

Two Stage Curvature Identification with Machine Learning: Causal Inference with Possibly Invalid Instrumental Variables

Zijian Guo and Peter Bühlmann

Abstract

Instrumental variables regression is a popular causal inference method for endogenous treatment. A significant concern in practical applications is the validity and strength of instrumental variables. This paper aims to perform causal inference when all instruments are possibly invalid. To do this, we propose a novel methodology called two stage curvature identification (TSCI) together with a generalized concept to measure the strengths of possibly invalid instruments: such invalid instruments can still be used for inference in our framework. We fit the treatment model with a general machine learning method and propose a novel bias correction method to remove the overfitting bias from machine learning methods. Among a collection of spaces of violation functions, we choose the best one by evaluating invalid instrumental variables' strength. We demonstrate our proposed TSCI methodology in a large-scale simulation study and revisit the important economics question on the effect of education on earnings.

1 Introduction

Observational studies are major sources for inferring causal effects when randomized experiments are not feasible. But such causal inference from observational studies requires strong assumptions and may be invalid due to the presence of unmeasured confounders. The instrumental variable (IV) regression is a practical and highly popular causal inference approach in the presence of unmeasured confounders. The IVs are required to satisfy three main assumptions: conditioning on the baseline covariates,

- (A1) the IVs are associated with the treatment variable;
- (A2) the IVs are not associated with the unmeasured confounders;
- (A3) the IVs do not directly affect the outcome variable.

Despite the popularity of the IV method, there is a major concern on whether the used IVs satisfy (A1)-(A3) in practical applications. Assumption (A1) requires the IV to be strongly associated with the treatment variable, which can be checked with an F test in a first stage linear regression model. Inference with the assumption (A1) being violated has been actively investigated under the name of weak IV. Notable works along this direction include [2, 44, 17, 52, 53, 51, 33, 42]. Assumptions (A2) and (A3) are the exclusion restriction assumptions, ensuring that the IV only affects the outcome through the treatment. We shall refer to an IV as an invalid IV if it violates either (A2) or (A3). Most empirical analyses rely on external knowledge to justify assumptions (A2) and (A3). We propose here a framework which does not rely on such additional assumptions as we believe that there is a great need to develop a causal inference method which is robust against IVs violating the classical assumptions.

1.1 Our results and contribution

We focus on the setting where all IVs are allowed to be invalid and the errors of the outcome and treatment models can be either homoscedastic or heteroscedastic. We propose a novel two stage curvature identification (TSCI) method to make inferences for the treatment effect. A key assumption is that the violations arise from different functional forms than the association between the instrument and the treatment; that is, we exclude “special coincidences” such as the violations as well as the association between instrument and treatment are all linear. An important operational step is to fit the treatment model with a machine learning algorithm, e.g., random forests, boosting, or deep neural network. This is helpful to capture a general non-linear relationship between treatment and the IVs. By generalizing the concentration parameter [53, 2, 25], we introduce a generalized IV strength measure for TSCI with invalid IVs.

We show that our developed TSCI methodology yields a consistent estimator of the treatment effect assuming the above mentioned “different functional form” of violations and instrument-treatment association. In addition, we show that the TSCI estimator is asymptotically normal, when the generalized IV strength measure is sufficiently large. The convergence rate is $1/\sqrt{n}$ for the favorable setting with sufficiently large generalized IV

strength.

We argue that an additional bias correction of the machine learning driven TSCI must be developed and performed. The reason is that a direct application of machine learning algorithms to the endogenous treatment model may lead to overfitting bias: the first-stage predicted values by machine learning algorithms (e.g., random forests) might still be endogenous due to the overfitting nature; see Remark 2.

Furthermore, we show that a data-adaptive construction of IV violation forms is important for reliable performance: a best set of basis functions for the violation form over various corresponding TSCI estimators is selected according to our novel generalized IV strength measure. We construct a confidence interval for the treatment effect with this best set of basis functions and establish that it achieves the desired coverage level asymptotically. We demonstrate its finite sample performance in a large-scale simulation study.

To sum up, the contribution of the current paper is two-fold:

1. Our proposed TSCI method is a robust instrumental variable approach allowing for invalid IVs. TSCI conducts self-checking of the IV assumptions by exploring a non-linear relationship between the treatment and IVs. TSCI leads to more reliable causal conclusions than existing methods by allowing for a broad class of invalid IVs.
2. We integrate machine learning algorithms into the TSCI method. A methodological novelty of TSCI is its bias correction in its second stage, which addresses the issue of overfitting bias due to the machine learning.

1.2 Comparison to existing literature

There is relevant literature on causal inference when the assumptions (A2) and (A3) are violated. We shall mention two related directions in the setting with multiple IVs. Firstly, [10] and [34] considered inference for the treatment effect when the direct effect of the instruments on the outcome is nearly orthogonal to the effect of the instruments on the treatment. Secondly, a recent line of research [24, 11, 31, 57, 23, 58, 37] considered the setting that a proportion of the candidate IVs are valid and made inference for the treatment effect by selecting valid IVs in a data-dependent way. Along this line, [22] proposed a uniform inference procedure that is robust to the selection errors. In contrast to assuming the orthogonality and existence of valid IVs, our proposed TSCI is effective even if *all* IVs are invalid and the effect orthogonality condition does not hold.

When all IVs are invalid, [35, 36, 54] proposed an identification strategy of the treatment effect by assuming the variance of the treatment model error is heteroscedastic and the covariance between the treatment and outcome model errors are homoscedastic. Our proposed TSCI is based on a completely different idea by exploring the non-linear structure between the treatment and the IVs not requiring anything on homo- and heteroscedasticity.

The construction of a non-parametric treatment model has been considered in the IV regression literature [32, 1, 43], with a low order polynomial transformation of IVs [32] or a nearest neighborhood and series approximation [43]. [8] proposed to construct the optimal IVs with a Lasso-type first stage estimator. The non-parametric treatment model is helpful in constructing optimal IVs and efficient IV estimators. More recently, machine learning algorithms have been integrated into the instrumental variable analysis to better capture the complex relationship in both the outcome and treatment models. [6] proposed the generalized random forests and applied it to make inference for heterogeneous treatment effect via IVs; [27] proposed the Deep IV method by fitting the outcome and treatment models with deep neural networks; [39] considered a simplified setting of [27] by assuming a linear outcome model; [9] studied the deep generalized method of moments with IVs; [59] applied the deep neural networks to define the non-linear features of the treatments and instruments. All of these works require prior knowledge of valid IVs, while our current paper focuses on the different robust IV framework with all IVs being possibly invalid.

Double machine learning [19] and regularizing double machine learning [20] were proposed to apply machine learning algorithms to construct estimators of nuisance parameters. The framework is still built on assuming valid IVs, and the effect identification only uses a linear association between the treatment and the IVs. Our current proposal is distinct because we allow for invalid IVs and utilize the possibly non-linear relationship between the treatment and IVs.

As a methodological application, our proposed TSCI methodology can be used to test the validity of IVs. This validity test is effective for the just identification case (e.g., one endogenous variable and one IV). This is different from the Sargan test or J test [26, 48] which are designed to test IV validity in the over-identification case.

Notation. For a matrix $X \in \mathbb{R}^{n \times p}$, a vector $x \in \mathbb{R}^n$, and a set $\mathcal{A} \subset \{1, \dots, n\}$, we use $X_{\mathcal{A}}$ to denote the submatrix of X whose row indices belong to \mathcal{A} , and $x_{\mathcal{A}}$ to denote the sub-vector of x with indices in \mathcal{A} . For a set \mathcal{A} , $|\mathcal{A}|$ denotes the cardinality of \mathcal{A} . For a vector $x \in \mathbb{R}^p$, the ℓ_q norm of x is defined as $\|x\|_q = (\sum_{l=1}^p |x_l|^q)^{\frac{1}{q}}$ for $q \geq 0$ with $\|x\|_0 = |\{1 \leq l \leq p : x_l \neq 0\}|$ and $\|x\|_{\infty} = \max_{1 \leq l \leq p} |x_l|$. We use c and C to denote generic

positive constants that may vary from place to place. For a sequence of random variables X_n indexed by n , we use $X_n \xrightarrow{p} X$ and $X_n \xrightarrow{d} X$ to represent that X_n converges to X in probability and in distribution, respectively. For two positive sequences a_n and b_n , $a_n \lesssim b_n$ means that $\exists C > 0$ such that $a_n \leq Cb_n$ for all n ; $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$, and $a_n \ll b_n$ if $\limsup_{n \rightarrow \infty} a_n/b_n = 0$. For a matrix M , we use $\text{Tr}[M]$ to denote its trace, $\text{rank}(M)$ to denote its rank, and $\|M\|_F$, $\|M\|_2$ and $\|M\|_\infty$ to denote its Frobenius norm, spectral norm and element-wise maximum norm, respectively. For a square matrix M , M^2 denotes the matrix multiplication of M by itself. For a symmetric matrix M , we use $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ to denote its maximum and minimum eigenvalues, respectively.

2 Invalid instruments: modeling and identification

We consider i.i.d. data $\{Y_i, D_i, Z_i, X_i\}_{1 \leq i \leq n}$, where for the i -th observation, $Y_i \in \mathbb{R}$ and $D_i \in \mathbb{R}$ respectively denote the outcome and the treatment and $Z_i \in \mathbb{R}^{p_z}$ and $X_i \in \mathbb{R}^{p_x}$ respectively denote the set of Instrumental Variables (IVs) and the set of measured covariates. We consider the outcome model: for $1 \leq i \leq n$,

$$Y_i = D_i\beta + g(Z_i, X_i) + \epsilon_i, \quad \text{with } \mathbf{E}(\epsilon_i | X_i, Z_i) = 0, \quad (1)$$

where $\beta \in \mathbb{R}$ is the constant effect of the treatment on the outcome and $g : \mathbb{R}^{p_x+p_z} \rightarrow \mathbb{R}$. We define $h(Z_i, X_i) = g(Z_i, X_i) - \phi(X_i)$ with $\phi(X_i) = \mathbf{E}[g(Z_i, X_i) | X_i]$ and rewrite the outcome model (1) as

$$Y_i = D_i\beta + h(Z_i, X_i) + \phi(X_i) + \epsilon_i, \quad \text{with } \mathbf{E}(\epsilon_i | X_i, Z_i) = 0. \quad (2)$$

For a valid IV Z_i , the function $g(Z_i, X_i)$ does not directly depend on the assignment of Z_i , which implies $h(\cdot) = 0$. We shall refer to this h function as the violation function, and a non-zero h function implies that the assumptions (A2) and (A3) are violated.

We consider the following treatment model: for $1 \leq i \leq n$,

$$D_i = f(Z_i, X_i) + \delta_i \quad \text{with } \mathbf{E}(\delta_i | Z_i, X_i) = 0. \quad (3)$$

The model (3) is flexible as f might be any unknown function of Z_i and X_i . Particularly, the treatment variable can be continuous, binary or a count variable.

In the following, we introduce the causal potential outcome interpretation of the model (1) (and analogously (2)) and demonstrate the violations of assumptions (A2) and (A3). For

the i -th subject with the baseline covariates X_i , we use $Y^{(z,d)}(X_i)$ to denote the potential outcome with the IVs and the treatment assigned to $z \in \mathbb{R}^{p_z}$ and $d \in \mathbb{R}$, respectively. For $1 \leq i \leq n$, we consider the potential outcome model [50, 46],

$$Y_i^{(z,d)}(X_i) = Y_i^{(0,0)}(X_i) + d\beta + g_1(z, X_i) \quad \text{and} \quad \mathbf{E}(Y_i^{(0,0)}(X_i) | Z_i, X_i) = g_2(Z_i, X_i), \quad (4)$$

where $\beta \in \mathbb{R}$ denotes the treatment effect, $g_1 : \mathbb{R}^{p_z+p_x} \rightarrow \mathbb{R}$, and $g_2 : \mathbb{R}^{p_z+p_x} \rightarrow \mathbb{R}$. If $g_1(z, x)$ changes with z , the IVs directly affect the outcome, violating the assumption (A2). If $g_2(z, x)$ changes with z , the IVs are associated with unmeasured confounders, violating the assumption (A3). The model (4) extends the Additive LInear, Constant Effects (ALICE) model of [28] by allowing for possibly invalid instruments. The model (4) also generalizes the invalid IV model of [49, 31, 23, 57] by allowing for non-linear $g_1(\cdot)$ and $g_2(\cdot)$. By the consistency assumption $Y_i = Y_i^{(Z_i, D_i)}(X_i)$, the potential outcome model (4) implies the outcome model (1) with $g(\cdot) = g_1(\cdot) + g_2(\cdot)$ and $\epsilon_i = Y_i^{(0,0)}(X_i) - g_2(Z_i, X_i)$.

For $j = 1, 2$, we define $h_j(Z_i, X_i) = g_j(Z_i, X_i) - \phi_j(X_i)$ with $\phi_j(X_i) = \mathbf{E}(g_j(Z_i, X_i) | X_i)$. For $j = 1, 2$, if $g_j(z, x)$ changes with the z value, there exists some z such that $h_j(z, x) \neq 0$. Consequently, the function h_j represents the possible violations of the classical IV assumptions: if both h_1 and h_2 are zero functions, the assumptions (A2) and (A3) are satisfied; otherwise, the assumptions are violated. We illustrate the effect of h_1 and h_2 in Figure 1. The potential outcome model (4) further implies the model (2) with $h(\cdot) = h_1(\cdot) + h_2(\cdot)$, $\phi(\cdot) = \phi_1(\cdot) + \phi_2(\cdot)$, and $\epsilon_i = Y_i^{(0,0)}(X_i) - g_2(Z_i, X_i)$.

As a remark, we can easily generalize the potential outcome model (4) by considering $Y_i^{(z,d)}(X_i) = Y_i^{(0,0)}(X_i) + d\beta_i + g_1(z, X_i)$, where β_i denotes the individual effect for the i -th individual. If we consider the random effect β_i with $\mathbf{E}\beta_i = \beta$ and $\beta_i - \beta$ being independent of (Z_i, X_i, D_i) , then we obtain the model (1) with $\epsilon_i = Y_i^{(0,0)}(X_i) - g_2(Z_i, X_i) + (\beta_i - \beta)D_i$ and our proposed method can be applied to make inference for $\beta = \mathbf{E}\beta_i$.

2.1 Structural equation model interpretation

We interpret the outcome model (2) and the treatment model (3) from the perspective of a Structural Equation Model (SEM) with hidden confounders. For expositional simplicity, we focus on a special setting with no covariates X_i . When there are no covariates, we have $h(Z_i) = g(Z_i) - \mathbf{E}(g(Z_i))$ and $h_j(Z_i) = g_j(Z_i) - \mathbf{E}(g_j(Z_i))$ for $j = 1, 2$.

For $1 \leq i \leq n$, we consider the following SEM for Y_i and D_i ,

$$Y_i \leftarrow a_0 + D_i\beta + h_1(Z_i) + \nu_1(H_i) + \epsilon_i^0, \quad \text{and} \quad D_i \leftarrow f_1(Z_i) + \nu_2(H_i) + \delta_i^0, \quad (5)$$

where H_i denotes hidden confounders, ϵ_i^0 and δ_i^0 are random errors independent of D_i, Z_i, H_i , and a_0 is the intercept such that $\mathbf{E}h_1(Z_i) = \mathbf{E}\nu_1(H_i) = 0$. The unmeasured confounders H_i might affect both the outcome and treatment. We define conditional expectations of $\nu_1(H_i)$ and $\nu_2(H_i)$ as $h_2(Z_i) = \mathbf{E}(\nu_1(H_i) | Z_i)$ and $f_2(Z_i) = \mathbf{E}(\nu_2(H_i) | Z_i)$. The above SEM model (5) is illustrated in Figure 1.

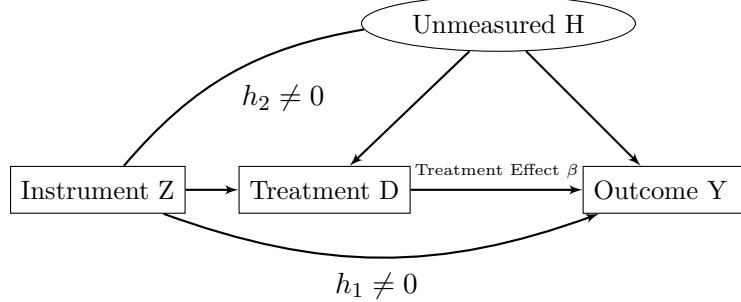


Figure 1: Illustration of $h_1 \neq 0$ and $h_2 \neq 0$ in (2) and the SEM (5).

The SEM (5) and Figure 1 demonstrate the meanings of h_1 and h_2 : h_1 denotes a direct effect of the treatment on the outcome and h_2 captures the association between the treatment and the unmeasured confounders. The SEM (5) together with the definitions of f_2 and h_2 imply the outcome model $Y_i = a_0 + D_i\beta + h(Z_i) + \epsilon_i$ with $h(Z_i) = h_1(Z_i) + h_2(Z_i)$ and $\epsilon_i = \epsilon_i^0 + \nu_1(H_i) - h_2(Z_i)$, which is the model (2) with no covariates. The treatment model $D_i = f(Z_i) + \delta_i$ arises with $f(Z_i) = f_1(Z_i) + f_2(Z_i)$ and $\delta_i = \delta_i^0 + \nu_2(H_i) - f_2(Z_i)$.

2.2 Two Stage Curvature Identification of β

We now present the identification of the treatment effect β in the outcome model (2) with a non-zero h . To illustrate the main idea, we consider again the special setting with no covariates X_i and will propose a general methodology allowing for X_i in the following Section 3. A combination of (2) and (3) leads to the following reduced form model,

$$Y_i = F(Z_i) + \epsilon_i + \beta\delta_i \quad \text{with} \quad F(Z_i) = \beta f(Z_i) + h(Z_i). \quad (6)$$

To facilitate the discussion, we define a Hilbert space $\mathcal{S} = \{g \mid g : \mathbb{R}^{p_z} \rightarrow \mathbb{R}\}$ for functions of $Z_1 \in \mathbb{R}^{p_z}$. For $g_1, g_2 \in \mathcal{S}$, we define the inner product $\langle g_1, g_2 \rangle = \int g_1(z)g_2(z)d\mu(z)$ and the induced norm $\|g_1\| = \sqrt{\langle g_1, g_1 \rangle}$, where μ is the probability measure of $Z_1 \in \mathbb{R}^{p_z}$.

We define $\mathcal{V} := \{v_1(\cdot), \dots, v_L(\cdot)\} \subset \mathcal{S}$ with the aim that $h(\cdot) \in \text{Span}(\mathcal{V})$, where $\text{Span}(\mathcal{V}) := \{\sum_{j=1}^L a_j v_j(\cdot) : a \in \mathbb{R}^L\}$ denotes the linear span of the basis set \mathcal{V} . For example, we may take $\{v_j(z)\}_{1 \leq j \leq L}$ as polynomials or other standard basis functions. For

$f \in \mathcal{S}$, define its projection to the space spanned by \mathcal{V} as $\mathcal{P}_{\mathcal{V}}f = \arg \min_{v \in \text{Span}(\mathcal{V})} \|f - v\|$, and $\mathcal{P}_{\mathcal{V}}^\perp f = f - \mathcal{P}_{\mathcal{V}}f$. The following proposition characterizes a necessary and sufficient condition of identifying β in (6) with $h \neq 0$.

Proposition 1 *Consider the reduced form model (6) and assume $h \in \text{Span}(\mathcal{V})$. If $f \notin \text{Span}(\mathcal{V})$, the treatment effect β is identifiable with $\beta = \langle \mathcal{P}_{\mathcal{V}}^\perp F, \mathcal{P}_{\mathcal{V}}^\perp f \rangle / \langle \mathcal{P}_{\mathcal{V}}^\perp f, \mathcal{P}_{\mathcal{V}}^\perp f \rangle$. If $f \in \text{Span}(\mathcal{V})$, there exists $h \in \text{Span}(\mathcal{V})$ such that β is not identifiable.*

Proposition 1 illustrates a simple yet useful identification strategy of β : we project out the possible violation function h and use the remaining component $\mathcal{P}_{\mathcal{V}}^\perp f$ to identify β . The effect identification is possible if $\|\mathcal{P}_{\mathcal{V}}^\perp f\| > 0$. Particularly, if \mathcal{V} is the space of linear functions, we are identifying the treatment effect by using the curvatures of F and f . Hence, we name such an identification strategy as two stage curvature identification (TSCI).

If $f \notin \mathcal{V}$, we apply $\mathcal{P}_{\mathcal{V}}^\perp$ to both sides of (6) and obtain $\mathcal{P}_{\mathcal{V}}^\perp F = \beta \mathcal{P}_{\mathcal{V}}^\perp f + \mathcal{P}_{\mathcal{V}}^\perp h = \beta \mathcal{P}_{\mathcal{V}}^\perp f$, which leads to the identification expression. If $f \in \mathcal{V}$, then, as a counter-example, $F(Z_i) = (\beta + 1)f(Z_i) + (h(Z_i) - f(Z_i))$ is not distinguished from $F(Z_i) = \beta f(Z_i) + h(Z_i)$ in (6).

In the more general setting with baseline covariates, we combine (2) and (3) and obtain,

$$Y_i = F(Z_i, X_i) + \epsilon_i + \beta \delta_i \quad \text{with} \quad F(Z_i, X_i) = \beta f(Z_i, X_i) + \phi(X_i) + h(Z_i, X_i).$$

The identification strategy in Proposition 1 can be directly extended to the situation here, where we project out both $\phi(X_i)$ and $h(X_i, Z_i)$ (or as a total $g(X_i, Z_i)$) out of the function $f(Z_i, X_i)$.

3 TSCI with Random Forests

A key component of the TSCI is to estimate $f(X_i, Z_i) = \mathbf{E}(D_i | X_i, Z_i)$. We develop in the following of this section TSCI with estimation of the conditional mean function by random forests, with extensions to general machine learning methods in Section 4.

We shall emphasize that a direct generalization of the two stage least squares estimator by replacing its first stage with random forests does not work, even though random forests has an excellent prediction performance. We show the deficiencies of such random forests or more general machine learning methods in the following Remarks 1 and 2.

Our proposed TSCI with random forests relies on viewing the split random forests as a weighting estimator, a technique that has been utilized in [38, 40]. Based on such a

weighting scheme, we propose a novel TSCI estimator with random forests in Section 3.2. The main novelty is to correct the overfitting bias caused by the random forests.

3.1 Split random forests: a weighting estimator

In the following, we review the results in [38, 40], showing that the sample split random forests can be expressed as a weighted average of the response variables. We randomly split the data into two disjoint subsets \mathcal{A}_1 and \mathcal{A}_2 with $|\mathcal{A}_1| = n_1 = \lfloor 2n/3 \rfloor$ and $|\mathcal{A}_2| = n - n_1$. Without loss of generality, we write $\mathcal{A}_1 = \{1, 2, \dots, n_1\}$. We construct the random forests with the data $\{D_i, X_i, Z_i\}_{i \in \mathcal{A}_2}$ and apply the constructed random forests together with the data $\{X_i, Z_i, D_i\}_{i \in \mathcal{A}_1}$ to estimate $f(Z_i, X_i)$ for $i \in \mathcal{A}_1$; see the following equation (8) for details. The sample splitting is essential for the success of our proposed TSCI; see Remark 2 for detailed discussions.

Random forests aggregates $S \geq 1$ decision trees, where each tree is built with a bootstrapped sample from $\{X_i, Z_i, D_i\}_{i \in \mathcal{A}_2}$, and at each splitting, only a small number of randomly sampled covariates are considered to be split. Let θ denote the random parameter that determines how a tree is grown. Each decision tree can be viewed as the partition of the whole covariate space $\mathbb{R}^{p_x + p_z}$ into disjoint subspaces $\{\mathcal{R}_l\}_{1 \leq l \leq J}$. For any given $(z^\top, x^\top)^\top \in \mathbb{R}^{p_x + p_z}$ and a given tree with the parameter θ , there exists an unique leaf $l(z, x, \theta)$ with $1 \leq l(z, x, \theta) \leq J$ such that $\mathcal{R}_{l(z, x, \theta)}$ contains $(z^\top, x^\top)^\top$. With the observations inside $\mathcal{R}_{l(z, x, \theta)}$, the decision tree predicts $f(z, x) = \mathbf{E}(D | Z = z, X = x)$ by

$$\hat{f}_\theta(z, x) = \sum_{j \in \mathcal{A}_1} \omega_j(z, x, \theta) D_j \quad \text{with} \quad \omega_j(z, x, \theta) = \frac{\mathbf{1}[(Z_j^\top, X_j^\top)^\top \in \mathcal{R}_{l(z, x, \theta)}]}{\sum_{k \in \mathcal{A}_1} \mathbf{1}[(Z_k^\top, X_k^\top)^\top \in \mathcal{R}_{l(z, x, \theta)}]}. \quad (7)$$

Random forests aggregate S partitions of the covariate space. We use $\{\theta_1, \dots, \theta_S\}$ to denote the parameters corresponding to the S trees of the random forests. The random forests estimator $\hat{f}(z, x) = \frac{1}{S} \sum_{s=1}^S \hat{f}_{\theta_s}(z, x)$ can be expressed as,

$$\hat{f}(z, x) = \sum_{j \in \mathcal{A}_1} \omega_j(z, x) D_j \quad \text{with} \quad w_j(z, x) = \frac{1}{S} \sum_{s=1}^S \omega_j(z, x, \theta_s),$$

where $\omega_j(z, x, \theta_s)$ is defined in (7). The weights $\{\omega_j(z, x)\}_{j \in \mathcal{A}_1}$ satisfy $\omega_j(z, x) \geq 0$ and $\sum_{j \in \mathcal{A}_1} \omega_j(z, x) = 1$. In a matrix notation, we estimate $f_{\mathcal{A}_1} = (f(Z_1, X_1), \dots, f(Z_{n_1}, X_{n_1}))^\top$ by

$$\hat{f}_{\mathcal{A}_1} = \Omega D_{\mathcal{A}_1} \quad \text{with} \quad \Omega_{ij} = \omega_j(Z_i, X_i) \quad \text{for } i, j \in \mathcal{A}_1. \quad (8)$$

The transformation matrix $\Omega \in \mathbb{R}^{n_1 \times n_1}$ serves as a similar role as the hat matrix in linear regression. However, the matrix Ω in (8) is not a projection matrix. Importantly, Ω is computed based on the covariate data $\{X_i, Z_i\}_{i \in \mathcal{A}_1}$ and the random forests constructed by the data $\{X_i, Z_i, D_i\}_{i \in \mathcal{A}_2}$. Its construction does not directly depend on $\{D_i\}_{i \in \mathcal{A}_1}$, which helps remove the endogeneity of the estimated $\hat{f}(Z_i, X_i)$; see Remark 2 for more discussions.

3.2 TSCI with random forests: correction of overfitting bias

For the outcome model (1), we approximate the function $g(X_i, Z_i) = h(X_i, Z_i) + \phi(X_i)$ by a set of basis functions. When the IVs are valid with $h(\cdot) = 0$, we then approximate the function $g(X_i, Z_i) = \phi(X_i)$ by $W_i \in \mathbb{R}^{p_w}$, which are basis transformations of the baseline covariates X_i . The first element of W_i is taken as the constant equal to 1. For example, if ϕ is linear, we set $W_{i,-1} = X_i$. In addition, if $\phi(X_i)$ is a summation of smooth functions on each coordinate of X_i , we can take $W_i \in \mathbb{R}^{p_w}$ as the union of basis transformations, e.g., the polynomial basis, the Fourier basis, or the B spline basis. In particular, for $1 \leq j \leq p_x$, we construct a set of basis functions $\{b_{j,l}(\cdot)\}_{1 \leq l \leq M_j}$ to approximate the function defined on $X_{i,j}$ where $M_j \geq 1$ is the number of basis functions. We define $W_{i,-1} = (b_{1,1}(X_{i,1}), \dots, b_{1,M_1}(X_{i,1}), b_{2,1}(X_{i,2}), \dots, b_{2,M_2}(X_{i,2}), \dots, b_{p_x,1}(X_{i,p_x}), \dots, b_{p_x,M_{p_x}}(X_{i,p_x}))$.

When the IV is invalid, we assume that $g(Z_i, X_i)$ can be approximated by W_i as above together with a set of basis transformation involved with both IVs and baseline covariates:

$$\mathcal{V} := \{v_1(z, x), \dots, v_L(z, x)\} \quad \text{with} \quad v_j : \mathbb{R}^{p_z + p_x} \rightarrow \mathbb{R} \quad \text{for} \quad 1 \leq j \leq L. \quad (9)$$

We consider the single IV setting and give two examples of \mathcal{V} .

1. Polynomial basis: if $g(z, x) = h(z) + \phi(x)$, the IVs and baseline covariates affect the outcome separately. We can set $v_j(z, x) = v_j(z)$ and take $v_1(z), \dots, v_L(z)$ as (piecewise) polynomials of various orders.
2. Interaction basis: if $g(z, x) = \sum_{j=1}^p \alpha_j z \cdot x_j + \phi(x)$, we set $L = p_x$ and $v_j(z, x) = z \cdot x_j$ for $1 \leq j \leq L$.

In the remaining part of this subsection, we illustrate the main idea by assuming that the basis functions in (9) are given and known. In applications, empirical researchers might apply the domain knowledge and construct the pre-specified set of basis functions in (9). For example, [4] apply the Maimonides' rule to construct the IV as the transformation of a variable "enrollment" and then adjust with the possible violation form generated by a

linear, quadratic, or piecewise linear transformation of “enrollment”. In Section 3.4, we will present a data-dependent way to choose \mathcal{V} .

