Journal of Machine Learning Research 27 (2026) 1-67 Submitted 4/24; Revised 12/25; Published 3/26

Robustness Against Weak or Invalid Instruments: Exploring Nonlinear Treatment Models with Machine Learning

Zijian Guo ZIJGUO@ZJU.EDU.CN
Center for Data Science
Zhejiang University
Hangzhou, Zhejiang 310058, China

Mengchu Zheng MZ666@STAT.RUTGERS.EDU
Department of Statistics
Rutgers University
Piscataway, NJ 08854, USA

Peter Bühlmann BUHLMANN@STAT.MATH.ETHZ.CH
Seminar für Statistik
ETH Zürich
Zürich, 8092, Switzerland

Editor: Joris Mooij

Abstract

We discuss causal inference for observational studies with possibly invalid instrumental variables. We propose a novel methodology called two-stage curvature identification (`TSCI`) by exploring the nonlinear treatment model with machine learning. The first-stage machine learning enables improving the instrumental variable's strength and adjusting for different forms of violating the instrumental variable assumptions. The success of `TSCI` requires the instrumental variable's effect on treatment to differ from its violation form. A novel bias correction step is implemented to remove bias resulting from the potentially high complexity of machine learning. Our proposed `TSCI` estimator is shown to be asymptotically unbiased and Gaussian even if the machine learning algorithm does not consistently estimate the treatment model. Furthermore, we design a data-dependent method to choose the best among several candidate violation forms. We apply `TSCI` to study the effect of education on earnings.

Keywords: generalized IV strength, confidence interval, random forests, boosting, neural network

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 assumptions: conditioning on the baseline covariates, (A1) the IVs are associated with the treatment; (A2) the IVs are not associated with the unmeasured confounders; (A3) the IVs do not directly affect the outcome.

Despite the popularity of the IV method, there is a significant concern about whether the used IVs satisfy (A1)-(A3) in practice. 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 (Stock et al., 2002; Staiger and Stock, 1997, e.g.,). Assumptions (A2) and (A3) ensure that the IV only affects the outcome through the treatment. If an IV violates either (A2) or (A3), we call it an invalid IV and define its functional form of violating (A2) and (A3) as the violation form, e.g., a linear violation. In the just-identification regime, most empirical analyses rely on external knowledge to argue about the validity of (A2) and (A3). However, there is a pressing need to develop a robust causal inference method against the proposed IVs violating the classical assumptions.

1.1 Our Results and Contribution

We aim to devise a robust IV framework that leads to causal identification even if the IVs proposed by domain experts may violate assumptions (A1) to (A3). Our framework provides a robustness guarantee even when all proposed IVs are invalid, including the most common regime with a single IV that is possibly invalid. It is well known that the treatment effect is not identifiable when there is no constraint on how the IVs violate assumptions (A2) and (A3). Our key identification assumption is that the violations of (A2) and (A3) arise from simpler forms than the association between the treatment and the IVs; that is, we exclude "special coincidences" such as the IVs violate (A2) and (A3) by linear forms and the conditional mean model of the treatment given IVs is also linear. This identification condition can be evaluated with the generalized IV strength measure introduced in Section 4.3 for a user-specific violation form.

We propose a novel two-stage curvature identification (`TSCI`) method to infer the treatment effect. An important operational step is to fit the treatment model with a machine learning (ML) algorithm, e.g., random forests, boosting, or deep neural networks. This first-stage ML model enhances the IV's strength by learning a general conditional mean model of treatment given the IVs, instead of restricting to a linear association model. In the second stage, we leverage the ML prediction and adjust for different IV violation forms. A main novelty of our proposal is to estimate the bias resulting from high complexity of the first-stage ML and implement a corresponding bias correction step; see more details in Section 4.2.

We show that the `TSCI` estimator is asymptotically Gaussian when the generalized IV strength measure is sufficiently large. We further devise a data-dependent way of choosing the most suitable violation form among a collection of IV violation forms. By including the valid IV setting into the selection, our proposed general `TSCI` methodology does not exclude the possibility of the IVs being valid but provides a robust IV method against different invalidity forms.

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

1. `TSCI` explores a nonlinear treatment model with ML and leads to more reliable causal conclusions than existing methods assuming valid IVs.

2. `TSCI` provides an efficient way of integrating the first-stage ML. A methodological novelty of `TSCI` is its bias correction in its second stage, which addresses the issue of high complexity and potential overfitting of ML.

The proposed method is implemented in the `R` package `TSCI` available from CRAN (Carl et al., 2025).

1.2 Comparison to Existing Literature

There is relevant literature on causal inference when assumptions (A2) and (A3) are violated. When all IVs are invalid, Lewbel (2012); Tchetgen et al. (2021) proposed an identification strategy by assuming that the treatment model has heteroscedastic error but the covariance between the treatment and outcome model errors is homoscedastic. Our proposed `TSCI` is based on a different idea by exploring the nonlinear structure between the treatment and the IVs, not requiring anything on homo- and heteroscedasticity. Bowden et al. (2015) and Kolesár et al. (2015) assumed that the direct effect of the IVs on the outcome is nearly orthogonal to the effect of the IVs on the treatment. Han (2008); Bowden et al. (2016); Kang et al. (2016); Windmeijer et al. (2019); Guo et al. (2018); Windmeijer et al. (2021) considered the setting that a proportion of IVs are valid and conducted causal inference by selecting valid IVs in a data-dependent way. Along this line, Guo (2023) proposed a uniform inference procedure that is robust to the IV selection error. More recently, Sun et al. (2023) leveraged the number of valid IVs and used the interaction of genetic markers to identify the causal effect. 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.

The nonlinear structure of the treatment model has been explored in the econometric and the more recent ML literature. In the IV literature, Kelejian (1971); Amemiya (1974); Newey (1990); Hausman et al. (2012) considered constructing a non-parametric treatment model and Belloni et al. (2012) proposed constructing the optimal IVs with a Lasso-type first-stage estimator. More recently, Chen et al. (2023); Liu et al. (2020) proposed constructing the IV as the first-stage prediction given by the ML algorithm. All of the above works are focused on the valid IV setting. In contrast, our current paper applies to the more robust regime where the IVs are allowed to violate assumptions (A2) or (A3) in a range of forms. Even in the context of valid IVs, our proposed `TSCI` may lead to a more accurate estimator than directly using the predicted value for the treatment model as the new IV (Chen et al., 2023; Liu et al., 2020); see Section 4.6 for details.

ML algorithms integrated into the IV analysis to better accommodate the complicated outcome model. In Sections 4.6 and 6.1, we provide detailed comparisons to Double Machine Learning (`DML`) proposed in Chernozhukov et al. (2018). Athey et al. (2019) proposed the generalized random forests to infer heterogeneous treatment effects. Both works are based on valid IVs, while our current paper assumes a homogeneous treatment effect and focuses on the different robust IV framework with possibly invalid IVs.

Finally, our proposed TSCI methodology can be used to check the IV validity for the just-identification case. This is distinct from the Sargan test, J test, or specification test (Hansen, 1982; Sargan, 1958; Woutersen and Hausman, 2019) which are used 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 $\mathcal{A} \subset \{1, \cdots, n\}$, 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 its cardinality. For a vector $x \in \mathbb{R}^p$, the ℓ_q norm of x is defined as $\|x\|_q = \left(\sum_{l=1}^p |x_l|^q\right)^{\frac{1}{q}}$ 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 $a_n \leq C b_n$ 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 $\limsup_{n\to\infty} a_n/b_n = 0$. For a matrix M, we use $\mathrm{Tr}[M]$ to denote its trace, $\mathrm{rank}(M)$ 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 Causal Interpretation

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 IVs and measured covariates.

2.1 Examples for Effect Identification with Nonlinear Treatment Models

To explain the main idea, we start with the special case with no covariates,

$$Y_i = D_i\beta + Z_i^\intercal \pi + \epsilon_i, \quad \text{and} \quad D_i = f(Z_i) + \delta_i, \quad \text{for} \quad 1 \leq i \leq n, \tag{1}$$

where the errors ϵ_i and δ_i satisfy $\mathbf{E}(\epsilon_i \mid Z_i) = 0$ and $\mathbf{E}(\delta_i \mid Z_i) = 0$. These errors ϵ_i and δ_i are typically correlated due to unobserved confounding. The outcome model in (1) is commonly used in the invalid IV literature (Small, 2007; Kang et al., 2016; Guo et al., 2018; Windmeijer et al., 2019), with $\pi \neq 0$ standing for the violation of IV assumptions (A2) and (A3). The treatment model in (1) is not a causal generating model between D_i and Z_i but only stands for a probability model between D_i and Z_i. Specifically, the treatment assignment might be directly influenced by unmeasured confounders, but the treatment model in (1) does not explicitly state this generating process. Instead, for the joint distribution of D_i and Z_i, we define the conditional expectation $f(Z_i) = \mathbf{E}(D_i \mid Z_i)$, which leads to the treatment model in (1). We discuss the structural equation model interpretation of the model (1) in Section 2.3, which explicitly characterizes how unmeasured confounders may affect the treatment assignment.

We introduce some projection notations to facilitate the discussion. For a random variable U_i, define its best linear approximation by Z_i as $\gamma^* = \arg\min_{\gamma\in\mathbb{R}^{p_z+1}} \mathbf{E}(U_i - \widetilde{Z}_i^\intercal\gamma)^2$ with $\widetilde{Z}_i = (1, Z_i^\intercal)^\intercal$. For $1 \leq i \leq n$, write $\mathcal{P}_{Z_i}U_i = \widetilde{Z}_i^\intercal\gamma^*$ and $\mathcal{P}_{Z_i}^\perp U_i = U_i - \widetilde{Z}_i^\intercal\gamma^*$. We apply $\mathcal{P}_{Z_i}^\perp$ to the model (1) and obtain $\mathcal{P}_{Z_i}^\perp Y_i = \mathcal{P}_{Z_i}^\perp D_i\beta + \mathcal{P}_{Z_i}^\perp\epsilon_i$, where the linear violation form $Z_i^\intercal\pi$ in (1) is removed after applying the transformation $\mathcal{P}_{Z_i}^\perp$. If f is nonlinear, then $\mathrm{Var}\left(\mathcal{P}_{Z_i}^\perp f(Z_i)\right) > 0$ and the adjusted variable $\mathcal{P}_{Z_i}^\perp f(Z_i)$ can be used as a valid IV to identify

the effect β via the following estimating equation

$$\mathbf{E}\left[\mathcal{P}_{Z_i}^{\perp} f(Z_i)(Y_i - D_i\beta)\right] = \mathbf{E}\left[\mathcal{P}_{Z_i}^{\perp} f(Z_i)(Z_i^{\intercal}\pi + \epsilon_i)\right] = 0. \tag{2}$$

Note that the last equality follows from the population orthogonality between $\mathcal{P}_{Z_i}^{\perp} f(Z_i)$ and Z_i, together with $\mathbf{E}(\epsilon_i \mid Z_i) = 0$.

The estimation equation in (2) reveals that the treatment effect β can be identified by exploring the nonlinear structure of f under the model (1). We propose to learn f using a ML prediction model and measure the variability of $\mathcal{P}_{Z_i}^{\perp} f(Z_i)$ by defining a notation of generalized IV strength in Section 4.3. We shall emphasize that using ML algorithms to learn the possibly nonlinear function f is critical in the current framework, allowing for invalid IVs. The ML algorithms help capture complicated structures in f and typically retain enough variability contained in the adjusted variable $\mathcal{P}_{Z_i}^{\perp} f(Z_i)$.

Nonlinear functional form identification. Nonlinear functional form identification has been proposed in various existing literature; see Lewbel (2019)[Sec.3.7] for a review. More concretely, Angrist and Pischke (2009)[Sec.4.6.1] discussed the identification of when the IV has a linear direct effect on the outcome, but the treatment model is a nonlinear parametric model of the IV. The Heckman selection model (Heckman, 1976) offered a practical and simple solution for estimating effects when the samples are not randomly selected. When no variables satisfy the exclusion restriction, one possible identification relies on the nonlinearity of the inverse Mills ratio. As pointed by Puhani (2000), such an identification is less reliable when the nonlinear model form is relatively weak in applications. Additionally, Shardell and Ferrucci (2016); Have et al. (2007) proposed the causal identification using the interaction of the (binary) IV and baseline covariates as the new IV. Lewbel (2019) commented that the proposed identification fails when the nonlinear functional form does not exist or is relatively weak. As a main difference, our proposal does not use hand-picked instruments or assume the data-generating distribution has some pre-specified nonlinear functional forms. Operationally, we apply the first-stage ML algorithm to learn the complex nonlinear treatment model, which is more powerful in detecting the nonlinear dependence based on exploring the data. However, including the first-stage ML requires a more delicate following-up statistical procedure, such as the bias correction step in (18). More importantly, instead of betting on the existence of nonlinearity, our proposed generalized IV strength measure in (21) guides whether there is enough nonlinear structure after adjusting for invalid forms.

2.2 Model with a General Class of Invalid IVs

We now introduce the general outcome model, allowing for more complicated violation forms than the linear violation in (1),

$$Y_i = D_i\beta + g(Z_i, X_i) + \epsilon_i, \quad \text{with} \quad \mathbf{E}(\epsilon_i \mid X_i, Z_i) = 0, \quad \text{for} \quad 1 \le i \le n, \tag{3}$$

where $\beta \in \mathbb{R}$ is the homogeneous treatment effect and the function $g : \mathbb{R}^{p_x+p_z} \to \mathbb{R}$ encodes how the IVs violate the assumptions (A2) and (A3); see more detailed explanations from the potential outcome model perspective in Section 2.3. The classical valid IV setting requires that the function $g(z, x)$ does not change with different assignment of z. In the invalid IV

literature, the commonly used outcome model with a linear violation form can be viewed as a special case of (3) when taking $g(Z_i, X_i) = Z_i^\intercal \pi_z + X_i^\intercal \pi_x$. In this paper, we focus on the model (3) but shall point out that (3) is restrictive by assuming a linear model and homogeneous effect β. The proposal in the current paper relies on such a partial linear model assumption while we provide a potential extension of the proposed procedure to a linear treatment model allowing for heterogeneous causal effects in Section 8.

For the treatment, we define its conditional mean function $f(Z_i, X_i) \coloneqq \mathbf{E}(D_i \mid Z_i, X_i)$ for $1 \le i \le n$, leading to the following treatment model:

$$D_i = f(Z_i, X_i) + \delta_i \quad \text{with} \quad \mathbf{E}(\delta_i \mid Z_i, X_i) = 0. \tag{4}$$

The model (4) is flexible as f might be any unknown function of Z_i and X_i, and the treatment variable can be continuous, binary, or a count variable. Similarly to (1), the model (4) is only a probability relation instead of a causal generating model and the treatment assignment is allowed to be influenced by unmeasured confounders.

It is well known that, for the outcome model (3), the identification of β is generally impossible without additional identifiability conditions on the function g. In this paper, we allow the existence of invalid IVs but constrain the functional form of violating (A2) and (A3) to a pre-specified class. An operational step is to specify a set of basis functions for $g(\cdot)$ in the outcome model (3).

In the following, we discuss approximating $g(z, x)$ by a class of basis functions defined in (5) and further express the outcome model (3) via the basis expansions in the following Equation (8). Define $\phi(X_i) = \mathbf{E}[g(Z_i, X_i) \mid X_i]$ and decompose g as $g(Z_i, X_i) = h(Z_i, X_i) + \phi(X_i)$ with $h(Z_i, X_i) = g(Z_i, X_i) - \phi(X_i)$. When the IVs Z_i are valid, $g(Z_i, X_i)$ does not directly depend on the assignment of Z_i, which implies $\phi(X_i) = g(Z_i, X_i)$ and $h(\cdot) = 0$. In this case, we just need to approximate $g(z, x) = \phi(x)$ by a set of basis functions $\overrightarrow{\mathbf{w}}(x) \in \mathbb{R}^{p_w}$. With $\overrightarrow{\mathbf{w}}(x)$, we approximate $g(Z_i, X_i) = \phi(X_i)$ by a linear combination of $W_i = \overrightarrow{\mathbf{w}}(X_i) \in \mathbb{R}^{p_w}$ for $1 \le i \le n$ and $\{\phi(X_i)\}_{1 \le i \le n}$ by the matrix $W = (W_1^\intercal, \ldots, W_n^\intercal)^\intercal$. For example, if $\phi(x)$ is linear, we set $\overrightarrow{\mathbf{w}}(x) = \{1, x_1, x_2, \cdots, x_{p_x}\}$. Furthermore, we consider the additive model $\phi(x) = \sum_{j=1}^{p_x} \phi_j(x_j)$ with smooth functions $\{\phi_j(\cdot)\}_{1 \le j \le p_x}$. For $1 \le j \le p_x$, we construct a set of basis functions $\overrightarrow{\mathbf{b}_j} = \{b_{j,l}(\cdot)\}_{1 \le l \le M_j}$ for $\phi_j(x_j)$ with $M_j \ge 1$ denoting the number of basis functions. Examples of the basis functions include the polynomial or B spline basis. We then approximate $\phi(x)$ with $\overrightarrow{\mathbf{w}}(x) = \{1, \overrightarrow{\mathbf{b}_1}(x_1), \cdots, \overrightarrow{\mathbf{b}_{p_x}}(x_{p_x})\}$.

When the IV is invalid with $h \not\equiv 0$, in addition to approximating $\phi(x)$ by $\overrightarrow{\mathbf{w}}(x)$, we further approximate the function $h(z, x)$ by the pre-specified basis functions $v_1(z, x)$, …, and $v_L(z, x)$ and then approximate $g(z, x)$ by

$$\mathcal{V} \coloneqq \{v_1(z, x), \cdots, v_L(z, x), \overrightarrow{\mathbf{w}}(x)\} \quad \text{with} \quad v_l : \mathbb{R}^{p_x + p_z} \to \mathbb{R} \quad \text{for} \quad 1 \le l \le L. \tag{5}$$

We now present different choices of $g(\cdot)$ in (3), or equivalently, the basis set $\mathcal{V}$ in (5). With them, the model (3) can accommodate a wide range of valid or invalid IV forms.

(1) **Linear violation.** We take $g(z, x) = \sum_{l=1}^{p_z} \pi_l z_l + \phi(x)$ and $\mathcal{V} = \{z_1, \cdots, z_{p_z}, \overrightarrow{\mathbf{w}}(x)\}$.

(2) **Polynomial violation (for a single IV).** We consider $g(z, x) = h(z) + \phi(x)$, where the IV and baseline covariates affect the outcome in an additive manner. We set $\mathcal{V} = \{v_1(z), \cdots, v_L(z), \overrightarrow{\mathbf{w}}(x)\}$ and may take $v_1(z), \cdots, v_L(z)$ as (piecewise) polynomials of various orders.

(3) **Interaction violation (for a single IV).** We consider $g(z,x) = \sum_{l=1}^{p_x} \alpha_l \cdot z \cdot x_l + \phi(x)$, allowing the IV to affect the outcome interactively with the base covariates. For such a violation form, we set $\mathcal{V} = \{z \cdot x_1, \cdots, z \cdot x_{p_x}, \overrightarrow{\mathbf{w}}(x)\}$.

In this paper, we mainly consider parametric basis sets for violation approximation, such as these listed above, while we discuss the feasibility of using nonparametric basis sets in our procedure in Section 4.5. We focus on the single IV setting for the polynomial and interaction violation to simplify the notations. It can be extended to the setting with multiple IVs; see Section D.3 in the supplement for more details.

Two important remarks are in order for the choice of g and $\mathcal{V}$. Firstly, the choice of g and the corresponding basis set $\mathcal{V}$ is part of the outcome model specification (3), reflecting the users' belief about how the IVs can potentially violate the assumptions (A2) and (A3). In applications, empirical researchers might apply domain knowledge and construct the pre-specified set of basis functions in (5). For example, Angrist and Lavy (1999) applied 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". Importantly, we do not have to assume the knowledge of $\mathcal{V}$ and will present in Section 4.4 a data-dependent way to choose the best $\mathcal{V}$ from a collection of candidate basis sets. Secondly, the linear violation form with $g(z,x) = \sum_{l=1}^{p_z} \pi_l z_l + \phi(x)$ is the most commonly used violation form in the context of the multiple IV framework (Small, 2007; Kang et al., 2016; Guo et al., 2018; Windmeijer et al., 2019). Most existing methods require at least a proportion of Z_i to be valid, i.e., the corresponding π_l being zero. In contrast, our framework allows all IVs to be invalid (i.e., all π_l to be nonzero) by taking $\mathcal{V} = \{z_1, \cdots, z_{p_z}, \overrightarrow{\mathbf{w}}(x)\}$.

With the set $\mathcal{V}$ of basis functions in (5), we define the matrix V which evaluates the functions in $\mathcal{V}$ at the observed variables:

$$V = \begin{pmatrix} V_1 & \cdots & V_n \end{pmatrix}^\intercal \in \mathbb{R}^{n \times (L+p_w)} \quad \text{with} \quad V_i = (v_1(Z_i, X_i), \cdots, v_L(Z_i, X_i), W_i^\intercal)^\intercal, \tag{6}$$

for $1 \le i \le n$. Note that we always use an intercept $W_{i,1} = 1$. Instead of assuming V_i perfectly generating $g(Z_i, X_i)$, we go with a broader framework by allowing an approximation error of $g(X_i, Z_i)$ by the best linear combination of V_i, defined as

$$R_i(V) \coloneqq g(Z_i, X_i) - V_i^\intercal \pi \quad \text{with} \quad \pi \coloneqq \arg\min_b \mathbf{E}\left[g(Z_i, X_i) - V_i^\intercal b\right]^2. \tag{7}$$

Define $R(V) = (R_1(V), \cdots, R_n(V)) \in \mathbb{R}^n$. We shall omit the dependence on V when there is no confusion. With (7), we express the outcome model (3) as

$$Y_i = D_i\beta + V_i^\intercal \pi + R_i(V) + \epsilon_i, \quad \text{for} \quad 1 \le i \le n. \tag{8}$$

If $g(Z_i, X_i)$ is well approximated by a linear combination of V_i, $\|R(V)\|_2/\sqrt{n}$ is close to zero or even $\|R(V)\|_2 = 0$.

