases}$$

The error terms ϵ_i are distributed i.i.d. standard Gaussian, and the running variables X_{1i} and X_{2i} are each drawn independently from a $2Beta(3,3) - 1$ distribution, so that the running variables have support on $[-1,1]^2$, similar to the simulations in IK.[28] I consider two versions of this DGP, with either constant treatment effects where I set $\tau(X_{1i}, X_{2i}) = 0.5$, or heterogeneous effects where I let $\tau(X_{1i}, X_{2i}) = 0.5 + X_{1i} - X_{2i}$. Figures 5a and 5b show the CEFs for these two different DGPs. For each different DGP, I run 100 simulations, each with 10,000 observations, and use Bayesian standard errors for MRD inference.

Figure 5: CEFs for DGPs in MRD Simulations



(a) Constant Treatment Effects

(b) Heterogeneous Treatment Effects

Notes: These figures show the conditional expectation functions (CEFs) for the two DGPs I consider in my MRD simulations. Panel A shows the CEF for the DGP with constant treatment effects, whereas panel B shows the CEF in the DGP with heterogeneous treatment effects.

The results for the DGP with constant treatment effects are shown in Table 1, where the MRD estimates obtained by integrating the MRD CATE estimates $\hat{\tau}(x)$ over the relevant part

[28] In fact, this is not the most favorable distribution choice for my thin plate spline-based estimator. In particular, the convergence results for thin plate splines often require that observations satisfy a quasi-uniform condition, so one would expect the spline-based estimator to perform even better in finite samples if we assumed the running variables were uniformly distributed. Hence, these simulations also provide an informal test of whether the theoretical results in the previous section are relatively robust to different distributions of the running variables.

of the treatment frontier with respect to the (estimated) distribution of the running variables. We would expect the single-dimensional methods to perform well in this case, given that they are restricted to estimating constant treatment effects by design. Panels A and B show the performances of different estimators of the average treatment effects over the two segments of the treatment frontier $\mathbb{F}$: the positive x_2-axis and the positive x_1-axis respectively (given that the single-dimensional estimators produce estimates corresponding to these two subsets of $\mathbb{F}$).[29]

We observe in the first two columns that the bias for the MRD estimators are sometimes larger than the other estimators, but that they also have smaller MSE. In the last two columns, we see that the 95 percent CIs for all estimators have roughly the correct coverage, but that the MRD CIs tend to be shorter than the CIs for other estimators, reflecting the efficiency gains from using all of the data for estimation simultaneously, compared to the single-dimensional methods which use only about half of the data for each estimation.

Table 1: MRD Simulation Results for DGP with Constant Treatment Effects

Panel A. Estimates of the Average Treatment Effect Over $\{X_1=0, X_2\geq 0\}$				
Estimator	Bias	MSE	Coverage	Average CI Length
IK	0.002	0.009	0.97	0.334
CCT	-0.015	0.017	0.87	0.397
KR	0.001	0.009	0.96	0.372
MRD (MSE-Optimal)	-0.030	0.006	0.94	0.318
MRD (Bias-Corrected)	-0.031	0.007	0.94	0.329
MRD (Undersmoothing)	-0.038	0.008	0.94	0.333
Panel B. Estimates of the Average Treatment Effect Over $\{X_1\geq 0, X_2=0\}$				
Estimator	Bias	MSE	Coverage	Average CI Length
IK	-0.005	0.010	0.93	0.350
CCT	0.022	0.014	0.91	0.397
KR	-0.007	0.008	0.98	0.361
MRD (MSE-Optimal)	-0.018	0.006	0.95	0.320
MRD (Bias-Corrected)	-0.018	0.007	0.95	0.331
MRD (Undersmoothing)	-0.015	0.007	0.95	0.335

Notes: The IK estimator is based on local linear regression with bandwidth selection according to IK (2012). The CCT estimator is based on local linear regression with bandwidth selection and bias correction according to CCT (2014). The KR estimator is based on the method introduced in KR (2018) with an assumption on the bound for the second derivative of the CEF. Three versions of the MRD estimator are considered in these simulations: an estimator using the MSE-optimal choice of penalty parameter for the thin plate regression splines (TPRS), a bias-corrected estimator using the MSE-optimal penalty parameter from higher-order TPRS, and an undersmoothed estimator using half of the MSE-optimal penalty parameter. The results shown in this table are based on 100 realizations of the DGP with constant treatment effects. Confidence intervals are based on a 5 percent significance level. See text for more details on these simulations.

[29] For the MRD estimator, I estimate $\tau(x)$ at 10 equally spaced grid points along the positive x_2-axis, and similarly for the positive x_1-axis (ranging from zero to the largest observed value of the relevant running variable in each realization of the DGP).

Simulation results for the DGP with heterogeneous treatment effects are shown in Table 2. There is no natural way to estimate heterogeneous treatment effects using single-dimensional methods, so instead, I apply them separately to different subsets of the treatment frontier in order to estimate heterogeneous effects. Specifically, in Panel A, I apply these methods separately to $\{X_{1i} = 0, X_{2i} \in [0, c_1]\}, \{X_{1i} = 0, X_{2i} \in (c_1, c_2]\}, ..., \{X_{1i} = 0, X_{2i} \in [c_9, c_{10}]\}$, where $c_0 = 0, c_1, ..., c_{10}$ are equally spaced grid points, with c_{10} being the largest value of X_{2i} observed in that particular realization of the DGP, thus yielding 10 different estimates of $\tau(x)$ over the positive x_2-axis. I then repeat this procedure to obtain 10 different estimates of $\tau(x)$ over the positive x_1-axis for the single-dimensional estimators in Panel B.

Columns 1 and 2 of Table 2 show that the MRD estimators tend to have larger bias but smaller integrated MSE compared to the other estimators, similar to the simulations with constant treatment effects.[30] However, columns 3 and 4 show a striking difference between the performance of the CIs for the MRD CATE estimates and those of the other estimators. Column 3 shows that the pointwise MRD CIs have close to nominal coverage (close to 95 percent), whereas those for the pointwise IK and KR CIs have much lower coverage (typically ranging between 50 and 75 percent). Moreover, this is despite the fact that the MRD CIs are roughly half the length of the other CIs, as shown in the last column.[31]

[30] The bias (or integrated MSE) in these simulations are computed as the weighted average of the difference (or squared difference) between the treatment effect estimate over a subset of the treatment frontier and the true average treatment effect over the same subset. The weights are computed based on the density of the running variables over these subsets.

[31] The underperformance of the IK, CCT, and KR CIs in the DGP with heterogeneous treatment effects is not a critique of these methods, given that they were designed for single-dimensional RD. Rather, it shows that using single-dimensional RD methods in an MRD setting in an ad hoc manner to estimate heterogeneous treatment effects may result in poor finite-sample performance (or in other words, they may require a substantially larger sample to perform well compared to the MRD estimator proposed in this paper).

Table 2: MRD Simulation Results for DGP with Heterogeneous Treatment Effects

Panel A. Estimates of the Treatment Effect Over $\{X_1=0, X_2\geq 0\}$				
Estimator	Bias	IMSE	Coverage	CI Length
IK	0.008	0.044	0.741	0.982
CCT	-0.006	0.080	0.751	1.216
KR	0.019	0.029	0.786	0.933
MRD (MSE-Optimal)	-0.028	0.017	0.927	0.498
MRD (Bias-Corrected)	-0.030	0.021	0.922	0.522
MRD (Undersmoothing)	-0.037	0.020	0.930	0.532
Panel B. Estimates of the Treatment Effect Over $\{X_1\geq 0, X_2=0\}$				
Estimator	Bias	IMSE	Coverage	CI Length
IK	-0.010	0.043	0.507	0.998
CCT	0.004	0.077	0.564	1.225
KR	-0.020	0.032	0.526	0.949
MRD (MSE-Optimal)	-0.019	0.015	0.944	0.502
MRD (Bias-Corrected)	-0.020	0.018	0.942	0.527
MRD (Undersmoothing)	-0.016	0.018	0.945	0.536

Notes: The IK estimator is based on local linear regression with bandwidth selection according to IK (2012). The CCT estimator is based on local linear regression with bandwidth selection and bias correction according to CCT (2014). The KR estimator is based on the method introduced in KR (2018) with an assumption on the bound for the second derivative of the CEF. Three versions of the MRD estimator are considered in these simulations: an estimator using the MSE-optimal choice of penalty parameter for the thin plate regression splines (TPRS), a bias-corrected estimator using the MSE-optimal penalty parameter from higher-order TPRS, and an undersmoothed estimator using half of the MSE-optimal penalty parameter. The results shown in this table are based on 100 realizations of the DGP with heterogeneous treatment effects. The bias and IMSE in these simulations are respectively computed as the weighted average of the difference and weighted average squared difference between the treatment effect estimate over a subset of the treatment frontier and the true average treatment effect over the same subset. The weights are based on the density of the running variables over these subsets. Confidence intervals are based on a 5 percent significance level, and confidence intervals and coverage rates are pointwise. See text for more details on these simulations.

One may wonder whether the MRD estimates are able to capture qualitative features of the treatment effect heterogeneity, namely that it is increasing in X_{1i} and decreasing in X_{2i}. Appendix Figure A.2 shows that the MRD estimates are indeed increasing in X_{1i} and decreasing in X_{2i}, and that they correspond relatively closely to the true treatment effects (shown in red).[32]

Finally, in Appendix Table A.1, I explore the performance of the MRD estimators in greater detail. In particular, I compare the pointwise CIs based on analytic (i.e., Bayesian) standard errors to pointwise CIs based on nonparametric bootstrap, and I also present results for the simultaneous confidence bands. These results indicate that pointwise CIs based on analytic standard errors and nonparametric bootstrap perform quite similarly in terms of both coverage rates and CI length, while the simultaneous confidence bands are (unsurprisingly) substantially wider, and tend to have conservative coverage rates.

[32] Moreover, if we plot a linear fit through the MRD estimates using weighted least squares (WLS) or feasible generalized least squares (FGLS), we observe that the resulting fit (shown as dashed green and blue lines for WLS and FGLS respectively) is very close to the true treatment effect, and the slope coefficient is not statistically different from the true slope for the treatment effect at the 5 percent significance level.

3.2 MRDK and MRK Simulations

Moving onto simulations for the MRDK and MRK estimators, I first consider simulations based on the UI examples covered in the previous section. I adopt the same assumptions on the distributions of the running variables X_i and error terms ϵ_i as in the MRD simulations, and normalize the running variables so that the cap is binding for individuals in the positive quadrant of the running variable space. In addition, I assume that weekly UI benefits has a (constant) negative effect $\tau < 0$ on job-finding Y_i, and that it is given by:

$$Y_i = \tau \mathcal{W}_i + \epsilon_i.$$

A subtlety involving MRK estimation for this UI example is that there are more than two kinks in the UI benefit schedule, so I estimate more than two thin plate splines in order to avoid fitting a spline directly over a kink.[33]

As a closer parallel to the DGP considered for the MRD simulations, I also consider a DGP for the MRK design which only requires the estimation of two separate thin plate splines. For this DGP, I assume the same relationship between Y_i and $\mathcal{W}_i$, and that $\mathcal{W}_i$ is a fifth-order polynomial in the running variables:

$$\mathcal{W}_i = \begin{cases} \sum_{p+q\leq 5} b_{p,q} X_{1i}^p X_{2i}^q & X_{1i} < 0 \text{ or } X_{2i} < 0, \\ 0 & \text{otherwise.} \end{cases}$$

In addition, I assume that the coefficients $b_{p,q}$ are chosen such that $b_{0,0} = 0$, $\sum_{q\leq 4} b_{1,q} X_{2i}^q \neq 0$, and $\sum_{p\leq 4} b_{p,1} X_{1i}^p \neq 0$, so as to ensure that $\mathcal{W}_i$ is continuous in X_i, and that Assumption 7 (i.e., the first stage assumption) for the MRK design is satisfied. I simulate these three DGPs with $N = 10,000$ observations over 100 replications, estimating τ in each replication using my MRDK and MRK estimators.[34]

The results in Table 3 show that coverage rates for the MRDK and MRK estimates are 100 percent, suggesting that the analytic standard errors are conservative in this setting. The bias, MSE, and average CI length for the MRDK and MRK estimates tend to be larger than those for the MRD estimates (which is unsurprising given the slower convergence rates for the estimation of kinks relative to discontinuities), but are of a similar order of magnitude in most cases. We also observe that the MRDK and MRK estimates using different criteria for selecting the penalty parameters are broadly similar.

[33] Specifically, I fit seperate thin plate regression splines over the regions: $\{(x_1, x_2)|x_1 \geq 0, x_2 \geq 0\}$, $\{(x_1, x_2)|x_1 < 0, x_2 \geq 0\}$, and $\{(x_1, x_2)| - x_1 \leq x_2 < 0\}$. Then, I take the difference between the partial derivatives of the first two thin plate regression splines at the their boundary with respect to x_1, and the difference between partial derivatives of the first and third splines at their boundary with respect to x_2.

[34] As discussed in the previous section, the bias for the MRK estimator tends to be greater than the MRD estimator, so I use a larger number of basis functions k_z for the thin plate regression splines (compared to the MRD simulations) for a more accurate approximation of the thin plate splines.

Table 3: MRDK and MRK Simulation Results

Panel A. Estimates of the Average Treatment Effect Over $\{X_1=0, X_2\geq 0\}$				
UI Example: MRDK (kink)	Bias	MSE	Coverage	Avg. CI Length
MRDK: MSE-Optimal	-0.006	0.008	1.00	0.216
MRDK: Bias-Corrected	-0.009	0.030	1.00	0.317
MRDK: Undersmoothing	-0.005	0.015	1.00	0.265
UI Example: MRK				
MRK: MSE-Optimal	-0.001	0.010	1.00	0.229
MRK: Bias-Corrected	-0.010	0.017	1.00	0.362
MRK: Undersmoothing	-0.005	0.014	1.00	0.282
Polynomial Specification: MRK				
MRK: MSE-Optimal	0.004	0.469	1.00	1.063
MRK: Bias-Corrected	-0.044	0.736	1.00	1.193
MRK: Undersmoothing	-0.078	0.735	1.00	1.221
Panel B. Estimates of the Average Treatment Effect Over $\{X_1\geq 0, X_2=0\}$				
UI Example: MRDK (discontinuity)	Bias	MSE	Coverage	Avg. CI Length
MRDK: MSE-Optimal	0.003	0.007	1.00	0.128
MRDK: Bias-Corrected	0.002	0.010	1.00	0.136
MRDK: Undersmoothing	0.002	0.008	1.00	0.134
UI Example: MRK				
MRK: MSE-Optimal	0.016	0.018	1.00	0.321
MRK: Bias-Corrected	0.052	0.103	1.00	0.579
MRK: Undersmoothing	0.014	0.025	1.00	0.386
Polynomial Specification: MRK				
MRK: MSE-Optimal	0.061	0.272	1.00	0.802
MRK: Bias-Corrected	0.029	0.453	1.00	0.902
MRK: Undersmoothing	0.018	0.438	1.00	0.922

Notes: The table contains results from the MRDK and MRK simulations based on data-generating processes described in the main text, with sample sizes of N=10,000 and 100 replications. Three versions of the MRDK and MRK estimators are considered in these simulations: an estimator using the MSE-optimal choice of penalty parameter for the thin plate regression splines, a bias-corrected estimator using the MSE-optimal penalty parameter from higher-order TPRS, and an undersmoothed estimator using half of the MSE-optimal penalty parameter. Confidence intervals (CIs) are constructed based on a 5 percent significance level.

Finally, we observe that in all the MRD, MRDK, and MRK simulations, estimators using different selection criteria for the thin plate regression splines' penalty parameters tend to perform similarly. Hence, in the following section on empirical applications, I report MRD estimates using MSE-optimal penalty parameters for most of my results. The exception is when I test the validity of the research designs (specifically, when I replace the outcome variable with a predetermined individual characteristic, or test for a discontinuity in the multivariate density function), considering that the point estimate is of secondary importance in these instances, and the focus is on testing whether the identifying assumptions for the research design are violated.

4 Empirical Applications

In this section, I present two empirical applications for my MRD estimator. The first application studies the effect of financial aid eligibility on college enrollment using an MRD design with test scores and family wealth as the running variables, while the second application studies the effect of campaign advertisements on voter turnout using a geographical MRD design.[35]

4.1 Effect of Financial Aid Eligibility on College Enrollment

In this subsection, I present an empirical application based on the *Ser Pilo Paga* (SPP) program in Colombia, using data from Londoño-Vélez, Rodríguez, and Sánchez (2020, henceforth LRS). The SPP is a merit-based financial aid program introduced in Colombia in 2014. Students who score above a threshold on a standardized high school test and whose families are poor enough are eligible for financial aid if they enroll in a university with High Quality Accreditation. The SPP provides loans that are forgivable upon graduation, as well as a biannual stipend while recipients attend college.

For convenience, in my analysis I normalize the running variables so that they have standard deviation one, and so that students with values of both running variables greater than zero are eligible for financial aid. Henceforth, I refer to these normalized running variables as the test score and inverse wealth index.[36] Histograms of the running variables based on my main sample are shown in Appendix Figure A.3, and we observe that the SPP program is much more selective on the academic dimension than on the wealth dimension, i.e. most students in the sample are poor enough to qualify, but relatively few score well enough on the standardized test to do so.