For the given v_1, \dots, v_L in (9), we define the violation matrix

$$V = \begin{pmatrix} V_1 & \cdots & V_n \end{pmatrix}^\top \in \mathbb{R}^{n \times L} \quad \text{with} \quad V_i = (v_1(Z_i, X_i), \dots, v_L(Z_i, X_i))^\top \quad \text{for } 1 \leq i \leq n. \quad (10)$$

We approximate $g(Z_i, X_i)$ by $V_i^\top \pi + W_i^\top \psi$ where

$$(\pi^\top, \psi^\top)^\top := \arg \min_{a \in \mathbb{R}^L, b \in \mathbb{R}^{pw}} \mathbf{E} [g(Z_i, X_i) - V_i^\top a - W_i^\top b]^2.$$

We define the approximation error vector $R(V) = (R_1(V), \dots, R_n(V)) \in \mathbb{R}^n$ with

$$R_i(V) := g(Z_i, X_i) - V_i^\top \pi - W_i^\top \psi, \quad \text{for } 1 \leq i \leq n.$$

We shall omit the dependence on V when there is no confusion. If $g(Z_i, X_i)$ is well approximated by a linear combination of W_i and V_i , then $\|R\|_2$ is close to zero or even $\|R\|_2 = 0$.

With the above basis approximation for $g(\cdot)$, we write the outcome model (1) as

$$Y_i = D_i \beta + V_i^\top \pi + W_i^\top \psi + R_i + \epsilon_i, \quad \text{for } 1 \leq i \leq n.$$

Applying the transformation Ω in (8) to the above model with data in \mathcal{A}_1 , we obtain

$$\widehat{Y}_{\mathcal{A}_1} = \widehat{f}_{\mathcal{A}_1} \beta + \widehat{V}_{\mathcal{A}_1} \pi + \widehat{W}_{\mathcal{A}_1} \psi + \widehat{R}_{\mathcal{A}_1} + \widehat{\epsilon}_{\mathcal{A}_1},$$

where $\widehat{Y}_{\mathcal{A}_1} = \Omega Y_{\mathcal{A}_1}$, $\widehat{f}_{\mathcal{A}_1} = \Omega D_{\mathcal{A}_1}$, $\widehat{V}_{\mathcal{A}_1} = \Omega V_{\mathcal{A}_1}$, $\widehat{W}_{\mathcal{A}_1} = \Omega W_{\mathcal{A}_1}$, $\widehat{R}_{\mathcal{A}_1} = \Omega R_{\mathcal{A}_1}$, and $\widehat{\epsilon}_{\mathcal{A}_1} = \Omega \epsilon_{\mathcal{A}_1}$. Based on the above expression, we estimate the effect β by

$$\widehat{\beta}_{\text{init}}(V) = \frac{\widehat{Y}_{\mathcal{A}_1}^\top P_{\widehat{V}_{\mathcal{A}_1}, \widehat{W}_{\mathcal{A}_1}}^\perp \widehat{f}_{\mathcal{A}_1}}{\widehat{f}_{\mathcal{A}_1}^\top P_{\widehat{V}_{\mathcal{A}_1}, \widehat{W}_{\mathcal{A}_1}}^\perp \widehat{f}_{\mathcal{A}_1}} = \frac{Y_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V) D_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V) D_{\mathcal{A}_1}} \quad \text{with} \quad \mathbf{M}_{\text{RF}}(V) = \Omega^\top P_{\widehat{V}_{\mathcal{A}_1}, \widehat{W}_{\mathcal{A}_1}}^\perp \Omega, \quad (11)$$

where V is the violation matrix defined in (10). This initial estimator suffers from a finite-sample bias

$$\frac{\epsilon_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V) \delta_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V) D_{\mathcal{A}_1}} \approx \frac{\text{Cov}(\delta_i, \epsilon_i) \cdot \text{Tr}[\mathbf{M}_{\text{RF}}(V)]}{D_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V) D_{\mathcal{A}_1}},$$

where the above approximation assumes the homoscedastic correlation $\text{Cov}(\epsilon_i, \delta_i | Z_i, X_i) = \text{Cov}(\epsilon_i, \delta_i)$, and $\text{Tr}[\mathbf{M}_{\text{RF}}(V)]$ denotes the trace of $\mathbf{M}_{\text{RF}}(V)$ defined in (11). Random forests might have a large degree of freedom, which leads to a relatively large value of $\text{Tr}[\mathbf{M}_{\text{RF}}(V)]$. This bias will be exacerbated for a small value of $D_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V) D_{\mathcal{A}_1}$, which is scaled to the

generalized IV strength measure introduced in the following (16). Consequently, $\widehat{\beta}_{\text{init}}(V)$ will be biased, and the corresponding confidence interval might not achieve the desired coverage for relatively small sample sizes or weak IVs; see also the numerical illustrations in Section 6.

To address this, we propose the following bias-corrected estimator,

$$\widetilde{\beta}_{\text{RF}}(V) = \widehat{\beta}_{\text{init}}(V) - \frac{\widehat{\text{Cov}}(\delta_i, \epsilon_i) \cdot \text{Tr}[\mathbf{M}_{\text{RF}}(V)]}{D_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V) D_{\mathcal{A}_1}},$$

where $\widehat{\beta}_{\text{init}}(V)$ is defined in (11) and the estimator of $\text{Cov}(\delta_i, \epsilon_i)$ is defined as,

$$\widehat{\text{Cov}}(\delta_i, \epsilon_i) = \frac{1}{n_1 - r} (D_{\mathcal{A}_1} - \widehat{f}_{\mathcal{A}_1})^\top P_{V_{\mathcal{A}_1}, W_{\mathcal{A}_1}}^\perp [Y_{\mathcal{A}_1} - D_{\mathcal{A}_1} \widehat{\beta}_{\text{init}}(V)],$$

with r denoting the rank of the matrix (V, W) . The correction in constructing $\widetilde{\beta}_{\text{RF}}(V)$ implicitly requires $\text{Cov}(\epsilon_i, \delta_i | Z_i, X_i) = \text{Cov}(\epsilon_i, \delta_i)$, which might limit practical applications.

To address this, we propose an estimator which is robust to heteroscedastic correlations,

$$\widehat{\beta}_{\text{RF}}(V) = \widehat{\beta}_{\text{init}}(V) - \frac{\sum_{i=1}^{n_1} [\mathbf{M}_{\text{RF}}(V)]_{ii} \widehat{\delta}_i [\widehat{\epsilon}(V)]_i}{D_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V) D_{\mathcal{A}_1}}, \quad (12)$$

with $\widehat{\delta}_{\mathcal{A}_1} = D_{\mathcal{A}_1} - \widehat{f}_{\mathcal{A}_1}$ and $\widehat{\epsilon}(V) = P_{V, W}^\perp [Y - D \widehat{\beta}_{\text{init}}(V)]$. We construct the confidence interval

$$\text{CI}_{\text{RF}}(V) = \left(\widehat{\beta}_{\text{RF}}(V) - z_{\alpha/2} \widehat{\text{SE}}(V), \widehat{\beta}_{\text{RF}}(V) + z_{\alpha/2} \widehat{\text{SE}}(V) \right), \quad (13)$$

where $\widehat{\beta}_{\text{RF}}(V)$ is defined in (12) and

$$\widehat{\text{SE}}(V) = \frac{\sqrt{\sum_{i=1}^{n_1} [\widehat{\epsilon}(V)]_i^2 [\mathbf{M}_{\text{RF}}(V) D_{\mathcal{A}_1}]_i^2}}{D_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V) D_{\mathcal{A}_1}} \quad \text{with} \quad \widehat{\epsilon}(V) = P_{V, W}^\perp [Y - D \widehat{\beta}_{\text{init}}(V)]. \quad (14)$$

Two important remarks are in order for our proposed methodology.

Remark 1 (TSLS with plugging-in Random Forests) One natural idea of combining TSLS with random forests is to replace the first stage of TSLS with random forests. We may calculate such a two stage estimator as $\widehat{\beta}_{\text{plug}}(V) = Y_{\mathcal{A}_1}^\top P_{V_{\mathcal{A}_1}, W_{\mathcal{A}_1}}^\perp \widehat{f}_{\mathcal{A}_1} / \widehat{f}_{\mathcal{A}_1}^\top P_{V_{\mathcal{A}_1}, W_{\mathcal{A}_1}}^\perp \widehat{f}_{\mathcal{A}_1}$, which is the least squares estimator of β by regressing $Y_{\mathcal{A}_1}$ on $\widehat{f}_{\mathcal{A}_1} = \Omega D_{\mathcal{A}_1}$, W and V . However, as shown in Section 6, the estimator $\widehat{\beta}_{\text{plug}}(V)$ suffers from a large estimation bias since the hat matrix Ω is not a projection matrix. The inconsistency of the TSLS with a non-linear first stage was pointed out in [3, 18]. One solution is to take the predicted value $\widehat{f}_{\mathcal{A}_1}^\top$ as the IV [18]. We generalize this idea to the invalid IV setting and estimate β by adjusting

possible violation forms and solving the estimating equation $\widehat{f}_{\mathcal{A}_1}^\top P_{V_{\mathcal{A}_1}, W_{\mathcal{A}_1}}^\perp (Y_{\mathcal{A}_1} - D_{\mathcal{A}_1}\beta) = 0$, which leads to the estimator $\widehat{\beta}_{\text{EE}}(V) = Y_{\mathcal{A}_1}^\top P_{V_{\mathcal{A}_1}, W_{\mathcal{A}_1}}^\perp \widehat{f}_{\mathcal{A}_1} / D_{\mathcal{A}_1}^\top P_{V_{\mathcal{A}_1}, W_{\mathcal{A}_1}}^\perp \widehat{f}_{\mathcal{A}_1}$. If the IV is strong after adjusting violation forms, $\widehat{\beta}_{\text{EE}}(V)$ is consistent. However, if the IVs become weak after adjusting the violation forms, we may even get negative values of $D_{\mathcal{A}_1}^\top P_{V_{\mathcal{A}_1}, W_{\mathcal{A}_1}}^\perp \widehat{f}_{\mathcal{A}_1}$, and the estimator $\widehat{\beta}_{\text{EE}}(V)$ tends to have an inflated bias and variance; see Section D.3 in the supplement.

Remark 2 (Overfitting bias from non-splitting methods) Without sample splitting, the TSCI estimator suffers from a significant bias due to overfitting. Similarly to Ω defined in (8), we define the transformation matrix Ω^{full} for random forests constructed with the full data. As a modification of (11), we consider the corresponding TSCI estimator,

$$\widehat{\beta}_{\text{full}}(V) = Y^\top \mathbf{M}_{\text{RF}}^{\text{full}}(V) D / D^\top \mathbf{M}_{\text{RF}}^{\text{full}}(V) D \quad \text{with} \quad \mathbf{M}_{\text{RF}}^{\text{full}}(V) = (\Omega^{\text{full}})^\top P_{\widetilde{V}, \widetilde{W}}^\perp \Omega^{\text{full}}, \quad (15)$$

where $\widetilde{V} = \Omega^{\text{full}} V$ and $\widetilde{W} = \Omega^{\text{full}} W$. The simulation results in Section 6 show that the estimator $\widehat{\beta}_{\text{full}}(V)$ in (15) suffers from a large bias and the resulting confidence does not achieve the desired coverage. This happens since the endogeneity of the treatment is not completely removed: the predicted values $\Omega^{\text{full}} D$ might be highly correlated with the errors δ due to the overfitting nature of the machine learning algorithm.

3.3 Generalized IV strength test

The IV strength is particularly important for identifying the treatment effect stably and the weak IV is a major concern in practical applications of IV-based methods [53, 2, 25]. The concentration parameter is a useful measure of the IV strength when the regression errors $\{\delta_i\}_{1 \leq i \leq n}$ in (3) are homoscedastic. For a given set of basis functions \mathcal{V} in (9), we introduce a generalized IV strength measure as,

$$\mu(V) := \frac{f_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V) f_{\mathcal{A}_1}}{\sum_{i \in \mathcal{A}_1} \text{Var}(\delta_i | X_i, Z_i) / |\mathcal{A}_1|}. \quad (16)$$

If $\text{Var}(\delta_i | X_i, Z_i) = \sigma_\delta^2$, then $\mu(V)$ is reduced to $f_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V) f_{\mathcal{A}_1} / \sigma_\delta^2$. A sufficiently large strength $\mu(V)$ will guarantee stable point and interval estimators defined in (12) and (13). Hence, we need to check whether $\mu(V)$ is sufficiently large. Since f is unknown, we estimate $f_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V) f_{\mathcal{A}_1}$ by its sample version $D_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V) D_{\mathcal{A}_1}$ and estimate $\mu(V)$ by

$$\widehat{\mu}(V) := D_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V) D_{\mathcal{A}_1} / \left[\|D_{\mathcal{A}_1} - \widehat{f}_{\mathcal{A}_1}\|_2^2 / n_1 \right]. \quad (17)$$

The consistency of our point estimator requires $\mu(V)$ to be much larger than $\text{Tr}[\mathbf{M}_{\text{RF}}(V)]$; see Condition (R2) in Section 5. Since [45, 53] suggest that the concentration parameter being larger than 10 as being “adequate”, we develop a bootstrap test for $\mu(V) \geq \max\{2\text{Tr}[\mathbf{M}_{\text{RF}}(V)], 10\}$. The test is based on the following error decomposition, $D_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V) D_{\mathcal{A}_1} - f_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V) f_{\mathcal{A}_1} = 2f_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V) \delta_{\mathcal{A}_1} + \delta_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V) \delta_{\mathcal{A}_1}$. We construct a probabilistic upper bound for the right-hand side of the above equation by the bootstrap method. For $1 \leq i \leq n_1$, we define $\widehat{\delta}_i = D_i - \widehat{f}_i$ and compute $\widetilde{\delta}_i = \widehat{\delta}_i - \bar{\mu}_\delta$ with $\bar{\mu}_\delta = \frac{1}{n_1} \sum_{i=1}^{n_1} \widehat{\delta}_i$. For $1 \leq l \leq L$, we generate $\delta_i^{[l]} = U_i^{[l]} \cdot \widetilde{\delta}_i$ for $1 \leq i \leq n_1$, where $\{U_i^{[l]}\}_{1 \leq i \leq n_1}$ are i.i.d. standard normal random variables. For $1 \leq l \leq L$, we compute

$$S^{[l]} = \left[2f_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V) \delta^{[l]} + (\delta^{[l]})^\top \mathbf{M}_{\text{RF}}(V) \delta^{[l]} \right] / \left[\|D_{\mathcal{A}_1} - \widehat{f}_{\mathcal{A}_1}\|_2^2 / n_1 \right],$$

and use $\mathcal{S}_{\alpha_0}(V)$ to denote the upper α_0 empirical quantile of $\{|S^{[l]}\|_{1 \leq l \leq L}\}$. We conduct the generalized IV strength test $\widehat{\mu(V)} \geq \max\{2\text{Tr}[\mathbf{M}_{\text{RF}}(V)], 10\} + \mathcal{S}_{\alpha_0}(V)$, with $\mathcal{S}_{\alpha_0}(V)$ capturing the estimation error $\widehat{\mu(V)} - \mu(V)$. We always use $\alpha_0 = 0.025$ throughout this paper. If the above generalized IV strength test is passed, the IV is claimed to be strong after adjusting for the violation matrix V defined in (10); otherwise, the IV is claimed to be weak after adjusting for the violation matrix V .

3.4 Data-dependent selection of \mathcal{V} and IV validity test

Our proposed TSCI estimator in (12) requires prior knowledge of the set of basis functions \mathcal{V} in (9) or the violation matrix V defined in (10). In the following, we consider the nested sets of basis functions $\mathcal{V}_0 \subset \mathcal{V}_1 \subset \dots \subset \mathcal{V}_Q$, where Q is a positive integer. We devise a data-dependent way to choose the best one among $\{\mathcal{V}_q\}_{0 \leq q \leq Q}$. The extension to the non-nested collection $\{\mathcal{V}_q\}_{0 \leq q \leq Q}$ is discussed in Section 3.6.

We define $\mathcal{V}_0 := \{0\}$ as the set of null function. For $q \geq 1$, define $\mathcal{V}_q := \{v_1(\cdot), \dots, v_{L_q}(\cdot)\}$ where $L_q \geq 1$ is the number of basis functions. For $0 \leq q \leq Q$, we define V_q as the violation matrix corresponding to \mathcal{V}_q , where the i -th row of V_q is defined as $(V_q)_i = (v_1(Z_i, X_i), \dots, v_{L_q}(Z_i, X_i))^\top \in \mathbf{R}^{L_q}$ for $1 \leq i \leq n$. We consider the single IV setting and give two examples of $\{\mathcal{V}_q\}_{1 \leq q \leq Q}$.

1. Polynomial violation: $\mathcal{V}_q = \{z, z^2, \dots, z^q\}$, for $1 \leq q \leq Q$.
2. Interaction violation: $\mathcal{V}_1 = \{z, z \cdot x_1, z \cdot x_2, \dots, z \cdot x_{p_x}\}$.

We implement the generalized IV strength test in Section 3.3 and define Q_{\max} as,

$$Q_{\max} = \arg \max_{q \geq 0} \left\{ \widehat{\mu(V_q)} \geq \max\{2\text{Tr}[\mathbf{M}_{\text{RF}}(V_q)], 10\} + \mathcal{S}_{\alpha_0}(V_q) \right\}, \quad (18)$$

where α_0 is set at 0.025 by default. For a larger violation matrix, we tend to adjust out more information and have relatively weaker IVs. Intuitively, $V_{Q_{\max}}$ denotes the largest violation matrix such that the IVs are still considered as having enough strength after adjusting for the corresponding violation matrix. With Q_{\max} , we shall choose among $\{V_q\}_{0 \leq q \leq Q_{\max}}$.

For any given $0 \leq q \leq Q_{\max}$, we apply the generalized estimator in (12) and construct

$$\widehat{\beta}_{\text{RF}}(V_q) = \widehat{\beta}_{\text{init}}(V_q) - \frac{\sum_{i=1}^{n_1} [\mathbf{M}_{\text{RF}}(V_q)]_{ii} \widehat{\delta}_i [\widehat{\epsilon}(V_{Q_{\max}})]_i}{D_{\mathcal{A}_1}^T \mathbf{M}_{\text{RF}}(V_q) D_{\mathcal{A}_1}}, \quad (19)$$

with $\widehat{\delta}_{\mathcal{A}_1} = D_{\mathcal{A}_1} - \widehat{f}_{\mathcal{A}_1}$, $\widehat{\beta}_{\text{init}}(\cdot)$ and $\mathbf{M}_{\text{RF}}(\cdot)$ defined in (11), and $\widehat{\epsilon}(\cdot)$ defined in (14).

We now test the difference between two violation matrices V_q and $V_{q'}$ for $0 \leq q \neq q' \leq Q_{\max}$. When both $R(V_q)$ and $R(V_{q'})$ are small, then

$$\widehat{\beta}_{\text{RF}}(V_{q'}) - \widehat{\beta}_{\text{RF}}(V_q) \approx \frac{f_{\mathcal{A}_1}^T \mathbf{M}_{\text{RF}}(V_{q'}) \epsilon_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^T \mathbf{M}_{\text{RF}}(V_{q'}) f_{\mathcal{A}_1}} - \frac{f_{\mathcal{A}_1}^T \mathbf{M}_{\text{RF}}(V_q) \epsilon_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^T \mathbf{M}_{\text{RF}}(V_q) f_{\mathcal{A}_1}}.$$

We estimate the conditional variance of the above term by

$$\begin{aligned} \widehat{H}(V_q, V_{q'}) &= \frac{\sum_{i=1}^{n_1} [\widehat{\epsilon}(V_{Q_{\max}})]_i^2 [\mathbf{M}_{\text{RF}}(V_{q'}) D_{\mathcal{A}_1}]_i^2}{[D_{\mathcal{A}_1}^T \mathbf{M}_{\text{RF}}(V_{q'}) D_{\mathcal{A}_1}]^2} + \frac{\sum_{i=1}^{n_1} [\widehat{\epsilon}(V_{Q_{\max}})]_i^2 [\mathbf{M}_{\text{RF}}(V_q) D_{\mathcal{A}_1}]_i^2}{[D_{\mathcal{A}_1}^T \mathbf{M}_{\text{RF}}(V_q) D_{\mathcal{A}_1}]^2} \\ &\quad - 2 \frac{\sum_{i=1}^{n_1} [\widehat{\epsilon}(V_{Q_{\max}})]_i^2 [\mathbf{M}_{\text{RF}}(V_{q'}) D_{\mathcal{A}_1}]_i [\mathbf{M}_{\text{RF}}(V_q) D_{\mathcal{A}_1}]_i}{[D_{\mathcal{A}_1}^T \mathbf{M}_{\text{RF}}(V_{q'}) D_{\mathcal{A}_1}] \cdot [D_{\mathcal{A}_1}^T \mathbf{M}_{\text{RF}}(V_q) D_{\mathcal{A}_1}]}, \end{aligned} \quad (20)$$

and define the following data-dependent way of choosing between V_q and $V_{q'}$,

$$\mathcal{C}^{\text{RF}}(V_q, V_{q'}) = \mathbf{1} \left(\left| \widehat{\beta}_{\text{RF}}(V_q) - \widehat{\beta}_{\text{RF}}(V_{q'}) \right| / \sqrt{\widehat{H}(V_q, V_{q'})} \geq z_{\alpha_0} \right) \quad (21)$$

where $\widehat{\beta}_{\text{RF}}(V_q)$ and $\widehat{\beta}_{\text{RF}}(V_{q'})$ are defined in (19) and z_{α_0} is the upper α_0 quantile of the standard normal random variable. If we set $q = 0$ in (21), then we are comparing the valid IV estimator and an estimator adjusting the user-specified violation matrix $V_{q'}$.

For $Q_{\max} \geq 2$, we generalize the pairwise comparison to multiple comparisons and then choose the best violation matrix. For $0 \leq q \leq Q_{\max} - 1$, we compare V_q to any larger violation matrix $V_{q'}$ with $q + 1 \leq q' \leq Q_{\max}$. For $0 \leq q \leq Q_{\max} - 1$, we define the test

$$\mathcal{C}^{\text{RF}}(V_q) = \mathbf{1} \left(\max_{q+1 \leq q' \leq Q_{\max}} \left[\left| \widehat{\beta}_{\text{RF}}(V_q) - \widehat{\beta}_{\text{RF}}(V_{q'}) \right| / \sqrt{\widehat{H}(V_q, V_{q'})} \right] \geq \widehat{\rho} \right), \quad (22)$$

where $\hat{\rho} > 0$ is a positive threshold to be determined. For completeness, we define $\mathcal{C}^{\text{RF}}(V_{Q_{\max}}) = 0$ as there is no larger matrix than $V_{Q_{\max}}$ that we might compare to. We interpret our test in (22) as follows: if none of the differences $\{|\hat{\beta}_{\text{RF}}(V_{q'}) - \hat{\beta}_{\text{RF}}(V_q)|\}_{q+1 \leq q' \leq Q_{\max}}$ is large, we conclude that the violation matrix V_q and W (approximately) generates the function $g(\cdot)$.

We shall choose $\hat{\rho} > 0$ in (22) by the bootstrap method. Note that the residues $\hat{\epsilon}(V_{Q_{\max}})$ are defined in (14) with $V = V_{Q_{\max}}$. For $1 \leq i \leq n_1$, we compute $\tilde{\epsilon}_i = [\hat{\epsilon}(V_{Q_{\max}})]_i - \bar{\mu}_\epsilon$ with $\bar{\mu}_\epsilon = \frac{1}{n_1} \sum_{i=1}^{n_1} [\hat{\epsilon}(V_{Q_{\max}})]_i$. For $1 \leq l \leq L$, we generate $\epsilon_i^{[l]} = U_i^{[l]} \cdot \tilde{\epsilon}_i$ for $1 \leq i \leq n_1$, where $\{U_i^{[l]}\}_{1 \leq i \leq n_1}$ are i.i.d. standard normal random variables. For $1 \leq l \leq L$, we compute

$$T^{[l]} = \max_{0 \leq q < q' \leq Q_{\max}} \frac{1}{\sqrt{\hat{H}(V_q, V_{q'})}} \left[\frac{f_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V_{q'}) \epsilon^{[l]}}{f_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V_{q'}) f_{\mathcal{A}_1}} - \frac{f_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V_q) \epsilon^{[l]}}{f_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V_q) f_{\mathcal{A}_1}} \right]. \quad (23)$$

We shall set $\hat{\rho} = \hat{\rho}(\alpha_0)$ to be the upper α_0 empirical quantile of $\{|T^{[l]}|\}_{1 \leq l \leq L}$, that is,

$$\hat{\rho}(\alpha_0) = \min \left\{ \rho \in \mathbb{R} : \frac{1}{L} \sum_{l=1}^L \mathbf{1}(|T^{[l]}| \leq \rho) \geq 1 - \alpha_0 \right\}, \quad (24)$$

where α_0 is set as 0.025 by default.

We choose the index $\hat{q}_c \in [0, Q_{\max}]$ as the smallest q such that the test statistics $\mathcal{C}^{\text{RF}}(V_q)$ is zero, that is, $\hat{q}_c = \arg \min_{0 \leq q \leq Q_{\max}} \{\mathcal{C}^{\text{RF}}(V_q) = 0\}$. The index \hat{q}_c is interpreted as follows: any violation matrix containing $V_{\hat{q}_c}$ as a submatrix does not lead to a substantially different estimator from that by $V_{\hat{q}_c}$. Here, the sub-index “c” represents the comparison as we compare different violation matrices to choose the best one. Note that \hat{q}_c must exist since $\mathcal{C}^{\text{RF}}(V_{Q_{\max}}) = 0$ by definition. In finite samples, there are chances that certain violations cannot be detected especially if the TSCI estimators given by $V_{\hat{q}_c}$ and $V_{\hat{q}_c+1}$ are not significantly different. We propose a more robust choice of the index as $\hat{q}_r = \min\{\hat{q}_c + 1, Q_{\max}\}$, where the sub index “r” denotes the robust selection; see Remark 4 for more discussions.

We summarize our proposed TSCI estimator and the IV validity test in Algorithm 1.

Algorithm 1 TSCI with random forests

Input: Data $X \in \mathbb{R}^{n \times p_x}$, $Z, D, Y \in \mathbb{R}^n$; Violation matrices $\{V_q\}_{q \geq 0}$;

Output: \hat{q}_c and \hat{q}_r ; $\hat{\beta}(V_{\hat{q}_c})$ and $\hat{\beta}(V_{\hat{q}_r})$; $\text{CI}(V_{\hat{q}_c})$ and $\text{CI}(V_{\hat{q}_r})$; IV Validity test I_{IV} .

- 1: Set $W = X$ or generate W as a user-specific transformation of X .
 - 2: Compute Q_{\max} as in (18);
 - 3: Compute $\hat{\epsilon}(V_{Q_{\max}})$ as in (14);
 - 4: Compute $\{\hat{\beta}_{RF}(V_q)\}_{0 \leq q \leq Q_{\max}}$ as in (19);
 - 5: Compute $\{\hat{H}(V_q, V_{q'})\}_{0 \leq q < q' \leq Q_{\max}}$ as in (20);
 - 6: Compute $\hat{\rho}$ as in (24);
 - 7: **if** $Q_{\max} \geq 1$ **then**
 - 8: Compute $\{\mathcal{C}^{RF}(V_q)\}_{0 \leq q \leq Q_{\max}-1}$ as in (22);
 - 9: **end if**
 - 10: Set $\mathcal{C}^{RF}(V_{Q_{\max}}) = 0$;
 - 11: Set $I_{IV} = 1$ if $\mathcal{C}^{RF}(V_0) = 1$ and $I_{IV} = 0$ otherwise; ▷ IV Invalidity test.
 - 12: Compute $\hat{q}_c = \min \{0 \leq q \leq Q_{\max} : \mathcal{C}^{RF}(V_q) = 0\}$; ▷ Comparison selection
 - 13: Compute $\hat{q}_r = \min \{\hat{q}_c + 1, Q_{\max}\}$; ▷ Robust selection
 - 14: Compute $\hat{\beta}_{RF}(V_{\hat{q}_c})$ and $\hat{\beta}_{RF}(V_{\hat{q}_r})$ as in (12) with $V = V_{\hat{q}_c}, V_{\hat{q}_r}$, respectively;
 - 15: Compute $\text{CI}_{RF}(V_{\hat{q}_c})$ and $\text{CI}_{RF}(V_{\hat{q}_r})$ as in (13) with $V = V_{\hat{q}_c}, V_{\hat{q}_r}$, respectively.
-

3.5 Finite-sample adjustment of uncertainty from data splitting

Our proposed TSCI estimator with random forests randomly splits the data into two subsamples. Even though our asymptotic theory in Section 5 is valid for any random sample splitting, the constructed point estimators and confidence intervals do vary with different sample splittings in finite samples. This randomness due to sample splittings has been also observed in double machine learning [19] and multi-splitting [41]. Following [19] and [41], we shall introduce a confidence interval which aggregates multiple confidence intervals due to different sample splittings.

Consider S random sample splittings and for the s -th splitting, we use $\hat{\beta}^s$ and $\widehat{\text{SE}}^s$ to denote the corresponding TSCI point and standard error estimators, respectively. Following Section 3.4 of [19], we introduce the median estimator and its estimated standard error as

$$\hat{\beta}^{\text{med}} = \text{median}\{\hat{\beta}^s\}_{1 \leq s \leq S} \quad \text{and} \quad \widehat{\text{SE}}^{\text{med}} = \text{median} \left\{ \sqrt{[\widehat{\text{SE}}^s]^2 + (\hat{\beta}^s - \hat{\beta}^{\text{med}})^2} \right\}_{1 \leq s \leq S},$$

and construct the median confidence interval as $(\widehat{\beta}^{\text{med}} - z_{\alpha/2} \widehat{\text{SE}}^{\text{med}}, \widehat{\beta}^{\text{med}} + z_{\alpha/2} \widehat{\text{SE}}^{\text{med}})$.

