REGRESSION DISCONTINUITY DESIGN WITH POTENTIALLY MANY COVARIATES

YOICHI ARAI, TAISUKE OTSU, AND MYUNG HWAN SEO

Abstract. This paper studies the case of possibly high-dimensional covariates in the regression discontinuity design (RDD) analysis. In particular, we propose estimation and inference methods for the RDD models with covariate selection which perform stably regardless of the number of covariates. The proposed methods combine the local approach using kernel weights with ℓ_1-penalization to handle high-dimensional covariates, and the combination is new in the literature. We provide theoretical and numerical results which illustrate the usefulness of the proposed methods. Theoretically, we present risk and coverage properties for our point estimation and inference methods, respectively. Numerically, our simulation experiments and empirical example show the robust behaviors of the proposed methods to the number of covariates in terms of bias and variance for point estimation and coverage probability and interval length for inference.

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

Key words and phrases. Regression discontinuity design; Covariate Selection; Many covariates; Local Lasso;
This paper is a developed version of the previous manuscript (https://sticerd.lse.ac.uk/dps/em/em601.pdf) inspired by the discussion with Matias Cattaneo. We are grateful to Matias Cattaneo and Sebastian Calonico for helpful comments and discussions. This research was supported by Grants-in-Aid for Scientific Research 20K01598 from the Japan Society for the Promotion of Science (Arai) and the ERC Consolidator Grant (SNP 615882) (Otsu). Financial support from the Center for Distributive Justice in the Institute of Economic Research of Seoul National University and the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2018S1A5A2A01033487) is gratefully acknowledged (Seo).

causal parameters of interest are identified by some contrasts of the left and right limits of the conditional mean functions. See e.g. Imbens and Lemieux (2008), Cattaneo, Titiunik and Vazquez-Bare (2020), an edited volume Cattaneo and Escanciano (2017), and references therein.

In the growing literature on the RDD analysis, this paper focuses on the RDDs where covariates are included in the estimation, which is extensively studied by Calonico, Cattaneo, Farrell and Titiunik (2019) (hereafter, CCFT). See also Frölich and Huber (2019) for an alternative estimation method based on kernel smoothing after localization around the cutoff. In practice, researchers often augment the regression models for the RDD analysis with various additional predetermined covariates such as demographic or socioeconomic characteristics for data units. For several RDD estimators using covariates based on local polynomial regression methods, CCFT investigated the MSE expansion, asymptotic efficiency, and data-driven bandwidth selection methods. Furthermore, CCFT developed asymptotic distributional approximations for those estimators and proposed valid inference procedures by constructing bias and variance estimators with covariate adjustment. These results may be considered as extensions of the analyses in Calonico, Cattaneo and Titiunik (2014) (hereafter, CCT) combined with robust bias correction methods in Calonico, Cattaneo and Farrell (2018, 2020) to incorporate covariates in the RDD analysis. See also Calonico, Cattaneo, Farrell and Titiunik (2017) for the statistical package on these methods.

This paper studies the case of possibly high-dimensional covariates in the RDD analysis. In empirical research, it is common to include covariates and their interaction terms, and the number of covariates can be pretty large. To accommodate many covariates, we propose estimation and inference methods for the RDD models with covariate selection by high-dimensional methods.

For point estimation on the causal effect parameter identified by the RDD, we consider the Lasso estimator and its post-selection estimator based on the local linear regression (i.e., eq. (2) of CCFT). To the best of our knowledge, such a combination of the localization using kernel weights and ℓ_1-penalization to deal with high-dimensional covariates is novel in the literature. Indeed this combination is particularly relevant for the RDD analysis, where the effective sample size would be typically small due to the localization so that the effect of dimensionality of covariates becomes severer. Theoretically, we derive the ℓ_1-risk properties of our "local Lasso" estimators and its post-selection version. Practically, based on our simulation study, we recommend the CCFT estimator after selecting covariates by the ℓ_1-penalization even for a relatively small number of covariates, which exhibits desirable MSE properties and stability across different setups.

For inference, we propose to select covariates with sufficiently large coefficient estimates. We show that the inference based on the selected covariates can be implemented in the same manner as in CCFT. Our simulation results demonstrate that the post-selection confidence interval exhibits robust performances in terms of both coverages and lengths, even for a relatively small number of covariates.

This paper also contributes to the large literature on high-dimensional methods in econometrics and statistics (see, e.g., Bühlmann and van de Geer, 2011, and Belloni *et al.*, 2018, for an overview) by combining the kernel localization with ℓ_1-penalization to handle high-dimensional covariates. 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 concerns the inference on a jump in a nonparametric regression model.

This paper is organized as follows. Section 2.1 introduces our basic setup and local Lasso estimator, and presents the ℓ_1-risk properties. In Section 2.2, we discuss the validity of CCFT's inference after selecting covariates by our Lasso procedure. To illustrate the proposed method, Section 4 conducts a simulation study, and Section 5 presents an empirical example based on the Head Start data.

2. Main result

2.1. **Setup and local Lasso estimator for covariate selection.** In this subsection, we present our basic setup and introduce the local Lasso estimator for the RDD with possibly high-dimensional covariates. For each unit $i = 1, \ldots, n$, we observe an indicator variable T_i for a treatment ($T_i = 1$ if treated and $T_i = 0$ otherwise), and outcome $Y_i = Y_i(0) \cdot (1 - T_i) + Y_i(1) \cdot T_i$, where $Y_i(0)$ and $Y_i(1)$ are potential outcomes for $T_i = 0$ and $T_i = 1$, 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 distributional aspects of the difference of potential outcomes $Y_i(1) - Y_i(0)$. The RDD analysis focuses on the case where the treatment assignment T_i is completely or partly determined by some observable covariate X_i, called the running 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, T_i is an indicator variable for the class size ($T_i = 0$ for one class and $T_i = 1$ for two classes), and X_i is the number of enrollments. For more examples, see e.g. Imbens and Lemieux (2008), Cattaneo, Titiunik and Vazquez-Bare (2020), Cattaneo and Escanciano (2017), and references therein.

Depending on the assignment rule for T_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. In the sharp RDD, the treatment is deterministically assigned as $T_i = \mathbb{I}\{X_i \geq \bar{x}\}$, where $\mathbb{I}\{\cdot\}$ is the indicator function and $\bar{x}$ is a known cutoff point. Throughout the paper, we normalize $\bar{x} = 0$ to simplify the presentation. A parameter of interest, in this case, is the average causal effect at the discontinuity point,

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = 0]. \tag{1}$$

Since the difference of potential outcomes $Y_i(1) - Y_i(0)$ is unobservable, we need a tractable representation of τ in terms of quantities that can be estimated by data. If the conditional mean functions $\mathbb{E}[Y_i(1)|X_i = x]$ and $\mathbb{E}[Y_i(0)|X_i = x]$ are continuous at the cutoff point $x = 0$, then the average causal effect τ can be identified as a contrast of the left and right limits of the conditional mean $\mathbb{E}[Y_i|X_i = x]$ at $x = 0$,

$$\tau = \lim_{x \downarrow 0} \mathbb{E}[Y_i|X_i = x] - \lim_{x \uparrow 0} \mathbb{E}[Y_i|X_i = x]. \tag{2}$$

As argued in CCFT, it is usually the case that practitioners have access to additional covariates (denoted by $Z_i \in \mathbb{R}^p$) and augment their empirical models with Z_i to estimate the causal effect τ of interest. This practically relevant setup is extensively studied in CCFT for the case where Z_i is low-dimensional. In this paper, we consider the case of possibly high-dimensional Z_i, and propose a new point estimation method for τ and an adjustment of CCFT's inference method.