2.3 Causal Interpretation: Structural Equation and Potential Outcome Models

We first provide the structural equation model (SEM) interpretation of the models (3) and (4). We consider the following SEM for Y_i and D_i and X_i,

$$Y_i \leftarrow D_i\beta + g_1(Z_i, X_i) + \nu_1(H_i) + \epsilon_i^0, \quad \text{and} \quad D_i \leftarrow f_1(Z_i, X_i) + \nu_2(H_i) + \delta_i^0, \tag{9}$$

where H_i denotes some unmeasured confounders, ϵ_i^0 and δ_i^0 are random errors independent of D_i, Z_i, X_i, H_i. Define $g_2(Z_i, X_i) = \mathbf{E}(\nu_1(H_i) \mid Z_i, X_i)$ and $f_2(Z_i, X_i) = \mathbf{E}(\nu_2(H_i) \mid Z_i, X_i)$. In (9), the unmeasured confounders H_i might affect both the outcome and treatment, g_1 encodes a direct effect of the IVs on the outcome, and g_2 captures the association between the IVs and the unmeasured confounders. The SEM (9) together with the definitions of f_2 and g_2 imply the outcome model (3) with $\epsilon_i = \epsilon_i^0 + \nu_1(H_i) - g_2(Z_i, X_i)$. The treatment model $D_i = f(Z_i, X_i) + \delta_i$ arises with $f(Z_i, X_i) = f_1(Z_i, X_i) + f_2(Z_i, X_i)$ and $\delta_i = \delta_i^0 + \nu_2(H_i) - f_2(Z_i, X_i)$.

We now discuss how to interpret the invalid IV model with the potential outcome perspective and explain why the model (3) allows the violation of both (A2) and (A3) assumptions. 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 $1 \le i \le n$, we consider the potential outcome model (Splawa-Neyman et al., 1990; Rubin, 1974),

$$Y_i^{(z,d)}(X_i) = Y_i^{(0,0)}(X_i) + d\beta + g_1(z, X_i) \quad \text{and} \quad \mathbf{E}(Y_i^{(0,0)}(X_i) \mid Z_i, X_i) = g_2(Z_i, X_i), \quad (10)$$

where $\beta \in \mathbb{R}$ denotes the treatment effect, $g_1 : \mathbb{R}^{p_z + p_x} \to \mathbb{R}$, and $g_2 : \mathbb{R}^{p_z + p_x} \to \mathbb{R}$. If $g_1(z, x)$ changes with z, the IVs directly affect the outcome, violating the assumption (A3). If $g_2(z, x)$ changes with z, the IVs are associated with unmeasured confounders, violating the assumption (A2). The model (10) extends the Additive LInear, Constant Effects (ALICE) model of Holland (1988) by allowing for a general class of invalid IVs. By the consistency assumption $Y_i = Y_i^{(Z_i, D_i)}(X_i)$, (10) implies (3) with $g(\cdot) = g_1(\cdot) + g_2(\cdot)$ and $\epsilon_i = Y_i^{(0,0)}(X_i) - g_2(Z_i, X_i)$.

We can easily generalize (10) 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 (3) 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 $\beta = \mathbf{E}\beta_i$.

2.4 Randomized Experiments with Non-Compliance and Invalid IVs

In this section, we take a slight detour to focus on the randomized experiments with noncompliance and consider an alternative identification strategy for the causal effect in the presence of invalid instrumental variables. We include this discussion here since it demonstrates in a slightly different regime that a nonlinear treatment model (with interactions) can enable causal identification even when instrumental variables are invalid.

In a randomized experiment, the ith subject is randomly assigned an assignment indicator Z_i, where $Z_i \in \{0, 1\}$ denotes assignment to control (0) or treatment (1). The subject may then decide whether to actually take the treatment, with $D_i \in \{0, 1\}$ indicating whether the subject ultimately does not take (0) or does take (1) the treatment. We denote by $X_i \in \{0, 1\}$ a binary covariate, and let $Y_i^{(z,d)}(x)$ denote the potential outcome for a subject with assignment z, treatment value d, and covariate value x.

A primary concern in the setting of randomized experiments with noncompliance is that the assignment indicator Z_i may be an invalid IV even if it is randomly assigned. To facilitate the discussion, we follow Imbens (2014) and adopt the example from McDonald

et al. (1992), which studies the effect of influenza vaccination on hospitalization for flu-related reasons. The researchers randomly sent letters to physicians encouraging them to vaccinate their patients. In this example, Z_i denotes the indicator that the physician receives the encouragement letter, while D_i stands for whether the patient is vaccinated. The outcome Y_i denotes whether the patient is hospitalized for a flu-related reason. As pointed out in Imbens (2014), a physician receiving the letter may "take actions that affect the likelihood of the patient getting infected with the flu other than simply administering the flu vaccine," implying that Z_i may have a direct effect on the outcome and thus violate the classical IV assumptions. In the following, we detail one possible identification strategy even when the assignment indicator Z_i has a direct effect on the outcome, by leveraging a nonlinear treatment model.

For a covariate x, we define the intention-to-treat (ITT) on the outcome as $\mathrm{ITT}_Y(x) = \mathbf{E}(Y_i \mid Z_i = 1, X_i = x) - \mathbf{E}(Y_i \mid Z_i = 0, X_i = x)$ and the ITT on the treatment as $\mathrm{ITT}_D(x) = \mathbf{E}(D_i \mid Z_i = 1, X_i = x) - \mathbf{E}(D_i \mid Z_i = 0, X_i = x)$. In the following proposition, we assume a constant effect and demonstrate the treatment effect identification even if the IV Z_i directly affects the outcome (violating assumption (A3)).

Proposition 1 *Suppose that* (I1) $\mathrm{ITT}_D(x = 1) - \mathrm{ITT}_D(x = 0) \neq 0$, (I2) Z_i *is randomly assigned,* (I3) $Y_i^{(z',d)}(x) - Y_i^{(z,d)}(x) = (z' - z)\pi$ *for* $x \in \{0, 1\}$, *and* (I4) $Y_i^{(z,d')}(x) - Y_i^{(z,d)}(x) = (d' - d)\beta$. *Then we identify* β *as* $\frac{\mathrm{ITT}_Y(x=1) - \mathrm{ITT}_Y(x=0)}{\mathrm{ITT}_D(x=1) - \mathrm{ITT}_D(x=0)}$.

Condition (I4) assumes a constant effect, while (I1)-(I3) are relaxations of the classical IV assumptions (A1)-(A3), respectively. In particular, the random assignment required by (I2) implies (A2) and (I2) is satisfied in randomized experiments with noncompliance.

Condition (I3) allows the IV to affect the outcome directly, but assumes that the IV's direct effect does not interact with the baseline covariate X_i. When there exists a direct effect, identification of β relies on Condition (I1). Condition (I1) essentially requires that the treatment assignment is determined interactively by Z_i and X_i: a special example is given by $D_i = \mathbf{1}(\gamma_0 + \gamma_z Z_i + \gamma_x X_i + \gamma Z_i \cdot X_i + \delta_i > 0)$ where δ_i denotes a random variable independent of Z_i and X_i. Such identification conditions have been proposed in Shardell and Ferrucci (2016); Have et al. (2007). In particular, Shardell and Ferrucci (2016) has pointed out that $\mathrm{ITT}_D(x = 1) - \mathrm{ITT}_D(x = 0)$ is referred to as the compliance score, and Condition (I1) requires that the compliance score changes with the covariate value x. We now provide some discussion on the plausibility of Condition (I1). We still use the vaccination example from McDonald et al. (1992) and imagine that we may have access to the covariate X_i, an indicator for whether the patients took regular flu shots in the past few years. Patients who regularly take flu shots (with $X_i = 1$) are more likely to follow the physician's encouragement than those who do not (with $X_i = 0$), suggesting $\mathrm{ITT}_D(x = 1) - \mathrm{ITT}_D(x = 0) \neq 0$.

3. TSCI: Overview of Identification

We propose in Section 4 a novel methodology called two-stage curvature identification (`TSCI`) to identify β under models (3) and (4). We first assume knowledge of the basis set $\mathcal{V}$ in (5) that spans the function $g(\cdot)$. The proposal consists of two stages: in the first stage, we employ ML algorithms to learn the conditional mean $f(\cdot)$ in the treatment model

(4); in the second stage, we adjust for the IV invalidity form encoded by $\mathcal{V}$ and construct the confidence interval for β by leveraging the first-stage ML fits.

The success of `TSCI` relies on the following identification condition.

Condition 1 *Define the best linear approximation of $f(Z_i, X_i)$ by V_i in population as $\gamma^* = \arg\min_\gamma \mathbf{E}(f(Z_i, X_i) - V_i^\intercal \gamma)^2$. We assume $\mathbf{E}(f(Z_i, X_i) - V_i^\intercal \gamma^*)^2 > 0$.*

The condition states that the basis set $\mathcal{V}$, used for generating $g(\cdot)$, does not fully span the conditional mean function $f(\cdot)$ of the treatment model. For any user-specified $\mathcal{V}$ in (5), Condition 1 can be partially examined by evaluating a generalized IV strength introduced in Section 4.3. When the generalized IV strength is sufficiently large, our proposed `TSCI` methodology leads to an estimator $\widehat{\beta}(\mathcal{V})$ in (18) such that

$$(\widehat{\beta}(\mathcal{V}) - \beta)/\widehat{\mathrm{SE}}(\mathcal{V}) \xrightarrow{d} N(0, 1),$$

with the estimated standard error $\widehat{\mathrm{SE}}(\mathcal{V})$ depending on the generalized IV strength. Importantly, the basis set $\mathcal{V}$ used for generating $g(\cdot)$ does not need to be fixed in advance: we discuss in Section 4.4 a data-driven strategy for choosing a "good" basis set $\mathcal{V}_{\widehat{q}}$ among a collection of candidate sets $\mathcal{V}_0 \subset \mathcal{V}_1 \subset \cdots \subset \mathcal{V}_Q$ for a positive integer Q, where $\mathcal{V}_0$ corresponds to the valid IV setting. Our proposed `TSCI` compares estimators assuming valid IVs and adjusting for violation forms generated from the family $\{\mathcal{V}_q\}_{1 \le q \le Q}$. Thus, we encompass the setting of valid IVs and through this comparison, the `TSCI` methodology provides more robust causal inference tools than directly assuming IVs' validity.

A central question in using our proposal is how to choose the IV and covariates so that Condition 1 is plausible. Intuitively, Condition 1 requires $f(\cdot)$ to be more nonlinear than $g(\cdot)$. We recommend that users adopt IVs identified by domain experts. Such IVs are more likely to satisfy Condition 1 and therefore provide a practical choice for applying the method. First, suppose these IVs are valid and satisfy the IV assumptions (A2) and (A3). Then the violation function $g(Z_i, X_i)$ does not depend on Z_i, which ensures that the function class used to generate $g(Z_i, X_i)$ is less complex than that of $f(Z_i, X_i)$, which does depend on Z_i. In such cases, Condition 1 is automatically satisfied. Second, suppose the IVs identified by domain experts are invalid. Even if these IVs violate assumptions (A2) and (A3), they tend to have a relatively weak direct effect on the outcome or are only weakly associated with the unmeasured confounders, because they are identified based on domain knowledge. Such weak violations of (A2) and (A3) make Condition 1 more plausible, because $g(Z_i, X_i)$ remains relatively simple compared with $f(Z_i, X_i)$. In this case, `TSCI` provides robust causal inference in the presence of invalid IVs, rather than requiring the IVs to be perfectly valid.

4. TSCI with Machine Learning

We discuss the first and second stages of our `TSCI` methodology in Sections 4.1 and 4.2, respectively. The main novelty of our proposal is to estimate a bias caused by the first-stage ML and implement a follow-up bias correction.

4.1 First Stage: Machine Learning Models for the Treatment Model

We estimate the conditional mean $f(\cdot)$ in the treatment model (4) by a general ML fit. We randomly split the data into two disjoint subsets $\mathcal{A}_1$ and $\mathcal{A}_2$. Throughout the paper we use $\mathcal{A}_1 = \{1, 2, \cdots, n_1\}$ with $n_1 = \lfloor 2n/3 \rfloor$ and set $\mathcal{A}_2 = \{n_1 + 1, \cdots, n\}$. Our results can be extended to any other splitting with $|\mathcal{A}_1| \asymp |\mathcal{A}_2|$. Sample splitting is essential for removing the endogeneity of the ML predicted values. Without sample splitting, the ML predicted value for the treatment can be close to the treatment (due to overfitting) and hence highly correlated with unmeasured confounders, leading to a `TSCI` estimator suffering from a significant bias.

The main step is to construct the ML prediction model with data belonging to $\mathcal{A}_2$ and express the ML estimator of $f_{\mathcal{A}_1} = (f(Z_1, X_1), \cdots, f(Z_{n_1}, X_{n_1}))^\intercal$ as

$$\widehat{f}_{\mathcal{A}_1} = \Omega D_{\mathcal{A}_1} \quad \text{for some matrix} \quad \Omega \in \mathbb{R}^{n_1 \times n_1}, \tag{11}$$

where the data belonging to $\mathcal{A}_2$ is used to train the ML algorithm and construct the transformation matrix Ω. Our proposed `TSCI` method mainly relies on expressing the first-stage prediction as the linear estimator in (11). A wide range of first-stage ML algorithms are shown to have the expression in (11). In the following, we follow the results in Lin and Jeon (2006); Meinshausen (2006); Wager and Athey (2018) and express the sample split random forests in the form (11). In Sections A.5, A.6, and A.7 in the supplement, we present the explicit definitions of Ω for boosting, deep neural network (DNN), and approximation with B-splines, respectively.

Importantly, the transformation matrix $\Omega \in \mathbb{R}^{n_1 \times n_1}$ has a related meaning as the hat matrix in linear regression procedures: however, Ω is not a projection matrix for nonlinear ML algorithms, which creates additional challenges in adopting the first-stage ML fit; see Section A.1.

Split random forests in (11)**.** To avoid the overfitting of random forests, we adopt the construction of honest trees and forests as in Wager and Athey (2018) and slightly modify it by excluding self-prediction for our purpose. We construct the random forests (RF) with the data $\{D_i, X_i, Z_i\}_{i \in \mathcal{A}_2}$ and estimate $f(Z_i, X_i)$ for $i \in \mathcal{A}_1$ by the constructed RF together with the data $\{X_i, Z_i, D_i\}_{i \in \mathcal{A}_1}$. RF aggregates $S \geq 1$ decision trees, with each decision tree being viewed as the partition of the whole covariate space $\mathbb{R}^{p_x + p_z}$ into disjoint subspaces $\{\mathcal{R}_l\}_{1 \leq l \leq J}$. Let θ denote the random parameter that determines how a tree is grown. For any given $(z^\intercal, x^\intercal)^\intercal \in \mathbb{R}^{p_x + p_z}$ 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 the subspace $\mathcal{R}_{l(z,x,\theta)}$ contains $(z^\intercal, x^\intercal)^\intercal$. With the observations inside $\mathcal{R}_{l(z,x,\theta)}$, the decision tree estimates $f(z, x) = \mathbf{E}(D \mid Z = z, X = x)$ by

$$\widehat{f}_\theta(z, x) = \sum_{j \in \mathcal{A}_1} \omega_j(z, x, \theta) D_j \quad \text{with} \quad \omega_j(z, x, \theta) = \frac{\mathbf{1}\left[(Z_j^\intercal, X_j^\intercal)^\intercal \in \mathcal{R}^0_{l(z,x,\theta)}\right]}{\sum_{k \in \mathcal{A}_1} \mathbf{1}\left[(Z_k^\intercal, X_k^\intercal)^\intercal \in \mathcal{R}^0_{l(z,x,\theta)}\right]}, \tag{12}$$

where $\mathcal{R}^0_{l(z,x,\theta)} \coloneqq \mathcal{R}_{l(z,x,\theta)}/(z, x)$ is defined as the region excluding the point (z, x). We use $\{\theta_1, \cdots, \theta_S\}$ to denote the parameters corresponding to the S trees. The estimator

$\widehat{f}(z,x) = \frac{1}{S}\sum_{s=1}^{S} \widehat{f}_{\theta_s}(z,x)$ can be expressed as

$$\widehat{f}(z,x) = \sum_{j\in\mathcal{A}_1} \omega_j(z,x) D_j \quad \text{where} \quad w_j(z,x) = \frac{1}{S}\sum_{s=1}^{S} \omega_j(z,x,\theta_s), \tag{13}$$

with $\omega_j(z,x,\theta_s)$ defined in (12). That is, the split RF estimator of $f_{\mathcal{A}_1}$ attains the form (11) with $\Omega_{ij} = \omega_j(Z_i, X_i)$ for $i,j \in \mathcal{A}_1$.

The construction of honest trees and removal of self-prediction help remove the endogeneity of the ML predicted values $\{\widehat{f}(Z_i,X_i)\}_{i\in\mathcal{A}_1}$. Firstly, due to the sample-splitting, the construction of Ω does not directly depend on $\{D_i\}_{i\in\mathcal{A}_1}$, which removes the endogeneity contained in the ML predicted values. Secondly, consider an extreme setting with Z_i and X_i not being predictive for D_i. In such a scenario, without removing the self-prediction, the estimator $\widehat{f}(Z_i,X_i)$ will be dominated by its own treatment observation D_i and $\widehat{f}(Z_i,X_i)$ is still highly correlated with the unmeasured confounders, leading to a biased `TSCI` estimator. The removal of self-prediction can be viewed as an ML version of the "leave-one-out" estimator for the treatment model, which was proposed in Angrist et al. (1999) to reduce the bias of `TSLS` with many IVs. In the numerical studies, we have observed that the exclusion of self-prediction leads to a more accurate estimator than the corresponding `TSCI` estimator with self-prediction, regardless of the IV strength; see Table 7 for the detailed comparison.

4.2 Second Stage: Adjusting for Violation Forms and Correcting Bias

In the second stage, we leverage the first-stage ML fits in the form of (11) and adjust for the possible IV invalidity forms. In the remaining of the construction, we consider the outcome model in the form of (8) and assume that $V_i^\intercal \pi$ provides an accurate approximation of $g(Z_i,X_i)$, resulting in zero or sufficiently small $R_i(V)$. Hence $R_i(V)$ does not enter in the following construction of the estimator for β. We quantify the effect of the error term $R_i(V)$ in the theoretical analysis in Section 5 and propose in Section 4.4 a data-dependent way of choosing V such that $R_i(V)$ is sufficiently small.

Applying the transformation Ω in (11) to the model (8) with data in $\mathcal{A}_1$, we obtain

$$\widehat{Y}_{\mathcal{A}_1} = \widehat{f}_{\mathcal{A}_1}\beta + \widehat{V}_{\mathcal{A}_1}\pi + \widehat{R}_{\mathcal{A}_1} + \widehat{\epsilon}_{\mathcal{A}_1}, \tag{14}$$

where $\widehat{Y}_{\mathcal{A}_1} = \Omega Y_{\mathcal{A}_1}, \widehat{f}_{\mathcal{A}_1} = \Omega D_{\mathcal{A}_1}, \widehat{V}_{\mathcal{A}_1} = \Omega V_{\mathcal{A}_1}, \widehat{R}_{\mathcal{A}_1} = \Omega R_{\mathcal{A}_1}$, and $\widehat{\epsilon}_{\mathcal{A}_1} = \Omega \epsilon_{\mathcal{A}_1}$. For a matrix V, we use P_V and $P_V^\perp$ to denote the projection to the column spaces of V and its orthogonal complement, respectively. Based on (14), we project out the columns of $\widehat{V}_{\mathcal{A}_1}$ and estimate the effect β by

$$\widehat{\beta}_{\text{init}}(V) := \frac{\widehat{Y}_{\mathcal{A}_1}^\intercal P^\perp_{\widehat{V}_{\mathcal{A}_1}} \widehat{f}_{\mathcal{A}_1}}{\widehat{f}_{\mathcal{A}_1}^\intercal P^\perp_{\widehat{V}_{\mathcal{A}_1}} \widehat{f}_{\mathcal{A}_1}} = \frac{Y_{\mathcal{A}_1}^\intercal \mathbf{M}(V) D_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\intercal \mathbf{M}(V) D_{\mathcal{A}_1}} \quad \text{with} \quad \mathbf{M}(V) = \Omega^\intercal P^\perp_{\widehat{V}_{\mathcal{A}_1}} \Omega, \tag{15}$$

When $R_i(V) = 0$, then we decompose the error of the above estimator as

$$\widehat{\beta}_{\text{init}}(V) - \beta = \frac{\epsilon_{\mathcal{A}_1}^\intercal \mathbf{M}(V) \delta_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\intercal \mathbf{M}(V) D_{\mathcal{A}_1}} + \frac{\epsilon_{\mathcal{A}_1}^\intercal \mathbf{M}(V) f_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\intercal \mathbf{M}(V) D_{\mathcal{A}_1}}. \tag{16}$$

In the above decomposition, the first term on the right-hand side is a bias component appearing due to the use of the first-stage ML. To explain this, we consider the homoscedastic correlation $\mathrm{Cov}(\epsilon_i, \delta_i \mid Z_i, X_i) = \mathrm{Cov}(\epsilon_i, \delta_i)$ and obtain the explicit bias expression,

$$\frac{\epsilon_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)\delta_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)D_{\mathcal{A}_1}} \approx \frac{\mathrm{Cov}(\delta_i, \epsilon_i)\cdot \mathrm{Tr}[\mathbf{M}(V)]}{D_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)D_{\mathcal{A}_1}}, \tag{17}$$

where $\mathrm{Tr}[\mathbf{M}(V)]$ denotes the trace of $\mathbf{M}(V)$ defined in (15). Note that $\mathrm{Tr}[\mathbf{M}(V)]$ can be viewed as degrees of freedom or complexity measure for the ML algorithm. It may become particularly large in the overfitting regime. The bias in (17) also becomes large when the denominator $D_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)D_{\mathcal{A}_1}$ is small, which means a weak generalized IV strength in the following (21). Thus, both the numerator and denominator in (17) can lead to a large bias of $\widehat{\beta}_{\mathrm{init}}(V)$. In Figure 1, we illustrate the bias of $\widehat{\beta}_{\mathrm{init}}(V)$ in settings with different values of the IV strength scaled to $D_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)D_{\mathcal{A}_1}$.