As an alternative, we follow the multi-splitting idea in [41], for any $\beta_0 \in \mathbb{R}$, we construct the p values $p^s(\beta_0) = 2(1 - \Phi(|\widehat{\beta}^s - \beta_0|/\widehat{\text{SE}}^s))$ for $1 \leq s \leq S$, where Φ is the CDF of the standard normal. We define the multi-splitting confidence interval as $\{\beta_0 \in \mathbb{R} : 2 \cdot \text{median}\{p^s(\beta_0)\}_{s=1}^S \leq \alpha\}$. See equation (2.2) of [41] for more details.

3.6 Multiple IVs and non-nested sets of basis functions

For a single IV, it is straightforward to generate the nested sets $\mathcal{V}_0 \subset \mathcal{V}_1 \subset \cdots \mathcal{V}_Q$; e.g., $\mathcal{V}_q = \{z, \dots, z^q\}$. When there are multiple IVs, there are more choices to specify the violation form. Moreover, the sequence of sets of basis functions $\{\mathcal{V}_q\}_{1 \leq q \leq Q}$ are not necessarily nested. For example, when we have two IVs z_1 and z_2 , we may set $\mathcal{V}_{q_1, q_2} = \{1, z_1, \dots, z_1^{q_1}, z_2, \dots, z_2^{q_2}\}$, and then $\{\mathcal{V}_{q_1, q_2}\}_{0 \leq q_1, q_2 \leq Q}$ are not nested. However, even in such a case, our proposed selection method in Section 3.4 is still applicable. Specifically, for any $0 \leq q_1, q_2 \leq Q$, we compare the estimator generated by \mathcal{V}_{q_1, q_2} to that by $\mathcal{V}_{q'_1, q'_2}$ with $q'_1 \geq q_1$ and $q'_2 \geq q_2$.

4 TSCI with general machine learning algorithms

We change the TSCI estimator with random forests proposed in Section 3 by incorporating other machine learning algorithms. The key step is to re-express formula (8) and write the first stage machine learning estimator as a linear transformation of $D_{\mathcal{A}_1}$, that is,

$$\widehat{f}_{\mathcal{A}_1} = \Omega D_{\mathcal{A}_1} \quad \text{for some matrix } \Omega \in \mathbb{R}^{n_1 \times n_1}. \quad (25)$$

We define a generalized transformation matrix

$$\mathbf{M}(V) = \Omega^\top P_{\widehat{V}_{\mathcal{A}_1}, \widehat{W}_{\mathcal{A}_1}}^\perp \Omega, \quad \text{with} \quad \widehat{V}_{\mathcal{A}_1} = \Omega V_{\mathcal{A}_1}, \quad \widehat{W}_{\mathcal{A}_1} = \Omega W_{\mathcal{A}_1}, \quad (26)$$

and modify the TSCI estimator with random forests in Section 3 by replacing $\mathbf{M}_{\text{RF}}(\cdot)$ with $\mathbf{M}(\cdot)$. In particular, we alter $\widehat{\beta}_{\text{RF}}(V)$ in (12) by

$$\widehat{\beta}(V) = \frac{Y_{\mathcal{A}_1}^\top \mathbf{M}(V) D_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\top \mathbf{M}(V) D_{\mathcal{A}_1}} - \frac{\sum_{i=1}^{n_1} [\mathbf{M}(V)]_{ii} \widehat{\delta}_i [\widehat{\epsilon}(V)]_i}{D_{\mathcal{A}_1}^\top \mathbf{M}(V) D_{\mathcal{A}_1}}. \quad (27)$$

We further replace (13) by

$$\text{CI}(V) = \left(\widehat{\beta}(V) - z_{\alpha/2} \widehat{\text{SE}}(V), \widehat{\beta}(V) + z_{\alpha/2} \widehat{\text{SE}}(V) \right), \quad \widehat{\text{SE}}(V) = \frac{\sqrt{\sum_{i=1}^{n_1} [\widehat{\epsilon}(V)]_i^2 [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2}}{D_{\mathcal{A}_1}^\top \mathbf{M}(V) D_{\mathcal{A}_1}} \quad (28)$$

where $\widehat{\epsilon}(V) = P_{V,W}^\perp \left[Y - D \frac{Y_{\mathcal{A}_1}^\top \mathbf{M}(V) D_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\top \mathbf{M}(V) D_{\mathcal{A}_1}} \right]$. With the above $\widehat{\epsilon}(V)$ and $\widehat{\delta}_{\mathcal{A}_1} = D_{\mathcal{A}_1} - \widehat{f}_{\mathcal{A}_1}$, we generalize $\widehat{\beta}_{\text{RF}}(V_q)$ in (19) as

$$\widehat{\beta}(V_q) = \frac{Y_{\mathcal{A}_1}^\top \mathbf{M}(V) D_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\top \mathbf{M}(V) D_{\mathcal{A}_1}} - \frac{\sum_{i=1}^{n_1} [\mathbf{M}(V_q)]_{ii} \widehat{\delta}_i [\widehat{\epsilon}(V_{Q_{\max}})]_i}{D_{\mathcal{A}_1}^\top \mathbf{M}(V_q) D_{\mathcal{A}_1}}. \quad (29)$$

We modify the definition of $\widehat{H}(V_q, V_{q'})$ in (20) and Algorithm 1 by replacing $\mathbf{M}_{\text{RF}}(\cdot)$ with $\mathbf{M}(\cdot)$, and define \widehat{q}_c and \widehat{q}_r accordingly. With the selected indices, we construct the point estimators and confidence intervals with the violation matrix $V_{\widehat{q}_c}$ or $V_{\widehat{q}_r}$.

In Sections 4.1, 4.2, and 4.3, we discuss three first-stage methods in the form of (25).

4.1 TSCI with boosting

In the following, we demonstrate how to express the L_2 boosting estimator [12, 13] in the form of (25). The boosting methods aggregate a sequence of base procedures $\{\widehat{g}^{[m]}(\cdot)\}_{m \geq 1}$. For $m \geq 1$, we construct the base procedure $\widehat{g}^{[m]}$ using the data in \mathcal{A}_2 and compute the estimated values given by the m -th base procedure $\widehat{g}_{\mathcal{A}_1}^{[m]} = (\widehat{g}^{[m]}(X_1, Z_1), \dots, \widehat{g}^{[m]}(X_{n_1}, Z_{n_1}))^\top$. With $\widehat{f}_{\mathcal{A}_1}^{[0]} = 0$ and $0 < \nu \leq 1$ as the step-length factor (the default being $\nu = 0.1$), we conduct the sequential updates,

$$\widehat{f}_{\mathcal{A}_1}^{[m]} = \widehat{f}_{\mathcal{A}_1}^{[m-1]} + \nu \widehat{g}_{\mathcal{A}_1}^{[m]} \quad \text{with} \quad \widehat{g}_{\mathcal{A}_1}^{[m]} = \mathcal{H}^{[m]}(D_{\mathcal{A}_1} - \widehat{f}_{\mathcal{A}_1}^{[m-1]}) \quad \text{for } m \geq 1,$$

where $\mathcal{H}^{[m]} \in \mathbb{R}^{n_1 \times n_1}$ is a hat matrix determined by the base procedures. In Section A.5 in the supplement, we give the exact expression of $\mathcal{H}^{[m]}$ for three different base procedures, including the pairwise regression, the pairwise thin plate, and the decision tree.

With the stopping time m_{stop} , the L_2 boosting estimator is $\widehat{f}^{\text{boo}} = \widehat{f}^{[m_{\text{stop}}]}$. We now compute the transformation matrix Ω . Set $\Omega^{[0]} = 0$ and for $m \geq 1$, we define

$$\widehat{f}_{\mathcal{A}_1}^{[m]} = \Omega^{[m]} D_{\mathcal{A}_1} \quad \text{with} \quad \Omega^{[m]} = \nu \mathcal{H}^{[m]} + (I - \nu \mathcal{H}^{[m]}) \Omega^{[m-1]}.$$

Define $\Omega^{\text{boo}} = \Omega^{[m_{\text{stop}}]}$ and write $\widehat{f}^{\text{boo}} = \Omega^{\text{boo}} D_{\mathcal{A}_1}$, which is the desired form in (25).

4.2 TSCI with deep neural network

In the following, we demonstrate how to calculate Ω for a deep neural network [29]. We define the first hidden layer as $H_{i,k}^{(1)} = \sigma \left(\omega_{k,0}^{(1)} + \sum_{l=1}^{p_z} \omega_{k,l}^{(1)} Z_{il} + \sum_{l=1}^{p_x} \omega_{k,l+p_z}^{(1)} X_{i,l} \right)$ for $1 \leq k \leq K_1$, where $\sigma(\cdot)$ is the activation function and $\{\omega_{k,l}^{(1)}\}_{1 \leq k \leq K_1, 1 \leq l \leq p_x + p_z}$ are parameters.

For $m \geq 2$, we define the m -th hidden layer as $H_{i,k}^{(m)} = \sigma\left(\omega_{k,0}^{(m)} + \sum_{l=1}^{K_{m-1}} \omega_{k,l}^{(m)} H_{i,l}^{(m-1)}\right)$ for $1 \leq k \leq K_m$, where $\{\omega_{k,l}^{(m)}\}_{1 \leq k \leq K_m, 1 \leq l \leq K_{m-1}}$ are unknown parameters. For given $M \geq 1$, we estimate the unknown parameters based on the data $\{X_i, Z_i, D_i\}_{i \in \mathcal{A}_2}$,

$$\left\{\widehat{\beta}, \{\widehat{\omega}^{(m)}\}_{1 \leq m \leq M}\right\} := \arg \min_{\{\beta_k\}_{0 \leq k \leq K_M}, \{\omega^{(m)}\}_{1 \leq m \leq M}} \sum_{i \in \mathcal{A}_2} \left(Y_i - \beta_0 - \sum_{k=1}^{K_M} \beta_k H_{i,k}^{(M)}\right)^2.$$

With $\{\widehat{\omega}^{(m)}\}_{1 \leq m \leq M}$, for $1 \leq m \leq M$ and $1 \leq i \leq n_1$, we define

$$\widehat{H}_{i,k}^{(m)} = \sigma\left(\widehat{\omega}_{k,0}^{(m)} + \sum_{l=1}^{K_{m-1}} \widehat{\omega}_{k,l}^{(m)} H_{i,l}^{(m-1)}\right) \quad \text{for } 1 \leq k \leq K_m$$

with $\widehat{H}_{i,k}^{(1)} = \sigma\left(\widehat{\omega}_{k,0}^{(1)} + \sum_{l=1}^{p_z} \widehat{\omega}_{k,l}^{(1)} Z_{il} + \sum_{l=1}^{p_x} \widehat{\omega}_{k,l+p_z}^{(1)} X_{il}\right)$ for $1 \leq k \leq K_1$. We use $\Omega^{\text{DNN}} = \widehat{H}^{(M)} \left([\widehat{H}^{(M)}]^\top \widehat{H}^{(M)}\right)^{-1} [\widehat{H}^{(M)}]^\top$ to denote the projection to the column space of the matrix $\widehat{H}^{(M)}$ and express the deep neural network estimator as $\Omega^{\text{DNN}} D_{\mathcal{A}_1}$. With $\widehat{V}_{\mathcal{A}_1} = \Omega^{\text{DNN}} V_{\mathcal{A}_1}$, and $\widehat{W}_{\mathcal{A}_1} = \Omega^{\text{DNN}} W_{\mathcal{A}_1}$, we define $\mathbf{M}_{\text{DNN}}(V) = [\Omega^{\text{DNN}}]^\top P_{\widehat{V}_{\mathcal{A}_1} \widehat{W}_{\mathcal{A}_1}}^\perp \Omega^{\text{DNN}}$, which is shown to be an orthogonal projection matrix in Lemma 5 in the supplement.

4.3 TSCI with basis approximation

As a simplification, we consider the additive model $\mathbf{E}(D_i | Z_i, X_i) = \gamma_1(Z_i) + \gamma_2(X_i)$, and assume that $\gamma_1(\cdot)$ can be well approximated by a set of basis functions $\{b_1(\cdot), \dots, b_M(\cdot)\}$. We define the matrix $B \in \mathbb{R}^{n \times M}$ with its i -th row $B_i = (b_1(Z_i), \dots, b_M(Z_i))^\top$. Without loss of generality, we approximate $\gamma_2(X_i)$ by $W_i \in \mathbb{R}^{p_w}$, which is generated by the same set of basis functions for $\phi(X_i)$. Define $\Omega^{\text{ba}} = P_{BW} \in \mathbb{R}^{n \times n}$ as the projection matrix to the space spanned by the columns of $B \in \mathbb{R}^{n \times k}$ and $W \in \mathbb{R}^{n \times p_w}$. We write the first-stage estimator as $\Omega^{\text{ba}} D$ and compute $\mathbf{M}_{\text{ba}}(V) = [\Omega^{\text{ba}}]^\top P_{\widehat{V}, W}^\perp \Omega^{\text{ba}}$ with $\widehat{V} = \Omega^{\text{ba}} V$. The transformation matrix $\mathbf{M}_{\text{ba}}(V)$ is a projection matrix with $M - \text{rank}(V) \leq \text{rank}[\mathbf{M}_{\text{ba}}(V)] \leq M$. When the basis number M is small and the degree of freedom $M + p_w$ is much smaller than n , sample splitting is not even needed for the basis approximation, which is different from the general machine learning algorithms.

5 Theoretical justification

We provide theoretical justifications for our proposed TSCI estimator with general machine learning algorithms as defined in (25) to (29).

5.1 Bias correction and asymptotic normality

We start with the required conditions. The first assumption is imposed on the regression errors in models (2) and (3) and the data $\{V_i, W_i, f_i\}_{1 \leq i \leq n}$ with $f_i = f(Z_i, X_i)$.

(R1) Conditioning on Z_i, X_i, ϵ_i and δ_i are sub-gaussian random variables, that is,

$$\sup_{Z_i, X_i} \max \{\mathbb{P}(|\epsilon_i| > t | Z_i, X_i), \mathbb{P}(|\delta_i| > t | Z_i, X_i)\} \leq \exp(-K^2 t^2/2),$$

where \sup_{Z_i, X_i} denotes the supremum taken over the support of the density of Z_i, X_i and K is the sub-gaussian norm. The random variables $\{\Psi_i, f_i\}_{1 \leq i \leq n_1}$ with $\Psi_i = (V_i^\top, W_i^\top)^\top$ and $f_i = f(Z_i, X_i)$ satisfy $\lambda_{\min}(\sum_{i=1}^{n_1} \Psi_i \Psi_i^\top / n_1) \geq c$, $\|\sum_{i=1}^{n_1} \Psi_i f_i / n_1\|_2 \leq C$, $\max_{1 \leq i \leq n_1} \{|f_i|, \|\Psi_i\|_2\} \leq C\sqrt{\log n_1}$, and $\|\sum_{i=1}^{n_1} \Psi_i [R(V)]_i / n_1\|_2 \leq C\|R(V)\|_\infty$, where $C > 0$ and $c > 0$ are constants independent of n and p . The matrix Ω defined in (25) satisfies $\lambda_{\max}(\Omega) \leq C$ for some positive constant $C > 0$.

The conditional sub-gaussian assumption is required to establish some concentration results. For the special case where ϵ_i and δ_i are independent of Z_i, X_i , it is sufficient to assume sub-gaussian errors ϵ_i and δ_i . The conditions on Ψ_i and f_i will be automatically satisfied with a high probability if $\mathbf{E}\Psi_i \Psi_i^\top$ is positive definite and Ψ_i and f_i are sub-gaussian random variables. The proof of Lemma 5 in the supplement shows that $\lambda_{\max}(\Omega) \leq 1$ for random forests, deep neural networks, and basis approximation methods.

The second assumption is imposed on the generalized IV strength $\mu(V)$ defined in (16). Throughout the paper, the asymptotics is taken as $n \rightarrow \infty$.

(R2) $f_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1}$ satisfies $f_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1} \rightarrow \infty$ and $f_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1} \gg \text{Tr}[\mathbf{M}(V)]$.

If $c \leq \text{Var}(\delta_i | X_i, Z_i) \leq C$ for some positive constants $C \geq c > 0$, then $\mu(V)$ is proportional to $f_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1}$. Our proposed test in Section 3.3 is designed to test the generalized IV strength, which is closely related to the above assumption.

Two important remarks are in order. Firstly, if we assume that the hat matrix Ω leads to an accurate estimator of f , then $\Omega f_{\mathcal{A}_1} \approx \Omega D_{\mathcal{A}_1} \approx f_{\mathcal{A}_1}$ and hence $f_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1} \approx f_{\mathcal{A}_1}^\top P_{\widehat{V}_{\mathcal{A}_1}, \widehat{W}_{\mathcal{A}_1}}^\perp f_{\mathcal{A}_1} = \sum_{i=1}^{n_1} [f_i - \tau^\top(\widehat{V}_i^\top, \widehat{W}_i^\top)]^2$ where \widehat{V} and \widehat{W} are defined in (26) and τ is the regression coefficient of regressing $f_{\mathcal{A}_1}$ on $\widehat{V}_{\mathcal{A}_1}$ and $\widehat{W}_{\mathcal{A}_1}$. (R2) essentially requires

$$\min_{1 \leq i \leq n_1} \mathbf{E} [f_i - \tau^\top(\widehat{V}_i^\top, \widehat{W}_i^\top)]^2 \gg \frac{\max\{1, \text{Tr}[\mathbf{M}(V)]\}}{n_1}, \quad \text{with } n_1 = \lceil 2n/3 \rceil,$$

where the expectation is taken conditioning on the data in \mathcal{A}_2 . This condition intuitively requires every of $\{f_i = f(Z_i, X_i)\}_{1 \leq i \leq n_1}$ to have enough variability after adjusting for the covariates and the violation form. More importantly, Condition (R2) does not require $\hat{f}_{\mathcal{A}_1}$ to be an accurate estimator of $f_{\mathcal{A}_1}$. Condition (R2) can be plausibly satisfied as long as the machine learning algorithms capture enough association between the treatment and the IVs. As an example, for TSCI with the basis approximation, $\text{Tr}[\mathbf{M}(V)] \leq r$ with r denoting the rank of the matrix (V, W) : Condition (R2) is satisfied as long as $f_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1} \gg r$. Note that $f_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1}$ can be of the scale n for the strong IV setting.

The following proposition shows that $\hat{\beta}_{\text{init}}(V)$ is consistent if Conditions (R1) and (R2) hold and the approximation errors $\{R_i(V)\}_{1 \leq i \leq n}$ with $R_i(V) = g(Z_i, X_i) - V_i^\top \pi - W_i^\top \phi$ are small.

Proposition 2 *Consider the model (2) and (3). Suppose that Conditions (R1) and (R2) hold and $f_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1} \gg \|R_{\mathcal{A}_1}(V)\|_2^2$, then $\hat{\beta}_{\text{init}}(V)$ defined in (11) with replacing $\mathbf{M}_{\text{RF}}(V)$ by $\mathbf{M}(V)$ satisfies $\hat{\beta}_{\text{init}}(V) \xrightarrow{p} \beta$.*

The condition $f_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1} \gg \|R_{\mathcal{A}_1}(V)\|_2^2$ will be satisfied if the g function is well approximated by the column space of V and W . In the extreme case, $R(V) = 0$ and this condition is automatically satisfied.

Theorem 1 *Suppose that the same conditions of Proposition 2 hold and*

$$\frac{\max_{1 \leq i \leq n_1} \sigma_i^2 \cdot [\mathbf{M}(V) f_{\mathcal{A}_1}]_i^2}{\sum_{i=1}^{n_1} \sigma_i^2 \cdot [\mathbf{M}(V) f_{\mathcal{A}_1}]_i^2} \rightarrow 0 \quad \text{with} \quad \sigma_i^2 = \mathbf{E}(\epsilon_i^2 | Z_i, X_i). \quad (30)$$

Then $\hat{\beta}(V)$ defined in (27) satisfies

$$\frac{1}{\text{SE}(V)} \left(\hat{\beta}(V) - \beta \right) = \mathcal{G}(V) + \mathcal{E}(V) \quad \text{with} \quad \text{SE}(V) = \frac{\sqrt{\sum_{i=1}^{n_1} \sigma_i^2 [\mathbf{M}(V) f_{\mathcal{A}_1}]_i^2}}{f_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1}}, \quad (31)$$

where $\mathcal{G}(V) \xrightarrow{d} N(0, 1)$ and there exist positive constants $t_0 > 0$ and $C > 0$ such that $\liminf_{n \rightarrow \infty} \mathbb{P}(|\mathcal{E}(V)| \leq CU_n(V)) \geq 1 - \exp(-t_0^2)$, with

$$U_n(V) = \frac{\sqrt{\log n} \cdot \eta_n(V) \cdot \text{Tr}[\mathbf{M}_{\text{RF}}(V)] + t_0 \sqrt{\text{Tr}([\mathbf{M}(V)]^2)} + \|R(V)\|_2}{\sqrt{f_{\mathcal{A}_1}^\top [\mathbf{M}(V)]^2 f_{\mathcal{A}_1}}}, \quad (32)$$

$$\eta_n(V) = \|f_{\mathcal{A}_1} - \hat{f}_{\mathcal{A}_1}\|_\infty + \left(|\beta - \hat{\beta}_{\text{init}}(V)| + \|R(V)\|_\infty + \frac{\log n}{\sqrt{n}} \right) (\sqrt{\log n} + \|f_{\mathcal{A}_1} - \hat{f}_{\mathcal{A}_1}\|_\infty). \quad (33)$$

The extra condition (30) is imposed such that not a single entry of the vector $\mathbf{M}(V)f_{\mathcal{A}_1}$ dominates all other entries, which is needed for verifying the Linderberg condition. The standard error $\text{SE}(V)$ defined in (31) relies on the generalized IV strength for a given violation matrix V . If $f_{\mathcal{A}_1}^\top \mathbf{M}(V)f_{\mathcal{A}_1}/n_1$ is of constant order, then $\text{SE}(V) \lesssim 1/\sqrt{n}$. In general, a larger violation matrix V will lead to a larger $\text{SE}(V)$ because the strength $f_{\mathcal{A}_1}^\top \mathbf{M}(V)f_{\mathcal{A}_1}$ decreases with a larger violation matrix V .

Remark 3 (Accuracy improvement with bias correction) We demonstrate the improvement of bias correction when $\text{Cov}(\epsilon_i, \delta_i \mid Z_i, X_i) = \text{Cov}(\epsilon_i, \delta_i)$. With this additional assumption, we show that (32) can be sharpened by replacing $\sqrt{\log n} \cdot \eta_n(V)$ with $\eta_n(V)$; see Theorem 5 in Section A.1 in the supplement. For the initial estimator $\widehat{\beta}_{\text{init}}(V)$, we can also establish $\frac{1}{\text{SE}(V)} (\widehat{\beta}_{\text{init}}(V) - \beta) = \mathcal{G}(V) + \widetilde{\mathcal{E}}(V)$, where

$$|\widetilde{\mathcal{E}}(V)| \leq \frac{\text{Cov}(\epsilon_i, \delta_i) \cdot \text{Tr}[\mathbf{M}_{\text{RF}}(V)] + \|R(V)\|_2}{\sqrt{f_{\mathcal{A}_1}[\mathbf{M}_{\text{RF}}(V)]^2 f_{\mathcal{A}_1}}} + \frac{\sqrt{\text{Tr}([\mathbf{M}_{\text{RF}}(V)]^2)}}{(f_{\mathcal{A}_1}[\mathbf{M}_{\text{RF}}(V)]^2 f_{\mathcal{A}_1})^{c_0}}. \quad (34)$$

When $\widehat{f}_{\mathcal{A}_1}$ accurately estimate $f_{\mathcal{A}_1}$ and g is well approximated by the column space of V and W , then $\eta_n(V) \ll \text{Cov}(\epsilon_i, \delta_i)$. By comparing (34) and (32), we observe our proposed method reduces the bias component $\text{Cov}(\epsilon_i, \delta_i) \cdot \text{Tr}[\mathbf{M}_{\text{RF}}(V)]/\sqrt{f_{\mathcal{A}_1}[\mathbf{M}_{\text{RF}}(V)]^2 f_{\mathcal{A}_1}}$ to $\eta_n(V) \cdot \text{Tr}[\mathbf{M}_{\text{RF}}(V)]/\sqrt{f_{\mathcal{A}_1}[\mathbf{M}_{\text{RF}}(V)]^2 f_{\mathcal{A}_1}}$. Even if the machine learning predicted values $\widehat{f}_{\mathcal{A}_1}$ do not estimate $f_{\mathcal{A}_1}$ well, then the bias correction will not lead to a worse estimator as long as the mild condition $\|f_{\mathcal{A}_1} - \widehat{f}_{\mathcal{A}_1}\|_2 \lesssim \sqrt{n}$ is satisfied.

For the statistical inference, we further impose a stronger IV strength condition.

(R2-Inf) $f_{\mathcal{A}_1}^\top [\mathbf{M}(V)]^2 f_{\mathcal{A}_1} \rightarrow \infty$, $f_{\mathcal{A}_1}^\top [\mathbf{M}(V)]^2 f_{\mathcal{A}_1} \gg \|R(V)\|_2^2$, and

$$f_{\mathcal{A}_1}^\top [\mathbf{M}(V)]^2 f_{\mathcal{A}_1} \gg \max \left\{ (\text{Tr}[\mathbf{M}(V)])^c, \log n \cdot \eta_n(V)^2 \cdot (\text{Tr}[\mathbf{M}(V)])^2 \right\},$$

where $c > 1$ is some positive constant and $\eta_n(\cdot)$ is defined in (33).

For the basis approximation and the deep neural network, we have $\mathbf{M}_{\text{ba}}^2 = \mathbf{M}_{\text{ba}}$ and $\mathbf{M}_{\text{DNN}}^2 = \mathbf{M}_{\text{DNN}}$. Hence, the condition (R2-Inf) is only slightly stronger than (R2) together with $f_{\mathcal{A}_1}^\top \mathbf{M}(V)f_{\mathcal{A}_1} \gg \|R_{\mathcal{A}_1}(V)\|_2^2$ required in Proposition 2. However, for random forests, we only have $f_{\mathcal{A}_1}^\top [\mathbf{M}(V)]^2 f_{\mathcal{A}_1} \leq f_{\mathcal{A}_1}^\top \mathbf{M}(V)f_{\mathcal{A}_1}$. Hence, in comparison to (R2), Condition (R2-Inf) is a stronger condition on the generalized IV strength.

Theorem 2 Consider the model (2) and (3). Suppose that Condition (R1), (R2-Inf) and (30) hold. Then $\widehat{\beta}(V)$ defined in (27) satisfies $\frac{1}{\text{SE}(V)} (\widehat{\beta}(V) - \beta) \xrightarrow{d} N(0, 1)$, where $\text{SE}(V)$ is defined in (31). If $\widehat{\text{SE}}(V)$ used in (28) satisfies $\widehat{\text{SE}}(V)/\text{SE}(V) \xrightarrow{p} 1$, then the confidence interval $\text{CI}(V)$ in (28) satisfies $\liminf_{n \rightarrow \infty} \mathbb{P}(\beta \in \text{CI}(V)) = 1 - \alpha$.

The consistency of $\widehat{\text{SE}}(V)$ is presented in Lemma 6 in the supplement.

5.2 Guarantee for Algorithm 1

We first present the asymptotic normality of the difference $\widehat{\beta}(V_q) - \widehat{\beta}(V_{q'})$, whose dominating component is $S^\top \epsilon_{\mathcal{A}_1}$ with

$$S = \left(\frac{1}{f_{\mathcal{A}_1}^\top \mathbf{M}(V_{q'}) f_{\mathcal{A}_1}} \mathbf{M}(V_{q'}) - \frac{1}{f_{\mathcal{A}_1}^\top \mathbf{M}(V_q) f_{\mathcal{A}_1}} \mathbf{M}(V_q) \right) f_{\mathcal{A}_1} \in \mathbb{R}^{n_1}. \quad (35)$$

Conditioning on the data in \mathcal{A}_2 and $\{X_i, Z_i\}_{i \in \mathcal{A}_1}$, $S^\top \epsilon_{\mathcal{A}_1}$ is of zero mean and variance

$$H(V_q, V_{q'}) = \frac{\sum_{i=1}^{n_1} \sigma_i^2 [\mathbf{M}(V_{q'}) f_{\mathcal{A}_1}]_i^2}{[f_{\mathcal{A}_1}^\top \mathbf{M}(V_{q'}) f_{\mathcal{A}_1}]^2} + \frac{\sum_{i=1}^{n_1} \sigma_i^2 [\mathbf{M}(V_q) f_{\mathcal{A}_1}]_i^2}{[f_{\mathcal{A}_1}^\top \mathbf{M}(V_q) f_{\mathcal{A}_1}]^2} - 2 \frac{\sum_{i=1}^{n_1} \sigma_i^2 [\mathbf{M}(V_{q'}) f_{\mathcal{A}_1}]_i [\mathbf{M}(V_q) f_{\mathcal{A}_1}]_i}{[f_{\mathcal{A}_1}^\top \mathbf{M}(V_{q'}) f_{\mathcal{A}_1}] \cdot [f_{\mathcal{A}_1}^\top \mathbf{M}(V_q) f_{\mathcal{A}_1}]}, \quad (36)$$

where $\sigma_i^2 = \mathbf{E}(\epsilon_i^2 \mid Z_i, X_i)$. In the following condition, we require that the difference's variance $H(V_q, V_{q'})$ dominates finite-sample approximation errors.

(R3) The variance $H(V_q, V_{q'})$ in (36) satisfies

$$\sqrt{H(V_q, V_{q'})} \gg \max_{V \in \{V_q, V_{q'}\}} \left\{ \frac{1}{\mu(V)} \left[1 + (1 + \sqrt{\log n} \cdot \eta_n(V_{Q_{\max}})) \cdot \text{Tr}[\mathbf{M}_{\text{RF}}(V)] \right] \right\},$$

with $\eta_n(\cdot)$ defined in (33). There exists a constant $c > 0$ such that $\text{Var}(\delta_i \mid Z_i, X_i) \geq c$.

