CAUSAL INFERENCE ON REGRESSION DISCONTINUITY DESIGNS BY HIGH-DIMENSIONAL METHODS

YOICHI ARAI, TAISUKE OTSU, AND MYUNG HWAN SEO

Abstract. In causal or treatment effect analysis, discontinuities in regression functions induced by an assignment variable can be utilized to retrieve causal effects. The regression discontinuity design (RDD) has been extensively employed in the literature to identify the average treatment effect at the discontinuity point. This paper proposes an estimation and inference method for causal parameters identified by the RDD based on high-dimensional statistical techniques. Our methods are practical and competitive with the existing kernel-based local methods.

1. Introduction

In causal or treatment effect analysis, discontinuities in regression functions induced by an assignment variable can provide useful information to identify certain causal effects. The regression discontinuity design (RDD) has been widely applied in observational studies to identify the average treatment effect at the discontinuity point. For the RDD, the causal parameters of interest are identified by some contrasts of the left and right limits of the conditional mean functions.

This paper proposes an estimation and inference method for causal parameters identified by the RDD based on high-dimensional statistical techniques. Our method is practical and competitive with the existing kernel-based local methods, such as Imbens and Kalyanaraman (2012). In particular, we interpret the estimation problem of the causal effect parameter in the sharp RDD as that of a slope coefficient in a partially linear model, and then estimate the parameter by the penalized least squares, such as the lasso. As discussed in Gelman and Imbens (2018), there are some problems of using global (and low-dimensional) polynomial regressions to estimate causal effects in the RDD. This paper argues that the penalized least square approach using relatively high-dimensional basis functions can be useful to alleviate those problems. Our high-dimensional approach is naturally extended to other setups, such as the fuzzy RDD and regression kink design (RKD) studied by e.g., Card, *et al.* (2015) and Ganong and Jäger (2018).

As theoretical contributions, this paper proposes asymptotic and bootstrap inference methods for the causal parameters in the RDD. Our inference problem can be formulated as the one for low-dimensional parameters in high-dimensional models. In statistics literature, many papers investigated this issue, such as Belloni, Chernozhukov and Hansen (2014), van de Geer, *et al.* (2014), and Zhang and Zhang (2014). However, these approaches are not directly applicable to the RDD context because the current problem is inference on a jump in a nonparametric regression model with a single regressor.

Otsu gratefully acknowledges financial support from the ERC Consolidator Grant (SNP 615882).

This paper also contributes to the growing literature for estimation and inference on the RDD. Imbens and Lemieux (2008) provided a comprehensive survey on early literature. Imbens and Kalyanaraman (2012) and Arai and Ichimura (2018) studied optimal bandwidth selection methods for kernel-based estimators. Calonico, Cattaneo and Titiunik (2014) proposed a robust confidence interval for kernel-based estimators. Calonico, Cattaneo and Titiunik (2015) developed an optimal data-driven method for the RDD plots. We emphasize that most methodology papers on the RDD analysis focus on the kernel-based local methods. On the other hand, this paper advocates an alternative series-based approach using high-dimensional methods.

This paper is organized as follows. Section 2.1 presents our basic setup and estimator for the sharp RDD. In Section 2.2, we propose an inference method on the causal effect in the sharp RDD. In Section 3, we discuss extensions to the fuzzy RDD (Section 3.1), the case of unknown discontinuity point (Section 3.2), and the RKD (Section 3.3). Section 4 presents some simulation results.

2. Main result

2.1. **Setup and point estimation.** In this section, we present our basic setup and point estimator for the RDD. For each unit $i = 1, \dots, n$, we observe an indicator variable W_i for a treatment ($W_i = 1$ if treated and $W_i = 0$ otherwise), and outcome $Y_i = \begin{cases} Y_i(0) & \text{if } W_i = 0, \\ Y_i(1) & \text{if } W_i = 1, \end{cases}$, where $Y_i(0)$ and $Y_i(1)$ are potential outcomes for $W_i = 0$ and $W_i = 1$, respectively. Note that we cannot observe $Y_i(0)$ and $Y_i(1)$ simultaneously. Our purpose is to make inference on the causal effect of the treatment, or more specifically, some distribution aspects of the difference of potential outcomes $Y_i(1) - Y_i(0)$. The RDD analysis focuses on the case where the treatment assignment W_i is completely or partly determined by some observable covariate X_i, called the forcing variable. For example, to study the effect of class size on pupils' achievements, it is reasonable to consider the following setup: the unit i is school, Y_i is an average exam score, W_i is an indicator variable for the class size ($W_i = 0$ for one class and $W_i = 1$ for two classes), and X_i is the number of enrollments.

Depending on the assignment rule for W_i based on X_i, we have two cases, called the sharp and fuzzy RDDs. In this section, we focus on the sharp RDD and discuss the fuzzy RDD in Section 3.1. In the sharp RDD, the treatment is deterministically assigned based on the value of X_i, i.e.

$$W_i = \mathbb{I}\{X_i \geq c\},$$

where $\mathbb{I}\{\cdot\}$ is the indicator function and c is a known cutoff point (the case of unknown cutoff will be briefly discussed in Section 3.2). A parameter of interest in this case is the average causal effect at the discontinuity point c,

$$\theta_0 = E[Y_i(1) - Y_i(0) | X_i = c].$$

Since the difference of potential outcomes $Y_i(1) - Y_i(0)$ is unobservable, we need a tractable representation of θ_0 in terms of quantities that can be estimated by data. If the conditional mean functions $E[Y_i(1)|X_i = x]$ and $E[Y_i(0)|X_i = x]$ are continuous at the cutoff point $x = c$,

then the average causal effect θ_0 can be identified as a contrast of the left and right limits of the conditional mean $E[Y_i|X_i = x]$ at $x = c$,

$$\theta_0 = \lim_{x \downarrow c} E[Y_i|X_i = x] - \lim_{x \uparrow c} E[Y_i|X_i = x]. \tag{2.1}$$

In the literature, it is common to employ some nonparametric kernel-based method to estimate this object. In contrast, we estimate θ_0 by the following regression model

$$Y_i = W_i\theta_0 + m(X_i) + \epsilon_i,$$

where ϵ_i is an error term satisfying $E[\epsilon_i|X_i] = 0$, and $m(\cdot)$ is a function continuous (but typically non-differentiable) at c. Note that this model is different from the conventional partially linear model because $W_i = \mathbb{I}\{X_i \geq c\}$ is also a function of X_i. We propose to estimate this model by the lasso regression

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left\{ Y_i - W_i\theta - \alpha - W_i M_i'\gamma_R - (1 - W_i)M_i'\gamma_L \right\}^2 + \lambda_n |\beta|_1, \tag{2.2}$$

where $\beta = (\alpha, \theta, \gamma_R', \gamma_L')'$ is a vector of parameters, $M_i = (m^{(1)}(X_i), \ldots, m^{(p)}(X_i))'$ is a vector of basis functions (or dictionaries) to approximate $m(\cdot)$ evaluated at X_i without intercept, $|\beta|_1 = \sum_{j=1}^{2p+2} |\beta_j|$ is the ℓ_1-norm of the parameter vector, and λ_n is a penalty level. The estimator $\hat{\theta}$ of the average causal effect θ_0 is obtained by the estimated coefficient of W_i.