We examine the case where the additional covariates Z_i are predetermined (in the sense that $Z_i = Z_i(0) \cdot (1 - T_i) + Z_i(1) \cdot T_i$ but $Z_i(0) =_d Z_i(1)$ for the potential covariates $Z_i(0)$ and $Z_i(1)$ for $T_i = 0$ and $T_i = 1$, respectively). Motivated by CCFT's recommended model (in their eq. (2)), we propose the local Lasso estimator $\hat{\theta} = (\hat{\alpha}, \hat{\tau}, \hat{\beta}_-, \hat{\beta}_+, \hat{\gamma}')'$ that solves

$$\min_{\theta} \frac{1}{nh} \sum_{i=1}^{n} K_i \left\{ Y_i - \alpha - T_i\tau - X_i\beta_- - T_iX_i\beta_+ - Z_i'\gamma \right\}^2 + \lambda_n|\theta|_1, \tag{3}$$

where $K_i = K(X_i/h)$ is the kernel weight to localize around the cutoff point $x = 0$ with a bandwidth denoted by h, $\theta = (\alpha, \tau, \beta_-, \beta_+, \gamma')'$ is a $(p+4)$-dimensional vector of parameters, $|\theta|_1 = \sum_{j=1}^{p+4} |\theta^{(j)}|$ is the ℓ_1-norm of the parameter vector ($\theta^{(j)}$ means the j-th element of θ), and λ_n is a penalty level. Popular choices for $K(\cdot)$ are the uniform and triangular kernels supported on $[-h, h]$. Then our point estimator for τ is given by $\hat{\tau}$.

It is often the case that researchers do not want to penalize some subset of parameters (particularly τ and perhaps (β_-, β_+)), denoted by θ_1. In this case, we consider the partially penalized estimator $\tilde{\theta} = (\tilde{\alpha}, \tilde{\tau}, \tilde{\beta}_-, \tilde{\beta}_+, \tilde{\gamma}')'$ that solves

$$\min_{\theta} \frac{1}{nh} \sum_{i=1}^{n} K_i \left\{ Y_i - \alpha - T_i \tau - X_i \beta_- - T_i X_i \beta_+ - Z_i' \gamma \right\}^2 + \lambda_n |\theta_2|_1, \tag{4}$$

where $\theta = (\theta_1', \theta_2')'$.

Our preliminary simulation results suggest that the local Lasso estimators for τ are somewhat biased in finite samples. Therefore, our recommendation for point estimation is to employ a post-selection method. Let $\bar{S} = \{j : |\hat{\theta}^{(j)}| > 0\}$ or $\{j : |\tilde{\theta}^{(j)}| > 0\}$ be the indices for selected covariates based on the local Lasso estimation in (3) or (4), respectively, $Z_{\bar{S}}$ be the vector of selected covariates, and $G_{\bar{S}} = (1, T, X, TX, Z_{\bar{S}}')'$. Then the local post-Lasso estimator $\bar{\theta}_{\bar{S}}$ is obtained by the local least square:

$$\min_{\theta_{\bar{S}}} \frac{1}{nh} \sum_{i=1}^{n} K_i \left\{ Y_i - G_{\bar{S},i}' \theta_{\bar{S}} \right\}^2, \tag{5}$$

and the estimator $\bar{\tau}$ for τ is given by the estimated coefficient of T.

To the best of our knowledge, such a combination of the localization using kernel weights K_i and ℓ_1-penalization to deal with high-dimensional covariates Z_i is novel in the literature (and also practically relevant in the RDD analysis). Several points are worthy of remark for this estimator. First, without the ℓ_1-penalization, our estimator reduces to the local linear-type estimator recommended by CCFT's eq. (2). Therefore, the proposed estimator is a natural generalization of CCFT's when the dimension of Z_i is high. Second, without the kernel weights K_i for localization, our estimator reduces to the conventional Lasso estimator. However, since our parameter of interest τ is identified as a local object in (2), it is crucial to introduce such localization to avoid misspecification bias of the conditional mean functions. Third, it is often the case that the kernel function $K(\cdot)$ has bounded support. In this case, the effective sample size would be typically of order nh. Thus even if the dimension of Z_i is relatively small compared to the original sample size n, the ℓ_1-penalization would be useful especially for small h.

We now present a risk property of the local Lasso estimators $\hat{\theta}$, $\tilde{\theta}$, and $\bar{\theta}$. Let $G_i = (1, T_i, X_i, T_i X_i, Z_i')'$ be the vector of regressors in (3), G_{ij} be the j-th element of G_i, $\Theta_n = \arg\min_\theta \mathbb{E}[K_i(Y_i - G_i'\theta)^2]$ be an argmin set, and $\hat{M} = \frac{1}{nh}\sum_{i=1}^n K_i G_i G_i'$. We impose the following assumptions.

Assumption 1. (1) *There exists* $\theta^* \in \Theta_n$ *such that* $|\theta^*|_0 \leq s^*$ *for some sequence* $s^* = o(n)$.

(2) *Let* $e_i = K_i^{1/2} Y_i - K_i^{1/2} G_i'\theta^*$. *There exists some* $C \in (0, \infty)$ *such that*

$$\mathbb{E}[|K_i G_{ij} e_i|^m] \leq h \frac{m! C^{m-2}}{2},$$

for all $j = 1, \dots, p$ *and* $m = 2, 3, \dots$.

(3) *Let* δ_j *be the* j*-th element of* δ. *There exists some* $\phi_* \in (0, \infty)$ *for* $S = S^* = \{j : \theta^{*(j)} \neq 0\}$ *such that*

$$\min_{\delta : |\delta_{S^c}|_1 \leq 3|\delta_S|_1} \frac{\delta' \hat{M} \delta}{|\delta_S|_1^2} |S| \geq \phi_*^2,$$

with probability approaching one, where for a given index set A, δ_A denotes the subvector of δ of indexes in A.

(4) $K : \mathbb{R} \to \mathbb{R}$ *is a bounded and symmetric second-order kernel function which is continuous with compact support. The bandwidth h is a positive sequence satisfying $h \to 0$ and $nh \to \infty$ as $n \to \infty$.*

The first condition defines θ^* as the best linear predictor with some sparsity feature. The second is a set of moment conditions, which appear in the Bernstein inequality in the high-dimensional literature except for the multiplicative factor "h". This extra factor is due to the presence of the kernel weight K_i. Similarly, the third condition is a localized version of the compatibility condition. The fourth condition contains standard assumptions on the kernel K and bandwidth h in the literature of nonparametric methods.

The ℓ_1-risk properties of the local Lasso estimators are obtained as follows.

Theorem 1. *Suppose $\sqrt{\log p/(nh)} = o(\lambda_n)$.*

(i): *Under Assumption 1, it holds*

$$|\hat{\theta} - \theta^*|_1 \leq 4\lambda_n \frac{s^*}{\phi_*^2}, \tag{6}$$

with probability approaching one.

(ii): *Under Assumption 1 with S^* containing all the indexes for θ_1, it holds*

$$|\tilde{\theta} - \theta^*|_1 \leq C\lambda_n \frac{s^*}{\phi_*^2}, \tag{7}$$

for some $C \in (0, \infty)$ with probability approaching one.