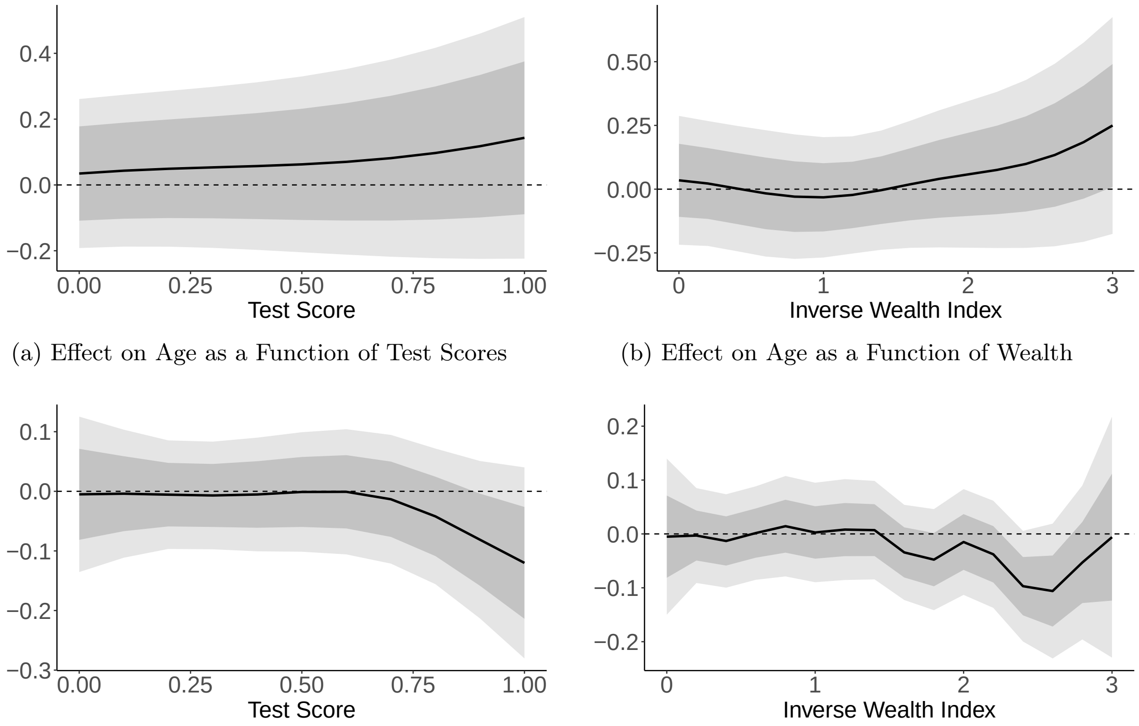
Before proceeding to the main estimation, I present tests of the validity of this MRD design, based on Lee's (2008) observation that if individuals close to the thresholds are able to precisely manipulate the values of their running variables to fall on the side of the thresholds they find desirable, this will likely invalidate the research design. This observation yields two testable implications: first, there should not be a discontinuity in pre-determined characteristics of individuals on either side of $\mathbb{F}$, and second, there should not be a discontinuity in the density of individuals at $\mathbb{F}$.

[35] Given the simulation evidence in the previous section that MRD estimates using different methods for selecting the penalty parameters tend to be quite similar, I report MRD estimates using MSE-optimal penalty parameters for most of my results, and only use the bias-corrected MRD estimates when testing the validity of the MRD research design (based on replacing the outcome variable with a predetermined individual characteristic, or based on a multidimensional McCrary test).

[36] In addition, I focus on the first cohort potentially eligible for the SPP (given that the program was announced two months after students took the standardized high school test, so there is little scope for test score manipulation), and drop observations with values of the running variables lower than the 1st percentile or greater than the 99th percentile (considering that these observations do not affect the CATE estimates at the boundary, and only add to the computational burden of density estimation).

I test the first of these conditions by estimating MRD designs replacing the outcome variable with age and gender, and plot the CATE estimates in Figure 6. The simultaneous confidence bands indicate that we are unable to reject that null that the CATE is zero along the entirety of $\mathbb{F}$ at the 5 percent significance level, which is consistent with the validity of the MRD design.

Figure 6: Placebo Test: Effect of SPP on Pre-Determined Characteristics



(a) Effect on Age as a Function of Test Scores

(b) Effect on Age as a Function of Wealth

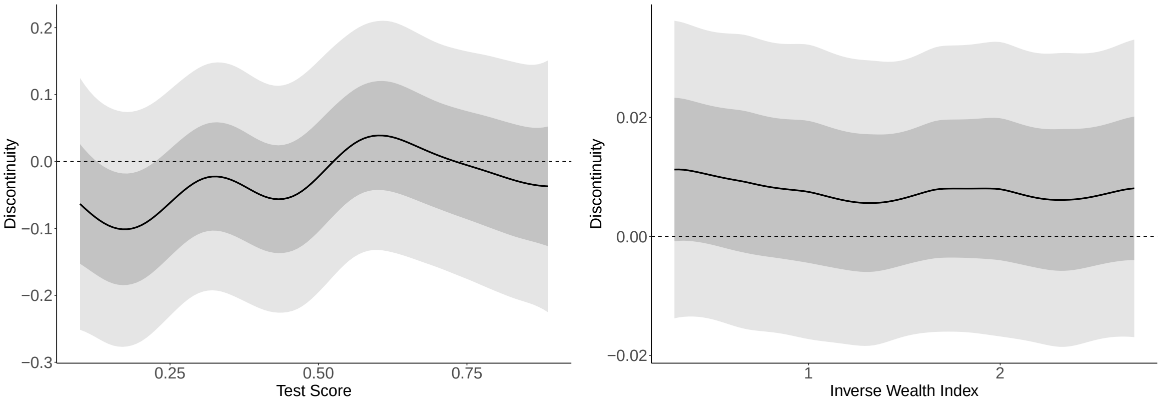
(c) Effect on Gender as a Function of Test Scores

(d) Effect on Gender as a Function of Wealth

Notes: The figures show MRD estimates of the CATE on the effect of financial aid on age and gender, as a function of test scores for students at the wealth threshold, and as a function of the inverse wealth index for students at the test score threshold. The light and dark shaded regions indicate 95 percent simultaneous confidence bands and 95 percent pointwise confidence intervals for the CATE estimates respectively.

The lack of any clear discontinuity in the univariate histograms of Appendix Figure A.3 is consistent with there being no precise manipulation, but does not necessarily imply that there are no discontinuities in the multivariate density (see Appendix section D for an example).[37,38] Hence, I develop a multidimensional "McCrary test" to formally test for discontinuities in the multivariate density function. At a high level, I first estimate a binned histogram, before applying the MRD estimator using the height of this histogram as the outcome variable. Theoretical properties for this multidimensional McCrary test are given in Appendix section D, where I also demonstrate via simulations that the test is able to detect discontinuities in the multivariate density even if the marginal univariate densities are smooth. On the other hand, when I apply this test to the SPP data, Figure 7 shows that estimates of the discontinuity are all statistically indistinguishable from zero at the 5 percent significance level, which supports the identification assumption of no precise manipulation.

Figure 7: Two-Dimensional "McCrary Tests"



(a) Test for Discontinuity at the Wealth Threshold (b) Test for Discontinuity at the Test Score Threshold

Notes: The figure shows two-dimensional "McCrary tests" for discontinuities in the multivariate density function along the treatment frontier $\mathbb{F}$. The shaded regions in light grey and dark grey represent the 95 percent pointwise confidence intervals and 95 percent simultaneous confidence bands respectively.

Having provided evidence on the credibility of this research design, I estimate the effect of the SPP on enrollment patterns. Tables 4 and 5 show the estimated effects of the program on enrollment in various types of college, for students with eligible wealth and test scores close to the threshold and vice versa, respectively. Panel A in these tables shows the MRD estimates of the treatment effects, whereas Panel B shows the original estimates from LRS, which were obtained by estimating single-dimensional RDs using CCT's method.

[37] For example, suppose that graders for the SABER 11 test believe that the SPP program should be targeted towards poor students to a greater extent. Then, it is possible that they may manipulate the grades of students who are far from the wealth threshold (i.e., very poor) to meet the threshold, but manipulate grades of students who are close to the wealth threshold (who are thus less poor) to fall just below the threshold. In this case, since manipulation occurs in both directions at the test score boundary, they may cancel out on average, and manipulation will not be detected from the univariate histogram for test scores.

[38] One may try to address this by plotting histograms for various subsets of the data as in Appendix Figures A.4, A.5, or plotting a two-dimensional histogram as in Appendix Figure A.6. However, slicing up the data may result in an underpowered test, and discontinuities may also be difficult to detect visually in a two-dimensional histogram.

We observe that the MRD point estimates and the original point estimates from LRS are quite similar qualitatively. Eligibility for the SPP increases overall college enrollment, an effect that is driven by increased enrollment in high quality private institutions; in fact, eligibility for SPP decreases enrollment in low quality colleges. This pattern can be explained by the fact that the SPP applies only for institutions with High Quality Accreditation.

In addition, we observe that the effects on enrollment (in any college, any high quality college, or any high quality private college) tend to be larger for students at the test score threshold than for students at the wealth threshold. One explanation for this pattern is that the test score threshold focuses on students with qualifying inverse wealth indices, who are on average much poorer than students at the wealth threshold with qualifying test scores. Hence, credit constraints may be more binding for the former set of students, thus explaining the larger effects on enrollment.[39]

While the MRD point estimates are not very different from the original LRS estimates, there are significant differences in their precision. This is especially pronounced in Table 5, which shows estimates of the average effect along the wealth threshold. We observe that MRD standard errors are sometimes less than half of the LRS standard errors, which makes sense given that LRS lose most of their sample when restricting the sample to students with eligible test scores for the analysis at the wealth threshold (due to the high test score threshold noted earlier) whereas the MRD estimator uses all of the data simultaneously.

[39] The only qualitative difference between the MRD and LRS estimates in these tables is that the MRD estimator finds that the SPP has a negative effect on enrollment in high quality public colleges in Table 4 for students along the test score threshold, whereas LRS find no effect for these students. Nonetheless, the negative effect is consistent with the general narrative in LRS that the SPP program induced students to substitute away from high quality public colleges to private colleges, and also agrees with LRS (and MRD) estimates of a negative effect on enrollment in high quality public colleges for students along the wealth threshold, as shown in Table 5.

Table 4: Effect of Financial Aid Eligibility on Enrollment: Students with Eligible Inverse Wealth Indices and Test Scores Close to the Threshold

Panel A. MRD Estimates

		High Quality Institutions			Low Quality Institutions		
	Any	Any	Private	Public	Any	Private	Public
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment Effect Estimate	0.334***	0.475***	0.478***	-0.01**	-0.139***	-0.062***	-0.07***
	(0.012)	(0.013)	(0.012)	(0.005)	(0.007)	(0.004)	(0.005)
Number of Observations	349,015	349,015	349,015	349,015	349,015	349,015	349,015

Panel B. Original Estimates from LRS

		High Quality Institutions			Low Quality Institutions		
	Any	Any	Private	Public	Any	Private	Public
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment Effect Estimate	0.320***	0.465***	0.466***	0.000	-0.154***	-0.063***	-0.087***
	(0.012)	(0.012)	(0.011)	(0.007)	(0.011)	(0.007)	(0.009)
Number of Observations	299,475	299,475	299,475	299,475	299,475	299,475	299,475

Notes: Standard errors are shown in parentheses.

Table 5: Effect of Financial Aid Eligibility on Enrollment: Students with Eligible Test Scores and with Inverse Wealth Indices Close to the Threshold

Panel A. MRD Estimates

		High Quality Institutions			Low Quality Institutions		
	Any	Any	Private	Public	Any	Private	Public
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment Effect Estimate	0.288***	0.429***	0.478***	-0.02***	-0.147***	-0.074***	-0.077***
	(0.021)	(0.023)	(0.019)	(0.005)	(0.011)	(0.006)	(0.005)
Number of Observations	349,015	349,015	349,015	349,015	349,015	349,015	349,015

Panel B. Original Estimates from LRS

		High Quality Institutions			Low Quality Institutions		
	Any	Any	Private	Public	Any	Private	Public
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment Effect Estimate	0.274***	0.396***	0.477***	-0.079***	-0.120***	-0.052***	-0.076***
	(0.027)	(0.024)	(0.020)	(0.018)	(0.022)	(0.015)	(0.016)
Number of Observations	23,132	23,132	23,132	23,132	23,132	23,132	23,132

Notes: Standard errors are shown in parentheses.

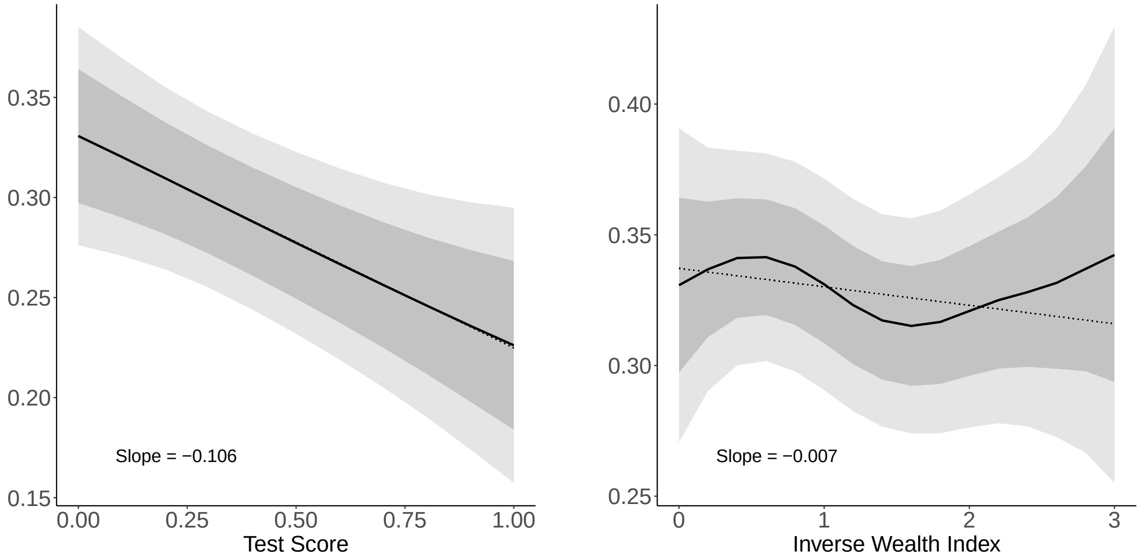
In addition to precision gains, the MRD estimator also allows us to recover heterogeneous treatment effects, which is especially interesting in this context given that it is unclear a priori whether the treatment effect should be increasing or decreasing in test scores and family wealth. Focusing on test scores first, on the one hand, we might expect that most high-ability students would have gone to college whether or not they were offered financial aid (e.g., perhaps because returns to college are increasing in academic ability), in which case there would be a "ceiling effect" and the treatment effect would be decreasing in test scores. On the other hand, several studies have uncovered a "reverse Roy" pattern of selection in education (Walters 2012; Kline and Walters 2016). If this were the case, high-ability students may require extra financial incentives to enroll in college, so the treatment effect may be increasing in test scores.

As for family wealth, one might expect the effect of the SPP to be increasing in the inverse wealth index, given that credit constraints may be more binding for poorer families. On the other hand, the test score threshold is rather high, so the marginal student is quite strong academically. So, it might also be the case that the labor market returns to a college degree are so high for these students that they are willing to borrow to go to college even at high interest rates.

MRD estimates of treatment effect heterogeneity shown in Figure 8 shed some light on these competing theories. The patterns are qualitatively consistent with a ceiling effect rather than a reverse Roy pattern: panel (a) shows that among students with wealth at the threshold and qualifying test scores, a one standard deviation higher test score is associated with a 10.6 percentage point smaller effect of financial aid on college enrollment, which is roughly one-third of the overall treatment effect. However, the confidence bands are rather wide, and a test for the null of constant treatment effects has a p-value of 0.224.[40] Panel (b) shows even less evidence of the treatment effect heterogeneity as a function of family wealth, with the treatment effect being only roughly 0.7 percentage points smaller for students with one standard deviation higher inverse wealth index on average (among students at the test score threshold with qualifying family wealth).

[40] This p-value is measured based on the MRD estimates using MSE-optimal penalty parameters for the underlying thin plate regression splines (TPRS). If we instead use the bias-corrected MRD estimates based on MSE-optimal penalty parameters from higher-order TPRS, or the undersmoothed TPRS using one-half of the MSE-optimal penalty parameters, we obtain p-values of 0.949 and 0.367 respectively. See Appendix section B for details about implementation of these tests.

Figure 8: Heterogeneity in the Effect of SPP on Enrollment in Any College



(a) As a Function of Test Scores

(b) As a Function of Inverse Wealth Index

Notes: Panel (a) shows MRD estimates of the CATE on the effect of financial aid on the probability of college enrollment as a function of test scores, for students with inverse wealth indices at the cutoff, whereas panel (b) shows MRD estimates of the CATE as a function of the inverse wealth index, for students with test scores at the cutoff. The shaded regions in light grey and dark grey represent the 95 percent pointwise confidence intervals and 95 percent simultaneous confidence bands respectively.

Appendix Figures A.8 and A.9 display analogous MRD estimates of treatment effect heterogeneity, focusing on enrollment in different types of colleges rather than in any type of college. We observe qualitatively similar patterns of treatment effect heterogeneity as for the results on enrollment in any college: focusing on the signs and magnitudes of the slope coefficients, the effect of financial aid on enrollment tends to be decreasing in test scores, and roughly constant as a function of family wealth.