Under mild regularity conditions, the lasso yields a consistent estimator for the nonparametric regression function without knowing a priori what are the most significant basis functions of the series estimation. The resulting estimation error bound for the regression function is as sharp as the oracle who knows the identity of the relevant basis functions up to a logarithmic factor (see, e.g., Bühlmann and van de Geer, 2011).

It may be possible to estimate θ_0 based on certain linear combination of the estimated coefficients of lasso regression from Y_i on $\{W_i(1, M_i), (1 - W_i)(1, M_i)\}$. Although this approach yields the same OLS estimate for θ_0 (for low-dimensional case), the lasso estimates are generally different. Our preliminary simulation results suggest that the lasso estimate based on such a linear combination is relatively unstable because the estimate of θ_0 depends on all the regression coefficients. Therefore, we recommend the parametrization in (2.2) because it directly estimates the parameter of interest by the coefficient of the single regressor W_i.

Regarding the choice of the basis functions M_i, there are several possibilities, such as polynomials, splines, Fourier series, wavelets, and mixed versions of them. We recommend to employ a sufficiently flexible M_i with large dimension and let the lasso choose the adequate subset.

An insightful paper by Gelman and Imbens (2018) pointed out the problems of using global (and low-dimensional) polynomial regressions estimated by the OLS. They raised three issues, and our method can be understood as a way to alleviate those issues. First, Gelman and Imbens (2018) argued that the weights in the weighted average representation of the estimator of θ_0 are highly sensitive to the order of polynomials. Although it is generally difficult to characterize the lasso estimator by the weighted average form, some shrinkage estimators, such as the ridge, may be interpreted as attempts to stabilize those weights. Second, Gelman and Imbens (2018)

illustrated that the estimate of θ_0 based on the OLS polynomial fits is too sensitive to their order. Again, since the lasso is a shrinkage estimator that attempts to stabilize the estimates by sacrificing the bias, our method can alleviate this issue. Third, Gelman and Imbens (2018) also argued that inference based on the OLS polynomial fits does not typically achieve the nominal level. Inference is another important issue for the RDD analysis, which will be discussed in the next section.

2.2. **Inference.** We now consider inference on the causal parameter θ_0. Let $F_i = (M_i'W_i, M_i'(1-W_i))'$ and define the demeaned variables $(y_i, w_i, f_i') = (Y_i - \bar{Y}, W_i - \bar{W}, (F_i - \bar{F})')$. We propose the following procedure to construct the confidence interval of θ_0, which is different from the existing methods, such as Zhang and Zhang (2014).

Asymptotic confidence interval for the RDD causal effect:

(1) Run the lasso regression from y_i on (w_i, f_i) to obtain

$$(\tilde{\theta}, \tilde{\gamma}') = \arg\min_{\theta,\gamma} \frac{1}{n}\sum_{i=1}^{n}(y_i - w_i\theta - f_i'\gamma)^2 + \lambda_n|\gamma|_1, \tag{2.3}$$

where $\lambda_n = A\sigma\sqrt{\frac{2\log(2p/\varsigma)}{n}}$ with $A > 1$, $\sigma^2 = E[\epsilon_i^2]$, and $0 < \varsigma \leq 1$, and the corresponding residual $\tilde{e}_i = y_i - w_i\tilde{\theta} - f_i'\tilde{\gamma}$.

(2) Compute the threshold lasso estimator $\hat{\gamma}^j = |\tilde{\gamma}^j|\mathbb{I}\{|\tilde{\gamma}^j| > a_n\}$ for each j and some $a_n \gg \lambda_n s$, where s is defined Assumption 1 in Appendix. Also let $S_n = \{j : |\tilde{\gamma}^j| > a_n\}$ and $f_{S_n,i}$ and γ_{S_n} be subvectors of f_i and γ that consist of the indexes in S_n. In practice, we may set $a_n = \lambda_n \sum_{j=1}^{2p}\mathbb{I}\{|\tilde{\gamma}^j| > 0\}\log n$.

(3) Run the OLS from w_i on $f_{S_n,i}$ to obtain the OLS estimates $\tilde{\delta}$ and residuals $\tilde{\zeta}_i = w_i - f_{S_n,i}'\tilde{\delta}$.

(4) Compute the bias corrected estimator

$$\bar{\theta} = \left(\sum_{i=1}^{n}\tilde{\zeta}_i w_i\right)^{-1}\sum_{i=1}^{n}\tilde{\zeta}_i(y_i - f_i'\hat{\gamma}), \tag{2.4}$$

and residuals $\bar{e}_i = y_i - w_i\bar{\theta} - f_i'\hat{\gamma}$.

(5) Report the $100(1-a)\%$ confidence interval for θ_0 as

$$\left[\bar{\theta} \pm z_{1-a/2}\left(\sum_{i=1}^{n}\tilde{\zeta}_i^2\right)^{-1}\left(\sum_{i=1}^{n}\tilde{\zeta}_i^2\bar{e}_i^2\right)^{1/2}\right], \tag{2.5}$$

where $z_{1-a/2}$ is the $(1-a/2)$-th quantile of the standard normal distribution.

Validity of this confidence interval is shown by the following theorem.

Theorem 1. *Under Assumptions 1-4 in Appendix,*

$$\left(\sum_{i=1}^{n}\tilde{\zeta}_i^2\right)\left(\sum_{i=1}^{n}\tilde{\zeta}_i^2\bar{e}_i^2\right)^{-1/2}(\bar{\theta} - \theta_0) \xrightarrow{d} N(0,1).$$

There are growing interests in developing inference methods for low-dimensional parameters in high-dimensional models, which are estimated by penalized estimation methods such as the (bias-corrected) lasso (see, e.g., Zhang and Zhang, 2014, and Belloni, Chernozhukov and Hansen,

2014). However, these approaches are not directly applicable to the current setup because we are concerned with a nonparametric regression problem with a single regressor. That is, the regressor of interest W_i is a function of X_i and can be approximated with an arbitrarily small approximation error by linear combinations of the other regressors. For example, the setup of Belloni, Chernozhukov and Hansen (2014, eq. (2.3)) does not cover ours, nor does Zhang and Zhang (2014, Theorem 1) that imposes a similar restriction (through the constant η_i in their notation).