(iii): *Under Assumption 1 with S^* containing all the indexes for θ_1, it holds*

$$|\bar{\theta}_{\bar{S}} - \tilde{\theta}_{\bar{S}}|_2 \leq \lambda_{\min}\left(\frac{1}{nh} G'_{\bar{S}} G_{\bar{S}}\right)^{-1} \lambda_n |\bar{S}|. \tag{8}$$

The proof of this theorem is presented in Appendix A.1. This theorem characterizes the risk properties of the estimators $\hat{\theta}$, $\tilde{\theta}$, and $\bar{\theta}$ around θ^*, and the bounds depend on the tuning parameter λ_n, the number of non-zero coefficients s^*, and the compatibility constant ϕ_*. Note that the decay rate of λ_n is bounded from below by $\sqrt{\log p/(nh)}$. Thus, the risk bounds of $\hat{\theta}$ and $\tilde{\theta}$ get worse as the number of covariates p increases or the effective sample size nh for the kernel localization decreases. The result for the post-selection estimator $\bar{\theta}$ shows that the deviation from the original Lasso estimator $\tilde{\theta}$ is small when tuning parameter λ_n or the number of selected covariates $|\bar{S}|$ is small, or the minimum eigenvalue of $\frac{1}{nh} G'_{\bar{S}} G_{\bar{S}}$ is large.

The above theorem is on estimation of the coefficients of the best linear predictor $\theta^* = (\alpha^*, \tau^*, \beta^*_-, \beta^*_+, \gamma^{*\prime})'$. Additionally suppose that the assumptions of Lemma 1 of CCFT hold true, and the covariates Z_i are predetermined. Then we can guarantee that τ^* coincides with the average causal effect τ in (1) so that Theorem 1 provides the conditions for the consistency and convergence rate of $\hat{\tau}$ to τ. If Z_i are not predetermined (i.e., $Z_i(0) \neq_d Z_i(1)$), then $\hat{\tau}$ typically converges to τ minus some bias component, which is obtained as a limit of CCFT's bias term in their Lemma 1.

Our estimators and above theorem can be extended to other regression models that contain the covariates $\{T_i Z_i, (1-T_i) Z_i\}$, $(Z_i - \bar{Z})$, or $\{T_i(Z_i - \bar{Z}), (1-T_i)(Z_i - \bar{Z})\}$ as in CCFT. However, as shown in Lemma 1 of CCFT, such estimators require more stringent conditions to guarantee the consistency for τ. Furthermore, the local lasso regression (3) can be extended to incorporate the polynomials of X_i and T_iX_i even though this paper focuses on the local linear model.

Finally, we discuss the choices of the localization bandwidth h and regularization parameter λ_n. We can use the MSE-optimal bandwidth based on the suggestion by CCFT and the regularization parameter λ_n using cross-validation by Friedman, Hastie and Tibshirani (2010) and the data-driven choice by Belloni, Chernozhukov and Hansen (2014) among others. Our recommendation is stated in Section 4.

2.2. **Inference.** We next consider interval estimation and hypothesis testing on the average causal effect τ. For finite or low-dimensional Z_i, we recommend to use CCFT's inference methods. This subsection argues that we can still apply CCFT's inference procedures for high-dimensional Z_i, provided that CCFT's conditions remain valid for S^* and $\theta^*_{S^*}$.

More precisely, based on the point estimator $\hat{\theta}$ in (3) or $\tilde{\theta}$ in (4), define the subsets of $\{1, \ldots, p\}$ as

$$\hat{S} = \left\{ j : |\hat{\gamma}^{(j)}| > \lambda_n \varrho_n \sum_{j=1}^{p} \mathbb{I}\{|\hat{\gamma}^{(j)}| > 0\} \right\}, \qquad \tilde{S} = \left\{ j : |\tilde{\gamma}^{(j)}| > \lambda_n \varrho_n \sum_{j=1}^{p} \mathbb{I}\{|\tilde{\gamma}^{(j)}| > 0\} \right\},$$

where we set $\varrho_n = \log\log\log n$. This choice of ϱ_n is based on the simulation experiments in Section 4 and it is not optimal in any sense but works reasonably well. The consistency of these selection procedures is presented as follows.

Theorem 2. *Suppose* $\sqrt{\log p/(nh)} = o(\lambda_n)$. *Under Assumption 1 and additionally* $|\theta^{*(j)}| > \lambda_n \varrho_n s^*(1+\varepsilon)$ *for each* $j \in S^*$ *for some* $\varepsilon > 0$, *it holds*

$$\mathbb{P}\{\hat{S} = S^*\} \to 1, \qquad \mathbb{P}\{\tilde{S} = S^*\} \to 1. \tag{9}$$

(For the second statement, S^* should contain θ_1.)

This theorem says that under the additional β-min type condition $|\theta^{*(j)}| > \lambda_n \varrho_n s^*(1+\varepsilon)$, our selectors $\hat{S}$ and $\tilde{S}$ consistently estimate the true set of non-zero coefficients S^*.

Let $\hat{Z}_i$ and $\tilde{Z}_i$ be subvectors of Z_i selected by $\hat{S}$ and $\tilde{S}$, respectively. If we additionally assume that $\hat{Z}_i$ or $\tilde{Z}_i$ satisfies the assumptions of Theorem 2 in CCFT (which include finite s^*), then the t-statistic in Theorem 2 of CCFT for the null hypothesis of $H_0 : \tau = 0$ using $\hat{Z}_i$ or $\tilde{Z}_i$ (denoted by $\hat{T}$ or $\tilde{T}$) satisfies

$$\hat{T}, \tilde{T} \xrightarrow{d} N(0,1). \tag{10}$$

Alternatively, we may conduct inference based on the double selection procedure (see, Belloni, Chernozhukov and Hansen, 2014) when some elements of γ may take smallish nonzero values. In particular, we run the additional local lasso regression from T_i on Z_i, i.e.,

$$\bar{\gamma} = \arg\min_{\gamma} \frac{1}{nh} \sum_{i=1}^{n} K_i (T_i - Z_i' \gamma)^2 + \lambda_n |\gamma|_1.$$

Letting $\bar{S} = \left\{ j : |\bar{\gamma}^{(j)}| > \lambda_n \varrho_n \sum_{j=1}^{p} \mathbb{I}\{|\bar{\gamma}^{(j)}| > 0\} \right\}$, CCFT's t-statistics can also be constructed by using the selected sets $\hat{S} \cup \bar{S}$ or $\tilde{S} \cup \bar{S}$.

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 still affects the treatment probability. In particular, the fuzzy RDD assumes that the conditional treatment probability $\mathbb{P}\{W_i = 1 | X_i = x\}$ jumps at the cutoff point $\bar{x}$. As in the last section, we normalize $\bar{x} = 0$. 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 $x = 0$. 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 $x = 0$,

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

Hahn, Todd and van der Klaauw (2001) showed that under mild conditions the parameter τ_f can be identified by the ratio of the jump in the conditional mean of Y_i at $x = 0$ to the jump in the conditional treatment probability at $X_i = 0$, i.e.,

$$\tau_f = \frac{\lim_{x \downarrow 0} \mathbb{E}[Y_i | X_i = x] - \lim_{x \uparrow 0} \mathbb{E}[Y_i | X_i = x]}{\lim_{x \downarrow 0} \mathbb{P}\{W_i = 1 | X_i = x\} - \lim_{x \uparrow 0} \mathbb{P}\{W_i = 1 | X_i = x\}}. \tag{11}$$