Finally, a policymaker may be more interested in a slightly different type of treatment effect heterogeneity: the marginal returns from changing the test and wealth thresholds (in terms of college enrollment), which are given by the discontinuities in the partial derivatives at the thresholds. I estimate discontinuities in the derivatives at the thresholds for the test score and wealth indices of -0.103 (0.025) and 0.031 (0.033) respectively, suggesting that there are increasing returns to lowering the test score threshold, and roughly constant returns to lowering the family wealth threshold. This is consistent with the qualitative evidence in Figure 8 that the treatment effect is decreasing in test scores and constant in the wealth index, although these two sets of results need not necessarily coincide.[41]

[41] Note that while the marginal returns from increasing the thresholds are loosely connected to treatment effect heterogeneity along the treatment frontier, they measure different quantities. In particular, the slopes

4.2 Effect of Campaign Advertisements on Voter Turnout

In this subsection, I consider an empirical example based on the effect of campaign advertisements on voter turnout during the 2008 presidential election. The research design is a geographical RD that leverages a discontinuity in the volume of political ads that voters in New Jersey on either side of a media-market boundary were exposed to. Similar sources of variation were used in Huber and Arceneaux (2007) and Krasno and Green (2008), but here I compare my estimation results to those from Keele and Titiunik (2015), and Cattaneo, Idrobo, and Titiunik (2023) since these papers estimate RD specifications.

Presidential campaigns often buy television advertisements by designated market areas (DMAs). These are labels for different areas chosen by Nielsen Media Research, and importantly for the purposes of this empirical application, the designation of DMAs have no clear relationship with political variables. Instead, they are defined by Nielson as "exclusive geographic area of counties in which the home market television stations hold a dominance of total hours viewed", and often straddle states. Here, we compare turnout of voters in New Jersey on either side of a boundary defining different DMAs, who were exposed to very different levels of political advertisements during the 2008 elections: on one side of the boundary, voters were exposed to an average of 177 presidential ads daily from September 1 until election day, whereas voters on the other side of the boundary were exposed to no ads at all during this period (Keele and Titiunik 2015). The treatment and control groups are illustrated graphically in Figure 9a, where the blue and red points correspond to the location of voters in the DMA with no ads or many ads respectively, and the black solid line is the geographical boundary between the two DMAs.

In Keele and Titiunik (2015), the authors implement an MRD design based on local linear regressions using distance to boundary as the running variable. Specifically, for any point $x \in \mathbb{F}$, they estimate the CATE $\tau(x)$ using:

$$\hat{\tau}^{KT}(x) = \hat{g}_1^{ll}(x) - \hat{g}_0^{ll}(x),$$

where is the local linear fit's predicted value based on observations that are treated or untreated ($z = 1$ or $z = 0$) respectively, with weights depending on distance from x, a bandwidth

for treatment effect heterogeneity along the positive x_2-axis and x_1-axis shown in Figure 8 measure:

$$\int_{X_2 \geq 0} \frac{\partial \tau(0, X_2)}{\partial X_2} dF_{X_2|X_2 \geq 0}(X_2) \text{ and } \int_{X_1 \geq 0} \frac{\partial \tau(X_1, 0)}{\partial X_1} dF_{X_1|X_1 \geq 0}(X_1),$$

respectively, whereas the marginal returns from increasing the thresholds for the test score and inverse wealth index respectively measure:

$$\int_{X_1 \geq 0} \frac{\partial \tau(X_1, 0)}{\partial X_2} dF_{X_1|X_1 \geq 0}(X_1) \text{ and } \int_{X_2 \geq 0} \frac{\partial \tau(0, X_2)}{\partial X_1} dF_{X_2|X_2 \geq 0}(X_2).$$

$h_z(x)$, and a kernel $K(\cdot)$:

$$(\hat{\alpha}_z^{ll}(x), \hat{\beta}_z^{ll}(x)) \equiv argmin_{\alpha,\beta} \sum_{i=1}^{n_z} (\alpha + \beta dist(x_i, x))\, w_{iz}(x),$$

$$w_{iz}(x) \equiv \frac{1}{h_z(x)} K\left(\frac{dist(x_i, x)}{h_z(x)}\right).$$

While Keele and Titiunik estimate $\hat{\tau}^{KT}(x)$ for three separate $x \in \mathbb{F}$, in principle they could estimate $\hat{\tau}^{KT}(x)$ along the entire boundary $\mathbb{F}$. Hence, to facilitate the comparison with my MRD estimates, I estimate $\hat{\tau}^{KT}(x)$ along the entire boundary $\mathbb{F}$.

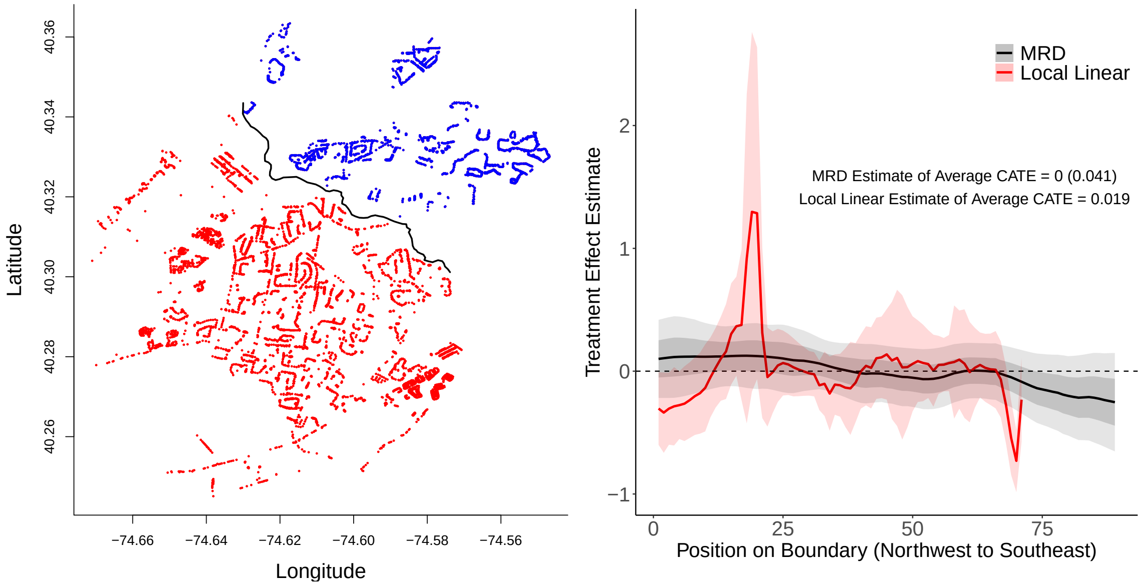
The results for my MRD CATE estimates (using longitude and latitude as the two running variables) and the local linear CATE estimates are shown by black and red solid lines respectively in Figure 9b. The dark shaded regions show the 95 percent pointwise confidence intervals for the CATE estimates, while the light shaded region shows the 95 percent simultaneous confidence bands for the MRD estimates.

First, we observe that the MRD estimate of the ATE along $\mathbb{F}$ is statistically indistinguishable from zero at the 5 percent significance level, consistent with Keele and Titiunik's (2015) finding that there is little evidence that political ads influenced voter turnout. The point estimate of the ATE for the local linear estimator is similarly small, but it is not easy to compute its standard error, given that for any $x, x' \in \mathbb{F}$, $\hat{\tau}^{KT}(x)$ and $\hat{\tau}^{KT}(x')$ are clearly correlated (since they were estimated using the same data) but were estimated in separate regressions.

Second, we see that the MRD CATE estimates $\{\hat{\tau}(x)\}_{x\in\mathbb{F}}$ tends to be smoother than the local linear CATE estimates $\{\hat{\tau}^{KT}(x)\}_{x\in\mathbb{F}}$, and that the local linear CATE estimates are sometimes unstable. For instance, we estimate that $\hat{\tau}^{KT}(x) > 1$ at a couple of points along the boundary, which does not make sense given that both treatment and outcome variables are binary, and for some points close to the edge of the boundary the local linear estimator fails to converge. In practice, an empirical researcher is likely to throw out estimates where $\hat{\tau}^{KT}(x) > 1$, and perhaps it also makes sense to ignore points close to the edge of the boundary since the data is sparser in that region. However, to the extent that subjective intervention by the researcher can be avoided, stability of the MRD CATE estimates is an attractive property.

Third, the pointwise confidence intervals for the MRD estimates tend to be narrower than the local linear estimates. In addition, I am able to estimate simultaneous confidence bands for the MRD estimate $\{\hat{\tau}(x)\}_{x\in\mathbb{F}}$, whereas it is less straightforward to do so for the local linear estimates given that $\hat{\tau}^{KT}(x)$ and $\hat{\tau}^{KT}(x')$ were estimated in separate regressions for $x, x' \in \mathbb{F}$, $x \neq x'$.

Figure 9: Effect of the Campaign Advertisements on Voter Turnout


(a) Treatment and Control Units in Bordering DMAs

(b) MRD Estimates

Notes: Panel (a) shows the locations of voters in neighboring DMAs, with the black line being the boundary between the DMAs. Locations of voters in the DMA where TV campaign ads were or were not broadcasted in the 2008 US presidential election are shown as red and blue points respectively. Panel (b) shows the MRD CATE estimates as well as the local linear CATE estimates, going from the northwest to southeast portion of the boundary. MRD and local linear estimates of the ATE along the boundary are also shown, as well as 95 percent pointwise confidence intervals for both sets of estimates, and simultaneous 95 percent confidence bands for the MRD estimates.

Finally, to illustrate why heterogeneity in the CATE in a geographical RD design may potentially be interesting, I project the MRD CATE estimates on surrounding characteristics. In particular, for each $x \in \mathbb{F}$, I find the mean age, voter registration status, income, and education of individuals close to x, and regress the MRD CATE estimates $\hat{\tau}(x)$ on these characteristics (weighting observations by the inverse of the variance of $\hat{\tau}(x)$). The results shown in Columns 2 and 3 of Appendix Table A.2 suggest that the effect of exposure to political ads on voter turnout is negatively associated with education and unemployment, but positively associated with poverty and age. However, given that we are unable to reject the null hypothesis that $\hat{\tau}(x) = 0$ for all $x \in \mathbb{F}$, these findings only hint at the possibility of treatment effect heterogeneity (stronger evidence of which may surface if one had a larger sample).

5 Conclusion

In this paper, I introduce a new method for estimating multidimensional RD and RK designs. This estimator allows the researcher to estimate heterogeneous treatment effects, and achieves efficiency gains relative to the common empirical approach of analyzing each running variable separately. I provide results on the theoretical properties of my estimator, and verify its performance in simulations. Finally, I demonstrate the utility of my estimator in two empirical applications. In the first application, my MRD estimation replicates the main findings of the original analysis with greater precision, and in addition, it reveals that the effect of financial eligibility on college enrollment is decreasing in students' test scores. In the second empirical application, my MRD estimates from a geographical RD setting replicate previous findings that political advertisements seemed to have little to no effect on voter turnout in the 2008 presidential elections, and these MRD estimates also tend to be more precise and stable than those produced by the local linear estimator.

References

[1] Abdulkadiroglu, Atila, Joshua D. Angrist, Yusuke Narita, and Parag Pathak, 2022. "Breaking Ties: Regression Discontinuity Design Meets Market Design." *Econometrica*, 90(1): 117–151.

[2] Altonji, Joseph G. and Rosa L. Matzkin, 2005. "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors," *Econometrica*, 73(4): 1053–1102.

[3] Angrist, Joshua D., and Victor Lavy, 1999. "Using Maimonides' rule to estimate the effect of class size on scholastic achievement." *The Quarterly Journal of Economics*, 114(2): 533–575.

[4] Angrist, Joshua D., Victor Lavy, Jetson Leder-Luis, and Adi Shany., 2019 "Maimonides' rule redux." *American Economic Review: Insights*, 1(3): 309–324.

[5] Armstrong, Timothy B., and Michal Kolesár, 2018. "Optimal inference in a class of regression models." *Econometrica*, 86(2): 655–683.

[6] Bertanha, Marinho, and Eunyi Chung, 2022. "Permutation Tests at Nonparametric Rates." *Journal of the American Statistical Association*, 0(0): 1–14.

[7] Caetano, António M., 2000. "Entropy Numbers of Embeddings Between Logarithmic Sobolev Spaces." *Portugaliae Mathematica*, 57(3): 355–379.

[8] Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik, 2014. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs". *Econometrica*, 82(6): 2295–2326.

[9] Canay, Ivan A., and Vishal Kamat, 2018. "Approximate Permutation Tests and Induced Order Statistics in the Regression Discontinuity Design." *Review of Economic Studies*, 85(3): 1577–1608.

[10] Card, David, David S. Lee, Zhuan Pei, and Andrea Weber, 2015. "Inference on causal effects in a generalized regression kink design." *Econometrica*, 83(6): 2453–2483.

[11] Castillo, Ismaël, and Richard Nickl, 2013. "Nonparametric Bernstein-von Mises Theorems in Gaussian White Noise." *The Annals of Statistics 2013*, 41(4): 1999–2028.

[12] Cattaneo, Matias D., Nicolás Idrobo, and Rocio Titiunik, 2023. "A Practical Introduction to Regression Discontinuity Designs: Extensions." arXiv preprint arXiv:2301.08958.

[13] Cattaneo, Matias D., Rocío Titiunik, Gonzalo Vazquez-Bare, and Luke Keele, 2016. "Interpreting regression discontinuity designs with multiple cutoffs." *The Journal of* Politics, 78(4): 1229–1248.

[14] Cattaneo, Matias D., and Rocio Titiunik, 2022. "Regression Discontinuity Designs." *Annual Review of Economics*, 14: 821–851.

[15] Dongarra, Jack J., Cleve Barry Moler, James R. Bunch, and Gilbert W. Stewart, 1979. "LINPACK users' guide." *Society for Industrial and Applied Mathematics.*

[16] Dell, Mellisa, 2010. "The Persistent Effects of Peru's Mining *Mita*." *Econometrica*, 78(6): 1863–1903.

[17] van Dijcke, David, and Florian Gunsilius, 2023, April 7. "Free Discontinuity Design". Retrieved from https://stat.mit.edu/calendar/gunsilius/ on April 23, 2023.

[18] Duchon, Jean, 1977. "Splines minimizing rotation-invariant semi-norms in Sobolev spaces." *Constructive theory of functions of several variables.* Springer, Berlin, Heidelberg: 85–100.

[19] Finkelstein, Amy, Nathaniel Hendren, and Mark Shepard, 2019. "Subsidizing health insurance for low-income adults: Evidence from Massachusetts." *American Economic Review*, 109(4): 1530–1567.

[20] Florens, Jean-Pierre, James J. Heckman, Costas Meghir, and Edward J. Vytlacil, 2008. "Identification of Treatment Effects Using Control Functions in Models With Continuous, Endogenous Treatment and Heterogeneous Effects." *Econometrica*, 76(5): 1191–1206.

[21] Freedman, David, 1999. "Wald Lecture on the Bernstein-Von Mises Theorem with Infinite-Dimensional Parameters." *The Annals of Statistics*, 27(4): 1119–1140.

[22] Ganong, Peter, and Simon Jäger, 2018. "A Permutation Test for the Regression Kink Design." *Journal of the American Statistical Association*, 113(522): 494–504.

[23] Gelman, Andrew, and Guido Imbens, 2019. "Why high-order polynomials should not be used in regression discontinuity designs". *Journal of Business & Economic Statistics*, 37(3): 447–456.

[24] Giné, Evarist, and Richard Nickl, 2010. "Adaptive estimation of a distribution function and its density in sup-norm loss by wavelet and spline projections." *Bernoulli*, 16(4): 1137–1163.

[25] Giné, Evarist, and Joel Zinn, 1984. "Some limit theorems for empirical processes." *The Annals of Probability*: 929–989.

[26] Giné, Evarist, and Joel Zinn, 1990. "Bootstrapping general empirical measures." *The Annals of Probability*: 851–869.

[27] Golub, Gene H., Michael Heath, and Grace Wahba, 1979. “Generalized cross-validation as a method for choosing a good ridge parameter.” *Technometrics*, 21(2): 215–223.

[28] Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw, 2001. “Identification and estimation of treatment effects with a regression-discontinuity design.” *Econometrica*, 69(1): 201–209.

[29] Huber, Gregory A., and Kevin Arceneaux, 2007. “Identifying the persuasive effects of presidential advertising.” *American Journal of Political Science*, 51(4): 957–977.

[30] Imbens, Guido W., and Joshua D. Angrist, 1994. “Identification and estimation of local average treatment effects.” *Econometrica*, 62(2): 467–475.

[31] Imbens, Guido, and Karthik Kalyanaraman, 2012. “Optimal Bandwidth Choice for the Regression Discontinuity Estimator.” *The Review of Economic Studies*, 79(3): 933–959.

[32] Imbens, Guido, and Stefan Wager, 2018. “Optimized Regression Discontinuity Designs.” Working paper.

[33] Kane, Thomas J., 2003. “A quasi-experimental estimate of the impact of financial aid on college-going.” NBER working paper 9703.

[34] Keele, Luke J., and Rocio Titiunik, 2015. “Geographic boundaries as regression discontinuities.” *Political Analysis*, 23(1): 127–155.

[35] Kline, Patrick, and Christopher R. Walters, 2016. “Evaluating Public Programs with Close Substitutes: the Case of Head Start.” *The Quarterly Journal of Economics*, 131(4): 1795–1848.

[36] Kolesár, Michal, and Chistoph Rothe, 2018. “Inference in Regression Discontinuity Designs with a Discrete Running Variable.” *American Economic Review*, 108(8): 2277–2304.

[37] Krasno, Jonathan S., and Donald P. Green, 2008. “Do televised presidential ads increase voter turnout? Evidence from a natural experiment.” *The Journal of Politics*, 70(1): 245–261.

