Econometric Theory, **41**, 2025, 1416–1451.
doi:10.1017/S0266466624000239

REGRESSION DISCONTINUITY DESIGN WITH POTENTIALLY MANY COVARIATES

YOICHI ARAI
Waseda University

TAISUKE OTSU
London School of Economics

MYUNG HWAN SEO
Seoul National University

This article examines high-dimensional covariates in regression discontinuity design (RDD) analysis. We introduce estimation and inference methods for the RDD models that incorporate covariate selection while maintaining stability across various numbers of covariates. The proposed methods combine a localization approach using kernel weights with ℓ_1-penalization to handle high-dimensional covariates. We provide both theoretical and numerical evidence demonstrating the efficacy of our methods. Theoretically, we present risk and coverage properties for our point estimation and inference methods. Conditions are given under which the proposed estimator becomes more efficient than the conventional covariate adjusted estimator at the cost of an additional sparsity condition. Numerically, our simulation experiments and empirical examples 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. Regression discontinuity design (RDD) has been widely applied in observational studies to identify the average treatment effect at the discontinuity point. For RDD, the causal parameters of interest are identified by differences in the left and right limits of the conditional mean functions (see, e.g.,

This article is a developed version of a previous manuscript (https://sticerd.lse.ac.uk/dps/em/em601.pdf) inspired by a 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 (Y.A.) and the ERC Consolidator Grant (SNP 615882) (T.O.). Financial support from the Center for National Competitiveness 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 (M.H.S.). Address correspondence to Taisuke Otsu, Department of Economics, London School of Economics, London, UK; e-mail: t.otsu@lse.ac.uk.

Imbens and Lemieux, 2008; Cattaneo, Titiunik, and Vazquez-Bare, 2020; Cattaneo and Titiunik, 2022; an edited volume by Cattaneo and Escanciano, 2017, and the references therein).

This article expands the growing literature on RDD by including covariates—an issue studied extensively by Calonico et al. (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 used in the RDD analysis with various additional predetermined covariates, such as demographic or socioeconomic characteristics for data units. For several RDD estimators that use covariates based on the local polynomial regression methods, CCFT investigated the Mean squared error (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 et al. (2017) for a statistical package on these methods.

In randomized controlled trials, regression adjustment using covariates is a common practice because it increases the asymptotic efficiency of the causal effect estimator, provided that a full set of treatment–covariate interactions is included (Lin, 2013). Also a recent article by Lei and Ding (2021) proposed a bias correction method for the regression adjustment estimator with a diverging number of covariates. On the other hand, RDD is a quasi-experimental design, so the efficiency gain from introducing covariates is not necessarily guaranteed; thus, CCFT provided a concrete guideline clarifying the conditions to achieve consistency and an efficiency gain for the covariate adjusted RDD estimator. Typically, efficiency improves when the projection coefficients of the covariates on the outcome are equal in the control and treatment groups. Since practitioners commonly incorporate covariates in RDD analysis, CCFT's guidelines have had a large impact on applied research. When we use covariates, it is common to use their transformations and interactions, leading to a potentially large number of covariates. This article adds a further guidance for practitioners who face a large number of covariates. First, the (weighted) Ordinary least squares (OLS) estimation in CCFT is not applicable when the number of covariates is larger than the sample size. Also, for the efficiency improvement in the above scenario, it is beneficial to augment CCFT's procedure with covariate selection using high-dimensional statistical methods particularly when the regression coefficients for the conditional mean function satisfy sparsity.

For point estimation on the causal effect parameter identified by RDD, we consider the Lasso estimator and its post-selection estimator based on the local linear regression (i.e., eqn. (2) of CCFT). The combination of localization using kernel weights and ℓ_1-penalization to deal with high-dimensional covariates is

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

particularly relevant for RDD analysis because the effective sample size is typically small due to the localization, so the cost of many covariates is exacerbated. Theoretically, we derive the ℓ_1-risk properties of our local Lasso estimator 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 as it exhibits desirable MSE properties and stability across different setups.

For inference, we propose to select covariates with the local Lasso. We show that inference based on the selected covariates can be implemented in the same manner as in CCFT. We also show that when the effect of the additional covariates on the potential outcomes with or without treatment is invariant, our approach can lead to improved efficiency at the cost of an additional sparsity condition. This sparsity condition is trivially satisfied when the set of active covariates is unknown but fixed. Our simulation results demonstrate that our post-selection confidence interval is robust in terms of both coverage and length, even for a relatively small number of covariates.

This article 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 kernel localization with ℓ_1-penalization to handle high-dimensional covariates. Our inference problem can be formulated in the same way as for low-dimensional parameters in high-dimensional models. In statistics literature, many articles have 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 in the RDD context because it concerns a jump in a nonparametric regression model.[1]

This article 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. In Section 3, we discuss some extensions of our approach. A step-by-step procedure to implement our method is described in Section 4. To illustrate the proposed method, Section 5 conducts a simulation study, and Section 6 presents an empirical example based on the Head Start data. Finally, Section 7 concludes.

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 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

[1] A recent article by Kreiß and Rothe (2023) investigates a similar estimator to ours, and discusses an inference method based on the approach by Armstrong and Kolesár (2018).

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

$T_i = 0$ 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 conduct inference on the causal effect of the treatment, or more specifically, some distributional aspects of the difference of the 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 a 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, for example, Imbens and Lemieux (2008), Cattaneo et al. (2020), Cattaneo and Escanciano (2017), and the 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 3.1. 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 discontinuity (cutoff) point. Throughout the article, we normalize $\bar{x} = 0$ to simplify our 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 $Y_i(1) - Y_i(0)$ is unobservable, we need a tractable representation of τ in terms of quantities that can be estimated by the 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 the difference between the left and right limits of the conditional mean $\mathbb{E}[Y_i|X_i = x]$ at $x = 0$, that is,

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

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 article, 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 eqn. (2)), we propose the local Lasso estimator $\hat{\theta} = (\hat{\alpha}, \hat{\tau}, \hat{\beta}_-, \hat{\beta}_+, \hat{\gamma}')'$ that solves

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

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

where $|\gamma|_1=\sum_{j=1}^{p}|\gamma_j|$ is the ℓ_1-norm of γ, γ_j means the jth element of γ, $K(\cdot)$ is a kernel function, b_n is a bandwidth, and λ_n is a penalty level. Popular choices for $K(\cdot)$ are the uniform and triangular kernels supported on $[-b_n,b_n]$. Based on (3), our point estimator for τ is given by $\hat{\tau}$.

Our preliminary simulation results suggest that the local Lasso estimator for τ is somewhat biased in finite samples. Therefore, our recommendation for point estimation is to employ a post-selection method. Let $\hat{S}=\{j:|\hat{\gamma}_j|\geq\zeta_n\}$ for a nonnegative sequence $\{\zeta_n\}$, and let $Z_{\hat{S},i}$ be a subvector of Z_i selected by $\hat{S}$. Then the local post-Lasso estimator $\bar{\theta}=(\bar{\alpha},\bar{\tau},\bar{\beta}_-,\bar{\beta}_+,\bar{\gamma}_{\hat{S}}')'$ is defined as a solution of the local least square:

$$\min_{\alpha,\tau,\beta_-,\beta_+,t} \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{X_i}{h_n}\right)(Y_i-\alpha-T_i\tau-X_i\beta_- - T_iX_i\beta_+ - Z_{\hat{S},i}'t)^2, \tag{4}$$

where h_n is another bandwidth, and the estimator for τ is given by $\bar{\tau}$.

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 eqn. (2). Therefore, the proposed estimator is a natural generalization of CCFT's when Z_i is high-dimensional. Second, without the kernel weights for localization, our estimator in (3) 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 orders nb_n and nh_n. 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 values of b_n and h_n. Finally, the trimming term ζ_n to obtain the set $\hat{S}$ is introduced to stabilize numerical results (see (7) for our recommended choice based on simulation studies), and theoretically we may set as $\zeta_n=0$.

We now present risk properties of the local Lasso estimators $\hat{\theta}$ and $\bar{\theta}$. Let $G_i=(1,T_i,X_i,T_iX_i,Z_i')'$ be the vector of regressors in (3), $G_{i,j}$ be the jth element of G_i, and $\Theta_n=\arg\min_\theta \mathbb{E}[K(X_i/b_n)(Y_i-G_i'\theta)^2]$ be an argmin set. We impose the following assumptions.

Assumption 1. There exists a sequence $\theta_n^*\in\Theta_n$ that satisfies the following conditions.

1. Let $\epsilon_i=\sqrt{K(X_i/b_n)}(Y_i-G_i'\theta_n^*)$. There exists some $C\in(0,\infty)$ such that

   $\mathbb{E}[|K(X_i/b_n)G_{i,j}\epsilon_i|^m]\leq b_n m!\,C^{m-2}/2,$

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

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

2. Let δ_A be the subvector of δ for an index set A, $S^* = \{j : \theta^*_{n,j} \neq 0\}$, and $(S^*)^c = \{j : \theta^*_{n,j} = 0\}$, where $\theta^*_{n,j}$ is the jth element of θ^*_n. There exists some $\phi^* \in (0, \infty)$ such that

$$\frac{s^*}{|\delta_{S^*}|_1^2} \cdot \min_{\delta : |\delta_{(S^*)^c}|_1 \leq 3|\delta_{S^*}|_1} \delta' \left(\frac{1}{nb_n} \sum_{i=1}^n K\left(\frac{X_i}{b_n}\right) G_i G_i' \right) \delta \geq (\phi^*)^2,$$

with probability approaching one, where $s^* = |S^*|$.

Assumption 2. $K : \mathbb{R} \to \mathbb{R}$ is a bounded and symmetric second-order kernel function that is continuous with a compact support. The bandwidth b_n is a positive sequence satisfying $b_n \to 0$ and $nb_n \to \infty$ as $n \to \infty$.

Assumption 1 defines θ^*_n as an approximate linear predictor or the linear projection on the set of included variables in the index set S^* since $\mathbb{E}[K(X_i/b_n)G_{i,j}\epsilon_i] = 0$ for all $j \in S^*$. This assumption is general enough to cover the setup in CCFT, which assumes p is fixed. In the RDD analyses, it is common to introduce many generated covariates, such as transformations of initial covariates like polynomials, interactions, and various basis functions, without knowing which of them are relevant a priori.[2]

Although it is beyond the scope of this article, the definition of θ^*_n could be modified to be an approximate minimizer which does not belong to Θ_n but gets closer to it at some suitable rate. Since this modification complicates the exposition and derivation as in Kreiß and Rothe (2023) or Belloni et al. (2014), we maintain this exact sparsity assumption. For example, such an extension to approximate sparsity will be useful to allow the situation where the conditional mean satisfies $\mathbb{E}[Y|X, Z] = \mathbb{E}[Y|X, Z_S]$ for some sparse set $\mathcal{S}$ but the conditional mean function $\mathbb{E}[Y|X, Z_S]$ is nonlinear in Z_S so that the exact sparsity assumption typically fails.

Assumption 1(1) contains a set of moment conditions, which are introduced to verify a local-type Bernstein inequality in Lemma A.1. The extra factor b_n is due to the presence of the kernel weight $K(X_i/b_n)$. Assumption 1(2) is a localized version of the compatibility condition. A sufficient condition for this is the so-called restricted eigenvalue condition. More specifically, $\min_{\beta : |\beta|_0 \leq s^*} \frac{1}{nb_n} \sum_{i=1}^n K\left(\frac{X_i}{b_n}\right) \frac{\beta' G_i G_i' \beta}{\beta' \beta}$, where $|\beta|_0$ denotes the cardinality of β, provides a lower bound for the compatibility constant $(\phi^*)^2$ (see, e.g., Sect. 6.13 of Bühlmann and van de Geer, 2011). If CCFT's method is feasible with each subset of covariates of dimension $2s^*$, then the restricted eigenvalue condition is indeed satisfied. Assumption 2 contains requirements on the kernel K and bandwidth b_n, which are standard in the literature of nonparametric methods. Note that since θ^*_n

[2] To motivate the use of generated covariates, it is insightful to note that the asymptotic variance of CCFT's RDD estimator is proportional to $\mathrm{Var}(\{(Y_i(1) - Y_i(0)) - (Z_i(1) - Z_i(0))'\gamma\}^2 | X_i = 0)$ for some γ, which is considered as the (conditional) variance of the linear projection error. Although CCFT considered the linear projection due to the constraint on dimensionality, it is clear that the asymptotic variance is minimized by employing the conditional expectation $\mathbb{E}[Y_i(1) - Y_i(0) | Z_i(1) - Z_i(0), X_i = 0]$ instead of the linear projection $(Z_i(1) - Z_i(0))'\gamma$. Therefore, it is natural to extend CCFT's approach to high-dimensional settings by employing generated covariates or series approximation for the conditional mean.

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

and S^* depend on b_n, Assumption 1 should be satisfied along each sequence $\{b_n\}$. Also, the deviation bounds on the prediction and estimation errors of $\hat{\theta}$ will be given as functions of s^*. While a precise condition on s^* is hard to specify and depends on the sampling distribution, it would be typically of smaller order than $\sqrt{nb_n}$ to satisfy the compatibility condition in Assumption 1(2).

Let $\hat{\gamma}_{\hat{S}}$ be the subvector of $\hat{\gamma}$ selected by $\hat{S}$, $\hat{\theta}_{\hat{S}} = (\hat{\alpha}, \hat{\tau}, \hat{\beta}_-, \hat{\beta}_+, \hat{\gamma}'_{\hat{S}})'$, $Z_{\hat{S},i}$ be the subvector of Z_i selected by $\hat{S}$, $G_{\hat{S},i} = (1, T_i, X_i, T_i X_i, Z'_{\hat{S},i})'$, and $m_n = \lambda_{\min}\left(\frac{1}{nh_n}\sum_{i=1}^n K(X_i/h_n) G_{\hat{S},i} G'_{\hat{S},i}\right)^{-1}$, where $\lambda_{\min}(A)$ means the minimum eigenvalue of a matrix A. Also let $\hat{S}_1 = \{i : 0 < |\hat{\gamma}_i| < \zeta_n\}$, $S_n = \hat{S} \cup \hat{S}_1 = \{i : \hat{\gamma}_i \neq 0\}$, $Z_{\hat{S}_1,i}$ be the subvector of Z_i selected by $\hat{S}_1$, $G_{\hat{S}_1,i} = (1, T_i, X_i, T_i X_i, Z'_{\hat{S}_1,i})'$, and $m_{1n} = \lambda_{\max}\left(\frac{1}{nh_n}\sum_{i=1}^n K(X_i/h_n) G_{\hat{S}_1,i} G'_{\hat{S}_1,i}\right)$, where $\lambda_{\max}(A)$ means the maximum eigenvalue of a matrix A. The ℓ_1-risk properties of the local Lasso and post-Lasso estimators (for the case of $h_n = b_n$) are obtained as follows.