If $\widehat{f}_{\mathcal{A}_1}$ accurately estimate $f_{\mathcal{A}_1}$ and g is well approximated by the column space of $V_{Q_{\max}}$ and W , then $\sqrt{\log n} \cdot \eta_n(V_{Q_{\max}}) \leq c$ for some positive constant $c > 0$. For the basis approximation or the DNN method with the homoscedastic errors $\text{Var}(\epsilon_i \mid Z_i, X_i) = \sigma_\epsilon^2$ and $\text{Var}(\delta_i \mid Z_i, X_i) = \sigma_\delta^2$, if $q \leq q'$, we have

$$H(V_q, V_{q'}) = \sigma_\epsilon^2 (1/f^\top \mathbf{M}(V_{q'}) f - 1/f^\top \mathbf{M}(V_q) f) \quad \text{and} \quad \mu(V_q) = f^\top \mathbf{M}(V_q) f / \sigma_\delta^2.$$

If we assume that $f^\top \mathbf{M}(V_{q'}) f = c_* f^\top \mathbf{M}(V_q) f$ for some $0 < c_* < 1$, then we have $H(V_q, V_{q'}) = \frac{1-c_*}{c_*} \frac{\sigma_\epsilon^2}{\sigma_\delta^2} \mu(V_q)$. In this case, Condition (R3) is satisfied if $\mu(V_q) \gg (\text{Tr}[\mathbf{M}(V_q)])^2$ and $\mu(V_{q'}) \gg (\text{Tr}[\mathbf{M}(V_{q'})])^2$, which are only slightly stronger than Condition (R2-Inf).

The following theorem establishes the asymptotic normality of $\widehat{\beta}(V_q) - \widehat{\beta}(V_{q'})$ under the null setting where both $R(V_q)$ and $R(V_{q'})$ are small. More theoretical results about comparing $\widehat{\beta}(V_q)$ and $\widehat{\beta}(V_{q'})$ can be found in Section A.2 in the supplement.

Theorem 3 Consider the model (2) and (3). Suppose that Condition (R1) holds, Condition (R2) holds for $V \in \{V_q, V_{q'}\}$, Condition (R3) holds, and the vector S defined in (35) satisfies $\max_{i \in \mathcal{A}_1} S_i^2 / (\sum_{i \in \mathcal{A}_1} S_i^2) \rightarrow 0$. If

$$\sqrt{H(V_q, V_{q'})} \gg \max_{V \in \{V_q, V_{q'}\}} \|R(V)\|_2 / \sqrt{\mu(V)}, \quad (37)$$

then we have $(\widehat{\beta}(V_q) - \widehat{\beta}(V_{q'})) / \sqrt{H(V_q, V_{q'})} \xrightarrow{d} N(0, 1)$, with $\widehat{\beta}(\cdot)$ defined in (29) and $H(V_q, V_{q'})$ defined in (36).

The condition (37) holds if the approximation errors $\|R(V_q)\|_2$ and $\|R(V_{q'})\|_2$ are small. When one of $\|R(V_q)\|_2$ and $\|R(V_{q'})\|_2$ is large, a leading component of the difference $\widehat{\beta}(V_q) - \widehat{\beta}(V_{q'})$ is

$$\mathcal{L}_n(V_q, V_{q'}) = \frac{1}{\sqrt{H(V_q, V_{q'})}} \left(\frac{D_{\mathcal{A}_1}^\top \mathbf{M}(V_q)[R(V_q)]_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\top \mathbf{M}(V_q) D_{\mathcal{A}_1}} - \frac{D_{\mathcal{A}_1}^\top \mathbf{M}(V_{q'})[R(V_{q'})]_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\top \mathbf{M}(V_{q'}) D_{\mathcal{A}_1}} \right). \quad (38)$$

With multiple normalized differences $\left\{ (\widehat{\beta}(V_q) - \widehat{\beta}(V_{q'})) / \sqrt{H(V_q, V_{q'})} \right\}_{0 \leq q < q' \leq Q_{\max}}$, we define the α_0 quantile for the maximum of multiple random error components in (35),

$$\mathbb{P} \left(\max_{0 \leq q < q' \leq Q_{\max}} \frac{1}{\sqrt{H(V_q, V_{q'})}} \left| \frac{f_{\mathcal{A}_1}^\top \mathbf{M}(V_{q'}) \epsilon_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^\top \mathbf{M}(V_{q'}) f_{\mathcal{A}_1}} - \frac{f_{\mathcal{A}_1}^\top \mathbf{M}(V_q) \epsilon_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^\top \mathbf{M}(V_q) f_{\mathcal{A}_1}} \right| \geq \rho(\alpha_0) \right) = \alpha_0. \quad (39)$$

We introduce the following condition for the selection among $\{\mathcal{V}_q\}_{0 \leq q \leq Q}$.

- (R4) For $\mathcal{V}_0 \subset \mathcal{V}_1 \subset \cdots \subset \mathcal{V}_Q$ and corresponding violation matrices $\{V_q\}_{0 \leq q \leq Q}$, there exists $Q^* \in \{0, 1, 2, \dots, Q\}$ such that $Q^* \leq Q_{\max}$ and $R(V_{Q^*}) = 0$. For any integer $q \in [0, Q^* - 1]$, there exists an integer $q' \in [q + 1, Q^*]$ such that

$$\mathcal{L}_n(V_q, V_{q'}) \geq A\rho(\alpha_0) \quad \text{for } A > 2, \quad (40)$$

where $\mathcal{L}_n(V_q, V_{q'})$ is defined in (38) and $\rho(\alpha_0)$ is defined in (39).

The above condition ensures that there exists \mathcal{V}_{Q^*} such that the function g is well approximated by the column space of V_{Q^*} and W . The well-separation condition (40) is

interpreted as follows: if g is not well approximated by the column space of V_q and W , then $\widehat{\beta}(V_q)$ has a large estimation bias. Note that $\mathcal{L}_n(V_q, V_{Q^*})$ defined in (38) is a measure of the bias of $\widehat{\beta}(V_q)$. The condition (40) requires that at least one of the bias measures $\{\mathcal{L}_n(V_q, V_{q'})\}_{q+1 \leq q' \leq Q^*}$ is sufficiently large; see more discussions in Remark 4.

The following theorem guarantees the coverage property for the CIs corresponding to $V_{\widehat{q}_c}$ and $V_{\widehat{q}_r}$ in Algorithm 1.

Theorem 4 *Consider the model (2) and (3). Suppose that Condition (R1) holds, Condition (R2-Inf) holds for $V \in \{V_q\}_{0 \leq q \leq Q_{\max}}$, Condition (R3) holds for any $0 \leq q < q' \leq Q_{\max}$, Condition (R4) holds, $\widehat{H}(V_q, V_{q'})/H(V_q, V_{q'}) \xrightarrow{p} 1$, and $\widehat{\rho}/\rho(\alpha_0) \xrightarrow{p} 1$ with $\widehat{\rho}$ and $\rho(\alpha_0)$ defined in (24) and (39) respectively. Then our proposed CI in Algorithm 1 with replacing $\mathbf{M}_{\text{RF}}(V)$ by $\mathbf{M}(V)$ satisfies*

$$\liminf_{n \rightarrow \infty} \mathbb{P}[\beta \in \text{CI}(V_{\widehat{q}})] \geq 1 - \alpha - 2\alpha_0 \quad \text{with } \widehat{q} = \widehat{q}_c \text{ or } \widehat{q}_r, \quad (41)$$

where α_0 is used in (39).

We shall remark that the CI based on $V_{\widehat{q}_r}$ is more robust. In particular, the well-separation condition (40) can be relaxed as follows: for any integer $q \in [0, Q^* - 2]$, there exists an integer $q' \in [q + 1, Q^*]$ such that (40) holds. That is, the statistical inference based on $V_{\widehat{q}_r}$ is still valid if we make mistakes for comparing V_{Q^*-1} and V_{Q^*} .

Remark 4 *If we apply TSCI with $\mathcal{V}_{Q_{\max}}$, this leads to a valid inference procedure without requiring the well-separation condition (40). However, such a confidence interval might be conservative by adjusting for a large violation matrix $V_{Q_{\max}}$.*

6 Simulation studies

We conduct a simulation study with the outcome model (1) and the treatment model (3). We set $\beta = 1$ as the treatment effect, fix $p = 20$ and vary the sample size n across $\{1000, 3000, 5000\}$. We generate $X_i^* \in \mathbb{R}^{p+1}$ following a multivariate normal distribution with zero mean and covariance matrix Σ where $\Sigma_{ij} = 0.5^{|i-j|}$ for $1 \leq i, j \leq p + 1$. With Φ denoting the standard normal cumulative distribution function, we define $X_{ij} = \Phi(X_{ij}^*)$ for $1 \leq j \leq p$. In the following, we focus on the continuous treatment and will investigate the performance of our proposed TSCI for binary treatments in Section D.2 in the supplement. We generate a continuous IV as $Z_i = 4(\Phi(X_{i,p+1}^*) - 0.5) \in (-2, 2)$ and consider the following two conditional mean models for the treatment,

- Model 1: $f(Z_i, X_i) = -\frac{25}{12} + Z_i + Z_i^2 + \frac{1}{8}Z_i^4 + Z_i \cdot (a \cdot \sum_{j=1}^5 X_{ij}) - 0.3 \cdot \sum_{j=1}^p X_{ij}$
- Model 2: $f(Z_i, X_i) = \sin(2\pi Z_i) + \frac{3}{2} \cos(2\pi Z_i) + Z_i \cdot (a \cdot \sum_{j=1}^5 X_{ij}) - 0.3 \cdot \sum_{j=1}^p X_{ij}$

The value a controls the interaction strength between Z_i and the first five variables of X_i , and when $a = 0$, the interaction term disappears. We will vary a across $\{0, 0.5, 1\}$. For Models 1 and 2, we consider two forms of the g function in (1): (a)Vio=1: $g(Z_i, X_i) = Z_i + 0.2 \cdot \sum_{j=1}^p X_{ij}$; (b)Vio=2: $g(Z_i, X_i) = Z_i + Z_i^2 - 1 + 0.2 \cdot \sum_{j=1}^p X_{ij}$.

To approximate the real data analysis in Section 7, we further generate a binary IV as $Z_i = \mathbf{1}(\Phi(X_{i,6}^*) > 0.6)$ and the covariates $X_{i,j} = X_{i,j}^*$ for $1 \leq i \leq n$ and $1 \leq j \leq 5$. We consider the following models for $f(Z_i, X_i)$ and $g(Z_i, X_i)$,

3. Model 3 (binary IV): $f(Z_i, X_i) = Z_i(1 + a \sum_{i=1}^4 X_{ij}(1 + X_{i,j+1})) - 0.3 \cdot \sum_{i=1}^5 X_{ij}$ and $g(Z_i, X_i) = Z_i + 0.5 \cdot Z_i \cdot (\sum_{i=1}^3 X_{ij})$.

In comparison to Models 1 and 2, the outcome model involves the interaction between Z_i and X_i while the treatment model involves a more complicated interaction term, whose strength is controlled by a . In Section D.1 in the supplement, we consider another setting of the binary IV with $p = 20$.

We consider two different distributions for the errors $\{(\delta_i, \epsilon_i)^\top\}_{1 \leq i \leq n}$ in (1) and (3),

- Error distribution 1. Generate $\{(\delta_i, \epsilon_i)^\top\}_{1 \leq i \leq n} \stackrel{i.i.d.}{\sim} N\left(\mathbf{0}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}\right)$.
- Error distribution 2. For $1 \leq i \leq n$, generate $\delta_i \sim N(0, Z_i^2 + 0.25)$ and

$$\epsilon_i = 0.6\delta_i + \sqrt{[1 - 0.6^2]/[0.86^4 + 1.38072^2]}(1.38072 \cdot \tau_{1,i} + 0.86^2 \cdot \tau_{2,i}), \quad (42)$$

where conditioning on Z_i , $\tau_{1,i} \sim N(0, Z_i^2 + 0.25)$, $\tau_{2,i} \sim N(0, 1)$, and $\tau_{1,i}$ and $\tau_{2,i}$ are independent of δ_i .

Error distribution 1 corresponds to homoscedastic while Error distribution 2 to heteroscedastic errors, where the generating method in (42) follows from [7].

For Models 1 and 2, we specify $\mathcal{V}_q = \{z, \dots, z^q\}$ for $1 \leq q \leq Q$; for Model 3, we specify $\mathcal{V}_1 = \{z, z \cdot x_1, \dots, z \cdot x_5\}$. We shall implement TSCI with random forests as detailed in Algorithm 1 and choose the best \mathcal{V}_q by the comparison and robust selection methods. As “naive” benchmarks, we compare different random forests based methods in the oracle setting, where the best \mathcal{V}_q used for approximating g is known a priori. We construct the

confidence intervals in the form of $(\hat{\beta} - z_{\alpha/2}\widehat{SE}, \hat{\beta} + z_{\alpha/2}\widehat{SE})$ and consider the following three specific estimators,

- RF-Init. Compute $\hat{\beta}$ as $\hat{\beta}_{\text{init}}$ in (11) and $\hat{\sigma}_\epsilon(V) = \sqrt{\|P_{V,W}^\perp[Y - D\hat{\beta}_{\text{init}}(V)]\|_2^2/(n - r)}$. Calculate $\widehat{SE} = \hat{\sigma}_\epsilon \sqrt{D_{\mathcal{A}_1}^\top [\mathbf{M}_{\text{RF}}(V)]^2 D_{\mathcal{A}_1}} / D_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V) D_{\mathcal{A}_1}$.
- RF-Plug. Compute $\hat{\beta} = \hat{\beta}_{\text{plug}}(V)$ and \widehat{SE} as

$$\hat{\beta}_{\text{plug}}(V) = \frac{Y_{\mathcal{A}_1}^\top P_{V,\mathcal{A}_1}^\perp \hat{f}_{\mathcal{A}_1}}{\hat{f}_{\mathcal{A}_1}^\top P_{V,\mathcal{A}_1}^\perp \hat{f}_{\mathcal{A}_1}} \quad \text{and} \quad \widehat{SE} = \sqrt{\frac{\|P_{V,W}^\perp(Y - \hat{\beta}_{\text{plug}}(V)D)\|_2^2}{|\mathcal{A}_1| \cdot \hat{D}_{\mathcal{A}_1}^\top P_{V,\mathcal{A}_1}^\perp \hat{D}_{\mathcal{A}_1}}}.$$

- RF-Full. Set $\hat{\beta} = \hat{\beta}_{\text{full}}(V)$ with $\hat{\beta}_{\text{full}}(V)$ defined in (15) and calculate

$$\widehat{SE} = \frac{\hat{\sigma}_{\text{full}}(V) \sqrt{D^\top [\mathbf{M}_{\text{RF}}^{\text{full}}(V)]^2 D}}{D^\top \mathbf{M}_{\text{RF}}^{\text{full}}(V) D} \quad \text{with} \quad \hat{\sigma}_{\text{full}}(V) = \sqrt{\|P_{V,W}^\perp(Y - \hat{\beta}_{\text{RF}}^{\text{full}}(V)D)\|_2^2/n}.$$

The code for implementing our proposed TSCI method and comparing with other random forests methods is available at <https://github.com/zijguo/TSCI-Replication>.

In table 1, we compare the coverage properties of the TSLS, the above three random forests methods, and our proposed TSCI with random forests and basis approximation. TSLS fails to have correct coverage probability 0.95 due to the existence of invalid IV; furthermore, we observe that RF-Full, and RF-Plug methods fail to achieve the desired coverage and the RF-Init method fails to have the desired coverage for vio=2. In contrast, our proposed TSCI achieves the desired coverage for the vio=1 setting and for the vio=2 setting with a relatively large interaction or large sample size.

We highlight two interesting observations. Firstly, \mathcal{V}_q selected by the comparison and robust methods tend to give valid coverage for vio=1; however, for vio=2, these methods work for more favorable settings, such as a larger interaction and sample size. For smaller interaction values and sample sizes, Q_{\max} is taken as 1 and hence the desired coverage is not achieved. However, the IV validity test can still report invalid IV in these challenging settings. Secondly, the RF-Init suffers from a larger bias in the vio=2 setting, which leads to its undercoverage. For vio=2, a larger set of basis functions ($\mathcal{V} = \{z, z^2\}$) is adjusted, and hence the generalized IV strength gets smaller, which results in a more significant estimation bias by the expression (34). For Model 1, we further report the mean absolute bias and the confidence interval length in Tables S5 and S6 in the supplement, respectively.

The observations for Model 2 are generally similar to those for Model 1, and Model 2 is slightly easier in the sense that the generalized IV strength remains relatively large

			TSCI-RF				TSCI-ba			TSLS	Other RF(oracle)		
vio	a	n	Oracle	Comp	Robust	Invalidity	Oracle	Comp	Robust		Init	Plug	Full
1	0.0	1000	0.93	0.93	0.93	1.00	0.94	0.93	0.91	0	0.93	0.02	0.79
		3000	0.95	0.95	0.95	1.00	0.95	0.94	0.95	0	0.94	0.00	0.74
		5000	0.96	0.95	0.95	1.00	0.94	0.93	0.95	0	0.96	0.00	0.66
	0.5	1000	0.96	0.96	0.96	1.00	0.95	0.94	0.94	0	0.96	0.03	0.80
		3000	0.95	0.93	0.92	1.00	0.95	0.94	0.95	0	0.95	0.00	0.68
		5000	0.95	0.93	0.93	1.00	0.95	0.94	0.93	0	0.95	0.00	0.60
	1.0	1000	0.94	0.93	0.90	1.00	0.96	0.95	0.95	0	0.93	0.02	0.79
		3000	0.94	0.94	0.93	1.00	0.96	0.95	0.95	0	0.94	0.00	0.60
		5000	0.94	0.93	0.93	1.00	0.95	0.93	0.94	0	0.94	0.00	0.47
2	0.0	1000	0.48	0.00	0.00	1.00	0.92	0.07	0.07	0	0.24	0.00	0.00
		3000	0.78	0.00	0.00	1.00	0.94	0.92	0.93	0	0.58	0.00	0.00
		5000	0.85	0.02	0.02	1.00	0.94	0.94	0.95	0	0.66	0.00	0.00
	0.5	1000	0.84	0.00	0.00	0.95	0.93	0.13	0.13	0	0.64	0.07	0.01
		3000	0.87	0.87	0.87	1.00	0.95	0.94	0.94	0	0.77	0.63	0.00
		5000	0.93	0.93	0.93	1.00	0.95	0.95	0.96	0	0.83	0.39	0.00
	1.0	1000	0.91	0.90	0.90	1.00	0.95	0.09	0.09	0	0.89	0.34	0.21
		3000	0.93	0.93	0.92	1.00	0.95	0.91	0.91	0	0.90	0.00	0.02
		5000	0.93	0.93	0.93	1.00	0.94	0.94	0.95	0	0.92	0.00	0.01

Table 1: Coverage (at nominal level 0.95) for Model 1 with Error distribution 1. The columns indexed with “TSCI-RF” “TSCI-ba” correspond to our proposed TSCI with the random forests and the basis approximation, where the columns indexed with “Oracle”, “Comp” and “Robust” correspond to the estimators with \mathcal{V}_q selected by the oracle knowledge, the comparison method, and the robust method. The column indexed with “Invalidity” reports the proportion of detecting the proposed IV as invalid. The columns index “TSLS” corresponds to the TSLS estimator. The columns indexed with “Init”, “Plug”, “Full” correspond to the RF estimators without bias-correction, the plug-in RF estimator and the no data-splitting RF estimator, with the oracle knowledge of the best \mathcal{V}_q .

			TSCI-RF									RF-Init		
			Bias			Length			Coverage			Invalidity	Bias	Coverage
vio	a	n	Orac	Comp	Robust	Orac	Comp	Robust	Orac	Comp	Robust		Orac	Orac
1	0.0	1000	0.03	0.03	0.03	0.31	0.31	0.31	0.83	0.83	0.83	1.00	0.09	0.69
		3000	0.00	0.00	0.00	0.13	0.13	0.12	0.95	0.95	0.94	1.00	0.01	0.94
		5000	0.00	0.00	0.00	0.08	0.08	0.08	0.96	0.96	0.95	1.00	0.01	0.93
	0.5	1000	0.01	0.01	0.01	0.18	0.18	0.17	0.92	0.92	0.92	1.00	0.04	0.84
		3000	0.00	0.00	0.00	0.08	0.08	0.08	0.95	0.95	0.95	1.00	0.01	0.93
		5000	0.00	0.00	0.00	0.06	0.06	0.06	0.96	0.96	0.95	1.00	0.00	0.94
	1.0	1000	0.00	0.00	0.00	0.16	0.16	0.15	0.93	0.94	0.92	1.00	0.02	0.90
		3000	0.00	0.00	0.00	0.08	0.08	0.08	0.94	0.94	0.94	1.00	0.01	0.92
		5000	0.00	0.00	0.00	0.06	0.06	0.06	0.94	0.94	0.94	1.00	0.01	0.92
2	0.0	1000	0.04	0.60	0.60	0.26	0.41	0.41	0.74	0.16	0.16	0.98	0.10	0.61
		3000	0.00	0.00	0.00	0.12	0.12	0.11	0.93	0.93	0.90	1.00	0.01	0.88
		5000	0.00	0.00	0.00	0.08	0.08	0.07	0.95	0.95	0.93	1.00	0.01	0.94
	0.5	1000	0.00	0.11	0.10	0.16	0.17	0.18	0.93	0.75	0.75	0.94	0.04	0.81
		3000	0.00	0.00	0.00	0.08	0.08	0.08	0.96	0.96	0.95	1.00	0.01	0.93
		5000	0.00	0.00	0.00	0.06	0.06	0.06	0.96	0.96	0.96	1.00	0.00	0.96
	1.0	1000	0.00	0.01	0.01	0.15	0.15	0.15	0.95	0.93	0.93	1.00	0.02	0.90
		3000	0.00	0.00	0.00	0.08	0.08	0.08	0.95	0.95	0.95	1.00	0.01	0.95
		5000	0.00	0.00	0.00	0.06	0.06	0.06	0.95	0.95	0.95	1.00	0.01	0.93

Table 2: Bias, length, and coverage (at nominal level 0.95) for Model 2 with Error distribution 2. The columns indexed with “TSCI-RF” corresponds to our proposed TSCI with the random forests, where the columns indexed with “Bias”, “Length”, and “Coverage” correspond to the absolute bias of the point estimator, the length and empirical coverage of the constructed confidence interval respectively. The columns indexed with “Oracle”, “Comp” and “Robust” correspond to the TSCI estimators with \mathcal{V}_q selected by the oracle knowledge, the comparison method, and the robust method. The column indexed with “Invalidity” reports the proportion of detecting the proposed IV as invalid. The columns indexed with “RF-Init” correspond to the RF estimators without bias-correction but with the oracle knowledge of the best \mathcal{V}_q .

even after adjusting for quadratic violation forms. For Model 2, we report the empirical coverage in Table S7 in the supplement.

We illustrate our proposed method for the settings with heteroscedastic errors. In Table 2, we report the performance of our proposed TSCI with random forests for Model 2. Our proposed TSCI achieves the desired coverage in most settings with vio=1 and vio=2 with $n \geq 3000$. For $n = 1000$ and vio=2, the TSCI with the oracle information of $\mathcal{V} = \{z, z^2\}$ tends to achieve reasonably good coverage, but the selection methods are not performing as well since Q_{\max} is set as 1 in many cases. However, we can still test the existence of invalid IVs in such settings. By comparing TSCI with random forests and RF-Init, we observe the improvement of our bias correction method, which leads to a better coverage property, especially for the relatively weak IV settings. The performance of our proposed estimator for Model 1 with Error distribution 2 is similar to those observed here, which is presented in Table S8 in the supplement.

		TSCI-RF									RF-Init		
		Bias			Length			Coverage			Invalidity	Bias	Coverage
a	n	Orac	Comp	Robust	Orac	Comp	Robust	Orac	Comp	Robust		Orac	Orac
0.25	1000	0.01	0.01	0.01	0.38	0.38	0.38	0.87	0.87	0.87	1	0.10	0.78
	3000	0.00	0.00	0.00	0.23	0.23	0.23	0.92	0.92	0.92	1	0.05	0.85
	5000	0.00	0.00	0.00	0.18	0.18	0.18	0.92	0.92	0.92	1	0.04	0.85
0.50	1000	0.00	0.00	0.00	0.22	0.22	0.22	0.92	0.92	0.92	1	0.03	0.89
	3000	0.00	0.00	0.00	0.12	0.12	0.12	0.93	0.93	0.93	1	0.02	0.90
	5000	0.00	0.00	0.00	0.09	0.09	0.09	0.95	0.95	0.95	1	0.01	0.91
0.75	1000	0.00	0.00	0.00	0.15	0.15	0.15	0.93	0.93	0.93	1	0.01	0.92
	3000	0.00	0.00	0.00	0.08	0.08	0.08	0.95	0.95	0.95	1	0.01	0.93
	5000	0.00	0.00	0.00	0.06	0.06	0.06	0.96	0.96	0.96	1	0.01	0.93

Table 3: Bias, length, and coverage (at nominal level 0.95) for Model 3 (binary IV) with Error distribution 2. The columns indexed with “TSCI-RF” corresponds to our proposed TSCI with the random forests, where the columns indexed with “Bias”, “Length”, and “Coverage” correspond to the absolute bias of the point estimator, the length and empirical coverage of the constructed confidence interval respectively. The columns indexed with “Oracle”, “Comp” and “Robust” correspond to the TSCI estimators with \mathcal{V}_q selected by the oracle knowledge, the comparison method, and the robust method. The column indexed with “Invalidity” reports the proportion of detecting the proposed IV as invalid. The columns indexed with “RF-Init” correspond to the RF estimators without bias-correction but with the oracle knowledge of the best \mathcal{V}_q .

We further illustrate our proposed method for Model 3 with a binary IV. Here the IV invalidity form involves the interaction between Z_i and X_i . We report the results in Table 3 with heteroscedastic errors (Error distribution 2). The results with the homoscedastic errors are similar and omitted here. The observations are coherent with those for Models 1 and 2, suggesting our TSCI method outputs confidence intervals with valid coverage and detect invalid IVs. We also observe that the bias correction is effective and improves the coverage when the interaction a is relatively small.

7 Real Data

We revisit the important economic question on the effect of education on income [15, 16]. We follow [14] and analyze the same data set from the National Longitudinal Survey of Young Men. The outcome is the log wages, and the treatment is the years of schooling. As argued in [14], there are various reasons that the treatment is endogenous. For example, the unobserved confounder “ability bias” may affect both the schooling years and wages, leading to the OLS estimator having a positive bias. [14] proposed an indicator for a nearby 4-year college in 1966 (`nearc4`) as the IV. As suggested in [14], we include the following baseline covariates: a quadratic function of potential experience (`exper`), a race indicator (`race`), and dummy variables for residence in a standard metropolitan statistical area in 1976 (`smsa`) and in 1966 (`smsa1966`), and the dummy variable for residence in the south in 1976 (`south`), and a full set of regional dummy variables. The data set consists of $n = 3010$ observations, which is made available by the R package `ivmodel` [30].

To implement algorithm 1, we include all baseline covariates as W and specify the interaction basis $\mathcal{V}_1 = \{\text{nearc4}, \text{nearc4} \cdot \text{exper}, \text{nearc4} \cdot \text{exper}^2, \text{nearc4} \cdot \text{race}, \text{nearc4} \cdot \text{sama}, \text{nearc4} \cdot \text{south}\}$, where interacted baseline covariates are the five most important variables reported by first stage random forests. After adjusting violation forms generated by \mathcal{V}_1 , our proposed TSCI method is robust even if the IV `nearc4` affects the outcome directly or through the interaction with other baseline covariates.

We report the point estimates and IV strength in Figure 2. Since the TSCI estimates depend on the specific sample splitting, we report 500 TSCI estimates due to 500 different splittings. On the leftmost panel of Figure 2, we compare the OLS, TSLS, and TSCI estimators, which uses $\mathcal{V}_{\hat{q}_c}$ reported in Algorithm 1. The median of these 500 TSCI estimators is 0.0592, 94% of the 500 TSCI estimates are smaller than the OLS estimate (0.0747), and 100% of TSCI estimates are smaller than the TSLS estimate (0.1315). In contrast to the

TSLS estimator, the TSCI estimators tend to be smaller than the OLS estimator, which helps correct the positive “ability bias”. We shall point out the IV strengths for the TSCI estimators are typically much larger than the TSLS (the concentration parameter is 13.26), which is illustrated in the rightmost panel of Figure 2.

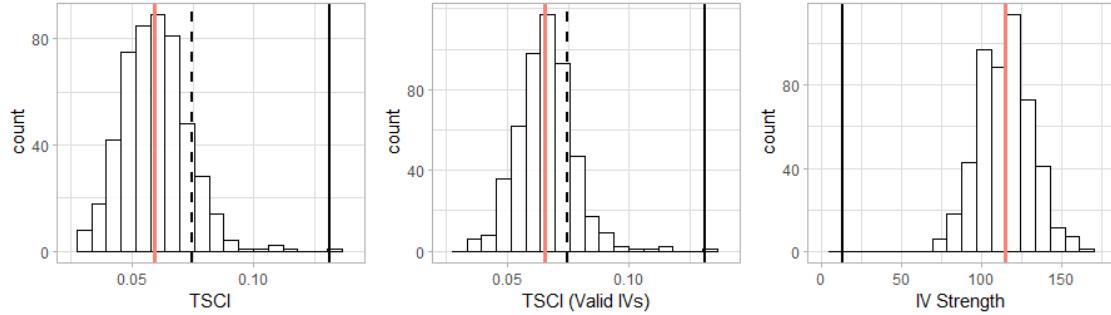


Figure 2: The leftmost and middle panels report the histograms of the TSCI (random forests) estimates, where the leftmost panel implements Algorithm 1 with the comparison method and the middle assumes the IV to be valid. On the leftmost and middle panels, the estimates differ due to the randomness of different 500 realized sample splittings; the solid red line corresponds to the median of the TSCI estimates while the solid and dashed black lines correspond to the TSLS and OLS estimates, respectively. The rightmost panel displays the histogram of the generalized IV strength (after adjusting for $\mathcal{V}_{\hat{q}_c}$) over the different 500 realized sample splittings; the solid red line denotes the median of all IV strength for TSCI while the solid black line denotes the IV strength of TSLS.