We now discuss how to select the tuning parameter λ_n. Since variable selection is embedded in the lasso regression, the standard error σ is difficult to estimate unlike in fixed-dimensional regressions. One way to deal with this issue is the iteration, in which the first step takes the sample standard deviation of y as an estimate of σ in the construction of λ_n and then take the standard deviation of the resulting residuals. Sometimes, however, the sample standard deviation of y is too large, which means λ_n is too large for the lasso to select any variables. In this case, we may adjust the initial choice for σ as λ_n/c for some $c > 1$. Another option is cross validation, which is widely used in practice.

Instead of the asymptotic critical value given by Theorem 1, we can also employ the following bootstrap procedure.

Bootstrap confidence interval for the RDD causal effect:

(1) The bootstrap resamples $\{y_i^*\}_{i=1}^n$ are generated from

$$y_i^* = w_i\bar{\theta} + f_i'\hat{\gamma} + \hat{\epsilon}_i\eta_i^*, \tag{2.6}$$

where $\hat{\epsilon}_i$ is the residual from the lasso regression in (2.3) and η_i^* is independent of the data and satisfies $E[\eta_i^*] = 0$ and $E[\eta_i^{*2}] = 1$.

(2) Follow Steps 1-4 to compute (2.4) using the bootstrap resample $\{y_i^*, x_i\}_{i=1}^n$. We obtain $\bar{\theta}^* = \left(\sum_{i=1}^n \tilde{\zeta}_i^{*2}\right)^{-1} \sum_{i=1}^n \tilde{\zeta}_i^*(y_i^* - f_i'\hat{\gamma}^*)$ and compute the residuals $\overline{e}_i^* = (y_i^* - f_i'\hat{\gamma}^*) - w_i\bar{\theta}^*$.

(3) Compute the bootstrap statistic

$$t^* = \left(\sum_{i=1}^n \tilde{\zeta}_i^{*2}\right)\left(\sum_{i=1}^n \tilde{\zeta}_i^{*2}\overline{e}_i^{*2}\right)^{-1/2}(\bar{\theta}^* - \bar{\theta}). \tag{2.7}$$

(4) Repeat Steps 1-2 many times and construct the $100(1-a)\%$ bootstrap confidence interval as

$$\left[\bar{\theta} - z_{1-a/2}^*\left(\sum_{i=1}^n \tilde{\zeta}_i^2\right)^{-1}\left(\sum_{i=1}^n \tilde{\zeta}_i^2\overline{e}_i^2\right)^{1/2}, \bar{\theta} - z_{a/2}^*\left(\sum_{i=1}^n \tilde{\zeta}_i^2\right)^{-1}\left(\sum_{i=1}^n \tilde{\zeta}_i^2\overline{e}_i^2\right)^{1/2}\right],$$

where z_a^* is the a-th empirical quantile of t^*.

Asymptotic validity of this procedure is established as follows.

Theorem 2. *Under the regularity conditions given in Appendix,*

$$\left(\sum_{i=1}^n \tilde{\zeta}_i^{*2}\right)\left(\sum_{i=1}^n \tilde{\zeta}_i^{*2}\overline{e}_i^{*2}\right)^{-1/2}(\bar{\theta}^* - \bar{\theta}) \xrightarrow{d} N(0,1) \qquad \text{in } P.$$

Chatterjee and Lahiri (2011) demonstrated that a naive application of the nonparametric bootstrap (i.e., resampling from the original sample $\{Y_i, X_i\}_{i=1}^n$ with equal weights) is invalid to replicate the asymptotic distribution of the lasso estimator when the dimension of the regression function is fixed. On the other hand, our bootstrap procedure employs the wild bootstrap with the bias corrected estimator $\bar{\theta}$ to ensure its validity.

3. Discussion

3.1. **Fuzzy RDD.** Although the discussion so far focuses on the sharp RDD analysis, it is possible to extend our approach to the fuzzy RDD analysis, where the forcing variable X_i is not informative enough to determine the treatment W_i but can affect on the treatment probability. In particular, the fuzzy RDD assumes that the conditional treatment probability $\Pr\{W_i = 1|X_i = x\}$ jumps at $x = c$. To define a reasonable parameter of interest for the fuzzy case, let $W_i(x)$ be a potential treatment for unit i when the cutoff level for the treatment was set at x, and assume that $W_i(x)$ is non-increasing in x at $x = c$. Using the terminology of Angrist, Imbens and Rubin (1996), unit i is called a complier if her cutoff level is X_i (i.e., $\lim_{x\downarrow X_i} W_i(x) = 0$ and $\lim_{x\uparrow X_i} W_i(x) = 1$). A parameter of interest in the fuzzy RDD, suggested by Hahn, Todd and van der Klaauw (2001), is the average causal effect for compliers at $x = c$,

$$\theta_f = E[Y_i(1) - Y_i(0)|i \text{ is complier, } X_i = c].$$

Hahn, Todd and van der Klaauw (2001) showed that under mild conditions the parameter θ_f can be identified by the ratio of the jump in the conditional mean of Y_i at $x = c$ to the jump in the conditional treatment probability at $X_i = c$, i.e.,

$$\theta_f = \frac{\lim_{x\downarrow c} E[Y_i|X_i = x] - \lim_{x\uparrow c} E[Y_i|X_i = x]}{\lim_{x\downarrow c} \Pr\{W_i = 1|X_i = x\} - \lim_{x\uparrow c} \Pr\{W_i = 1|X_i = x\}}. \tag{3.1}$$

In this case, letting $T_i = \mathbb{I}\{X_i \geq c\}$, the numerator of (3.1) can be estimated as in (2.2) by replacing W_i with T_i. Also, the denominator of (3.1) can be estimated as in (2.2) by replacing (Y_i, W_i) with (W_i, T_i). We expect that analogous results to the sharp RDD case can be established.

3.2. **Unknown discontinuity point.** In some applications, the discontinuity point c for the RDD analysis is unknown typically due to privacy or ethical reasons, and needs to be estimated. Such examples include the threshold for scholarship offers and tipping point for dynamics of segregation (Card, Mas and Rothstein, 2008). See also Porter and Yu (2015).

If the discontinuity point c is unknown, we can jointly estimate the slope parameters in (2.2) and c by

$$\min_{\beta,c} \frac{1}{n} \sum_{i=1}^{n} \left\{Y_i - \mathbb{I}\{X_i \geq c\}\theta - \alpha - \mathbb{I}\{X_i \geq c\}M_i'\gamma_R - \mathbb{I}\{X_i < c\}M_i'\gamma_L\right\}^2 + \lambda_n \sum_{j=1}^{\dim \beta} |d_j(c)\beta_j|,$$

where $\beta = (\alpha, \theta, \gamma_R', \gamma_L')'$ and $d_j(c)$ is the empirical L_2-norm of the corresponding regressor. This approach was considered in Lee, Seo and Shin (2016), and they showed that the discontinuity point c can be estimated fast enough to ensure certain oracle property. Thus analogous inference methods on θ_0 as in Section 2.2 can be developed.