[38] Landais, Camille, 2015. “Assessing the welfare effects of unemployment benefits using the regression kink design.” *American Economic Journal: Economic Policy*, 7(4): 243–278.

[39] Lee, David S., 2008. “Randomized experiments from non-random selection in U.S. House elections.” *Journal of Econometrics,* 142: 675–697.

[40] Londoño-Vélez, Juliana, Catherine Rodríguez, and Fabio Sánchez, 2020. “Upstream and Downstream Impacts of College Merit-Based Financial Aid for Low-Income Students: Ser Pilo Paga in Colombia.” *American Economic Journal: Economic Policy.*

[41] Marcus, David J, 1985. "Relationships between Donsker classes and Sobolev spaces." *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 69(3): 323–330.

[42] Marra, Giampiero, and Simon N. Wood, 2011. "Coverage Properties of Confidence Intervals for Generalized Additive Model Components." *Scandinavian Journal of Statistics*, 39(1): 53–74.

[43] Matsudaira, Jordan D., 2008. "Mandatory summer school and student achievement." *Journal of Econometrics,* 142(2): 829–850.

[44] Nychka, Douglas, 1988. "Bayesian Confidence Intervals for Smoothing Splines." *Journal of the American Statistical Association*, 83(404): 1134–1143.

[45] Papay, John P., John B. Willett, and Richard J. Murnane, 2011. "Extending the regression-discontinuity approach to multiple assignment variables." *Journal of Econometrics,* 161(2): 203–207.

[46] Shang, Zuofeng, and Guang Cheng, 2013. "Local and Global Asymptotic Inference in Smoothing Spline Models." *The Annals of Statistics*, 41(5): 2608–2638.

[47] Simonsen, Marianne, Lars Skipper, and Niels Skipper, 2016. "Price sensitivity of demand for prescription drugs: exploiting a regression kink design." *Journal of Applied Econometrics*, 31(2): 320–337.

[48] Snider, Connan, and Jonathan W. Williams, 2015. "Barriers to Entry in the Airline Industry: A Multidimensional Regression-Discontinuity Analysis of AIR-21." *Review of Economics and Statistics*, 97(5): 1002–1022.

[49] Taylor, Michael 2018. "PDE Course, Chapter 4: Sobolev Spaces." Lectures notes from Michael Taylor's website: https://mtaylor.web.unc.edu/notes/pde-course/

[50] Thistlethwaite, Donald L., and Donald T. Campbell, 1960. "Regression-discontinuity analysis: An alternative to the ex post facto experiment." *Journal of Educational Psychology*, 51(6): 309–317.

[51] Utreras, Florencio I. 1987. "On Generalized Cross-Validation for Multivariate Smoothing Spline Functions." *SIAM Journal on Scientific and Statistical Computing*, 8(4): 630–643.

[52] Utreras, Florencio I., 1988. "Convergence Rates for Multivariate Smoothing Spline Functions." *Journal of Approximation Theory*, 52: 1–27.

[53] Van der Vaart, Aad W., 2000. "Asymptotic statistics." Vol. 3. Cambridge university press.

[54] Wahba, Grace, 1985. "A Comparison of GCV and GML for Choosing the Smoothing Parameter in the Generalized Spline Smoothing Problem." *The Annals of Statistics*, 13(4): 1378–1402.

[55] Wahba, Grace, 1990. "Spline models for observational data." Vol. 59, Siam.

[56] Wood, Simon, 2003. "Thin plate regression splines." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1), 95–114.

[57] Wood, Simon, 2006. *Generalized Additive Models: An Introduction with R.* CRC Press.

[58] Zajonc, Tristan, 2012. "Essays on Causal Inference for Public Policy." Doctoral dissertation.

Appendix

A Background on Thin Plate Splines

In this section, I provide some general background on thin plate splines that may be useful for discussions in subsequent sections.[42] For notational simplicity, I drop the z subscripts/superscripts and focus on the fitting of a single thin plate spline in this section, before reintroducing these subscripts/superscripts in later sections when I discuss the estimators proposed in this paper (which involve fitting multiple thin plate splines).

We can view thin plate splines as the solution to a special case of a more general problem. Let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS), which is a Hilbert space of functions on Ω such that for each $x \in \Omega$, the evaluation functional L_x defined by $L_x u = u(x)$ for any $u \in \mathcal{H}$ is a bounded linear functional. By the Rietz representation theorem, for each $x \in \Omega$, there exists an element R_x in $\mathcal{H}$ such that:

$$L_x u = \langle R_x, u \rangle = u(x), u \in \mathcal{H},$$

which we will call the Rietz representer of L_x, where $\langle \cdot, \cdot \rangle$ denotes the inner product associated with $\mathcal{H}$. We also have a (unique positive definite) reproducing kernel R associated with $\mathcal{H}$ such that:

$$\langle R_s, R_t \rangle = \langle R(s, \cdot), R(t, \cdot) \rangle = R(s, t).$$

We can decompose the RKHS $\mathcal{H}$ into:

$$\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1,$$

where $\mathcal{H}_0$ is the null space of $\mathcal{H}$, and $\dim(\mathcal{H}_0) = M$. Denoting $L_i \equiv L_{x_i}$ for $i \in \{1, ..., n\}$, and letting P_1 be the orthogonal projection of $u \in \mathcal{H}$ onto $\mathcal{H}_1$, we consider the following problem:

$$g_\lambda \equiv \min_{u \in \mathcal{H}} \sum_{i=1}^{n} (y_i - L_i u)^2 + \lambda ||P_1 u||^2.$$

Thin plate splines are a special case of this minimization problem, where we consider an RKHS with $\mathcal{H} = H^m(\Omega)$ being the Sobolev space containing functions with bounded weak derivatives up to order m. In addition, assume that $\Omega \subseteq \mathbb{R}^d$, and $2m > d$. For any $u \in \mathcal{H}$, we can write:

$$u = u_0 + u_1,$$

[42] For an in-depth introduction to thin plate splines, see Wahba (1990).

where $u_0 \in \mathcal{H}_0$ and $u_1 \in \mathcal{H}_1$. Since the null space $\mathcal{H}_0$ is spanned by $M = \begin{pmatrix} m+d-1 \\ d \end{pmatrix}$ linearly independent polynomials $(\varphi_1, ..., \varphi_M)$ of degree less than m, this decomposition is equivalent to an application of the exact Taylor expansion of u, with u_0 being the $(m-1)$th order Taylor approximation, and u_1 being the remainder term.

Letting η_i be the Rietz representer of L_i, we can show that the thin plate spline problem may be written as the solution to:

$$\min_{c,d} ||Y - (\Sigma c + Sd)||^2 + \lambda c' \Sigma c, \tag{11}$$

where S is the $n \times M$ matrix defined by $S_{i,\nu} = L_i \varphi_i$, $\Sigma = \{\langle \xi_i, \xi_j \rangle\}$, and $\xi_i = P_1 \eta_i$. The solution to this problem can be expressed as:

$$\begin{aligned} g_\lambda &= \sum_{\nu=1}^{M} \alpha_\nu^* \varphi_\nu + \sum_{i=1}^{n} \delta_i^* \xi_i, \text{ where:} \\ d^* &= (S'\mathcal{M}^{-1}S)^{-1} S' \mathcal{M}^{-1} y, \\ c^* &= \mathcal{M}^{-1}(I - S(S'\mathcal{M}^{-1}S)^{-1}S'\mathcal{M}^{-1})y, \\ \mathcal{M} &= \Sigma + \lambda I. \end{aligned}$$

For notational simplicity, sometimes I also write: $g_\lambda(x) = s'(x)\beta$, where $s(x)$ is a $K \times 1$ vector where $K = M + n$, the first M elements of which correspond to the basis functions for $\mathcal{H}_0$, and the remaining terms correspond to the basis functions for $\mathcal{H}_1$.

The influence matrix is defined as the matrix $A(\lambda)$ that satisfies:

$$\begin{pmatrix} L_1 g_\lambda \\ \vdots \\ L_n g_\lambda \end{pmatrix} = A(\lambda) y.$$

Consider the QR decomposition of T (Dongarra, Bunch, Moler, and Stewart 1979):

$$T = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} \mathcal{R} \\ 0 \end{bmatrix},$$

where Q_1 is $n \times M$, Q_2 is $n \times (n - M)$, Q is orthogonal, and $\mathcal{R}$ is upper triangular, with $T'Q = 0$. Then, an explicit formula for $A(\lambda)$ is given by:

$$A(\lambda) = I - \lambda Q_2 (Q_2' \mathcal{M} Q_2)^{-1} Q_2' y.$$

B Testing Hypotheses About the CATE Function

The construction of simultaneous confidence bands allows us to test various null hypotheses about the CATE function $\tau(x)$, including:

- $H_0 : \tau(x) = 0$ for all $x \in \mathbb{F}$.
- $H_0 : \tau(x) = \bar{\tau}$ for all $x \in \mathbb{F}$, for some constant $\bar{\tau}$.
- $H_0 : \tau(x)$ is a linear function of x, for $x \in \mathbb{F}$.

Denote the simultaneous confidence band of level α for $\{\tau(x)\}_{x \in \mathbb{F}}$ as:

$$\{\bar{C}(\hat{\tau}(x), \hat{se}(\hat{\tau}(x)), 1 - \alpha, \mathbb{F})\}_{x \in \mathbb{F}}.$$

It is clear how to test the null hypothesis of whether the CATE contains zero: we can simply check whether $0 \in \bar{C}(\hat{\tau}(x), \hat{se}(\hat{\tau}(x)), 1 - \alpha, \mathbb{F})$ for any $x \in \mathbb{F}$. A p-value for this null can also be obtained by setting the p-value as α^*, which is defined as:

$$\alpha^* \equiv \inf\{\alpha | 0 \in \bar{C}(\hat{\tau}(x), \hat{se}(\hat{\tau}(x)), 1 - \alpha, \mathbb{F}) \text{ for some } x \in \mathbb{F}\}.$$

Testing the null hypothesis of constant treatment effects is relatively straightforward as well. Recall that our confidence band takes the form:

$$\bar{C}(\hat{\tau}(x), \hat{se}(\hat{\tau}(x)), 1 - \alpha, \mathbb{F}) = \left(\tau(x) - \bar{c}_{1-\alpha,\mathbb{F}} \hat{se}(\hat{\tau}(x)), \tau(x) + \bar{c}_{1-\alpha}(\mathbb{F}) \hat{se}(\hat{\tau}(x))\right).$$

An equivalent way to test whether there exists some $\bar{\tau}$ such that $\bar{\tau} \in \bar{C}(\hat{\tau}(x), \hat{se}(\hat{\tau}(x)), 1 - \alpha, \mathbb{F})$ for all $x \in \mathbb{F}$ is to check if:

$$\sup_{x \in \mathbb{F}} \left\{\tau(x) - \bar{c}_{1-\alpha}(\mathbb{F}) \hat{se}(\hat{\tau}(x))\right\} > \inf_{x \in \mathbb{F}} \left\{\tau(x) + \bar{c}_{1-\alpha}(\mathbb{F}) \hat{se}(\hat{\tau}(x))\right\},$$

in which case we will reject the null hypothesis at a significance level of α. We can also obtain a p-value by computing $\bar{c}_{1-\alpha}(\mathbb{F})$ for different values of α, and setting the p-value as the value of α^* that satisfies:

$$\sup_{x \in \mathbb{F}} \left\{\tau(x) - \bar{c}_{1-\alpha^*}(\mathbb{F}) \hat{se}(\hat{\tau}(x))\right\} = \inf_{x \in \mathbb{F}} \left\{\tau(x) + \bar{c}_{1-\alpha^*}(\mathbb{F}) \hat{se}(\hat{\tau}(x))\right\}.$$

Finally, to test whether the function is linear, we can employ a grid search. For example, suppose that the treatment frontier consists of the positive x_1-axis and x_2-axis. Then, to test whether $\tau(x)$ can be written as a linear function of x_1 along the positive x_1-axis, we can choose grids for the constant $a_1 \in A_1 \subseteq \bar{C}(\hat{\tau}(0, 0), \hat{se}(\hat{\tau}(0, 0)), 1 - \alpha, \mathbb{F})$ and slope $b_1 \in B_1$, and test whether $a_1 + b_1 x_1 \in \bar{C}(\hat{\tau}(x_1, 0), \hat{se}(\hat{\tau}(x_1, 0)), 1 - \alpha, \mathbb{F})$ for all x_1, and similarly for the x_2-axis.

Empirical applications of some of these tests can be found in Section 4. For example, tests for the validity of the research design — using predetermined variables as the outcome variable and the multidimensional McCrary test — are essentially tests of null "placebo" treatment effects along the treatment frontier.

A more interesting application was when we tested the hypothesis of whether the treatment effect of financial aid eligibility on college enrollment is increasing or decreasing in test scores (respectively, the inverse wealth index) for students at the wealth threshold with qualifying test scores, $\mathbb{F}_2$ (respectively, for students at the test score threshold with qualifying wealth, $\mathbb{F}_1$). Appendix Figure A.10 provides a visual illustration of how the p-value for these tests are constructed, by plotting:

$$\inf_x \left\{\tau(x) + \bar{c}_{1-\alpha}(\mathbb{F}_d)\hat{se}(\hat{\tau}(x))\right\} - \sup_x \left\{\tau(x) - \bar{c}_{1-\alpha}(\mathbb{F}_d)\hat{se}(\hat{\tau}(x))\right\},$$

as a function α, for α between zero and one. If the curve crosses zero, the p-value is given by the value of α at this intersection; if instead the curve is entirely above (respectively, below) zero, then the p-value is equal to one (zero).

C Ridge Formulation of MRD and MRK Estimators

For notational simplicity, I will suppress the dependence of various terms on x throughout most of this section.

C.1 Sharp MRD Estimation as Ridge Regression

To compute $\hat{\tau}(x)$, recall that we are fitting two thin plate splines over Ω_1 and Ω_0 separately, and evaluating the difference at $x \in \mathbb{F}$. Now consider the formulation of the thin plate spline problem given in equation (11) but applied separately to observations in regions Ω_1 and Ω_0, and where we translate the coordinates by x. Here, we are solving two problems:

$$\min_{c_z, d_z} ||Y_z - (\Sigma_z c_z + S_z d_z)||^2 + \lambda_z c_z' \Sigma_z c_z,$$

for $z = 0, 1$. Let us denote $\check{X}_z$ as the $n_z \times K_z$ matrix given by $\check{X}_z \equiv \begin{bmatrix} S_z & \Sigma_z \end{bmatrix}$, and let $\check{\beta}_z \equiv (d_z', c_z')$, and $\check{\Sigma}_z \equiv \begin{bmatrix} 0^{M\times M} & 0^{M\times n_z} \\ 0^{n_z\times M} & \Sigma_z \end{bmatrix}$. Then, we can rewrite the problem as:

$$\min_{\check{\beta}_z} ||Y_z - \check{X}_z \check{\beta}_z||^2 + \lambda_r \check{\beta}_z' \check{\Sigma}_z \check{\beta}_z',$$

and simple algebra reveals that the solution is given by:

$$\check{\beta}_r^* = \left(\check{X}_z'\check{X}_z + \lambda_z\check{\Sigma}_z\right)^{-1}\check{X}_z'Y_z.$$

Now, define:

$$\mathbf{X}(x)_i \equiv \begin{pmatrix} 1 \\ W_i \\ Z_i \cdot s_{1,2}(x_i - x) \\ \vdots \\ Z_i \cdot s_{1,K_1}(x_i - x) \\ (1 - Z_i) \cdot s_{0,2}(x_i - x) \\ \vdots \\ (1 - Z_i) \cdot s_{0,K_0}(x_i - x) \end{pmatrix} \in \mathbb{R}^{K_1+K_0},$$

$$\mathbf{M} \equiv \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0^{(K_1-1)\times 1} & \check{\Sigma}_1^{\sim 1} \\ 0^{(K_0-1)\times 1} & \check{\Sigma}_0^{\sim 1} \end{bmatrix} \in \mathbb{R}^{(K_1+K_0)\times(K_1+K_0)}, \mathbf{X} \equiv \begin{bmatrix} \mathbf{X}_1' \\ \vdots \\ \mathbf{X}'_n \end{bmatrix},$$

and consider the problem:

$$\min_b \sum_{i=1}^n (y_i - \mathbf{X}_i'b)^2 + (\lambda^{1/2} * b)'\mathbf{M}(\lambda)(\lambda^{1/2} * b),$$

where $\mathbf{M}(\lambda)$ is equal to $\mathbf{M}$ with rows 3 to $K_1 + 1$ multiplied by λ_1, and rows $K_1 + 2$ onwards multiplied by λ_0, and:

$$\lambda^{1/2} \equiv (0, \sqrt{\lambda_1}\iota_{n_1}', \sqrt{\lambda_0}\iota_{n_0}')',$$

where the symbol $*$ denotes element-wise multiplication, and ι_{n_z} is a vector of ones that is of length n_z.

Due to the multiplication of most elements in $\mathbf{X}_i$ by Z_i and $1 - Z_i$, this problem essentially fits two thin plate splines over Ω_1 and Ω_0 using only observations from the corresponding regions. The definition of $\lambda^{1/2}$ and $\mathbf{M}(\lambda)$ ensures that the intercept and linear terms are not penalized, and that the other basis functions for the two thin plate splines are penalized by their respective penalty terms λ_1 and λ_0.