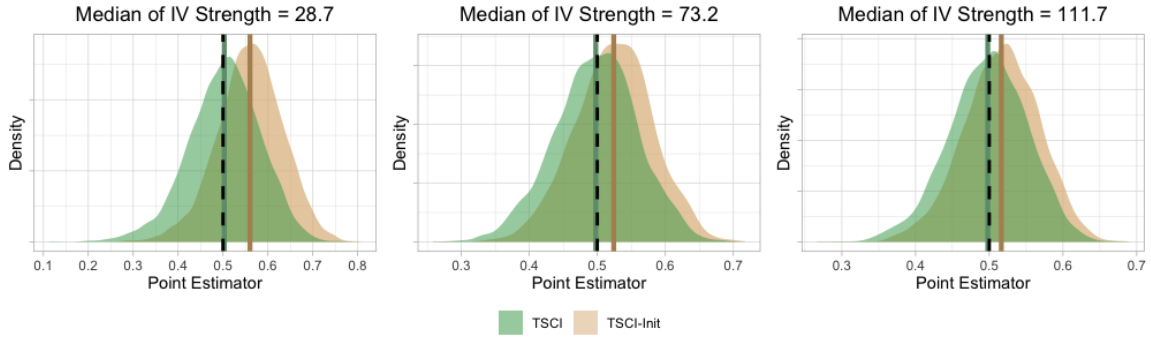

Figure 1: Density plot of `TSCI` and `TSCI-Init` estimators (with 500 simulations), which are after and before bias correction, respectively. The three panels from left to right correspond to settings with increasing IV strengths; see Section D.1 in the supplement for details. The black dashed line represents the true value $\beta = 0.5$. The green and brown solid lines indicate the means of `TSCI` and `TSCI-Init` estimators.

To address this finite-sample bias of $\widehat{\beta}_{\mathrm{init}}(V)$, we propose the following bias-corrected estimator,

$$\widehat{\beta}(V) = \frac{Y_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)D_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)D_{\mathcal{A}_1}} - \frac{\sum_{i=1}^{n_1}[\mathbf{M}(V)]_{ii}\widehat{\delta}_i[\widehat{\epsilon}(V)]_i}{D_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)D_{\mathcal{A}_1}}, \tag{18}$$

with $\widehat{\delta}_i = D_i - \widehat{f}_i$ for $1 \le i \le n_1$ and $\widehat{\epsilon}(V) = P_V^{\perp}[Y - D\widehat{\beta}_{\mathrm{init}}(V)]$. We show in Figure 1 that our proposal effectively corrects the bias of $\widehat{\beta}_{\mathrm{init}}(V)$. Importantly, our proposed bias correction is effective for both homoscedastic and heteroscedastic correlations. We present a simplified bias correction assuming homoscedastic correlation in Section A.9 in the supplement.

Centering at $\widehat{\beta}_{\mathrm{RF}}(V)$ defined in (18), we construct the confidence interval

$$\mathrm{CI}(V) = \left(\widehat{\beta}(V) - z_{\alpha/2}\widehat{\mathrm{SE}}(V), \widehat{\beta}(V) + z_{\alpha/2}\widehat{\mathrm{SE}}(V)\right), \tag{19}$$

with $z_{\alpha/2}$ denoting the upper $\alpha/2$ quantile of the standard normal distribution and

$$\widehat{\mathrm{SE}}(V) = \frac{\sqrt{\sum_{i=1}^{n_1}[\widehat{\epsilon}(V)]_i^2[\mathbf{M}(V)D_{\mathcal{A}_1}]_i^2}}{D_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)D_{\mathcal{A}_1}} \quad \text{with} \quad \widehat{\epsilon}(V) = P_V^{\perp}\left[Y - D\widehat{\beta}_{\mathrm{init}}(V)\right], \tag{20}$$

Remark 1 The bias component in (17) results from the correlation between ϵ_i and δ_i for $i \in \mathcal{A}_1$. In the regime of many IVs, Angrist et al. (1999) proposed the "leave-one-out" jackknife-type IV estimator, that is, instead of fitting the first stage model with all the data, the treatment model for the i-th observation is fitted without using itself. Such an estimator effectively removes the bias for `TSLS` due to the correlation between ϵ_i and δ_i. Our proposed removal of self-prediction is of a similar spirit to the jackknife. However, even after removing the self-prediction, the diagonal of $\mathbf{M}(V)$ is not zero, and hence the correlation between ϵ_i and δ_i still leads to a bias component, which requires the bias correction as in (18).

4.3 Generalized IV Strength: Detection of Under-Fitted Machine Learning

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 (Stock et al., 2002; Hansen et al., 2008). With a larger basis set $\mathcal{V}$, the IV strength will generally decrease as the information contained in $\mathcal{V}$ is projected out from the first-stage ML fit. It is crucial to introduce a generalized IV strength measure in consideration of invalid IV forms and the ML algorithm. Similarly to the classical setup, good performance of our proposed `TSCI` estimator also requires a relatively large generalized IV strength. We introduce a generalized IV strength measure as,

$$\mu(V) \coloneqq f_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)f_{\mathcal{A}_1} / \left[\sum_{i\in\mathcal{A}_1} \mathrm{Var}(\delta_i \mid X_i, Z_i)/n_1\right]. \tag{21}$$

If $\mathrm{Var}(\delta_i \mid X_i, Z_i) = \sigma_\delta^2$, then $\mu(V)$ is reduced to $f_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)f_{\mathcal{A}_1}/\sigma_\delta^2$. A sufficiently large strength $\mu(V)$ will guarantee stable point and interval estimators defined in (18) and (19). Hence, we need to check whether $\mu(V)$ is sufficiently large. Since f is unknown, we estimate $f_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)f_{\mathcal{A}_1}$ by its sample version $D_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)D_{\mathcal{A}_1}$ and estimate $\mu(V)$ by $\widehat{\mu(V)} \coloneqq D_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)D_{\mathcal{A}_1} / \left[\|D_{\mathcal{A}_1} - \widehat{f}_{\mathcal{A}_1}\|_2^2/n_1\right]$.

In Section 5, we show that our point estimator in (18) is consistent when the IV strength $\mu(V)$ is much larger than $\mathrm{Tr}\left[\mathbf{M}(V)\right]$; see Condition (R2). We now develop a bootstrap test to provide an empirical assessment of this IV strength requirement. Since Rothenberg (1984) and Stock et al. (2002) suggested the concentration parameter being larger than 10 as being "adequate", we develop a bootstrap test for $\mu(V) \geq \max\{2\mathrm{Tr}\left[\mathbf{M}(V)\right], 10\}$. We apply the wild bootstrap method and construct a probabilistic upper bound for the estimation error $D_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)D_{\mathcal{A}_1} - f_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)f_{\mathcal{A}_1} = 2f_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)\delta_{\mathcal{A}_1} + \delta_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)\delta_{\mathcal{A}_1}$. 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 $1 \leq i \leq n_1$, with $\{U_i^{[l]}\}_{1\leq i\leq n_1}$ generated as i.i.d. standard normal random variables, and compute $S^{[l]} = \left[2f_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)\delta^{[l]} + (\delta^{[l]})^{\intercal}\mathbf{M}(V)\delta^{[l]}\right] / \left[\|D_{\mathcal{A}_1} - \widehat{f}_{\mathcal{A}_1}\|_2^2/n_1\right]$. We

use $\mathcal{S}_{\alpha_0}(V)$ to denote the upper α_0 empirical quantile of $\{|S^{[l]}|\}_{1\le l\le L}$. We conduct the generalized IV strength test $\widehat{\mu(V)} \ge \max\{2\mathrm{Tr}\left[\mathbf{M}(V)\right], 10\} + \mathcal{S}_{\alpha_0}(V)$, with $\mathcal{S}_{\alpha_0}(V)$ being a high probability upper bound for $|\widehat{\mu(V)} - \mu(V)|$. We 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 matrix V defined in (6); otherwise, the IV is claimed to be weak after adjusting for the matrix V. Empirically, we observe reliable inference properties when the estimated generalized IV strength $\widehat{\mu(V)}$ is above 40; see Figure 6 for details.

4.4 Data-Dependent Selection of $\mathcal{V}$

Our proposed `TSCI` estimator in (18) requires prior knowledge of the basis set $\mathcal{V}$, which generates the function $g(\cdot)$. In the following, we consider the nested sets of basis functions $\mathcal{V}_0 \subset \mathcal{V}_1 \subset \cdots \subset \mathcal{V}_Q$, where Q is a positive integer. We devise a data-dependent way to choose the best one among $\{\mathcal{V}_q\}_{0\le q\le Q}$.

We define $\mathcal{V}_0 \coloneqq \{\overrightarrow{\mathbf{w}}(x)\}$ as the set of basis functions for the valid IV setting. For $q \ge 1$, define $\mathcal{V}_q \coloneqq \{v_1(\cdot), \cdots, v_{L_q}(\cdot), \overrightarrow{\mathbf{w}}(x)\}$ as the basis set for different invalid IV forms, where $L_q \ge 1$ is the number of basis functions. We present two examples of $\{\mathcal{V}_q\}_{1\le q\le Q}$ for the single IV setting.

(1) Polynomial violation: $\mathcal{V}_q = \{z, z^2, \cdots, z^q, \overrightarrow{\mathbf{w}}(x)\}$, for $1 \le q \le Q$.

(2) Interaction violation: $\mathcal{V}_1 = \left\{z, z\cdot x_1, z\cdot x_2, \cdots, z\cdot x_{p_x}, \overrightarrow{\mathbf{w}}(x)\right\}$.

For $0 \le q \le Q$, we define the matrix $V_q \in \mathbb{R}^{n\times(L_q+p_w)}$ with its i-th row defined as $(V_q)_i = \left(v_1(Z_i, X_i), \cdots, v_{L_q}(Z_i, X_i), W_i^\intercal\right)^\intercal$ with $W_i = \overrightarrow{\mathbf{w}}(X_i)$ for $1 \le i \le n$.

For estimating $q \in \{0, 1, \ldots\}$ from data, we proceed as follows. We implement the generalized IV strength test in Section 4.3 and define $Q_{\max}$ as,

$$Q_{\max} \coloneqq \max_{q\ge 0}\left\{q : \widehat{\mu(V_q)} \ge \max\{2\mathrm{Tr}\left[\mathbf{M}(V_q)\right], 10\} + \mathcal{S}_{\alpha_0}(V_q)\right\}, \tag{22}$$

where α_0 is set at 0.025 by default. For a larger q, we tend to adjust out more information and have relatively weaker IVs. Intuitively, $Q_{\max}$ denotes the largest index such that the IVs still have enough strength after adjusting for V_q. With $Q_{\max}$, we shall choose among $\{V_q\}_{0\le q\le Q_{\max}}$. As a remark, when there is no $q \ge 0$ satisfying (22), this corresponds to the weak IV regime. In such a scenario, we shall implement the valid IV estimator and output a warning of weak IV.

For any given $0 \le q \le Q_{\max}$, we apply the generalized estimator in (18) and construct

$$\widehat{\beta}(V_q) = \frac{Y_{\mathcal{A}_1}^\intercal \mathbf{M}(V_q) D_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\intercal \mathbf{M}(V_q) D_{\mathcal{A}_1}} - \frac{\sum_{i=1}^{n_1}[\mathbf{M}(V_q)]_{ii}\widehat{\delta}_i[\widehat{\epsilon}(V_{Q_{\max}})]_i}{D_{\mathcal{A}_1}^\intercal \mathbf{M}(V_q) D_{\mathcal{A}_1}}, \tag{23}$$

with $\widehat{\delta}_{\mathcal{A}_1} = D_{\mathcal{A}_1} - \widehat{f}_{\mathcal{A}_1}$, $\mathbf{M}(\cdot)$ defined in (15), and $\widehat{\epsilon}(\cdot)$ defined in (20).

The selection of the best V_q among $\{V_q\}_{0\le q\le Q_{\max}}$ relies on comparing the estimators $\{\widehat{\beta}(V_q)\}_{0\le q\le Q_{\max}}$. We start with comparing the difference between the estimators $\widehat{\beta}_{V_q}$ and $\widehat{\beta}_{V_{q'}}$ with $0 \le q < q' \le Q_{\max}$. When $\{g(X_i, Z_i)\}_{1\le i\le n}$ are well approximated by both V_q and $V_{q'}$, the approximation errors $R(V_q)$ and $R(V_{q'})$ defined in (7) are small and the main

difference $\widehat{\beta}(V_{q'}) - \widehat{\beta}(V_q)$ is approximately due to the randomness of the errors $\epsilon_{\mathcal{A}_1}$. In such cases. we then have $\widehat{\beta}(V_{q'}) - \widehat{\beta}(V_q)$ is approximated centered at zero with the following conditional variance,

$$\begin{aligned}\widehat{H}(V_q, V_{q'}) = & \frac{\sum_{i=1}^{n_1}[\widehat{\epsilon}(V_{Q_{\max}})]_i^2[\mathbf{M}(V_{q'})D_{\mathcal{A}_1}]_i^2}{[D_{\mathcal{A}_1}^\intercal \mathbf{M}(V_{q'})D_{\mathcal{A}_1}]^2} + \frac{\sum_{i=1}^{n_1}[\widehat{\epsilon}(V_{Q_{\max}})]_i^2[\mathbf{M}(V_q)D_{\mathcal{A}_1}]_i^2}{[D_{\mathcal{A}_1}^\intercal \mathbf{M}(V_q)D_{\mathcal{A}_1}]^2} \\ & - 2\frac{\sum_{i=1}^{n_1}[\widehat{\epsilon}(V_{Q_{\max}})]_i^2[\mathbf{M}(V_{q'})D_{\mathcal{A}_1}]_i[\mathbf{M}(V_q)D_{\mathcal{A}_1}]_i}{[D_{\mathcal{A}_1}^\intercal \mathbf{M}(V_{q'})D_{\mathcal{A}_1}] \cdot [D_{\mathcal{A}_1}^\intercal \mathbf{M}(V_q)D_{\mathcal{A}_1}]}.\end{aligned} \tag{24}$$

Based on the above approximation, we further conduct the following test for whether $\widehat{\beta}(V_q)$ is significantly different from $\widehat{\beta}(V_{q'})$,

$$\mathcal{C}(V_q, V_{q'}) = \mathbf{1}\left(\left|\widehat{\beta}(V_q) - \widehat{\beta}(V_{q'})\right| / \sqrt{\widehat{H}(V_q, V_{q'})} \geq z_{\alpha_0}\right), \tag{25}$$

where $\widehat{\beta}(V_q)$ and $\widehat{\beta}(V_{q'})$ are defined in (23). On the other hand, if the smaller matrix V_q does not provide a good approximation of $\{g(Z_i, X_i)\}_{1\leq i\leq n}$, the difference $\left|\widehat{\beta}(V_q) - \widehat{\beta}(V_{q'})\right|$ tends to be much larger than the threshold in (25) and hence $\mathcal{C}(V_q, V_{q'}) = 1$, indicating that V_q does not fully generate $\{g(Z_i, X_i)\}_{1\leq i\leq n}$.

For $Q_{\max} \geq 2$, we generalize the pairwise comparison to multiple comparisons. For $0 \leq q \leq Q_{\max} - 1$, we compare $\widehat{\beta}(V_q)$ to any $\widehat{\beta}(V_{q'})$ with $q + 1 \leq q' \leq Q_{\max}$. Particularly, for $0 \leq q \leq Q_{\max} - 1$, we define the test

$$\mathcal{C}(V_q) = \mathbf{1}\left(\max_{q+1\leq q'\leq Q_{\max}}\left[\left|\widehat{\beta}(V_q) - \widehat{\beta}(V_{q'})\right| / \sqrt{\widehat{H}(V_q, V_{q'})}\right] \geq \widehat{\rho}\right), \tag{26}$$

where the thresholding $\widehat{\rho}$ is chosen by the wild bootstrap; see more details in Section A.2 in the supplement. We define $\mathcal{C}(V_{Q_{\max}}) = 0$ as there is no index larger than $Q_{\max}$ that we might compare to. We interpret $\mathcal{C}(V_q) = 0$ as follows: none of the differences $\{|\widehat{\beta}(V_q) - \widehat{\beta}(V_{q'})|\}_{q+1\leq q'\leq Q_{\max}}$ is large, indicating that V_q (approximately) generates the function $g(\cdot)$ and $\widehat{\beta}(V_q)$ is a reliable estimator.

We choose the index $\widehat{q}_c \in [0, Q_{\max}]$ as $\widehat{q}_c = \min_{0\leq q\leq Q_{\max}}\{q : \mathcal{C}(V_q) = 0\}$, which is the smallest q such that $\mathcal{C}(V_q)$ is zero. The index $\widehat{q}_c$ is interpreted as follows: any matrix containing $V_{\widehat{q}_c}$ as a submatrix does not lead to a substantially different estimator from that by $V_{\widehat{q}_c}$. This provides evidence for $V_{\widehat{q}_c}$ forming a good approximation of $\{g(Z_i, X_i)\}_{1\leq i\leq n}$. Here, the sub-index "c" represents the comparison as we compare different $\{V_q\}_{0\leq q\leq Q_{\max}}$ to choose the best one. Note that $\widehat{q}_c$ must exist since $\mathcal{C}(V_{Q_{\max}}) = 0$ by definition. In finite samples, there are chances that certain violations cannot be detected, especially if the `TSCI` estimators given by $V_{\widehat{q}_c}$ and $V_{\widehat{q}_c+1}$ are not significantly different. We propose a more robust choice of the index as $\widehat{q}_r = \min\{\widehat{q}_c + 1, Q_{\max}\}$, where the sub index "r" denotes the robust selection; see the discussion after Theorem 3. We summarize our proposed `TSCI` estimator in Algorithm 1.

As a remark, $\mathcal{C}(V_q, V_{q'})$ in (25) can be used to test the IV validity. For any $q' \geq 1$, $\mathcal{C}(V_0, V_{q'}) = 1$ represents that the estimator assuming valid IV is sufficiently different from an estimator allowing for the violation form generated by $V_{q'}$, indicating that the valid IV assumption is violated.

Algorithm 1 `TSCI` with machine learning

Input: Data $X \in \mathbb{R}^{n \times p_x}, Z, D, Y \in \mathbb{R}^n$; Sets of basis $\{\mathcal{V}_q\}_{q \geq 0}$ for approximating $g(\cdot)$;
Output: $\widehat{q}_c$ and $\widehat{q}_r$; $\widehat{\beta}(V_{\widehat{q}_c})$ and $\widehat{\beta}(V_{\widehat{q}_r})$; $\mathrm{CI}(V_{\widehat{q}_c})$ and $\mathrm{CI}(V_{\widehat{q}_r})$.

Generate matrices V_q based on $\mathcal{V}_q$ for $q \geq 0$ as in (6);
Compute $Q_{\max}$ as in (22) and $\widehat{\epsilon}(V_{Q_{\max}})$ as in (20);
Compute $\{\widehat{\beta}(V_q)\}_{0 \leq q \leq Q_{\max}}$ as in (23) and $\{\widehat{H}(V_q, V_{q'})\}_{0 \leq q < q' \leq Q_{\max}}$ as in (24);
if $Q_{\max} \geq 1$ **then**
    Compute $\{\mathcal{C}(V_q)\}_{0 \leq q \leq Q_{\max}-1}$ as in (26);
end if
Set $\mathcal{C}(V_{Q_{\max}}) = 0$;
Compute $\widehat{q}_c = \min\{0 \leq q \leq Q_{\max} : \mathcal{C}(V_q) = 0\}$; ▷ Comparison selection
Compute $\widehat{q}_r = \min\{\widehat{q}_c + 1, Q_{\max}\}$; ▷ Robust selection
Compute $\widehat{\beta}_{\mathrm{RF}}(V_{\widehat{q}_c})$ and $\widehat{\beta}_{\mathrm{RF}}(V_{\widehat{q}_r})$ as in (18) with $V = V_{\widehat{q}_c}, V_{\widehat{q}_r}$, respectively;
Compute $\mathrm{CI}_{\mathrm{RF}}(V_{\widehat{q}_c})$ and $\mathrm{CI}_{\mathrm{RF}}(V_{\widehat{q}_r})$ as in (19) with $V = V_{\widehat{q}_c}, V_{\widehat{q}_r}$, respectively.

4.5 Nonparametric Basis for Violation Approximation

We can also use nonparametric bases to approximate the violation in `TSCI`, such as splines or Chebyshev polynomial bases, if there is sufficient strength of IVs; see the last paragraph of this subsection. Recall that the violation basis matrix V is constructed with a set of basis functions $\{v_1(z,x), \ldots, v_L(z,x)\}$ as in (6). For example, when using B-splines for violation approximation, we specify $p_x + 1$ functions $v_j(z,x)$ for $1 \leq j \leq p_x + 1$ as

$$v_1(z,x) = \overrightarrow{\mathbf{b}}_1(x_1), \quad \cdots, \quad v_{p_x}(z,x) = \overrightarrow{\mathbf{b}}_{p_x}(x_{p_x}), \quad v_{p_x+1}(z,x) = \overrightarrow{\mathbf{b}}_{p_x+1}(z),$$

where $\overrightarrow{\mathbf{b}}_j = \{b_{j,l}(\cdot)\}_{1 \leq l \leq M_j}$ with $M_j \geq 1$ denoting the number of B-spline basis functions, as defined in Section 2.2. If other basis functions (such as Chebyshev polynomial basis) are preferred, we can proceed analogously.

Such extensions, however, require sufficiently strong IVs. The reason is that we essentially project out all IV transformations generated by the nonparametric basis, which may significantly weaken the remaining IV-treatment association. In practice, the remaining IV-treatment association after this projection may be too weak and fail to pass the generalized IV strength test introduced in Section 4.3. When this occurs, the inference may not be reliable after the violation adjustment. However, if the remaining IV-treatment association passes the generalized IV strength test, then our proposal can be naturally extended to allow for nonparametric bases. In Section 6.4, we investigate our proposal's performance when integrated with different choices of basis functions for the violation functions.