3.3. **Regression kink design.** Our high-dimensional method can be extended for the RKDs. For each unit $i = 1, \ldots, n$, we observe continuous outcome and explanatory variables denoted by Y_i and X_i, respectively. The RKD analysis is concerned with the following nonseparable model

$$Y = f(B, X, U),$$

where U is an error term (possibly multivariate) and $B = b(X)$ is a continuous policy variable of interest with known $b(\cdot)$. In general, even though we know the function $b(\cdot)$, we are not able to identify the treatment effect by the policy variable B. However, it is often the case that the policy function $b(\cdot)$ has some kinks (but is continuous). For instance, suppose Y is duration of unemployment and X is earnings before losing the job. We are interested in the effect of unemployment benefits $B = b(X)$. In many unemployment insurance systems (e.g., the one in Austria), $b(\cdot)$ is specified by a piecewise linear function. In such a scenario, one may exploit changes of slopes in the conditional mean $E[Y|X = x]$ to identify a treatment effect of B. Suppose $b(\cdot)$ is kinked at c. In particular, Card, *et al.* (2015) have shown that a treatment on treated parameter $\tau_0 = \int \frac{\partial f(b,x,u)}{\partial b} dF_{U|B=b,X=x}(u)$ is identified as

$$\tau_0 = \frac{\lim_{x \downarrow c} \frac{d}{dx} E[Y|X = x] - \lim_{x \uparrow c} \frac{d}{dx} E[Y|X = x]}{\lim_{x \downarrow c} \frac{d}{dx} b(x) - \lim_{x \uparrow c} \frac{d}{dx} b(x)}. \tag{3.2}$$

To estimate τ_0, we propose the following lasso regression

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left\{ Y_i - \alpha - W_i X_i \theta - X_i \gamma - W_i M_i' \gamma_R - (1 - W_i) M_i' \gamma_L \right\}^2 + \lambda_n |\beta|_1, \tag{3.3}$$

where $\beta = (\alpha, \theta, \gamma, \gamma_R', \gamma_L')'$ is a vector of parameters, $M_i = (m^{(1)}(X_i), \ldots, m^{(p-1)}(X_i))'$ is a vector of basis functions (or dictionaries) evaluated at X_i without the intercept and linear terms, and λ_n is a penalty level. Let $\hat{\theta}$ be the lasso estimator of θ by (3.3). Since the denominator $b_0 = \lim_{x \downarrow c} \frac{d}{dx} b(x) - \lim_{x \uparrow c} \frac{d}{dx} b(x)$ in (3.2) is assumed to be known, the estimator of τ_0 is given by $\hat{\tau} = \hat{\theta}/b_0$. For the properties of the point estimator, similar comments to the previous section apply. To conduct inference on the causal parameter τ_0, it is possible to adapt the inference methods in Section 2 to the RKD setup.

4. Simulation

Here we evaluate finite sample performance of our high-dimensional method for RDDs. We adopt the simulation design of Imbens and Kalyanaraman (2012) and generate

$$Y_i = m(X_i) + U_i,$$

for $i = 1, \ldots, n$, where $X_i = 2Z_i - 1$ with $Z_i \sim Beta(2, 4)$ (so that X_i is supported on $[-1, 1]$) and $U_i \sim N(0, 0.1295^2)$. For the conditional mean function $m(\cdot)$, we consider four cases:

$$
\begin{aligned}
1 \quad &: \quad m_1(x) = \begin{cases} 0.48 + 1.27x + 7.18x^2 + 20.21x^3 + 21.54x^4 + 7.33x^5 & \text{if } x < 0, \\ 0.52 + 0.84x - 3.00x^2 + 7.99x^3 - 9.01x^4 + 3.56x^5 & \text{if } x \geq 0, \end{cases} \\
2 \quad &: \quad m_2(x) = [4\mathbb{I}\{x \geq 0\} + 3\mathbb{I}\{x < 0\}]x^2, \\
3 \quad &: \quad m_3(x) = 0.42 + 0.1\mathbb{I}\{x \geq 0\} + 0.84x - 3.00x^2 + 7.99x^3 - 9.01x^4 + 3.56x^5, \\
4 \quad &: \quad m_4(x) = 0.42 + 0.1\mathbb{I}\{x \geq 0\} + 0.84x + 7.99x^3 - 9.01x^4 + 3.56x^5.
\end{aligned}
$$

All functions have jumps at $x = 0$.

To estimate the causal effect θ_0 in (2.1), we implement the lasso regression in (2.2) using cubic spline basis functions. In particular, we consider two sequences of knots $(-0.9, -0.8, \ldots, -0.1, 0.1, \ldots, 0.9)$ (called Lasso 1) and $(-0.99, -0.98, \ldots, -0.01, 0.01, \ldots, 0.99)$ (called Lasso 2). Note that we do not put a knot at 0, the discontinuity point of interest. The penalty level λ_n is chosen by the cross validation. Finally, for comparison with existing methods, we compute the nonparametric kernel estimator for θ_0 using the bandwidth selection by Imbens and Kalyanaraman (2012) (called IK).

Table 1 reports the bias and root mean squared error (RMSE) of these point estimates based on 1,000 Monte Carlo replications. Overall performance of the proposed Lasso estimator is comparable to that of IK. For example, in terms of the RMSE, Lasso 1 works equally well as IK for Design 1 and outperforms IK for Designs 2 and 3. Lasso 2, which has finer knots, tends to exhibit larger bias than Lasso 1.

We next consider interval estimation of θ_0. In particular, we compare the lasso-based confidence interval in (2.5) (called Lasso-CI) with the asymptotic conventional confidence interval based on the kernel estimator using Imbens and Kalyanaraman's (2012) bandwidth (called Kernel-CI).[1]

Table 2 presents the empirical coverages (EC) and average lengths (AL) of these confidence intervals across 1,000 Monte Carlo replications. The nominal level is 0.95. Overall the proposed lasso-based confidence interval is favorably comparable with the kernel-based confidence interval. For Designs 1 and 2, our method works significantly better than the kernel method in terms of the empirical coverages. For Design 3 and 4, our lass-based method works better for the smaller sample size.

[1]We also tried to implement a robust confidence interval for the RDD causal effect proposed by Calonico, Cattaneo and Titiunik (2014). However, in our simulation, we often encountered numerical problems to compute their standard error (on p. 2320) particularly for the case of $n = 200$.

Appendix A. Mathematical Appendix

A.1. **Assumptions.** We impose the following assumptions.

Assumption.