Finally, due to the recentering of the observations at x, the value of the thin plate splines over regions Ω_1 and Ω_0 at $x \in \mathbb{F}$ are given by $b_1 + b_0$ and b_0 respectively. Hence, $\hat{\tau}(x)$ is the second term in the solution vector:

$$\hat{\beta}(x) = (\mathbf{X}'\mathbf{X} + \mathbf{M}(\lambda))^{-1}\mathbf{X}'\mathbf{Y}.$$

C.2 Fuzzy MRD Estimation as Ridge 2SLS

The argument for the formulation of fuzzy MRD as 2SLS is similar, but we additionally define:

$$\mathbf{Z}(x)_i \equiv \begin{pmatrix} 1 \\ Z_i \\ Z_i \cdot s_{1,2}(x_i - x) \\ \vdots \\ Z_i \cdot s_{1,K_1}(x_i - x) \\ (1 - Z_i) \cdot s_{0,2}(x_i - x) \\ \vdots \\ (1 - Z_i) \cdot s_{0,K_0}(x_i - x) \end{pmatrix} \in \mathbb{R}^{K_1 + K_0}$$

which we use as instruments for $\mathbf{X}(x)_i$.

We note that the thin plate splines for the denominator of the Wald ratio (i.e., the first stage) are solutions to the problem:

$$\min_{c_z, d_z} ||\vec{W}_z - (\Sigma_z c_z + S_z d_z)||^2 + \lambda_{z,h} c_z' \Sigma_z c_z.$$

Hence, the first stage can be formulated:

$$\min_b \sum_{i=1}^n (W_i - \mathbf{X}_i' b)^2 + (\lambda_h^{1/2} * b)' \mathbf{M} (\lambda_h^{1/2} * b),$$

where $\lambda_h^{1/2}$ is defined similarly $\lambda^{1/2}$ except with $\lambda_{z,h}$ in place of λ_z, and the solution is given by:

$$\hat{\beta}^{first} = (\mathbf{Z}'\mathbf{Z} + \mathbf{M}(\lambda_h))^{-1} \mathbf{Z}'\mathbf{X}.$$

Next, we consider the "second stage" ridge regression:

$$\min_b \sum_{i=1}^n (y_i - \hat{\mathbf{X}}_i' b)^2 + (\lambda^{1/2} * b)' \mathbf{M}(\lambda) (\lambda^{1/2} * b),$$

where the fitted values are given by $\hat{\mathbf{X}} \equiv \mathbf{Z}(\mathbf{Z}'\mathbf{Z} + \mathbf{M}(\lambda_h))^{-1}\mathbf{Z}'\mathbf{X}$. From this, we obtain:

$$\begin{aligned} \hat{\beta}_{2SLS}(x) &= (\hat{\mathbf{X}}'\hat{\mathbf{X}} + \mathbf{M}(\lambda))^{-1} \hat{\mathbf{X}}'\mathbf{Y}, \\ &= (\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z} + \mathbf{M}(\lambda_h))^{-1}\mathbf{Z}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z} + \mathbf{M}(\lambda_h))^{-1}\mathbf{Z}'\mathbf{X} + \mathbf{M}(\lambda))^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z} + \mathbf{M}(\lambda_h))^{-1}\mathbf{Z}'\mathbf{Y}, \end{aligned}$$

where $\mathbf{M}(\lambda_h)$ is defined in the same way as $\mathbf{M}(\lambda)$, except with the penalty parameters for the thin plate splines in the denominator $\lambda_{z,h}$ in place of λ_z. The fuzzy MRD estimate is given by the second element of this solution vector.

C.3 Fuzzy MRK Estimation as Seemingly Unrelated Ridge Regression

Finally, we can formulate FMRK estimation as a seemingly unrelated ridge regression. In particular, let us write:

$$X(x)_i^{FMRK} \equiv \begin{pmatrix} Z_i \\ 1 - Z_i \\ Z_i \cdot s_{1,2}(x_i - x) \\ \vdots \\ Z_i \cdot s_{1,K_1}(x_i - x) \\ (1 - Z_i) \cdot s_{0,2}(x_i - x) \\ \vdots \\ (1 - Z_i) \cdot s_{0,K_0}(x_i - x) \end{pmatrix} \in \mathbb{R}^{K_1+K_0}, \mathbf{X}(x)_1^{FMRK} = \begin{bmatrix} X(x)_1^{FMRK\prime} \\ \vdots \\ X(x)_n^{FMRK\prime} \end{bmatrix}.$$

Then, we can estimate the following equation:

$$\underbrace{\begin{bmatrix} Y \\ \mathcal{W} \end{bmatrix}}_{\equiv Y^{FMRK}} = \underbrace{\begin{bmatrix} \mathbf{X}_1^{FMRK} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_1^{FMRK} \end{bmatrix}}_{\equiv \mathbf{X}^{FMRK}} \underbrace{\begin{bmatrix} \beta^{reduced} \\ \beta^{first} \end{bmatrix}}_{\equiv \beta^{FMRK}} + \underbrace{\begin{bmatrix} \epsilon^{reduced} \\ \epsilon^{first} \end{bmatrix}}_{\equiv \epsilon^{FMRK}},$$

where we impose a ridge penalty on the parameters.

Specifically, consider the following minimization problem:

$$\begin{aligned} \hat{\beta}^{MRK} \equiv \min_{b \in \mathbb{R}^{2(K_0+K_1)}} & \sum_{i=1}^{n} (y_i - b^{reduced} X(x)_i^{FMRK})^2 \\ & + \sum_{i=1}^{n} (\mathcal{W}_i - b^{first} X(x)_i^{FMRK})^2 + (\lambda_{FMRK}^{1/2} * b)' \mathbf{M}^{FMRK}(\lambda)(\lambda^{1/2} * b), \end{aligned}$$

where $\lambda_{FMRK}^{1/2} \equiv (0, 0, \sqrt{\lambda_1} \cdot \iota_{n_1}', \sqrt{\lambda_0} \cdot \iota_{n_0}', 0, 0, \sqrt{\lambda_{h,1}} \cdot \iota_{n_1}, \sqrt{\lambda_{h,0}} \cdot \iota_{n_0})'$, and $\mathbf{M}^{FMRK}(\lambda)$ is equal to the block-diagonal matrix in $\mathbb{R}^{2(K_0+K_1) \times 2(K_0+K_1)}$ with $\mathbf{M}(\lambda)$ and $\mathbf{M}(\lambda_h)$ on the block diagonals. From this, we obtain the solution:

$$\hat{\beta}^{FMRK}(x) = \left(\mathbf{X}^{FMRK} \mathbf{X}^{FMRK\prime} + \mathbf{M}^{FMRK}(\lambda)\right)^{-1} \mathbf{X}^{FMRK\prime} Y^{FMRK},$$

and we can estimate its conditional variance using:

$$\begin{aligned}
&\hat{Var}(\hat{\beta}^{FMRK}(x)|\mathbf{X}^{FMRK}) \\
&\quad=(\mathbf{X}^{FMRK\prime}\mathbf{X}^{FMRK}+\mathbf{M}^{FMRK}(\lambda))^{-1}\mathbf{X}^{FMRK\prime} \\
&\qquad\hat{\Omega}^{FMRK}\mathbf{X}^{FMRK}(\mathbf{X}^{FMRK\prime}\mathbf{X}^{FMRK}+\mathbf{M}^{FMRK}(\lambda))^{-1}.
\end{aligned}$$

Now, define:

$$X_{v,\sim 0}^{FMRK}(x) \equiv \begin{pmatrix} D_v s_{1,\sim 1}(0) \\ -D_v s_{0,\sim 1}(0) \end{pmatrix},$$

where $s_{z,\sim 1}(\cdot)$ is equal to $s_z(\cdot)$ with the first element removed, and partition $\hat{\beta}^{FMRK}(x)$ into components corresponding to the "reduced form" and "first stage" of the FMRK:

$$\hat{\beta}^{FMRK}(x)' \equiv \left(\hat{\beta}_{reduced}^{FMRK\prime}, \hat{\beta}_{first}^{FMRK\prime}\right).$$

Also, let $\hat{\beta}_{reduced,\sim 0}^{FMRK}$ and $\hat{\beta}_{first,\sim 0}^{FMRK}$ be $\hat{\beta}_{reduced}^{FMRK}$ and $\hat{\beta}_{first}^{FMRK}$ respectively with the first two elements removed. Then, we can write the FMRK estimator as:

$$\hat{\tau}^{FMRK}(x) = \frac{X_{v,\sim 0}^{FMRK}(x)'\hat{\beta}_{reduced,\sim 0}^{FMRK}(x)}{X_{v,\sim 0}^{FMRK}(x)'\hat{\beta}_{first,\sim 0}^{FMRK}(x)}.$$

Given that we have an estimate of the entire (conditional) covariance matrix of $\hat{\beta}^{FMRK}(x)$, we can estimate the standard error of $\hat{\tau}^{FMRK}(x)$ using the delta method.

D Multidimensional McCrary Test

D.1 Theory

For simplicity, let:

$$\Omega_1 = \{(x_1, ..., x_d)|x_k \in [0,1)\ \forall k = 1, ..., d\}, \Omega_0 = (-1,1)^d \backslash \Omega_1, \tag{12}$$

and assume that the density of the running variables $f_z(x)$ has strictly positive density over these supports.

Divide Ω_1 and Ω_0 into cubes of side length b, and denote these cubes by C_g. We define the histogram density estimator by:

$$\hat{Y}_z^b(x) = \frac{1}{n_z b^d}\sum_{i=1}^{n_z}\mathbb{I}[X_i \in C_g(x)],$$

where $C_g(x)$ is the cube containing x, and denote the thin plate spline fit of $\hat{Y}_z^b(X_g)$ as a function of X_g by $\hat{f}_z$, where X_g is the center of the cube $C_g(x)$.

Finally, for $x \in \mathbb{F}$, we define:

$$\Delta f(x) \equiv \lim_{\epsilon \to 0^+} f_1(x + \epsilon v) - \lim_{\epsilon' \to 0^+} f_0(x + \epsilon' v'),$$

for vectors v and v' where $x + \epsilon v \in \Omega_1$ and $x + \epsilon' v' \in \Omega_0$ for sufficiently small $\epsilon > 0$ and $\epsilon' > 0$.

Proposition 8. *Suppose Ω_1 and Ω_0 are defined as in equation (12), as well as that condition F is satisfied and the distributions f_z of the running variables satisfy $f_z \in H_m(\Omega_z)$. In addition, suppose that the penalty parameters λ_z are chosen such that $\lambda_z = o(n_z^{-2m/(2m+d)})$ and $\lambda_z^{-1} = o(n_z^{2m/d})$, and the bin size is chosen such that $b = o(n^{-1/2}\lambda^{-d/(4m)})$, and $b^{-1} = o(\lambda^{-1/2m})$. Then,*

1. $\Delta \hat{f}_z(x) - \Delta f_z(x) \to^p 0$.
2. $\sqrt{n\lambda^{d/2m}} \left(\Delta \hat{f}_z(x) - \Delta f_z(x) \right) \to^d N(0, \sigma^2_{\Delta f}(x))$ for some constant $\sigma^2_{\Delta f}(x)$.
3. The Bayesian confidence sets for $\Delta \hat{f}_z(x)$ have correct coverage rate asymptotically, i.e.:

$$Pr\left(\Delta f_z(x) \in C(\Delta \hat{f}_z(x), \hat{se}(\Delta \hat{f}_z(x)), 1 - \alpha) \right) \to 1 - \alpha,$$

where:

$$C(\Delta \hat{f}_z(x), \hat{se}(\Delta \hat{f}_z(x)), 1-\alpha) \equiv \\ \left[\Delta \hat{f}_z(x) - q_{1-\alpha/2} \cdot \hat{se}\left(\Delta \hat{f}_z(x)\right), \Delta \hat{f}_z(x) + q_{1-\alpha/2} \cdot \hat{se}\left(\Delta \hat{f}_z(x)\right) \right],$$

and $\hat{se}(\Delta \hat{f}_z(x))$ is the standard error of $\Delta \hat{f}_z(x)$ based on the posterior distribution of the thin plate spline estimates.

D.2 Simulation Example

Suppose that $d = 2$, X_1 is distributed uniform on $(-1, 1)$, and that $X_2 | X_1 = x_1$ has mean zero but variance that depends on x_1. Specifically, assume that:

$$X_2 | X_1 = x_1 \sim \begin{cases} N(0, \sigma_2^2) & x_1 < 0, \\ \Phi^{-1}(U) \cdot (\mathbb{I}[U < 0]\sigma_2 + (1 - \mathbb{I}[U \geq 0])(a + bx_1)\sigma_2) & x_1 \geq 0, \end{cases}$$

where Φ^{-1} denotes the quantile function for a standard normal distribution, U is a continuous random variable distributed uniform on $(0, 1)$, $a = 1/2$, and b is the (strictly) positive solution of the equation:

$$log(b + 1/2) - log(1/2) = b.$$

In other words, X_2 has Gaussian distribution with constant variance σ_2^2 if $X_1 < 0$ (which corresponds to a region that is always untreated no matter the value of X_2. On the other hand, when X_1 is positive, informally if X_2 "is negative" it is drawn from the negative part of a Gaussian distribution with variance σ_2^2, whereas if X_2 "is positive", it is drawn from the positive part of a Gaussian distribution whose variance is increasing in x_1.

As a result, there is a discontinuity in the multivariate density along the part of the frontier $\mathbb{F}$ where $x_2 = 0$ and $x_1 > 0$ (which I will denote by $\mathbb{F}_2$). For smaller values of x_1, the density is higher on the right hand side of $x_2 = 0$, whereas for larger values of x_1, the density is higher on the left hand side of $x_2 = 0$. However, when plotting a single-dimensional histogram of the running variable x_2 (for values of x_1 greater than zero), we are averaging over $\mathbb{F}_2$ and consequently these positive and negative discontinuities "balance out". To see this, note that the average height of the limit of the density function from the left of $\mathbb{F}_2$ is $(2\pi)^{-1/2}\sigma_2^{-1}$, whereas the average height of the limit of the density function from the right of $\mathbb{F}_2$ is also:

$$\begin{aligned}(2\pi)^{-1/2}\sigma_2^{-1}\int_0^1 \frac{1}{a+bx_1}dx_1 &= (2\pi)^{-1/2}\sigma_2^{-1}(1/b)\left(ln(a+bx_1)\right)|_{x_1=0}^{x_1=1} \\ &= (2\pi)^{-1/2}\sigma_2^{-1}(1/b)\left(ln(b+1/2)-ln(1/2)\right) \\ &= (2\pi)^{-1/2}\sigma_2^{-1}.\end{aligned}$$

The simulation results from this data-generating process are shown in Appendix Figure A.7. In panel (a), we observe that the standard McCrary test does not detect a discontinuity in the univariate density (at the 5 percent significance level).[43] On the other hand, in panel (b) we observe that the two-dimensional "McCrary test" reveals clear discontinuities in the multivariate density function, with a positive discontinuity for smaller values of X_1 and a negative discontinuity for larger values of X_2, consistent with the DGP described above.

E Proofs

Proof of Theorem 1. If the quasi-uniform condition holds, then Theorem 1.1 in Utreras (1988) implies that there exists constants $P_{0,z}$ and $Q_{0,z}$ such that:

$$\mathbb{E}\left[|\hat{g}_z(x) - g_z(x)|_{j,\Omega_z}^2\right] \leq P_{0,z}\lambda_z^{(m-j)/m}|g_z(x)|_{m,\Omega}^2 + \frac{Q_{0,z}\nu_z^2}{n_z\lambda_z^{(2j+d)/2m}}, \tag{13}$$

[43] This is not a critique of the McCrary test given that it is designed to detect a discontinuity in a univariate density function for which there is none (as the calculations above show). In fact, in failing to reject the null, the test is working as intended.

where $|\cdot|_{k,\Omega}$ denotes the Sobolev seminorm defined by:

$$|u|^2_{k,\Omega} \equiv \mathbb{E}\left[\sum_{i_1,\ldots,i_k=1}^{d} \int_\Omega |\frac{\partial^k u(x)}{\partial x_{i_1} \ldots \partial x_{i_k}}|^2 dx\right],$$

for $\lambda_z \leq \lambda_{0,z}$ and $n_z \lambda_z^{d/2m} \geq 1$. The quasi-uniform condition holds with high probability under condition F, so the inequality above is satisfied with probability approaching one as $n_z \to \infty$. Taking $j = 0$, we obtain an expression for the MSE:

$$\mathbb{E}\left[\int_{\Omega_z} |\hat{g}_z(x) - g_z(x)|^2 dx\right] \leq P_{0,z}\lambda_z |g_z(x)|^2_{m,\Omega} + \frac{Q_0 \nu_z^2}{n_z \lambda_z^{d/2m}}. \tag{14}$$

From this expression, we observe that in order for $\mathbb{E}\left[\int |\hat{\tau}(x) - \tau(x)|^2 dx\right] \to 0$, we need both $\lambda_z = o(1)$ and $n_z \lambda_z^{d/2m} \to \infty$, which are precisely the rate conditions given in the statement of the theorem, thus proving part 1 of the theorem.

For any $x \in \Omega_z$ and $u \in H_z^m(\Omega)$, let L_x^z be the evaluation functional at x and let f_L^z be the unique Rietz representer of L_x^z so that $L_x^z = \langle \cdot, f_L^z \rangle$. We may choose a prior as described in Wahba (1990) and Wood (2006) which guarantees a Gaussian posterior distribution, thus satisfying the weak Bernstein-von Mises phenomenon (as defined in Definition 1 of Castillo and Nickl (2013)) trivially. Moreover, the evaluation functional is linear, which combined with the assumptions in the theorem, imply that:

$$\beta_{\mathbb{R}}\left(\Pi_{n_z}^{L_x^z} \circ (\theta^z_{L(\mathbb{X}^{(n_z)})})^{-1}, N(0, ||f_{g_z}||_2^2\right) \to 0,$$

for $z = 0, 1$ after applying Gram-Schmidt to the basis functions for the thin plate splines estimators,[44] using Theorem 3 of Castillo and Nickl (2013).[45] Applying this argument to the estimators for g_0 and g_1, we obtain asymptotic normality of $\hat{\tau}(x)$ as $n_0 \to \infty$ and $n_1 \to \infty$, as desired.