We apply our proposed IV validity test, and out of the total 500 splits, 302 data splits detect a significant difference between TSCI with \mathcal{V}_1 and TSCI assuming a valid IV. This test suggests that the proximity to the college might not be a good IV (and our TSCI is robust against its invalidity). We report the TSCI estimates (assuming valid IVs) in the middle panel of Figure 2 and observe that the magnitude of the treatment effect is further reduced by adjusting the possible violation form specified in \mathcal{V}_1 .

In Figure 3, we further compare different CIs. The TSLS CI is $(0.0238, 0.2392)$, which is much broader than the CIs by OLS and TSCI. This wide interval results from the relatively weak IV illustrated in Figure 2. The CI by TSCI with random forests varies with the specific sample splitting. We implement the methods described in Section 3.5 to adjust the finite-sample variation due to sample splitting. The two adjustment methods lead to similar results: the multi-split CI $(0.0282, 0.0898)$ and the median CI $(0.0271, 0.0915)$, both based on 500 different realized sample splittings. As reported in Figure 3, the CIs based

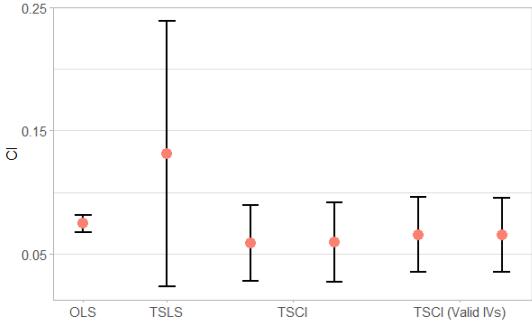


Figure 3: Confidence intervals (CIs) plots. The CIs with labels “OLS”, “TSLS”, “TSCI”, “TSCI (Valid IVs)” correspond to the CIs constructed by OLS, TSLS, the TSCI in Algorithm 1 with the comparison method, and the TSCI assuming valid IVs. For the two intervals indexed with “TSCI” or “TSCI (Valid IVs)”, the left is the multi-splitting CI while the right is the median CI.

on the TSCI estimator with random forests are pushed to the lower part of the wide CI by TSLS. The CIs by our proposed TSCI in Algorithm 1 tend to shift down in comparison to the TSCI assuming valid IVs.

8 Conclusion and discussion

We integrate modern machine learning algorithms into the framework of instrumental variable analysis. We devise a novel TSCI methodology which provides reliable causal conclusions even with invalid IVs. Our proposed generalized IV strength measure helps to understand when our proposed method is reliable and supports the basis functions’ selection in approximating the violation form of possibly invalid instruments. The current methodology is focused on inference for a linear and constant treatment effect. An interesting future research direction is about inference for heterogeneous treatment effects [5, 56].

9 Proofs

In Section 9.1, we establish Proposition 2. In Section 9.2, we establish Theorems 1 and 2. We prove Theorem 3 in Section A.2. We shall use \mathcal{O} to denote the set of random variables $\{Z_i, X_i\}_{1 \leq i \leq n}$ and $\{D_i\}_{i \in \mathcal{A}_2}$. We introduce the following lemma about the concentration of

quadratic forms, which is Theorem 1.1 in [47].

Lemma 1 (Hanson-Wright inequality) *Let $\epsilon \in \mathbb{R}^n$ be a random vector with independent sub-gaussian components ϵ_i with zero mean and sub-gaussian norm K . Let A be an $n \times n$ matrix. For every $t \geq 0$, $\mathbf{P}(|\epsilon^\top A \epsilon - \mathbf{E}\epsilon^\top A \epsilon| > t) \leq 2 \exp\left[-c \min\left(\frac{t^2}{K^4 \|A\|_F^2}, \frac{t}{K^2 \|A\|_2}\right)\right]$.*

Note that $2\epsilon^\top A\delta = (\epsilon + \delta)^\top A(\epsilon + \delta) - \epsilon^\top A\epsilon - \delta^\top A\delta$. If both ϵ_i and δ_i are sub-gaussian, we apply the union bound, Lemma 1 for both ϵ and δ and then establish,

$$\mathbf{P}(|\epsilon^\top A\delta - \mathbf{E}\epsilon^\top A\delta| > t) \leq 6 \exp\left[-c \min\left(\frac{t^2}{K^4 \|A\|_F^2}, \frac{t}{K^2 \|A\|_2}\right)\right]. \quad (43)$$

We define the conditional covariance matrices $\Lambda, \Sigma^\delta, \Sigma^\epsilon \in \mathbb{R}^{n_1 \times n_1}$ as $\Lambda = \mathbf{E}(\delta_{\mathcal{A}_1} \epsilon_{\mathcal{A}_1}^\top | X_{\mathcal{A}_1}, Z_{\mathcal{A}_1})$, $\Sigma^\delta = \mathbf{E}(\delta_{\mathcal{A}_1} \delta_{\mathcal{A}_1}^\top | X_{\mathcal{A}_1}, Z_{\mathcal{A}_1})$, and $\Sigma^\epsilon = \mathbf{E}(\epsilon_{\mathcal{A}_1} \epsilon_{\mathcal{A}_1}^\top | X_{\mathcal{A}_1}, Z_{\mathcal{A}_1})$. For any $i, j \in \mathcal{A}_1$ and $i \neq j$, we have $\Lambda_{i,j} = \mathbf{E}[\delta_j \mathbf{E}(\epsilon_i | X_{\mathcal{A}_1}, Z_{\mathcal{A}_1}, \delta_j) | X_{\mathcal{A}_1}, Z_{\mathcal{A}_1}]$. Since $\mathbf{E}(\epsilon_i | X_{\mathcal{A}_1}, Z_{\mathcal{A}_1}, \delta_j) = \mathbf{E}(\epsilon_i | X_i, Z_i) = 0$ for $i \neq j$, we have $\Lambda_{i,j} = 0$ and Λ is a diagonal matrix. Similarly, we can show Σ^δ and Σ^ϵ are diagonal matrices. The conditional sub-gaussian condition in (R1) implies that $\max_{1 \leq j \leq n_1} \max\{|\Lambda_{j,j}|, |\Sigma_{j,j}^\delta|, |\Sigma_{j,j}^\epsilon|\} \leq K^2$.

9.1 Proof of Proposition 2

Recall that we are analyzing $\widehat{\beta}_{\text{init}}(V)$ defined in (11) with replacing $\mathbf{M}_{\text{RF}}(V)$ by $\mathbf{M}(V)$. We decompose the estimation error $\widehat{\beta}_{\text{init}}(V) - \beta$ as

$$[\epsilon_{\mathcal{A}_1}^\top \mathbf{M}(V) \delta_{\mathcal{A}_1} + \epsilon_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1} + R_{\mathcal{A}_1}(V)^\top \mathbf{M}(V) D_{\mathcal{A}_1}] / D_{\mathcal{A}_1}^\top \mathbf{M}(V) D_{\mathcal{A}_1}. \quad (44)$$

The following Lemma controls the key components of the above decomposition, whose proof can be found in Section C.1 in the supplement.

Lemma 2 *Under Condition (R1), with probability larger than $1 - \exp(-c \min\{t_0^2, t_0\})$ for some positive constants $c > 0$ and $t_0 > 0$,*

$$\left| \epsilon_{\mathcal{A}_1}^\top \mathbf{M}(V) \delta_{\mathcal{A}_1} - \text{Tr}[\mathbf{M}(V) \Lambda] \right| \leq t_0 K^2 \sqrt{\text{Tr}([\mathbf{M}(V)]^2)}, \quad \left| \epsilon_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1} \right| \leq t_0 K \sqrt{f_{\mathcal{A}_1}^\top [\mathbf{M}(V)]^2 f_{\mathcal{A}_1}} \quad (45)$$

$$\left| \frac{D_{\mathcal{A}_1}^\top \mathbf{M}(V) D_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1}} - 1 \right| \lesssim \frac{K^2 \text{Tr}[\mathbf{M}(V)] + t_0 K^2 \sqrt{\text{Tr}([\mathbf{M}(V)]^2)} + t_0 K \sqrt{f_{\mathcal{A}_1}^\top [\mathbf{M}(V)]^2 f_{\mathcal{A}_1}}}{f_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1}}. \quad (46)$$

where $\Lambda = \mathbf{E}(\delta_{\mathcal{A}_1} \epsilon_{\mathcal{A}_1}^\top | X_{\mathcal{A}_1}, Z_{\mathcal{A}_1})$.

Define $\tau_n = f_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1}$. Together with (62) and the generalized IV strength condition (R2), we apply (46) with $t_0 = \tau_n^{1/2-c_0}$ for some $0 < c_0 < 1/2$. Then there exists positive constants $c > 0$ and $C > 0$ such that with probability larger than $1 - \exp(-c\tau_n^{1/2-c_0})$,

$$\left| \frac{D_{\mathcal{A}_1}^\top \mathbf{M}(V) D_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1}} - 1 \right| \leq C \frac{K^2 \text{Tr}[\mathbf{M}(V)]}{f_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1}} + C \frac{K + K^2}{(f_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1})^{c_0}} \leq 0.1. \quad (47)$$

Together with (44), we establish

$$|\widehat{\beta}_{\text{init}}(V) - \beta| \lesssim \frac{|\epsilon_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1}|}{f_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1}} + \frac{|\epsilon_{\mathcal{A}_1}^\top \mathbf{M}(V) \delta_{\mathcal{A}_1}|}{f_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1}} + \sqrt{\frac{R_{\mathcal{A}_1}(V)^\top \mathbf{M}(V) R_{\mathcal{A}_1}(V)}{f_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1}}}. \quad (48)$$

By applying the decomposition (48) and the upper bounds (45), and (45) with $t_0 = (f_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1})^{1/2-c_0}$, we establish that, conditioning on \mathcal{O} , with probability larger than $1 - \exp(-c\tau_n^{1/2-c_0})$ for some positive constants $c > 0$ and $c_0 \in (0, 1/2)$,

$$|\widehat{\beta}_{\text{init}}(V) - \beta| \lesssim \frac{K^2 \text{Tr}[\mathbf{M}(V)]}{f_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1}} + \frac{K + K^2}{(f_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1})^{c_0}} + \frac{\|R_{\mathcal{A}_1}(V)\|_2}{\sqrt{f_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1}}}. \quad (49)$$

The above concentration bound, the assumption (R2), and the assumption $f_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1} \gg \|R_{\mathcal{A}_1}(V)\|_2^2$ imply $\mathbb{P}(|\widehat{\beta}_{\text{init}}(V) - \beta| \geq c | \mathcal{O}) \rightarrow 0$. Then for any constant $c > 0$, we have $\mathbb{P}(|\widehat{\beta}_{\text{init}}(V) - \beta| \geq c) = \mathbf{E}(\mathbb{P}(|\widehat{\beta}_{\text{init}}(V) - \beta| \geq c | \mathcal{O}))$. We apply the bounded convergence theorem and establish $\mathbb{P}(|\widehat{\beta}_{\text{init}}(V) - \beta| \geq c) \rightarrow 0$ and hence $\widehat{\beta}_{\text{init}}(V) \xrightarrow{p} \beta$.

9.2 Proof of Theorems 1 and 2

In the following, we establish Theorem 1, which implies Theorem 2 together with Condition (R2-Inf). We start with the following error decomposition,

$$\widehat{\beta}_{\text{RF}}(V) - \beta = \frac{\epsilon_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1} + R_{\mathcal{A}_1}(V)^\top \mathbf{M}(V) D_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\top \mathbf{M}(V) D_{\mathcal{A}_1}} - \frac{\text{Err}_1 + \text{Err}_2}{D_{\mathcal{A}_1}^\top \mathbf{M}(V) D_{\mathcal{A}_1}}, \quad (50)$$

with $\text{Err}_1 = \sum_{1 \leq i \neq j \leq n_1} [\mathbf{M}(V)]_{ij} \delta_i \epsilon_j$ and $\text{Err}_2 = \sum_{i=1}^{n_1} [\mathbf{M}(V)]_{ii} (f_i - \widehat{f}_i) (\epsilon_i + [\widehat{\epsilon}(V)]_i - \epsilon_i) + \sum_{i=1}^{n_1} [\mathbf{M}(V)]_{ii} \delta_i ([\widehat{\epsilon}(V)]_i - \epsilon_i)$. Define

$$\mathcal{G}(V) = \frac{1}{\text{SE}(V)} \frac{\epsilon_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\top \mathbf{M}(V) D_{\mathcal{A}_1}}, \quad \mathcal{E}(V) = \frac{1}{\text{SE}(V)} \frac{R_{\mathcal{A}_1}(V)^\top \mathbf{M}(V) D_{\mathcal{A}_1} - \text{Err}_1 - \text{Err}_2}{D_{\mathcal{A}_1}^\top \mathbf{M}(V) D_{\mathcal{A}_1}}.$$

Then the decomposition (50) implies (31). We apply (47) together with the generalized IV strength condition (R2) and establish $\frac{D_{\mathcal{A}_1}^\top \mathbf{M}(V) D_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1}} \xrightarrow{p} 1$. Condition (30) implies the Linderberg condition. Hence, we establish $\mathcal{G}(V) | \mathcal{O} \xrightarrow{d} N(0, 1)$, and $\mathcal{G}(V) \xrightarrow{d} N(0, 1)$.

Since $\mathbf{E}(\text{Err}_1 \mid \mathcal{O}) = 0$, we apply the similar argument as in (45) and obtain that, conditioning on \mathcal{O} , with probability larger than $1 - \exp(-ct_0^2)$ for some constants $c > 0$ and $t_0 > 0$, $|\text{Err}_1| \leq t_0 K^2 \sqrt{\sum_{1 \leq i \neq j \leq n_1} [\mathbf{M}(V)]_{ij}^2} \leq t_0 K^2 \sqrt{\text{Tr}([\mathbf{M}(V)]^2)}$. Then we establish

$$\mathbb{P}\left(|\text{Err}_1| \geq t_0 K^2 \sqrt{\text{Tr}([\mathbf{M}(V)]^2)}\right) = \mathbf{E}\left[\mathbb{P}\left(|\text{Err}_1| \geq t_0 K^2 \sqrt{\text{Tr}([\mathbf{M}(V)]^2)} \mid \mathcal{O}\right)\right] \leq \exp(-ct_0^2). \quad (51)$$

We introduce the following Lemma to control Err_2 , whose proof is presented in Section C.2 in the supplement.

Lemma 3 *If Condition (R2) holds, then with probability larger than $1 - n_1^{-c}$,*

$$\max_{1 \leq i \leq n_1} |[\widehat{\epsilon}(V)]_i - \epsilon_i| \leq C \sqrt{\log n} \left(\|R(V)\|_\infty + |\beta - \widehat{\beta}_{\text{init}}(V)| + \frac{\log n}{\sqrt{n}} \right).$$

The sub-gaussianity of ϵ_i implies that, with probability larger than $1 - n_1^{-c}$ for some positive constant $c > 0$, $\max_{1 \leq i \leq n_1} |\epsilon_i| + \max_{1 \leq i \leq n_1} |\delta_i| \leq C \sqrt{\log n}$ for some positive constant $C > 0$. Since $\mathbf{M}(V)$ is positive definite, we have $[\mathbf{M}(V)]_{ii} \geq 0$ for any $1 \leq i \leq n_1$. We have

$$|\text{Err}_2| \lesssim \sum_{i=1}^{n_1} [\mathbf{M}(V)]_{ii} \left[|f_i - \widehat{f}_i| \left(\sqrt{\log n} + |[\widehat{\epsilon}(V)]_i - \epsilon_i| \right) + \sqrt{\log n} |[\widehat{\epsilon}(V)]_i - \epsilon_i| \right]. \quad (52)$$

By Lemma 3, we have $|\text{Err}_2| \lesssim \sqrt{\log n} \cdot \eta_n(V) \cdot \text{Tr}[\mathbf{M}(V)]$ with $\eta_n(V)$ defined in (33). Together with (51), we establish $\liminf_{n \rightarrow \infty} \mathbb{P}(|\mathcal{E}(V)| \leq CU_n(V)) \geq 1 - \exp(-t_0^2)$.

9.3 Proofs of Theorem 3

By conditional sub-gaussianity and $\text{Var}(\delta_i \mid Z_i, X_i) \geq c$, we have $c \leq \text{Var}(\delta_i \mid Z_i, X_i) \leq C$. Hence, we have $\mu(V) \asymp f_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1}$. With $H(V_q, V_{q'})$ defined in (36), we have

$$\frac{\widehat{\beta}(V_q) - \widehat{\beta}(V_{q'})}{\sqrt{H(V_q, V_{q'})}} = \mathcal{G}_n(V_q, V_{q'}) + \mathcal{L}_n(V_q, V_{q'}) + \frac{1}{\sqrt{H(V_q, V_{q'})}} \left| \widetilde{\mathcal{E}}(V_q) - \widetilde{\mathcal{E}}(V_{q'}) \right|, \quad (53)$$

where $\mathcal{L}_n(V_q, V_{q'})$ is defined in (38),

$$\mathcal{G}_n(V_q, V_{q'}) = \frac{1}{\sqrt{H(V_q, V_{q'})}} \left(\frac{f_{\mathcal{A}_1}^\top \mathbf{M}(V_q) \epsilon_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\top \mathbf{M}(V_q) D_{\mathcal{A}_1}} - \frac{f_{\mathcal{A}_1}^\top \mathbf{M}(V_{q'}) \epsilon_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\top \mathbf{M}(V_{q'}) D_{\mathcal{A}_1}} \right), \quad (54)$$

$$\widetilde{\mathcal{E}}(V_q) = \frac{\sum_{i=1}^{n_1} [\mathbf{M}(V_q)]_{ii} \widehat{\delta}_i [\widehat{\epsilon}(V_{Q_{\max}})]_i - \delta_{\mathcal{A}_1}^\top \mathbf{M}(V_q) \epsilon_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\top \mathbf{M}(V_q) D_{\mathcal{A}_1}}.$$

The remaining proof relies on the following lemma, whose proof can be found at Section C.3 in the supplement.

Lemma 4 Suppose that Conditions (R1), (R2), and (R3) hold, then

$$\frac{1}{\sqrt{H(V_q, V_{q'})}} \left| \tilde{\mathcal{E}}(V_q) - \tilde{\mathcal{E}}(V_{q'}) \right| \xrightarrow{p} 0, \quad \text{and} \quad \mathcal{G}_n(V_q, V_{q'}) \xrightarrow{d} N(0, 1).$$

By applying (47), we establish that, conditioning on \mathcal{O} , with probability larger than $1 - \exp(-c\tau_n^{1/2-c_0})$, $|\mathcal{L}_n(V_q, V_{q'})| \lesssim \frac{1}{\sqrt{H(V_q, V_{q'})}} \left(\frac{\|[R(V_q)]_{\mathcal{A}_1}\|_2}{\sqrt{\mu(V_q)}} + \frac{\|[R(V_{q'})]_{\mathcal{A}_1}\|_2}{\sqrt{\mu(V_{q'})}} \right)$. Then we apply the above inequality and the condition (37) and establish that $|\mathcal{L}_n(V_q, V_{q'})| \xrightarrow{p} 0$. Together with the decomposition (53), and Lemma 4, we establish the asymptotic limiting distribution of $\hat{\beta}(V_q) - \hat{\beta}(V_{q'})$ in Theorem 3.

References

- [1] Takeshi Amemiya. The nonlinear two-stage least-squares estimator. *Journal of econometrics*, 2(2):105–110, 1974. [4](#)
- [2] Donald Andrews and James H Stock. Inference with weak instruments, 2005. [2](#), [13](#)
- [3] Joshua D Angrist and Alan B Krueger. Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic perspectives*, 15(4):69–85, 2001. [12](#)
- [4] Joshua D Angrist and Victor Lavy. Using maimonides’ rule to estimate the effect of class size on scholastic achievement. *The Quarterly journal of economics*, 114(2):533–575, 1999. [10](#)
- [5] Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016. [34](#)
- [6] Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019. [4](#)
- [7] Paul A Bekker and Federico Crudu. Jackknife instrumental variable estimation with heteroskedasticity. *Journal of econometrics*, 185(2):332–342, 2015. [27](#)
- [8] Alexandre Belloni, Daniel Chen, Victor Chernozhukov, and Christian Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429, 2012. [4](#)

- [9] Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. *Advances in neural information processing systems*, 32, 2019. [4](#)
- [10] Jack Bowden, George Davey Smith, and Stephen Burgess. Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. *International journal of epidemiology*, 44(2):512–525, 2015. [3](#)
- [11] Jack Bowden, George Davey Smith, Philip C Haycock, and Stephen Burgess. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic epidemiology*, 40(4):304–314, 2016. [3](#)
- [12] Peter Bühlmann and Torsten Hothorn. Boosting algorithms: Regularization, prediction and model fitting. *Statistical science*, 22(4):477–505, 2007. [19](#)
- [13] Peter Bühlmann and Bin Yu. Sparse boosting. *Journal of Machine Learning Research*, 7(6), 2006. [19](#)
- [14] David Card. Using geographic variation in college proximity to estimate the return to schooling, 1993. [32](#)
- [15] David Card. The causal effect of education on earnings. In *Handbook of labor economics*, volume 3, pages 1801–1863. Elsevier, 1999. [32](#)
- [16] David Card. Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica*, 69(5):1127–1160, 2001. [32](#)
- [17] John C Chao and Norman R Swanson. Consistent estimation with a large number of weak instruments. *Econometrica*, 73(5):1673–1692, 2005. [2](#)
- [18] Jiafeng Chen, Daniel L Chen, and Greg Lewis. Mostly harmless machine learning: learning optimal instruments in linear iv models. *arXiv preprint arXiv:2011.06158*, 2020. [12](#)
- [19] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning. *The Econometrics Journal*, 21(1), 2018. [4, 17](#)

- [20] Corinne Emmenegger and Peter Bühlmann. Regularizing double machine learning in partially linear endogenous models. *Electronic Journal of Statistics*, 15(2):6461–6543, 2021. 4
- [21] Peter J Green and Bernard W Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach*. Crc Press, 1993. 4
- [22] Zijian Guo. Post-selection problems for causal inference with invalid instruments: A solution using searching and sampling. *arXiv preprint arXiv:2104.06911*, 2021. 3
- [23] Zijian Guo, Hyunseung Kang, T Tony Cai, and Dylan S Small. Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):793–815, 2018. 3, 6
- [24] Chirok Han. Detecting invalid instruments using l1-gmm. *Economics Letters*, 101(3):285–287, 2008. 3
- [25] Christian Hansen, Jerry Hausman, and Whitney Newey. Estimation with many instrumental variables. *Journal of Business & Economic Statistics*, 26(4):398–422, 2008. 2, 13
- [26] Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the econometric society*, pages 1029–1054, 1982. 4
- [27] Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep iv: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, pages 1414–1423. PMLR, 2017. 4
- [28] Paul W Holland. Causal inference, path analysis and recursive structural equations models. *ETS Research Report Series*, 1988(1):i–50, 1988. 6
- [29] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013. 19
- [30] Hyunseung Kang, Yang Jiang, Qingyuan Zhao, and Dylan S Small. Ivmodel: an r package for inference and sensitivity analysis of instrumental variables models with one endogenous variable. *Observational Studies*, 7(2):1–24, 2021. 32

- [31] Hyunseung Kang, Anru Zhang, T Tony Cai, and Dylan S Small. Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association*, 111(513):132–144, 2016. [3](#), [6](#)
- [32] Harry H Kelejian. Two-stage least squares and econometric systems linear in parameters but nonlinear in the endogenous variables. *Journal of the American Statistical Association*, 66(334):373–374, 1971. [4](#)
- [33] Frank Kleibergen. Pivotal statistics for testing structural parameters in instrumental variables regression. *Econometrica*, 70(5):1781–1803, 2002. [2](#)
- [34] Michal Kolesár, Raj Chetty, John Friedman, Edward Glaeser, and Guido W Imbens. Identification and inference with many invalid instruments. *Journal of Business & Economic Statistics*, 33(4):474–484, 2015. [3](#)
- [35] Arthur Lewbel. Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models. *Journal of Business & Economic Statistics*, 30(1):67–80, 2012. [4](#)
- [36] Arthur Lewbel. Identification and estimation using heteroscedasticity without instruments: The binary endogenous regressor case. *Economics Letters*, 165:10–12, 2018. [4](#)
- [37] Sai Li and Zijian Guo. Causal inference for nonlinear outcome models with possibly invalid instrumental variables. *arXiv preprint arXiv:2010.09922*, 2020. [3](#)
- [38] Yi Lin and Yongho Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590, 2006. [8](#), [9](#)
- [39] Ruiqi Liu, Zuofeng Shang, and Guang Cheng. On deep instrumental variables estimate. *arXiv preprint arXiv:2004.14954*, 2020. [4](#)
- [40] Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999, 2006. [8](#), [9](#)
- [41] Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 2009. [17](#), [18](#)

- [42] Marcelo J Moreira. A conditional likelihood ratio test for structural models. *Econometrica*, 71(4):1027–1048, 2003. [2](#)
- [43] Whitney K Newey. Efficient instrumental variables estimation of nonlinear models. *Econometrica: Journal of the Econometric Society*, pages 809–837, 1990. [4](#)
- [44] Whitney K Newey and Frank Windmeijer. Generalized method of moments with many weak moment conditions. *Econometrica*, 77(3):687–719, 2009. [2](#)
- [45] Thomas J Rothenberg. Approximating the distributions of econometric estimators and test statistics. *Handbook of econometrics*, 2:881–935, 1984. [14](#)
- [46] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974. [6](#)
- [47] Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013. [35](#)
- [48] John D Sargan. The estimation of economic relationships using instrumental variables. *Econometrica: Journal of the Econometric Society*, pages 393–415, 1958. [4](#)
- [49] Dylan S Small. Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association*, 102(479):1049–1058, 2007. [6](#)
- [50] Jerzy Splawa-Neyman, Dorota M Dabrowska, and TP Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472, 1990. [6](#)
- [51] Douglas Staiger and James H Stock. Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586, 1997. [2](#)
- [52] James H Stock and Jonathan H Wright. Gmm with weak identification. *Econometrica*, 68(5):1055–1096, 2000. [2](#)
- [53] James H Stock, Jonathan H Wright, and Motohiro Yogo. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 20(4):518–529, 2002. [2, 13, 14](#)
- [54] Eric Tchetgen Tchetgen, BaoLuo Sun, and Stefan Walter. The genius approach to robust mendelian randomization inference. *Statistical Science*, 36(3):443–464, 2021. [4](#)

- [55] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010. [10](#), [11](#), [14](#)
- [56] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. [34](#)
- [57] Frank Windmeijer, Helmut Farbmacher, Neil Davies, and George Davey Smith. On the use of the lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association*, 114(527):1339–1350, 2019. [3](#), [6](#)
- [58] Frank Windmeijer, Xiaoran Liang, Fernando P Hartwig, and Jack Bowden. The confidence interval method for selecting valid instrumental variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(4):752–776, 2021. [3](#)
- [59] Liyuan Xu, Yutian Chen, Siddarth Srinivasan, Nando de Freitas, Arnaud Doucet, and Arthur Gretton. Learning deep features in instrumental variable regression. *arXiv preprint arXiv:2010.07154*, 2020. [4](#)

A Additional Discussions

A.1 Homoscedastic correlation

We present a simplified version of Theorem 1 by assuming homoscedastic correlation. We shall only present the results for the TSCI with random forests but the extension to the general machine learning methods is straightforward.

Theorem 5 *Consider the model (2) and (3) with $\text{Cov}(\epsilon_i, \delta_i \mid Z_i, X_i) = \text{Cov}(\epsilon_i, \delta_i)$. Suppose that Conditions (R1) and (R2) hold, then*

$$\left| \widehat{\text{Cov}}(\delta_i, \epsilon_i) - \text{Cov}(\delta_i, \epsilon_i) \right| \leq \eta_n^0, \quad (55)$$

where

$$\eta_n^0(V) = \frac{\|f_{\mathcal{A}_1} - \widehat{f}_{\mathcal{A}_1}\|_2}{\sqrt{n}} + \sqrt{\frac{\log n}{n}} + \left(|\beta - \widehat{\beta}_{\text{init}}(V)| + \frac{\|R(V)\|_2}{\sqrt{n}} \right) \left(1 + \frac{\|f_{\mathcal{A}_1} - \widehat{f}_{\mathcal{A}_1}\|_2}{\sqrt{n}} \right). \quad (56)$$

Furthermore, if we assume (30) holds, then $\widetilde{\beta}_{\text{RF}}(V)$ defined in (12) satisfies

$$\frac{1}{\text{SE}(V)} \left(\widetilde{\beta}_{\text{RF}}(V) - \beta \right) = \mathcal{G}(V) + \mathcal{E}(V), \quad (57)$$

where $\text{SE}(V)$ is defined in (31), $\mathcal{G}(V) \xrightarrow{d} N(0, 1)$ and there exist positive constants $C > 0$ and $t_0 > 0$ such that

$$\liminf_{n \rightarrow \infty} \mathbb{P} \left(|\mathcal{E}(V)| \leq C \frac{\eta_n^0(V) \cdot \text{Tr}[\mathbf{M}_{\text{RF}}(V)] + \|R(V)\|_2 + t_0 \sqrt{\text{Tr}([\mathbf{M}_{\text{RF}}(V)]^2)}}{\sqrt{f_{\mathcal{A}_1}[\mathbf{M}_{\text{RF}}(V)]^2 f_{\mathcal{A}_1}}} \right) \geq 1 - \exp(-t_0^2), \quad (58)$$

with $\eta_n^0(V)$ defined in (56).