In this case, letting $T_i = \mathbb{I}\{X_i \geq 0\}$, the numerator and denominator of (11) can be estimated by the local Lasso estimators $\hat{\theta}_Y$ and $\hat{\theta}_W$, which solve

$$\min_{\theta_Y} \frac{1}{nh} \sum_{i=1}^{n} K_i \left\{ Y_i - \alpha_Y - T_i \tau_Y - X_i \beta_{Y-} - T_i X_i \beta_{Y+} - Z_i' \gamma_Y \right\}^2 + \lambda_n |\theta_Y|_1,$$

$$\min_{\theta_W} \frac{1}{nh} \sum_{i=1}^{n} K_i \left\{ W_i - \alpha_W - T_i \tau_W - X_i \beta_{W-} - T_i X_i \beta_{W+} - Z_i' \gamma_W \right\}^2 + \lambda_n |\theta_W|_1,$$

respectively. Then, based on the union of the selected covariates from the above local Lasso regressions (i.e., $\bar{S} = \{ j : |\hat{\theta}_Y^{(j)}| > 0, \text{ or } |\hat{\theta}_W^{(j)}| > 0 \}$), we implement the local least squares as in CCFT:

$$\min_{\theta_{Y,\bar{S}}} \frac{1}{nh} \sum_{i=1}^{n} K_i \left\{ Y_i - G_{\bar{S},i}' \theta_{Y,\bar{S}} \right\}^2, \qquad \min_{\theta_{W,\bar{S}}} \frac{1}{nh} \sum_{i=1}^{n} K_i \left\{ W_i - G_{\bar{S},i}' \theta_{W,\bar{S}} \right\}^2,$$

where $G_{\bar{S}} = (1, T, X, TX, Z'_{\bar{S}})'$. The numerator and denominator of (11) are given by the estimated coefficients of τ_Y and τ_W for the above minimizations, respectively. If the treatment variable W satisfies analogous conditions for Theorems 1 and 2 (by replacing W with Y), we expect that analogous results to the sharp RDD case can be established.

3.2. **Regression kink design.** Our high-dimensional method can be extended for the RKDs. For each unit $i = 1, \dots, n$, 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 $\mathbb{E}[Y|X = x]$ to identify a treatment effect of B. Suppose $b(\cdot)$ is kinked at 0. Otherwise, we redefine X by subtracting the kink point c from X. In particular, Card, *et al.* (2015) have shown that a treatment on treated parameter $\tau_k = \int \frac{\partial f(b,x,u)}{\partial b} dF_{U|B=b,X=x}(u)$ is identified as

$$\tau_k = \frac{\lim_{x \downarrow 0} \frac{d}{dx}\mathbb{E}[Y|X = x] - \lim_{x \uparrow 0} \frac{d}{dx}\mathbb{E}[Y|X = x]}{\lim_{x \downarrow 0} \frac{d}{dx} b(x) - \lim_{x \uparrow 0} \frac{d}{dx} b(x)}. \tag{12}$$

To estimate τ_k, we propose the following local lasso regression

$$\min_{\theta} \frac{1}{nh} \sum_{i=1}^{n} K_i \left\{ Y_i - \alpha - T_i X_i \delta - X_i \beta - T_i X_i^2 \zeta - X_i^2 \eta - Z_i' \gamma \right\}^2 + \lambda_n |\theta|_1, \qquad (13)$$

where $\theta = (\alpha, \delta, \beta, \zeta, \eta, \gamma')'$ is a vector of parameters. Let $\hat{\delta}$ be the lasso estimator of δ by (13). Since the denominator $b_0 = \lim_{x \downarrow c} \frac{d}{dx} b(x) - \lim_{x \uparrow c} \frac{d}{dx} b(x)$ in (12) is assumed to be known, the estimator of τ_k is given by $\hat{\tau}_k = \hat{\delta}/b_0$. Under analogous conditions to Theorems 1 and 2 (by setting $\beta_+ = 0$ in Assumption 1), we expect that analogous results to the sharp RDD case can be established.

4. Simulation

In this section, we conduct simulation experiments to investigate finite sample properties of our covariate selection approach for estimation and inference on the sharp RDD analysis. We consider three simulation designs based on CCFT with introducing additional covariates. The additional covariates are generated based on the simulation designs in Belloni, Chernozhukov and Hansen (2014). Let $\mathcal{B}(a, b)$ be a beta distribution with parameters a and b. The data generating process (DGP) is specified as follows

$$Y = \mu_1(X) + \mu_2(Z) + \mu_3(W) + \varepsilon_y, \quad X \sim 2\mathcal{B}(2, 4) - 1, \quad Z = \mu_z(X) + \varepsilon_z,$$

$$\mu_z(x) = \begin{cases} 0.49 + 1.06x + 5.74x^2 + 17.14x^3 + 19.75x^5 + 7.47x^5 & \text{for } x < 0, \\ 0.49 + 0.61x + 0.23x^2 - 3.46x^3 + 6.43x^4 - 3.48x^5 & \text{for } x \geq 0, \end{cases}$$

$W = (W_1, W_2, \ldots, W_p)' \sim N(0, \Sigma_W)$ with $\mathbb{E}(W_h^2) = 1$ and $Cov(W_h, W_l) = 0.5^{|h-l|}$, and

$$\begin{pmatrix} \varepsilon_y \\ \varepsilon_z \end{pmatrix} \sim N(0, \Sigma), \quad \Sigma = \begin{pmatrix} \sigma_y^2 & \rho \sigma_y \sigma_z \\ \rho \sigma_y \sigma_z & \sigma_z^2 \end{pmatrix},$$

with $\sigma_y = 0.1295$ and $\sigma_z = 0.1353$.

For the functions μ_1, μ_2, and μ_3, we consider three cases. For DGP1, we set $\rho = 0.2692$, all the coefficients of $\mu_2(z)$ and $\mu_3(w)$ to be zero, and

$$\mu_1(x) = \begin{cases} 0.48 + 1.27x + 7.18x^2 + 20.21x^3 + 21.54x^4 + 7.33x^5 & \text{for } x < 0, \\ 0.52 + 0.84x - 3.00x^2 + 7.99x^3 - 9.01x^4 + 3.56x^5 & \text{for } x \geq 0, \end{cases}$$

For DGP2, we set $\rho = 0.2692$,

$$\mu_1(x) = \begin{cases} 0.36 + 0.96x + 5.47x^2 + 15.28x^3 + 15.87x^4 + 5.14x^5 & \text{for } x < 0, \\ 0.38 + 0.62x - 2.84x^2 + 8.42x^3 - 10.24x^4 + 4.31x^5 & \text{for } x \geq 0, \end{cases}$$

$$\mu_2(z) = \begin{cases} 0.22z & \text{for } x < 0, \\ 0.28z & \text{for } x \geq 0, \end{cases}$$

and $\mu_3(w) = \sum_{h=1}^{p} \pi_h w_h$ with $\pi_h = 0.2^h$. For DGP3, we set $\mu_1(x)$ and $\mu_2(z)$ as in DGP2, and $\mu_3(w) = \sum_{h=1}^{p} \pi_h w_h$ with $\pi_h = 0.5^h$. The sample size is set as $n = 500$ for all the cases. The number of the covariates p varies from 5 to 500. The results are based on 1,000 Monte Carlo replications.

Table 1 shows the biases and RMSEs of the four point estimation methods. For the bandwidth h, the first two methods use the MSE-optimal bandwidth without covariates proposed by CCT. The third method uses the MSE-optimal bandwidth with covariates proposed by CCFT. The fourth method, called "Adaptive", is the bandwidth for our covariate selection approach which uses the MSE optimal bandwidth without covariates for the covariate selection stage and uses that with covariates in the estimation stage. For estimation methods, the first method uses the standard RD estimation method without covariates by CCT. The second and the third methods use the RD estimation with covariates by CCFT. The fourth method uses the RD estimation with the selected covariates. The Lasso procedure applied in the covariate selection stage employs the data-driven penalty level by Belloni, Chernozhukov and Hansen (2014) and the MSE-optimal bandwidth without covariates by CCT. For the RD estimation, we employ the MSE-optimal bandwidth with covariates by CCFT.