THEOREM 1. *Suppose* $\lambda_n^{-1}\sqrt{\log p/(nb_n)} \to 0$.

(i) *Under Assumptions 1 and 2, it holds*

$$|\hat{\theta} - \theta_n^*|_1 \leq C\frac{\lambda_n s^*}{\phi^{*2}}, \tag{5}$$

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

(ii) *Under Assumptions 1 and 2 and* $h_n = b_n$, *it holds*

$$|\bar{\theta}_{S_n} - \hat{\theta}_{S_n}|_1 \leq (m_n|S_n|\lambda_n) \vee (|\hat{S}_1|\zeta_n) \vee (m_{1n}|\hat{S}_1|\zeta_n^2/\lambda_n). \tag{6}$$

The proof of this theorem is presented in Appendix A.1. This theorem characterizes the risk properties of the estimators $\hat{\theta}$ and $\bar{\theta}$ around θ_n^*. The risk bound of $\hat{\theta}$ depends on the tuning parameter λ_n, the number of nonzero coefficients s^*, and the compatibility constant ϕ^*. Note that the decay rate of λ_n is bounded from below by $\sqrt{\log p/(nb_n)}$. Thus, the risk bound of $\hat{\theta}$ gets worse as the number of covariates p increases or the effective sample size nb_n due to the kernel localization decreases. The result (6) for the post-selection estimator $\bar{\theta}$ shows that the deviation from the original Lasso estimator $\hat{\theta}$ is small when tuning parameter λ_n or the number of selected covariates $|S_n|$ is small, or the minimum eigenvalue of $\frac{1}{nh_n}\sum_{i=1}^n K(X_i/b_n) G_{\hat{S},i} G'_{\hat{S},i}$ is large. If the trimming parameter ζ_n is of similar magnitude of λ_n, then the three terms in the bounds are of similar magnitude. If ζ_n is smaller order of magnitude than λ_n, then the first term will dominate. We suggest some practical choice of the trimming term ζ_n in Section 5 based on our simulation studies.

The above theorem is about estimation of the coefficients of the best linear predictor θ_n^* defined in Assumption 1(1). Additionally suppose that the assumptions of Lemma 1 in CCFT hold true, and that the covariates Z_i are predetermined. Then we can guarantee that the second element of θ_n^* coincides with the average causal

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

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 the 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 polynomials of X_i and $T_i X_i$ even though this article focuses on the local linear model.

Finally, we discuss the choices of the localization bandwidths b_n and h_n 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) or a data-driven choice by Belloni et al. (2014) among others. See Theorem 2 in Section 2.2 for justification of these choices, and Section 4 for details on our practical recommendation.

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 bias-corrected inference method. This subsection argues that we can still apply CCFT's inference procedure for high-dimensional Z_i, provided that CCFT's conditions remain valid for S^* and the subvector θ^*_{n,S^*} of θ^*_n selected by S^*.

In this subsection, we specify the tuning constant ζ_n to obtain $\hat{S} = \{j : |\hat{\gamma}_j| \geq \zeta_n\}$ as

$$\zeta_n = \lambda_n \varrho_n \sum_{j=1}^{p} \mathbb{I}\{\hat{\gamma}_j \neq 0\}, \tag{7}$$

where we set $\varrho_n = \log\log\log n$. This choice of ϱ_n is based on the simulation experiments in Section 5, and it is not shown to be optimal but works reasonably well. Based on the selected covariates by $\hat{S}$ with ζ_n in (7), we apply CCFT's bias-corrected t-ratio to conduct statistical inference on the causal effect parameter τ.

Consider the local post-Lasso estimator $\bar{\tau}$ defined by (4). As shown in Appendix A.2, the dominant term of $\bar{\tau}$ can be characterized as

$$\bar{\tau} = e_2' \left(\frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{X_i}{h_n}\right) G_{1i} G_{1i}' \right)^{-1} \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{X_i}{h_n}\right) G_{1i} \xi_i + o_p((nh_n)^{-1/2}), \tag{8}$$

where $e_2 = (0,1,0,0)'$, $G_{1i} = (1, T_i, X_i, T_i X_i)'$, $\xi_i = T_i \xi_i(1) + (1 - T_i)\xi_i(0)$ with $\xi_i(t) = Y_i(t) - Z_i(t)'\gamma_Y$, and

$$\gamma_Y = \arg \min_{\gamma : \gamma_j = 0 \text{ for } j \notin S^*} \mathbb{E}[(\tilde{Y} - Z'\gamma)^2 | X = 0], \tag{9}$$

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

with $\tilde{Y} = Y(1) - Y(0) - \mathbb{E}[Y(1) - Y(0)|X = 0]$. Indeed, the asymptotic linear form in (8) is analogous to the one derived for the case of fixed-dimensional Z_i in CCFT (except that $\gamma_{Y,j} = 0$ for $j \notin S^*$). Therefore, the pre-asymptotic bias and variance of $\bar{\tau}$ can be analogously written as

$$\begin{aligned}
&\mathcal{B} = \frac{1}{2} e_1' \Gamma_-^{-1} \vartheta_-(q' \mu_-^{(2)}) + \frac{1}{2} e_1' \Gamma_+^{-1} \vartheta_+(q' \mu_+^{(2)}), \\
&\mathcal{V} = (q \otimes P_-' e_1)' \Sigma_-(q \otimes P_-' e_1) + (q \otimes P_+' e_1)' \Sigma_+(q \otimes P_+' e_1),
\end{aligned} \tag{10}$$

respectively, where $e_1 = (1, 0)'$, $R = \begin{bmatrix} 1 & \cdots & 1 \\ X_1/h_n & \cdots & X_n/h_n \end{bmatrix}'$, $q = (1, -\gamma^{*\prime})'$, and

$$\begin{aligned}
&K_- = h_n^{-1} \mathrm{diag}(\mathbb{I}\{X_1 < 0\} K(X_1/h_n), \ldots, \mathbb{I}\{X_n < 0\} K(X_n/h_n)), \\
&K_+ = h_n^{-1} \mathrm{diag}(\mathbb{I}\{X_1 \geq 0\} K(X_1/h_n), \ldots, \mathbb{I}\{X_n \geq 0\} K(X_n/h_n)), \\
&\Gamma_- = n^{-1} R' K_- R, \qquad \Gamma_+ = n^{-1} R' K_+ R, \\
&\vartheta_- = n^{-1} R' K_- [X_1^2/h_n^2, \ldots, X_n^2/h_n^2]', \qquad \vartheta_+ = n^{-1} R' K_+ [X_1^2/h_n^2, \ldots, X_n^2/h_n^2]', \\
&\mu_-^{(2)} = \left[\left. \frac{\partial^2 \mathbb{E}[Y_i(0)|X_i = x]}{\partial x^2} \right|_{x=0}, \left. \frac{\partial^2 \mathbb{E}[Z_i^*(0)'|X_i = x]}{\partial x^2} \right|_{x=0} \right]', \\
&\mu_+^{(2)} = \left[\left. \frac{\partial^2 \mathbb{E}[Y_i(1)|X_i = x]}{\partial x^2} \right|_{x=0}, \left. \frac{\partial^2 \mathbb{E}[Z_i^*(1)'|X_i = x]}{\partial x^2} \right|_{x=0} \right]', \\
&P_- = \sqrt{\frac{h}{n}} \Gamma_-^{-1} R' K_-, \qquad P_+ = \sqrt{\frac{h}{n}} \Gamma_+^{-1} R' K_+, \\
&\Sigma_- = \mathrm{Var}(\mathrm{vec}(\mathbf{Y}(0), \mathbf{Z}^*(0))|\mathbf{X} = 0), \qquad \Sigma_+ = \mathrm{Var}(\mathrm{vec}(\mathbf{Y}(1), \mathbf{Z}^*(1))|\mathbf{X} = 0),
\end{aligned} \tag{11}$$

with $\mathbf{Y}(0) = (Y_1(0), \ldots, Y_n(0))'$, $\mathbf{Y}(1) = (Y_1(1), \ldots, Y_n(1))'$, $\mathbf{X} = (X_1, \ldots, X_n)'$, $\mathbf{Z}^*(0) = (Z_{S^*,1}(0), \ldots, Z_{S^*,n}(0))'$, and $\mathbf{Z}^*(1) = (Z_{S^*,1}(1), \ldots, Z_{S^*,n}(1))'$.

By estimating the unknown components, the pre-asymptotic bias and variance can be estimated as

$$\begin{aligned}
&\bar{\mathcal{B}} = \frac{1}{2} e_1' \Gamma_-^{-1} \vartheta_-(\bar{q}' \bar{\mu}_-^{(2)}) + \frac{1}{2} e_1' \Gamma_+^{-1} \vartheta_+(\bar{q}' \bar{\mu}_+^{(2)}), \\
&\bar{\mathcal{V}} = (\bar{q} \otimes P_-' e_1)' \bar{\Sigma}_-(\bar{q} \otimes P_-' e_1) + (\bar{q} \otimes P_+' e_1)' \bar{\Sigma}_+(\bar{q} \otimes P_+' e_1),
\end{aligned}$$

respectively, where $\bar{q}' = (1, -\bar{\gamma}_{\hat{S}}')'$, $\bar{\mu}_-^{(2)}$ and $\bar{\mu}_+^{(2)}$ are local polynomial estimators of $\mu_-^{(2)}$ and $\mu_+^{(2)}$ for the elements corresponding to $Z_{\hat{S},i}$, respectively, and $\bar{\Sigma}_+$ and $\bar{\Sigma}_-$ are conditional variance estimators of Σ_- and Σ_+, respectively, such as the nearest neighborhood or plug-in estimator in Section 7.9 of CCFT's supplement. Based on these estimators, the t-ratio for τ is obtained as

$$T_\tau = \frac{\bar{\tau} - h_n^2 \bar{\mathcal{B}} - \tau}{\sqrt{(nh_n)^{-1} \bar{\mathcal{V}}}}, \tag{12}$$

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

which is exactly the same as the t-ratio in Theorem 2 of CCFT but using the selected covariates $Z_{\hat{S},i}$. By extending the theoretical developments in CCFT, we obtain the following result.

THEOREM 2. *Suppose Assumptions 1 and 2 hold true. Suppose for all x in a neighborhood of 0 and $t = 0, 1$, the density of X_i is continuous and bounded away from zero, $\mathbb{E}[(Y_i(t), Z_i(t)')|X_i = x]$ is three times continuously differentiable, $\partial^2\mathbb{E}[Z_i(0)|X_i = x]/\partial x^2 = \partial^2\mathbb{E}[Z_i(1)|X_i = x]/\partial x^2$, $\mathbb{E}[Z_i(t)Y_i(t)|X_i = x]$ is continuously differentiable, $\mathrm{Var}((Y_i(t), Z_i(t)')|X_i = x)$ is continuously differentiable and invertible, and $\mathbb{E}[|(Y_i(t), Z_i(t)')|^4|X_i = x]$ is continuous. For all j,k and positive integers m, and some finite C, $\mathbb{E}[|K(X_i/h_n)G_{1i,j}Z_{i,k}|^m] \leq h_n m! C^{m-2}/2$. Finally, assume $h_n \to 0$, $nh_n \to \infty$, $\lambda_n^{-1}\sqrt{\log p/(nb_n)} \to 0$, and $(\sqrt{\log p} + \sqrt{nh_n}h_n^2)(h_n^2 + b_n^2 + \lambda_n + \zeta_n + \zeta_n^2/\lambda_n)s^* \to 0$. Then the conditional MSE expansion of the first term in (8) (denoted by $\bar{\tau}_1$) is obtained as*

$$\mathbb{E}[(\bar{\tau}_1 - \tau)^2|\mathbf{X}] = h_n^4\mathcal{B}^2\{1 + o_p(1)\} + \frac{1}{nh_n}\mathcal{V}. \tag{13}$$

Furthermore, if we additionally assume $\sqrt{\frac{nh_n^5}{\mathcal{V}}}(\bar{\mathcal{B}} - \mathcal{B}) \xrightarrow{p} 0$ and $\frac{\bar{\mathcal{V}}}{\mathcal{V}} \xrightarrow{p} 1$, then

$$T_\tau \xrightarrow{d} N(0, 1). \tag{14}$$

The results in (13) and (14) are analogous to CCFT's Theorems 1 and 2, respectively. This theorem theoretically supports employing the bias correction and bandwidth selection methods by CCFT based on the selected covariates $Z_{\hat{S},i}$ (see Section 4 for our practical recommendation). The assumption $\partial^2\mathbb{E}[Z_i(0)|X_i = x]/\partial x^2 = \partial^2\mathbb{E}[Z_i(1)|X_i = x]/\partial x^2$ is natural for predetermined covariates but may be relaxed by introducing additional regularity conditions (see, Kreiß and Rothe, 2023). Other assumptions except for the last one are also imposed in CCFT. The assumption $(\sqrt{\log p} + \sqrt{nh_n}h_n^2)(h_n^2 + b_n^2 + \lambda_n + \zeta_n + \zeta_n^2/\lambda_n)s^* \to 0$ is used to control the remainder term in (8), and can be considered as a sparsity assumption to restrict the growth rate of s^*.[3] Although the conditions $\sqrt{\frac{nh_n^5}{\mathcal{V}}}(\bar{\mathcal{B}} - \mathcal{B}) \xrightarrow{p} 0$ and $\frac{\bar{\mathcal{V}}}{\mathcal{V}} \xrightarrow{p} 1$ are high level, these are typically satisfied for the bias and variance estimators discussed in CCFT.[4] See Remark 4 for a specific example of the variance estimator $\bar{\mathcal{V}}$.