□

Proof of Proposition 2. The rate condition for λ_z that achieves the optimal rate of convergence can be be easily solved for by minimizing the right-hand-side of the inequality (14) as a function

[44] In fact, in the case of Schoenberg spaces with equally spaced knots, one can obtain the Battle-Lemarie wavelets by using Gram-Schmidt orthogonalization on the spline basis (Giné and Nickl 2010).

[45] The results in Castillo and Nickl (2013) are written for $\Omega_z \subseteq \mathbb{R}$, but the proofs extends to the multidimensional case with notational changes, so they apply to $\Omega_z \subseteq \mathbb{R}^d$ satisfying the conditions in the theorems in the present paper. For example, the requirement in Theorem 3 of Castillo and Nickl (2013) that we are considering functions in the Sobolev space $H^s_{(z)}$ for $s > 1/2$, where the standard definition of Sobolev spaces is extended to non-integer s using a wavelet-type basis. For the multidimensional case, we modify this requirement to functions in the Sobolev space $H^s_{(z)}$ for $s > d/2$, which guarantees that functions in this space are bounded and continuous (Taylor 2018). This is indeed satisfied given that we required that $2m > d$ for the Sobolev space H^m_z in our definition of thin plate splines.

of λ_z. From this solution, we also obtain the optimal rate of convergence.

To show that $\hat{\lambda}^{GCV}_{m,z} = O(\lambda^{opt}_{m,z})$, we observe that (under the conditions of Proposition 2), Utreras (1987) shows that the asymptotic inefficiency of the GCV choice of λ_z is $I_{GCV/OPT} = 1 + o(1)$ where λ_{OPT} is the MSE-optimal choice of λ.

□

Proof of Theorem 3. Examination of inequality (14) reveals that under the rate condition given in the theorem for λ_z, the asymptotic bias (the first term on the right-hand-side) tends to zero when $\hat{g}_z(x) - g_z(x)$ is scaled by $\sqrt{n_z \lambda_z^{d/2m}}$. Hence, the bias term $b(x)$ in Theorem 1 is zero, and to complete the proof, we simply need to show that the standard error $\hat{se}(\hat{\tau}(x))$ from the posterior distribution of $\hat{\tau}(x)$ is valid. This follows from Theorem 3 of Castillo and Nickl (2013), based on the same argument as the proof of the second part of Theorem 1.

□

Proof of Proposition 4. Suppose that the conditions in the proposition are satisfied. Then, by the same argument as in the proof of Theorem 2, the MSE-optimal penalty parameter for the thin plate splines of order m and $m+1$ satisfy $\lambda^{opt}_{m,z} = O(n_z^{-2m/(2m+d)})$ and $\lambda^{opt}_{m+1,z} = O(n_z^{-2(m+1)/(2(m+1)+d)})$ respectively. Given that $2(m+1)/(2(m+1)+d) > 2m/(2m+d)$, we have that $\lambda^{opt}_{m+1,z} = o(n_z^{-2m/(2m+d)})$. Finally, as noted in the proof of Proposition 2, we have $\hat{\lambda}^{GCV}_{m+1,z} = O(\lambda^{opt}_{m+1,z})$ under the maintained assumptions, hence completing the proof.

□

Proof of Proposition 5. The proof that $\Delta\hat{g}(x) \to^p \Delta g(x)$ and $\Delta\hat{h}(x) \to^p \Delta h(x)$ follows exactly as in the proof of part 1 of Theorem 1. Then, using the continuous mapping theorem (in conjunction with Assumption 4), we obtain $\hat{\tau}^{FMRD}(x) \to^p \tau^{FMRD}(x)$, as desired.

□

Proof of Theorem 6. For parts 1–4, given that the denominator $\Delta D_v h(x)$ is a known constant (for each $x \in \mathbb{F}$), we only need to consider the numerator for the proofs. Part 1 follows when we use inequality (13) with $j = 1$. Similarly, to show part 2, we just follow the proof of Theorem 1, noting that the derivative with respect to the direction v is a linear operator. Part 3 follows immediately if we choose λ_z to minimize the right-hand-side of inequality (13) with $j = 1$ as a function of n_z. The proof of part 4 is essentially the same as for Theorem 3 except with $j = 1$ instead of $j = 0$.

Finally, to show part 5, we use inequality (13) with $j = 1$ for $\hat{h}_z(x)$. Then, the result follows immediately from assumption 7 and the continuous mapping theorem.

□

Proof of Corollary 7. Combining part 3 and 4 of Theorem 6 with the inequality (13), with $j = 1$ as well as m and $m + 1$, we obtain the desired result.

□

Proof of Proposition 8. The proof essentially follows those of Theorems 1 and 3, the only difference being that we need to account for the approximation error originating from the need for an additional bandwidth choice for the histogram density estimator before applying the standard MRD estimator to the histogram density estimates.

To show that the approximation error is negligible under the rates for the bandwidth in the histogram density estimator given in the theorem, I first show that the error is negligible for $\hat{f}_z(x)$. Note that we can write:

$$\hat{f}_z(x) - f_z(x) = \left(\hat{f}_z(x) - \hat{Y}^b(x)\right) + \left(\hat{Y}^b(x) - f_z(x)\right),$$

where the first term is the approximation error from fitting the thin plate spline, and the second term is the approximation error from using the histogram density estimate instead of the density itself as the outcome variable in the thin plate spline estimation.

Consistency, asymptotic normality, and required undersmoothing rates for the first term were already studied earlier, so now we show that the approximation error from the second term is negligible relative to the first. First, we can compute the expectation and variance of $\hat{Y}_b(x)$. Let D be a vector of dummy variables of length d, and writing the sum of the elements as $|D|$, we have:

$$\begin{aligned}
\mathbb{E}[\hat{Y}_b(x)] &= \mathbb{E}\left[\frac{1}{n_z b^d}\sum_{i=1}^{n_z}\mathbb{I}[X_i \in C_g(x)]\right] \\
&= \mathbb{E}\left[\frac{1}{n_z b^d}\cdot n_z \cdot Pr\left(X_i \in C_g(x)\right)\right] \\
&= \mathbb{E}\left[\frac{1}{b^d}\cdot \sum_{k=0}^{d}(-1)^k \sum_{|D|=k} F_z(x + bD)\right] \\
&\to \frac{\partial^d F_z(x)}{\partial x_1 ... \partial x_d} = f_z(x)
\end{aligned}$$

where F_z is the CDF for f_z, so we have consistency. The variance is given by:

$$\begin{aligned}
Var[\hat{Y}_b(x)] &= \frac{n_z Pr\left(X_i \in C_g(x)\right)\left[1 - Pr\left(X_i \in C_g(x)\right)\right]}{n_z^2 b^{2d}} \\
&= \mathbb{E}[\hat{Y}_b(x)] \cdot \frac{1 - Pr\left(X_i \in C_g(x)\right)}{n_z b^d},
\end{aligned}$$

where $1 - Pr\left(X_i \in C_g(x)\right) \to 1$ as $b \to 0$.

Now, we consider $\sqrt{\lambda^{d/2m}n}\left(\hat{Y}_b(x) - f_z(x)\right)$. First, taking the expectation:

$$\mathbb{E}\left[\sqrt{n_z\lambda^{d/2m}}\left(\hat{Y}_b(x) - f_z(x)\right)\right] = \sqrt{n_z\lambda^{d/2m}}\left(Mb + o(b)\right),$$

we find that the expectation goes to zero if and only if $b = o(n^{-1/2}\lambda^{-d/4m})$, as is assumed in the proposition. Turning next to the variance, we find that:

$$\begin{aligned} Var\left[\sqrt{n_z\lambda^{d/2m}}\left(\hat{Y}_b(x) - f_z(x)\right)\right] &= \lambda^{d/2m}n_z\mathbb{E}[\hat{Y}_b(x)] \cdot \frac{1 - Pr\left(X_i \in C_g(x)\right)}{n_z b^d} \\ &= n_z\lambda^{d/2m}\left(O(n_z^{-1}b^{-d})\right) \\ &= O(\lambda^{d/2m}b^{-d}), \end{aligned}$$

and this tends to zero if and only if b tends to zero at a slower rate than $\lambda^{1/2m}$, i.e., $b^{-1} = o(\lambda^{-1/2m})$, which is also a condition listed in the proposition. Therefore, the approximation error from using $\hat{Y}^b(x)$ instead of $f_z(x)$ is negligible relative to the approximation error from the thin plate spline estimation. Finally, this argument applies to both $\hat{f}_1(x)$ and $\hat{f}_0(x)$, so the same applies to $\Delta\hat{f}_z(x)$, as desired.

□

F An Approximation of Thin Plate Splines

In practice, fitting thin plate splines can be computationally intensive in multivariate cases. Hence, for implementation, I approximate thin plate splines with thin plate regression splines (TPRS), which I define in this section. The description in this section borrows heavily from Wood (2003, 2006).

For simplicity of exposition, I assume for this discussion that no two observations have an identical combination of covariate values, and I use y_z to denote the vector of outcome variable values for observations lying in Ω_z. Recall that a thin plate spline $\hat{g}_z(x)$ is the solution to the following minimization problem:

$$\min_{u \in \mathcal{H}} \sum_{i=1}^{n_z}(y_{z,i} - u(x_i^z))^2 + \lambda J_{md}(u),$$

where the definition J_{md} can be found in the main text. Assuming $2m > d$ is satisfied, one way to write the solution is:

$$\hat{g}_z(x) = \sum_{i=1}^{n_z}\delta_{z,i}^*\eta_{md}(||x - x_i^z||) + \sum_{j=1}^{M}\alpha_{j,z}^*\phi_j(x),$$

where the orthogonality constraint $T_z'\delta_z^* = 0$ is satisfied, with T_z defined by $(T_z)_{ij} \equiv \phi_j(x_i^z)$, $M \equiv \binom{m+d-1}{d}$, and the function η_{md} is defined by:

$$\eta_{md}(q) = \begin{cases} \dfrac{(-1)^{m+1+d/2}}{2^{2m-1}\pi^{d/2}(m-1)!(m-d/2)!} q^{2m-d}\log(q) & \text{if } d \text{ is even} \\ \\ \dfrac{\Gamma(d/2-m)}{2^{2m}\pi^{d/2}(m-1)!} q^{2m-d} & \text{if } d \text{ is odd.} \end{cases}$$

The M functions ϕ_j are linearly independent polynomials that span the space of polynomials of degree less than m, and are thus not penalized at all by the penalty term J_{md}.

Now, defining E_z by $E_{ij}^z \equiv \eta_{md}(||x_i^z - x_j^z||)$, the minimization problem that defines the thin plate spline can alternatively be written as:

$$\min_{\delta_z,\alpha_z} \; ||y_z - E_z\delta_z - T_z\alpha_z||^2 + \lambda\delta_z' E_z\delta_z \quad \text{s.t.} \quad T_z'\delta_z = 0. \tag{15}$$

Leaving the basis for the unpenalized functions untouched, the TPRS focuses on truncating the basis for the penalized terms in a way that perturbs the minimization problem as little as possible. To elaborate, let k_z be the basis dimension for the TPRS chosen by the user. Instead of searching for the value of δ_z over the entire space $\mathbb{R}^{n_z}$ that (along with α_z) minimizes the objective function and satisfies the orthogonality constraint, the minimization problem that defines the TPRS only considers possible values of δ_z within a k_z-dimensional subspace, $\mathbf{W}_z$ of $\mathbb{R}^{n_z}$.

To make precise how the subspace $\mathbf{W}_z$ is chosen for TPRS (for a given value of k_z), I introduce the following notation. Given a k_z-dimensional subspace $\mathbf{W}_z$ of $\mathbb{R}^{n_z}$, let Γ_{k_z} be an $n_z \times k_z$ matrix of rank k_z with columns that form an orthonormal basis for $\mathbf{W}_z$. The TPRS minimization problem can then be written as:

$$\min_{\delta_k^z,\alpha_z} \; ||y_z - E_z\Gamma_{k_z}\delta_{k_z}^r - T_z\alpha_z||^2 + \lambda\delta_{k_r,z}'\Gamma_{k_z}'E\Gamma_{k_z}\delta_{k_r}^z \quad \text{s.t.} \quad T_z\Gamma_{k_z}\delta_{k_z}^z = 0$$

where $\delta_{k_z}^z \in \mathbb{R}^{k_z}$.

In order to express this in a form closer to that of the minimization problem for the thin plate spline in (15), I define the $n_z \times n_z$ matrices $\tilde{E}_{k_z} \equiv E_z\Gamma_{k_z}\Gamma_{k_z}'$ and $\hat{E}_{k_z} \equiv \Gamma_{k_z}\Gamma'_{k_z}E_z\Gamma_{k_z}\Gamma_{k_z}'$. This allows me to write the TPRS minimization problem as:

$$\min_{\delta_z,\alpha_z} \; ||y_z - \tilde{E}_z\delta_z - T_z\alpha_z||^2 + \lambda\delta_z'\hat{E}_{k_z}\delta_r \quad \text{s.t.} \quad T_z'\delta_z = 0, \tag{16}$$

since $\delta_z \in \mathbf{W}_z$ if and only if $\Gamma_{k_z}\delta_{k_z}^z = \delta_z$ for some $\delta_{k_z}^z \in \mathbb{R}^{k_z}$, by definition of $\mathbf{W}_z$.

Now, the goal of TPRS is to choose $\mathbf{W}_z$, or equivalently Γ_{k_z}, so that replacing the matrix

E_z in problem (15) by $\tilde{E}_{k_z}$ and $\hat{E}_{k_z}$ in problem (16) perturbs problem (15) as little as possible. Unfortunately, there is no k_z-dimensional subspace that minimizes the change in objective value for *all* possible values of δ_z. Hence, TPRS instead chooses Γ_{k_z} based on a minimax criterion, i.e., to minimize the *worst* possible change in objective value. In other words, Γ_{k_z} is taken to be the orthonormal basis matrix of rank k_z in $\mathbb{R}^{n_z \times k_z}$ that simultaneously minimizes:

$$\epsilon_{k_z} \equiv \max_{\delta_z \neq 0} \left\{ \frac{||(E_z - \tilde{E}_{k_z})\delta_z||}{||\delta_z||^2} \right\} \quad \text{and} \quad e_{k_z} \equiv \max_{\delta_z} \left\{ \frac{\delta_z'||(E_z - \hat{E}_{k_z})\delta_z||}{||\delta_z||^2} \right\},$$

where ϵ_k and e_k correspond to the worst possible change in the least squares and penalty terms respectively.

It turns out that the solution that simultaneously minimizes ϵ_{k_z} and e_{k_z} is a truncated eigenbasis of E_z. To elaborate, write the spectral decomposition of E_z as $E_z = U_z D_z U_z'$ where D_z is the diagonal matrix containing the eigenvalues of E_z, arranged in decreasing order of magnitude, i.e. $|D^z_{ii}| \geq |D^z_{i+1,i+1}|$ for $i = 1, ..., n_z - 1$.[46] Then, the solution Γ_{k_z} is the first k_z columns of U_z, appropriately scaled so that the columns are orthonormal. One may also verify that this solution results in $\tilde{E}_{k_z} = \hat{E}_{k_z}$.

G Details on Standard Error Calculations

For concreteness, in this section I discuss the standard error calculations for the sharp MRD case, but the procedure can be adapted for fuzzy MRD and MRK. Recall that the MRD estimate of the CATE is based on the difference between two fitted surfaces:

$$\hat{\tau}(x) = \hat{g}_1(x) - \hat{g}_0(x),$$

where $\hat{g}_1$ and $\hat{g}_0$ were estimated based on observations in the treated and untreated regions respectively. Generally, these functions can be written in the following form:

$$\hat{g}_z(x) = \sum_{k=1}^{K_z} \hat{\beta}_{z,k} s_{z,k}(x),$$

where $s_{z,k}(x)$ is the kth basis function of the thin plate spline that is fit to the region where individuals receive treatment z. Adopting a Bayesian perspective, for appropriately chosen priors, the parameter vectors $\hat{\beta}_{1,k}$ and $\hat{\beta}_{0,k}$ both have Gaussian posterior distributions, so let us denote their posterior covariance matrices by Σ_1 and Σ_0 respectively.[47] Also, if we assume i.i.d. error terms, then the fact that we fit $\hat{g}_1$ and $\hat{g}_0$ using separate data implies that the two parameter vectors are independent from each other, so that the entire posterior covariance

[46] This decomposition is possible because E_z is a real symmetric matrix by definition.

[47] See Wahba (1990) and Wood (2006) for choices of priors that lead to these posterior distributions.

matrix Σ is block diagonal with diagonal blocks Σ_1 and Σ_0.