Our findings are summarized as follows. First, the RMSEs of the covariate-adjusted estimation get larger irrespective of the bandwidths as the number of covariates increases across all DPGs. These increases in the RMSEs are due to inflated standard errors caused by a large number of covariates. This result clearly indicates the need for covariate selection. Second, the covariate selection approach shows excellent performances for all cases. Both the biases and RMSEs are stable for different values of p for all designs. Finally, all methods work equally well for DGP1, where all the additional covariates are irrelevant. However, for DGP2 and DGP3, we find substantial efficiency loss of the standard method. Overall, we recommend the covariate selection even for relatively small p.

Table 2 reports the number of selected covariates for our covariate selection approach. It is quite natural that the number of selected covariates increases when the number of non-zero coefficients of covariates increases. It is interesting to note that the average number of selected covariates decreases as the number of covariates increase.

Table 3 shows the coverage probabilities and interval lengths of the robust confidence intervals for the causal effect. The nominal coverage level is 0.95. The following points are notable. First, the performances of our covariate selection approach are stable for all DGPs, although the coverage probabilities tend to be a little bit smaller than the nominal level. Second, for the covariate-adjusted approaches, the coverage probabilities decrease and the interval length gets shorter as p increases. Third, the coverage of CCT is more stable and better than other methods, especially for DGP2 and DGP3. However, the average lengths of CCT are substantially longer than the other methods. Overall, the covariate selection approach is promising for inference as well since it exhibits robust performances in both coverages and lengths for different number of covariates and DGPs.

Finally, Table 4 presents the properties of the MSE-optimal bandwidths. We can observe that the MSE-optimal bandwidth without covariates and the adaptive one are very stable while the MSE optimal bandwidth with covariates shrinks as p increases. This is possibly the main source of the increased RMSE and under-coverages.

Table 1. Simulation: Point estimation

MSE-Optimal bandwidths:		w/o Covariates				w/ Covariates		Adaptive	
Estimation methods:		Standard		Covariate adjusted		Covariate adjusted		Covariate selection	
	p	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
DGP1	5	0.019	0.063	0.023	0.064	0.019	0.065	0.019	0.064
	10	0.020	0.063	0.022	0.068	0.019	0.069	0.019	0.061
	20	0.020	0.061	0.019	0.066	0.014	0.073	0.016	0.064
	30	0.018	0.063	0.022	0.074	0.011	0.093	0.020	0.062
	40	0.015	0.064	0.022	0.080	0.004	0.141	0.019	0.064
	50	0.022	0.065	0.027	0.101	0.012	0.397	0.019	0.060
	100	0.019	0.062	0.038	0.387	0.000	0.130	0.023	0.065
	250	0.023	0.064	0.026	0.082	0.012	0.108	0.019	0.064
	500	0.022	0.065	0.028	0.072	0.012	0.096	0.022	0.059
DGP2	5	0.004	0.526	0.000	0.058	0.030	0.082	0.029	0.095
	10	-0.007	0.527	-0.001	0.061	0.027	0.087	0.028	0.100
	20	0.032	0.507	-0.007	0.064	0.031	0.098	0.025	0.097
	30	0.017	0.529	-0.006	0.070	0.026	0.194	0.026	0.101
	40	0.008	0.532	-0.012	0.080	0.032	0.223	0.027	0.100
	50	0.001	0.519	-0.009	0.087	0.011	0.420	0.026	0.101
	100	0.021	0.516	-0.032	0.462	0.047	0.509	0.025	0.107
	250	0.003	0.542	-0.042	0.310	-0.114	4.310	0.024	0.109
	500	0.021	0.543	-0.044	0.402	0.021	0.806	0.020	0.114
DGP3	5	-0.005	0.673	-0.004	0.058	0.030	0.082	0.027	0.109
	10	-0.012	0.689	-0.005	0.060	0.027	0.087	0.020	0.177
	20	0.033	0.672	-0.011	0.064	0.031	0.098	0.024	0.178
	30	0.015	0.701	-0.011	0.070	0.026	0.194	0.015	0.185
	40	-0.005	0.674	-0.017	0.080	0.032	0.222	0.015	0.185
	50	-0.004	0.695	-0.015	0.085	-0.005	0.460	0.027	0.195
	100	-0.007	0.689	-0.036	0.301	0.009	0.621	0.019	0.190
	250	-0.013	0.689	-0.055	0.347	0.009	0.861	0.012	0.201
	500	0.031	0.663	-0.050	0.485	0.095	1.030	0.014	0.216

Table 2. Simulation: Number of selected covariates

	p	Average	Min	Max
DGP1	5	0.409	0	1
	10	0.371	0	1
	20	0.379	0	1
	30	0.377	0	1
	40	0.347	0	2
	50	0.338	0	1
	100	0.342	0	2
	250	0.322	0	1
	500	0.310	0	1
DGP2	5	2.869	2	3
	10	2.789	2	3
	20	2.717	1	3
	30	2.647	1	3
	40	2.573	1	3
	50	2.564	1	3
	100	2.461	1	3
	250	2.392	1	3
	500	2.325	1	3
DGP3	5	3.915	2	5
	10	3.080	2	5
	20	3.013	2	4
	30	2.98	2	5
	40	2.944	1	4
	50	2.940	1	5
	100	2.866	1	5
	250	2.725	1	4
	500	2.548	0	4

Table 3. Simulation: Inference

MSE-Optimal bandwidths:		w/o Covariates				w/ Covariates		Adaptive	
Estimation methods:		Standard		Covariate adjusted		Covariate adjusted		Covariate selection	
	p	CP	Length	CP	Length	CP	Length	CP	Length
DGP1	5	0.923	0.205	0.898	0.246	0.892	0.241	0.915	0.231
	10	0.910	0.226	0.855	0.211	0.851	0.219	0.930	0.188
	20	0.930	0.172	0.815	0.259	0.804	0.173	0.911	0.190
	30	0.915	0.215	0.734	0.148	0.644	0.165	0.911	0.226
	40	0.911	0.214	0.649	0.182	0.458	0.168	0.923	0.335
	50	0.898	0.309	0.54	0.152	0.258	0.074	0.919	0.223
	100	0.918	0.218	0.194	0.342	0.193	0.176	0.896	0.161
	250	0.908	0.344	0.108	0.040	0.172	0.068	0.906	0.225
	500	0.912	0.195	0.138	0.020	0.188	0.097	0.931	0.262
DGP2	5	0.924	2.472	0.661	0.232	0.852	0.278	0.842	0.548
	10	0.925	2.449	0.627	0.209	0.786	0.234	0.845	0.466
	20	0.929	2.456	0.591	0.261	0.653	0.167	0.872	0.490
	30	0.928	2.305	0.517	0.136	0.462	0.178	0.880	0.465
	40	0.926	2.006	0.472	0.170	0.286	0.128	0.890	0.438
	50	0.927	2.276	0.407	0.127	0.191	0.762	0.876	0.548
	100	0.933	2.563	0.156	0.069	0.172	0.754	0.883	0.284
	250	0.933	1.647	0.135	0.103	0.177	0.195	0.890	0.640
	500	0.919	1.820	0.137	0.112	0.174	0.431	0.900	0.509
DGP3	5	0.934	3.350	0.615	0.219	0.852	0.278	0.861	0.329
	10	0.911	2.965	0.601	0.202	0.786	0.234	0.891	0.704
	20	0.925	3.395	0.564	0.236	0.653	0.167	0.905	0.721
	30	0.920	2.951	0.495	0.151	0.462	0.178	0.899	0.713
	40	0.940	2.848	0.456	0.174	0.288	0.128	0.903	0.740
	50	0.928	2.654	0.397	0.132	0.188	0.071	0.885	0.697
	100	0.933	2.954	0.182	0.073	0.169	0.407	0.918	0.910
	250	0.927	3.446	0.119	0.139	0.176	0.529	0.895	0.765
	500	0.939	2.820	0.130	0.158	0.173	0.389	0.908	0.751