[3] In the standard Lasso literature, we typically impose $\frac{s^*\log p}{\sqrt{n}} \to 0$ and the minimal penalty level requirement on λ_n. For comparison, consider the following standard setting for tuning parameters, where $b_n \sim h_n \sim n^{-1/5}$, $\lambda_n = a_n\sqrt{\log p/(nb_n)}$ with a slowly diverging a_n, and $\zeta_n = O(\lambda_n)$. Then the condition on s^* reduces to $\frac{s^* a_n \log p}{\sqrt{nh_n}} \to 0$, which is analogous to the standard case.

[4] Under the assumption $\partial^2\mathbb{E}[Z_i(0)|X_i = x]/\partial x^2 = \partial^2\mathbb{E}[Z_i(1)|X_i = x]/\partial x^2$, the components $q'\mu_-^{(2)}$ and $q'\mu_+^{(2)}$ in $\mathcal{B}$ become $\mu_{Y-}^{(2)} = \partial^2\mathbb{E}[Y_i(0)|X_i = x]/\partial x^2\big|_{x=0}$ and $\mu_{Y+}^{(2)} = \partial^2\mathbb{E}[Y_i(1)|X_i = x]/\partial x^2\big|_{x=0}$, respectively. Thus, in this case, the convergence rates of the conventional local polynomial estimators for $\mu_{Y-}^{(2)}$ and $\mu_{Y+}^{(2)}$ guarantee $\sqrt{\frac{nh_n^5}{\mathcal{V}}}(\bar{\mathcal{B}} - \mathcal{B}) \xrightarrow{p} 0$ (see, e.g., Fan and Gijbels, 1992; Ruppert and Wand, 1994).

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

Remark 1. [Efficiency comparison] It should be noted that the asymptotic variance $\mathcal{V}$ in (10) of the estimator $\bar{\tau}$ (or the bias-corrected version $\bar{\tau} - h_n^2\bar{\mathcal{B}}$) takes the same form as the one in CCFT even when the dimension of Z_{S^*} grows as n increases. As investigated in CCFT, we can see that the relative efficiency of $\bar{\tau}$ compared to the conventional RDD estimator without covariates (say, $\hat{\tau}_{\text{unadjusted}}$) is

$$\frac{\mathcal{V}}{\mathcal{V}_{\text{unadjusted}}} = \frac{\sum_{t=0}^{1} \text{Var}(Y_i(t) - Z_i(t)'\gamma_Y | X_i = 0)}{\sum_{t=0}^{1} \text{Var}(Y_i(t)|X_i = 0)}, \tag{15}$$

where $\mathcal{V}_{\text{unadjusted}}$ is the asymptotic variance of $\hat{\tau}_{\text{unadjusted}}$ and γ_Y is defined in (9). Generally, there is no clear ranking for these asymptotic variances. However, letting γ_{Y,S^*} be the s^*-dimensional subvector of γ_Y selected by S^*, in an important special case where

$$\begin{aligned}\gamma_{Y,S^*} = {} & \text{Var}(Z_{S^*,i}(t)|X_i = 0)^{-1}\mathbb{E}[(Z_{S^*,i}(t) \\ & - \mathbb{E}[Z_{S^*,i}(t)|X_i])Y_i(t)|X_i = 0] \quad \text{for } t = 0 \text{ and } 1,\end{aligned} \tag{16}$$

the coefficient vector γ_{Y,S^*} becomes the best linear approximation by $Z_{S^*}(t)$ for each group and thus $\bar{\tau}$ is asymptotically more efficient than $\hat{\tau}_{\text{unadjusted}}$.

The relative efficiency in (15) is also insightful to illustrate the merit of our Lasso approach compared to CCFT. Let $Z = [Z_{S^*} : Z_{S^{*c}}]$ and Z_C be covariates employed to apply the CCFT estimator $\hat{\tau}_{\text{CCFT}}$ (without the Lasso covariates selection). Consider the special case in (16). As far as Z_C contains Z_{S^*}, $\bar{\tau}$ and $\hat{\tau}_{\text{CCFT}}$ achieve the same asymptotic efficiency $\mathcal{V}$. On the other hand, if Z_C does not contain some elements of Z_{S^*}, then under (16), the CCFT estimator $\hat{\tau}_{\text{CCFT}}$ is asymptotically less efficient than the post-Lasso estimator $\bar{\tau}$. Therefore, when researchers are less certain whether all elements of Z_{S^*} are included in Z_C typically due to too many candidates in Z or too small effective sample sizes used for estimation, our Lasso-based approach may be more attractive to achieve asymptotic efficiency $\mathcal{V}$ in broader situations. We emphasize that such an efficiency gain of our estimator $\bar{\tau}$ can be achieved at the cost of the additional sparsity condition in Assumption 1, which is trivially satisfied when the set of active covariates is unknown but fixed.

Remark 2. [Finite sample comparison] Although there is no gain in using $\bar{\tau}$ instead of $\hat{\tau}_{\text{CCFT}}$ in terms of asymptotic efficiency as long as Z_C contains Z_{S^*}, the estimator $\bar{\tau}$ may exhibit better finite sample performance even in such a scenario. To see this point, let $(\hat{\beta}_C', \hat{\gamma}_C')$ be the OLS estimator for the regression of $K^{1/2}Y$ on $K^{1/2}(1, T, X, TX)$ and $K^{1/2}Z_C$ so that $\hat{\tau}_{\text{CCFT}}$ is the second element of $\hat{\beta}_C$. As can be seen from CCFT's supplement, one of the remainder terms of $\sqrt{\frac{nh_n}{\mathcal{V}}}(\hat{\tau}_{\text{CCFT}} - \tau)$ involves a linear combination of the estimation error $\hat{\gamma}_C - (\gamma^{*\prime}, 0')'$. Since the ℓ_2-convergence rate of $\hat{\gamma}_C - (\gamma^{*\prime}, 0')'$ is typically of order $\sqrt{\dim Z_C/n}$, this remainder term is of larger order than $\hat{\gamma}^* - \gamma_Y$, where $(\hat{\beta}^{*\prime}, \hat{\gamma}^{*\prime})$ is the OLS estimator for the regression of $K^{1/2}Y$ on $K^{1/2}(1, T, X, TX)$ and $K^{1/2}Z_{S^*}$. Even though these remainder terms do not appear in the first-order asymptotic distribution, they contribute to the

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

finite sample behaviors of $\bar{\tau}$ and $\hat{\tau}_{\text{CCFT}}$. Another finite sample issue we encounter in our simulation study below is that as the dimension of Z_C increases, the values of the MSE-optimal bandwidth tend to be smaller (due to larger bias estimates but relatively stable variance estimates). Thus the effective sample size used for RDD estimation tends to be smaller so that we observe larger variations in the resulting RDD estimates across simulation draws. Finally, our simulation results in Section 5 (particularly DGPs 2 and 3 with large p) illustrate that the covariate selection approach exhibits smaller standard deviations than CCFT in finite samples even when the asymptotic variances may be equivalent.

Remark 3. [Selection consistency] Although Theorem 2 is our main result on inference of the causal effect τ, it is also possible to derive the consistency of the selection procedure (i.e., $\mathbb{P}\{\hat{S} = S^*\} \to 1$) under some additional β-min type condition (i.e., there exists some $\varepsilon > 0$ such that $|\gamma_j^*| > \lambda_n \varrho_n s^*(1+\varepsilon)$ for each $j \in S^*$). See a working paper version of this article (Arai, Otsu, and Seo, 2021) for more details on the selection consistency.

Remark 4. [Estimation of $\mathcal{V}$] An example of the estimator $\bar{\mathcal{V}}$ for the asymptotic variance $\mathcal{V}$ is the nearest neighborhood estimator, which is employed in CCT, CCFT, and our numerical illustrations. Let

$$\bar{\varepsilon}_{V-,i} = \mathbb{I}\{X_i < 0\}\sqrt{\frac{J}{J+1}}\left(V_i - \frac{1}{J}\sum_{j=1}^{J} V_{\ell_{-,j}(i)}\right),$$

$$\bar{\varepsilon}_{V+,i} = \mathbb{I}\{X_i \geq 0\}\sqrt{\frac{J}{J+1}}\left(V_i - \frac{1}{J}\sum_{j=1}^{J} V_{\ell_{+,j}(i)}\right),$$

for $V \in \{Y, Z_1, \ldots, Z_p\}$ and a fixed positive integer J, where $\ell_{-,j}(i)$ is the index of the jth closest unit to unit i among $\{i : X_i < 0\}$, and $\ell_{+,j}(i)$ is the index of the jth closest unit to unit i among $\{i : X_i \geq 0\}$. The nearest neighborhood estimators of Σ_- and Σ_+ in (11) are defined as

$$\bar{\Sigma}_-^{NN} = \begin{bmatrix} \bar{\Sigma}^{NN}_{YY-} & \bar{\Sigma}^{NN}_{YZ_1-} & \cdots & \bar{\Sigma}^{NN}_{YZ_{|\hat{S}|}-} \\ \bar{\Sigma}^{NN}_{Z_1Y-} & \bar{\Sigma}^{NN}_{Z_1Z_1-} & & \\ \vdots & & \ddots & \\ \bar{\Sigma}^{NN}_{Z_{|\hat{S}|}Y-} & & & \bar{\Sigma}^{NN}_{Z_{|\hat{S}|}Z_{|\hat{S}|}-} \end{bmatrix},$$

$$\bar{\Sigma}_+^{NN} = \begin{bmatrix} \bar{\Sigma}^{NN}_{YY+} & \bar{\Sigma}^{NN}_{YZ_1+} & \cdots & \bar{\Sigma}^{NN}_{YZ_{|\hat{S}|}+} \\ \bar{\Sigma}^{NN}_{Z_1Y+} & \bar{\Sigma}^{NN}_{Z_1Z_1+} & & \\ \vdots & & \ddots & \\ \bar{\Sigma}^{NN}_{Z_{|\hat{S}|}Y+} & & & \bar{\Sigma}^{NN}_{Z_{|\hat{S}|}Z_{|\hat{S}|}+} \end{bmatrix},$$

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

where $\bar{\Sigma}^{NN}_{VW-}$ and $\bar{\Sigma}^{NN}_{VW+}$ are $n \times n$ matrices whose (i,j)th elements are

$$[\bar{\Sigma}^{NN}_{VW-}]_{i,j} = \mathbb{I}\{X_i < 0\}\mathbb{I}\{X_j < 0\}\mathbb{I}\{i = j\}\bar{\varepsilon}_{V-,i}\bar{\varepsilon}_{W-,j},$$
$$[\bar{\Sigma}^{NN}_{VW+}]_{i,j} = \mathbb{I}\{X_i \geq 0\}\mathbb{I}\{X_j \geq 0\}\mathbb{I}\{i = j\}\bar{\varepsilon}_{V+,i}\bar{\varepsilon}_{W+,j},$$

for $i,j = 1,\ldots,n$ and $V, W \in \{Y, Z_1, \ldots, Z_p\}$. Then nearest neighborhood estimator of $\mathcal{V}$ is defined as

$$\bar{\mathcal{V}}^{NN} = (\bar{q} \otimes P'_- e_1)' \bar{\Sigma}^{NN}_- (\bar{q} \otimes P'_- e_1) + (\bar{q} \otimes P'_+ e_1)' \bar{\Sigma}^{NN}_+ (\bar{q} \otimes P'_+ e_1).$$

By adapting the proof of Section S.2.4 in the supplement of CCT, the consistency of this variance estimator is obtained as follows.

PROPOSITION 1. *Suppose that the assumptions of Theorem 2 hold true. Additionally assume that* $\mathrm{Var}((Y_i(t), Z_i(t)')|X_i = x)$ *is Lipschitz continuous in a neighborhood of* 0*, and*

$$\varpi_n^{-1}\{(m_n|S_n|\lambda_n) \vee (|\hat{S}_1|\zeta_n) \vee (m_{1n}|\hat{S}_1|\zeta_n^2/\lambda_n)\} \xrightarrow{p} 0,$$
$$(\varpi_n n h_n)^{-1} s^* \sqrt{\log s^*} \xrightarrow{p} 0, \qquad \zeta_n^{-1} \frac{\lambda_n s^*}{\phi^{*2}} \to 0,$$

where ϖ_n *is defined in (A.11). Then* $\frac{\bar{\mathcal{V}}^{NN}}{\mathcal{V}} \xrightarrow{p} 1$.