Based on the discussion above, standard error estimates of the CATE estimates at various points of the treatment frontier can be computed. Specifically, for $x \in \mathbb{F}$, we have:

$$\begin{aligned} \hat{Var}\left(\hat{\tau}_m(x)\right) &= Var\left(\sum_{k=1}^{K_1} \hat{\beta}_{1,k} s_{1,k}(x) - \sum_{k=1}^{K_0} \hat{\beta}_{0,,k} s_{0,k}(x)\right) \\ &= Var\left(\sum_{k=1}^{K_1} \hat{\beta}_{1,k} s_{1,k}(x)\right) + Var\left(\sum_{k=1}^{K_0} \hat{\beta}_{0,k} s_{0,k}(x)\right) \\ &= s_1(x)'\hat{\Sigma}_1 s_1(x) + s_0(x)'\hat{\Sigma}_0 s_0(x). \end{aligned}$$

Moreover, for a finite collection of points $\{x_g\}_{g\in\mathcal{G}}$, $x_g \in \mathbb{F}$, we can compute an estimate of the covariance matrix for the CATE at these points. In particular, we have:

$$\hat{Var}\left(\{\hat{\tau}_m(x_g)\}\right)_{g,g'} = s_1(x^g)'\hat{\Sigma}_1 s_1(x^{g'}) + s_0(x^g)'\hat{\Sigma}_0 s_0(x^{g'}),$$

which can be written in matrix form:

$$\hat{Var}\left(\{\hat{\tau}_m(x_g)\}\right) = S_1'\hat{\Sigma}_1 S_1 + S_0'\hat{\Sigma}_0 S_0,$$

where S_z is the $K \times |\mathcal{G}|$ matrix with (k, g)th element equal to $s_{z,k}(x^g)$.

Suppose we want to compute the average treatment effect over a subset of the treatment frontier $\mathbb{F}$. This can be done via numerical integration based on a discrete grid of points $\{x_g\}_{g\in\mathcal{G}}$, $x_g \in \mathbb{F}$. If we want to average the treatment effect using deterministic weights $w_1, ..., w_{|\mathcal{G}|}$ that are positive and sum to one (e.g. based on some known counterfactual population of interest), then we can simply compute the average effect as $\sum_{g\in\mathcal{G}} w_g \hat{\tau}(x^g)$, and estimate the variance as $w'\hat{Var}\left(\{\hat{\tau}(x_g)\}\right)w$.

If we instead want to compute the average effect over the population (from which the estimation sample is randomly drawn from), but the distribution of the running variables for this population is unknown, then we need to estimate the density. Many methods for density estimation yield estimates $\{\hat{f}(x^g)\}_{g\in\mathcal{G}}$ that are asymptotically normal, with some covariance matrix $\hat{\Sigma}^f$. To use these as weights, we need to scale them so that they sum to one. Hence, we can estimate the average treatment effect using:

$$\sum_{g\in\mathcal{G}} \frac{\hat{f}(x^g)}{\sum_{g'\in\mathcal{G}} \hat{f}(x^{g'})} \hat{\tau}(x_g).$$

Denoting $\hat{\Sigma}^\tau \equiv \hat{Var}\left(\{\hat{\tau}_m(x_g)\}_{g\in\mathcal{G}}\right)$, and assuming independence between the estimates $\{\hat{\tau}(x_g)\}_{g\in\mathcal{G}}$ and $\{\hat{f}(x_g)\}_{g\in\mathcal{G}}$, we can obtain an estimate of the asymptotic variance of average treatment effect estimate using the delta method.

Specifically, consider the function:

$$t(f,\tau) \equiv \sum_{g\in\mathcal{G}} \frac{f^g}{\sum_{g'\in\mathcal{G}} f^{g'}}\tau^g,$$

for $f \in \mathbb{R}_+^{|\mathcal{G}|}$, $\tau \in \mathbb{R}^{|\mathcal{G}|}$. The gradient of this function is:

$$\nabla t(\tau, f) = \begin{pmatrix} D_\tau t(\tau, f) \\ D_f t(\tau, f) \end{pmatrix},$$

where:

$$D_\tau t(\tau, f) = \begin{pmatrix} \dfrac{f^1}{\sum_{g\in\mathcal{G}} f^g} \\ \vdots \\ \dfrac{f^{|\mathcal{G}|}}{\sum_{g\in\mathcal{G}} f^g} \end{pmatrix}, \; D_f t(\tau, f) = \begin{pmatrix} \sum_{g\in\mathcal{G}} \dfrac{f^g}{\left(\sum_{g'\in\mathcal{G}} f^{g'}\right)^2}\left(\tau^1 - \tau^g\right) \\ \vdots \\ \sum_{g\in\mathcal{G}} \dfrac{f^g}{\left(\sum_{g'\in\mathcal{G}} f^{g'}\right)^2}\left(\tau^{|\mathcal{G}|} - \tau^g\right) \end{pmatrix}.$$

So, we can estimate the asymptotic variance of the average treatment effect estimate over $\mathbb{F}$ via the delta method using the following formula:

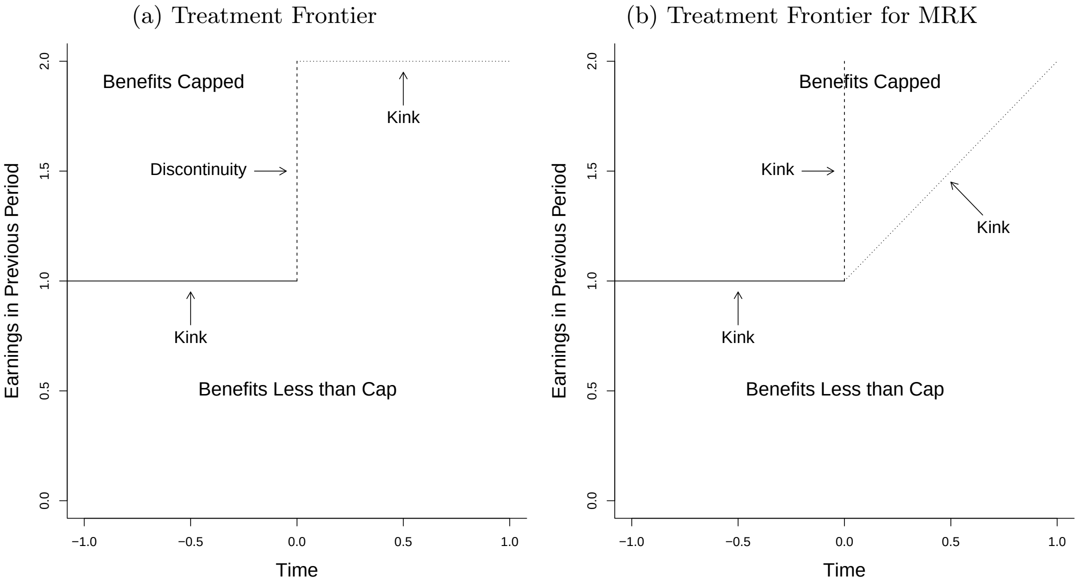
$$\begin{aligned} \nabla t(\hat{\tau}, \hat{f})' \begin{pmatrix} \hat{\Sigma}^\tau & 0 \\ 0 & \hat{\Sigma}^f \end{pmatrix} \nabla t(\hat{\tau}, \hat{f}) &= \nabla t(\hat{\tau}, \hat{f})' \begin{pmatrix} \hat{\Sigma}^\tau D_\tau t(\hat{\tau}, \hat{f}) \\ \hat{\Sigma}^f D_f t(\tau, f) \end{pmatrix} \\ &= D_f t(\hat{\tau}, \hat{f})' \hat{\Sigma}^f D_f t(\hat{\tau}, \hat{f}) + D_\tau t(\hat{\tau}, \hat{f})' \hat{\Sigma}^\tau D_\tau t(\hat{\tau}, \hat{f}). \end{aligned}$$

More generally, this same methodology can be used to estimate average treatment effects over other subsets of the treatment frontier $\mathbb{F}$ (using estimates of the density corresponding to those subsets), and to obtain standard errors for these estimates. Furthermore, if we were interested in the weighted average treatment effect over the treatment frontier with respect to counterfactual distributions of individuals (e.g., f^* instead of f), then we need not estimate f and account for this uncertainty when estimating the standard error.

Finally, we assumed independence between the estimates $\{\hat{\tau}_m(x_g)\}_{g\in\mathcal{G}}$ and $\{\hat{f}(x_g)\}_{g\in\mathcal{G}}$ in our derivation of the standard error for the average CATE estimate above. A potential concern with this assumption is that individuals may have varying tastes for treatment based on unobservables and may exert differential amounts of effort in manipulating their running variables in order to receive treatment, thus leading to a correlation between the density estimates and the CATE estimates. However, such a scenario is somewhat unlikely given the local randomization interpretation of RD designs (Lee 2008), which posits that individuals have imprecise control over the running variables.

Appendix Figures and Tables

Figure A.1: Treatment Frontiers for MRDK and MRK Designs

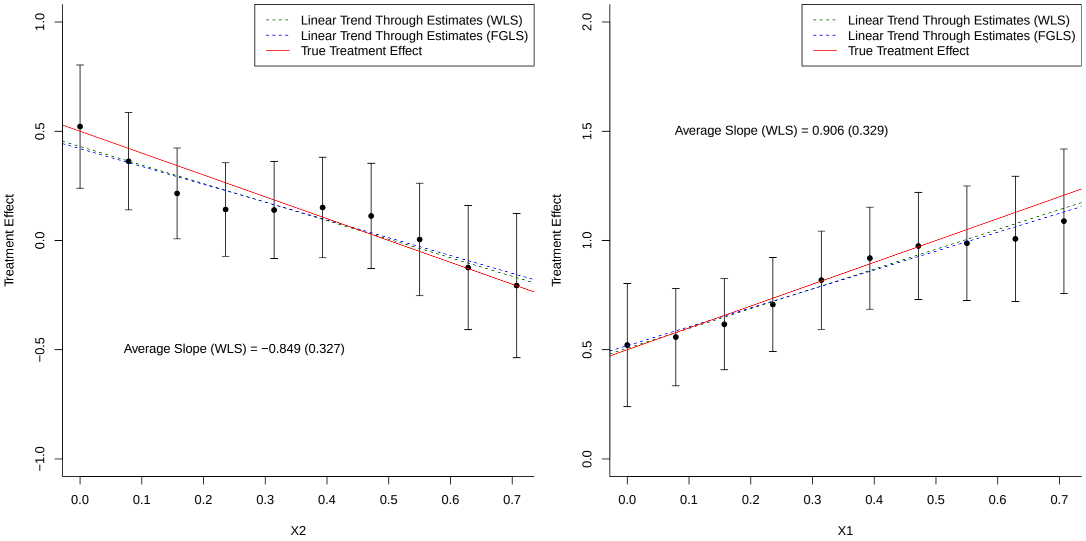


(a) Treatment Frontier

(b) Treatment Frontier for MRK

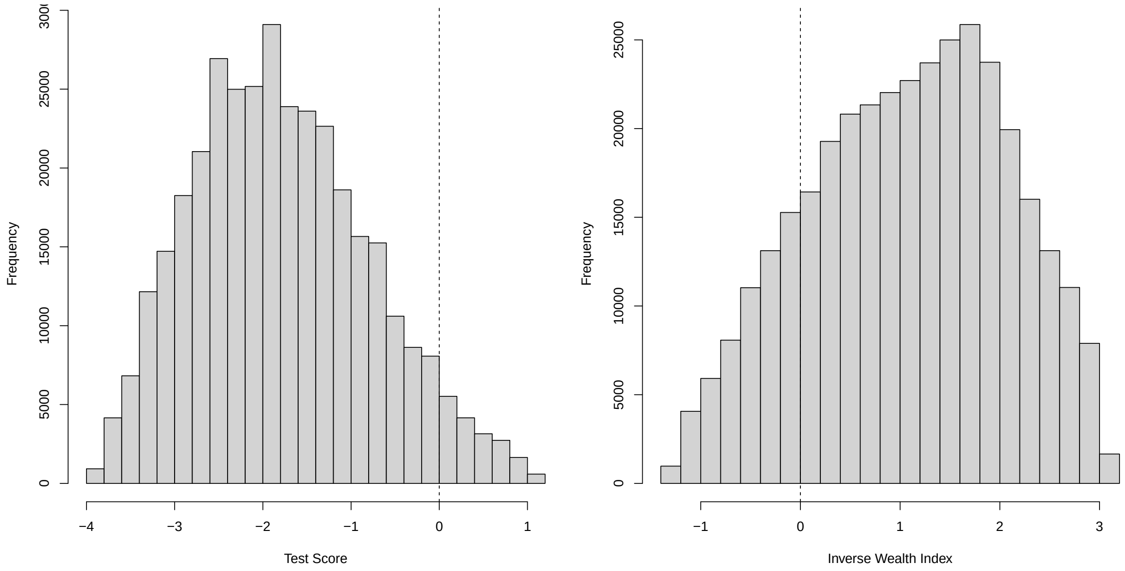
Notes: Panel B shows the treatment frontier $\mathbb{F}$ as well as regions where individuals are or are not subject to the cap (Ω_1 and Ω_0 respectively) for the MRK design.

Figure A.2: MRD Estimates of Heterogeneous Treatment Effects in the Simulations



Notes: The figures show MRD estimates of the CATE from the DGP with heterogeneous treatment effects, separately from the portion of the frontier $\mathbb{F}$ where $X_{1i} = 0$ in the left figure, and the portion where $X_{2i} = 0$ in the right figure. The true tratment effect is shown as a solid red line. Error bars indicate 95 percent pointwise confidence intervals. Standard errors for the slope coefficients for the WLS fits are computed via bootstrap. Standard errors for the slope coefficients for FGLS fits are analytic standard errors.

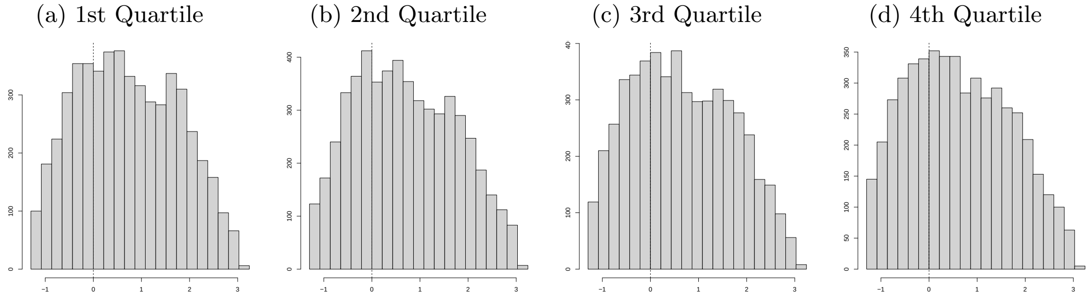
Figure A.3: Histograms of Running Variables


(a) Normalized SABER 11 Score

(b) Normalized SISBEN Wealth Index

Notes: The figure shows univariate histograms of the running variables.

Figure A.4: Histograms of Wealth Index for Students from Different Quartiles of Test Scores


(a) 1st Quartile

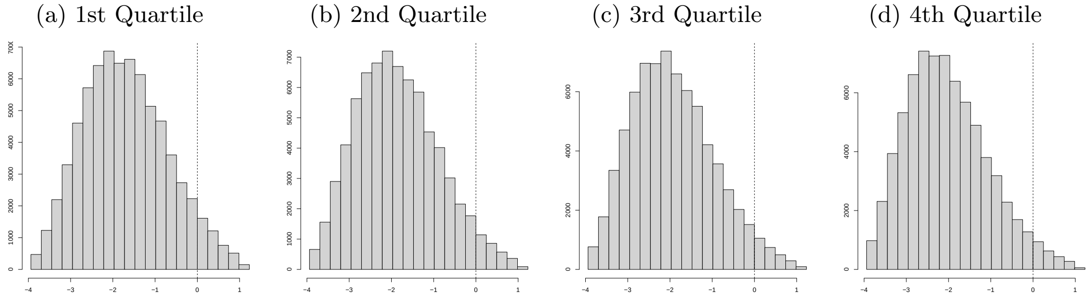
(b) 2nd Quartile

(c) 3rd Quartile

(d) 4th Quartile

Notes: The figures show histograms of the inverse wealth index, for observations from different quartiles of test scores. The threshold is shown by the vertical line.

Figure A.5: Histograms of Test Scores for Students from Different Quartiles of Wealth Index


(a) 1st Quartile

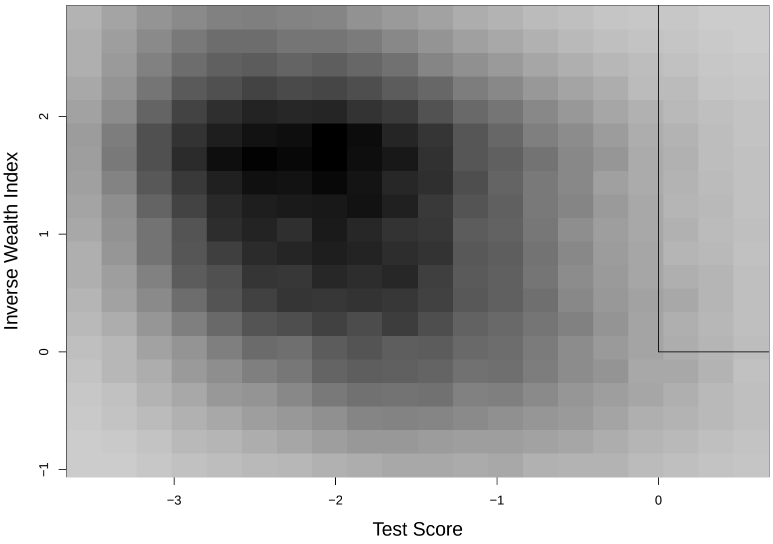
(b) 2nd Quartile

(c) 3rd Quartile

(d) 4th Quartile

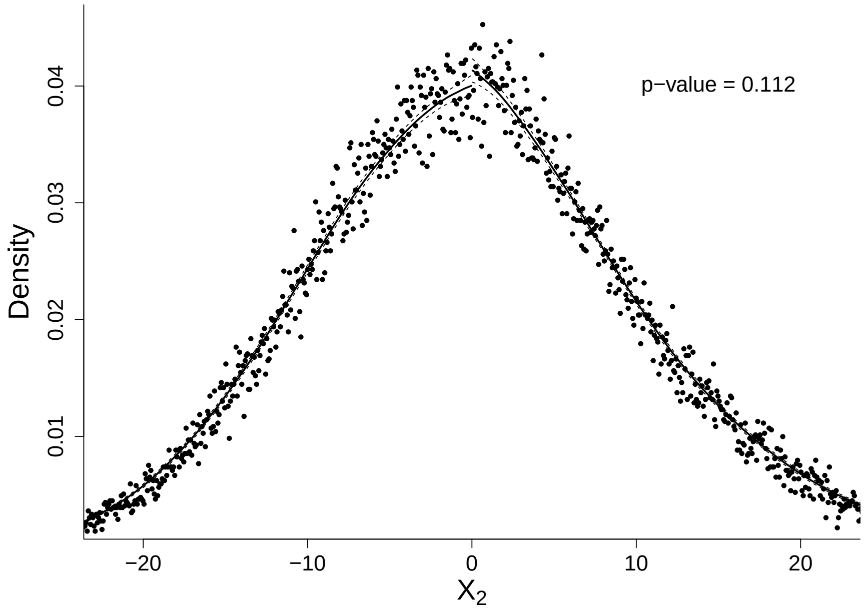
Notes: The figures show histograms of test scores, for observations from different quartiles of the inverse wealth index. The threshold is shown by the vertical line.