(1) *There exists a representation of the conditional expectation of y given x satisfying $E[y_i|x_i] = w_i\theta_0 + f_i'\gamma_0 + r(x_i)$, where $S = \{j : \gamma_{0j} \neq 0\}$ $s = |\gamma_0|_0 = o(n)$, $\min\{|\gamma_{0j}| : \gamma_{0j} \neq 0\} \geq 2a_n$, $\sup_x |r(x)| = o(1)$, and $\sum_{i=1}^n r(x_i)^2 = o_p(1)$.*

(2) *Let $\varepsilon_i = y_i - E[y_i|x_i]$ and $\zeta_i = w_i - f_{S,i}'\delta_0$, where δ_0 denotes the projection coefficient of w_i on $f_{S,i}$, Then, there exist c and C such that*

$$0 < c \leq \frac{\sum_{i=1}^n \zeta_i^2}{\sum_{i=1}^n \zeta_i^2|\varepsilon_i|^q}, \frac{\sum_{i=1}^n \zeta_i^2}{\sum_{i=1}^n \zeta_i^2|f_{ij}|^2} \leq C < \infty, \qquad \text{for } q = 1, 2,$$

for all $j \in S$, with probability approaching one.

(3) *$E[\varepsilon_i^2|x_i]$ is bounded above and bounded away from zero, and*

$$\max_j E\left[\exp(|\varepsilon_i f_{ij}|) + \exp(|\zeta_i f_{ij}|) + \exp(f_{ij}^2)\right] \leq C$$

for some $C < \infty$.

(4) *Let $\Sigma_f = E[f_i f_i']$ and $\Sigma = E[z_i z_i']$ with $z_i = (w_i, f_i')'$. There exists a positive constants ϕ such that*

$$(1 + |S|)\frac{b'\Sigma b}{\{|b_1| + \sum_{j\in S}|b_{j+1}|)\}^2} \geq \phi^2,$$

for any $b \in \mathbb{R}^{p+1}$ satisfying $\sum_{j\in S^C}|b_{j+1}| \leq 3\{|b_1| + \sum_{j\in S}|b_{j+1}|\}$.

This a set of regularity conditions on the regression model and moments of the variables. Assumption (1) states that the regression function $E[y_i|x_i]$ can be well-approximated by s terms of basis functions $\{f_i\}$ and the minimal signal strength is bounded below by $2a_n$, which is known as a *betamin* condition. Assumption (2) controls the linear dependence between w_i and f_i. Because both w_i and $f_{S,i}$ are functions of the covariate x_i, we want to take a good care of the choice of f_i to ensure the presence of enough variation in the projection error ζ_i. The other conditions are common in the high-dimensional regression. Assumption (3) is a set of moment conditions, and Assumption (4) is the so-called *compatibility condition* for the design matrix Σ to yield an oracle result for the lasso estimator. See, e.g., Bühlmann and van de Geer (2011, Section 6.13) for more discussions on its relation to the restricted eigenvalue condition by Bickel, Ritov and Tsybakov (2009) and many other related concepts.

A.2. **Proof of Theorem 1.** By the definition of $\bar{\theta}$ and Assumption (1), we can decompose

$$\begin{aligned}\frac{\sum_{i=1}^n \tilde{\zeta}_i w_i}{\sqrt{\sum_{i=1}^n \tilde{\zeta}_i^2 \bar{e}_i^2}}(\bar{\theta} - \theta_0) &= \frac{\sum_{i=1}^n \tilde{\zeta}_i \varepsilon_i}{\sqrt{\sum_{i=1}^n \tilde{\zeta}_i^2 \bar{e}_i^2}} + \frac{\sum_{i=1}^n \tilde{\zeta}_i r(x_i)}{\sqrt{\sum_{i=1}^n \tilde{\zeta}_i^2 \bar{e}_i^2}} - \frac{\sum_{i=1}^n \tilde{\zeta}_i f_i'(\hat{\gamma} - \gamma_0)}{\sqrt{\sum_{i=1}^n \tilde{\zeta}_i^2 \bar{e}_i^2}} \\ &\equiv T_1 + T_2 - T_3.\end{aligned}$$

Thus, it is enough to show that $T_1 \xrightarrow{d} N(0,1)$ and $T_2, T_3 \xrightarrow{p} 0$.

For T_1, Lemma A.1, Assumption (2), and the central limit theorem imply

$$T_1 = \frac{\sum_{i=1}^n \tilde{\zeta}_i \varepsilon_i}{\sqrt{\sum_{i=1}^n \zeta_i^2 \varepsilon_i^2}} + o_p(1) = \frac{\sum_{i=1}^n \zeta_i \varepsilon_i}{\sqrt{\sum_{i=1}^n \zeta_i^2 \varepsilon_i^2}} + o_p(1) \xrightarrow{d} N(0,1).$$

For T_2, the Cauchy-Schwarz inequality and Assumption (1) ($\sum_{i=1}^n r(x_i)^2 = o_p(1)$) imply

$$|T_2| \leq \sqrt{\frac{\sum_{i=1}^n \tilde{\zeta}_i^2}{\sum_{i=1}^n \tilde{\zeta}_i^2 \bar{e}_i^2}} o_p(1) = \sqrt{\frac{\sum_{i=1}^n \zeta_i^2}{\sum_{i=1}^n \zeta_i^2 \varepsilon_i^2}} o_p(1) \xrightarrow{p} 0,$$

where the equality follows from Lemma A.1 and the convergence follows from Assumption (2).

For T_3, note that $\tilde{\zeta}_i$ is orthogonal to the elements of f_i that belong to S_n and thus

$$|T_3| = \frac{\left|\sum_{i=1}^n \tilde{\zeta}_i f'_{S_n^C,i} \gamma_{0,S_n^C}\right|}{\sqrt{\sum_{i=1}^n \tilde{\zeta}_i^2 \bar{e}_i^2}} \xrightarrow{p} 0,$$

since $P\{S_n = S\} \to 1$ due to Lemma A.2.

Lemma A.1. *Under the assumptions above,*

$$\frac{\sum_{i=1}^n \tilde{\zeta}_i^2 \bar{e}_i^2}{\sum_{i=1}^n \zeta_i^2 \varepsilon_i^2} \xrightarrow{p} 1, \qquad \frac{\sum_{i=1}^n \tilde{\zeta}_i^2}{\sum_{i=1}^n \zeta_i^2} \xrightarrow{p} 1, \quad \textit{and} \quad \sum_{i=1}^n \tilde{\zeta}_i \varepsilon_i = \sum_{i=1}^n \zeta_i \varepsilon_i + o_p\left(\left|\sum_{i=1}^n \zeta_i^2\right|^{1/2}\right).$$