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 is set at x, and assume that $W_i(x)$ is nonincreasing in x at $x = 0$. Using the terminology of Angrist, Imbens, and Rubin (1996), unit i is called a complier if i's 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 et al. (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 $x = 0$, that is,

$$\tau_f = \frac{\lim_{x \downarrow 0} \mathbb{E}[Y_i|X_i = x] - \lim_{x \uparrow 0} \mathbb{E}[Y_i|X_i = x]}{\lim_{x \downarrow 0} \mathbb{P}\{W_i = 1|X_i = x\} - \lim_{x \uparrow 0} \mathbb{P}\{W_i = 1|X_i = x\}}. \tag{17}$$

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

In this case, letting $T_i = \mathbb{I}\{X_i \geq 0\}$, the numerator and denominator of (17) can be estimated by the local Lasso estimators $\hat{\vartheta}_1 = (\hat{\alpha}_1, \hat{\tau}_1, \hat{\beta}_{1-}, \hat{\beta}_{1+}, \hat{\gamma}_1')'$ and $\hat{\vartheta}_2 = (\hat{\alpha}_2, \hat{\tau}_2, \hat{\beta}_{2-}, \hat{\beta}_{2+}, \hat{\gamma}_2')'$, which solve

$$\min_{\alpha_1,\tau_1,\beta_{1-},\beta_{1+},\gamma_1} \frac{1}{nb_{1n}} \sum_{i=1}^{n} K\left(\frac{X_i}{b_{1n}}\right)(Y_i - \alpha_1 - T_i\tau_1 - X_i\beta_{1-} - T_iX_i\beta_{1+} - Z_{1i}'\gamma_1)^2 + \lambda_{1n}|\gamma_1|_1,$$

$$\min_{\alpha_2,\tau_2,\beta_{2-},\beta_{2+},\gamma_2} \frac{1}{nb_{2n}} \sum_{i=1}^{n} K\left(\frac{X_i}{b_{2n}}\right)(W_i - \alpha_2 - T_i\tau_2 - X_i\beta_{2-} - T_iX_i\beta_{2+} - Z_{2i}'\gamma_2)^2 + \lambda_{2n}|\gamma_2|_1,$$

respectively, where Z_{1i} and Z_{2i} are predetermined covariates used for the first and second regressions. Then, based on the selected covariates from the above local Lasso regressions (i.e., $\tilde{S}_1 = \{j : |\hat{\gamma}_{1j}| > \zeta_n\}$ with a subvector $Z_{1\tilde{S},i}$ of Z_{1i} selected by $\tilde{S}_1$, and $\tilde{S}_2 = \{j : |\hat{\gamma}_{2j}| > \zeta_n\}$ with a subvector $Z_{2\tilde{S},i}$ of Z_{2i} selected by $\tilde{S}_2$), we implement the local least squares as in CCFT:

$$\min_{\alpha_1,\tau_1,\beta_{1-},\beta_{1+},t_1} \frac{1}{nh_{1n}} \sum_{i=1}^{n} K\left(\frac{X_i}{h_{1n}}\right)(Y_i - \alpha_1 - T_i\tau_1 - X_i\beta_{1-} - T_iX_i\beta_{1+} - Z_{1\tilde{S},i}'t_1)^2,$$

$$\min_{\alpha_2,\tau_2,\beta_{2-},\beta_{2+},t_2} \frac{1}{nh_{2n}} \sum_{i=1}^{n} K\left(\frac{X_i}{h_{2n}}\right)(W_i - \alpha_2 - T_i\tau_2 - X_i\beta_{2-} - T_iX_i\beta_{2+} - Z_{2\tilde{S},i}'t_2)^2.$$

The numerator and denominator of (17) are estimated by the estimated coefficients of τ_1 and τ_2 for the above minimizations, respectively. If the treatment variable W satisfies analogous conditions for Theorem 1 (by replacing Y with W), 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 to regression kink designs. For each unit $i = 1, \ldots, n$, we observe a continuous outcome and explanatory variables denoted by Y_i and X_i, respectively. Regression kink design analysis is concerned with the following nonseparable model

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

where U is an error term (possibly multivariate) and $Q = q(X)$ is a continuous policy variable of interest with known $q(\cdot)$. In general, even though we know the function $q(\cdot)$, we are not able to identify the treatment effect by the policy variable Q. However, it is often the case that the policy function $q(\cdot)$ has some kinks (but is continuous). For instance, suppose Y is duration of unemployment and X is earnings before losing one's job. We are interested in the effect of unemployment benefits $Q = q(X)$. In many unemployment insurance systems (e.g., the one in

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

Austria), $q(\cdot)$ 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 Q. Suppose $q(\cdot)$ is kinked at 0. Otherwise, we redefine X by subtracting the kink point c from X. In particular, Card et al. (2015) showed that a treatment on treated parameter $\tau_k = \int \frac{\partial f(q,x,u)}{\partial q} dF_{U|Q=q,X=x}(u)$ is identified as

$$\tau_k = \frac{\lim_{x\downarrow 0} \frac{d}{dx}\mathbb{E}[Y|X=x] - \lim_{x\uparrow 0} \frac{d}{dx}\mathbb{E}[Y|X=x]}{\lim_{x\downarrow 0} \frac{d}{dx} q(x) - \lim_{x\uparrow 0} \frac{d}{dx} q(x)}. \tag{18}$$

To estimate τ_k, we suggest the local Lasso regression

$$\min_{\alpha,\delta,\beta,\zeta,\eta,\gamma} \frac{1}{nb_n} \sum_{i=1}^{n} K\left(\frac{X_i}{b_n}\right) (Y_i - \alpha - T_i X_i \delta - X_i \beta - T_i X_i^2 \zeta - X_i^2 \eta - Z_i' \gamma)^2 + \lambda_n |\gamma|_1. \tag{19}$$

Let $\hat{\delta}$ be the Lasso estimator of δ computed by (19). Since the denominator $q_0 = \lim_{x\downarrow c} \frac{d}{dx} q(x) - \lim_{x\uparrow c} \frac{d}{dx} q(x)$ in (18) is assumed to be known, our estimator of τ_k is given by $\hat{\tau}_k = \hat{\delta}/q_0$. Under analogous conditions to Theorem 1, we expect that analogous results to the sharp RDD case can be established.

4. RECOMMENDATION FOR IMPLEMENTATION

We recommend the following steps to implement our method in R. The most important difference from CCFT is the covariate selection step (Step 2). This procedure is employed in our numerical studies in the following sections.

1. Without using data on covariates, obtain the MSE-optimal bandwidth $b_n = b_n^{\mathrm{CCT}}$ developed by CCT. This can be implemented by the R package, `Rdrobust` (Calonico, Cattaneo, and Titiunik, 2015b).
2. Using the data on covariates, implement the Lasso estimation in (3) to compute $\hat{\theta}$ by setting $b_n = b_n^{\mathrm{CCT}}$ and choosing λ_n in a data-driven way. This can be implemented by the R package, `hdm` (https://cran.r-project.org/web/packages/hdm/index.html) for example.
3. Based on the Lasso estimates $\hat{\theta}$ obtained in Step 2, select covariates $Z_{\hat{S},i}$ by using the trimming term ζ_n in (7). Then implement CCFT's RDD estimation by (4) and inference procedure including the bias correction and bandwidth selection. For this step, we recommend following the procedures detailed in CCFT's supplement by using `Rdrobust`.

The MSE-optimal bandwidth with the full set of covariates can be too narrow when the number of covariates is large relative to the number of observations and many of the covariates are not relevant. Given that we do not know the exact identities of S^* a priori, we instead employ the MSE-optimal bandwidth by CCT in Step 1, which does not include any covariates. This yields more robust results as shown in the simulation in the next section. Furthermore, we note that CCFT's bandwidth selection in Step 3 is derived under the assumption of fixed-dimensional

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

covariates even though the dimension of the selected covariates $Z_{\hat{S},i}$ may grow. We recommend the above procedure because of its practicality given the available packages. Further theoretical analysis for an optimal choice of the bandwidth parameters and other tuning constants in the present setup is beyond the scope of this article. We note that the choice described in the above procedure works well for all data generating processes (DGPs) considered in our numerical studies.

5. SIMULATION

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

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

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

$W = (W_1, \ldots, W_p)' \sim N(0, \Sigma_W)$ with $\mathbb{E}(W_l^2) = 1$ and $\mathrm{Cov}(W_l, W_{l_1}) = 0.5^{|l-l_1|}$, 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 $\rho = 0.2692$,

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

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

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

and $\mu_3(w) = \sum_{l=0}^{p-1} \pi_l w_l$ with $\pi_l = 0.5^l$.[5] DGP1 is the baseline model considered by CCFT which contains no covariates. DGP2 introduces additional covariates to Model 2 of CCFT.[6] For DGP3, we set $\mu_1(x)$ and $\mu_2(z)$ as in DGP2, and $\mu_3(w) = \sum_{l=1}^{p} \pi_l w_l$ with $\pi_l = 0.2l^{-1}$. DGP3 is considered to see how all the approaches behave when the (approximate) sparsity assumption fails. The sample size is set as $n = 500$ for all cases. The number of covariates p varies from 5 to 500. Our simulations are based on 1,000 Monte Carlo replications.

Table 1 reports the biases and RMSEs of the following four-point estimation methods. For the bandwidth h_n, 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 RDD estimation method without covariates by CCT. The second and third methods use RDD estimation with covariates by CCFT. The fourth method uses RDD estimation with selected covariates. See Section 4 for more details on the fourth method.

Our findings are summarized as follows. First, the RMSEs of the covariate adjusted methods 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 performance for all cases. Both the biases and RMSEs are stable for different values of p for all DGPs. Finally, all methods work equally well for DGP1, where all the additional covariates are irrelevant. However, for DGP2, we find substantial efficiency loss when using the standard method. The results for DGP3 show that the standard and covariate selection approaches perform stably even when the sparsity assumption is not satisfied. However, the covariate adjusted approaches by CCFT perform unstably in terms of the RMSE. Overall, we recommend our covariate selection approach 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 nonzero coefficients of covariates increases. It is interesting to note that the average number of selected covariates decreases as the number of covariates increases.

Table 3 reports the coverage probabilities and interval lengths of the robust confidence intervals for the causal effect. The nominal coverage level is 0.95.

[5] Strictly speaking, DGP2 does not satisfy the sparsity assumption (when p is large) because the coefficients π_l are not exactly equal to zero even for very large l. However, our unreported simulation results show that the covariate selection approach produces identical results to Tables 1–4 when DGP2 is modified to satisfy the exact sparsity assumption, where π_l are set to zero for $l > n^{9/10}$. Because the covariate adjusted approaches are unstable, their results differ but are similar qualitatively.

[6] In our preliminary simulation, we also study the case of $\pi_l = 0.2^l$ but the results are similar to the ones for DGP2.

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

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	0.022	0.062	0.023	0.064	0.022	0.065	0.022	0.062
	10	0.021	0.065	0.022	0.068	0.019	0.070	0.020	0.065
	50	0.021	0.065	0.027	0.101	0.003	0.316	0.021	0.065
	100	0.019	0.063	0.038	0.387	0.004	0.133	0.018	0.063
	250	0.018	0.065	0.026	0.082	0.006	0.102	0.017	0.065
	500	0.021	0.064	0.028	0.072	0.009	0.100	0.020	0.064
DGP2	5	0.013	0.708	−0.004	0.058	0.029	0.081	0.028	0.103
	10	−0.010	0.681	−0.005	0.060	0.026	0.084	0.030	0.179
	50	0.038	0.686	−0.015	0.085	0.042	0.784	0.017	0.190
	100	0.048	0.722	−0.036	0.301	0.045	0.605	0.013	0.201
	250	−0.019	0.668	−0.055	0.347	−0.021	0.957	0.009	0.216
	500	−0.004	0.707	−0.050	0.485	0.015	0.965	0.014	0.217
DGP3	5	0.028	0.180	0.018	0.066	0.029	0.081	0.022	0.131
	10	0.021	0.187	0.015	0.068	0.026	0.084	0.019	0.185
	50	0.029	0.198	0.011	0.110	−0.026	1.491	0.022	0.194
	100	0.036	0.206	−0.053	0.853	−0.024	0.470	0.032	0.203
	250	0.017	0.194	−0.039	0.547	−0.072	0.578	0.013	0.190
	500	0.020	0.196	−0.026	0.549	−0.053	0.627	0.016	0.194

The following points are notable. First, the performance of our covariate selection approach is 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 lengths get shorter as p increases. Third, the coverage of CCT is more stable and better than other methods, especially for DGP2. However, the average interval lengths of CCT are substantially longer than when using the other methods. Fourth, even when the sparsity assumption fails as in DGP3, the standard and covariate selection approaches perform stably but the covariate adjusted approach exhibits erratic behaviors in terms of the coverages and interval lengths. Overall, the covariate selection approach is promising for inference as well since it exhibits robust performances in both coverages and interval lengths for different numbers of covariates and DGPs.

Finally, Table 4 presents the properties of the MSE-optimal bandwidths. We 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

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

TABLE 2. Simulation: Number of selected covariates

	p	Average	Min	Max
DGP1	5	0.376	0	1
	10	0.362	0	1
	50	0.379	0	1
	100	0.350	0	1
	250	0.331	0	1
	500	0.264	0	1
DGP2	5	3.910	3	5
	10	3.063	2	5
	50	2.920	1	5
	100	2.850	1	4
	250	2.683	1	4
	500	2.547	1	4
DGP3	5	1.771	1	3
	10	0.983	0	3
	50	0.730	0	3
	100	0.680	0	3
	250	0.593	0	3
	500	0.479	0	3

increases. This is possibly a main source of the increased RMSEs and under-coverages.

We note that the unstable performance of the covariate adjusted approaches when the number of covariates is large can be because optimal bandwidths are not available for the case where the number of covariates is large. It would be more appropriate to develop optimal bandwidths for the case of high-dimensional covariates and make comparisons, but this is beyond the scope of our article.

6. EMPIRICAL ILLUSTRATION: HEAD START DATA

To illustrate our covariate 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 an RDD with the poverty index as the running variable where the cut-off value is set as $\bar{x} = 59.1984$. Ludwig and Miller (2007) conducted their RDD analysis using no covariates, and CCFT examined the impact of the covariate adjustment. CCFT employed nine pre-intervention covariates from

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

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	0.918	0.270	0.898	0.246	0.901	0.246	0.916	0.270
	10	0.892	0.213	0.855	0.211	0.856	0.205	0.891	0.213
	50	0.915	0.246	0.540	0.152	0.247	0.076	0.917	0.246
	100	0.913	0.246	0.194	0.342	0.186	0.040	0.911	0.246
	250	0.900	0.252	0.108	0.040	0.174	0.048	0.904	0.233
	500	0.905	0.188	0.138	0.020	0.155	0.065	0.903	0.188
DGP2	5	0.926	3.173	0.615	0.219	0.845	0.283	0.870	0.456
	10	0.932	2.243	0.601	0.202	0.781	0.292	0.905	0.724
	50	0.927	2.462	0.397	0.132	0.189	0.082	0.888	0.444
	100	0.912	2.750	0.182	0.073	0.164	0.116	0.898	0.846
	250	0.931	3.578	0.119	0.139	0.175	0.362	0.895	0.745
	500	0.922	3.235	0.130	0.158	0.157	0.255	0.892	0.635
DGP3	5	0.910	1.083	0.807	0.269	0.845	0.283	0.899	0.525
	10	0.924	0.619	0.762	0.253	0.781	0.292	0.912	0.624
	50	0.920	0.666	0.375	0.113	0.232	0.121	0.914	0.673
	100	0.914	0.714	0.306	0.289	0.471	0.835	0.914	0.714
	250	0.916	0.918	0.677	1.547	0.705	1.436	0.920	0.882
	500	0.910	0.821	0.853	1.862	0.791	3.522	0.910	0.810

the U.S. Census, which include total population, the percentage of population, percentages of the black and urban population, and levels and percentages of the population in three age groups (children aged 3–5, children aged 14–17, and adults older than 25). A main finding by CCFT is that their covariate adjusted RDD inference yields shorter confidence intervals while their RDD point estimates remain stable.