Table 4. Simulation: MSE-Optimal Bandwidths

MSE-Optimal bandwidths:		w/o Covariates		w/ Covariates		Adaptive	
	p	Mean	SD	Mean	SD	Mean	SD
DGP1	5	0.194	0.044	0.187	0.042	0.194	0.046
	10	0.195	0.043	0.181	0.041	0.195	0.045
	20	0.195	0.045	0.162	0.037	0.196	0.045
	30	0.196	0.045	0.146	0.031	0.196	0.045
	40	0.195	0.045	0.127	0.026	0.196	0.046
	50	0.195	0.045	0.108	0.024	0.196	0.045
	100	0.198	0.044	0.073	0.023	0.197	0.047
	250	0.199	0.045	0.070	0.020	0.195	0.046
	500	0.197	0.045	0.070	0.020	0.198	0.044
DGP2	5	0.176	0.025	0.108	0.011	0.111	0.014
	10	0.176	0.025	0.107	0.011	0.112	0.014
	20	0.176	0.025	0.104	0.011	0.114	0.015
	30	0.176	0.024	0.102	0.012	0.116	0.015
	40	0.176	0.026	0.098	0.012	0.117	0.015
	50	0.176	0.024	0.092	0.013	0.117	0.015
	100	0.177	0.025	0.079	0.020	0.119	0.015
	250	0.176	0.025	0.077	0.022	0.12	0.015
	500	0.175	0.024	0.076	0.021	0.121	0.015
DGP3	5	0.182	0.028	0.108	0.011	0.117	0.014
	10	0.182	0.028	0.107	0.011	0.138	0.017
	20	0.182	0.028	0.104	0.011	0.139	0.016
	30	0.181	0.028	0.102	0.012	0.139	0.015
	40	0.182	0.029	0.098	0.012	0.140	0.016
	50	0.183	0.028	0.093	0.014	0.140	0.016
	100	0.183	0.029	0.078	0.020	0.141	0.016
	250	0.182	0.029	0.076	0.022	0.142	0.017
	500	0.182	0.028	0.074	0.021	0.144	0.019

5. Empirical illustration: Head Start data

To illustrate our variable selection approach, we revisit the problem of the Head Start program first studied by Ludwig and Miller (2007) where they investigate the effect of the Head Start program on various outcomes related to health and schooling. The federal government provided grant-writing assistance to the 300 poorest counties based on the poverty index to apply for the Head Start program. This leads to the RD design with the poverty index as a running variable where the cut-off value is set as $\bar{x} = 59.1984$. Ludwig and Miller (2007) conducted their RDD analysis using no covariate, and CCFT examined the impact of the covariance-adjustment. CCFT employed nine pre-intervention covariates from the U.S. Census, which include total population, percentages of population, percentages of black and urban population, and levels and percentages of population in three age groups (children aged 3 to 5, children aged 14 to 17, and adults older than 25). The main finding by CCFT is that the covariate-adjusted RD inference yields shorter confidence intervals while the RD point estimates remain stable.

We conduct the empirical exercises of CCFT by applying our variable selection approach with two extensions. First, we introduce 36 interaction terms in addition to the nine original covariates. Second, we also implement those estimation and inference for subsamples to see the effect of changes in the ratio of the number of covariates (p) to that of observations (n). Hereafter, as in CCFT, we focus on child mortality among many outcome variables.

Table 5 shows the results of our empirical illustration. Four columns correspond to four estimation procedures which are the same as those used in the simulation experiments. The first panel shows the full sample results ($n = 2799$ and $p/n = 0.016$). The RD causal effect estimates are presented in the first row. The next three rows show 95% confidence intervals, their percentage length changes relative to the one in the first column, and their associated p-values where these are obtained without restriction on the MSE optimal bandwidth for the local linear regression (h) and the pilot bandwidth (b). See CCT and CCFT for more details on the robust inference methods. These results are also obtained under the restriction $h/b = 1$, which are reported in the following three rows. The last two rows in the same panel present the bandwidths (h, b), and effective sample sizes (n_-, n_+) used for the RD estimates. The effective sample sizes are the number of observations with the running variable in the intervals $[\bar{x} - h, \bar{x}]$ and $[\bar{x}, \bar{x} + h]$. We also report the selected covariates for our covariate selection approach. We use subsamples of the first 1000 and 500 observations for the second and third panels, leading to $p/n = 0.045$ and .0.090, respectively.

For the full sample case, the covariate-adjusted estimates mildly deviate from the standard one while our estimate based on the variable selection is identical to the standard one. Although the confidence intervals of the covariate-adjusted approaches are shorter than the standard one, this might induce under-coverages for the case of many covariates as illustrated in the simulation experiment. As the sample size gets smaller, the observations made here are amplified. In contrast, we can see the stable performance of the variable selection approach and its mild contribution to shorten the confidence intervals.

Table 5. Empirical illustration: Head Start data

	MSE-Optimal bandwidths:	w/o Covariates		w/ Covariates	Adaptive
	Estimation methods:	Standard	Cov-adjusted	Cov-adjusted	Variable selection
$n = 2779$	Point estimate	-2.41	-2.19	-3.14	-2.41
$p/n = 0.016$	h/b unrestricted				
	Robust 95% CI	$[-5.46, -0.1]$	$[-4.7, -0.27]$	$[-5.59, -0.56]$	$[-5.46, -0.1]$
	CI length change (%)		-17.42	-6.18	0
	Robust p-value	0.042	0.028	0.017	0.042
	$h/b = 1$				
	Robust 95% CI	$[-6.41, -1.09]$	$[-5.75, -1.07]$	$[-6.37, -0.31]$	$[-6.41, -1.09]$
	CI length change (%)		-12.05	-13.82	0
	Robust p-value	0.006	0.004	0.031	0.006
	h, b	6.81, 10.73	6.81, 10.73	3.26, 6.05	6.81, 10.73
	n_-, n_+	234, 180	234, 180	99, 94	234, 180
	Selected covariates				None
$n = 1000$	Point estimate	-1.68	-3.1	-4.1	-1.48
$p/n = 0.045$	h/b unrestricted				
	Robust 95% CI	$[-5.45, 1.75]$	$[-6.35, -1.1]$	$[-7.7, -2.28]$	$[-5.08, 1.79]$
	CI length change (%)		-26.92	-24.6	-4.49
	Robust p-value	0.314	0.005	0.000	0.347
	$h/b = 1$				
	Robust 95% CI	$[-8.26, 0.22]$	$[-8.27, -1.87]$	$[-8.74, -1.85]$	$[-7.94, 0.27]$
	CI length change (%)		-24.52	-18.78	-3.28
	Robust p-value	0.063	0.002	0.003	0.070
	h, b	6.52, 10.23	6.52, 10.23	5.26, 8.07	5.26, 8.07
	n_-, n_+	74, 77	79, 79	64, 69	79, 79
	Selected covariates				% of adult population
$n = 500$	Point estimate	-2.35	-4.51	-7.1	-2.22
$p/n = 0.090$	h/b unrestricted				
	Robust 95% CI	$[-7.25, 2.48]$	$[-8.59, -2.38]$	$[-11.57, -5.28]$	$[-6.93, 2.25]$
	CI length change (%)		-36.26	-35.47	-5.74
	Robust p-value	0.337	0.001	0.000	0.317
	$h/b = 1$				
	Robust 95% CI	$[-10.22, 1.42]$	$[-10.67, -3.35]$	$[-12.33, -4.45]$	$[-10, 1.07]$
	CI length change (%)		-37.1	-32.34	-4.94
	Robust p-value	0.139	0.000	0.000	0.110
	h, b	6.37, 9.16	6.37, 9.16	4.31, 6.82	4.31, 6.82
	n_-, n_+	60, 56	61, 56	42, 42	61, 56
	Selected covariates				% of adult population

Appendix A. Mathematical appendix

A.1. **Proof of Theorem 1.** We use the following modification of Bernstein's inequality.