4.6 Comparison to Double Machine Learning and Machine Learning IV

We compare our proposed `TSCI` with the double machine learning (`DML`) estimator (Chernozhukov et al., 2018) and the machine learning IV (`MLIV`) estimator (Chen et al., 2023; Liu et al., 2020). As the most significant difference, `TSCI` provides a valid inference that is robust to a certain class of invalid IVs while `DML` and `MLIV` are designed for valid IV regimes

only. We focus on the valid IV setting in the following and compare our proposal to DML and MLIV.

When comparing our proposal to Double Machine Learning with instrumental variables, we use the standard version (Chernozhukov et al., 2018). We note that efficiency improvements may be possible when the treatment exhibits a nonlinear association with the instruments (Vansteelandt and Didelez, 2018), but such a version has not been rigorously justified in the context of double machine learning. Chernozhukov et al. (2018) considered the outcome model $Y_i = D_i\beta + g(X_i) + \epsilon_i$, which is a special case of our outcome model (3) by assuming valid IVs. The population parameter β in DML can be identified through the following expression (Chernozhukov et al., 2018; Emmenegger and Bühlmann, 2021)

$$\beta = \frac{\mathbf{E}\left[Y_i - \mathbf{E}(Y_i \mid X_i)\right]\left[Z_i - \mathbf{E}(Z_i \mid X_i)\right]}{\mathbf{E}\left[D_i - \mathbf{E}(D_i \mid X_i)\right]\left[Z_i - \mathbf{E}(Z_i \mid X_i)\right]}, \tag{27}$$

where $\mathbf{E}(Y_i \mid X_i)$, $\mathbf{E}(Z_i \mid X_i)$, and $\mathbf{E}(D_i \mid X_i)$ are fitted by machine learning algorithms.

As a fundamental difference, we apply machine learning algorithms to capture the nonlinear relation between D_i and Z_i, X_i while (27) identifies β based on the linear association $\mathbf{E}\left[D_i - \mathbf{E}(D_i \mid X_i)\right]\left[Z_i - \mathbf{E}(Z_i \mid X_i)\right]$. Due to the first-stage ML, our proposed TSCI estimator is generally more efficient than the DML estimator. On the other hand, we approximate $g(z, x)$ by a set of basis functions, which might provide an inaccurate estimator when the basis functions are misspecified. In contrast, DML learns the conditional mean model by general machine learning algorithms and does not particularly require such a specification of basis functions. We compare the performance of DML and TSCI with valid IVs in Section 6.1 and with invalid IVs in Section 6.3.

Chen et al. (2023); Liu et al. (2020) proposed to use the ML prediction values $\widehat{f}_{\mathcal{A}_1}$ in (11) as the IV, referred to as the MLIV in the current paper. With this MLIV, the standard TSLS estimator can be implemented: in the first stage, use OLS regression of $D_{\mathcal{A}_1}$ on the MLIV $\widehat{f}_{\mathcal{A}_1}$ and the baseline covariates $V_{\mathcal{A}_1}$ and construct the predicted value $\widehat{D}_{\mathcal{A}_1} = c_f\widehat{f}_{\mathcal{A}_1} + V_{\mathcal{A}_1}c_v$ with $c_f \in \mathbb{R}$ and the vector c_v denoting the OLS regression coefficients; in the second stage, use the outcome regression by replacing $D_{\mathcal{A}_1}$ with $\widehat{D}_{\mathcal{A}_1}$. The TSLS estimator with MLIV can then be expressed as

$$\widehat{\beta}_{\texttt{MLIV}} := Y_{\mathcal{A}_1}^{\intercal} P_{V_{\mathcal{A}_1}}^{\perp} \widehat{D}_{\mathcal{A}_1} / \widehat{D}_{\mathcal{A}_1}^{\intercal} P_{V_{\mathcal{A}_1}}^{\perp} \widehat{D}_{\mathcal{A}_1} = \beta + \epsilon_{\mathcal{A}_1}^{\intercal} P_{V_{\mathcal{A}_1}}^{\perp} \widehat{f}_{\mathcal{A}_1} / [c_f \cdot \widehat{f}_{\mathcal{A}_1}^{\intercal} P_{V_{\mathcal{A}_1}}^{\perp} \widehat{f}_{\mathcal{A}_1}]. \tag{28}$$

The last equality of (28) holds by plugging in the outcome model $Y_i = D_i\beta + V_i^{\intercal}\pi + \epsilon_i$ and $\widehat{D}_{\mathcal{A}_1} = c_f\widehat{f}_{\mathcal{A}_1} + V_{\mathcal{A}_1}c_v$ and applying the orthogonality between $D_{\mathcal{A}_1} - \widehat{D}_{\mathcal{A}_1}$ and $\{\widehat{f}_{\mathcal{A}_1}, V_{\mathcal{A}_1}\}$. Note that a dominating term of $\widehat{\beta}_{\text{init}}(V) - \beta$ in (16) is $\widehat{\epsilon}_{\mathcal{A}_1}^{\intercal} P_{\widehat{V}_{\mathcal{A}_1}}^{\perp} \widehat{f}_{\mathcal{A}_1} / \widehat{f}_{\mathcal{A}_1}^{\intercal} P_{\widehat{V}_{\mathcal{A}_1}}^{\perp} \widehat{f}_{\mathcal{A}_1}$. In comparison, the last term in (28) is roughly inflated by a factor of $1/c_f$, appearing due to the additional first-stage regression using the MLIV. When the IV is relatively strong, then this factor c_f is around 1, and MLIV and $\widehat{\beta}_{\text{init}}(V)$ are of similar performance for valid IV settings. However, in the presence of relatively weak IVs, the coefficient c_f in front of $\widehat{f}_{\mathcal{A}_1}$ has a large fluctuation due to the weak association between $D_{\mathcal{A}_1}$ and $\widehat{f}_{\mathcal{A}_1}$. The weak IV problem deteriorates with the extra first stage regression using the MLIV. This explains why the estimator $\widehat{\beta}_{\texttt{MLIV}}$ in (28) has a much larger bias and standard error than our proposed TSCI estimator in (18) when the IVs are relatively weak; see Section D.2 in the supplementary for the detailed comparison.

5. Theoretical Justification

We first establish the asymptotic normality of $\widehat{\beta}(V)$ for a given V. To begin with, we present the required conditions. The first assumption is imposed on the regression errors in models (3) and (4) and the data $\{V_i, f_i\}_{1\le i\le n}$ 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, there exists a positive constant $K > 0$ such that

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

with $\sup_{Z_i,X_i}$ denoting the supremum over the support of the density of Z_i, X_i. The random variables $\{V_i, f_i\}_{1\le i\le n_1}$ satisfy $\lambda_{\min}\left(\sum_{i=1}^{n_1} V_iV_i^\intercal/n_1\right) \ge c$, $\|\sum_{i=1}^{n_1} V_if_i/n_1\|_2 \le C$, $\max_{1\le i\le n_1}\{|f_i|, \|V_i\|_2\} \le C\sqrt{\log n_1}$, and $\|\sum_{i=1}^{n_1} V_i[R(V)]_i/n_1\|_2 \le C\|R(V)\|_\infty$, where $C > 0$ and $c > 0$ are constants independent of n and p. The matrix Ω defined in (11) satisfies $\lambda_{\max}(\Omega) \le C$ 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 assuming sub-gaussian errors ϵ_i and δ_i. The sub-gaussian conditions on regression errors may be relaxed to moment conditions. The conditions on V_i and f_i will be automatically satisfied with high probability if $\mathbf{E}V_iV_i^\intercal$ is positive definite and V_i and f_i are sub-gaussian random variables, where the sub-gaussianality conditions on $\{V_i, f_i\}_{1\le i\le n}$ may be relaxed to moment conditions. The proof of Lemma 1 in the supplement shows that $\lambda_{\max}(\Omega) \le 1$ for random forests and deep neural networks.

The second assumption is imposed on the generalized IV strength $\mu(V)$ defined in (21). Throughout the paper, the asymptotics is taken as $n \to \infty$.

(R2) $f_{\mathcal{A}_1}^\intercal \mathbf{M}(V) f_{\mathcal{A}_1}$ satisfies $f_{\mathcal{A}_1}^\intercal \mathbf{M}(V) f_{\mathcal{A}_1} \to \infty$ and $f_{\mathcal{A}_1}^\intercal \mathbf{M}(V) f_{\mathcal{A}_1} \gg \mathrm{Tr}[\mathbf{M}(V)]$, with $\mathbf{M}(V)$ defined in (15).

The above condition is closely related to Condition 1, where $f_{\mathcal{A}_1}^\intercal \mathbf{M}(V) f_{\mathcal{A}_1}$ measures the variability of the estimated $\widehat{f}$ after adjusting for V used to approximate $\{g(Z_i, X_i)\}_{1\le i\le n}$. Intuitively, a larger value of $f_{\mathcal{A}_1}^\intercal \mathbf{M}(V) f_{\mathcal{A}_1}$ indicates that $\mathbf{E}(f(Z_i, X_i) - V_i^\intercal\gamma^*)^2 > 0$ in Condition 1 holds more plausibly. The generalized IV strength $\mu(V)$ introduced in (21) is proportional to $f_{\mathcal{A}_1}^\intercal \mathbf{M}(V) f_{\mathcal{A}_1}$ when $\mathrm{Var}(\delta_i \mid X_i, Z_i)$ is of a constant scale. Our proposed test in Section 4.3 is designed to test whether $\mu(V)$ is sufficiently large for `TSCI`. For the setting with $\mathbf{E}(f(Z_i, X_i) - V_i^\intercal\gamma^*)^2$ in Condition 1 being a positive constant, $f_{\mathcal{A}_1}^\intercal \mathbf{M}(V) f_{\mathcal{A}_1}$ can be of the order n, making Condition (R2) a mild assumption.

The following proposition establishes the consistency of $\widehat{\beta}_{\text{init}}(V)$ if Conditions (R1) and (R2) hold and the approximation errors $\{R_i(V) = g(Z_i, X_i) - V_i^\intercal\pi\}_{1\le i\le n}$ are small.

Proposition 2 *Consider the models* (3) *and* (4). *If Conditions* (R1) *and* (R2) *hold and* $\|R_{\mathcal{A}_1}(V)\|_2^2 \ll f_{\mathcal{A}_1}^\intercal \mathbf{M}(V) f_{\mathcal{A}_1}$, *then* $\widehat{\beta}_{\text{init}}(V)$ *defined in* (15) *satisfies* $\widehat{\beta}_{\text{init}}(V) \xrightarrow{p} \beta$.

5.1 Improvement with Bias Correction

In the following, we present the distributional properties of the bias-corrected `TSCI` estimator $\widehat{\beta}(V)$ defined in (18) and demonstrate the advantage of this extra bias correction step. For establishing the asymptotic normality, we further impose a stronger generalized IV strength condition than (R2).

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

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

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

$$\eta_n(V) = \|f_{\mathcal{A}_1} - \widehat{f}_{\mathcal{A}_1}\|_\infty + \left(|\beta - \widehat{\beta}_{\text{init}}(V)| + \|R(V)\|_\infty + \frac{\log n}{\sqrt{n}}\right)(\sqrt{\log n} + \|f_{\mathcal{A}_1} - \widehat{f}_{\mathcal{A}_1}\|_\infty). \quad (29)$$

The above condition can be viewed as requiring the IV to be sufficiently strong, where $f_{\mathcal{A}_1}^\intercal[\mathbf{M}(V)]^2 f_{\mathcal{A}_1}$ can be viewed as another measure of IV strength. If the treatment model is fitted with neural network, we have $f_{\mathcal{A}_1}^\intercal[\mathbf{M}(V)]^2 f_{\mathcal{A}_1} = f_{\mathcal{A}_1}^\intercal \mathbf{M}(V) f_{\mathcal{A}_1}$. For random forests, we only have $f_{\mathcal{A}_1}^\intercal[\mathbf{M}(V)]^2 f_{\mathcal{A}_1} \le f_{\mathcal{A}_1}^\intercal \mathbf{M}(V) f_{\mathcal{A}_1}$. $\eta_n(V)$ defined in (29) depends on the accuracy of the ML prediction model $\widehat{f}$. We have $\eta_n(V) \to 0$ if $\|f_{\mathcal{A}_1} - \widehat{f}_{\mathcal{A}_1}\|_\infty \to 0$, $\widehat{\beta}_{\text{init}}(V) \xrightarrow{p} \beta$, and $\|R(V)\|_\infty \to 0$. However, even for inconsistent $\widehat{f}$, $\eta_n(V)$ is of a smaller order of magnitude than $\sqrt{\log n}$ as long as $\|f_{\mathcal{A}_1} - \widehat{f}_{\mathcal{A}_1}\|_\infty$ is of a constant scale.

The following theorem establishes the asymptotic normality of the `TSCI` estimator.

Theorem 1 *Consider the models* (3) *and* (4). *Suppose that Condition* (R1), (R2-I) *hold and*

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

Then $\widehat{\beta}(V)$ *defined in* (18) *satisfies*

$$\frac{1}{\mathrm{SE}(V)}\left(\widehat{\beta}(V) - \beta\right) \xrightarrow{d} N(0,1), \quad \textit{with} \quad \mathrm{SE}(V) = \frac{\sqrt{\sum_{i=1}^{n_1} \sigma_i^2 [\mathbf{M}(V) f_{\mathcal{A}_1}]_i^2}}{f_{\mathcal{A}_1}^\intercal \mathbf{M}(V) f_{\mathcal{A}_1}}.$$

If $\widehat{\mathrm{SE}}(V)$ *used in* (19) *satisfies* $\widehat{\mathrm{SE}}(V)/\mathrm{SE}(V) \xrightarrow{p} 1$, *then the confidence interval* $\mathrm{CI}(V)$ *in* (19) *satisfies* $\liminf_{n\to\infty} \mathbb{P}(\beta \in \mathrm{CI}(V)) = 1 - \alpha$.

We emphasize that the validity of our proposed confidence interval does not require the ML prediction model $\widehat{f}$ to be a consistent estimator of f. Particularly, Condition (R2) and (R2-I) can be plausibly satisfied as long as the ML algorithms capture enough association between the treatment and the IVs. The 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 $\mathrm{SE}(V)$ relies on the generalized IV strength for a given matrix V. If $f_{\mathcal{A}_1}^\intercal \mathbf{M}(V) f_{\mathcal{A}_1}/n_1$ is of constant order, then $\mathrm{SE}(V) \lesssim 1/\sqrt{n}$. A larger matrix V will generally lead to a larger $\mathrm{SE}(V)$ because $f_{\mathcal{A}_1}^\intercal \mathbf{M}(V) f_{\mathcal{A}_1}$ decreases after adjusting for more information contained in V. The consistency of $\widehat{\mathrm{SE}}(V)$ is presented in Lemma 2 in the supplement.

We now explain the effectiveness of the bias correction for the homoscedastic setting with $\mathrm{Cov}(\epsilon_i, \delta_i \mid Z_i, X_i) = \mathrm{Cov}(\epsilon_i, \delta_i)$. For the initial estimator $\widehat{\beta}_{\text{init}}(V)$, we can establish that $\frac{1}{\mathrm{SE}(V)}\left(\widehat{\beta}_{\text{init}}(V) - \beta\right) = \mathcal{G}(V) + \widetilde{\mathcal{E}}(V)$, where $\mathcal{G}(V) \xrightarrow{d} N(0,1)$ and

$$\left|\widetilde{\mathcal{E}}(V)\right| \leq \frac{\mathrm{Cov}(\epsilon_i, \delta_i) \cdot \mathrm{Tr}[\mathbf{M}(V)] + \|R(V)\|_2}{\sqrt{f_{\mathcal{A}_1}[\mathbf{M}(V)]^2 f_{\mathcal{A}_1}}} + \frac{\sqrt{\mathrm{Tr}([\mathbf{M}(V)]^2)}}{(f_{\mathcal{A}_1}[\mathbf{M}(V)]^2 f_{\mathcal{A}_1})^{c_0}}, \tag{31}$$

for some positive constant $c_0 > 0$. Our proposed bias-corrected estimator $\widehat{\beta}(V)$ is effective in reducing the bias component $\widetilde{\mathcal{E}}(V)$. Particularly, Theorem 4 in Section A.9 in the supplement establishes that the term $\mathrm{Cov}(\epsilon_i, \delta_i) \cdot \mathrm{Tr}[\mathbf{M}(V)]/\sqrt{f_{\mathcal{A}_1}[\mathbf{M}(V)]^2 f_{\mathcal{A}_1}}$ in (31) is reduced to $\eta_n(V) \cdot \mathrm{Tr}[\mathbf{M}(V)]/\sqrt{f_{\mathcal{A}_1}[\mathbf{M}(V)]^2 f_{\mathcal{A}_1}}$. If $\eta_n(V) \to 0$, which can be achieved for a consistent ML prediction $\widehat{f}$, the bias-corrected `TSCI` estimator effectively reduces the finite-sample or higher-order bias. However, even if $\widehat{f}_{\mathcal{A}_1}$ is inconsistent with $\|f_{\mathcal{A}_1} - \widehat{f}_{\mathcal{A}_1}\|_2 \lesssim \sqrt{n}$, the bias correction will not lead to a worse estimator.

5.2 Guarantee for Algorithm 1

We now justify the validity of the confidence interval from the `TSCI` Algorithm 1, which is based on the asymptotic normality of $\widehat{\beta}(V_{\widehat{q}})$ with a data-dependent index $\widehat{q}$. The property of $\widehat{q}$ relies on a careful analysis of the difference $\widehat{\beta}(V_q) - \widehat{\beta}(V_{q'})$ for $0 \leq q < q' \leq Q_{\max}$.

We start with the setting where both V_q and $V_{q'}$ provide good approximations to $\{g(Z_i, X_i)\}_{1 \leq i \leq n}$, leading to sufficiently small $\|R(V_q)\|_2$ and $\|R(V_{q'})\|_2$. In this case, the dominating component of $\widehat{\beta}(V_q) - \widehat{\beta}(V_{q'})$ is $S^\intercal \epsilon_{\mathcal{A}_1}$ with

$$S = \left(\frac{1}{f_{\mathcal{A}_1}^\intercal \mathbf{M}(V_{q'}) f_{\mathcal{A}_1}} \mathbf{M}(V_{q'}) - \frac{1}{f_{\mathcal{A}_1}^\intercal \mathbf{M}(V_q) f_{\mathcal{A}_1}} \mathbf{M}(V_q)\right) f_{\mathcal{A}_1} \in \mathbb{R}^{n_1}. \tag{32}$$

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

$$H(V_q, V_{q'}) = \frac{\sum_{i=1}^{n_1} \sigma_i^2 [\mathbf{M}(V_{q'}) f_{\mathcal{A}_1}]_i^2}{[f_{\mathcal{A}_1}^\intercal \mathbf{M}(V_{q'}) f_{\mathcal{A}_1}]^2} + \frac{\sum_{i=1}^{n_1} \sigma_i^2 [\mathbf{M}(V_q) f_{\mathcal{A}_1}]_i^2}{[f_{\mathcal{A}_1}^\intercal \mathbf{M}(V_q) f_{\mathcal{A}_1}]^2} - 2\frac{\sum_{i=1}^{n_1} \sigma_i^2 [\mathbf{M}(V_{q'}) f_{\mathcal{A}_1}]_i [\mathbf{M}(V_q) f_{\mathcal{A}_1}]_i}{[f_{\mathcal{A}_1}^\intercal \mathbf{M}(V_{q'}) f_{\mathcal{A}_1}] \cdot [f_{\mathcal{A}_1}^\intercal \mathbf{M}(V_q) f_{\mathcal{A}_1}]}, \tag{33}$$

with $\sigma_i^2 = \mathbf{E}(\epsilon_i^2 \mid Z_i, X_i)$. The following condition requires that the variance of $S^\intercal f_{\mathcal{A}_1}$ dominates other components of $\widehat{\beta}(V_q) - \widehat{\beta}(V_{q'})$, which are mainly finite-sample approximation errors.

(R3) The variance $H(V_q, V_{q'})$ in (33) satisfies

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

with $\eta_n(\cdot)$ defined in (29). There exists $c > 0$ such that $\mathrm{Var}(\delta_i \mid Z_i, X_i) \geq c$.

When the first stage is fitted with the basis method or neural network and $\mathrm{Var}(\epsilon_i \mid Z_i, X_i) = \sigma_\epsilon^2, \mathrm{Var}(\delta_i \mid Z_i, X_i) = \sigma_\delta^2$, we have $H(V_q, V_{q'}) = \sigma_\epsilon^2 \left(1/f^\intercal \mathbf{M}(V_{q'}) f - 1/f^\intercal \mathbf{M}(V_q) f\right)$ and $\mu(V_q) = f^\intercal \mathbf{M}(V_q) f / \sigma_\delta^2$ for $q \leq q'$. If we assume that $f^\intercal \mathbf{M}(V_{q'}) f = c_* f^\intercal \mathbf{M}(V_q) f$ for some

$0 < c_* < 1$, 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 $\mu(V_q) \gg (\mathrm{Tr}[\mathbf{M}(V_q)])^2$ and $\mu(V_{q'}) \gg (\mathrm{Tr}[\mathbf{M}(V_{q'})])^2$ up to some polynomial order of $\log n$. In this case, Condition (R3) is slightly stronger than Condition (R2-I).

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.