An important aspect of the Head Start example is that it is unclear which covariates become useful to improve efficiency mainly due to the lack of economic theories behind the problem. We revisit the empirical exercises of CCFT by applying our covariate selection approach with two extensions. First, we introduce 36 interaction terms in addition to the nine original covariates. Second, we apply the same estimation and inference methods to subsamples of sizes of $n = 500$ and $1,000$ 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 reports the results of our empirical illustration. The last four columns correspond to the same estimation and inference procedures used in the simulation

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

TABLE 4. Simulation: MSE-optimal bandwidths

MSE-optimal bandwidths		w/o Covariates		w/ Covariates		Adaptive	
	p	Mean	SD	Mean	SD	Mean	SD
DGP1	5	0.198	0.046	0.190	0.043	0.197	0.046
	10	0.195	0.046	0.179	0.041	0.195	0.046
	50	0.194	0.045	0.108	0.024	0.193	0.045
	100	0.196	0.047	0.075	0.024	0.196	0.047
	250	0.194	0.045	0.069	0.019	0.194	0.045
	500	0.197	0.043	0.071	0.020	0.196	0.043
DGP2	5	0.182	0.028	0.108	0.011	0.117	0.014
	10	0.183	0.029	0.107	0.011	0.138	0.015
	50	0.183	0.029	0.092	0.013	0.140	0.015
	100	0.183	0.029	0.078	0.021	0.141	0.016
	250	0.181	0.028	0.076	0.022	0.142	0.018
	500	0.183	0.029	0.074	0.021	0.143	0.017
DGP3	5	0.141	0.015	0.108	0.011	0.126	0.016
	10	0.144	0.015	0.107	0.011	0.142	0.015
	50	0.144	0.016	0.092	0.013	0.143	0.016
	100	0.145	0.015	0.104	0.021	0.143	0.016
	250	0.144	0.016	0.132	0.025	0.143	0.016
	500	0.144	0.015	0.150	0.025	0.143	0.016

experiments. The first panel shows the full sample results ($n = 2,799$ and $p/n = 0.016$). The 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 restricting the MSE-optimal bandwidths 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 next two rows in the same panel present the bandwidths (h, b), and effective sample sizes (n_-, n_+) used for the RDD estimation. The effective sample sizes are the numbers of observations of 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 1,000 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 covariate 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

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

TABLE 5. Empirical illustration: Head Start data (45 covariates)

	MSE-optimal bandwidths	w/o Covariates		w/ Covariates	Adaptive
	Estimation methods	Standard	Cov-adjusted	Cov-adjusted	Covariate selection
$n = 2,779$	Point estimate	−2.41	−1.82	−3.64	−2.41
$p/n = 0.016$	h/b unrestricted				
	Robust 95% CI	[−5.46, −0.1]	[−3.85, −0.07]	[−6.05, −1.24]	[−5.46, −0.1]
	CI length change (%)		−29.57	−10.31	0
	Robust p-value	0.042	0.042	0.003	0.042
	$h/b = 1$				
	Robust 95% CI	[−6.41, −1.09]	[−5.44, −1.10]	[−6.55, −1.14]	[−6.41, −1.09]
	CI length change (%)		−18.37	1.64	0
	Robust p-value	0.006	0.003	0.005	0.006
	h, b	6.81, 10.73	6.81, 10.73	3.02, 5.40	6.81, 10.73
	n_-, n_+	234, 180	234, 180	96, 84	234, 180
	Selected covariates				None
$n = 1,000$	Point estimate	−1.68	−2.40	−3.44	−1.48
$p/n = 0.045$	h/b unrestricted				
	Robust 95% CI	[−5.45, 1.75]	[−5.44, −0.29]	[−6.14, −1.34]	[−5.08, 1.79]
	CI length change (%)		−28.38	−33.23	−4.49
	Robust p-value	0.314	0.029	0.002	0.347
	$h/b = 1$				
	Robust 95% CI	[−8.26, 0.22]	[−6.84, −0.47]	[−5.46, .0.15]	[−7.94, 0.27]
	CI length change (%)		−24.95	−33.87	−3.28
	Robust p-value	0.063	0.025	0.064	0.070
	h, b	6.52, 10.23	6.52, 10.23	3.47, 6.00	5.26, 8.07
	n_-, n_+	74, 77	79, 79	40, 46	79, 79
	Selected covariates				% of adult population
$n = 500$	Point estimate	−2.35	−4.23	−5.68	−2.22
$p/n = 0.090$	h/b unrestricted				
	Robust 95% CI	[−7.25, 2.48]	[−7.44, −2.06]	[−9.64, −4.63]	[−6.93, 2.25]
	CI length change (%)		−44.77	−48.58	−5.74
	Robust p-value	0.337	0.001	0.000	0.317
	$h/b = 1$				
	Robust 95% CI	[−10.22, 1.42]	[−9.03, −3.43]	[−9.62, −4.29]	[−10, 1.07]
	CI length change (%)		−51.93	−54.16	−4.94
	Robust p-value	0.139	0.000	0.000	0.110
	h, b	6.37, 9.16	6.37, 9.16	4.96, 7.49	4.31, 6.82
	n_-, n_+	60, 56	61, 56	49, 47	61, 56
	Selected covariates				% of adult population

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

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 covariate selection approach and its mild contribution to shorten the confidence intervals. Table 6 presents the corresponding results for the case of the nine covariates as considered by CCFT. The results are essentially similar to those of Table 5, although the influence of the covariates is less dramatic.

To better understand the proposed method in this article, we investigate the behaviors of three approaches through the Head Start example. The approach denoted by "Cov-adjusted 1" corresponds to the covariate adjusted approach using the CCT MSE-optimal bandwidths, and the one denoted by "Cov-adjusted 2" to the one using the CCFT MSE-optimal bandwidths. We denote the approach proposed in this article by "Covariate selection." We construct 50 subsamples of k observations out of the Head Start data where k is set to 1,000 and 500. We estimate RDD treatment effects based on the subsamples using nine or forty-five covariates where the nine covariates are those considered in CCFT and the forty-five covariates are those considered in Table 5. Figures 1–4 show boxplots of the results on estimation (left panel) and inference (right panel).[7] The top and bottom of the boxes are the third and first quartiles, and the top and bottom bars show the maximum and minimum values less than (the third quartiles + 1.5 times the interquartile range) and greater than (the first quartile—1.5 times the interquartile range), respectively. The left panels in Figures 1–4 show the differences in RDD treatment effects between the three approaches and the standard one where the standard one is based on CCT. The right panel in each figure shows the corresponding differences in the confidence interval lengths.[8]

First, the left panels of Figures 1 and 2 show the point estimates by the covariate adjusted approaches deviate from the standard one by a large extent when the number of covariates is large, and the differences get larger as the sample size becomes small. We observe a similar tendency for the case of nine covariates, while differences are less dramatic. Second, we note that the point estimates based on the standard and covariate selection approaches are very close, and they are stable. Third, we observe that the covariate adjusted approaches tend to produce large decreases in the interval length, which can reflect under-coverages as observed in Table 3. Fourth, the covariate selection approach leads to reductions of around 5%, which are still nonnegligible. Fifth, the results for the nine covariates given in Figures 3 and 4 are similar to those for the forty-five covariates, although the magnitude of changes for the nine covariates is not as extreme as that for the forty-five covariates. These results show the usefulness of our covariate selection approach not only for the situation where the number of covariates is large but also

[7]The results shown in Figures 1–3 are based on the bandwidths with h/b unrestricted, and those with the bandwidths with $h/b = 1$ are qualitatively similar.

[8]The covariate selection approach produced an identical result to the standard one about 20 times out of 50 for all setups. The boxplots are drawn based on the nonidentical results.

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

TABLE 6. Empirical illustration: Head Start data (nine covariates)

	MSE-optimal bandwidths	w/o Covariates		w/ Covariates	Adaptive
	Estimation methods	Standard	Cov-adjusted	Cov-adjusted	Covariate selection
$n = 2,779$	Point estimate	−2.41	−2.51	2.47	-2.41
$p/n = 0.003$	h/b unrestricted				
	Robust 95% CI	[−5.46, −0.1]	[−5.37, −0.45]	[−5.21, −0.37]	[−5.46, −0.1]
	CI length change (%)		−8.24	−9.74	0
	Robust p-value	0.042	0.021	0.024	0.042
	$h/b = 1$				
	Robust 95% CI	[−6.41, −1.09]	[−6.63, −1.46]	[−6.54, −1.39]	[−6.41, −1.09]
	CI length change (%)		−2.87	−3.23	0
	Robust p-value	0.006	0.002	0.003	0.006
	h, b	6.81, 10.73	6.81, 10.73	6.98, 11.64	6.81, 10.73
	n_-, n_+	234, 180	234, 180	240, 184	234, 180
	Selected covariates				None
$n = 1,000$	Point estimate	−1.68	−2.05	−1.53	−1.48
$p/n = 0.009$	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 (%)		−7.12	−20.0	−4.49
	Robust p-value	0.314	0.141	0.237	0.347
	$h/b = 1$				
	Robust 95% CI	[−8.26, 0.22]	[−8.26, −0.63]	[−5.93, −0.88]	[−7.94, 0.27]
	CI length change (%)		−10.20	−19.75	−3.28
	Robust p-value	0.063	0.022	0.145	0.070
	h, b	6.52, 10.23	6.52, 10.23	9.14, 13.93	9.26, 8.07
	n_-, n_+	74, 77	79, 79	121, 98	79, 79
	Selected covariates				% of adult population
$n = 500$	Point estimate	−2.35	−2.54	−2.39	−2.22
$p/n = 0.018$	h/b unrestricted				
	Robust 95% CI	[−7.25, 2.48]	[−7.25, 1.28]	[−6.75, 1.6]	[−6.93, 2.25]
	CI length change (%)		−12.36	−14.24	−5.74
	Robust p-value	0.337	0.170	0.227	0.317
	$h/b = 1$				
	Robust 95% CI	[−10.22, 1.42]	[−9.97, −0.03]	[−9.74, 0.03]	[−10, 1.07]
	CI length change (%)		−14.58	−16.10	−4.94
	Robust p-value	0.139	0.049	0.051	0.110
	h, b	6.37, 9.16	6.37, 9.16	6.58, 9.88	4.31, 6.82
	n_-, n_+	60, 56	61, 56	63, 58	61, 56
	Selected covariates				% of adult population

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

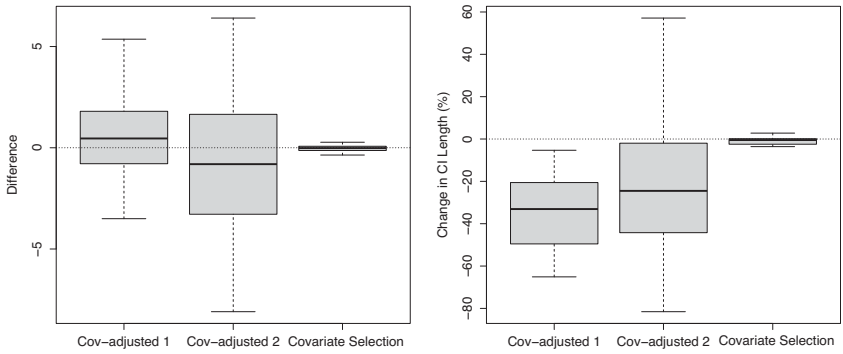

FIGURE 1. Comparison of three approaches ($n = 1,000$, 45 covariates) for point estimates (left panel) and CI lengths (right panel).

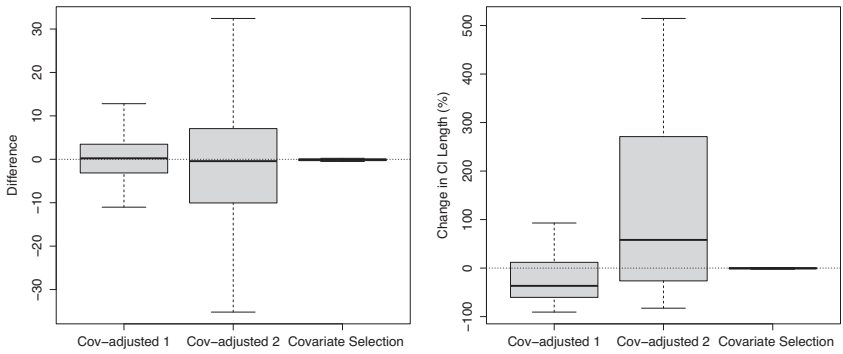

FIGURE 2. Comparison of three approaches ($n = 500$, 45 covariates) for point estimates (left panel) and CI lengths (right panel).

for that where the number of covariates is relatively small, when the sample size is moderate.

We often encounter situations where a number of potentially useful covariates are available. Furthermore, it is common to employ transformations of covariates such as interaction and quadratic terms. However, we are typically uncertain about which covariates contribute to improving efficiency possibly due to the lack of economic theory. RDD analysis is local in nature, and the number of covariates relative to the effective sample size can be large. We have seen that the covariate adjusted approaches can become misleading in several examples. The steady performance of our covariate selection approach under various circumstances is noteworthy, and it can provide an essential second opinion for the standard and covariate adjusted approaches.