Lemma A.1. *Under Assumption 1, it holds*

$$\mathbb{E}\left[\max_{1\leq j\leq p}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}K_iG_{ij}e_i\right|^m\right]\leq 2h^{m/2}\log^{m/2}p,$$

for $m\leq 1+\log p$.

Proof of Lemma A.1. Note that $\mathbb{E}[K_iG_ie_i]=0$ by construction. From Bernstein's inequality (e.g., Lemma 14.12 in BG)

$$\mathbb{E}\left[\max_{1\leq j\leq p}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}K_iG_{ij}e_i\right|^m\right]\leq 2h^{m/2}\log^{m/2}p,$$

for $m\leq 1+\log p$. □

A.1.1. *Proof of (i).* Under Assumption 1-2, Lemma A.1 implies

$$\mathbb{E}\left[\max_{1\leq j\leq p}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}K_iG_{ij}e_i\right|\right]\leq 2h^{1/2}\log^{1/2}p.$$

Thus, we have

$$\mathbb{P}\{\mathcal{A}_n\}:=\mathbb{P}\left\{\frac{4}{nh}\left|\sum_{i=1}^{n}K_iG_ie_i\right|_\infty\leq\lambda_n\right\}\to 1, \tag{14}$$

as $n\to\infty$, provided that $\sqrt{\log p/(nh)}=o(\lambda_n)$.

Let Y, e, and G denote the vector collections of $Y_iK_i^{1/2}$, $e_iK_i^{1/2}$, and $G_i'K_i^{1/2}$, respectively. Since $\hat{\theta}$ is a minimizer, we have

$$\frac{1}{nh}|Y-G\hat{\theta}|_2^2+\lambda_n|\hat{\theta}|_1\leq\frac{1}{nh}|Y-G\theta^*|_2^2+\lambda_n|\theta^*|_1.$$

By plugging $Y = G\theta^* + e$ into the above, we obtain

$$\begin{aligned}\frac{2}{nh}|G(\hat{\theta}-\theta^*)|_2^2 &\leq \frac{4}{nh}e'G(\hat{\theta}-\theta^*) + 2\lambda_n|\theta^*|_1 - 2\lambda_n|\hat{\theta}|_1 \\ &\leq \frac{4}{nh}|e'G|_\infty|\hat{\theta}-\theta^*|_1 + 2\lambda_n(|\theta^*|_1 - |\hat{\theta}|_1), \\ &\leq 3\lambda_n|\hat{\theta}_{S^*} - \theta^*_{S^*}|_1 - \lambda_n|\hat{\theta}_{S^*_c}|_1, \end{aligned} \tag{15}$$

conditionally on $\mathcal{A}_n$, where S^*_c is the complement of S^*, the second inequality follows from the Hölder inequality, and the third inequality follows from the definition of $\mathcal{A}_n$ and the following facts

$$|\hat{\theta}-\theta^*|_1 = |\hat{\theta}_{S_*} - \theta^*_{S_*}|_1 + |\hat{\theta}_{S^*_c}|_1, \tag{16}$$

$$|\theta^*|_1 - |\hat{\theta}|_1 = |\theta^*_{S^*}|_1 - |\hat{\theta}_{S^*}|_1 - |\hat{\theta}_{S^*_c}|_1 \leq |\hat{\theta}_{S^*} - \theta^*_{S^*}|_1 - |\hat{\theta}_{S^*_c}|_1,$$

due to the triangle inequality. Thus, (15) implies $3|\hat{\theta}_{S^*} - \theta^*_{S^*}|_1 \geq |\hat{\theta}_{S^*_c}|_1$ and

$$\frac{2}{nh}|G(\hat{\theta}-\theta^*)|_2^2 + \lambda_n|\hat{\theta}-\theta^*|_1 \leq 4\lambda_n|\hat{\theta}_{S^*} - \theta^*_{S^*}|_1, \tag{17}$$

by using (16).

Now, Assumption 1-3 implies

$$4\lambda_n|\hat{\theta}_{S^*} - \theta^*_{S^*}|_1 \leq 4\lambda_n\sqrt{\frac{s^*}{nh\phi_*^2}}|G(\hat{\theta}-\theta^*)|_2 \leq \frac{1}{nh}|G(\hat{\theta}-\theta^*)|_2^2 + 4\lambda_n^2\frac{s^*}{\phi_*^2},$$

with probability approaching one, where we note that $2ab \leq a^2 + b^2$ for the second inequality. Combining this with (17), we have

$$\frac{1}{nh}|G(\hat{\theta}-\theta^*)|_2^2 + \lambda_n|\hat{\theta}-\theta^*|_1 \leq 4\lambda_n^2\frac{s^*}{\phi_*^2},$$

and the conclusion in (6) follows.

A.1.2. *Proof of (ii).* The compatibility condition for G_1 is implied by the usual full column rank condition in the classical linear regression since $|\theta_1|_1^2 \leq \dim(\theta_1)|\theta_1|_2^2$.

Note that the result in (14) still holds. Since $\tilde{\theta}$ is a minimizer, we have

$$\frac{1}{nh}|Y - G\tilde{\theta}|_2^2 + \lambda_n|\tilde{\theta}_2|_1 \leq \frac{1}{nh}|Y - G\theta^*|_2^2 + \lambda_n|\theta_2^*|_1.$$

By plugging $Y = G\theta^* + e$ into the above,

$$\begin{aligned}
\frac{2}{nh}|G(\tilde{\theta} - \theta^*)|_2^2 &\leq \frac{4}{nh}e'G(\tilde{\theta} - \theta^*) + 2\lambda_n|\theta_2^*|_1 - 2\lambda_n|\tilde{\theta}_2|_1 \\
&\leq \frac{4}{nh}|e'G|_\infty|\tilde{\theta} - \theta^*|_1 + 2\lambda_n(|\theta_2^*|_1 - |\tilde{\theta}_2|_1) \\
&\leq \lambda_n|\tilde{\theta}_1 - \theta_1^*|_1 + 3\lambda_n|\tilde{\theta}_{2,S^*} - \theta_{2,S^*}^*|_1 - \lambda_n|\tilde{\theta}_{2,S_c^*}|_1, \qquad (18)
\end{aligned}$$

conditionally on $\mathcal{A}_n$, where the second inequality follows from the Hölder inequality, and the third inequality follows from the definition of $\mathcal{A}_n$ and the following facts

$$\begin{aligned}
|\tilde{\theta} - \theta^*|_1 &= |\tilde{\theta}_{S^*} - \theta_{S^*}^*|_1 + |\tilde{\theta}_{S_c^*}|_1, \qquad (19)\\
|\theta^*|_1 - |\tilde{\theta}|_1 &= |\theta_{S^*}^*|_1 - |\tilde{\theta}_{S^*}|_1 - |\tilde{\theta}_{S_c^*}|_1 \leq |\tilde{\theta}_{S^*} - \theta_{S^*}^*|_1 - |\tilde{\theta}_{S_c^*}|_1,
\end{aligned}$$

due to the triangle inequality. Thus, (18) implies $3|\tilde{\theta}_{S_*} - \theta_{S_*}^*|_1 \geq |\tilde{\theta}_{S_*^c}|_1$ and

$$\frac{2}{nh}|G(\tilde{\theta} - \theta^*)|_2^2 + \lambda_n|\tilde{\theta}_2 - \theta_2^*|_1 \leq \lambda_n|\tilde{\theta}_1 - \theta_1^*|_1 + 4\lambda_n|\tilde{\theta}_{2,S^*} - \theta_{2,S^*}^*|_1, \qquad (20)$$

by using (19).