Theorem 2 *Consider the model* (3) *and* (4). *Suppose that* (R1) *and* (R3) *hold,* (R2) *holds for* $V \in \{V_q, V_{q'}\}$, *and* S *defined in* (32) *satisfies* $\max_{i\in\mathcal{A}_1} S_i^2/(\sum_{i\in\mathcal{A}_1} S_i^2) \to 0$. *If* $\sqrt{H(V_q, V_{q'})} \gg \max_{V\in\{V_q,V_{q'}\}} \|R(V)\|_2/\sqrt{\mu(V)}$, *then* $\left(\widehat{\beta}(V_q) - \widehat{\beta}(V_{q'})\right)/\sqrt{H(V_q, V_{q'})} \xrightarrow{d} N(0,1)$, *with* $\widehat{\beta}(\cdot)$ *and* $H(V_q, V_{q'})$ *defined in* (23) *and* (33), *respectively.*

For small approximation errors $\|R(V_q)\|_2$ and $\|R(V_{q'})\|_2$, the condition $\sqrt{H(V_q, V_{q'})} \gg \max_{V\in\{V_q,V_{q'}\}} \|R(V)\|_2/\sqrt{\mu(V)}$ holds. In this case, Theorem 2 establishes that the difference $\widehat{\beta}(V_q) - \widehat{\beta}(V_{q'})$ is centered at zero and has an asymptotic normal distribution.

We now move on to the case that at least one of V_q and $V_{q'}$ does not approximate $\{g(Z_i, X_i)\}_{1\le i\le n}$ well. In this case, Theorem 2 does not hold and we establish Theorem 5 in the supplement that $(\widehat{\beta}(V_q) - \widehat{\beta}(V_{q'}))/\sqrt{H(V_q, V_{q'})}$ is centered at

$$\mathcal{L}_n(V_q, V_{q'}) = \frac{1}{\sqrt{H(V_q, V_{q'})}}\left(\frac{D_{\mathcal{A}_1}^\intercal \mathbf{M}(V_q)[R(V_q)]_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\intercal \mathbf{M}(V_q) D_{\mathcal{A}_1}} - \frac{D_{\mathcal{A}_1}^\intercal \mathbf{M}(V_{q'})[R(V_{q'})]_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\intercal \mathbf{M}(V_{q'}) D_{\mathcal{A}_1}}\right). \tag{34}$$

To simultaneously quantify the errors of $\{\widehat{\beta}(V_q) - \widehat{\beta}(V_{q'})\}_{0\le q<q'\le Q_{\max}}$, we define the α_0 quantile for the maximum of multiple random error components in (32),

$$\mathbb{P}\left(\max_{0\le q<q'\le Q_{\max}} \frac{1}{\sqrt{H(V_q, V_{q'})}}\left|\frac{f_{\mathcal{A}_1}^\intercal \mathbf{M}(V_{q'})\epsilon_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^\intercal \mathbf{M}(V_{q'}) f_{\mathcal{A}_1}} - \frac{f_{\mathcal{A}_1}^\intercal \mathbf{M}(V_q)\epsilon_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^\intercal \mathbf{M}(V_q) f_{\mathcal{A}_1}}\right| \ge \rho(\alpha_0)\right) = \alpha_0. \tag{35}$$

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

(R4) For $\mathcal{V}_0 \subset \mathcal{V}_1 \subset \cdots \subset \mathcal{V}_Q$ and corresponding matrices $\{V_q\}_{0\le q\le Q}$, there exists $q^* \in \{0, 1, 2, \cdots, Q\}$ such that $q^* \le Q_{\max}$ and $R(V_{q^*}) = 0$ with $Q_{\max}$ defined in (22). For any integer $q \in [0, q^* - 1]$, there exists an integer $q' \in [q+1, q^*]$ such that

$$\mathcal{L}_n(V_q, V_{q'}) \ge A\rho(\alpha_0) \quad \text{with} \quad A > 2, \tag{36}$$

where $\mathcal{L}_n(V_q, V_{q'})$ is defined in (34) and $\rho(\alpha_0)$ is defined in (35).

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^*}. The well-separation condition (36) is interpreted as follows: if g is not well approximated by the column space of V_q, then the estimation bias of $\widehat{\beta}(V_q)$ is larger than the uncertainty $\rho(\alpha_0)$ due to the multiple random error components. Note that $\mathcal{L}_n(V_q, V_{q^*})$ defined in (34) is a measure of the bias of $\widehat{\beta}(V_q)$.

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 3 *Consider the model* (3) *and* (4). *Suppose that Conditions* (R1) *and* (R4) *hold, Condition* (R2-I) *holds for* $V \in \{V_q\}_{0\le q\le Q_{\max}}$, *Condition* (R3) *holds for any* $0 \le q < q' \le Q_{\max}$, $\widehat{H}(V_q, V_{q'})/H(V_q, V_{q'}) \xrightarrow{p} 1$, *and* $\widehat{\rho}/\rho(\alpha_0) \xrightarrow{p} 1$ *with* $\widehat{\rho}$ *used in* (26) *and* $\rho(\alpha_0)$ *defined in* (35) *respectively. Our proposed CI in Algorithm 1 satisfies* $\liminf_{n\to\infty} \mathbb{P}\left[\beta \in \mathrm{CI}(V_{\widehat{q}})\right] \ge 1-\alpha-2\alpha_0$ *with* $\widehat{q} = \widehat{q}_c$ *or* $\widehat{q}_r$ *where* α_0 *is used in* (35).

The condition (36) of (R4) is critical to guarantee the selection consistency $\widehat{q}_c = q^*$, ensuring that $\mathrm{CI}(V_{\widehat{q}_c})$ achieves the desired coverage. However, in finite samples, we may make mistakes in conducting the selection among $\{V_q\}_{0\le q\le Q_{\max}}$. The selection method $\widehat{q}_r$ is more robust in the sense that statistical inference based on $V_{\widehat{q}_r}$ is still valid even if we cannot separate V_{q^*-1} and V_{q^*}. To achieve the uniform inference without requiring the well-separation condition (36), we may simply apply `TSCI` with $\mathcal{V}_{Q_{\max}}$. However, such a confidence interval might be conservative by adjusting for a large $V_{Q_{\max}}$.

6. Simulation Studies

In Section 6.1, we consider the valid IV settings and compare our proposal to the `DML` estimator proposed in Chernozhukov et al. (2018). In Sections 6.2 and 6.3, we demonstrate our proposal for general invalid IV settings and also compare it with the `DML` estimator. In Section D.3 in the supplement, we further consider multiple IVs settings and compare `TSCI` with existing methods based on the majority rule. In Section 6.4, we demonstrate the sensitivity of our proposal to different choices of basis functions, including polynomial basis, B-spline basis and Chebyshev polynomial basis. We compute all measures using 500 simulations throughout the numerical studies. The code for replicating the numerical results is available at `https://github.com/zijguo/TSCI-Replication`.

6.1 Comparison with DML in Valid IV Settings

We focus on the valid IV settings and illustrate the advantages and limitations of both `TSCI` and `DML`. Because the violation function $g(z, x)$ does not depend on the IV z in valid IV settings, we simply use a covariates' basis $\mathcal{V}_0$ to approximation $g(\cdot)$ and skip the violation selection step introduced in Section 4.4. We adopt the data generative setting used in the R package `dmlalg` (Emmenegger, 2021) and implement `DML` by the R package `DoubleML` (Bach et al., 2024). We set the sample size as $n = 3000$ and generate the baseline covariate $X_i \in \mathbb{R}$ following the uniform distribution on $[-\pi, \pi]$. Conditioning on X_i, we generate the IV Z_i and the hidden confounder H_i following $Z_i \mid X_i \sim N(3 \cdot \tanh(2X_i - 1), 1)$ and $H_i \mid X_i \sim N(2 \cdot \sin(X_i), 1)$, respectively. We generate the outcome Y_i and treatment D_i using the SEM in (9). Particularly, we generate the outcome Y_i following $Y_i = D_i + X_i^2/2 - 3 \cdot \cos(\pi H_i/4) + e_{i,2}$ and consider the following two treatment models,

- Setting S1: $D_i = -a \cdot |Z_i| - 2 \cdot \tanh(X_i) - H_i + e_{i,1}$,
- Setting S2: $D_i = a \cdot Z_i^2/2 - 2 \cdot \tanh(X_i) - H_i + e_{i,1}$,

where a controls the nonlinearity of the treatment model and strength of the IV and $e_{i,1} \sim N(0,1)$ and $e_{i,2} \sim N(0,1)$ are independent noises. Note that a larger value of a in the treatment model leads to a more nonlinear dependence between D_i and Z_i. The treatment

model and the outcome model are consistent to (4) and (3) with δ_i and ϵ_i being composed of a hidden confounder term and an independent random noise.

In addition to the above two settings, we consider another setting where $g(X_i)$ in the outcome model is more complicated. We set $p_x = 5$ and for $1 \le i \le n$, we generate $\{X_{i,j}\}_{1\le j\le p_x}$ which are independently distributed, and each of them follows a uniform distribution on $[-1, 1]$. We generate Z_i and H_i following $Z_i \mid X_i \sim N(3 \cdot \tanh(2X_{i,1} - 1), 1)$ and $H_i \mid X_i \sim N(2 \cdot \sin(X_{i,1}), 1)$. We generate $\{D_i, Y_i\}_{1\le i\le n}$ following

- Setting S3: $D_i = b \cdot Z_i + \sin(2\pi Z_i) + 3/2 \cdot \cos(2\pi Z_i) - 1/5 \cdot \sum_{j=1}^{p_x} X_{i,j} - 1/2 \cdot H_i + e_{i,1}$ and $Y_i = D_i/2 + g(X_i) - \cos(\pi H_i/4) + e_{i,2}$ with $g(X_i) = \sum_{1\le j\ne k\le p_x} X_{i,j}X_{i,k}$,

where the value of b controls the linear association strength in Setting S3.

We implement Algorithm 1 with random forests by specifying a collection of basis functions for $g(x)$. Since the IV is valid here, we use the set of basis functions $\mathcal{V}_0 = \{1, \overrightarrow{\mathbf{b}}_1(x_1), \cdots, \overrightarrow{\mathbf{b}}_{p_x}(x_{p_x})\}$ with $\overrightarrow{\mathbf{b}}_j(x_j)$ denoting the B-spline basis with 5 knots of x_j for $1 \le j \le p_x$. As illustrated in Algorithm 1, we use $\mathcal{V}_0$ to obtain the `TSCI` estimator in (18) and construct CIs as in (19).

In Figure 2, we compare `DML` and `TSCI` under Settings S1, S2, and S3. In the top two panels, `TSCI` achieves the desired coverage in Setting S2 while it is slightly undercovered in Setting S1. This is caused by the fact that there still exists remaining bias even after the bias correction step. This problem is alleviated by increasing the IV strength in Setting S2. Even though `DML` achieves desired coverage in setting S1, the RMSE and CI length of `DML` is uniformly larger than those of our proposed `TSCI`. As discussed in Section 4.6, this happens since `DML` only uses the linear association in the treatment model while `TSCI` can increase the IV strength by applying ML algorithms to fit nonlinear treatment models. We have designed setting S2 as a less favorable setting for `DML`, where the linear association between the treatment and IV is nearly zero but the nonlinear association is strong. Consequently, the CI length and RMSE of the `DML` procedure is much larger than our proposed `TSCI`. In Setting S3, `TSCI` does not show the advantage over `DML` and has an under-coverage problem. This is caused by the misspecification of $g(x)$ in the outcome model. The approximation error of $g(x)$ contributes to the additional variance which is not captured by the estimated standard error defined in (20). `DML` is not exposed to such a problem because the ML algorithm fits $E(Y_i \mid X_i)$ well with machine learning algorithms. We also implement `TSLS` in Settings S1, S2 and S3, but we do not include it in the figures because it has low coverage due to misspecification of nonlinear $g(x)$.

6.2 `TSCI` with Invalid IVs

We now demonstrate our `TSCI` method under the general setting with possibly invalid IVs. In the following, we focus on the continuous treatment and will investigate the performance of `TSCI` for binary treatment in Section D.5 in the supplement. We generate $X_i^* \in \mathbb{R}^{p_x+1}$ following a multivariate normal distribution with zero mean and covariance matrix Σ where $\Sigma_{ij} = 0.5^{|i-j|}$ for $1 \le i, j \le p_x + 1$. With Φ denoting the standard normal cumulative distribution function, we define $X_{i,j} = \Phi(X_{ij}^*)$ for $1 \le j \le p_x$. We generate a continuous IV as $Z_i = 4(\Phi(X_{i,p_x+1}^*) - 0.5) \in (-2, 2)$.

We consider the following two conditional mean models for the treatment,

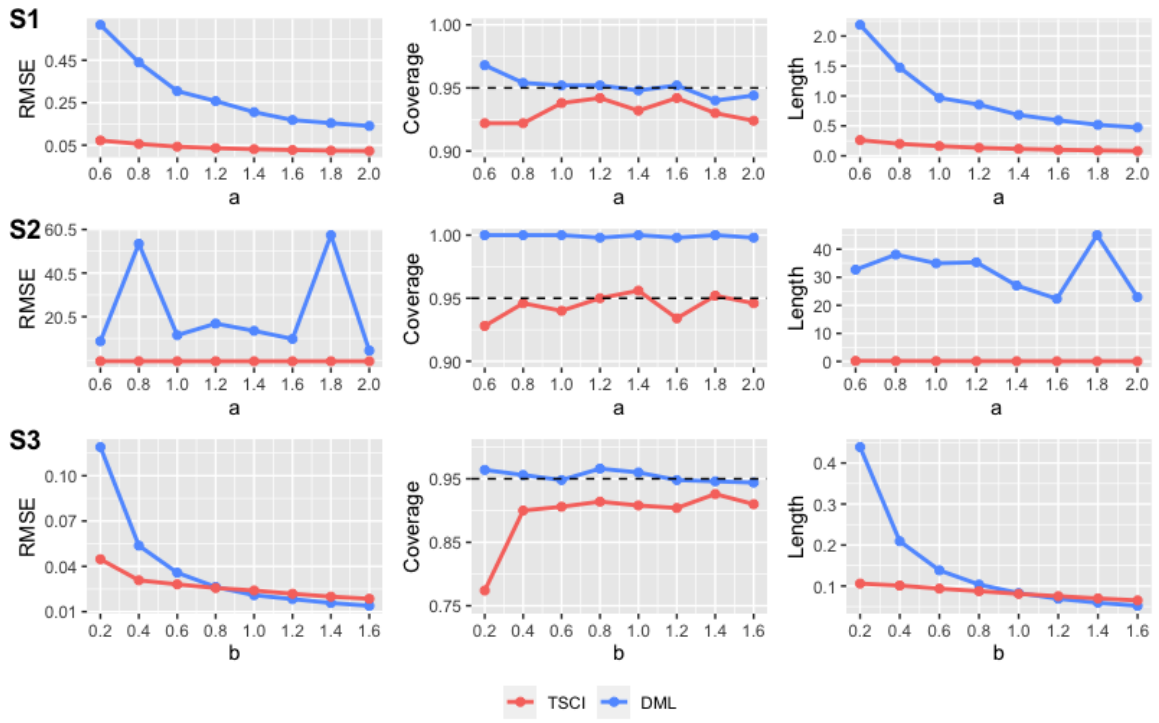

Figure 2: Comparison of `DML` and `TSCI` in terms of RMSE, CI coverage, and length under Settings S1, S2, and S3. The larger value of the constant a in Settings S1 and S2 stands for a higher nonlinearity level in the treatment model, and the larger value of b in Setting S3 for a higher linearity level. In addition, a larger value of a and b indicate larger generalized IV strength.

- Setting B1: $f(Z_i, X_i) = -\frac{25}{12} + Z_i + Z_i^3/3 + Z_i \cdot (a \cdot \sum_{j=1}^{5} X_{i,j}) - \frac{3}{10} \cdot \sum_{j=1}^{p} X_{i,j}$,
- Setting B2: $f(Z_i, X_i) = \sin(2\pi Z_i) + \frac{3}{2}\cos(2\pi Z_i) + Z_i \cdot (a \cdot \sum_{j=1}^{5} X_{i,j}) - \frac{3}{10} \cdot \sum_{j=1}^{p} X_{i,j}$.

The value of the constant 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 vary a across $\{0, 0.5, 1\}$ and n across $\{1000, 3000, 5000\}$ to observe the performance.

We consider the outcome model (3) with two forms of $g(Z_i, X_i)$: (a)Vio=1: $g(Z_i, X_i) = Z_i + 1/5 \cdot \sum_{j=1}^{p_x} X_{i,j}$; (b)Vio=2: $g(Z_i, X_i) = Z_i + Z_i^2 - 1 + 1/5 \cdot \sum_{j=1}^{p_x} X_{i,j}$. We set the errors $\{(\delta_i, \epsilon_i)^\intercal\}_{1 \le i \le n}$ as heteroscedastic following Bekker and Crudu (2015): for $1 \le i \le n$, generate $\delta_i \sim N(0, Z_i^2 + 0.25)$ and $\epsilon_i = 0.6\delta_i + \sqrt{[1 - 0.6^2]/[0.86^4 + 1.38072^2]}(1.38072 \cdot \tau_{1,i} + 0.86^2 \cdot \tau_{2,i})$, where conditioning on Z_i, $\tau_{1,i}$ and $\tau_{2,i}$ are generated to be independent of δ_i, with $\tau_{1,i} \sim N(0, Z_i^2 + 0.25)$ and $\tau_{2,i} \sim N(0, 1)$.

We shall implement `TSCI` with random forests as detailed in Algorithm 1. Since the IV is possibly invalid, we consider four possible basis sets $\{\mathcal{V}_q\}_{0 \le q \le 3}$ to approximate $g(z, x)$, where $\mathcal{V}_0 = \{\overrightarrow{\mathbf{w}}(x)\}$ and $\mathcal{V}_q = \{z, \cdots, z^q, \overrightarrow{\mathbf{w}}(x)\}$ for $1 \le q \le 3$ with $\overrightarrow{\mathbf{w}}(x) = \{1, x_1, \cdots, x_{p_x}\}$.

Our proposed TSCI is designed to choose the best $\mathcal{V}_q$. We consider both the comparison method and robust method to select the best $\mathcal{V}_q$ (step 8 and 9 in Algorithm 1) and denote these two selections as $\mathcal{V}_{\widehat{q}_c}$ and $\mathcal{V}_{\widehat{q}_r}$, respectively. With the selected basis set $\mathcal{V}_{\widehat{q}_c}$ or $\mathcal{V}_{\widehat{q}_r}$, we calculate the TSCI estimator and construct the CI following (18) and (19).

As "naive" benchmarks, we compare TSCI with three different random forests based methods in the oracle setting, where the best $\mathcal{V}_q$ is known a priori. In particular, RF-Init denotes the TSCI estimator without bias correction; RF-Full is to implement TSCI without data-splitting; RF-Plug is the estimator of directly plugging the ML fitted treatment in the outcome model, as a naive generalization of TSLS. We give the detailed expressions of these three estimators and their CIs in Section A.1 in the supplement.

			TSCI-RF				Proportions of selection				TSLS	Other RF(oracle)		
vio	a	n	Oracle	Comp	Robust	Invalidity	$\widehat{q}_c=0$	$\widehat{q}_c=1$	$\widehat{q}_c=2$	$\widehat{q}_c=3$		Init	Plug	Full
1	0.0	1000	0.91	0.01	0.01	0.01	0.99	0.01	0.00	0.00	0.00	0.80	0.38	0.01
		3000	0.92	0.92	0.92	1.00	0.00	0.84	0.16	0.00	0.00	0.91	0.64	0.00
		5000	0.91	0.92	0.93	1.00	0.00	0.85	0.15	0.00	0.00	0.89	0.74	0.00
	0.5	1000	0.91	0.23	0.25	0.24	0.76	0.24	0.01	0.00	0.00	0.84	0.56	0.02
		3000	0.95	0.94	0.94	1.00	0.00	0.97	0.02	0.01	0.00	0.91	0.43	0.00
		5000	0.92	0.92	0.91	1.00	0.00	0.98	0.01	0.01	0.00	0.88	0.09	0.00
	1.0	1000	0.96	0.92	0.92	0.95	0.05	0.93	0.01	0.00	0.00	0.91	0.52	0.08
		3000	0.94	0.94	0.95	1.00	0.00	0.99	0.01	0.00	0.00	0.92	0.00	0.00
		5000	0.94	0.94	0.94	1.00	0.00	0.98	0.02	0.01	0.00	0.92	0.00	0.00

Table 1: Coverage for Setting B1 with Vio=1. The columns indexed with "TSCI-RF" correspond to our proposed TSCI with random forests, 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 indexed with "Proportions of selection" reports the proportions of basis sets $\mathcal{V}_q$ for $0 \leq q \leq 3$ selected by TSCI-RF. The column indexed with "TSLS" corresponds to the TSLS estimator. The columns indexed with "Init", "Plug", and "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 $\mathcal{V}_q$.

In Table 1, we compare the coverage properties of our proposed TSCI, TSLS, and the above three random forests based methods with Vio=1 in the outcome model. We observe that TSLS fails to have desired coverage probability 95% due to the existence of invalid IV; furthermore, RF-Full and RF-Plug have coverages far below 95% and RF-Init has slightly lower coverage due to the finite-sample bias. In contrast, our proposed TSCI achieves the desired coverage with a relatively strong interaction or large sample size. This is mainly due to its correct selection of $\mathcal{V}_q$ in those settings. In some settings with small interaction and small sample size (like $a = 0$, $n = 1000$), TSCI fails to select the correct basis set, and the coverage is below 95%. For the outcome model with Vio=2, TSCI can have desired coverage as well with a relatively strong interaction or large sample size; see Table 8 in the supplement.

For Setting B1, we further report the absolute bias and CI length in Tables 9 and 10 in the supplement, respectively. In Section D.4, we consider a binary IV setting, denoted as B3, and observe that Settings B2 and B3 exhibit a similar pattern to those of B1. Setting B2 is easier than B1 in the sense that the generalized IV strength remains relatively large even after adjusting for quadratic violation forms in the basis set $\mathcal{V}_2$.

6.3 TSCI with Various Nonlinearity Levels

In the following, we explore the performance of our proposed TSCI when this identification condition (Condition 1) fails to hold, that is, the conditional mean model $f(z, x)$ is not more complex than $g(z, x)$. To approximate such a regime, we consider the outcome model $Y_i = D_i/2+Z_i+\sum_{i=1}^{p_x} X_{i,j}^2-\cos(\pi H_i/4)+e_{i,2}$ and the treatment model $D_i = f(Z_i, X_i)-H_i/2+e_{i,1}$ with different specification of f detailed in Table 2 where $e_{i,1}$,$e_{i,2}$ are independent random noises following the standard normal distribution. In such a generating model, when a gets close to zero, $f(z, x)$ becomes a linear function in z, and Condition 1 fails since $g(z, x)$ is also linear in z. We fix $n = 3000$ and $p_x = 5$, generate Z_i and X_i as detailed in Table 2, and generate H_i as $H_i \mid X_i \sim N(2 \cdot \sin(X_{i,1}), 1)$.