Proof. Since the proof is similar for the first two, we focus on the first statement. By Assumption (1) and the fact that $P\{S_n = S\} \to 1$, we can write without loss of generality that $\tilde{\zeta}_i = \zeta_i - f'_{S,i}(\tilde{\delta} - \delta_0)$ and $\bar{e}_i = \varepsilon_i + r(x_i) - w_i(\bar{\theta} - \theta_0) - f'_i(\hat{\gamma} - \gamma_0)$. The conclusion follows by plugging in these expressions to $\sum_{i=1}^n \tilde{\zeta}_i^2 \bar{e}_i^2$ and showing all terms except $\sum_{i=1}^n \zeta_i^2 \varepsilon_i^2$ are negligible. First, by Assumption (1)-(2),

$$\frac{\sum_{i=1}^n \zeta_i^2 r(x_i)^2}{\sum_{i=1}^n \zeta_i^2 \varepsilon_i^2} \leq \{\sup_x |r(x)|\}^2 \frac{\sum_{i=1}^n \zeta_i^2}{\sum_{i=1}^n \zeta_i^2 \varepsilon_i^2} \xrightarrow{p} 0.$$

Second, by $|w_i| \leq 1$ and Assumption (2),

$$\frac{\sum_{i=1}^n \zeta_i^2 w_i^2}{\sum_{i=1}^n \zeta_i^2 \varepsilon_i^2} (\bar{\theta} - \theta_0)^2 \leq C(\bar{\theta} - \theta_0)^2 \xrightarrow{p} 0.$$

Third, by Assumption (2),

$$\frac{\sum_{i=1}^n \zeta_i^2 \{f'_i(\hat{\gamma} - \gamma_0)\}^2}{\sum_{i=1}^n \zeta_i^2 \varepsilon_i^2} \leq C \left(\max_{1 \leq i \leq n} |f'_i(\hat{\gamma} - \gamma_0)|\right)^2 \xrightarrow{p} 0,$$

whose convergence follows from

$$\max_{1 \leq i \leq n} |f'_i(\hat{\gamma} - \gamma_0)| \leq \left(\max_{1 \leq i \leq n} \max_{1 \leq j \leq 2p} |f_{ij}|\right) (|\tilde{\gamma} - \gamma_0|_1 + |\tilde{\gamma} - \hat{\gamma}|_1) = O_p(\lambda_n s \log(np)) = o_p(1),$$

where the first inequality is due to the Hölder inequality and the triangle inequality and the subsequent equality follows from the deviation bound for $\tilde{\gamma}$ in Lemma A.2 (and its implication on the perfect selection on S), $|\tilde{\gamma} - \hat{\gamma}|_1 = \sum_j |\tilde{\gamma}_j| \mathbb{I}\{|\tilde{\gamma}_j| \leq a_n\} \leq O_p(s\lambda_n \log n)$, and Bühlmann

and van de Geer (2011, Lemma 14.12). Finally, by Assumption (2),

$$\frac{\sum_{i=1}^n \varepsilon_i^2 \{f'_{S,i}(\tilde{\delta} - \delta_0)\}^2}{\sum_{i=1}^n \zeta_i^2 \varepsilon_i^2} \leq \frac{\sum_{i=1}^n \varepsilon_i^2}{\sum_{i=1}^n \zeta_i^2 \varepsilon_i^2} \left(\max_{1 \leq i \leq n} |f'_{S,i}(\tilde{\delta} - \delta_0)| \right)^2 \xrightarrow{p} 0.$$

The convergence follows from Assumption (2), specifically, by the condition that $\sum_{i=1}^n \zeta_i^2 \varepsilon_i^2$ is the same order of magnitude as $\sum_{i=1}^n \zeta_i^2$, and from

$$\max_{1 \leq i \leq n} |f'_{S,i}(\tilde{\delta} - \delta_0)| \leq \left(\max_{1 \leq i \leq n} \max_{j \in S} |f_{ij}| \right) |\tilde{\delta} - \delta_0|_1 = O_p(n^{-1/2} s \log^{1/2} s \log(ns)),$$

for which we recall that $\tilde{\delta}$ is the OLS estimator and thus its l_1 deviation bound is $n^{-1/2} s \log^{1/2} s$. By similar arguments, we can show that other terms are also negligible.

Turning to the last statement, write $\sum_{i=1}^n \tilde{\zeta}_i \varepsilon_i = \sum_{i=1}^n \zeta_i \varepsilon_i - \sum_{i=1}^n \varepsilon_i f'_{S,i}(\tilde{\delta} - \delta_0)$ and note that

$$\begin{aligned} \frac{\left| \sum_{i=1}^n \varepsilon_i f'_{S,i}(\tilde{\delta} - \delta_0) \right|}{\left| \sum_{i=1}^n \zeta_i^2 \right|^{1/2}} &\leq \left| \sum_{i=1}^n \varepsilon_i f'_{S,i} (\sum_{i=1}^n f_{S,i} f'_{S,i})^{-1} \right|_1 C \sup_{j \in S} \frac{|\sum_{i=1}^n \zeta_i f_{ij}|}{\left| \sum_{i=1}^n \zeta_i^2 f_{ij}^2 \right|^{1/2}} \\ &\leq O_p(s/\sqrt{n}) O_p(\sqrt{\log s}) = o_p(1), \end{aligned}$$

where the Hölder inequality and Assumption (2), and then the Bernstein inequality are applied for the first and second inequalities, respectively.

Lemma A.2. *Under the assumptions above, it holds* $|(\tilde{\theta}, \tilde{\gamma}') - (\theta_0, \gamma_0')|_1 = O_p(\lambda_n s)$. *Furthermore,* $\Pr\{S_n = S\} \to 1$.

Proof. The regression model estimated in (2.3) can be written as $Y = g + \varepsilon$, where $Y = (y_1, \ldots, y_n)'$, $g = (E[y_1|x_1], \ldots, E[y_n|x_n])'$, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$. By Assumption (1), g is written as $g = Z\beta_0 + R$, where $Z = \left[\begin{pmatrix} w_1 \\ f_1 \end{pmatrix}, \ldots, \begin{pmatrix} w_n \\ f_n \end{pmatrix} \right]'$, $\beta_0 = (\theta_0, \gamma_0')'$, and $R = (r(x_1), \ldots, r(x_n))'$. We assume that the columns in Z are scale normalized (i.e., $\sum_{i=1}^n (z_i^{(j)})^2 = n$ for each j). Bühlmann and van de Geer (2011, Lemma 14.12) implies that if $\sqrt{n^{-1/2} \log(\max\{n, p\})} = o(\lambda_n)$, then

$$\Pr\{\mathcal{A}_n\} \equiv \Pr\left\{ \frac{8}{n} \left| \sum_{i=1}^n Z_i \varepsilon_i \right|_\infty \leq \lambda_n \right\} \to 1, \tag{A.1}$$

as $n \to \infty$. Let $\tilde{\beta} = (\tilde{\theta}, \tilde{\gamma}')'$.