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

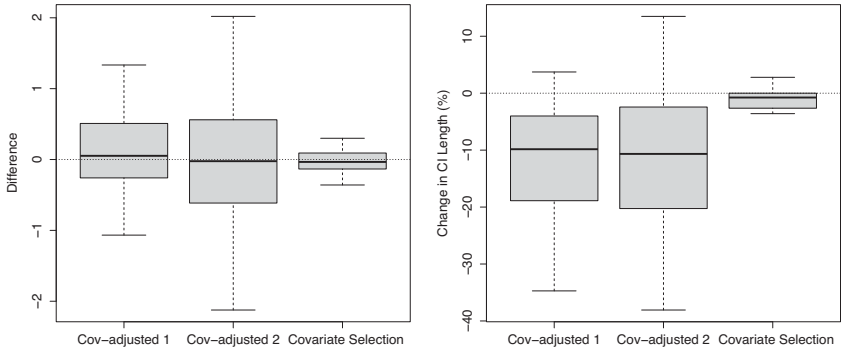

FIGURE 3. Comparison of three approaches ($n = 1,000$, 9 covariates) for point estimates (left panel) and CI lengths (right panel).

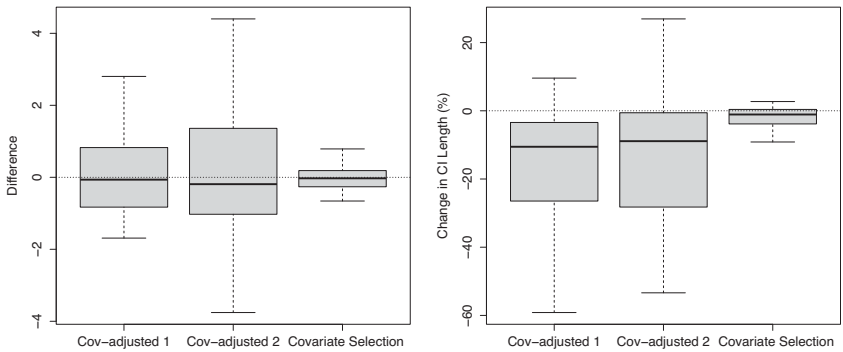

FIGURE 4. Comparison of three approaches ($n = 500$, nine covariates) for point estimates (left panel) and CI lengths (right panel).

7. CONCLUSION

In this article, we propose estimation and inference methods for regression discontinuity analysis with possibly high-dimensional covariates. The point estimator, obtained by combining a kernel-based localization and ℓ_1-penalization to deal with high-dimensional covariates, exhibits reasonable theoretical and finite sample properties including an efficiency gain compared with the conventional covariate adjusted estimator under some relevant scenarios at the cost of an additional sparsity condition. Furthermore, we show that the inference methods by Calonico et al. (2019) can be applied to post-selection RDD models by our estimation method, and illustrate their robust finite sample properties by simulation and empirical examples. As a direction of future research, it is important to develop optimal choices for the bandwidth and tuning parameters.

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

A. MATHEMATICAL APPENDIX

A.1. Proof of Theorem 1

We use the following modification of Bernstein's inequality.

LEMMA A.1. *Under Assumptions 1 and 2, it holds*

$$\mathbb{E}\left[\max_{1\le j\le p}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{K\left(\frac{X_i}{b_n}\right)G_{i,j}\epsilon_i-\mathbb{E}\left[K\left(\frac{X_i}{b_n}\right)G_{i,j}\epsilon_i\right]\right\}\right|^m\right]\le 2C^m b_n^{m/2}\log^{m/2}p,$$

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

Proof of Lemma A.1. It follows from Bernstein's inequality (e.g., Lemma 14.12 in Bühlmann and van de Geer, 2011) that

$$\mathbb{E}\left[\max_{1\le j\le p}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{K\left(\frac{X_i}{b_n}\right)G_{i,j}\epsilon_i-\mathbb{E}\left[K\left(\frac{X_i}{b_n}\right)G_{i,j}\epsilon_i\right]\right\}\right|^m\right]\le 2C^m b_n^{m/2}\log^{m/2}p,$$

for $m\le 1+\log p$. □

To make the proof more accessible and comparable to the standard Lasso, we begin with $\tilde{\theta}$ as the solution of

$$\min_{\theta}\frac{1}{nb_n}\sum_{i=1}^{n}K\left(\frac{X_i}{b_n}\right)(Y_i-\alpha-T_i\tau-X_i\beta_--T_iX_i\beta_+-Z_i'\gamma)^2+\lambda_n|\theta|_1. \tag{A.1}$$

A.1.1. *Deviation Bounds for* $\tilde{\theta}$. Under Assumptions 1 and 2, Lemma A.1 implies

$$\mathbb{E}\left[\max_{1\le j\le p}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}K\left(\frac{X_i}{b_n}\right)G_{ij}\epsilon_i\right|\right]\le 2b_n^{1/2}\log^{1/2}p.$$

Thus, we have

$$\mathbb{P}\{\mathcal{A}_n\}:=\mathbb{P}\left\{\frac{4}{nb_n}\left|\sum_{i=1}^{n}K\left(\frac{X_i}{b_n}\right)G_i\epsilon_i\right|_\infty\le\lambda_n\right\}\to 1, \tag{A.2}$$

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

Let $\mathbf{Y}=(Y_1K_1^{1/2},\ldots,Y_nK_n^{1/2})'$, $\mathbf{e}=(\epsilon_1K_1^{1/2},\ldots,\epsilon_nK_n^{1/2})'$, and $\mathbf{G}=(G_1K_1^{1/2},\ldots,G_nK_n^{1/2})'$. Since $\tilde{\theta}$ is a minimizer, we have

$$\frac{1}{nb_n}|\mathbf{Y}-\mathbf{G}\tilde{\theta}|_2^2+\lambda_n|\tilde{\theta}|_1\le\frac{1}{nb_n}|\mathbf{Y}-\mathbf{G}\theta^*|_2^2+\lambda_n|\theta^*|_1.$$

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

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

$$\begin{aligned}\frac{2}{nb_n}|\mathbf{G}(\tilde{\theta}-\theta^*)|_2^2 &\leq \frac{4}{nb_n}\mathbf{e}'\mathbf{G}(\tilde{\theta}-\theta^*) + 2\lambda_n|\theta^*|_1 - 2\lambda_n|\tilde{\theta}|_1 \\ &\leq \frac{4}{nb_n}|\mathbf{e}'\mathbf{G}|_\infty|\tilde{\theta}-\theta^*|_1 + 2\lambda_n(|\theta^*|_1 - |\tilde{\theta}|_1) \\ &\leq 3\lambda_n|\tilde{\theta}_{S^*} - \theta^*_{S^*}|_1 - \lambda_n|\tilde{\theta}_{S_c^*}|_1, \end{aligned} \tag{A.3}$$

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:

$$\begin{aligned}&|\tilde{\theta}-\theta^*|_1 = |\tilde{\theta}_{S_*} - \theta^*_{S_*}|_1 + |\tilde{\theta}_{S_c^*}|_1, \\ &|\theta^*|_1 - |\tilde{\theta}|_1 = |\theta^*_{S^*}|_1 - |\tilde{\theta}_{S^*}|_1 - |\tilde{\theta}_{S_c^*}|_1 \leq |\tilde{\theta}_{S^*} - \theta^*_{S^*}|_1 - |\tilde{\theta}_{S_c^*}|_1,\end{aligned} \tag{A.4}$$

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

$$\frac{2}{nb_n}|\mathbf{G}(\tilde{\theta}-\theta^*)|_2^2 + \lambda_n|\tilde{\theta}-\theta^*|_1 \leq 4\lambda_n|\tilde{\theta}_{S^*} - \theta^*_{S^*}|_1, \tag{A.5}$$

by using (A.4).

Now, Assumption 1 implies that

$$4\lambda_n|\tilde{\theta}_{S^*} - \theta^*_{S^*}|_1 \leq 4\lambda_n\sqrt{\frac{s^*}{nb_n\phi^{*2}}}|\mathbf{G}(\tilde{\theta}-\theta^*)|_2 \leq \frac{1}{nb_n}|\mathbf{G}(\tilde{\theta}-\theta^*)|_2^2 + 4\lambda_n^2\frac{s^*}{\phi^{*2}},$$

with probability approaching one, where we note that $2ab \leq a^2 + b^2$ for the second inequality. Combining this with (A.5), we have

$$\frac{1}{nb_n}|\mathbf{G}(\tilde{\theta}-\theta^*)|_2^2 + \lambda_n|\tilde{\theta}-\theta^*|_1 \leq 4\lambda_n^2\frac{s^*}{\phi^{*2}},$$

with probability approaching one.

Also note that this result implies that the number of nonzero elements in $\tilde{\theta}$ is $O(s^*)$ by the argument given in Bickel, Ritov, and Tsybakov (2009, eqn. (B.3)).

A.1.2. *Proof of (i).* Let $G_i = (G_{1i}', G_{2i}')'$, where $G_{2i} = K_i^{1/2}Z_i$ and G_{1i} is the other in the partition of G_i. Define $\mathbf{G}_1 = (G_{11}, \ldots, G_{1n})'$ and $\mathbf{G}_2 = (G_{21}, \ldots, G_{2n})'$ and partition θ accordingly. The compatibility condition for $\mathbf{G}_1$ 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 (A.2) still holds. Since $\hat{\theta}$ is a minimizer, we have

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

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

By plugging $\mathbf{Y} = \mathbf{G}\theta^* + \mathbf{e}$ into the above, we have

$$\begin{aligned}\frac{2}{nb_n}|\mathbf{G}(\hat{\theta}-\theta^*)|_2^2 &\le \frac{4}{nb_n}\mathbf{e}'\mathbf{G}(\hat{\theta}-\theta^*) + 2\lambda_n|\theta_2^*|_1 - 2\lambda_n|\hat{\theta}_2|_1 \\ &\le \frac{4}{nb_n}|\mathbf{e}'\mathbf{G}|_\infty|\hat{\theta}-\theta^*|_1 + 2\lambda_n(|\theta_2^*|_1 - |\hat{\theta}_2|_1) \\ &\le \lambda_n|\hat{\theta}_1 - \theta_1^*|_1 + 3\lambda_n|\hat{\theta}_{2,S^*} - \theta_{2,S^*}^*|_1 - \lambda_n|\hat{\theta}_{2,S_c^*}|_1, \end{aligned} \tag{A.6}$$

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:

$$|\hat{\theta}-\theta^*|_1 = |\hat{\theta}_{S^*} - \theta_{S^*}^*|_1 + |\hat{\theta}_{S_c^*}|_1, \tag{A.7}$$

$$|\theta^*|_1 - |\hat{\theta}|_1 = |\theta_{S^*}^*|_1 - |\hat{\theta}_{S^*}|_1 - |\hat{\theta}_{S_c^*}|_1 \le |\hat{\theta}_{S^*} - \theta_{S^*}^*|_1 - |\hat{\theta}_{S_c^*}|_1,$$

due to the triangle inequality. Thus, (A.6) implies $3|\hat{\theta}_{S_*} - \theta_{S_*}^*|_1 \ge |\hat{\theta}_{S_*^c}|_1$ and

$$\frac{2}{nb_n}|\mathbf{G}(\hat{\theta}-\theta^*)|_2^2 + \lambda_n|\hat{\theta}_2 - \theta_2^*|_1 \le \lambda_n|\hat{\theta}_1 - \theta_1^*|_1 + 4\lambda_n|\hat{\theta}_{2,S^*} - \theta_{2,S^*}^*|_1, \tag{A.8}$$

by using (A.7).

Now Assumption 1 implies

$$4\lambda_n|\hat{\theta}_{2,S^*} - \theta_{2,S^*}^*|_1 \le 4\lambda_n\sqrt{\frac{s^*}{nb_n\phi^{*2}}}|\mathbf{G}_2(\hat{\theta}_2 - \theta_2^*)|_2 \le \frac{1}{nb_n}|\mathbf{G}_2(\hat{\theta}_2 - \theta_2^*)|_2^2 + 4\lambda_n^2\frac{s^*}{\phi^{*2}},$$

$$\lambda_n|\hat{\theta}_1 - \theta_1^*|_1 \le \lambda_n\sqrt{\frac{s^*}{nb_n\phi^{*2}}}|\mathbf{G}_1(\hat{\theta}_1 - \theta_1^*)|_2 \le \frac{1}{nb_n}|\mathbf{G}_1(\hat{\theta}_1 - \theta_1^*)|_2^2 + \lambda_n^2\frac{s^*}{\phi^{*2}},$$

with probability approaching one. Combining these inequalities with (A.8), we have

$$\frac{1}{nb_n}|\mathbf{G}(\hat{\theta}-\theta^*)|_2^2 + \lambda_n|\hat{\theta}_2 - \theta_2^*|_1 \le 5\lambda_n^2\frac{s^*}{\phi^{*2}},$$

which implies

$$|\hat{\theta}_2 - \theta_2^*|_1 \le 5\lambda_n\frac{s^*}{\phi^{*2}}. \tag{A.9}$$

Turning to the finite-dimensional component $\hat{\theta}_1$, note that

$$\begin{aligned}\hat{\theta}_1 - \theta_1^* &= \left(\frac{1}{nb_n}\mathbf{G}_1'\mathbf{G}_1\right)^{-1}\left(\frac{1}{nb_n}\mathbf{G}_1'e - \frac{1}{nb_n}\mathbf{G}_1'\mathbf{G}_2(\hat{\theta}_2 - \theta_2^*)\right) \\ &= O_p((nb_n)^{-1/2}) + O_p(|\hat{\theta}_2 - \theta_2^*|_1).\end{aligned}$$

Combining this with (A.9) yields the conclusion in (5).