As a remark, if $\|R(V)\|_2/\sqrt{n} \rightarrow 0$, $\widehat{\beta}_{\text{init}}(V) \xrightarrow{p} \beta$, and $\|f_{\mathcal{A}_1} - \widehat{f}_{\mathcal{A}_1}\|_2/\sqrt{n} \rightarrow 0$, then we have $\eta_n^0(V) \rightarrow 0$.

A.2 Theoretical justification for comparing two violation matrices

As a generalization of (21) and (20), we define the following data-dependent way of choosing between V_q and $V_{q'}$

$$\mathcal{C}(V_q, V_{q'}) = \begin{cases} 0 & \text{if } \frac{|\widehat{\beta}(V_q) - \widehat{\beta}(V_{q'})|}{\sqrt{\widehat{H}(V_q, V_{q'})}} \leq z_{\alpha_0}, \\ 1 & \text{otherwise} \end{cases}, \quad (59)$$

where z_{α_0} is the upper α_0 quantile of the standard normal random variable and

$$\begin{aligned}\widehat{H}(V_q, V_{q'}) &= \frac{\sum_{i=1}^{n_1} [\widehat{\epsilon}(V_{Q_{\max}})]_i^2 [\mathbf{M}(V_{q'}) D_{\mathcal{A}_1}]_i^2}{[D_{\mathcal{A}_1}^\top \mathbf{M}(V_{q'}) D_{\mathcal{A}_1}]^2} + \frac{\sum_{i=1}^{n_1} [\widehat{\epsilon}(V_{Q_{\max}})]_i^2 [\mathbf{M}(V_q) D_{\mathcal{A}_1}]_i^2}{[D_{\mathcal{A}_1}^\top \mathbf{M}(V_q) D_{\mathcal{A}_1}]^2} \\ &\quad - 2 \frac{\sum_{i=1}^{n_1} [\widehat{\epsilon}(V_{Q_{\max}})]_i^2 [\mathbf{M}(V_q) D_{\mathcal{A}_1}]_i [\mathbf{M}(V_{q'}) D_{\mathcal{A}_1}]_i}{[D_{\mathcal{A}_1}^\top \mathbf{M}(V_{q'}) D_{\mathcal{A}_1}] \cdot [D_{\mathcal{A}_1}^\top \mathbf{M}(V_q) D_{\mathcal{A}_1}]}.\end{aligned}$$

We generalize the multiple comparison in (22) as

$$\mathcal{C}(V_q) = \begin{cases} 0 & \text{if } \max_{q+1 \leq q' \leq Q_{\max}} \frac{|\widehat{\beta}(V_q) - \widehat{\beta}(V_{q'})|}{\sqrt{\widehat{H}(V_q, V_{q'})}} \leq \widehat{\rho} \\ 1 & \text{otherwise} \end{cases}, \quad (60)$$

where $\widehat{\rho} = \widehat{\rho}(\alpha_0)$ is defined in (24) with $\mathbf{M}_{RF}(\cdot)$ replaced by $\mathbf{M}(\cdot)$. As a direct application of Theorem 3, we establish the type I error control in the following corollary.

Corollary 1 Suppose that Conditions of Theorem 3 hold and $\widehat{H}(V_q, V_{q'})/H(V_q, V_{q'}) \xrightarrow{p} 1$, then the type I error of $\mathcal{C}(V_q, V_{q'})$ in (59) with corresponding α_0 is controlled as $\limsup_{n \rightarrow \infty} \mathbf{P}(\mathcal{C}(V_q, V_{q'}) = 1) = 2\alpha_0$.

The following corollary characterizes the power of the comparison test $\mathcal{C}(V_q, V_{q'})$ in (59).

Corollary 2 Suppose that the Conditions of Theorem 3 hold, $\widehat{H}(V_q, V_{q'})/H(V_q, V_{q'}) \xrightarrow{p} 1$, and $\mathcal{L}_n(V_q, V_{q'})$ defined in (38) satisfies $\mathcal{L}_n(V_q, V_{q'}) \xrightarrow{p} \mathcal{L}^*(V_q, V_{q'})$ for $\mathcal{L}^*(V_q, V_{q'}) \in \mathbb{R} \cup \{-\infty, \infty\}$, then the test $\mathcal{C}(V_q, V_{q'})$ in (59) with corresponding α_0 satisfies,

$$\lim_{n \rightarrow \infty} \mathbf{P}(\mathcal{C}(V_q, V_{q'}) = 1) = 1 - \Phi(z_{\alpha_0} - \mathcal{L}^*(V_q, V_{q'})) + \Phi(-z_{\alpha_0} - \mathcal{L}^*(V_q, V_{q'})). \quad (61)$$

With $|\mathcal{L}^*(V_q, V_{q'})| = \infty$, then we have $\liminf_{n \rightarrow \infty} \mathbf{P}(\mathcal{C}(V_q, V_{q'}) = 1) = 1$.

With the decomposition (53), we establish (61) in Corollary 2 by Lemma 4 in the main paper and applying the condition $\mathcal{L}_n(V_q, V_{q'}) \xrightarrow{p} \mathcal{L}^*(V_q, V_{q'})$ and $\widehat{H}(V_q, V_{q'}) \xrightarrow{p} H(V_q, V_{q'})$.

A.3 Properties of $\mathbf{M}(V)$

The following lemma is about the property of the transformation matrix $\mathbf{M}(V)$, whose proof can be found in Section C.4.

Lemma 5 *The transformation matrix $\mathbf{M}_{\text{RF}}(V)$ satisfies*

$$\lambda_{\max}(\mathbf{M}_{\text{RF}}(V)) \leq 1 \quad \text{and} \quad b^{\top}[\mathbf{M}_{\text{RF}}(V)]^2 b \leq b^{\top} \mathbf{M}_{\text{RF}}(V) b \quad \text{for any } b \in \mathbb{R}^{n_1}. \quad (62)$$

As a consequence, we establish $\text{Tr}([\mathbf{M}_{\text{RF}}(V)]^2) \leq \text{Tr}[\mathbf{M}_{\text{RF}}(V)]$. The transformation matrixs $\mathbf{M}_{\text{ba}}(V)$ and $\mathbf{M}_{\text{DNN}}(V)$ are orthogonal projection matrices with

$$[\mathbf{M}_{\text{ba}}(V)]^2 = \mathbf{M}_{\text{ba}}(V) \quad \text{and} \quad [\mathbf{M}_{\text{DNN}}(V)]^2 = \mathbf{M}_{\text{DNN}}(V). \quad (63)$$

The transformation matrix $\mathbf{M}_{\text{boo}}(V)$ satisfies

$$\lambda_{\max}(\mathbf{M}_{\text{boo}}(V)) \leq \|\Omega^{\text{boo}}\|_2^2, \quad \text{and} \quad b^{\top}[\mathbf{M}_{\text{boo}}(V)]^2 b \leq \|\Omega^{\text{boo}}\|_2^2 \cdot b^{\top} \mathbf{M}_{\text{boo}}(V) b, \quad (64)$$

for any $b \in \mathbb{R}^{n_1}$. As a consequence, we establish $\text{Tr}([\mathbf{M}_{\text{boo}}(V)]^2) \leq \|\Omega^{\text{boo}}\|_2^2 \cdot \text{Tr}[\mathbf{M}(V)]$.

A.4 Consistency of variance estimators

The following lemma controls the variance consistency, whose proof can be found in Section C.5.

Lemma 6 *Suppose that Conditions (R1) and (R2) hold and $\kappa_n(V)^2 + \sqrt{\log n} \kappa_n(V) \xrightarrow{p} 0$, with $\kappa_n(V) = \sqrt{\log n} \left(\|R(V)\|_{\infty} + |\beta - \widehat{\beta}_{\text{init}}(V)| + \log n / \sqrt{n} \right)$. If*

$$\frac{\sum_{i=1}^{n_1} \sigma_i^2 [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2}{\sum_{i=1}^{n_1} \sigma_i^2 [\mathbf{M}(V) f_{\mathcal{A}_1}]_i^2} \xrightarrow{p} 1, \quad \text{and} \quad \frac{\max_{1 \leq i \leq n_1} [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2}{\sum_{i=1}^{n_1} [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2} \xrightarrow{p} 0,$$

then we have $\widehat{\text{SE}}(V)/\text{SE}(V) \xrightarrow{p} 1$.

A.5 Calculation of Ω for the boosting method

In the following, we give three examples on the construction of the base procedures and provide the detailed expression for $\mathcal{H}^{[m]}$ used in the boosting method in Section 4.1 in the main paper. We write $C_i = (X_i^{\top}, Z_i^{\top})^{\top} \in \mathbb{R}^p$ and define the matrix $C \in \mathbb{R}^{n \times p}$ with its i -th row as C_i^{\top} . An important step of building the base procedures $\{\widehat{g}^{[m]}(Z_i, X_i)\}_{m \geq 1}$ is to conduct the variable selection. That is, each base procedure is only constructed based on a subset of covariates $C_i = (X_i^{\top}, Z_i^{\top}) \in \mathbb{R}^p$.

In Algorithm 2, we describe the construction of Ω^{boo} with the pairwise regression and the pairwise thin plate as the base procedure. For both base procedures, we need to specify

how to construct the basis functions in steps 3 and 8. For the pairwise regression, we set the first element of C_i as 1 and define $S_i^{j,l} = C_{ij}C_{il}$ for $1 \leq i \leq n$. Then for step 3, we define $\mathcal{P}^{j,l}$ as the projection matrix to the vector $S_{\mathcal{A}_2}^{j,l}$; for step 8, we define $\mathcal{H}^{[m]} = (S_{\mathcal{A}_1}^{\widehat{j}_m, \widehat{l}_m})^\top / \|S_{\mathcal{A}_1}^{\widehat{j}_m, \widehat{l}_m}\|_2^2$.

For the pairwise thin plate, we follow chapter 7 of [21] to construct the projection matrix. For $1 \leq j < l \leq p$, we define the matrix $T^{j,l} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ X_{1,j} & X_{2,j} & \cdots & X_{n,j} \\ X_{1,l} & X_{2,l} & \cdots & X_{n,l} \end{pmatrix}$. For $1 \leq s \leq n$, define $E_{s,s}^{j,l} = 0$; for $1 \leq s \neq t \leq n$, define

$$E_{s,t}^{j,l} = \frac{1}{16\pi} \|(X_{s,(j,l)} - X_{t,(j,l)})\|_2^2 \log(\|(X_{s,(j,l)} - X_{t,(j,l)})\|_2^2).$$

Then we define $\bar{\mathcal{P}}^{j,l} = \begin{pmatrix} E_{\mathcal{A}_2, \mathcal{A}_2}^{j,l} & (T_{\cdot, \mathcal{A}_2}^{j,l})^\top \\ T_{\cdot, \mathcal{A}_2}^{j,l} & 0 \end{pmatrix}^{-1} \in \mathbb{R}^{n_2 \times (n_2+3)}$. In step 3, compute $\mathcal{P}^{j,l} \in \mathbb{R}^{n_2 \times n_2}$ as the first n_2 columns of $\bar{\mathcal{P}}^{j,l}$. For step 8, we compute

$$\bar{\mathcal{H}}^{j,l} = \begin{pmatrix} E_{\mathcal{A}_1, \mathcal{A}_1}^{j,l} & (T_{\cdot, \mathcal{A}_1}^{j,l})^\top \\ T_{\cdot, \mathcal{A}_1}^{j,l} & 0 \end{pmatrix}^{-1} \in \mathbb{R}^{n_1 \times (n_1+3)},$$

and set $\mathcal{H}^{[m]} \in \mathbb{R}^{n_1 \times n_1}$ as the first n_1 columns of $\bar{\mathcal{H}}^{\widehat{j}_m, \widehat{l}_m}$.

In Algorithm 3, we describe the construction of Ω^{boo} with the decision tree as the base procedure.

B Additional proofs

We establish Theorem 4 in Section B.1 and prove Theorem 5 in Section B.2.

B.1 Proof of Theorem 4

Recall that the test $\mathcal{C}(V_q)$ is defined in (60) and note that

$$\{\widehat{q}_c = Q^*\} = \{Q_{\max} \geq Q^*\} \cap \left(\bigcap_{q=0}^{Q^*-1} \{\mathcal{C}(V_q) = 1\} \right) \cap \{\mathcal{C}(V_{Q^*}) = 0\}. \quad (65)$$

Define the events

$$\begin{aligned} \mathcal{B}_1 &= \left\{ \max_{0 \leq q < q' \leq Q_{\max}} |\mathcal{G}_n(V_q, V_{q'})| \leq \rho(\alpha_0) \right\} \quad \text{and} \quad \mathcal{B}_2 = \{|\widehat{\rho}/\rho(\alpha_0) - 1| \leq \tau_0\}, \\ \mathcal{B}_3 &= \left\{ \max_{0 \leq q < q' \leq Q_{\max}} \frac{1}{\sqrt{H(V_q, V_{q'})}} \left| \widetilde{\mathcal{E}}(V_q) - \widetilde{\mathcal{E}}(V_{q'}) \right| \leq \tau_1 \rho(\alpha_0) \right\}. \end{aligned}$$

Algorithm 2 Construction of Ω in Boosting with non-parametric pairwise regression

Input: Data $C \in \mathbb{R}^{n \times p}, D \in \mathbb{R}^n$; the step-length factor $0 < \nu \leq 1$; the stoping time m_{stop} .
Output: Ω^{boo}

- 1: Randomly split the data into disjoint subsets $\mathcal{A}_1, \mathcal{A}_2$ with $n_1 = \lfloor \frac{2n}{3} \rfloor$ and $|\mathcal{A}_2| = n - n_1$;
 - 2: Set $m = 0$, $\hat{f}_{\mathcal{A}_2}^{[0]} = 0$, and $\Omega^{[0]} = 0$.
 - 3: For $1 \leq j, l \leq p$, compute $\mathcal{P}^{j,l}$ as the projection matrix to a set of basis functions generated by $C_{\mathcal{A}_2,j}$ and $C_{\mathcal{A}_2,l}$.
 - 4: **for** $1 \leq m \leq m_{\text{stop}}$ **do**
 - 5: Compute the adjusted outcome $U_{\mathcal{A}_2}^{[m]} = D_{\mathcal{A}_2} - \hat{f}_{\mathcal{A}_2}^{[m-1]}$
 - 6: Implement the following base procedure on $\{C_i, U_i^{[m]}\}_{i \in \mathcal{A}_2}$
$$(\hat{j}_m, \hat{l}_m) = \arg \min_{1 \leq j, l \leq p} \left\| U_{\mathcal{A}_2}^{[m]} - \mathcal{P}^{j,l} U_{\mathcal{A}_2}^{[m]} \right\|^2.$$
 - 7: Update $\hat{f}_{\mathcal{A}_2}^{[m]} = \hat{f}_{\mathcal{A}_2}^{[m-1]} + \nu \mathcal{P}^{\hat{j}_m, \hat{l}_m} (D_{\mathcal{A}_2} - \hat{f}_{\mathcal{A}_2}^{[m-1]})$
 - 8: Construct $\mathcal{H}^{[m]}$ in the same way as $\mathcal{P}^{\hat{j}_m, \hat{l}_m}$ but with the data in \mathcal{A}_1 .
 - 9: Compute $\Omega^{[m]} = \nu \mathcal{H}^{[m]} + (I - \nu \mathcal{H}^{[m]}) \Omega^{[m-1]}$
 - 10: **end for**
 - 11: Return Ω^{boo}
-

Algorithm 3 Construction of Ω in Boosting with Decision tree

Input: Data $C \in \mathbb{R}^{n \times p}, D \in \mathbb{R}^n$; the step-length factor $0 < \nu \leq 1$; the stoping time m_{stop} .
Output: Ω^{boo}

- 1: Randomly split the data into disjoint subsets $\mathcal{A}_1, \mathcal{A}_2$ with $n_1 = \lfloor \frac{2n}{3} \rfloor$ and $|\mathcal{A}_2| = n - n_1$;
- 2: Set $m = 0$, $\hat{f}_{\mathcal{A}_2}^{[0]} = 0$, and $\Omega^{[0]} = 0$.
- 3: **for** $1 \leq m \leq m_{\text{stop}}$ **do**
- 4: Compute the adjusted outcome $U_{\mathcal{A}_2}^{[m]} = D_{\mathcal{A}_2} - \hat{f}_{\mathcal{A}_2}^{[m-1]}$
- 5: Run the decision tree on $\{C_i, U_i^{[m]}\}_{i \in \mathcal{A}_2}$ and partition \mathbb{R}^p as leaves $\{\mathcal{R}_1^{[m]}, \dots, \mathcal{R}_L^{[m]}\}$;
- 6: For any $j \in \mathcal{A}_2$, we identify the leaf $\mathcal{R}_{l(C_j)}^{[m]}$ containing C_j and compute

$$\mathcal{P}_{j,t}^{[m]} = \frac{\mathbf{1} [C_t \in \mathcal{R}_{l(C_j)}^{[m]}]}{\sum_{k \in \mathcal{A}_2} \mathbf{1} [C_k \in \mathcal{R}_{l(C_j)}^{[m]}]} \quad \text{for } t \in \mathcal{A}_2$$

- 7: Compute the matrix $(\mathcal{P}_{j,t}^{[m]})_{j,t \in \mathcal{A}_2}$ and update

$$\hat{f}_{\mathcal{A}_2}^{[m]} = \hat{f}_{\mathcal{A}_2}^{[m-1]} + \nu \mathcal{P}^{[m]} \left(D_{\mathcal{A}_2} - \hat{f}_{\mathcal{A}_2}^{[m-1]} \right)$$

- 8: Construct the matrix $(\mathcal{H}_{j,t}^{[m]})_{j,t \in \mathcal{A}_1}$ as $\mathcal{H}_{j,t}^{[m]} = \frac{\mathbf{1} [C_t \in \mathcal{R}_{l(C_j)}^{[m]}]}{\sum_{k \in \mathcal{A}_1} \mathbf{1} [C_k \in \mathcal{R}_{l(C_j)}^{[m]}]}$.
- 9: Compute $\Omega^{[m]} = \nu \mathcal{H}^{[m]} + (I - \nu \mathcal{H}^{[m]}) \Omega^{[m-1]}$

10: **end for**

- 11: Return Ω^{boo}
-

where $\rho(\alpha_0)$ is defined in (39), $\widehat{\rho}$ is defined in (24) with \mathbf{M}_{RF} replaced by \mathbf{M} , and $\mathcal{G}_n(V_q, V_{q'})$ is defined in (54). By the definition of $\rho(\alpha_0)$ in (39) and the following (84), we control the probability of the event $\lim_{n \rightarrow \infty} \mathbf{P}(\mathcal{B}_1) = 1 - 2\alpha_0$. By the following (4) and $\widehat{\rho}/\rho(\alpha_0) \xrightarrow{P} 1$, we establish that, for any positive constants $\tau_0 > 0$ and $\tau_1 > 0$, $\liminf_{n \rightarrow \infty} \mathbf{P}(\mathcal{B}_2 \cap \mathcal{B}_3) = 1$. Combing the above two equalities, we have

$$\liminf_{n \rightarrow \infty} \mathbf{P}(\mathcal{B}_1 \cap \mathcal{B}_2 \cap \mathcal{B}_3) \geq 1 - 2\alpha_0. \quad (66)$$

For any $0 \leq q \leq Q^* - 1$, the condition (R4) implies that there exists some $q+1 \leq q' \leq Q^*$ such that $|\mathcal{L}_n(V_q, V_{q'})| \geq A\rho(\alpha_0)$. On the event $\mathcal{B}_1 \cap \mathcal{B}_2 \cap \mathcal{B}_3$, we apply the expression (53) and obtain

$$\left| [\widehat{\beta}(V_q) - \widehat{\beta}(V_{q'})] / \sqrt{H(V_q, V_{q'})} \right| \geq A\rho(\alpha_0) - \rho(\alpha_0) - \tau_1\rho(\alpha_0) \geq (A - 1 - \tau_1)(1 - \tau_0)\widehat{\rho}. \quad (67)$$

For any $Q^* \leq q' \leq Q_{\max}$, we have $R(V_{q'}) = 0$. Then on the event $\mathcal{B}_1 \cap \mathcal{B}_2 \cap \mathcal{B}_3$, we apply the expression (53) and obtain that,

$$\left| [\widehat{\beta}(V_{Q^*}) - \widehat{\beta}(V_{q'})] / \sqrt{H(V_{Q^*}, V_{q'})} \right| \leq \rho(\alpha_0) + \tau_1\rho(\alpha_0) \leq (1 + \tau_1)(1 + \tau_0)\widehat{\rho} \leq 1.01\widehat{\rho}.$$

Together with (67), the event $\mathcal{B}_1 \cap \mathcal{B}_2 \cap \mathcal{B}_3$ implies $\bigcap_{q=0}^{Q^*-1} \{\mathcal{C}(V_q) = 1\} \cap \{\mathcal{C}(V_{Q^*}) = 0\}$. By (66), we establish $\liminf_{n \rightarrow \infty} \mathbf{P}\left[\left(\bigcap_{q=0}^{Q^*-1} \{\mathcal{C}(V_q) = 1\}\right) \cap \{\mathcal{C}(V_{Q^*}) = 0\}\right] \geq 1 - 2\alpha_0$. Together with (65), we have $\liminf_{n \rightarrow \infty} \mathbf{P}(\widehat{q}_c \neq Q^*) \leq 2\alpha_0$. To control the coverage probability, we decompose $\mathbf{P}\left(\frac{1}{\text{SE}(V_{\widehat{q}_c})} \left| \widehat{\beta}(V_{\widehat{q}_c}) - \beta \right| \geq z_{\alpha/2}\right)$ as

$$\begin{aligned} & \mathbf{P}\left(\left\{\frac{1}{\text{SE}(V_{\widehat{q}_c})} \left| \widehat{\beta}(V_{\widehat{q}_c}) - \beta \right| \geq z_{\alpha/2}\right\} \cap \{\widehat{q}_c = Q^*\}\right) \\ & + \mathbf{P}\left(\left\{\frac{1}{\text{SE}(V_{\widehat{q}_c})} \left| \widehat{\beta}(V_{\widehat{q}_c}) - \beta \right| \geq z_{\alpha/2}\right\} \cap \{\widehat{q}_c \neq Q^*\}\right). \end{aligned} \quad (68)$$

Note that

$$\mathbf{P}\left(\left\{\frac{1}{\text{SE}(V_{\widehat{q}_c})} \left| \widehat{\beta}(V_{\widehat{q}_c}) - \beta \right| \geq z_{\alpha/2}\right\} \cap \{\widehat{q}_c = Q^*\}\right) \leq \mathbf{P}\left(\left\{\frac{1}{\text{SE}(V_{Q^*})} \left| \widehat{\beta}(V_{Q^*}) - \beta \right| \geq z_{\alpha/2}\right\}\right) \leq \alpha,$$

and $\mathbf{P}\left(\left\{\frac{1}{\text{SE}(V_{\widehat{q}_c})} \left| \widehat{\beta}(V_{\widehat{q}_c}) - \beta \right| \geq z_{\alpha/2}\right\} \cap \{\widehat{q}_c \neq Q^*\}\right) \leq \mathbf{P}(\widehat{q}_c \neq Q^*) \leq 2\alpha_0$. By the decomposition (68), we combine the above two inequalities and establish

$$\mathbf{P}\left(\frac{1}{\text{SE}(V_{\widehat{q}_c})} \left| \widehat{\beta}(V_{\widehat{q}_c}) - \beta \right| \geq z_{\alpha/2}\right) \leq \alpha + 2\alpha_0,$$

which implies (41) with $\widehat{q} = \widehat{q}_c$. By the definition $\widehat{q}_r = \min\{\widehat{q}_c + 1, Q_{\max}\}$, we apply a similar argument and establish (41) with $\widehat{q} = \widehat{q}_r$.

B.2 Proof of Theorem 5

If $\text{Cov}(\epsilon_i, \delta_i | Z_i, X_i) = \text{Cov}(\epsilon_i, \delta_i)$, then (45) further implies

$$|\epsilon_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V) \delta_{\mathcal{A}_1} - \text{Cov}(\epsilon_i, \delta_i) \cdot \text{Tr}[\mathbf{M}_{\text{RF}}(V)]| \leq t_0 K^2 \sqrt{\text{Tr}([\mathbf{M}_{\text{RF}}(V)]^2)}. \quad (69)$$

Proof of (55). We decompose the error $\widehat{\text{Cov}}(\delta_i, \epsilon_i) - \text{Cov}(\delta_i, \epsilon_i)$ as,

$$\begin{aligned} & \frac{1}{n_1 - r} (f_{\mathcal{A}_1} - \widehat{f}_{\mathcal{A}_1} + \delta_{\mathcal{A}_1})^\top P_{V_{\mathcal{A}_1}, W_{\mathcal{A}_1}}^\perp [\epsilon_{\mathcal{A}_1} + D_{\mathcal{A}_1}(\beta - \widehat{\beta}_{\text{init}}(V)) + R_{\mathcal{A}_1}(V)] - \text{Cov}(\delta_i, \epsilon_i) \\ &= T_1 + T_2 + T_3, \end{aligned} \quad (70)$$

where $T_1 = \frac{1}{n_1 - r} \left[(\delta_{\mathcal{A}_1})^\top P_{V_{\mathcal{A}_1}, W_{\mathcal{A}_1}}^\perp \epsilon_{\mathcal{A}_1} - (n_1 - r) \text{Cov}(\delta_i, \epsilon_i) \right]$, and

$$\begin{aligned} T_2 &= \frac{1}{n_1 - r} (\delta_{\mathcal{A}_1})^\top P_{V_{\mathcal{A}_1}, W_{\mathcal{A}_1}}^\perp \left[D_{\mathcal{A}_1}(\beta - \widehat{\beta}_{\text{init}}(V)) + R_{\mathcal{A}_1}(V) \right], \\ T_3 &= \frac{1}{n_1 - r} (f_{\mathcal{A}_1} - \widehat{f}_{\mathcal{A}_1})^\top P_{V_{\mathcal{A}_1}, W_{\mathcal{A}_1}}^\perp [\epsilon_{\mathcal{A}_1} + D_{\mathcal{A}_1}(\beta - \widehat{\beta}_{\text{init}}(V)) + R_{\mathcal{A}_1}(V)]. \end{aligned}$$

We apply (43) with $A = P_{V_{\mathcal{A}_1}, W_{\mathcal{A}_1}}^\perp$ and $t = t_0 K^2 \sqrt{n_1 - r}$ for some $0 < t_0 \leq \sqrt{n_1 - r}$, and establish

$$\mathbb{P} \left(\left| \epsilon_{\mathcal{A}_1}^\top P_{V_{\mathcal{A}_1}, W_{\mathcal{A}_1}}^\perp \delta_{\mathcal{A}_1} - \text{Cov}(\epsilon_i, \delta_i) \cdot \text{Tr} \left(P_{V_{\mathcal{A}_1}, W_{\mathcal{A}_1}}^\perp \right) \right| \geq t_0 K^2 \sqrt{n_1 - r} \mid \mathcal{O} \right) \leq 6 \exp(-ct_0^2).$$

By choosing $t_0 = \log(n_1 - r)$, we establish that, with probability larger than $1 - (n_1 - r)^{-c}$ for some positive constant $c > 0$, then

$$|T_1| \lesssim \sqrt{\log(n_1 - r)/(n_1 - r)}. \quad (71)$$

We apply (43) with $A = P_{V_{\mathcal{A}_1}, W_{\mathcal{A}_1}}^\perp$ and $t = t_0 K^2 \sqrt{n_1 - r}$ for some $0 < t_0 \leq \sqrt{n_1 - r}$,

$$\mathbb{P} \left(\left| \epsilon_{\mathcal{A}_1}^\top P_{V_{\mathcal{A}_1}, W_{\mathcal{A}_1}}^\perp D_{\mathcal{A}_1} - \text{Cov}(\epsilon_i, D_i) \cdot \text{Tr} \left(P_{V_{\mathcal{A}_1}, W_{\mathcal{A}_1}}^\perp \right) \right| \geq t_0 K^2 \sqrt{n_1 - r} \mid \mathcal{O} \right) \leq 6 \exp(-ct_0^2).$$

The above concentration bound implies

$$\mathbb{P} \left(\frac{1}{n_1 - r} \left| \epsilon_{\mathcal{A}_1}^\top P_{V_{\mathcal{A}_1}, W_{\mathcal{A}_1}}^\perp D_{\mathcal{A}_1} \right| \geq C \left(1 + \sqrt{\frac{\log(n_1 - r)}{n_1 - r}} \right) \mid \mathcal{O} \right) \leq 6(n_1 - r)^{-c}. \quad (72)$$

Hence, we establish that, with probability larger than $1 - (n_1 - r)^{-c}$,

$$|T_2| \lesssim |\beta - \widehat{\beta}_{\text{init}}(V)| + \|R(V)\|_2 / \sqrt{n_1 - r}, \quad (73)$$

where the last inequality follows from (49). Regarding T_3 , we have

$$\begin{aligned} |T_3| &= \left| \frac{1}{n_1 - r} (f_{\mathcal{A}_1} - \hat{f}_{\mathcal{A}_1})^\top P_{V_{\mathcal{A}_1}, W_{\mathcal{A}_1}}^\perp [\epsilon_{\mathcal{A}_1} + D_{\mathcal{A}_1}(\beta - \hat{\beta}_{\text{init}}(V)) + R_{\mathcal{A}_1}(V)] \right| \\ &\lesssim \frac{\|P_{V_{\mathcal{A}_1}, W_{\mathcal{A}_1}}^\perp (f_{\mathcal{A}_1} - \hat{f}_{\mathcal{A}_1})\|_2}{\sqrt{n_1 - r}} \frac{\|P_{V_{\mathcal{A}_1}, W_{\mathcal{A}_1}}^\perp \epsilon_{\mathcal{A}_1}\|_2 + \|P_{V_{\mathcal{A}_1}, W_{\mathcal{A}_1}}^\perp D_{\mathcal{A}_1}\|_2 \cdot |\beta - \hat{\beta}_{\text{init}}(V)| + \|R_{\mathcal{A}_1}(V)\|_2}{\sqrt{n_1 - r}}. \end{aligned} \quad (74)$$

By a similar argument as in (72), we establish that, with probability larger than $1 - (n_1 - r)^{-c}$, $\frac{1}{\sqrt{n_1}} \|P_{V_{\mathcal{A}_1}, W_{\mathcal{A}_1}}^\perp \epsilon_{\mathcal{A}_1}\|_2 + \frac{1}{\sqrt{n_1}} \|P_{V_{\mathcal{A}_1}, W_{\mathcal{A}_1}}^\perp D_{\mathcal{A}_1}\|_2 \leq C$, for some positive constant $C > 0$. Together with (74), we establish that, with probability larger than $1 - (n_1 - r)^{-c}$,

$$|T_3| \lesssim \frac{\|P_{V_{\mathcal{A}_1}, W_{\mathcal{A}_1}}^\perp (f_{\mathcal{A}_1} - \hat{f}_{\mathcal{A}_1})\|_2}{\sqrt{n_1 - r}} \cdot \left(1 + |\beta - \hat{\beta}_{\text{init}}(V)| + \frac{\|R(V)\|_2}{\sqrt{n_1 - r}} \right).$$

Together with (70) and the upper bounds (71) and (73), we establish (55).