Settings	Distribution Z_i and X_i	Treatment model
C1	Same as Setting B1 in Section 6.2	$f(Z_i, X_i) = Z_i + 1/2 \cdot Z_i \cdot (a \cdot \sum_{i=1}^{p_x} X_{i,j}) - 25/12 - 3/10 \cdot \sum_{i=1}^{p_x} X_{i,j}$
C2	Same as Setting S3 in Section 6.1	
C3	Same as Setting B1 in Section 6.2	$f(Z_i, X_i) = Z_i + a \cdot (\sin(2\pi Z_i) + 3/2 \cdot \cos(2\pi Z_i) - 3/10 \cdot \sum_{i=1}^{p_x} X_{i,j}$

Table 2: Different simulation settings, where $f(z, x)$ becomes a linear function of z with $a \to 0$.

We implement TSCI detailed in Algorithm 1 by specifying $\{\mathcal{V}_q\}_{0\le q\le 3}$ as $\mathcal{V}_q = \{z, \ldots, z^q, \overrightarrow{\mathbf{b}}(x)\}$ with $\overrightarrow{\mathbf{b}}(x) = \{1, \overrightarrow{\mathbf{b}}_1(x_1), \cdots, \overrightarrow{\mathbf{b}}_{p_x}(x_{p_x})\}$, where $\overrightarrow{\mathbf{b}}_j(x_j)$ denotes the B-spline basis functions defined in Section 6.1. We follow step 8 in Algorithm 1 to select the best basis set $\mathcal{V}_{\widehat{q}_c}$. We then construct the TSCI estimator in (18) and CIs as in (19) with the selected basis.

In Figure 3, we compare TSCI to DML and TSLS when treatment conditional mean models change from linear ($a = 0$) to nonlinear ($a > 0$); Note that in addition, a larger value of a increases the generalized IV strength. We vary the value of a in $\{0, 0.1, 0.2, \cdots, 1.2\}$ where larger values of a represent more nonlinear dependence between D_i and Z_i. When a is 0, Condition 1 fails to hold, and hence TSCI cannot identify the treatment effect when there are invalid IVs. In such settings, TSCI has similar performance to DML and TSLS. However, when a gets slightly larger than 0, TSCI has a better RMSE than DML and TSLS, especially for settings C2 and C3. When a is sufficiently large (e.g., $a \ge 0.1$ in Setting C2), TSCI detects the invalid IV and achieves the desired coverage. However, DML and TSLS do not have coverage since they are designed for valid IV settings.

We shall point out that, in Setting C2, the confidence intervals by TSCI are shorter than that of DML, even after adjusting for the linear invalid IV forms. This happens since the remaining nonlinear IV strength after adjusting for the invalidity form is even larger than the IV strength due to the linear association.

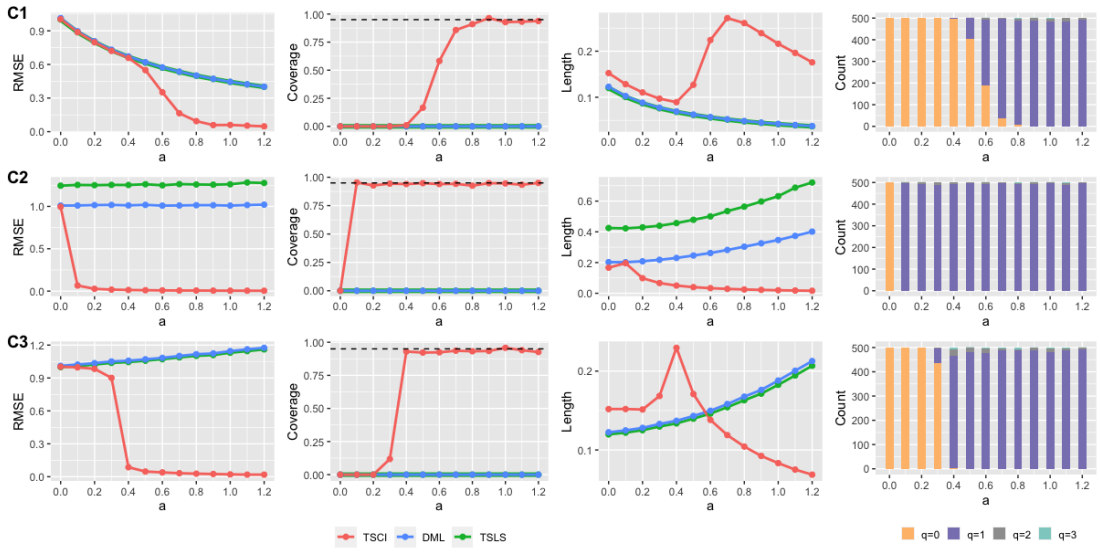

Figure 3: Comparison of `TSCI`, `DML` and `TSLS` in terms of RMSE, CI coverage, and length in Settings C1, C2 and C3, where a controls the nonlinearity level in the treatment model. The stacked bar charts show the basis selection of `TSCI`, where $q = 1$ corresponds to the correct selection.

6.4 Sensitivity to Basis Specification for Violation Approximation

Next we examine the sensitivity of `TSCI` to the choice of basis functions used to approximate violations. The main purposes of this section are: (1) to illustrate a crucial trade-off when selecting a basis: a more flexible basis function can capture more complex violation forms $g(Z_i, X_i)$ but may substantially weaken the IV strength; and (2) to highlight that when the basis function strikes a balance in this trade-off—that is to capture the violation $g(Z_i, X_i)$ without fully spanning $f(Z_i, X_i)$—`TSCI` is insensitive to basis choices. By showing these two points, this section provides further support for using expert-chosen IVs, as recommended in Section 3. Because these IVs are identified with domain knowledge and tend to have weak violation forms, a parametric basis is sufficient to approximate the violation without fully spanning the IV-treatment association. This, in turn, allows `TSCI` to deliver valid inference.

We consider the same data generation process as in Setting B1 in Section 6.2 to generate covariates X_i, IV Z_i and noises. To explore the trade-off in basis selection, we design three settings of generating treatment models and violation forms, as summarized in Table 3. Based on the functions $f(\cdot)$ and $g(\cdot)$ specified in Table 3, we generate the treatment D_i following $D_i = f(Z_i, X_i) + \delta_i$ and the outcome Y_i following $Y_i = D_i + g(Z_i, X_i) + \epsilon_i$. Across Settings D1-D3, the treatment model is designed to be more complex than the violation form with respect to the IV Z_i. For example, in Settings D1-D2, the treatment models contain additional interaction terms compared to the violation forms. In Setting D3, the treatment model includes a sine component, which is more complex than the polynomial structure of the corresponding violation function.

Settings	Treatment model	Violation form
D1	$f(Z_i, X_i) = -\frac{25}{12} + Z_i + Z_i^3/3 + Z_i \cdot \sum_{j=1}^{5} X_{i,j} - \frac{3}{10}\sum_{j=1}^{p} X_{i,j}$	$g(Z_i, X_i) = 4\sin(2Z_i) + Z_i^2 + \frac{1}{5}\sum_{j=1}^{p} X_{i,j}$
D2		$g(Z_i, X_i) = 2Z_i + Z_i^2 + \frac{1}{5}\sum_{j=1}^{p_x} X_{i,j}$
D3	$f(Z_i, X_i) = -\frac{25}{12} + 4\sin(2Z_i) + Z_i^3/3 - \frac{3}{10}\sum_{j=1}^{p_x} X_{i,j}$	

Table 3: Simulation settings with varying complexities in treatment models and violations.

We implement `TSCI` with random forests as detailed in Algorithm 1 and use different basis functions to approximate the violation form in the following. Particularly, we consider three hierarchically structured sets of basis functions for the violation forms as outlined in Table 4. The basis set $\mathcal{V}_0$ only includes the intercept and the covariates as it corresponds to the valid IV regime; The basis set $\mathcal{V}_1$ augments the basis set $\mathcal{V}_0$ with different basis functions, including polynomial basis up to degree 3, B-spline basis with 10 knots and Chebyshev polynomial basis with degree 11 of z; the largest violation space $\mathcal{V}_2$ further includes interaction functions between IV and covariates. For each basis specification, `TSCI` selects the best violation space $\mathcal{V}_{\widehat{q}_c}$ as in step 8 of Algorithm 1 and proceeds with the estimator defined in (18) and CI in (19).

	$\mathcal{V}_0$	$\mathcal{V}_1$	$\mathcal{V}_2$
Basis 1	$\{1, x_1, \ldots, x_{p_x}\}$	$\mathcal{V}_0 \cup \{\overrightarrow{\mathbf{poly}}(z)\}$	$\mathcal{V}_1 \cup z \cdot \{x_1, \ldots, x_{p_x}\}$
Basis 2		$\mathcal{V}_0 \cup \{\overrightarrow{\mathbf{B}}(z)\}$	
Basis 3		$\mathcal{V}_0 \cup \{\overrightarrow{\mathbf{cheb}}(z)\}$	

Table 4: Different basis specifications used to implement `TSCI`, where $\overrightarrow{\mathbf{poly}}(z)$ is the polynomial basis up to degree 3, $\overrightarrow{\mathbf{B}}(z)$ is the B-spline basis with 10 knots of z, and $\overrightarrow{\mathbf{cheb}}(z)$ is the Chebyshev polynomial basis with degree 11 of z.

In Table 5, we report the `TSCI` results obtained by applying Bases 1-3 to Settings D1-D3. We analyze the results in each setting separately. In Setting D1, a polynomial basis is not flexible enough to approximate the sine component in the violation, leading to undercoverage under Basis 1. Both B-spline and Chebyshev bases, however, capture the violation well without fully spanning the treatment model due to interaction terms, thereby achieving the desired balance of the trade-off between violation approximation and preserving IV strength. Consequently, `TSCI` produces robust inference under both Bases 2 and 3. In Setting D2, the violation form is polynomial based, and thus all three bases approximate it well. Because the treatment model still includes interaction terms that none of the bases can fully capture, all three bases achieve the balance of the trade-off, and `TSCI` is robust across Bases 1-3. In Setting D3, only Basis 1 achieves the balance. Both the B-spline and Chebyshev polynomials fully capture the IV-treatment association—where the interaction terms are replaced by a sine function—leading to weak IVs after adjustment. Consequently, the IV strength test fails under the set $\mathcal{V}_1$, forcing `TSCI` to revert to the valid-IV assumption and base inference on $\mathcal{V}_0$. The resulting inference is similar to those from

Settings	n	Basis 1 (polynomial)				Basis 2 (B-spline)				Basis 3 (Chebyshev polynomial)			
		Coverage	Frequency of selection			Coverage	Frequency of selection			Coverage	Frequency of selection		
			$\widehat{q}_c = 0$	$\widehat{q}_c = 1$	$\widehat{q}_c = 2$		$\widehat{q}_c = 0$	$\widehat{q}_c = 1$	$\widehat{q}_c = 2$		$\widehat{q}_c = 0$	$\widehat{q}_c = 1$	$\widehat{q}_c = 2$
D1	1000	0.422	283	217	0	0.200	397	103	0	0.194	399	101	0
	3000	0.902	11	489	0	0.918	19	481	0	0.920	17	483	0
	5000	0.890	2	498	0	0.930	1	499	0	0.930	2	498	0
D2	1000	0.676	140	360	0	0.408	276	224	0	0.408	279	221	0
	3000	0.928	0	500	0	0.938	0	500	0	0.928	0	500	0
	5000	0.928	0	500	0	0.930	0	500	0	0.930	0	500	0
D3	1000	0.932	3	492	5	0.058	500	0	0	0.068	500	0	0
	3000	0.956	0	499	1	0.002	500	0	0	0.002	500	0	0
	5000	0.952	0	492	8	0.000	500	0	0	0.000	500	0	0

Table 5: Empirical coverage and violation-space selection frequencies of `TSCI` across 500 simulations. The columns under "Frequency of selection" report the frequencies with which `TSCI` selected each violation space $\mathcal{V}_q$ (where $0 \leq q \leq 2$) for the different basis specifications detailed in Table 4.

`TSLS`, leading to substantial undercoverage under Bases 2 and 3. The polynomial basis, however, preserves the IV strength by being too coarse to fit the sine component. This explains why `TSCI` achieves the desired coverage only under Basis 1.

Based on the findings in this section, the selected basis functions need to be flexible enough to capture the violation term $g(Z_i, X_i)$ without fully spanning the conditional mean function $f(Z_i, X_i)$ in the treatment model. When `TSCI` is implemented with IVs identified by domain experts, as recommended in Section 3, the resulting violation forms tend to be simple and can be approximated by parametric basis functions. Such parametric bases are typically not flexible enough to fully span the nonlinear IV-treatment association $f(Z_i, X_i)$, thus preserving sufficient IV strength after violation adjustment for valid downstream inference. Therefore, in practice, we recommend using parametric basis functions together with expert-chosen IVs to implement the proposed `TSCI` for robust causal inference.

7. Real Data Analysis

We revisit the question on the effect of education on income (Card, 1993). We follow Card (1993) and analyze the same data set from the National Longitudinal Survey of Young Men. The outcome is the log wages (`lwage`), and the treatment is the years of schooling (`educ`). As argued in Card (1993), 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. Following Card (1993), we use the indicator for a nearby 4-year college in 1966 (`nearc4`) as the IV and include the following covariates: a quadratic function of potential experience (`exper` and `expersq`), a race indicator (`black`), and dummy variables for residence in a standard metropolitan statistical area in 1976 (`smsa`) and in 1966 (`smsa66`), and the dummy variable for residence in the south in 1976 (`south`), and a full set of 8 regional dummy variables(`reg1` to `reg8`). The data set consists of $n = 3010$ observations, which is made available by the R package `ivmodel` (Kang et al., 2021).

7.1 Analysis by TSCI

We implement random forests for the treatment model and report the variable importance in Figure 9 in the supplement, where the IV `nearc4` ranks the seventh in terms of variable importance, after the covariates `exper,expersq`, `black,south,smsa,smsa66.`

We allow for different IV violation forms by approximating $g(z, x)$ with various violation spaces $\{\mathcal{V}_q\}_{q=0,1,2}$ detailed in Table 6. Particularly, since $\mathcal{V}_0$ does not involve the IV `nearc4`, $\mathcal{V}_0$ corresponds to the valid IV setting as assumed in the analysis of Card (1993); moreover, $\mathcal{V}_1$ and $\mathcal{V}_2$ correspond to invalid IV settings by allowing the IV `nearc4` to affect the outcome directly or interactively with the baseline covariates. The main difference is that $\mathcal{V}_1$ includes the interaction with the six most important covariates while $\mathcal{V}_2$ includes all fourteen covariates. In Section E.1 of the supplement, we consider several alternative choices of $\mathcal{V}_2$, including versions that add the IV's cubic interaction with `exper` and versions that incorporate two-way interaction terms between the IV, `exper`, and key demographic indicators. Across all these alternatives, we observe that the resulting inference remains consistent with the one obtained by using violation spaces specified in Table 6.

$\mathcal{V}_0$	{exper,expersq,black,south,smsa,smsa66,reg1,reg2,reg3,reg4,reg5,reg6,reg7,reg8}
$\mathcal{V}_1$	$\mathcal{V}_0 \cup$ nearc4 $\cdot$ {1, exper, expersq, black, south, smsa, smsa66}
$\mathcal{V}_2$	$\mathcal{V}_0 \cup$ nearc4 $\cdot$ {1, exper, expersq, black, south, smsa, smsa66, reg1, reg2, reg3, reg4, reg5, reg6, reg7, reg8}

Table 6: Definitions of $\mathcal{V}_0, \mathcal{V}_1$ and $\mathcal{V}_2$, which are used to approximate $g(\cdot)$.

We implement Algorithm 1 to choose the best from $\{\mathcal{V}_0, \mathcal{V}_1, \mathcal{V}_2\}$, and construct the `TSCI` estimator. 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 4, we compare estimates of `OLS`, `TSLS`, and `TSCI`, which uses $\mathcal{V}_{\widehat{q}_c}$ reported from Algorithm 1. The median of these 500 `TSCI` estimators is 0.0604, 87.2% of these 500 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".

On the rightmost panel of Figure 4, we compare different confidence intervals (CIs). The `TSLS` CI is $(0.0238, 0.2393)$, and the `DML` CI is $(0.0061, 0.1907)$, which are both much wider than the CIs by `TSCI`. This wide interval results from the relatively weak IV because `TSLS` and `DML` only consider the linear association between IV and the treatment. The CI by `TSCI` with random forests varies with the specific sample splitting. We follow Meinshausen et al. (2009) and implement the multi-split method to adjust the finite-sample variation due to sample splitting; see Section A.3 in the supplement. The multi-split CI is (0.0294, 0.0914). The CIs based 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.

We shall point out that the IV strengths for the `TSCI` estimators, even after adjusting for the selected basis functions, are typically much larger than the `TSLS` (the concentration parameter is 13.33), which is illustrated in the middle panel of Figure 4. This happens since

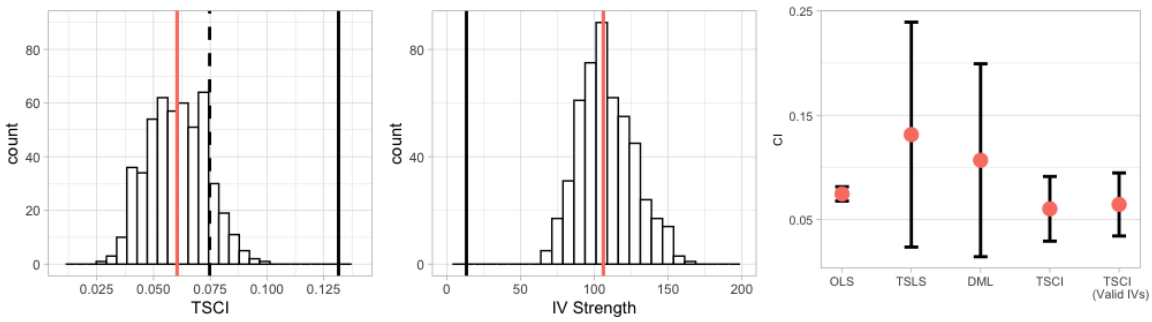

Figure 4: The leftmost panel reports the histograms of the `TSCI` (random forests) estimates with the comparison method, where the estimates differ due to the randomness of different 500 realized sample splittings; the solid red line corresponds to the median of the `TSCI` estimates, while the solid and dashed black lines correspond to the `TSLS` and `OLS` estimates, respectively. The middle panel displays the histogram of the generalized IV strength (after adjusting for $\mathcal{V}_{\widehat{q}_c}$) 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`. The rightmost panel compares different confidence intervals (CIs) produced by `OLS`, `TSLS`, `DML`, and our proposed `TSCI` with $\mathcal{V}_{\widehat{q}_c}$, and `TSCI` assuming valid IVs. The CIs of TSCI are adjusted by the multi-splitting method due to finite samples; see Appendix A.3.

the first-stage ML captures much more associations than the linear model in `TSLS`, even after adjusting for the possible IV violations captured by $\mathcal{V}_{\widehat{q}_c}$.

Among 500 sample splittings, the proportion of choosing $\mathcal{V}_0, \mathcal{V}_1$, and $\mathcal{V}_2$ are 59.2%, 38.2%, and 2.6%, respectively, where around 41% of splittings report `nearc4` as invalid IVs. Importantly, 93.6% out of the 500 splittings report $Q_{\max} > \widehat{q}_c$, which represents that $\widehat{\beta}_{\mathcal{V}_{\widehat{q}_c}}$ is not statistically different from $\widehat{\beta}_{\mathcal{V}_{Q_{\max}}}$ produced by the largest $\mathcal{V}_{Q_{\max}}$. This indicates that $\mathcal{V}_{\widehat{q}_c}$ already provides a reasonably good approximation of $g(Z_i, X_i)$ in the outcome model.

7.2 Falsification Argument of Condition 1

Next we provide a falsification argument to further assess the plausibility of the `TSCI` results. Specifically, if the argument fails, this provides evidence that the `TSCI` identification condition is violated; otherwise, there is no clear evidence against the condition. As demonstrated in the following, our application supports the falsification argument, which strengthens the credibility of the empirical findings in Section 7.1.

We now describe the main intuition behind our falsification argument. If Condition 1 holds for a specific $\mathcal{V}$, it means that $\mathcal{V}$ approximates the violation function $g(\cdot)$ well without fully spanning the conditional mean function $f(\cdot)$ of the treatment. We use random forests as a benchmark algorithm for fitting $g(\cdot)$ and $f(\cdot)$ and compare their mean squared errors (MSEs) to that obtained when fitting $g(\cdot)$ (or $f(\cdot)$) using the basis functions $\mathcal{V}$. We

say that Condition 1 passes for a specific $\mathcal{V}$ if it yields (1) a similar MSE for fitting the violation function $g(\cdot)$ compared to random forests, but (2) a much larger MSE for fitting the conditional treatment model $f(\cdot)$ compared to random forests.

Based on the above intuition, we evaluate the falsification argument for the chosen $\mathcal{V}_{\widehat{q}_c}$. To begin with, we construct the pseudo outcome $\widetilde{Y} = Y - D\widehat{\beta}_{\mathcal{V}_{\widehat{q}_c}}$ using the `TSCI` estimator $\widehat{\beta}_{\mathcal{V}_{\widehat{q}_c}}$, where the pseudo outcomes $\{\widetilde{Y}_i\}_{1\le i\le n}$ can be used as proxies for $\{g(Z_i, X_i)\}_{1\le i\le n}$ if $\widehat{\beta}_{\mathcal{V}_{\widehat{q}_c}}$ is a reasonably good estimator of β. Then we implement two OLS regressions of the pseudo outcome on the covariates in $\mathcal{V}_1$ and $\mathcal{V}_2$ as defined in Table 6. In addition, we build a random forest prediction model for the pseudo outcome with predictors Z_i and X_i. To evaluate performance, we split the data into a training set $\mathcal{A}_2$ and a test set $\mathcal{A}_1$ as in Section 4.1, where the test set $\mathcal{A}_1$ is used to estimate the out-of-sample MSE of the model constructed with the training set $\mathcal{A}_2$. We randomly split the data 500 times and report the violin plot of the 500 MSEs in the left panel of Figure 5. Since our specified basis set $\mathcal{V}_1$ leads to prediction performance similar to that of the random forest prediction model, this suggests that $\mathcal{V}_1$ provides a good approximation of the function $g(\cdot)$.