We now prove the following statements:

(a): Conditionally on $\mathcal{A}_n$, it holds

$$\frac{4}{n} |Z\tilde{\beta} - g|_2^2 + 3\lambda_n |\tilde{\beta}_{S^c}|_1 \leq \frac{4}{n} |R|_2^2 + 5\lambda_n |\tilde{\beta}_{S^c} - \beta_0|_1. \tag{A.2}$$

Also, additionally, if $\hat{\Sigma} = n^{-1} Z'Z$ is compatible for S with a compatibility constant $\phi > 0$, then

$$\frac{2}{n} |Z\tilde{\beta} - g|_2^2 + \lambda_n |\tilde{\beta} - \beta_0|_1 \leq \frac{6}{n} |R|_2^2 + 24\lambda_n^2 s_\lambda \phi^{-2}. \tag{A.3}$$

(b): $\hat{\Sigma}$ is compatible for S with a compatibility parameter $\phi \geq \phi_\Sigma / \sqrt{2}$, where ϕ_Σ is the compatibility parameter of $\Sigma = E[z_i z_i']$ for S.

First, we prove (a). Since $\tilde{\beta}$ is a minimizer, we have

$$\frac{1}{n}|Y - Z\tilde{\beta}|_2^2 + \lambda_n|\tilde{\beta}|_1 \leq \frac{1}{n}|Y - Z\beta_0|_2^2 + \lambda_n|\beta_0|_1.$$

By plugging $Y = Z\beta_0 + R + \varepsilon$ into the above,

$$n^{-1}|Z\tilde{\beta} - g|_2^2 + \lambda_n|\tilde{\beta}|_1 \leq n^{-1}|R|_2^2 + 2n^{-1}\varepsilon' Z(\tilde{\beta} - \beta_0) + \lambda_n|\beta_0|.$$

By the Hölder inequality, $|n^{-1}\varepsilon' Z(\tilde{\beta} - \beta_0)| \leq |n^{-1}\varepsilon' Z|_\infty|\tilde{\beta} - \beta_0|_1$. Thus, conditionally on $\mathcal{A}_n$, we can proceed as in Bühlmann and van de Geer (2011, Lemma 6.3) and the statement in (A.2) follows. Also, the statement in (A.3) is given by Bühlmann and van de Geer (2011, Theorem 6.2).

Next, we show (b). Bühlmann and van de Geer (2011, Lemma 14.12) guarantees

$$|\hat{\Sigma} - \Sigma|_\infty = O_p(n^{-1/2}\log p) \leq \frac{\phi_\Sigma^2}{32|S|},$$

with probability approaching one. Therefore, the statement (b) follows by Bühlmann and van de Geer (2011, Corollary 6.8).

Note that Assumption (1) guarantees $\frac{6}{n}|R|_2^2 = o_p(\lambda_n^2 s_\lambda \phi^{-2})$. Therefore combining (A.1) and (A.3) based on statement (b) implies the first conclusion of the lemma.

Finally, to prove that $\Pr\{S_n = S\} \to 1$, we note that the deviation bound we obtained above is asymptotically negligible to the threshold a_n, implying that if $|\gamma_{0j}| = 0$, $\hat{\gamma}_j = 0$ as well with probability approaching one since $|\tilde{\gamma}_j|$ cannot exceed the threshold a_n and that if $|\gamma_{0j}| > 2a_n$ then $|\tilde{\gamma}_j|$ must exceed the threshold a_n.

A.3. **Proof of Theorem 2.** In this proof, the superscript * indicates the bootstrap counterpart of the original statistic.

To show the validity of the above bootstrap, we begin with deriving the deviation bounds for the bootstrap lasso estimate

$$(\tilde{\theta}^*, \tilde{\gamma}^{*\prime}) = \arg\min_\beta \frac{1}{n}\sum_{i=1}^n (y_i^* - w_i\theta - f_i'\gamma)^2 + \lambda_n^*|\gamma|_1, \tag{A.4}$$

Following the same line of arguments in Lemma A.2, however, we obtain the deviation bounds for $|\tilde{\theta}^* - \bar{\theta}| + |\tilde{\gamma}^* - \hat{\gamma}_1|$ at the same rate for $\tilde{\theta}$ and $\tilde{\gamma}$ and the selection consistency that $P\{S_n^* = S\}$ converges to one, where S_n^* denotes the indexes of selected variables from a bootstrap sample by the thresholded lasso. In particular, all the other steps are identical other than an analogous bound to (A.1) for the bootstrap sample, that is

$$\begin{aligned}
& P\left\{\frac{8}{n}\left|\sum_{i=1}^n Z_i\bar{e}_i\eta_i^*\right|_\infty > \lambda_n\right\} \\
\leq\ & P\left\{\max_j\left|\frac{8}{n}\sum_{i=1}^n f_{ij}\varepsilon_i\eta_i^*\right| > \frac{\lambda_n}{4}\right\} + P\left\{\max_j\left|\frac{8}{n}\sum_{i=1}^n f_{ij}\eta_i^* w_i(\bar{\theta} - \theta_0)\right| > \frac{\lambda_n}{4}\right\} \\
& + P\left\{\max_j\left|\frac{8}{n}\sum_{i=1}^n f_{ij}\eta_i^* r_i\right| > \frac{\lambda_n}{4}\right\} + P\left\{\max_j\left|\frac{8}{n}\sum_{i=1}^n f_{ij}\eta_i^* f_i'(\hat{\gamma} - \gamma_0)\right| > \frac{\lambda_n}{4}\right\} \\
\to\ & 0
\end{aligned}$$

for which we apply in sequel the union bound and the decomposition of $\bar{e}_i$ for the inequality and Bernstein inequality for the terms after the inequality in conjunction with the deviation bounds. Specifically, we illustrate the last term

$$\begin{aligned}\max_j \left| \frac{8}{n} \sum_{i=1}^n f_{ij} \eta_i^* f_i'(\hat{\gamma} - \gamma_0) \right| &\leq \max_j \max_t \left| \frac{8}{n} \sum_{i=1}^n f_{ij} \eta_i^* f_{it} \right| |\hat{\gamma} - \gamma_0|_1 \\ &= O_p\left(\frac{\log 2p}{\sqrt{n}}\right) O_p(s\lambda_n).\end{aligned}$$

Second, the preceding selection consistency implies that $\tilde{\zeta}_i^* = \tilde{\zeta}_i$ with probability approaching one and thus we can write without loss of generality that

$$\bar{\theta}^* = \frac{\sum_{i=1}^n \tilde{\zeta}_i (y_i^* - f_i' \hat{\gamma}^*)}{\sum_{i=1}^n \tilde{\zeta}_i w_i} = \bar{\theta} - \frac{\sum_{i=1}^n \tilde{\zeta}_i f_i' (\hat{\gamma}^* - \hat{\gamma})}{\sum_{i=1}^n \tilde{\zeta}_i w_i} + \frac{\sum_{i=1}^n \tilde{\zeta}_i \bar{e}_i \eta_i}{\sum_{i=1}^n \tilde{\zeta}_i w_i}.$$