A.1.3. *Proof of (ii).* Note that we assume $h_n = b_n$. Let $\mathbf{G}_{\hat{S}} = (G_{\hat{S},1}, \ldots, G_{\hat{S},n})'$. Since $\bar{\theta}_{\hat{S}}$ is the OLS estimator that minimizes the sum of the squared residuals in the regression of $\mathbf{Y}$ on $\mathbf{G}_{\hat{S}}$, it holds

$$|\mathbf{Y} - \mathbf{G}_{\hat{S}}\bar{\theta}_{\hat{S}}|_2^2 \le |\mathbf{Y} - \mathbf{G}_{\hat{S}}\hat{\theta}_{\hat{S}}|_2^2.$$

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

Let $\hat{S}_1 = \{i : 0 < |\hat{\theta}_i| < \zeta_n\}$ and $S_n = \hat{S} \cup \hat{S}_1$. Note also that

$$|\mathbf{Y} - \mathbf{G}_{S_n}\bar{\theta}_{S_n}|_2^2 = |\mathbf{Y} - \mathbf{G}_{\hat{S}}\bar{\theta}_{\hat{S}}|_2^2 \leq |\mathbf{Y} - \mathbf{G}_{\hat{S}}\hat{\theta}_{\hat{S}}|_2^2.$$

Then, we get

$$\begin{aligned}\frac{1}{nh_n}|\mathbf{G}_{S_n}(\bar{\theta}_{S_n} - \hat{\theta}_{S_n})|_2^2 &\leq \frac{1}{nh_n}|\mathbf{G}_{\hat{S}_1}\hat{\theta}_{\hat{S}_1}|_2^2 + \frac{2}{nh_n}|\hat{\mathbf{e}}'\mathbf{G}_{\hat{S}_1}\hat{\theta}_{\hat{S}_1}| + \frac{2}{nh_n}|\hat{\mathbf{e}}'\mathbf{G}_{S_n}(\bar{\theta}_{S_n} - \hat{\theta}_{S_n})| \\ &\leq \frac{1}{nh_n}|\mathbf{G}_{\hat{S}_1}\hat{\theta}_{\hat{S}_1}|_2^2 + \lambda_n|\hat{\theta}_{\hat{S}_1}|_1 + \lambda_n|\bar{\theta}_{S_n} - \hat{\theta}_{S_n}|_1, \qquad \textbf{(A.10)}\end{aligned}$$

due to the Hölder inequality and the Karush–Kuhn–Tucker condition for $\hat{\mathbf{e}}$. On the other hand,

$$\lambda_{\min}\left(\frac{1}{nh_n}\mathbf{G}_S'\mathbf{G}_S\right)|\bar{\theta}_S - \hat{\theta}_S|_2^2 \leq \frac{1}{nh_n}|\mathbf{G}_S(\bar{\theta}_S - \hat{\theta}_S)|_2^2.$$

Also note that $|a|_1 \leq \sqrt{s}|a|_2$ for an s-dimensional vector a. Then, we consider the three cases. First, let the last term in (A.10) be the biggest among the three terms. Then, the bound becomes $\lambda_{\min}\left(\frac{1}{nh_n}\mathbf{G}_{S_n}'\mathbf{G}_{S_n}\right)^{-1}|S_n|\lambda_n$. If the second is the biggest, then $|\hat{S}_1|\zeta_n \geq |\hat{\theta}_{\hat{S}_1}|_1 \geq |\bar{\theta}_{S_n} - \hat{\theta}_{S_n}|_1$. Finally, if the first term is the biggest, then $\lambda_n|\bar{\theta}_{S_n} - \hat{\theta}_{S_n}|_1 \leq \frac{1}{nh_n}|\mathbf{G}_{\hat{S}_1}\hat{\theta}_{\hat{S}_1}|_2^2 \leq \lambda_{\max}\left(\frac{1}{nh_n}\mathbf{G}_{\hat{S}_1}'\mathbf{G}_{\hat{S}_1}\right)|\hat{S}_1|\zeta_n^2$. Therefore, we get the desired bound under $h_n = b_n$ and Assumption 1.

A.2. Proof of Theorem 2

First, we show the asymptotic expansion in (8). Let $\bar{\gamma}$ be a coefficient vector of Z_i, where the elements correspond to $Z_{\hat{S},i}$ is $\bar{\gamma}_{\hat{S}}$ in (4), and other elements are zero. Observe that

$$\begin{aligned}\bar{\tau} &= e_2'\left(\sum_{i=1}^n K\left(\frac{X_i}{h_n}\right)G_{1i}G_{1i}'\right)^{-1}\sum_{i=1}^n K\left(\frac{X_i}{h_n}\right)G_{1i}(Y_i - Z_i'\bar{\gamma}) \\ &= \hat{\tau}_\xi - e_2'\left(\sum_{i=1}^n K\left(\frac{X_i}{h_n}\right)G_{1i}G_{1i}'\right)^{-1}\sum_{i=1}^n K\left(\frac{X_i}{h_n}\right)G_{1i}Z_i'(\bar{\gamma} - \gamma_Y),\end{aligned}$$

where $\hat{\tau}_\xi = e_2'\left(\sum_{i=1}^n K(X_i/h_n)G_{1i}G_{1i}'\right)^{-1}\sum_{i=1}^n K(X_i/h_n)G_{1i}\xi_i$. Thus it is sufficient for (8) to show that the second term is of order $o_p((nh_n)^{-1/2})$.

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

Note that

$$\begin{aligned}
&\left\| e_2' \left(\frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{X_i}{h_n}\right) G_{1i} G_{1i}' \right)^{-1} \frac{1}{\sqrt{nh_n}} \sum_{i=1}^{n} K\left(\frac{X_i}{h_n}\right) G_{1i} Z_i' \right\|_\infty \\
&\quad \leq \lambda_{\min} \left(\frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{X_i}{h_n}\right) G_{1i} G_{1i}' \right)^{-1} \\
&\qquad \times \left\| \frac{1}{\sqrt{nh_n}} \sum_{i=1}^{n} \left\{ K\left(\frac{X_i}{h_n}\right) G_{1i} Z_i' - \mathbb{E}\left[K\left(\frac{X_i}{h_n}\right) G_{1i} Z_i' \right] \right\} \right\|_\infty + O_p(\sqrt{nh_n} h_n^2) \\
&\quad = O_p(\sqrt{\log p}) + O_p(\sqrt{nh_n} h_n^2),
\end{aligned}$$

where the inequality follows from the fact that $\left(\sum_{i=1}^n K(X_i/h_n) G_{1i} G_{1i}'\right)^{-1} \sum_{i=1}^n K(X_i/h_n) G_{1i} Z_i'$ is a vector of the local linear RDD estimator for the outcome Z_i and its bias is of order $O(h_n^2)$ from Lemma SA-2 of CCFT, and the equality follows from the local Bernstein's inequality in Lemma A.1.

Turing to bound $\bar{\gamma} - \gamma_Y$, recall that $\theta_{S^*}^* = \arg\min_{\theta_{S^*}} \mathbb{E}[K(X_i/b_n)(Y_i - G_{S^*i}' \theta_{S^*})^2/b_n]$ while $\gamma_Y = \arg\min_\gamma \mathbb{E}[(\tilde{Y} - Z_{S^*}'\gamma)^2 | X = 0]$, where $\tilde{Y} = Y(1) - Y(0) - \mathbb{E}[Y(1) - Y(0) | X = 0]$ and $Z_{S^*} = Z_{S^*}(1) - Z_{S^*}(0)$. Due to van der Vaart and Wellner (1996, Thm. 3.4.1), we can bound the contrast $|\gamma_Y - \gamma^*|$ by the difference in the criteria, which is of order $O(s^* b_n^2)$ by the standard bias calculation. For the same reasoning, we obtain $|\bar{\gamma}(b_n) - \bar{\gamma}(h_n)| = O_p(s^*(h_n^2 + b_n^2))$, where we highlight by $\bar{\gamma}(b_n)$ the dependence on the bandwidth used to compute $\bar{\gamma}$.

Since $|\bar{\gamma}(b_n) - \gamma^*|_1 = O_p(\lambda_n s^*)$ by Theorem 1 and thus $|\gamma_Y - \bar{\gamma}(h_n)|_1 = O_p(s^*(h_n^2 + b_n^2) + \lambda_n s^*)$ due to the preceding derivation, Hölder's inequality and the assumption $(\sqrt{\log p} + \sqrt{nh_n} h_n^2)(s^*(h_n^2 + b_n^2) + \lambda_n s^*) \to 0$ guarantee (8).

Second, we note that $\hat{\tau}_\xi$ is the conventional local linear RDD estimator without covariates for the outcome variable ξ_i. Thus, the proof of CCFT's Theorem 1 is directly applicable and the MSE expansion in (13) follows.

Finally, we show (14). By (8), we have

$$T_\tau = \sqrt{\frac{\mathcal{V}}{\bar{\mathcal{V}}}} \sqrt{\frac{nh_n}{\mathcal{V}}} (\hat{\tau}_\xi - h_n^2 \mathcal{B} - \tau) - \sqrt{\frac{\mathcal{V}}{\bar{\mathcal{V}}}} \sqrt{\frac{nh_n}{\mathcal{V}}} h_n^2 (\bar{\mathcal{B}} - \mathcal{B}).$$

Since $\hat{\tau}_\xi$ is the conventional local linear RDD estimator without covariates for the outcome variable ξ_i, Lemma SA-10 of CCFT yields $\sqrt{\frac{nh_n}{\mathcal{V}}} (\hat{\tau}_\xi - h_n^2 \mathcal{B} - \tau) \xrightarrow{d} N(0,1)$. Therefore, the assumptions $\sqrt{\frac{nh_n^5}{\mathcal{V}}} (\bar{\mathcal{B}} - \mathcal{B}) \xrightarrow{p} 0$ and $\frac{\bar{\mathcal{V}}}{\mathcal{V}} \xrightarrow{p} 1$ imply the conclusion in (14).

A.3. Proof of Proposition 1

Observe that $\bar{\mathcal{V}}^{NN}$ can be written as

$$\bar{\mathcal{V}}^{NN} = \bar{r}_-' \bar{\Psi}_-^{NN} \bar{r}_- + \bar{r}_+' \bar{\Psi}_+^{NN} \bar{r}_+,$$

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

where $\bar{r}_- = \bar{q} \otimes \Gamma_-^{-1} e_1$, $\bar{r}_+ = \bar{q} \otimes \Gamma_+^{-1} e_1$, and

$$\bar{\Psi}_-^{NN} = \begin{bmatrix} \bar{\Psi}_{YY-}^{NN} & \bar{\Psi}_{YZ_1-}^{NN} & \cdots & \bar{\Psi}_{YZ_{|\hat{S}|}-}^{NN} \\ \bar{\Psi}_{Z_1Y-}^{NN} & \bar{\Psi}_{Z_1Z_1-}^{NN} & & \\ \vdots & & \ddots & \\ \bar{\Psi}_{Z_{|\hat{S}|}Y-}^{NN} & & & \bar{\Psi}_{Z_{|\hat{S}|}Z_{|\hat{S}|}-}^{NN} \end{bmatrix},$$

$$\bar{\Psi}_+^{NN} = \begin{bmatrix} \bar{\Psi}_{YY+}^{NN} & \bar{\Psi}_{YZ_1+}^{NN} & \cdots & \bar{\Psi}_{YZ_{|\hat{S}|}+}^{NN} \\ \bar{\Psi}_{Z_1Y+}^{NN} & \bar{\Psi}_{Z_1Z_1+}^{NN} & & \\ \vdots & & \ddots & \\ \bar{\Psi}_{Z_{|\hat{S}|}Y+}^{NN} & & & \bar{\Psi}_{Z_{|\hat{S}|}Z_{|\hat{S}|}+}^{NN} \end{bmatrix},$$

$$\bar{\Psi}_{VW-}^{NN} = \frac{1}{nh_n} \sum_{i=1}^{n} \mathbb{I}\{X_i < 0\} K(X_i/h_n)^2 \begin{bmatrix} 1 & X_i/h_n \\ X_i/h_n & (X_i/h_n)^2 \end{bmatrix} \bar{\varepsilon}_{V-,i} \bar{\varepsilon}_{W-,i},$$

$$\bar{\Psi}_{VW+}^{NN} = \frac{1}{nh_n} \sum_{i=1}^{n} \mathbb{I}\{X_i \geq 0\} K(X_i/h_n)^2 \begin{bmatrix} 1 & X_i/h_n \\ X_i/h_n & (X_i/h_n)^2 \end{bmatrix} \bar{\varepsilon}_{V+,i} \bar{\varepsilon}_{W+,i},$$

for $V, W \in \{Y, Z_1, \ldots, Z_p\}$. Similarly, the asymptotic variance $\mathcal{V}$ can be written as

$$\mathcal{V} = r_-' \Psi_- r_- + r_+' \Psi_+ r_+,$$

where $r_- = q \otimes \Gamma_-^{-1} e_1$, $r_+ = q \otimes \Gamma_+^{-1} e_1$, and

$$\Psi_- = \begin{bmatrix} \Psi_{YY-} & \Psi_{YZ_1-} & \cdots & \Psi_{YZ_{s^*}-} \\ \Psi_{Z_1Y-} & \Psi_{Z_1Z_1-} & & \\ \vdots & & \ddots & \\ \Psi_{Z_{s^*}Y-} & & & \Psi_{Z_{s^*}Z_{s^*}-} \end{bmatrix},$$

$$\Psi_+ = \begin{bmatrix} \Psi_{YY+} & \Psi_{YZ_1+} & \cdots & \Psi_{YZ_{s^*}+} \\ \Psi_{Z_1Y+} & \Psi_{Z_1Z_1+} & & \\ \vdots & & \ddots & \\ \Psi_{Z_{s^*}Y+} & & & \Psi_{Z_{s^*}Z_{s^*}+} \end{bmatrix},$$

$$\Psi_{VW-} = \frac{1}{nh_n} \sum_{i=1}^{n} \mathbb{I}\{X_i < 0\} K(X_i/h_n)^2 \begin{bmatrix} 1 & X_i/h_n \\ X_i/h_n & (X_i/h_n)^2 \end{bmatrix} \mathrm{Cov}(V_i(0), W_i(0)|X_i),$$

$$\Psi_{VW+} = \frac{1}{nh_n} \sum_{i=1}^{n} \mathbb{I}\{X_i \geq 0\} K(X_i/h_n)^2 \begin{bmatrix} 1 & X_i/h_n \\ X_i/h_n & (X_i/h_n)^2 \end{bmatrix} \mathrm{Cov}(V_i(1), W_i(1)|X_i),$$

for $V, W \in \{Y, Z_1, \ldots, Z_p\}$.