In comparison, we replace the pseudo outcome with the treatment variable and compare the MSEs of the three prediction models. In the right panel of Figure 5, the random forest prediction model performs much better than the OLS models with covariates in $\mathcal{V}_1$ and $\mathcal{V}_2$, indicating that $\mathcal{V}_1$ does not provide a good approximation for $f(\cdot)$ in the treatment model.

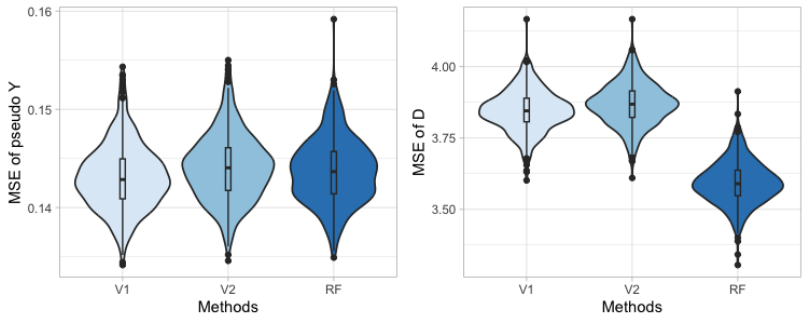

Figure 5: Violin plot of MSE for different prediction models, where the left and right panels correspond to using $\widetilde{Y}_i = Y_i - D_i\widehat{\beta}_{\mathcal{V}_{\widehat{q}_c}}$ and D_i as the regression outcome variables, respectively. In both panels, "V1", "V2", and "RF" respectively stand for OLS regression with $\mathcal{V}_1$, $\mathcal{V}_2$, and random forests with Z_i and X_i, where $\mathcal{V}_1$ and $\mathcal{V}_2$ are specified in Table 6.

8. Conclusion and Discussion

We integrate modern ML algorithms into the framework of instrumental variable analysis. We devise a novel `TSCI` methodology which provides reliable causal conclusions even with a wide range of violation forms. 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 IVs.

The current paper focuses on inference for a linear and constant treatment effect. An interesting direction is to extend our proposal to regimes with heterogeneous treatment effects (Athey and Imbens, 2016; Wager and Athey, 2018) but allowing for potentially invalid instruments. Specifically, we can consider the outcome model

$$Y_i = D_i \cdot \tau(Z_i, X_i) + g(Z_i, X_i) + \epsilon_i,$$

where $\tau(Z_i, X_i)$ represents the heterogeneous treatment effect and $g(Z_i, X_i)$ encodes violations of the exclusion restriction (A3) and independence (A2) assumptions. To estimate the heterogeneous treatment effect $\tau(Z_i, X_i)$ in the above model, we could approximate both functions τ and g using basis functions such that $\tau(Z_i, X_i) \approx V_i^\intercal \gamma$ and $g(Z_i, X_i) \approx V_i^\intercal \pi$ where V_i is the basis derived from Z_i and X_i. The outcome regression would then include terms such as $D \cdot V\gamma$, and we may adapt our proposal with control function methods. However, a rigorous methodology of combining the control function approach with machine learning and violation adjustment in the current context has not yet been developed and is left for future research.

Acknowledgments

We acknowledge valuable suggestions from Xu Cheng, Tirthankar Dasgupta, Qingliang Fan, Michal Kolesár, Greg Lewis, Yuan Liao, Molei Liu, Zhonghua Liu, Zuofeng Shang, and Frank Windmeijer. P. Bühlmann received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 786461).

Appendix A. Additional Methods and Theories

This section provides additional methodological and theoretical details for `TSCI`. We first discuss the naive plug-in estimators, practical implementations, and finite-sample adjustments used alongside the main procedure. We then show how the `TSCI` framework extends to other first-stage learners, including boosting, deep neural networks, and B-splines. We also discusses additional properties and asymptotic results for the homoscadastic-correlation setting, the variance estimator, and the comparison test.

A.1 Pitfalls with the Naive Machine Learning

As the benchmarks in Section 6.2, we also include the point estimators $\widehat{\beta}_{\text{init}}$, $\widehat{\beta}_{\text{plug}}$, and $\widehat{\beta}_{\text{full}}$ together with their corresponding standard errors. We construct the 95% confidence interval as $(\widehat{\beta} - 1.96 * \widehat{\text{SE}}, \widehat{\beta} + 1.96 * \widehat{\text{SE}})$.

RF-Init. We compute $\widehat{\beta}$ as $\widehat{\beta}_{\text{init}}$ in (15) and $\widehat{\sigma}_\epsilon(V) = \sqrt{\|P_V^\perp[Y - D\widehat{\beta}_{\text{init}}(V)]\|_2^2/(n-r)}$. Calculate $\widehat{\text{SE}} = \widehat{\sigma}_\epsilon \sqrt{D_{\mathcal{A}_1}^\intercal [\mathbf{M}(V)]^2 D_{\mathcal{A}_1}} / D_{\mathcal{A}_1}^\intercal \mathbf{M}(V) D_{\mathcal{A}_1}$.

RF-Plug and its pitfall. A simple idea of combining `TSLS` with ML is to directly use $\widehat{f}_{\mathcal{A}_1}$ in the second stage. We may calculate such a two-stage estimator as $\widehat{\beta}_{\text{plug}}(V) = Y_{\mathcal{A}_1}^\intercal P_{V_{\mathcal{A}_1}}^\perp \widehat{f}_{\mathcal{A}_1} / \widehat{f}_{\mathcal{A}_1}^\intercal P_{V_{\mathcal{A}_1}}^\perp \widehat{f}_{\mathcal{A}_1}$, where the second-stage regression is implemented by regressing

$Y_{\mathcal{A}_1}$ on $\widehat{f}_{\mathcal{A}_1} = \Omega D_{\mathcal{A}_1}$ and $V_{\mathcal{A}_1}$. We compute the standard error $\widehat{\mathrm{SE}} = \sqrt{\frac{\|P_V^{\perp}(Y-\widehat{\beta}_{\mathrm{plug}}(V)D)\|_2^2}{|\mathcal{A}_1|\cdot \widehat{D}_{\mathcal{A}_1}^{\intercal} P_{V_{\mathcal{A}_1}}^{\perp}}}$. Angrist and Pischke (2009)[Sec.4.6.1] criticized such use of the nonlinear first-stage model under the name of "forbidden regression" and Chen et al. (2023) pointed out that $\widehat{\beta}_{\mathrm{plug}}(V)$ is biased since Ω in (11) is not a projection matrix for random forests.

RF-Full and its pitfall. Sample splitting is essential for removing the endogeneity of the ML-predicted values. Without sample splitting, the ML predicted value for the treatment can be close to the treatment (due to overfitting) and hence highly correlated with unmeasured confounders, leading to a `TSCI` estimator suffering from a significant bias. In the following, we take the non-split Random Forests as an example. Similarly to Ω defined in (11), we define the transformation matrix Ω^{full} for random forests constructed with the full data. As a modification of (15), we consider the corresponding `TSCI` estimator,

$$\widehat{\beta}_{\mathrm{full}}(V) = Y^{\intercal}\mathbf{M}^{\mathrm{full}}(V)D/D^{\intercal}\mathbf{M}^{\mathrm{full}}(V)D \quad \text{with} \quad \mathbf{M}_{\mathrm{RF}}^{\mathrm{full}}(V) = (\Omega^{\mathrm{full}})^{\intercal}P_{\widetilde{V}}^{\perp}\Omega^{\mathrm{full}}, \tag{37}$$

where $\widetilde{V} = \Omega^{\mathrm{full}}V$. We calculate its standard error as

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

The simulation results in Section 6 show that the estimator $\widehat{\beta}_{\mathrm{full}}(V)$ in (37) suffers from a large bias and the resulting confidence interval does not achieve the desired coverage.

A.2 Choice of $\widehat{\rho}$ in (26)

In the following, we present the choice of $\widehat{\rho} = \widehat{\rho}(\alpha_0) > 0$ used in (26) by the wild bootstrap. For $1 \le i \le n_1$, we compute $\widetilde{\epsilon}_i = [\widehat{\epsilon}(V_{Q_{\max}})]_i - \bar{\mu}_\epsilon$ where $\widehat{\epsilon}(V_{Q_{\max}})$ is defined in (20) with $V = V_{Q_{\max}}$ and $\bar{\mu}_\epsilon = \frac{1}{n_1}\sum_{i=1}^{n_1}[\widehat{\epsilon}(V_{Q_{\max}})]_i$. We generate $\epsilon_i^{[l]} = U_i^{[l]} \cdot \widetilde{\epsilon}_i$ for $1 \le i \le n_1$ and $1 \le l \le L$, where $\{U_i^{[l]}\}_{1\le i\le n_1}$ are i.i.d. standard normal random variables. We define

$$T^{[l]} = \max_{0\le q<q'\le Q_{\max}} \frac{1}{\sqrt{\widehat{H}(V_q, V_{q'})}}\left[\frac{f_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V_{q'})\epsilon^{[l]}}{f_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V_{q'})f_{\mathcal{A}_1}} - \frac{f_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V_q)\epsilon^{[l]}}{f_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V_q)f_{\mathcal{A}_1}}\right] \text{ for } 1 \le l \le L.$$

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

$$\widehat{\rho}(\alpha_0) = \min\left\{\rho \in \mathbb{R} : \frac{1}{L}\sum_{l=1}^{L}\mathbf{1}\left(|T^{[l]}| \le \rho\right) \ge 1-\alpha_0\right\}, \tag{38}$$

where α_0 is set as 0.025 by default.

A.3 Finite-Sample Adjustment of Uncertainty from Data Splitting

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 (Chernozhukov et al., 2018) and multi-splitting (Meinshausen et al., 2009). Following Chernozhukov et al. (2018) and Meinshausen et al. (2009), we introduce a confidence interval which aggregates multiple confidence intervals due to different splittings. Consider S random splittings and for the s-th splitting, we use $\widehat{\beta}^s$ and $\widehat{\mathrm{SE}}^s$ to denote the corresponding `TSCI` point and standard error estimators, respectively. Following Section 3.4 of Chernozhukov et al. (2018), we introduce the median estimator $\widehat{\beta}^{\mathrm{med}} = \mathrm{median}\{\widehat{\beta}^s\}_{1\le s\le S}$ together with $\widehat{\mathrm{SE}}^{\mathrm{med}} = \mathrm{median}\left\{\sqrt{[\widehat{\mathrm{SE}}^s]^2 + (\widehat{\beta}^s - \widehat{\beta}^{\mathrm{med}})^2}\right\}_{1\le s\le S}$, and construct the median confidence interval as $(\widehat{\beta}^{\mathrm{med}} - z_{\alpha/2}\widehat{\mathrm{SE}}^{\mathrm{med}}, \widehat{\beta}^{\mathrm{med}} + z_{\alpha/2}\widehat{\mathrm{SE}}^{\mathrm{med}})$. Alternatively, following Meinshausen et al. (2009), we construct the p values $p^s(\beta_0) = 2(1 - \Phi(|\widehat{\beta}^s - \beta_0|/\widehat{\mathrm{SE}}^s))$ for $\beta_0 \in \mathbb{R}$ and $1 \le s \le S$, where Φ is the CDF of the standard normal. We define the multi-splitting confidence interval as $\{\beta_0 \in \mathbb{R} : 2 \cdot \mathrm{median}\{p^s(\beta_0)\}_{s=1}^S \ge \alpha\}$. See equation (2.2) of Meinshausen et al. (2009).

A.4 Choices of Violation Basis Functions for Multiple IVs

When there are multiple IVs, there are more choices to specify the violation form, and $\{\mathcal{V}_q\}_{1\le q\le Q}$ are not necessarily nested. For example, when we have two IVs z_1 and z_2, we may set $\mathcal{V}_{q_1,q_2} = \{1, z_1, \cdots, z_1^{q_1}, z_2, \cdots, z_2^{q_2}\}$, and then $\{\mathcal{V}_{q_1,q_2}\}_{0\le q_1,q_2\le Q}$ are not nested. However, even in such a case, our proposed selection method is still applicable. Specifically, for any $0 \le q_1, q_2 \le Q$, we compare the estimator generated by $\mathcal{V}_{q_1,q_2}$ to that by $\mathcal{V}_{q_1',q_2'}$ with $q_1' \ge q_1$ and $q_2' \ge q_2$.

A.5 `TSCI` with Boosting

In the following, we demonstrate how to express the L_2 boosting estimator (Bühlmann and Hothorn, 2007; Bühlmann and Yu, 2006) in the form of (11). The boosting methods aggregate a sequence of base procedures $\{\widehat{g}^{[m]}(\cdot)\}_{m\ge1}$. For $m \ge 1$, 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 $\widehat{g}_{\mathcal{A}_1}^{[m]} = \big(\widehat{g}^{[m]}(X_1, Z_1), \cdots, \widehat{g}^{[m]}(X_{n_1}, Z_{n_1})\big)^\intercal$. With $\widehat{f}_{\mathcal{A}_1}^{[0]} = 0$ and $0 < \nu \le 1$ as the step-length factor (the default being $\nu = 0.1$), we conduct the sequential updates,

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

where $\mathcal{H}^{[m]} \in \mathbb{R}^{n_1\times n_1}$ is a hat matrix determined by the base procedures.

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

$$\widehat{f}_{\mathcal{A}_1}^{[m]} = \Omega^{[m]} D_{\mathcal{A}_1} \quad \text{with} \quad \Omega^{[m]} = \nu\mathcal{H}^{[m]} + (\mathrm{I} - \nu\mathcal{H}^{[m]})\Omega^{[m-1]}. \tag{39}$$

Define $\Omega^{\mathrm{boo}} = \Omega^{[m_{\mathrm{stop}}]}$ and write $\widehat{f}^{\mathrm{boo}} = \Omega^{\mathrm{boo}} D_{\mathcal{A}_1}$, which is the desired form in (11).

In the following, we give two examples of the construction of the base procedures and provide the detailed expression for $\mathcal{H}^{[m]}$ used in (39). We write $C_i = (X_i^\intercal, Z_i^\intercal)^\intercal \in \mathbb{R}^p$ and define the matrix $C \in \mathbb{R}^{n\times p}$ with its i-th row as $C_i^\intercal$. An important step of building the base procedures $\{\widehat{g}^{[m]}(Z_i, X_i)\}_{m\ge1}$ is to conduct the variable selection. That is, each base procedure is only constructed based on a subset of covariates $C_i = (X_i^\intercal, Z_i^\intercal) \in \mathbb{R}^p$.

Pairwise regression. 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 $1 \le i \le n$. 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 $\mathcal{H}^{[m]} = (S_{\mathcal{A}_1}^{\widehat{j}_m,\widehat{l}_m})^\intercal / \|S_{\mathcal{A}_1}^{\widehat{j}_m,\widehat{l}_m}\|_2^2$.

For the pairwise thin plate, we follow chapter 7 of Green and Silverman (1993) to construct the projection matrix. For $1 \le j < l \le p$, 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 $1 \le s \le n$, define $E_{s,s}^{j,l} = 0$; for $1 \le s \ne t \le n$, define

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

Then we define $\bar{\mathcal{P}}^{j,l} = \begin{pmatrix} E_{\mathcal{A}_2,\mathcal{A}_2}^{j,l} & (T_{\cdot,\mathcal{A}_2}^{j,l})^\intercal \end{pmatrix} \begin{pmatrix} E_{\mathcal{A}_2,\mathcal{A}_2}^{j,l} & (T_{\cdot,\mathcal{A}_2}^{j,l})^\intercal \\ T_{\cdot,\mathcal{A}_2}^{j,l} & 0 \end{pmatrix}^{-1} \in \mathbb{R}^{n_2 \times (n_2+3)}$. 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

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

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

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

A.6 `TSCI` with Deep Neural Network

In the following, we demonstrate how to calculate Ω for a deep neural network (James et al., 2013). 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 $1 \le k \le K_1$, where $\sigma(\cdot)$ is the activation function and $\{\omega_{k,l}^{(1)}\}_{1 \le k \le K_1, 1 \le l \le p_x+p_z}$ are parameters. For $m \ge 2$, 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 $1 \le k \le K_m$, where $\{\omega_{k,l}^{(m)}\}_{1 \le k \le K_m, 1 \le l \le K_{m-1}}$ are unknown parameters. For given $M \ge 1$, we estimate the unknown parameters based on the data $\{X_i, Z_i, D_i\}_{i \in \mathcal{A}_2}$,

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

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

$$\widehat{H}_{i,k}^{(m)} = \sigma\left(\widehat{\omega}_{k0}^{(m)} + \sum_{l=1}^{K_{m-1}} \widehat{\omega}_{kl}^{(m)} H_{i,l}^{(m-1)}\right) \quad \text{for} \quad 1 \le k \le K_m,$$

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 $m = 0$, $\widehat{f}_{\mathcal{A}_2}^{[0]} = 0$, 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 $U_{\mathcal{A}_2}^{[m]} = D_{\mathcal{A}_2} - \widehat{f}_{\mathcal{A}_2}^{[m-1]}$
6: Implement the following base procedure on $\{C_i, U_i^{[m]}\}_{i\in\mathcal{A}_2}$

$$(\widehat{j}_m, \widehat{l}_m) = \underset{1\leq j,l\leq p}{\arg\min} \left\| U_{\mathcal{A}_2}^{[m]} - \mathcal{P}^{j,l} U_{\mathcal{A}_2}^{[m]} \right\|^2.$$

7: Update $\widehat{f}_{\mathcal{A}_2}^{[m]} = \widehat{f}_{\mathcal{A}_2}^{[m-1]} + \nu \mathcal{P}^{\widehat{j}_m, \widehat{l}_m} \left(D_{\mathcal{A}_2} - \widehat{f}_{\mathcal{A}_2}^{[m-1]} \right)$
8: Construct $\mathcal{H}^{[m]}$ in the same way as $\mathcal{P}^{\widehat{j}_m, \widehat{l}_m}$ but with the data in $\mathcal{A}_1$.
9: Compute $\Omega^{[m]} = \nu\mathcal{H}^{[m]} + (\mathrm{I} - \nu\mathcal{H}^{[m]})\Omega^{[m-1]}$
10: **end for**
11: Return Ω^{boo}

with $\widehat{H}_{i,k}^{(1)} = \sigma\left(\widehat{\omega}_{k0}^{(1)} + \sum_{l=1}^{p_z} \widehat{\omega}_{kl}^{(1)} Z_{il} + \sum_{l=1}^{p_x} \widehat{\omega}_{k,l+p_z}^{(1)} X_{il}\right)$ for $1 \leq k \leq K_1$. We use $\Omega^{\text{DNN}} = \widehat{H}^{(M)} \left([\widehat{H}^{(M)}]^\intercal \widehat{H}^{(M)} \right)^{-1} [\widehat{H}^{(M)}]^\intercal$ 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 $\widehat{V}_{\mathcal{A}_1} = \Omega^{\text{DNN}} V_{\mathcal{A}_1}$, we define $\mathbf{M}_{\text{DNN}}(V) = \left[\Omega^{\text{DNN}}\right]^\intercal P_{\widehat{V}_{\mathcal{A}_1}}^{\perp} \Omega^{\text{DNN}}$, which is shown to be an orthogonal projection matrix in Lemma 1 in the supplement.

A.7 `TSCI` with B-Spline

As a simplification, we consider the additive model $\mathbf{E}(D_i \mid Z_i, X_i) = \gamma_1(Z_i) + \gamma_2(X_i)$, and assume that $\gamma_1(\cdot)$ can be well approximated by a set of B-spline functions $\{b_1(\cdot), \cdots, b_M(\cdot)\}$. In practice, the number M can be chosen by cross-validation. We define the matrix $B \in \mathbb{R}^{n\times M}$ with its i-th row $B_i = (b_1(Z_i), \cdots, b_M(Z_i))^\intercal$. 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 $\mathbf{M}_{\text{ba}}(V) = \left[\Omega^{\text{ba}}\right]^\intercal P_{\widehat{V},W}^{\perp} \Omega^{\text{ba}}$ with $\widehat{V} = \Omega^{\text{ba}} V$. The transformation matrix $\mathbf{M}_{\text{ba}}(V)$ is a projection matrix with $M - \text{rank}(V) \leq \text{rank}\left[\mathbf{M}_{\text{ba}}(V)\right] \leq M$. 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 if the B-spline is used for fitting the treatment model, which is different from the general machine learning algorithms.

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 $m = 0$, $\widehat{f}^{[0]}_{\mathcal{A}_2} = 0$, and $\Omega^{[0]} = 0$.
3: **for** $1 \leq m \leq m_{\text{stop}}$ **do**
4: Compute the adjusted outcome $U^{[m]}_{\mathcal{A}_2} = D_{\mathcal{A}_2} - \widehat{f}^{[m-1]}_{\mathcal{A}_2}$
5: Run the decision tree on $\{C_i, U^{[m]}_i\}_{i\in\mathcal{A}_2}$ and partition $\mathbb{R}^p$ as leaves $\{\mathcal{R}^{[m]}_1, \cdots, \mathcal{R}^{[m]}_L\}$;
6: For any $j \in \mathcal{A}_2$, we identify the leaf $\mathcal{R}^{[m]}_{l(C_j)}$ containing C_j and compute

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

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

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

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

A.8 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. Recall that

$$\mathbf{M}_{\text{RF}}(V) = \Omega^\intercal P^{\perp}_{\widehat{V}_{\mathcal{A}_1}}\Omega \quad \text{with} \quad \Omega_{ij} = \omega_j(Z_i, X_i),$$

where $\omega_j(z, x)$ is defined in (13).