For the same reasoning as Lemma A.1, we conclude that

$$\frac{\sum_{i=1}^n \tilde{\zeta}_i^{*2} \bar{e}_i^{*2}}{\sum_{i=1}^n \zeta_i^2 \varepsilon_i^2 \eta_i^2} \xrightarrow{d} 1, \quad \text{and} \quad \sum_{i=1}^n \tilde{\zeta}_i \bar{e}_i \eta_i = \sum_{i=1}^n \zeta_i \varepsilon_i \eta_i + o_p\left(\left(\sum_{i=1}^n \zeta_i^2 \right)^{1/2} \right).$$

Furthermore, note that $\sum_{i=1}^n \tilde{\zeta}_i^{*2} = \sum_{i=1}^n \tilde{\zeta}_i^* w_i$ as $\tilde{\zeta}_i^*$ is the OLS residual and proceed as in the proof of Theorem 1 to conclude

$$\begin{aligned}\frac{\sum_{i=1}^n \tilde{\zeta}_i^{*2}}{\sqrt{\sum_{i=1}^n \tilde{\zeta}_i^{*2} \bar{e}_i^{*2}}} (\bar{\theta}^* - \bar{\theta}) &= \frac{\sum_{i=1}^n \zeta_i \varepsilon_i \eta_i}{\sqrt{\sum_{i=1}^n \zeta_i^2 \varepsilon_i^2 \eta_i^2}} + o_p(1) \\ &\xrightarrow{d} N(0,1) \qquad \text{in } P,\end{aligned}$$

by the conditional multiplier central limit theorem, see e.g. p.176 in van der Vaart and Wellner (1996).

Appendix B. Tables

Table 1. Bias and RMSE of point estimators for RDD

		$n = 200$		$n = 500$	
Design	Method	Bias	RMSE	Bias	RMSE
1	IK	0.040	0.072	0.038	0.054
	Lasso 1	0.008	0.073	0.006	0.050
	Lasso 2	-0.010	0.072	-0.029	0.056
2	IK	0.010	0.062	0.004	0.039
	Lasso 1	-0.001	0.019	0.000	0.010
	Lasso 2	-0.018	0.049	-0.010	0.020
3	IK	-0.022	0.093	-0.014	0.058
	Lasso 1	-0.019	0.065	-0.012	0.051
	Lasso 2	-0.100	0.100	-0.100	0.100
4	IK	-0.020	0.087	-0.011	0.056
	Lasso 1	-0.090	0.095	-0.097	0.098
	Lasso 2	-0.100	0.100	-0.117	0.118

Table 2. Empirical Coverage (EC) and Average Length (AL) for RDD

		$n = 200$		$n = 500$	
Design	Method	EC	AL	EC	AL
1	Kernel-CI	0.770	0.178	0.663	0.113
	Lasso-CI	0.824	0.369	0.880	0.252
2	Kernel-CI	0.862	0.187	0.885	0.120
	Lasso-CI	0.957	0.249	0.962	0.124
3	Kernel-CI	0.824	0.248	0.841	0.163
	Lasso-CI	0.835	0.329	0.778	0.214
4	Kernel-CI	0.819	0.248	0.840	0.164
	Lasso-CI	0.822	0.505	0.776	0.311

References

[1] Angrist, J. D., Imbens, G. W. and D. B. Rubin (1996) Identification of causal effects using instrumental variables, *Journal of the American Statistical Association*, 91, 444-455.

[2] Arai, Y. and H. Ichimura (2018) Simultaneous selection of optimal bandwidths for the sharp regression discontinuity estimator, *Quantitative Economics*, 9, 441-482.

[3] Belloni, A., Chernozhukov, V. and C. Hansen (2014) Inference on treatment effects after selection amongst high-dimensional controls, *Review of Economic Studies*, 81, 608-650.

[4] Bickel, P. J., Ritov, Y. A. and A. B. Tsybakov (2009) Simultaneous analysis of Lasso and Dantzig selector, *Annals of Statistics*, 37, 1705-1732.

[5] Bühlmann, P. and S. van de Geer (2011) *Statistics for High-Dimensional Data*, Springer.

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

[7] Calonico, S., Cattaneo, M. D. and R. Titiunik (2015) Optimal data-driven regression discontinuity plots, *Journal of the American Statistical Association*, 110, 1753-1769.

[8] Card, D., Lee, D. S., Pei, Z. and A. Weber (2015) Inference on causal effects in a generalized regression kink design, *Econometrica*, 83, 2453-2483.

[9] Card, D., Mas, A. and J. Rothstein (2008) Tipping and the dynamics of segregation, *Quarterly Journal Economics*, 123, 177–218.

[10] Chatterjee, A. and S. N. Lahiri (2011) Bootstrapping lasso estimators, *Journal of the American Statistical Association*, 106-608-625.

[11] Ganong, P. and S. Jäger (2018) A permutation test for the regression kink design, *Journal of American Statistical Association*, 113, 494-504.

[12] Gelman, A. and G. W. Imbens (2018) Why high-order polynomials should not be used in regression discontinuity designs, forthcoming in *Journal of Business & Economic Statistics*.

[13] Hahn, J., Todd, P. and W. van der Klaauw (2001) Identification and estimation of treatment effects with a regression-discontinuity design, *Econometrica*, 69, 201-209.

[14] Imbens, G. W. and K. Kalyanaraman (2012) Optimal bandwidth choice for the regression discontinuity estimator, *Review of Economic Studies*, 79, 933-959.

[15] Imbens, G. W. and T. Lemieux (2008) Regression discontinuity designs: a guide to practice, *Journal of Econometrics*, 142, 615-635.

[16] Lee, S., Seo, M. H. and Y. Shin (2016) The lasso for high dimensional regression with a possible change point, *Journal of the Royal Statistical Society*, B, 78, 193-210.

[17] Porter, J. and P. Yu (2015) Regression discontinuity designs with unknown discontinuity points: Testing and estimation, *Journal of Econometrics*, 189, 132-147.

[18] van de Geer, S., Bühlmann, P., Ritov, Y. and R. Dezeure (2014) On asymptotically optimal confidence regions and tests for high-dimensional models, *Annals of Statistics*, 42, 1166-1202.

[19] Zhang, C.-H. and S. S. Zhang (2014) Confidence intervals for low dimensional parameters in high dimensional linear models, *Journal of the Royal Statistical Society*, B, 76, 217-242.

School of Social Sciences, Waseda University, 1-6-1 Nishiwaseda, Shinjuku-ku, Tokyo 169-8050, Japan.

E-mail address: yarai@waseda.jp

Department of Economics, London School of Economics, Houghton Street, London, WC2A 2AE, UK.

E-mail address: t.otsu@lse.ac.uk

Department of Economics, Seoul National University, 1 Gwankro Gwanakgu, Seoul, 08826, Korea.

E-mail address: myunghseo@snu.ac.kr