Proof of (58). By (44), we obtain the following decomposition for $\tilde{\beta}_{\text{RF}}(V)$ in (12),

$$\begin{aligned} \hat{\beta}_{\text{RF}}(V) - \beta &= \frac{\epsilon_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V) f_{\mathcal{A}_1} - [\widehat{\text{Cov}}(\delta_i, \epsilon_i) - \text{Cov}(\delta_i, \epsilon_i)] \text{Tr}[\mathbf{M}_{\text{RF}}(V)]}{D_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V) D_{\mathcal{A}_1}} \\ &\quad + \frac{\epsilon_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V) \delta_{\mathcal{A}_1} - \text{Cov}(\delta_i, \epsilon_i) \text{Tr}[\mathbf{M}_{\text{RF}}(V)]}{D_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V) D_{\mathcal{A}_1}} + \frac{R_{\mathcal{A}_1}(V)^\top \mathbf{M}_{\text{RF}}(V) D_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V) D_{\mathcal{A}_1}}. \end{aligned} \quad (75)$$

The above decomposition implies (31) with $\mathcal{G}(V) = \frac{1}{\text{SE}(V)} \frac{\epsilon_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V) f_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V) D_{\mathcal{A}_1}}$, and

$$\begin{aligned} \text{SE}(V) \cdot \mathcal{E}(V) &= \frac{R_{\mathcal{A}_1}(V)^\top \mathbf{M}_{\text{RF}}(V) D_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V) D_{\mathcal{A}_1}} \\ &\quad + \frac{[\text{Cov}(\delta_i, \epsilon_i) - \widehat{\text{Cov}}(\delta_i, \epsilon_i)] \text{Tr}[\mathbf{M}_{\text{RF}}(V)] + \epsilon_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V) \delta_{\mathcal{A}_1} - \text{Cov}(\delta_i, \epsilon_i) \text{Tr}[\mathbf{M}_{\text{RF}}(V)]}{D_{\mathcal{A}_1}^\top \mathbf{M}_{\text{RF}}(V) D_{\mathcal{A}_1}}. \end{aligned}$$

We establish $\mathcal{G}(V) \xrightarrow{d} N(0, 1)$ by applying the same arguments as in Section 9.2 in the main paper. We establish (58) by combining (55), (47), and (69).

C Proof of Extra Lemmas

C.1 Proof of Lemma 2

Note that $\mathbf{E} [\epsilon_{\mathcal{A}_1}^\top \mathbf{M}(V) \delta_{\mathcal{A}_1} | \mathcal{O}] = \text{Tr}(\mathbf{M}(V)\Lambda)$. We apply (43) with $A = \mathbf{M}(V)$ and $t = t_0 K^2 \|\mathbf{M}(V)\|_F$ for some $t_0 > 0$ and establish

$$\begin{aligned} & \mathbb{P} (|\epsilon_{\mathcal{A}_1}^\top \mathbf{M}(V) \delta_{\mathcal{A}_1} - \text{Tr}(\mathbf{M}(V)\Lambda)| \geq t_0 K^2 \|\mathbf{M}(V)\|_F | \mathcal{O}) \\ & \leq 6 \exp \left(-c \min \left\{ t_0^2, t_0 \frac{\|\mathbf{M}(V)\|_F}{\|\mathbf{M}(V)\|_2} \right\} \right) \leq 6 \exp (-c \min \{t_0^2, t_0\}), \end{aligned} \quad (76)$$

where the last inequality follows from $\|\mathbf{M}(V)\|_F \geq \|\mathbf{M}(V)\|_2$. The above concentration bound implies (45) by taking an expectation with respect to \mathcal{O} .

Since $\mathbf{E} [\delta_{\mathcal{A}_1}^\top \mathbf{M}(V) \delta_{\mathcal{A}_1} | \mathcal{O}] = \text{Tr}(\mathbf{M}(V)\Sigma^\delta)$, we apply a similar argument to (76) and establish

$$\mathbb{P} (|\delta_{\mathcal{A}_1}^\top \mathbf{M}(V) \delta_{\mathcal{A}_1} - \text{Tr}(\mathbf{M}(V)\Sigma^\delta)| \geq t_0 K^2 \|\mathbf{M}(V)\|_F | \mathcal{O}) \leq 2 \exp (-c \min \{t_0^2, t_0\}). \quad (77)$$

Note that $\mathbf{E} [\delta_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1} | \mathcal{O}] = 0$, and conditioning on \mathcal{O} , $\{\delta_i\}_{i \in \mathcal{A}_1}$ are independent sub-gaussian random variables. We apply Proposition 5.16 of [55] and establish that, with probability larger than $1 - \exp(-t_0^2)$ for some positive constant $c > 0$,

$$\mathbb{P} \left(\delta_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1} \geq C t_0 K \sqrt{f_{\mathcal{A}_1}^\top [\mathbf{M}(V)]^2 f_{\mathcal{A}_1}} | \mathcal{O} \right) \leq \exp(-ct_0^2). \quad (78)$$

The above concentration bound implies (45) by taking an expectation with respect to \mathcal{O} .

By the decomposition $D_{\mathcal{A}_1}^\top \mathbf{M}(V) D_{\mathcal{A}_1} - f_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1} = \delta_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1} + \delta_{\mathcal{A}_1}^\top \mathbf{M}(V) \delta_{\mathcal{A}_1}$, we establish (46) by applying the concentration bounds (77) and (78).

C.2 Proof of Lemma 3

Under the model $Y_i = D_i \beta + V_i^\top \pi + W_i^\top \psi + R_i + \epsilon_i$, we have

$$Y_{\mathcal{A}_1} - \widehat{\beta}_{\text{init}}(V) D_{\mathcal{A}_1} = V_{\mathcal{A}_1} \pi + W_{\mathcal{A}_1} \psi + R_{\mathcal{A}_1} + D_{\mathcal{A}_1} (\beta - \widehat{\beta}_{\text{init}}(V)) + \epsilon_{\mathcal{A}_1}. \quad (79)$$

Define the matrix $\Psi = (V_{\mathcal{A}_1}, W_{\mathcal{A}_1})$. The least square estimator of $(\pi^\top, \psi^\top)^\top$ is expressed as, $(\widehat{\pi}^\top, \widehat{\psi}^\top) = (\Psi^\top \Psi)^{-1} \Psi^\top \left(Y_{\mathcal{A}_1} - \widehat{\beta}_{\text{init}}(V) D_{\mathcal{A}_1} \right)$. Note that

$$\begin{aligned} & (\Psi^\top \Psi)^{-1} \Psi^\top \left(Y_{\mathcal{A}_1} - \widehat{\beta}_{\text{init}}(V) D_{\mathcal{A}_1} \right) - (\pi^\top, \psi^\top)^\top \\ &= (\Psi^\top \Psi)^{-1} \Psi^\top \left([R(V)]_{\mathcal{A}_1} + D_{\mathcal{A}_1}(\beta - \widehat{\beta}_{\text{init}}(V)) + \epsilon_{\mathcal{A}_1} \right) \\ &= \left(\sum_{i=1}^{n_1} \Psi_i \Psi_i^\top \right)^{-1} \sum_{i=1}^{n_1} \Psi_i \left([R(V)]_i + D_i(\beta - \widehat{\beta}_{\text{init}}(V)) + \epsilon_i \right). \end{aligned}$$

Note that $\mathbf{E} \Psi_i \epsilon_i = 0$ and conditioning on \mathcal{O} , $\{\epsilon_i\}_{i \in \mathcal{A}_1}$ are independent sub-gaussian random variables. By Proposition 5.16 of [55], there exist positive constants $C > 0$ and $c > 0$ such that $\mathbb{P} \left(\frac{1}{n_1} \left| \sum_{i=1}^{n_1} \Psi_{ij} \epsilon_i \right| \geq C t_0 \sqrt{\sum_{i=1}^{n_1} \Psi_{ij}^2} / n_1 \mid \mathcal{O} \right) \leq \exp(-ct_0^2)$. By Condition (R1), we have $\max_{i,j} \Psi_{ij}^2 \leq C \log n$ and apply the union bound and establish

$$\mathbb{P} \left(\left\| \sum_{i=1}^{n_1} \Psi_i \epsilon_i / n_1 \right\|_\infty \geq C \log n_1 / \sqrt{n_1} \right) \leq n_1^{-c}. \quad (80)$$

Similarly to (80), we establish $\mathbb{P} \left(\left\| \frac{1}{n_1} \sum_{i=1}^{n_1} \Psi_i \delta_i \right\|_\infty \geq C \frac{\log n_1}{\sqrt{n_1}} \right) \leq n_1^{-c}$. Together with the expression $\frac{1}{n_1} \sum_{i=1}^{n_1} \Psi_i D_i = \frac{1}{n_1} \sum_{i=1}^{n_1} \Psi_i f_i + \frac{1}{n_1} \sum_{i=1}^{n_1} \Psi_i \delta_i$, we apply Condition (R1) and establish $\mathbb{P} \left(\left| \frac{1}{n_1} \sum_{i=1}^{n_1} \Psi_i D_i \right| \geq C \right) \leq n_1^{-c}$. Together with (80), and Condition (R1), we establish that, with probability larger than $1 - n_1^{-c}$,

$$\|(\widehat{\pi}^\top, \widehat{\psi}^\top) - (\pi^\top, \psi^\top)^\top\|_2 \lesssim \|R(V)\|_\infty + |\beta - \widehat{\beta}_{\text{init}}(V)| + \log n / \sqrt{n}. \quad (81)$$

Our proposed estimator $\widehat{\epsilon}(V)$ defined in (14) has the following equivalent expression,

$$\begin{aligned} \widehat{\epsilon}(V) &= P_{V_{\mathcal{A}_1}, W_{\mathcal{A}_1}}^\perp [Y_{\mathcal{A}_1} - D_{\mathcal{A}_1} \widehat{\beta}_{\text{init}}(V)] \\ &= Y_{\mathcal{A}_1} - D_{\mathcal{A}_1} \widehat{\beta}_{\text{init}}(V) - \Psi(\Psi^\top \Psi)^{-1} \Psi^\top \left(Y_{\mathcal{A}_1} - \widehat{\beta}_{\text{init}}(V) D_{\mathcal{A}_1} \right) \\ &= Y_{\mathcal{A}_1} - D_{\mathcal{A}_1} \widehat{\beta}_{\text{init}}(V) - V_{\mathcal{A}_1} \widehat{\pi} - W_{\mathcal{A}_1} \widehat{\psi}. \end{aligned}$$

Then we apply (79) and obtain $[\widehat{\epsilon}(V)]_i - \epsilon_i = D_i(\beta - \widehat{\beta}_{\text{init}}(V)) + V_i^\top (\pi - \widehat{\pi}) + W_i^\top (\psi - \widehat{\psi}) + R_i(V)$. Together with Condition (R1) and (81), we establish Lemma 3.

C.3 Proof of Lemma 4

Similarly to (50), we decompose $\widetilde{\mathcal{E}}(V_q)$ as $\widetilde{\mathcal{E}}(V_q) = \frac{\text{Err}_1 + \text{Err}_2}{D_{\mathcal{A}_1}^\top \mathbf{M}(V_q) D_{\mathcal{A}_1}}$, where $\text{Err}_1 = \sum_{1 \leq i \neq j \leq n_1} [\mathbf{M}(V_q)]_{ij} \delta_i \epsilon_j$, and

$$\text{Err}_2 = \sum_{i=1}^{n_1} [\mathbf{M}(V_q)]_{ii} (f_i - \widehat{f}_i) (\epsilon_i + [\widehat{\epsilon}(V_{Q_{\max}})]_i - \epsilon_i) + \sum_{i=1}^{n_1} [\mathbf{M}(V_q)]_{ii} \delta_i ([\widehat{\epsilon}(V_{Q_{\max}})]_i - \epsilon_i).$$

We apply the same analysis as that of (52) and establish

$$|\text{Err}_2| \lesssim \sum_{i=1}^{n_1} [\mathbf{M}(V)]_{ii} \left[|f_i - \hat{f}_i| \left(\sqrt{\log n} + |[\widehat{\epsilon}(V_{Q_{\max}})]_i - \epsilon_i| \right) + \sqrt{\log n} |[\widehat{\epsilon}(V_{Q_{\max}})]_i - \epsilon_i| \right].$$

By Lemma 3 with $V = V_{Q_{\max}}$, we apply the similar argument as that of (32) and establish that

$$\mathbb{P} \left(\left| \widetilde{\mathcal{E}}(V_q) \right| \geq C \frac{\sqrt{\log n} \cdot \eta_n(V_{Q_{\max}}) \cdot \text{Tr}[\mathbf{M}(V_q)] + t_0 \sqrt{\text{Tr}([\mathbf{M}(V_q)]^2)}}{f_{\mathcal{A}_1} \mathbf{M}(V_q) f_{\mathcal{A}_1}} \right) \geq 1 - \exp(-t_0^2)$$

where $\eta_n(V)$ is defined in (33). We establish $\left| \widetilde{\mathcal{E}}(V_q) \right| / \sqrt{H(V_q, V_{q'})} \xrightarrow{p} 0$ under Condition (R3). Similarly, we establish $\left| \widetilde{\mathcal{E}}(V_{q'}) \right| / \sqrt{H(V_q, V_{q'})} \xrightarrow{p} 0$ under Condition (R3). That is, we establish $\frac{1}{\sqrt{H(V_q, V_{q'})}} \left| \widetilde{\mathcal{E}}(V_q) - \widetilde{\mathcal{E}}(V_{q'}) \right| \xrightarrow{p} 0$.

Now we prove $\mathcal{G}_n(V_q, V_{q'}) \xrightarrow{d} N(0, 1)$ and start with the decomposition,

$$\begin{aligned} \frac{f_{\mathcal{A}_1}^T \mathbf{M}(V_{q'}) \epsilon_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^T \mathbf{M}(V_{q'}) D_{\mathcal{A}_1}} - \frac{f_{\mathcal{A}_1}^T \mathbf{M}(V_q) \epsilon_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^T \mathbf{M}(V_q) D_{\mathcal{A}_1}} &= \frac{f_{\mathcal{A}_1}^T \mathbf{M}(V_{q'}) \epsilon_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^T \mathbf{M}(V_{q'}) f_{\mathcal{A}_1}} - \frac{f_{\mathcal{A}_1}^T \mathbf{M}(V_q) \epsilon_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^T \mathbf{M}(V_q) f_{\mathcal{A}_1}} \\ &+ \frac{f_{\mathcal{A}_1}^T \mathbf{M}(V_{q'}) \epsilon_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^T \mathbf{M}(V_{q'}) f_{\mathcal{A}_1}} \left(\frac{f_{\mathcal{A}_1}^T \mathbf{M}(V_{q'}) f_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^T \mathbf{M}(V_{q'}) D_{\mathcal{A}_1}} - 1 \right) - \frac{f_{\mathcal{A}_1}^T \mathbf{M}(V_q) \epsilon_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^T \mathbf{M}(V_q) f_{\mathcal{A}_1}} \left(\frac{f_{\mathcal{A}_1}^T \mathbf{M}(V_q) f_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^T \mathbf{M}(V_q) D_{\mathcal{A}_1}} - 1 \right). \end{aligned} \quad (82)$$

Since the vector S defined in (35) satisfies $\max_{i \in \mathcal{A}_1} S_i^2 / \sum_{i \in \mathcal{A}_1} S_i^2 \rightarrow 0$, we verify the Linderberg condition and establish

$$\frac{1}{\sqrt{H(V_q, V_{q'})}} \left(\frac{f_{\mathcal{A}_1}^T \mathbf{M}(V_{q'}) \epsilon_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^T \mathbf{M}(V_{q'}) f_{\mathcal{A}_1}} - \frac{f_{\mathcal{A}_1}^T \mathbf{M}(V_q) \epsilon_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^T \mathbf{M}(V_q) f_{\mathcal{A}_1}} \right) \xrightarrow{d} N(0, 1). \quad (83)$$

We apply (45) and (46) and establish that with probability larger than $1 - \exp(-c \min\{t_0^2, t_0\})$ for some positive constants $c > 0$ and $t_0 > 0$,

$$\left| \frac{f_{\mathcal{A}_1}^T \mathbf{M}(V_q) \epsilon_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^T \mathbf{M}(V_q) f_{\mathcal{A}_1}} \right| \lesssim \frac{t_0}{\sqrt{\mu(V_q)}}, \quad \text{and} \quad \left| \frac{f_{\mathcal{A}_1}^T \mathbf{M}(V_q) f_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^T \mathbf{M}(V_q) D_{\mathcal{A}_1}} - 1 \right| \lesssim \frac{\text{Tr}[\mathbf{M}(V_q)]}{\mu(V_q)} + \frac{t_0}{\sqrt{\mu(V_q)}}.$$

We apply the above inequalities and establish that with probability larger than $1 - \exp(-c \min\{t_0^2, t_0\})$ for some positive constants $c > 0$ and $t_0 > 0$,

$$\begin{aligned} &\frac{1}{\sqrt{H(V_q, V_{q'})}} \left| f_{\mathcal{A}_1}^T \mathbf{M}(V_{q'}) \epsilon_{\mathcal{A}_1} \left(\frac{f_{\mathcal{A}_1}^T \mathbf{M}(V_{q'}) f_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^T \mathbf{M}(V_{q'}) D_{\mathcal{A}_1}} - 1 \right) \right| \\ &\lesssim \frac{1}{\sqrt{H(V_q, V_{q'})}} \frac{t_0}{\sqrt{\mu(V_q)}} \left(\frac{\text{Tr}[\mathbf{M}(V_q)]}{\mu(V_q)} + \frac{t_0}{\sqrt{\mu(V_q)}} \right). \end{aligned}$$

Under the condition (R3), we establish

$$\frac{1}{\sqrt{H(V_q, V_{q'})}} \left| \frac{f_{\mathcal{A}_1}^\top \mathbf{M}(V_q) \epsilon_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^\top \mathbf{M}(V_{q'}) f_{\mathcal{A}_1}} \left(\frac{f_{\mathcal{A}_1}^\top \mathbf{M}(V_q) f_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\top \mathbf{M}(V_q) D_{\mathcal{A}_1}} - 1 \right) \right| \xrightarrow{p} 0.$$

Similarly, we have $\frac{1}{\sqrt{H(V_q, V_{q'})}} \left| \frac{f_{\mathcal{A}_1}^\top \mathbf{M}(V_{q'}) \epsilon_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^\top \mathbf{M}(V_{q'}) f_{\mathcal{A}_1}} \left(\frac{f_{\mathcal{A}_1}^\top \mathbf{M}(V_{q'}) f_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\top \mathbf{M}(V_{q'}) D_{\mathcal{A}_1}} - 1 \right) \right| \xrightarrow{p} 0$. The above inequalities and the decomposition (82) imply

$$\left| \mathcal{G}_n(V_q, V_{q'}) - \frac{1}{\sqrt{H(V_q, V_{q'})}} \left(\frac{f_{\mathcal{A}_1}^\top \mathbf{M}(V_{q'}) \epsilon_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^\top \mathbf{M}(V_{q'}) f_{\mathcal{A}_1}} - \frac{f_{\mathcal{A}_1}^\top \mathbf{M}(V_q) \epsilon_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^\top \mathbf{M}(V_q) f_{\mathcal{A}_1}} \right) \right| \xrightarrow{p} 0. \quad (84)$$

Together with (83), we establish $\mathcal{G}_n(V_q, V_{q'}) \xrightarrow{d} N(0, 1)$.

C.4 Proof of Lemma 5

Note that $[\mathbf{M}_{\text{RF}}(V)]^2 = (\Omega)^\top P_{\widehat{\mathcal{V}}_{\mathcal{A}_1}, \widehat{W}_{\mathcal{A}_1}}^\perp \Omega (\Omega)^\top P_{\widehat{\mathcal{V}}_{\mathcal{A}_1}, \widehat{W}_{\mathcal{A}_1}}^\perp \Omega$, and $\mathbf{M}_{\text{RF}}(V)$ and $[\mathbf{M}_{\text{RF}}(V)]^2$ are positive definite. For a vector $b \in \mathbb{R}^n$, we have

$$b^\top [\mathbf{M}_{\text{RF}}(V)]^2 b = (P_{\widehat{\mathcal{V}}_{\mathcal{A}_1}, \widehat{W}_{\mathcal{A}_1}}^\perp \Omega b)^\top \Omega (\Omega)^\top P_{\widehat{\mathcal{V}}_{\mathcal{A}_1}, \widehat{W}_{\mathcal{A}_1}}^\perp \Omega b \leq \|P_{\widehat{\mathcal{V}}_{\mathcal{A}_1}, \widehat{W}_{\mathcal{A}_1}}^\perp \Omega b\|_2^2 \|\Omega\|_2^2. \quad (85)$$

Let $\|\Omega\|_1$ and $\|\Omega\|_\infty$ denote the matrix 1 and ∞ norm, respectively. Since $\|\Omega\|_1 = 1$ and $\|\Omega\|_\infty = 1$ for the random forests setting, we have the upper bound for the spectral norm $\|\Omega\|_2^2 \leq \|\Omega\|_1 \cdot \|\Omega\|_\infty \leq 1$. Together with (85), we establish (62). Note that $b^\top \mathbf{M}_{\text{RF}}(V) b = b^\top (\Omega)^\top P_{\widehat{\mathcal{V}}_{\mathcal{A}_1}, \widehat{W}_{\mathcal{A}_1}}^\perp \Omega b \leq \|\Omega\|_2^2 \|b\|_2^2$, we establish that $\lambda_{\max}(\mathbf{M}_{\text{RF}}(V)) \leq 1$.

We apply the minimax expression of eigenvalues and obtain that

$$\begin{aligned} \lambda_k([\mathbf{M}_{\text{RF}}(V)]^2) &= \max_{U: \dim(U)=k} \min_{u \in U} \frac{u^\top [\mathbf{M}_{\text{RF}}(V)]^2 u}{\|u\|_2^2} \\ &\leq \max_{U: \dim(U)=k} \min_{u \in U} \frac{u^\top \mathbf{M}_{\text{RF}}(V) u}{\|u\|_2^2} = \lambda_k(\mathbf{M}_{\text{RF}}(V)). \end{aligned}$$

where the inequality follows from (62). The above inequality leads to $\text{Tr}([\mathbf{M}_{\text{RF}}(V)]^2) \leq \text{Tr}[\mathbf{M}_{\text{RF}}(V)]$. Since $P_{BW}^\perp P_{V_{BW}, W}^\perp = P_{BW}^\perp$, we have

$$\begin{aligned} [\mathbf{M}_{\text{ba}}(V)]^2 &= P_{BW} P_{V_{BW}, W}^\perp P_{BW} P_{V_{BW}, W}^\perp P_{BW} \\ &= P_{BW} P_{V_{BW}, W}^\perp (I - P_{BW}^\perp) P_{V_{BW}, W}^\perp P_{BW} = P_{BW} P_{V_{BW}, W}^\perp P_{BW} = \mathbf{M}_{\text{ba}}(V) \end{aligned}$$

Similar to the above proof, we establish $[\mathbf{M}_{\text{DNN}}(V)]^2 = \mathbf{M}_{\text{DNN}}(V)$. The proof of (64) is the same as that of (62) by replacing (85) with $b^\top [\mathbf{M}_{\text{RF}}(V)]^2 b \leq b^\top \mathbf{M}_{\text{RF}}(V) b \cdot \|\Omega\|_2^2$.

C.5 Proof of Lemma 6

By rewriting Lemma 3, we have that, with probability larger than $1 - n^{-c}$,

$$\max_{1 \leq i \leq n_1} |[\hat{\epsilon}(V)]_i - \epsilon_i| \leq C\kappa_n(V), \quad \kappa_n(V) = \sqrt{\log n} \left(\|R(V)\|_\infty + |\beta - \hat{\beta}_{\text{init}}(V)| + \frac{\log n}{\sqrt{n}} \right). \quad (86)$$

It is sufficient to show

$$\frac{(f_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1})^2}{\sum_{i=1}^{n_1} \sigma_i^2 [\mathbf{M}(V) f_{\mathcal{A}_1}]_i^2} \frac{\sum_{i=1}^{n_1} [\hat{\epsilon}(V)]_i^2 [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2}{(D_{\mathcal{A}_1}^\top \mathbf{M}(V) D_{\mathcal{A}_1})^2} \xrightarrow{p} 1. \quad (87)$$

Note that

$$\frac{\sum_{i=1}^{n_1} [\hat{\epsilon}(V)]_i^2 [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2}{\sum_{i=1}^{n_1} \sigma_i^2 [\mathbf{M}(V) f_{\mathcal{A}_1}]_i^2} = \frac{\sum_{i=1}^{n_1} \sigma_i^2 [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2}{\sum_{i=1}^{n_1} \sigma_i^2 [\mathbf{M}(V) f_{\mathcal{A}_1}]_i^2} \cdot \frac{\sum_{i=1}^{n_1} [\hat{\epsilon}(V)]_i^2 [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2}{\sum_{i=1}^{n_1} \sigma_i^2 [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2}. \quad (88)$$

We further decompose

$$\begin{aligned} & \frac{\sum_{i=1}^{n_1} [\hat{\epsilon}(V)]_i^2 [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2}{\sum_{i=1}^{n_1} \sigma_i^2 [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2} - 1 = \frac{\sum_{i=1}^{n_1} [\hat{\epsilon}(V) - \epsilon_i + \epsilon_i]^2 [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2}{\sum_{i=1}^{n_1} \sigma_i^2 [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2} - 1 \\ &= \frac{\sum_{i=1}^{n_1} [\hat{\epsilon}(V) - \epsilon_i]^2 [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2}{\sum_{i=1}^{n_1} \sigma_i^2 [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2} + \frac{\sum_{i=1}^{n_1} \epsilon_i [\hat{\epsilon}(V) - \epsilon_i] [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2}{\sum_{i=1}^{n_1} \sigma_i^2 [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2} + \frac{\sum_{i=1}^{n_1} (\epsilon_i^2 - \sigma_i^2) [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2}{\sum_{i=1}^{n_1} \sigma_i^2 [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2} \end{aligned} \quad (89)$$

Note that $\left| \frac{\sum_{i=1}^{n_1} (\epsilon_i^2 - \sigma_i^2) [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2}{\sum_{i=1}^{n_1} \sigma_i^2 [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2} \right| \lesssim \left| \frac{\sum_{i=1}^{n_1} (\epsilon_i^2 - \sigma_i^2) [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2}{\sum_{i=1}^{n_1} [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2} \right|$. Define the vector $a \in \mathbb{R}^{n_1}$ with $a_i = \frac{[\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2}{\sum_{i=1}^{n_1} [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2}$ and we have $\|a\|_1 = 1$ and $\|a\|_2^2 \leq \|a\|_\infty$. By applying Proposition 5.16 of [55], we establish that,

$$\mathbb{P} \left(\left| \sum_{i=1}^{n_1} a_i (\epsilon_i^2 - \sigma_i^2) \right| \geq t_0 K \|a\|_\infty \mid \mathcal{O} \right) \leq \exp(-c \min\{t_0^2, t_0\}). \quad (90)$$

By the condition $\|a\|_\infty \xrightarrow{p} 0$ and $\kappa_n(V)^2 + \sqrt{\log n} \kappa_n(V) \xrightarrow{p} 0$, we establish

$$\sum_{i=1}^{n_1} [\hat{\epsilon}(V)]_i^2 [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2 / \sum_{i=1}^{n_1} \sigma_i^2 [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2 \xrightarrow{p} 1.$$

By (88) and $\sum_{i=1}^{n_1} \sigma_i^2 [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2 / \sum_{i=1}^{n_1} \sigma_i^2 [\mathbf{M}(V) f_{\mathcal{A}_1}]_i^2 \xrightarrow{p} 1$, and $\frac{(f_{\mathcal{A}_1}^\top \mathbf{M}(V) f_{\mathcal{A}_1})^2}{(D_{\mathcal{A}_1}^\top \mathbf{M}(V) D_{\mathcal{A}_1})^2} \xrightarrow{p} 1$, we establish (87).