Since $\mathbb{P}\{\hat{S} \subset S^*\} \to 1$ due to the deviation bound in Theorem 1, the following statements are all conditional on the event $\{\hat{S} \subset S^*\}$. Without loss of generality, suppose $\hat{S} = \{1, \ldots, |\hat{S}|\}$ and $S^* = \{1, \ldots, |\hat{S}|, |\hat{S}| + 1, \ldots, s^*\}$. Let

$$\varpi_n = \min\{\lambda_{\min}(\Psi_-), \lambda_{\min}(\Psi_+)\}. \tag{A.11}$$

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

We decompose

$$\begin{aligned}\frac{\bar{\mathcal{V}}^{NN}}{\mathcal{V}} - 1 &= \frac{1}{\mathcal{V}}\begin{bmatrix} \bar{r}_- \\ 0 \end{bmatrix}'\left(\begin{bmatrix} \bar{\Psi}_-^{NN} & 0 \\ 0 & 0 \end{bmatrix} - \Psi_-\right)\begin{bmatrix} \bar{r}_- \\ 0 \end{bmatrix} \\ &\quad + \frac{1}{\mathcal{V}}\left(\begin{bmatrix} \bar{r}_- \\ 0 \end{bmatrix} + r_-\right)'\Psi_-\left(\begin{bmatrix} \bar{r}_- \\ 0 \end{bmatrix} + r_-\right) \\ &\quad + \frac{1}{\mathcal{V}}\begin{bmatrix} \bar{r}_+ \\ 0 \end{bmatrix}'\left(\begin{bmatrix} \bar{\Psi}_+^{NN} & 0 \\ 0 & 0 \end{bmatrix} - \Psi_+\right)\begin{bmatrix} \bar{r}_+ \\ 0 \end{bmatrix} \\ &\quad + \frac{1}{\mathcal{V}}\left(\begin{bmatrix} \bar{r}_+ \\ 0 \end{bmatrix} + r_+\right)'\Psi_+\left(\begin{bmatrix} \bar{r}_+ \\ 0 \end{bmatrix} + r_+\right). \\ &=: T_{1-} + T_{2-} + T_{1+} + T_{2+}.\end{aligned}$$

For T_{2-}, we have

$$\begin{aligned}|T_{2-}| &\le \varpi_n^{-1}\left|\begin{bmatrix} \bar{r}_- \\ 0 \end{bmatrix} + r_-\right|_1\left|\Psi_-\left(\begin{bmatrix} \bar{r}_- \\ 0 \end{bmatrix} + r_-\right)\right|_\infty \\ &= O_p\left(\varpi_n^{-1}\{(m_n|S_n|\lambda_n) \vee (|\hat{S}_1|\zeta_n) \vee (m_{1n}|\hat{S}_1|\zeta_n^2/\lambda_n)\}\right),\end{aligned}$$

where the inequality follows from $\mathcal{V} \ge \varpi_n$ and Hölder's inequality, and the equality follows from Theorem 1(ii). A similar argument yields that T_{2+} is of the same stochastic order.

For T_{1-}, letting Ψ_-^{11} be the first $|\hat{S}| \times |\hat{S}|$ block component of Ψ_-, we have

$$|T_{1-}| \le \varpi_n^{-1}\lambda_{\max}(\bar{\Psi}_-^{NN} - \Psi_-^{11})\bar{r}_-'\bar{r}_- \le \max_{1\le j_1,j_2\le 2\hat{S}}|\bar{\Psi}_{-,l_1,l_2}^{NN} - \Psi_{-,l_1,l_2}^{11}|O_p(\varpi_n^{-1}s^*).$$

So it remains to obtain the stochastic order of $\max_{1\le l_1,l_2\le 2\hat{S}}|\bar{\Psi}_{-,l_1,l_2}^{NN} - \Psi_{-,l_1,l_2}^{11}|$. From page 36 of the supplement of CCT, we can decompose

$$\bar{\Psi}_{-,VW}^{NN} - \Psi_{-,VW}^{11} = \eta_1^{VW} + \eta_2^{VW} + \eta_3^{VW},$$

for $V, W \in \{Y, Z_1, \ldots, Z_p\}$, where $\varepsilon_{V,i} = V_i - \mathbb{E}[V_i|X_i]$, $\varepsilon_{W,i} = W_i - \mathbb{E}[W_i|X_i]$, $\mu_{V-}(x) = \mathbb{E}[V_i(0)|X_i = x]$, $\mu_{W-}(x) = \mathbb{E}[W_i(0)|X_i = x]$, and

$$\begin{aligned}\eta_1^{VW} &= \frac{1}{J(J+1)}\sum_{j=1}^{J}\eta_{1,j}^{VW}, \\ \eta_{1,j}^{VW} &= \frac{1}{nh_n}\sum_{i=1}^{n}\mathbb{I}\{X_i < 0\}K(X_i/h_n)^2\begin{bmatrix} 1 & X_i/h_n \\ X_i/h_n & (X_i/h_n)^2 \end{bmatrix} \\ &\quad \times (\varepsilon_{V,\ell_{-,j}(i)}\varepsilon_{W,\ell_{-,j}(i)} - \varepsilon_{V,i}\varepsilon_{W,i}), \\ \eta_2^{VW} &= \frac{2}{J(J+1)}\sum_{1\le j,k\le J}\eta_{2,j,k}^{VW}, \\ \eta_{2,j,k}^{VW} &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{X_i < 0\}K(X_i/h_n)^2\begin{bmatrix} 1 & X_i/h_n \\ X_i/h_n & (X_i/h_n)^2 \end{bmatrix}\varepsilon_{V,\ell_{-,j}(i)}\varepsilon_{W,\ell_{-,k}(i)},\end{aligned}$$

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

$$\eta_3^{VW} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{X_i<0\}K(X_i/h_n)^2\begin{bmatrix}1 & X_i/h_n\\ X_i/h_n & (X_i/h_n)^2\end{bmatrix}$$
$$\times\{(\hat\sigma_{3,i}^{VW})^2-(\hat\sigma_{4,i}^{VW})^2-(\hat\sigma_{5,i}^{VW})^2+(\hat\sigma_{6,i}^{VW})^2+(\hat\sigma_{7,i}^{VW})^2-(\hat\sigma_{8,i}^{VW})^2-(\hat\sigma_{9,i}^{VW})^2\},$$

$$(\hat\sigma_{3,i}^{VW})^2 = \frac{1}{J(J+1)}\left(\sum_{j=1}^{J}\{\mu_{V-}(X_i)-\mu_{V-}(X_{\ell_{-,j}(i)})\}\right)\left(\sum_{j=1}^{J}\{\mu_{W-}(X_i)-\mu_{W-}(X_{\ell_{-,j}(i)})\}\right),$$

$$(\hat\sigma_{4,i}^{VW})^2 = \varepsilon_{V,i}\frac{1}{J+1}\sum_{j=1}^{J}\varepsilon_{W,\ell_{-,j}(i)},\qquad (\hat\sigma_{5,i}^{VW})^2 = \varepsilon_{W,i}\frac{1}{J+1}\sum_{j=1}^{J}\varepsilon_{V,\ell_{-,j}(i)},$$

$$(\hat\sigma_{6,i}^{VW})^2 = \varepsilon_{V,i}\frac{1}{J+1}\sum_{j=1}^{J}\{\mu_{W-}(X_i)-\mu_{W-}(X_{\ell_{-,j}(i)})\},$$

$$(\hat\sigma_{7,i}^{VW})^2 = \varepsilon_{W,i}\frac{1}{J+1}\sum_{j=1}^{J}\{\mu_{V-}(X_i)-\mu_{V-}(X_{\ell_{-,j}(i)})\},$$

$$(\hat\sigma_{8,i}^{VW})^2 = \frac{1}{J+1}\sum_{j=1}^{J}\varepsilon_{V,\ell_{-,j}(i)}\sum_{k=1}^{J}\{\mu_{W-}(X_i)-\mu_{W-}(X_{\ell_{-,j}(i)})\},$$

$$(\hat\sigma_{9,i}^{VW})^2 = \frac{1}{J+1}\sum_{j=1}^{J}\varepsilon_{W,\ell_{-,j}(i)}\sum_{k=1}^{J}\{\mu_{V-}(X_i)-\mu_{V-}(X_{\ell_{-,j}(i)})\}.$$

Since the argument is similar to the one in Section S.2.4 of CCT's supplement, we only present the result for η_1^{VW}. Since J is fixed, take any j. We want to obtain the stochastic order of

$$\max_{V,W\in\{Y,Z_1,\ldots,Z_{|\hat S|}\}}\left|\frac{1}{nh_n}\sum_{i=1}^{n}\mathbb{I}\{X_i<0\}K(X_i/h_n)^2(\varepsilon_{V,\ell_{-,j}(i)}\varepsilon_{W,\ell_{-,j}(i)}-\varepsilon_{V,i}\varepsilon_{W,i})\right|.$$

Note that

$$\max_{V,W\in\{Y,Z_1,\ldots,Z_{|\hat S|}\}}\left|\frac{1}{nh_n}\sum_{i=1}^{n}\mathbb{I}\{X_i<0\}K(X_i/h_n)^2(\varepsilon_{V,\ell_{-,j}(i)}\varepsilon_{W,\ell_{-,j}(i)}-\varepsilon_{V,i}\varepsilon_{W,i})\right|$$
$$\le \max_{V,W\in\{Y,Z_1,\ldots,Z_{|\hat S|}\}}\left|\frac{1}{nh_n}\sum_{i=1}^{n}\mathbb{I}\{X_i<0\}K(X_i/h_n)^2\begin{Bmatrix}(\varepsilon_{V,\ell_{-,j}(i)}\varepsilon_{W,\ell_{-,j}(i)}-\varepsilon_{V,i}\varepsilon_{W,i})\\ -(\sigma_{VW}^2(X_{\ell_{-,j}(i)})-\sigma_{VW}^2(X_i))\end{Bmatrix}\right|$$
$$+\max_{V,W\in\{Y,Z_1,\ldots,Z_{|\hat S|}\}}\left|\frac{1}{nh_n}\sum_{i=1}^{n}\mathbb{I}\{X_i<0\}K(X_i/h_n)^2(\sigma_{VW}^2(X_{\ell_{-,j}(i)})-\sigma_{VW}^2(X_i))\right|$$
$$=: T_{11}+T_{12}.$$

For T_{11}, we can apply the localized Bernstein inequality in Lemma A.1, which implies

$$T_{11} = O_p\left(\sqrt{\frac{\log s^*}{nh_n}}\right).$$

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

For T_{12}, note that

$$\begin{aligned} T_{12} &= \max_{V,W\in\{Y,Z_1,\ldots,Z_{|\hat{S}|}\}} \left| \frac{1}{nh_n} \sum_{i=1}^{n} \mathbb{I}\{X_i < 0\} K(X_i/h_n)^2 (\sigma^2_{VW}(X_{\ell_j^-(i)}) - \sigma^2_{VW}(X_i)) \right| \\ &\leq \max_{1\leq i\leq n} |X_{\ell_j^-(i)} - X_i| \max_{V,W\in\{Y,Z_1,\ldots,Z_{|\hat{S}|}\}} |C_{VW}| \left| \frac{1}{nh_n} \sum_{i=1}^{n} \mathbb{I}\{X_i < 0\} K(X_i/h_n)^2 \right| \\ &= O_p\left(\max_{1\leq i\leq n} |X_{\ell_j^-(i)} - X_i| \right) = O_p(n^{-1}), \end{aligned}$$

where C_{VW} is the Lipschitz constant for σ^2_{VW} (and we assume $\max_{V,W\in\{Y,Z_1,\ldots,Z_{|\hat{S}|}\}} |C_{VW}|$ is bounded), and the second equality follows from Abadie and Imbens (2006, Thm. 1). Combining these results, we have

$$T_{1-} = O_p\left(\frac{s^* \sqrt{\log s^*}}{\varpi_n n h_n} \right).$$

REFERENCES

Abadie, A., & Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74, 253–267.

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

Arai, Y., Otsu, T. and Seo, M. (2021). Regression discontinuity design with potentially many covariates. Preprint, arXiv:2109.08351.

Armstrong, T. B., & Kolesár, M. (2018). Optimal inference in a class of regression models. *Econometrica*, 86, 655–683.

Belloni, A., Chernozhukov, V., Chetverikov, D., Hansen, C., & Kato, K. (2018). High-dimensional econometrics and regularized GMM. Massachusetts Institute of Technology, Cambridge MA, Working Paper.

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

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

Bühlmann, P., & van de Geer, S. (2011). *Statistics for high-dimensional data*. Springer.

Calonico, S., Cattaneo, M. D., & Farrell, M. H. (2018). On the effect of bias estimation on coverage accuracy in nonparametric inference. *Journal of the American Statistical Association*, 113, 767–779.

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

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2017). Rdrobust: Software for regression discontinuity designs. *Stata Journal*, 17, 372–404.

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

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

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2015b). Rdrobust: An R package for robust inference in regression discontinuity design. *R Journal*, 7, 38–51.

https://doi.org/10.1017/S0266466624000239 Published online by Cambridge University Press

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

Cattaneo, M. D., & Escanciano, J. C. (2017). *Regression discontinuity designs: Theory and applications*, in Advances in Econometrics, vol. 38, Emerald Group Publishing.

Cattaneo, M. D., & Titiunik, R. (2022). Regression discontinuity designs. *Annual Review of Economics*, 14, 821–851.

Cattaneo, M. D., Titiunik, R., & Vazquez-Bare, G. (2020). The regression discontinuity design. In L. Curini & R. J. Franzese (Eds.), *Handbook of research methods in political science and international relations* (Ch. 44, pp. 835–857). Sage Publications.

Fan, J., & Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *Annals of Statistics*, 20, 2008–2036.

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

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

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

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

Kreiß, A., & Rothe, C. (2023). Inference in regression discontinuity designs with high-dimensional covariates. *Econometrics Journal*, 26, 105–123.

Lei, L., & Ding, P. (2021). Regression adjustment in completely randomized experiments with a diverging number of covariates. *Biometrika*, 108, 815–828.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Annals of Applied Statistics*, 7, 295–318.

Ludwig, J., & Miller, D. L. (2007). Does head start improve children's life chances? Evidence from a regression discontinuity design. *Quarterly Journal of Economics*, 122, 159–208.

Ruppert, D., & Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Annals of Statistics*, 22, 1346–1370.

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

van der Vaart, A. W., & Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer.

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