Now Assumption 1-3 implies

$$\begin{aligned}4\lambda_n|\tilde\theta_{2,S^*}-\theta^*_{2,S^*}|_1 &\leq 4\lambda_n\sqrt{\frac{s^*}{nh\phi_*^2}}|G_2(\tilde\theta_2-\theta_2^*)|_2 \leq \frac{1}{nh}|G_2(\tilde\theta_2-\theta_2^*)|_2^2+4\lambda_n^2\frac{s^*}{\phi_*^2},\\ \lambda_n|\tilde\theta_1-\theta_1^*|_1 &\leq \lambda_n\sqrt{\frac{s^*}{nh\phi_*^2}}|G_1(\tilde\theta_1-\theta_1^*)|_2 \leq \frac{1}{nh}|G_1(\tilde\theta_1-\theta_1^*)|_2^2+\lambda_n^2\frac{s^*}{\phi_*^2},\end{aligned}$$

with probability approaching one. Combining these inequalities with (20), we have

$$\frac{1}{nh}|G(\tilde\theta-\theta^*)|_2^2+\lambda_n|\tilde\theta_2-\theta_2^*|_1 \leq 5\lambda_n^2\frac{s^*}{\phi_*^2},$$

which implies

$$|\tilde\theta_2-\theta_2^*|_1 \leq 5\lambda_n\frac{s^*}{\phi_*^2}. \tag{21}$$

Turning to the finite dimensional component $\tilde\theta_1$, note that

$$\begin{aligned}\tilde\theta_1-\theta_1^* &= \left(\frac{1}{nh}G_1'G_1\right)^{-1}\left(\frac{1}{nh}G_1'e-\frac{1}{nh}G_1'G_2(\tilde\theta_2-\theta_2^*)\right)\\ &= O_p((nh)^{-1/2})+O_p(|\tilde\theta_2-\theta_2^*|_1).\end{aligned}$$

Combining this with (21) yields the conclusion in (7).

A.1.3. *Proof of (iii).* Since $\bar\theta_{\bar S}$ is the weighted OLS estimate that minimizes the sum of the squared residuals in the regression of Y on $G_{\bar S}$ and $\bar\theta_{\bar S^c}=\tilde\theta_{\bar S^c}=0$, it holds

$$|Y-G_{\bar S}\bar\theta_{\bar S}|_2^2 \leq |Y-G_{\bar S}\tilde\theta_{\bar S}|_2^2,$$

and thus,

$$\frac{1}{nh}|G_{\bar S}(\bar\theta_{\bar S}-\tilde\theta_{\bar S})|_2^2 \leq \frac{2}{nh}|\hat e' G_{\bar S}(\bar\theta_{\bar S}-\tilde\theta_{\bar S})| \leq \lambda_n|\bar\theta_{\bar S}-\tilde\theta_{\bar S}|_1, \tag{22}$$

due to the Hölder inequality and KKT condition. On the other hand,

$$\lambda_{\min}\left(\frac{1}{nh}G'_{\bar{S}}G_{\bar{S}}\right)|\bar{\theta}_{\bar{S}}-\tilde{\theta}_{\bar{S}}|_2^2 \le \frac{1}{nh}|G_{\bar{S}}(\bar{\theta}_{\bar{S}}-\tilde{\theta}_{\bar{S}})|_2^2.$$

Since$|a|_1 \le \sqrt{s}|a|_2$ for an s-dimensional vector a, we conclude that

$$|\bar{\theta}_{\bar{S}}-\tilde{\theta}_{\bar{S}}|_2 \le \lambda_{\min}\left(\frac{1}{nh}G'_{\bar{S}}G_{\bar{S}}\right)^{-1}\lambda_n|\bar{S}|.$$

A.2. **Proof of Theorem 2.** Let $a_n = \lambda_n \varrho_n \sum_{j=1}^{p} \mathbb{I}\{|\hat{\gamma}^{(j)}| > 0\}$. To prove that $\mathbb{P}\{\hat{S} = S^*\} \to 1$, we note that the deviation bound in Theorem 1 is asymptotically negligible to the threshold a_n. This implies that if $|\gamma^{*(j)}| = 0$, then $|\hat{\gamma}^{(j)}| \le 4s^*\lambda_n/\phi_*^2 = o(a_n)$ with probability approaching one, and otherwise $|\hat{\gamma}^{(j)}|$ must exceed the threshold a_n. The proof of $\mathbb{P}\{\tilde{S} = S^*\} \to 1$ is similar.

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] Belloni, A., Chernozhukov, V., Chetverikov, D., Hansen, C. and K. Kato (2018) High-dimensional econometrics and regularized GMM, Working paper.

[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 M. H. Farrell (2018) On the effect of bias estimation on coverage accuracy in nonparametric inference, *Journal of the American Statistical Association*, 113, 767-779.

[7] Calonico, S., Cattaneo, M. D. and M. H. Farrell (2020) Coverage error optimal confidence intervals for local polynomial regression, arXiv:1808.01398.

[8] Calonico, S., Cattaneo, M. D., Farrell, M. H., and R. Titiunik (2017) rdrobust: Software for regression discontinuity designs, *Stata Journal*, 17, 372-404.

[9] Calonico, S., Cattaneo, M. D., Farrell, M. H., and R. Titiunik (2019) Regression discontinuity designs using covariates, *Review of Economics and Statistics,* 101, 442-451.

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

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

[12] 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.

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

[14] Cattaneo, M. D. and J. C. Escanciano (2017) *Regression Discontinuity Designs: Theory and Applications*, in *Advances in Econometrics*, vol. 38, Emerald Group Publishing.

[15] Cattaneo, M. D., Titiunik, R. and G. Vazquez-Bare (2020) The regression discontinuity design, in Curini, L. and R. J. Franzese (eds.), *Handbook of Research Methods in Political Science and International Relations*, ch. 44, pp. 835-857, Sage Publications.

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

[17] Friedman, J. H., Hastie, T. and R. Tibshirani (2010) Regularization paths for generalized linear models via coordinate descent, *Journal of Statistical Software*, 33, 1-22.

[18] Frölich, M. and M. Huber (2019) Including covariates in the regression discontinuity design, *Journal of Business & Economic Statistics*, 37, 736-748.

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

[20] 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.*

[21] 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.

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

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

[24] 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.

[25] Ludwig, J. and D. L. Miller (2007) Does Head Start improve children's life chances? Evidence from a regression discontinuity design, *Quarterly Journal of Economics*, 122, 159-208.

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

[27] 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.

[28] 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.

Email address: yarai@waseda.jp

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

Email address: t.otsu@lse.ac.uk

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

Email address: myunghseo@snu.ac.kr