Figure A.6: Contour Plot for Joint Distribution of Test Scores and Inverse Wealth Index

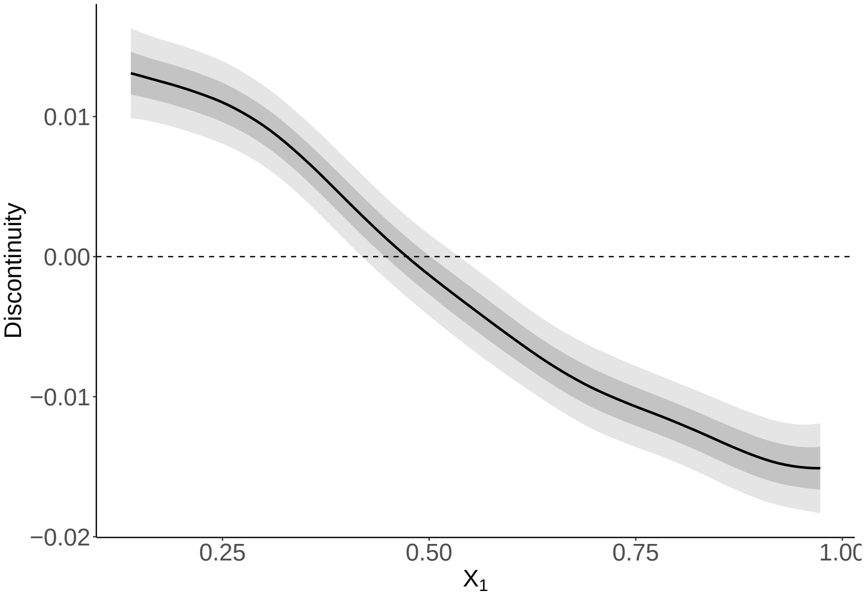


Notes: The figure shows a contour plot from a two-dimensional histogram of test scores and the inverse wealth index. The treatment frontier $\mathbb{F}$ is shown by the solid lines.

Figure A.7: Simulation Results for Discontinuity in the Multivariate Density



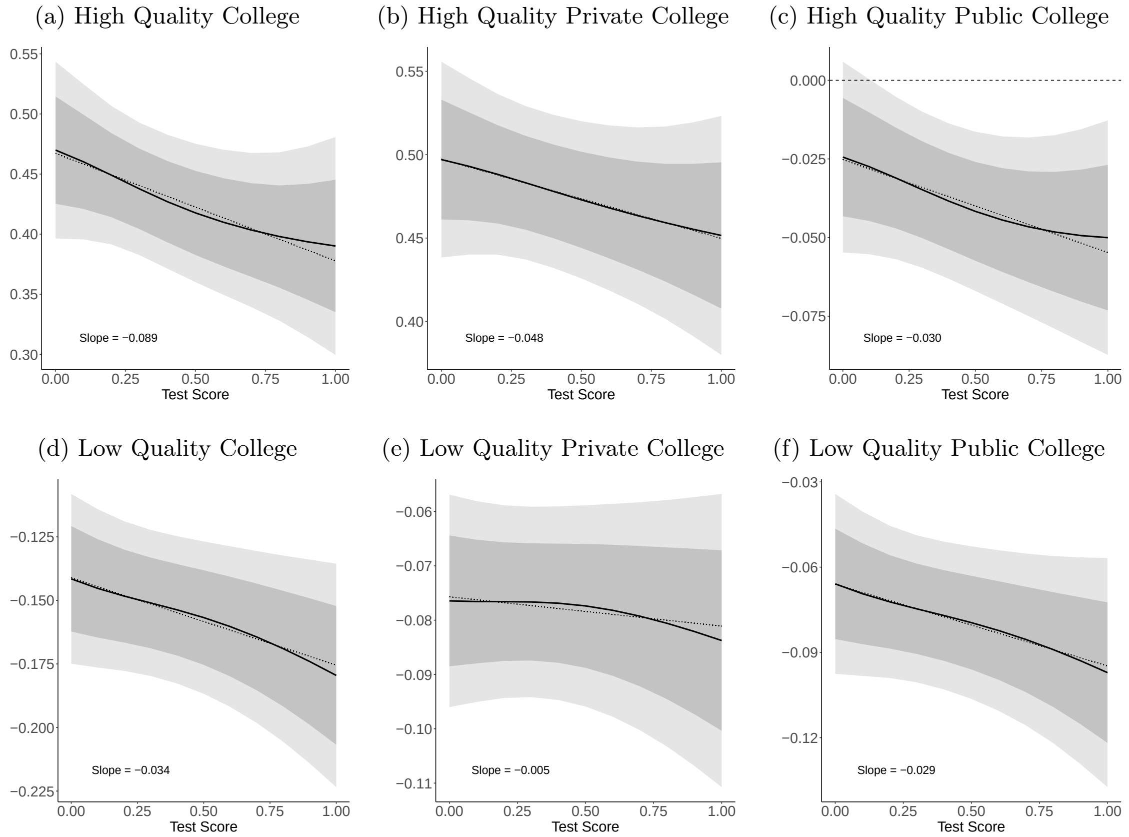
(a) Single-Dimensional McCrary Test



(b) Two-Dimensional McCrary Test

Notes: The first figure shows the result from a single-dimensional McCrary test from the data-generating process described in Appendix section D, whereas the second figure shows the results from the two-dimensional McCrary test described in the same section. The shaded regions in light grey and dark grey in the second figure represent the 95 percent pointwise confidence intervals and 95 percent simultaneous confidence bands respectively.

Figure A.8: Effect of the SPP on Enrollment as a Function of Test Scores: Different Types of Colleges



(a) High Quality College

(b) High Quality Private College

(c) High Quality Public College

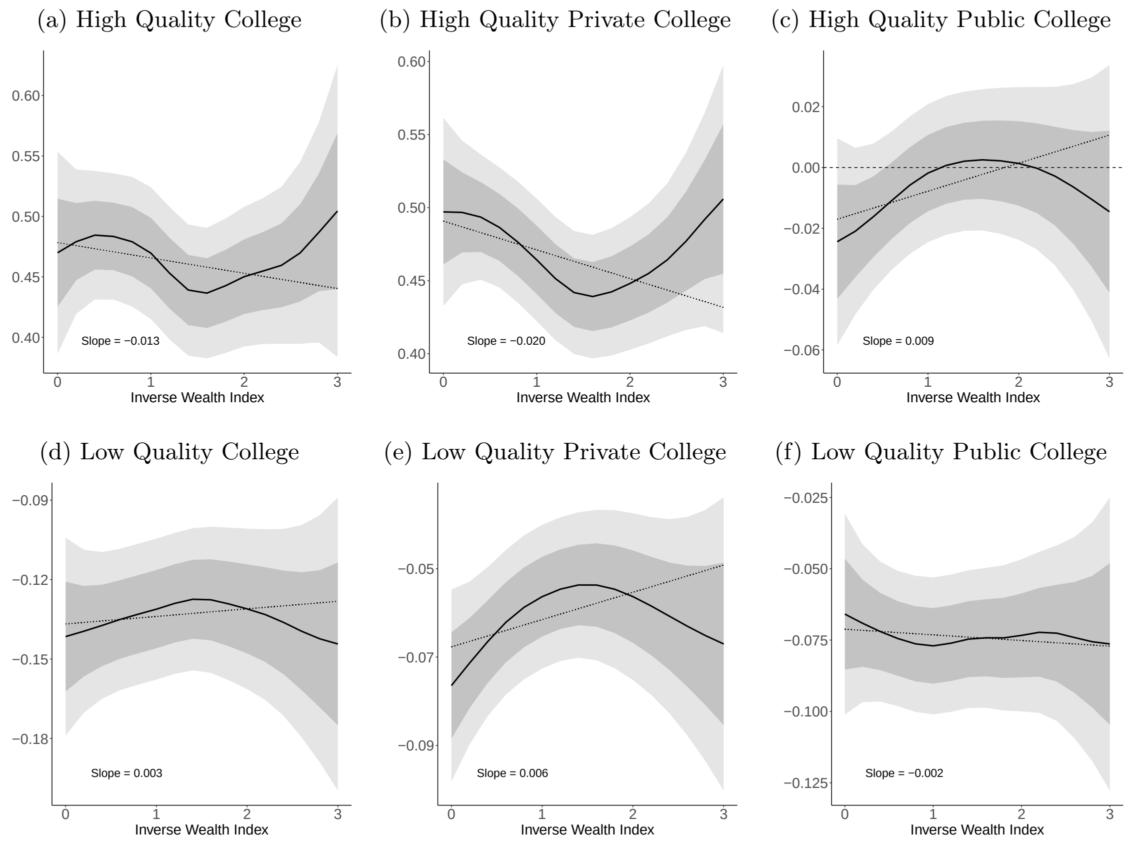
(d) Low Quality College

(e) Low Quality Private College

(f) Low Quality Public College

Notes: The figures show MRD estimates of the CATE on the effect of financial aid on the probability of enrollment in different types of colleges as a function of test scores, for students with inverse wealth indices at the cutoff. The shaded regions in light grey and dark grey represent the 95 percent pointwise confidence intervals and 95 percent simultaneous confidence bands respectively.

Figure A.9: Effect of the SPP on Enrollment as a Function of Inverse Wealth Index: Different Types of Colleges



(a) High Quality College

(b) High Quality Private College

(c) High Quality Public College

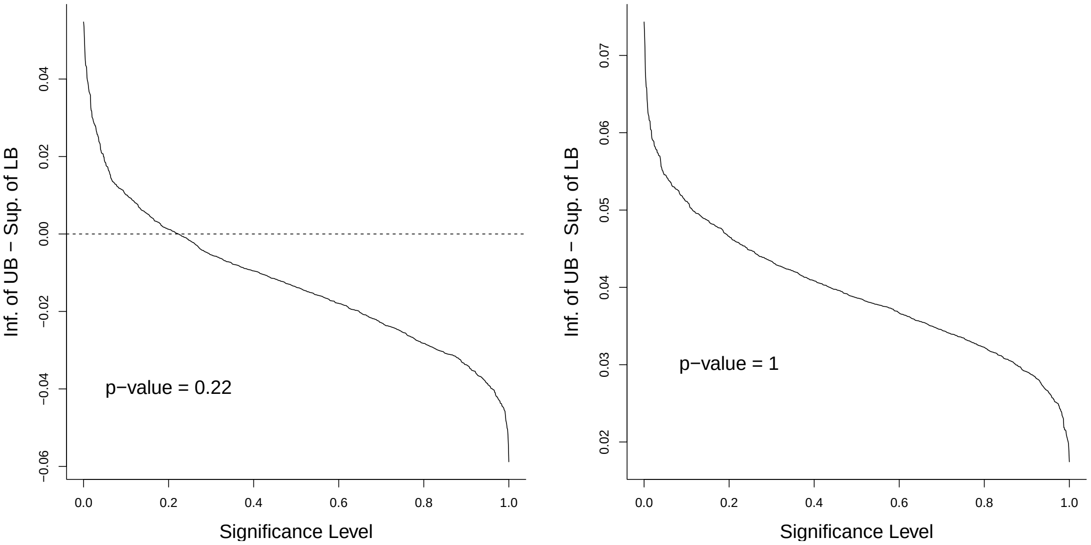
(d) Low Quality College

(e) Low Quality Private College

(f) Low Quality Public College

Notes: The figures show MRD estimates of the CATE on the effect of financial aid on the probability of enrollment in different types of colleges as a function of the inverse wealth index, for students with test scores at the cutoff. The shaded regions in light grey and dark grey represent the 95 percent pointwise confidence intervals and 95 percent simultaneous confidence bands respectively.

Figure A.10: Tests of Constant Treatment Effects



(a) H_0: $\tau(x)$ is Constant in Test Scores for all $x \in \mathbb{F}$

(b) H_0: $\tau(x)$ is Constant in Wealth for all $x \in \mathbb{F}$

Notes: The figures plot $\inf_x \{\tau(x) + \bar{c}_{1-\alpha}(\mathbb{F}_d)\hat{se}(\hat{\tau}(x))\} - \sup_x \{\tau(x) - \bar{c}_{1-\alpha}(\mathbb{F}_d)\hat{se}(\hat{\tau}(x))\}$ as a function of α for values of x in the segment of the treatment frontier corresponding to the wealth threshold in panel (a), and for values of x in the segment in the corresponding to the test score threshold in panel (b).

Table A.1: MRD Simulation Results: Pointwise Confidence Intervals and Simultaneous Confidence Bands

Panel A. Estimates of the Treatment Effect (TE) Over $\{X_1=0, X_2\geq 0\}$

		Pointwise (Analytic)		**Pointwise (Bootstrap)**		**Simultaneous Confidence Bands**	
MRD Estimator	DGP	Coverage	Length	Coverage	Length	Coverage	Length
MSE-Optimal	Constant TE	0.94	0.318	0.95	0.321	0.99	0.851
Bias-Corrected	Constant TE	0.94	0.329	0.97	0.356	0.99	0.893
Undersmoothing	Constant TE	0.94	0.333	0.94	0.337	0.99	0.911
MSE-Optimal	Heterogeneous TE	0.927	0.498	0.935	0.508	0.99	0.851
Bias-Corrected	Heterogeneous TE	0.922	0.522	0.959	0.586	0.99	0.893
Undersmoothing	Heterogeneous TE	0.930	0.532	0.935	0.544	0.99	0.911

Panel B. Estimates of the Treatment Effect (TE) Over $\{X_1\geq 0, X_2=0\}$

		Pointwise (Analytic)		**Pointwise (Bootstrap)**		**Simultaneous Confidence Bands**	
MRD Estimator	DGP	Coverage	Length	Coverage	Length	Coverage	Length
MSE-Optimal	Constant TE	0.95	0.320	0.95	0.323	0.99	0.858
Bias-Corrected	Constant TE	0.95	0.331	0.96	0.360	0.99	0.901
Undersmoothing	Constant TE	0.95	0.335	0.95	0.340	0.99	0.918
MSE-Optimal	Heterogeneous TE	0.944	0.502	0.948	0.510	0.99	0.858
Bias-Corrected	Heterogeneous TE	0.942	0.527	0.957	0.589	0.99	0.899
Undersmoothing	Heterogeneous TE	0.945	0.536	0.945	0.547	0.99	0.917

Notes: Three versions of the MRD estimator are considered in these simulations: an estimator using the MSE-optimal choice of penalty parameter for the thin plate regression splines (TPRS), a bias-corrected estimator using the MSE-optimal penalty parameter from higher-order TPRS, and an undersmoothed estimator using half of the MSE-optimal penalty parameter. The results shown in this table are based on 100 realizations of the DGP with either constant or heterogeneous treatment effects. Analytic standard errors are based on the posterior distribution of the thin plate spline estimates, whereas bootstrap confidence intervals are constructed using nonparametric bootstrap. Pointwise confidence intervals and simultaneous confidence bands are based on a 5 percent significance level. See text for more details on these simulations.

Table A.2: Heterogeneity in the Effect of Political Ads on Voter Turnout

	Turnout	Effect of Ads on Turnout	Effect of Ads on Turnout (Weighted)
	(1)	(2)	(3)
Education (Years)	0.005	-0.111***	-0.066**
	(0.007)	(0.036)	(0.032)
Unemployment	0.383*	-11.085***	-9.598***
	(0.198)	(1.185)	(1.023)
Poverty	0.184*	2.688***	2.500***
	(0.125)	(0.618)	(0.577)
Income (in thousands)	0.001***	0.000	-0.001
	(0.000)	(0.001)	(0.001)
Age	0.010***	0.001**	0.001**
	(0.000)	(0.000)	(0.000)
Voter Registration	0.170***	0.050	0.074*
	(0.015)	(0.061)	(0.064)
Constant	-0.111*	1.923***	1.228**
	(0.105)	(0.538)	(0.478)
Number of Observations	24,460	89	89

Notes: The first two columns are OLS regressions, and the third column is a weighted least squares regression with weights equal to the inverse of the variance of the MRD CATE estimates. Standard errors are shown in parentheses.