D Additional Simulation Results

D.1 Additional Results for the Binary IV

We consider an additional simulation setting with a binary IV. We set $p = 20$, $\beta = 1$, and generate $X_i \in \mathbb{R}^p$ in the same way as Models 1 and 2. We generate a binary IV as $Z_i = \mathbf{1}(\Phi(X_{i,6}^*) > 0.6)$ and consider the following way of generating f and g .

- Model 4 (binary IV): $f(Z_i, X_i) = Z_i \cdot (1 + a \cdot \sum_{j=1}^5 X_{ij}) - \sum_{j=1}^p 0.3X_{ij}$, and $g(Z_i, X_i) = Z_i + 0.2 \cdot \sum_{j=1}^p X_{ij}$.

We specify $\mathcal{V}_1 = \{z\}$ and implement TSCI with random forests as detailed in Algorithm 1 in the main paper.

In Table S3, we demonstrate our proposed TSCI method for Model 4 with both homoscedastic and heteroscedastic errors. The observations are coherent with those for Models 1 and 2 in Section 6 in the main paper. The main difference between the binary IV (Model 4) and the continuous IV (Models 1 and 2) is that the treatment effect is not identifiable for $a = 0$, which happens only for the binary IV setting. However, with a non-zero interaction and a relatively large sample size, our proposed TSCI methods achieve the desired coverage.

D.2 Binary Treatment

We consider the binary treatment setting and explore the finite sample performance of our proposed TSCI method. We consider the outcome model $Y_i = D_i\beta + Z_i + 0.2 \cdot \sum_{j=1}^p X_{ij} + \epsilon_i$ with $\beta = 1$ and $p = 20$. We generate ϵ_i and δ_i following Error distribution 1. We generate the binary treatment D_i with the conditional mean as

$$\mathbf{E}(D_i | Z_i, X_i) = \frac{\exp(f(Z_i, X_i) + \delta_i)}{1 + \exp(f(Z_i, X_i) + \delta_i)},$$

where $f(Z_i, X_i)$ is specified in the following two ways.

- Model 1 (continuous IV): generate Z_i, X_i and $f(Z_i, X_i)$ as Model 1 in the main paper.
- Model 4 (binary IV): we generate $f(Z_i, X_i) = Z_i \cdot (1 + a \cdot \sum_{j=1}^5 X_{ij}) - \sum_{j=1}^p 0.3X_{ij}$, which is the same as Model 4 in Section D.1.

			TSCI-RF									RF-Init		
			Bias			Length			Coverage			Invalidity	Bias	Coverage
Error	a	n	Oracle	Comp	Robust	Oracle	Comp	Robust	Oracle	Comp	Robust		Oracle	Oracle
1	0.0	1000	0.49	0.51	0.49	0.32	0.31	0.32	0.02	0.02	0.02	0.92	0.50	0.00
		3000	0.50	0.51	0.50	0.30	0.29	0.30	0.02	0.02	0.02	0.95	0.50	0.00
		5000	0.49	0.50	0.49	0.27	0.27	0.27	0.01	0.01	0.01	0.98	0.50	0.00
	0.5	1000	0.13	0.17	0.13	0.40	0.32	0.40	0.66	0.59	0.66	0.68	0.25	0.31
		3000	0.04	0.04	0.04	0.30	0.30	0.30	0.82	0.82	0.82	0.99	0.14	0.52
		5000	0.02	0.02	0.02	0.25	0.25	0.25	0.82	0.82	0.82	1.00	0.11	0.58
	1.0	1000	0.00	0.02	0.00	0.31	0.28	0.31	0.88	0.80	0.88	0.85	0.06	0.84
		3000	0.00	0.00	0.00	0.19	0.19	0.19	0.93	0.93	0.93	1.00	0.02	0.91
		5000	0.00	0.00	0.00	0.15	0.15	0.15	0.93	0.93	0.93	1.00	0.02	0.89
2	0.0	1000	0.60	0.63	0.60	0.36	0.33	0.36	0.01	0.01	0.01	0.80	0.60	0.00
		3000	0.60	0.61	0.60	0.31	0.29	0.31	0.01	0.01	0.01	0.90	0.60	0.00
		5000	0.60	0.61	0.60	0.28	0.26	0.28	0.00	0.00	0.00	0.92	0.60	0.00
	0.5	1000	0.11	0.16	0.11	0.41	0.32	0.41	0.69	0.62	0.69	0.68	0.26	0.31
		3000	0.03	0.03	0.03	0.29	0.29	0.29	0.83	0.83	0.83	1.00	0.14	0.48
		5000	0.01	0.01	0.01	0.24	0.24	0.24	0.88	0.88	0.88	1.00	0.11	0.57
	1.0	1000	0.00	0.02	0.00	0.30	0.28	0.30	0.93	0.87	0.93	0.90	0.06	0.88
		3000	0.00	0.00	0.00	0.18	0.18	0.18	0.94	0.94	0.94	1.00	0.03	0.89
		5000	0.00	0.00	0.00	0.14	0.14	0.14	0.96	0.96	0.96	1.00	0.02	0.91

Table S1: Bias, length, and coverage (at nominal level 0.95) for Model 4 (binary IV) with Error distributions 1 and 2. The columns indexed with “TSCI-RF” corresponds to our proposed TSCI with the random forests, where the columns indexed with “Bias”, “Length”, and “Coverage” correspond to the absolute bias of the point estimator, the length and empirical coverage of the constructed confidence interval respectively. The columns indexed with “Oracle”, “Comp” and “Robust” correspond to the TSCI estimators with \mathcal{V}_q selected by the oracle knowledge, the comparison method, and the robust method. The column indexed with “Invalidity” reports the proportion of detecting the proposed IV as invalid. The columns indexed with “RF-Init” correspond to the RF estimators without bias-correction but with the oracle knowledge of the best \mathcal{V}_q .

Table S2:

			TSCI-RF									RF-Init		TSCI-RF	
			Bias			Length			Coverage			Invalidity	Bias	Coverage	Weak IV
Setting	a	n	Orac	Comp	Robust	Orac	Comp	Robust	Orac	Comp	Robust		Orac	Orac	
1	0.0	1000	0.00	0.00	0.00	1.08	1.08	1.08	0.92	0.92	0.92	1.00	0.05	0.94	0.99
		3000	0.01	0.01	0.00	0.59	0.59	0.61	0.94	0.93	0.93	1.00	0.00	0.95	0.00
		5000	0.00	0.00	0.01	0.44	0.45	0.89	0.94	0.93	0.95	1.00	0.01	0.94	0.00
	0.5	1000	0.03	0.03	0.03	1.12	1.12	1.12	0.90	0.90	0.90	1.00	0.07	0.94	0.99
		3000	0.00	0.00	0.00	0.62	0.62	0.62	0.93	0.93	0.93	1.00	0.01	0.94	0.00
		5000	0.00	0.00	0.01	0.45	0.45	0.68	0.94	0.93	0.93	1.00	0.01	0.94	0.00
	1.0	1000	0.01	0.01	0.01	1.22	1.22	1.22	0.92	0.92	0.92	1.00	0.04	0.95	0.99
		3000	0.01	0.01	0.00	0.66	0.66	0.67	0.95	0.95	0.95	1.00	0.01	0.96	0.00
		5000	0.00	0.00	0.00	0.49	0.49	0.62	0.93	0.93	0.93	1.00	0.01	0.94	0.00
	1.5	1000	0.01	0.01	0.01	1.30	1.30	1.30	0.89	0.89	0.89	1.00	0.06	0.94	1.00
		3000	0.00	0.00	0.00	0.73	0.73	0.74	0.95	0.95	0.95	1.00	0.01	0.96	0.01
		5000	0.00	0.00	0.01	0.53	0.54	0.82	0.93	0.92	0.93	1.00	0.01	0.94	0.00
4	0.0	1000	0.37	0.37	0.37	1.62	1.62	1.62	0.66	0.66	0.66	0.93	0.41	0.78	1.00
		3000	0.41	0.41	0.41	1.25	1.25	1.25	0.60	0.60	0.60	1.00	0.42	0.72	1.00
		5000	0.35	0.35	0.35	1.05	1.05	1.05	0.61	0.61	0.61	1.00	0.40	0.65	1.00
	0.5	1000	0.30	0.30	0.30	1.21	1.21	1.21	0.65	0.65	0.65	1.00	0.36	0.77	1.00
		3000	0.21	0.21	0.21	1.18	1.18	1.18	0.73	0.73	0.73	1.00	0.29	0.83	1.00
		5000	0.10	0.10	0.10	1.09	1.09	1.09	0.82	0.82	0.82	1.00	0.22	0.89	1.00
	1.0	1000	0.16	0.20	0.16	1.35	1.30	1.35	0.77	0.77	0.77	0.93	0.26	0.87	1.00
		3000	0.03	0.03	0.03	1.11	1.10	1.11	0.88	0.88	0.88	0.99	0.11	0.94	1.00
		5000	0.00	0.00	0.00	0.92	0.92	0.92	0.89	0.89	0.89	1.00	0.07	0.93	0.57
	1.5	1000	0.06	0.22	0.06	1.59	1.31	1.59	0.83	0.67	0.83	0.75	0.18	0.93	1.00
		3000	0.01	0.02	0.01	1.14	1.13	1.14	0.90	0.90	0.90	0.98	0.08	0.93	0.87
		5000	0.00	0.00	0.00	0.91	0.91	0.91	0.93	0.93	0.93	1.00	0.04	0.95	0.02

Table S3: Bias, length, and coverage (at nominal level 0.95) for binary treatment model with Model 1 (continuous IV) and Model 4 (binary IV). The columns indexed with “TSCI-RF” corresponds to our proposed TSCI with the random forests, where the columns indexed with “Bias”, “Length”, and “Coverage” correspond to the absolute bias of the point estimator, the length and empirical coverage of the constructed confidence interval respectively. The columns indexed with “Oracle”, “Comp” and “Robust” correspond to the TSCI estimators with \mathcal{V}_q selected by the oracle knowledge, the comparison method, and the robust method. The column indexed with “Invalidity” reports the proportion of detecting the proposed IV as invalid. The columns indexed with “RF-Init” correspond to the RF estimators without bias-correction but with the oracle knowledge of the best \mathcal{V}_q . The column indexed with “Weak IV” stands for the proportion out of 500 simulations reporting $Q_{\max} < 1$.

The binary treatment result is summarized in Table D.2. The main observation is similar to those for the continuous treatment reported in the main paper. We shall point out the major differences in the following. The settings with the binary treatment are in general more challenging since the IV strength is relatively weak. To measure this, we have reported the column indexed with “weak IV” standing for the proportion of simulations with $Q_{\max} < 1$. For settings where our proposed generalized IV strength is strong such that $Q_{\max} \geq 1$, our proposed TSCI method achieves the desired coverage level. Even when the generalized IV strength leads to $Q_{\max} < 1$, our proposed (oracle) TSCI may still achieve the desired coverage level for setting 1. As a remark, even if $Q_{\max} < 1$, we still set $Q_{\max} = 1$ and implement Algorithm 1 in the main paper to select the index by comparison or robust methods.

D.3 Comparison with $\widehat{\beta}_{\text{EE}}(V)$

We compare our proposed estimator with

$$\widehat{\beta}_{\text{EE}}(V) = \frac{Y_{\mathcal{A}_1}^T P_{V_{\mathcal{A}_1}, W_{\mathcal{A}_1}}^\perp \widehat{f}_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^T P_{V_{\mathcal{A}_1}, W_{\mathcal{A}_1}}^\perp \widehat{f}_{\mathcal{A}_1}}. \quad (91)$$

by varying the generalized IV strength. We consider the data being generated in the same way as Model 1 and Error distribution 1 in the main paper, with the only difference as setting $f(Z_i, X_i)$ as

$$f(Z_i, X_i) = -\frac{25}{12} + Z_i + b \cdot (Z_i^2 + \frac{1}{8}Z_i^4) + Z_i \cdot (a \cdot \sum_{j=1}^5 X_{ij}) - 0.3 \cdot \sum_{j=1}^p X_{ij}. \quad (92)$$

We consider both settings with vio=1 and vio=2. For vio=1, we consider the no interaction setting with $a = 0$ and vary b across $\{0.1, 0.125, 0.15, 0.20, 1\}$. For vio=2, we set $b = 1$ and vary the interaction value a across $\{0, 0.5, 1\}$. In general, the larger values of a and b indicate a larger generalized IV strength. The comparison is reported in Table S4. If the IVs become weak after adjusting the violation, $\widehat{\beta}_{\text{EE}}(V)$ has an inflated bias and variance while our proposed TSCI method performs much more stably. When the IV becomes strong, then $\widehat{\beta}_{\text{EE}}(V)$ has a comparable performance to our proposed TSCI estimator.

D.4 Homoscedastic and heteroscedastic settings

We present the additional results for homoscedastic regression errors. For Model 1, we further report the mean absolute bias and the confidence interval length in Tables S5 and

Vio=1, a = 0											
		Bias			Coverage			Length		IV Strength	Weak IV
b	n	TSCI-RF	RF-Init	RF-EE	TSCI-RF	RF-Init	RF-EE	TSCI-RF	RF-EE	TSCI-RF	TSCI-RF
0.10	1000	0.10	0.23	90.61	0.70	0.53	0.91	0.53	34426.03	2.83	1.00
	3000	0.02	0.11	0.01	0.88	0.79	0.97	0.44	0.87	17.55	0.96
	5000	0.00	0.07	0.02	0.93	0.88	0.98	0.37	0.60	35.56	0.56
0.125	1000	0.06	0.18	0.57	0.79	0.65	0.97	0.51	47.63	3.64	1.00
	3000	0.01	0.08	0.01	0.90	0.84	0.97	0.39	0.59	24.46	0.67
	5000	0.00	0.05	0.01	0.92	0.87	0.97	0.32	0.41	51.03	0.07
0.15	1000	0.04	0.15	0.00	0.82	0.72	0.97	0.48	1.20	4.53	1.00
	3000	0.00	0.06	0.01	0.92	0.87	0.97	0.35	0.44	31.83	0.31
	5000	0.01	0.05	0.00	0.92	0.89	0.95	0.28	0.32	68.29	0.00
0.20	1000	0.02	0.10	0.01	0.89	0.79	0.97	0.44	0.63	6.65	1.00
	3000	0.01	0.03	0.00	0.93	0.90	0.96	0.29	0.30	47.94	0.01
	5000	0.01	0.02	0.00	0.92	0.91	0.94	0.23	0.22	106.80	0.00
1	1000	0.00	0.00	0.00	0.93	0.93	0.96	0.10	0.09	279.07	0.00
	3000	0.00	0.00	0.00	0.95	0.94	0.97	0.05	0.05	2273.12	0.00
	5000	0.00	0.00	0.00	0.96	0.96	0.95	0.04	0.04	4911.53	0.00

Vio = 2, b = 1

		Bias			Coverage			Length		IV Strength	Weak IV
a	n	TSCI-RF	RF-Init	RF-EE	TSCI-RF	RF-Init	RF-EE	TSCI-RF	RF-EE	TSCI-RF	TSCI-RF
0.0	1000	0.24	0.33	1930.84	0.48	0.24	0.98	0.48	6441243.82	2.62	0.00
	3000	0.07	0.19	3.99	0.78	0.58	0.98	0.43	1020.11	13.60	0.00
	5000	0.03	0.14	0.00	0.85	0.66	0.96	0.37	1.03	27.94	0.00
0.5	1000	0.04	0.12	0.13	0.84	0.64	0.98	0.31	6.76	7.72	0.00
	3000	0.01	0.05	0.00	0.87	0.77	0.94	0.19	0.25	57.10	0.00
	5000	0.00	0.04	0.00	0.93	0.83	0.96	0.15	0.17	129.58	0.00
1.0	1000	0.00	0.03	0.00	0.91	0.89	0.96	0.17	0.24	28.42	0.00
	3000	0.00	0.01	0.00	0.93	0.90	0.96	0.09	0.10	272.83	0.00
	5000	0.00	0.01	0.00	0.93	0.92	0.96	0.07	0.07	660.27	0.00

Table S4: Comparison of TSCI and $\widehat{\beta}_{\text{EE}}(V)$ in terms of Bias, length, and coverage (at nominal level 0.95) for the model (92). The columns indexed with “TSCI-RF”, “RF-EE”, and “RF-Init” corresponds to our proposed TSCI with the random forests, $\widehat{\beta}_{\text{EE}}(V)$ defined in (91), and the RF estimators without bias-correction, respectively. All methods are implemented with the best \mathcal{V}_q determined by the oracle knowledge. The columns indexed with “Bias”, “Length”, and “Coverage” correspond to the absolute bias of the point estimator, the length and empirical coverage of the constructed confidence interval respectively. The columns indexed with “IV strength” stands for the average generalized IV strength over 500 simulations. The column indexed with “Weak IV” stands for the proportion out of 500 simulations reporting $Q_{\max} < 1$.

[S6](#), respectively. TSLS has a large bias due to the existence of invalid IVs; even in the oracle setting with the prior knowledge of the best \mathcal{V}_q approximating g , RF-Full, and RF-Plug have a large bias, and RF-Init has a large bias for vio=2. Our proposed TSCI method corrects the bias effectively. This explains why TSLS, RF-Full, RF-Plug, and RF-Init fail to have correct coverage probability 0.95. In table [S6](#), the lengths of confidence intervals by TSCI-RF are in general smaller than those by TSCI-ba due to the reason that TSCI-RF tends to capture the more complex relationship and leads to stronger IVs. The robust selection methods typically lead to longer confidence intervals since more violation forms might be adjusted with the robust selection. For Model 2, we report the empirical coverage in Table [S7](#). The observations for Model 2 are generally similar to those for Model 1.

In table [S8](#), we consider Model 1 with Error distribution 2 and compare the performance of RF-Init and our proposed TSCI with random forests. The results are similar to that for Model 2 in table [2](#). We note that TSCI effectively corrects the bias of the RF-Init.

			TSCI-RF			TSCI-ba			TSLS	Other RF(oracle)		
vio	a	n	Oracle	Comp	Robust	Oracle	Comp	Robust		Init	Plug	Full
1	0.0	1000	0.00	0.00	0.00	0.00	0.00	0.01	1.00	0.00	0.14	0.02
		3000	0.00	0.00	0.00	0.00	0.00	0.02	1.00	0.00	0.08	0.01
		5000	0.00	0.00	0.00	0.00	0.00	0.01	1.00	0.00	0.06	0.01
	0.5	1000	0.00	0.00	0.00	0.00	0.00	0.01	0.44	0.00	0.12	0.02
		3000	0.00	0.00	0.00	0.00	0.00	0.01	0.44	0.00	0.08	0.01
		5000	0.00	0.00	0.00	0.00	0.00	0.01	0.45	0.00	0.07	0.01
	1.0	1000	0.00	0.00	0.00	0.00	0.00	0.01	0.29	0.00	0.12	0.02
		3000	0.00	0.00	0.00	0.00	0.00	0.01	0.29	0.00	0.10	0.01
		5000	0.00	0.00	0.00	0.00	0.00	0.01	0.29	0.00	0.09	0.01
2	0.0	1000	0.24	0.69	0.69	0.05	0.65	0.65	1.00	0.33	0.98	0.41
		3000	0.07	0.69	0.69	0.02	0.04	0.03	1.00	0.19	0.81	0.37
		5000	0.03	0.67	0.67	0.02	0.01	0.01	1.00	0.14	0.74	0.35
	0.5	1000	0.04	0.60	0.60	0.04	0.56	0.55	0.44	0.12	0.55	0.21
		3000	0.01	0.01	0.01	0.01	0.02	0.01	0.44	0.05	0.03	0.17
		5000	0.00	0.00	0.00	0.01	0.01	0.01	0.44	0.04	0.11	0.16
	1.0	1000	0.00	0.01	0.01	0.03	0.53	0.53	0.29	0.03	0.18	0.08
		3000	0.00	0.00	0.00	0.01	0.04	0.03	0.29	0.01	0.39	0.06
		5000	0.00	0.00	0.00	0.01	0.00	0.00	0.29	0.01	0.37	0.06

Table S5: Mean absolute bias for Model 1 with Error distribution 1. The columns indexed with “TSCI-RF” and “TSCI-ba” correspond to our proposed TSCI with the random forests and the basis approximation, where the columns indexed with “Oracle”, “Comp” and “Robust” correspond to the estimators with \mathcal{V}_q selected by the oracle knowledge, the comparison method, and the robust method. The columns index “TSLS” corresponds to the TSLS estimator. The columns indexed with “Init”, “Plug”, “Full” correspond to the RF estimators without bias-correction, the plug-in RF estimator and the no data-splitting RF estimator, with the oracle knowledge of the best \mathcal{V}_q .

			TSCI-RF			TSCI-ba			TSLS	Other RF(oracle)		
vio	a	n	Oracle	Comp	Robust	Oracle	Comp	Robust		Init	Plug	Full
1	0.0	1000	0.10	0.10	0.10	0.07	0.08	0.12	0.25	0.10	0.10	0.07
		3000	0.05	0.05	0.05	0.04	0.05	0.46	0.14	0.05	0.06	0.04
		5000	0.04	0.04	0.05	0.03	0.03	0.36	0.11	0.04	0.04	0.03
	0.5	1000	0.09	0.09	0.09	0.07	0.07	0.14	0.07	0.09	0.09	0.07
		3000	0.05	0.05	0.19	0.04	0.04	0.38	0.04	0.05	0.05	0.04
		5000	0.04	0.04	0.15	0.03	0.03	0.30	0.03	0.04	0.04	0.03
	1.0	1000	0.08	0.09	0.17	0.06	0.06	0.10	0.04	0.08	0.08	0.06
		3000	0.05	0.05	0.09	0.04	0.04	0.31	0.02	0.05	0.04	0.03
		5000	0.03	0.03	0.07	0.03	0.03	0.25	0.02	0.03	0.03	0.03
2	0.0	1000	0.48	0.09	0.09	0.79	0.10	0.11	0.14	0.48	0.82	0.23
		3000	0.43	0.05	0.05	0.46	0.45	0.46	0.08	0.43	0.49	0.15
		5000	0.37	0.04	0.04	0.36	0.36	0.37	0.06	0.37	0.39	0.12
	0.5	1000	0.31	0.09	0.09	0.66	0.13	0.13	0.05	0.31	0.58	0.18
		3000	0.19	0.19	0.19	0.38	0.38	0.38	0.03	0.19	0.26	0.10
		5000	0.15	0.15	0.15	0.30	0.30	0.30	0.02	0.15	0.18	0.08
	1.0	1000	0.17	0.17	0.17	0.56	0.10	0.10	0.04	0.17	0.27	0.11
		3000	0.09	0.09	0.09	0.32	0.31	0.31	0.02	0.09	0.13	0.06
		5000	0.07	0.07	0.07	0.25	0.25	0.25	0.02	0.07	0.09	0.05

Table S6: Average confidence interval length for Model 1 with Error distribution 1. The columns indexed with “TSCI-RF” “TSCI-ba” correspond to our proposed TSCI with the random forests and the basis approximation, where the columns indexed with “Oracle”, “Comp” and “Robust” correspond to the estimators with \mathcal{V}_q selected by the oracle knowledge, the comparison method, and the robust method. The columns index “TSLS” corresponds to the TSLS estimator. The columns indexed with “Init”, “Plug”, “Full” correspond to the RF estimators without bias-correction, the plug-in RF estimator and the no data-splitting RF estimator, with the oracle knowledge of the best \mathcal{V}_q .

			TSCI-RF			Invalidity	TSCI-ba			TSLS	Other RF(oracle)		
vio	a	n	Oracle	Comp	Robust	TSCI-RF	Oracle	Comp	Robust		Init	Plug	Full
1	0.0	1000	0.86	0.86	0.86	1.00	0.92	0.92	0.91	0.42	0.82	0.07	0.06
		3000	0.96	0.96	0.95	1.00	0.95	0.95	0.95	0.03	0.96	0.00	0.11
		5000	0.95	0.95	0.95	1.00	0.95	0.95	0.95	0.00	0.93	0.00	0.07
	0.5	1000	0.92	0.92	0.92	1.00	0.95	0.95	0.94	0.00	0.91	0.00	0.23
		3000	0.93	0.93	0.94	1.00	0.93	0.93	0.93	0.00	0.93	0.00	0.13
		5000	0.95	0.95	0.95	1.00	0.95	0.95	0.94	0.00	0.95	0.00	0.08
	1.0	1000	0.94	0.93	0.93	1.00	0.96	0.96	0.95	0.00	0.94	0.00	0.23
		3000	0.95	0.95	0.95	1.00	0.95	0.95	0.95	0.00	0.95	0.00	0.09
		5000	0.94	0.94	0.94	1.00	0.95	0.95	0.95	0.00	0.93	0.00	0.03
2	0.0	1000	0.82	0.16	0.16	0.99	0.94	0.94	0.94	0.41	0.69	0.09	0.06
		3000	0.95	0.95	0.96	1.00	0.94	0.94	0.95	0.03	0.94	0.00	0.09
		5000	0.95	0.95	0.96	1.00	0.94	0.94	0.93	0.00	0.94	0.00	0.11
	0.5	1000	0.94	0.89	0.88	0.97	0.93	0.93	0.94	0.00	0.92	0.00	0.24
		3000	0.93	0.93	0.93	1.00	0.94	0.94	0.95	0.00	0.93	0.00	0.15
		5000	0.96	0.96	0.95	1.00	0.95	0.95	0.95	0.00	0.95	0.00	0.07
	1.0	1000	0.92	0.92	0.92	1.00	0.95	0.95	0.95	0.00	0.91	0.00	0.19
		3000	0.95	0.95	0.95	1.00	0.95	0.95	0.94	0.00	0.94	0.00	0.06
		5000	0.95	0.95	0.94	1.00	0.97	0.97	0.96	0.00	0.93	0.00	0.03

Table S7: Empirical coverage for Model 2 with Error distribution 1. The columns indexed with “TSCI-RF” “TSCI-ba” correspond to our proposed TSCI with the random forests and the basis approximation, where the columns indexed with “Oracle”, “Comp” and “Robust” correspond to the estimators with \mathcal{V}_q selected by the oracle knowledge, the comparison method, and the robust method. The column indexed with “Invalidity” reports the proportion of detecting the proposed IV as invalid. The columns index “TSLS” corresponds to the TSLS estimator. The columns indexed with “Init”, “Plug”, “Full” correspond to the RF estimators without bias-correction, the plug-in RF estimator and the no data-splitting RF estimator, with the oracle knowledge of the best \mathcal{V}_q .

			TSCI-RF									RF-Init		
			Bias			Length			Coverage			Invalidity	Bias	Coverage
vio	a	n	Oracle	Comp	Robust	Oracle	Comp	Robust	Oracle	Comp	Robust		Oracle	Oracle
1	0.0	1000	0.00	0.00	0.00	0.15	0.15	0.15	0.92	0.92	0.92	1.00	0.01	0.92
		3000	0.00	0.00	0.00	0.08	0.08	0.08	0.96	0.96	0.96	1.00	0.00	0.95
		5000	0.00	0.00	0.00	0.06	0.06	0.06	0.94	0.93	0.93	1.00	0.00	0.94
	0.5	1000	0.00	0.00	0.00	0.13	0.13	0.13	0.94	0.94	0.94	1.00	0.01	0.93
		3000	0.00	0.00	0.01	0.08	0.08	0.25	0.94	0.90	0.88	1.00	0.00	0.94
		5000	0.00	0.00	0.00	0.06	0.06	0.20	0.95	0.92	0.89	1.00	0.00	0.95
	1.0	1000	0.00	0.00	0.01	0.11	0.12	0.20	0.92	0.90	0.91	1.00	0.01	0.92
		3000	0.00	0.00	0.00	0.06	0.06	0.13	0.95	0.94	0.92	1.00	0.00	0.94
		5000	0.00	0.00	0.00	0.05	0.05	0.10	0.95	0.94	0.94	1.00	0.00	0.94
2	0.0	1000	0.38	0.69	0.69	0.47	0.11	0.11	0.28	0.00	0.00	1.00	0.47	0.09
		3000	0.17	0.69	0.69	0.47	0.07	0.07	0.59	0.00	0.00	1.00	0.32	0.26
		5000	0.09	0.68	0.68	0.44	0.05	0.05	0.70	0.00	0.00	1.00	0.25	0.38
	0.5	1000	0.07	0.61	0.61	0.36	0.11	0.11	0.75	0.00	0.00	0.89	0.20	0.40
		3000	0.02	0.03	0.03	0.25	0.25	0.25	0.90	0.88	0.89	1.00	0.09	0.68
		5000	0.00	0.00	0.00	0.20	0.20	0.20	0.89	0.89	0.89	1.00	0.06	0.75
	1.0	1000	0.01	0.12	0.12	0.23	0.20	0.20	0.91	0.71	0.71	1.00	0.05	0.85
		3000	0.00	0.00	0.00	0.13	0.13	0.13	0.94	0.94	0.94	1.00	0.02	0.89
		5000	0.00	0.00	0.00	0.10	0.10	0.10	0.92	0.92	0.92	1.00	0.01	0.91

Table S8: Bias, length, and coverage for Model 1 with Error distribution 2. The columns indexed with “TSCI-RF” corresponds to our proposed TSCI with the random forests, where the columns indexed with “Bias”, “Length”, and “Coverage” correspond to the absolute bias of the point estimator, the length and empirical coverage of the constructed confidence interval respectively. The columns indexed with “Oracle”, “Comp” and “Robust” correspond to the TSCI estimators with \mathcal{V}_q selected by the oracle knowledge, the comparison method, and the robust method. The column indexed with “Invalidity” reports the proportion of detecting the proposed IV as invalid. The columns indexed with “RF-Init” correspond to the RF estimators without bias-correction but with the oracle knowledge of the best \mathcal{V}_q .