Lemma 1 *The transformation matrix $\mathbf{M}_{\text{RF}}(V)$ satisfies*

$$\lambda_{\max}(\mathbf{M}_{\text{RF}}(V)) \leq 1 \quad \textit{and} \quad b^\intercal[\mathbf{M}_{\text{RF}}(V)]^2 b \leq b^\intercal\mathbf{M}_{\text{RF}}(V)b \quad \textit{for any} \quad b \in \mathbb{R}^{n_1}. \tag{40}$$

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

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

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

$$\lambda_{\max}(\mathbf{M}_{\text{boo}}(V)) \leq \|\Omega^{\text{boo}}\|_2^2, \quad \textit{and} \quad b^\intercal[\mathbf{M}_{\text{boo}}(V)]^2 b \leq \|\Omega^{\text{boo}}\|_2^2 \cdot b^\intercal\mathbf{M}_{\text{boo}}(V)b, \tag{41}$$

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

A.9 Homoscadastic Correlation

In the following, we discuss the simplified method and theory in the homoscadastic correlation setting, that is, $\mathrm{Cov}(\epsilon_i, \delta_i \mid Z_i, X_i) = \mathrm{Cov}(\epsilon_i, \delta_i)$. With this extra assumption, we present the following bias-corrected estimator as an alternative to the estimator defined in (18),

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

where $\widehat{\beta}_{\mathrm{init}}(V)$ is defined in (15) and the estimator of $\mathrm{Cov}(\delta_i, \epsilon_i)$ is defined as,

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

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

We present a simplified version of Theorem 6 by assuming homoscadastic 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 4 *Consider the model* (3) *and* (4) *with* $\mathrm{Cov}(\epsilon_i, \delta_i \mid Z_i, X_i) = \mathrm{Cov}(\epsilon_i, \delta_i)$. *Suppose that Conditions* (R1) *and* (R2) *hold, then*

$$\left| \widehat{\mathrm{Cov}}(\delta_i, \epsilon_i) - \mathrm{Cov}(\delta_i, \epsilon_i) \right| \leq \eta_n^0(V) \leq \eta_n(V), \tag{42}$$

where $\eta_n(V)$ *is defined in* (29) *and*

$$\eta_n^0(V) = \frac{\|f_{\mathcal{A}_1} - \widehat{f}_{\mathcal{A}_1}\|_2}{\sqrt{n}} + \sqrt{\frac{\log n}{n}} + \left(|\beta - \widehat{\beta}_{\mathrm{init}}(V)| + \frac{\|R(V)\|_2}{\sqrt{n}} \right) \left(1 + \frac{\|f_{\mathcal{A}_1} - \widehat{f}_{\mathcal{A}_1}\|_2}{\sqrt{n}} \right). \tag{43}$$

Furthermore, if we assume (30) *holds, then* $\widetilde{\beta}_{\mathrm{RF}}(V)$ *satisfies*

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

where $\mathrm{SE}(V)$ *is defined in* (55), $\mathcal{G}(V) \xrightarrow{d} N(0,1)$ *and there exist positive constants* $C > 0$ *and* $t_0 > 0$ *such that*

$$\liminf_{n \to \infty} \mathbb{P} \left(|\mathcal{E}(V)| \leq C \frac{\eta_n^0(V) \cdot \mathrm{Tr}[\mathbf{M}(V)] + \|R(V)\|_2 + t_0 \sqrt{\mathrm{Tr}([\mathbf{M}_{\mathrm{RF}}(V)]^2)}}{\sqrt{f_{\mathcal{A}_1} [\mathbf{M}(V)]^2 f_{\mathcal{A}_1}}} \right) \geq 1 - \exp(-t_0^2), \tag{44}$$

with $\eta_n^0(V)$ *defined in* (43).

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

A.10 Consistency of Variance Estimators

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

Lemma 2 *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*

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

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

A.11 Size and Power of $\mathcal{C}(V_q, V_{q'})$

The limiting distribution in Theorem 2 implies the size and power of the comparison test $\mathcal{C}(V_q, V_{q'})$ in (25), stated in the following theorem.

Theorem 5 *Suppose that the Conditions of Theorem 2 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* (34) *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* (25) *with corresponding* α_0 *satisfies,*

$$\lim_{n\to\infty} \mathbf{P}\left(\mathcal{C}(V_q, V_{q'}) = 1\right) = 1 - \Phi\left(z_{\alpha_0} - \mathcal{L}^*(V_q, V_{q'})\right) + \Phi\left(-z_{\alpha_0} - \mathcal{L}^*(V_q, V_{q'})\right). \tag{45}$$

With $\left|\mathcal{L}^*(V_q, V_{q'})\right| = 0$, *then we have* $\lim_{n\to\infty} \mathbf{P}\left(\mathcal{C}(V_q, V_{q'}) = 1\right) = 2\alpha_0$. *With* $\left|\mathcal{L}^*(V_q, V_{q'})\right| = \infty$, *then we have* $\lim_{n\to\infty} \mathbf{P}\left(\mathcal{C}(V_q, V_{q'}) = 1\right) = 1$.

Appendix B. Proofs

We establish Proposition 1 in Section B.1. In Section B.2, we establish Proposition 2. In Section B.3, we establish Theorems 6 and 1. We prove Theorems 2 and 5 in Section B.4. We establish Theorem 3 in Section B.5 and prove Theorem 4 in Section B.6.

We define the conditional covariance matrices $\Lambda, \Sigma^\delta, \Sigma^\epsilon \in \mathbb{R}^{n_1\times n_1}$ as $\Lambda = \mathbf{E}(\delta_{\mathcal{A}_1}\epsilon_{\mathcal{A}_1}^\intercal \mid X_{\mathcal{A}_1}, Z_{\mathcal{A}_1})$, $\Sigma^\delta = \mathbf{E}(\delta_{\mathcal{A}_1}\delta_{\mathcal{A}_1}^\intercal \mid X_{\mathcal{A}_1}, Z_{\mathcal{A}_1})$, and $\Sigma^\epsilon = \mathbf{E}(\epsilon_{\mathcal{A}_1}\epsilon_{\mathcal{A}_1}^\intercal \mid X_{\mathcal{A}_1}, Z_{\mathcal{A}_1})$. For any $i, j \in \mathcal{A}_1$ and $i \neq j$, we have $\Lambda_{i,j} = \mathbf{E}\left[\delta_j \mathbf{E}(\epsilon_i \mid X_{\mathcal{A}_1}, Z_{\mathcal{A}_1}, \delta_j) \mid X_{\mathcal{A}_1}, Z_{\mathcal{A}_1}\right]$. Since $\mathbf{E}(\epsilon_i \mid X_{\mathcal{A}_1}, Z_{\mathcal{A}_1}, \delta_j) = \mathbf{E}(\epsilon_i \mid X_i, Z_i) = 0$ 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 $\max_{1\le j\le n_1} \max\left\{|\Lambda_{j,j}|, |\Sigma^\delta_{j,j}|, |\Sigma^\epsilon_{j,j}|\right\} \le K^2$. We shall use $\mathcal{O}$ to denote the set of random variables $\{Z_i, X_i\}_{1\le i\le n}$ 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 Rudelson and Vershynin (2013).

Lemma 3 (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* $t \ge 0$, $\mathbf{P}\left(|\epsilon^\intercal A\epsilon - \mathbf{E}\epsilon^\intercal A\epsilon| > t\right) \le 2\exp\left[-c\min\left(\frac{t^2}{K^4\|A\|_F^2}, \frac{t}{K^2\|A\|_2}\right)\right]$.

Note that $2\epsilon^\intercal A\delta = (\epsilon+\delta)^\intercal A(\epsilon+\delta) - \epsilon^\intercal A\epsilon - \delta^\intercal A\delta$. If both ϵ_i and δ_i are sub-gaussian, we apply the union bound, Lemma 3 for both ϵ and δ and then establish,

$$\mathbf{P}\left(|\epsilon^\intercal A\delta - \mathbf{E}\epsilon^\intercal A\delta| > t\right) \le 6\exp\left[-c\min\left(\frac{t^2}{K^4\|A\|_F^2}, \frac{t}{K^2\|A\|_2}\right)\right]. \tag{46}$$

B.1 Proof of Proposition 1

Note that

$$\begin{aligned}
&\mathbf{E}(Y_i \mid Z_i = z, X_i = 1)\\
&= \mathbf{E}\left[D_i^{(z)}(x=1)Y_i^{(1,z)}(x=1) + (1 - D_i^{(z)}(x=1))Y_i^{(0,z)}(x=1) \mid Z_i = z, X_i = 1\right]\\
&= \mathbf{E}\left[D_i^{(z)}(x=1)Y_i^{(1,z)}(x=1) + (1 - D_i^{(z)}(x=1))Y_i^{(0,z)}(x=1) \mid X_i = 1\right].
\end{aligned}$$

We assume that $Y_i^{(1,z)}(x=1) - Y_i^{(1,w)}(x=1) = Y_i^{(1,z)}(x=0) - Y_i^{(1,w)}(x=0) = (z-w)\pi$ and obtain

$$\begin{aligned}
&\mathrm{ITT}_Y(x=1) \coloneqq \mathbf{E}(Y_i \mid Z_i = z, X_i = 1) - \mathbf{E}(Y_i \mid Z_i = w, X_i = 1)\\
&= \mathbf{E}\left[(D_i^{(z)}(x=1) - D_i^{(w)}(x=1))(Y_i^{(1,w)}(x=1) - Y_i^{(0,w)}(x=1)) \mid X_i = 1\right] + (z-w)\pi\\
&= \beta\mathbf{E}\left[D_i^{(z)}(x=1) - D_i^{(w)}(x=1) \mid X_i = 1\right] + (z-w)\pi\\
&= \beta\left(\mathbf{E}(D_i \mid Z_i = z, X_i = 1) - \mathbf{E}(D_i \mid Z_i = w, X_i = 1)\right) + (z-w)\pi\\
&= \beta\cdot\mathrm{ITT}_D(x=1) + (z-w)\pi.
\end{aligned} \tag{47}$$

Similarly, we have

$$\mathrm{ITT}_Y(x=0) = \beta\cdot\mathrm{ITT}_D(x=0) + (z-w)\pi. \tag{48}$$

We combine (47) and (48) to establish the results.

B.2 Proof of Proposition 2

Recall that we are analyzing $\widehat{\beta}_{\text{init}}(V)$ defined in (15) with replacing $\mathbf{M}(V)$ by $\mathbf{M}(V)$. We have decomposed the estimation error $\widehat{\beta}_{\text{init}}(V) - \beta$ as

$$\widehat{\beta}_{\text{init}}(V) - \beta = \frac{\epsilon_{\mathcal{A}_1}^\intercal\mathbf{M}(V)\delta_{\mathcal{A}_1} + \epsilon_{\mathcal{A}_1}^\intercal\mathbf{M}(V)f_{\mathcal{A}_1} + R_{\mathcal{A}_1}(V)^\intercal\mathbf{M}(V)D_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\intercal\mathbf{M}(V)D_{\mathcal{A}_1}}. \tag{49}$$

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

Lemma 4 *Under Condition* (R1)*, with probability larger than* $1-\exp\left(-c\min\left\{t_0^2, t_0\right\}\right)$ *for some positive constants* $c > 0$ *and* $t_0 > 0$,

$$\left|\epsilon_{\mathcal{A}_1}^\intercal\mathbf{M}(V)\delta_{\mathcal{A}_1} - \mathrm{Tr}[\mathbf{M}(V)\Lambda]\right| \le t_0K^2\sqrt{\mathrm{Tr}([\mathbf{M}(V)]^2)},\ \left|\epsilon_{\mathcal{A}_1}^\intercal\mathbf{M}(V)f_{\mathcal{A}_1}\right| \le t_0K\sqrt{f_{\mathcal{A}_1}^\intercal[\mathbf{M}(V)]^2f_{\mathcal{A}_1}}; \tag{50}$$

$$\left|\frac{D_{\mathcal{A}_1}^\intercal\mathbf{M}(V)D_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^\intercal\mathbf{M}(V)f_{\mathcal{A}_1}} - 1\right| \lesssim \frac{K^2\mathrm{Tr}[\mathbf{M}(V)] + t_0K^2\sqrt{\mathrm{Tr}([\mathbf{M}(V)]^2)} + t_0K\sqrt{f_{\mathcal{A}_1}^\intercal[\mathbf{M}(V)]^2f_{\mathcal{A}_1}}}{f_{\mathcal{A}_1}^\intercal\mathbf{M}(V)f_{\mathcal{A}_1}}. \tag{51}$$

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

Define $\tau_n = f_{\mathcal{A}_1}^\intercal \mathbf{M}(V) f_{\mathcal{A}_1}$. Together with (40) and the generalized IV strength condition (R2), we apply (51) 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 $1 - \exp\left(-c\tau_n^{1/2-c_0}\right)$,

$$\left| \frac{D_{\mathcal{A}_1}^\intercal \mathbf{M}(V) D_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^\intercal \mathbf{M}(V) f_{\mathcal{A}_1}} - 1 \right| \leq C \frac{K^2 \mathrm{Tr}[\mathbf{M}(V)]}{f_{\mathcal{A}_1}^\intercal \mathbf{M}(V) f_{\mathcal{A}_1}} + C \frac{K + K^2}{(f_{\mathcal{A}_1}^\intercal \mathbf{M}(V) f_{\mathcal{A}_1})^{c_0}} \leq 0.1. \tag{52}$$

Together with (49), we establish

$$\left| \widehat{\beta}_{\mathrm{init}}(V) - \beta \right| \lesssim \frac{\left| \epsilon_{\mathcal{A}_1}^\intercal \mathbf{M}(V) f_{\mathcal{A}_1} \right|}{f_{\mathcal{A}_1}^\intercal \mathbf{M}(V) f_{\mathcal{A}_1}} + \frac{\left| \epsilon_{\mathcal{A}_1}^\intercal \mathbf{M}(V) \delta_{\mathcal{A}_1} \right|}{f_{\mathcal{A}_1}^\intercal \mathbf{M}(V) f_{\mathcal{A}_1}} + \sqrt{\frac{R_{\mathcal{A}_1}(V)^\intercal \mathbf{M}(V) R_{\mathcal{A}_1}(V)}{f_{\mathcal{A}_1}^\intercal \mathbf{M}(V) f_{\mathcal{A}_1}}}. \tag{53}$$

By applying the decomposition (53) and the upper bounds (50), and (50) with $t_0 = (f_{\mathcal{A}_1}^\intercal \mathbf{M}(V) f_{\mathcal{A}_1})^{1/2-c_0}$, we establish that, conditioning on $\mathcal{O}$, with probability lager than $1 - \exp\left(-c\tau_n^{1/2-c_0}\right)$ for some positive constants $c > 0$ and $c_0 \in (0, 1/2)$,

$$|\widehat{\beta}_{\mathrm{init}}(V) - \beta| \lesssim \frac{K^2 \mathrm{Tr}[\mathbf{M}(V)]}{f_{\mathcal{A}_1}^\intercal \mathbf{M}(V) f_{\mathcal{A}_1}} + \frac{K + K^2}{(f_{\mathcal{A}_1}^\intercal \mathbf{M}(V) f_{\mathcal{A}_1})^{c_0}} + \frac{\|R_{\mathcal{A}_1}(V)\|_2}{\sqrt{f_{\mathcal{A}_1}^\intercal \mathbf{M}(V) f_{\mathcal{A}_1}}}. \tag{54}$$

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

B.3 Proof of Theorem 1

In the following, we shall prove Theorem 6, which implies Theorem 1 together with Condition (R2-I).

Theorem 6 *Suppose that the same conditions of Proposition 2 and* (30) *hold. Then* $\widehat{\beta}(V)$ *defined in* (18) *satisfies*

$$\frac{1}{\mathrm{SE}(V)} \left(\widehat{\beta}(V) - \beta \right) = \mathcal{G}(V) + \mathcal{E}(V) \quad \textit{with} \quad \mathrm{SE}(V) = \frac{\sqrt{\sum_{i=1}^{n_1} \sigma_i^2 [\mathbf{M}(V) f_{\mathcal{A}_1}]_i^2}}{f_{\mathcal{A}_1}^\intercal \mathbf{M}(V) f_{\mathcal{A}_1}}, \tag{55}$$

where $\mathcal{G}(V) \xrightarrow{d} N(0,1)$ *and there exist positive constants* $t_0 > 0$ *and* $C > 0$ *such that*

$$\liminf_{n \to \infty} \mathbb{P}\left(|\mathcal{E}(V)| \leq \frac{\sqrt{\log n} \cdot \eta_n(V) \cdot \mathrm{Tr}[\mathbf{M}(V)] + t_0 \sqrt{\mathrm{Tr}([\mathbf{M}(V)]^2)} + \|R(V)\|_2}{\sqrt{f_{\mathcal{A}_1} [\mathbf{M}(V)]^2 f_{\mathcal{A}_1}}} \right) \geq 1 - \exp(-t_0^2), \tag{56}$$

with $\eta_n(V)$ *defined in* (29).

We start with the following error decomposition,

$$\widehat{\beta}_{\mathrm{RF}}(V) - \beta = \frac{\epsilon_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)f_{\mathcal{A}_1} + R_{\mathcal{A}_1}(V)^{\intercal}\mathbf{M}(V)D_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)D_{\mathcal{A}_1}} + \frac{\mathrm{Err}_1 + \mathrm{Err}_2}{D_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)D_{\mathcal{A}_1}}, \tag{57}$$

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

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

Then the decomposition (57) implies (55). We apply (52) together with the generalized IV strength condition (R2) and establish $\frac{D_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)D_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)f_{\mathcal{A}_1}} \xrightarrow{p} 1$. Condition (30) implies the Linderberg condition. Hence, we establish $\mathcal{G}(V) \mid \mathcal{O} \xrightarrow{d} N(0,1)$, and $\mathcal{G}(V) \xrightarrow{d} N(0,1)$.

Since $\mathbf{E}(\mathrm{Err}_1 \mid \mathcal{O}) = 0$, we apply the similar argument as in (50) and obtain that, conditioning on $\mathcal{O}$, with probability larger than $1 - \exp\left(-ct_0^2\right)$ for some constants $c > 0$ and $t_0 > 0$, $|\mathrm{Err}_1| \le t_0K^2\sqrt{\sum_{1\le i\neq j\le n_1}[\mathbf{M}(V)]_{ij}^2} \le t_0K^2\sqrt{\mathrm{Tr}([\mathbf{M}(V)]^2)}$. Then we establish

$$\mathbb{P}\left(|\mathrm{Err}_1| \ge t_0K^2\sqrt{\mathrm{Tr}([\mathbf{M}(V)]^2)}\right) = \mathbf{E}\left[\mathbb{P}\left(|\mathrm{Err}_1| \ge t_0K^2\sqrt{\mathrm{Tr}([\mathbf{M}(V)]^2)} \mid \mathcal{O}\right)\right] \le \exp\left(-ct_0^2\right). \tag{58}$$

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

Lemma 5 *If Condition* (R2) *holds, then with probability larger than* $1 - n_1^{-c}$,

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

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

$$|\mathrm{Err}_2| \lesssim \sum_{i=1}^{n_1}[\mathbf{M}(V)]_{ii}\left[|f_i - \widehat{f}_i|\left(\sqrt{\log n} + |[\widehat{\epsilon}(V)]_i - \epsilon_i|\right) + \sqrt{\log n}\,|[\widehat{\epsilon}(V)]_i - \epsilon_i|\right]. \tag{59}$$

By Lemma 5, we have $|\mathrm{Err}_2| \lesssim \sqrt{\log n}\cdot\eta_n(V)\cdot\mathrm{Tr}[\mathbf{M}(V)]$ with $\eta_n(V)$ defined in (29). Together with (58), we establish $\liminf_{n\to\infty}\mathbb{P}\left(|\mathcal{E}(V)| \le CU_n(V)\right) \ge 1 - \exp(-t_0^2)$.

B.4 Proofs of Theorems 2 and 5

By conditional sub-gaussianity and $\mathrm{Var}(\delta_i \mid Z_i, X_i) \ge c$, we have $c \le \mathrm{Var}(\delta_i \mid Z_i, X_i) \le C$. Hence, we have $\mu(V) \asymp f_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)f_{\mathcal{A}_1}$. With $H(V_q, V_{q'})$ defined in (33), we have

$$\frac{\widehat{\beta}(V_q) - \widehat{\beta}(V_{q'})}{\sqrt{H(V_q, V_{q'})}} = \mathcal{G}_n(V_q, V_{q'}) + \mathcal{L}_n(V_q, V_{q'}) + \frac{1}{\sqrt{H(V_q, V_{q'})}}\left(\widetilde{\mathcal{E}}(V_q) - \widetilde{\mathcal{E}}(V_{q'})\right), \tag{60}$$

where $\mathcal{L}_n(V_q, V_{q'})$ is defined in (34),

$$\mathcal{G}_n(V_q, V_{q'}) = \frac{1}{\sqrt{H(V_q, V_{q'})}} \left(\frac{f_{\mathcal{A}_1}^\intercal \mathbf{M}(V_q) \epsilon_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\intercal \mathbf{M}(V_q) D_{\mathcal{A}_1}} - \frac{f_{\mathcal{A}_1}^\intercal \mathbf{M}(V_{q'}) \epsilon_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\intercal \mathbf{M}(V_{q'}) D_{\mathcal{A}_1}} \right), \tag{61}$$

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

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

Lemma 6 *Suppose that Conditions* (R1), (R2), *and* (R3) *hold, then*

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

By applying (52), we establish that, conditioning on $\mathcal{O}$, with probability larger than $1 - \exp\left(-c\tau_n^{1/2-c_0}\right)$, $\left|\mathcal{L}_n(V_q, V_{q'})\right| \lesssim \frac{1}{\sqrt{H(V_q, V_{q'})}} \left(\frac{\|[R(V_q)]_{\mathcal{A}_1}\|_2}{\sqrt{\mu(V_q)}} + \frac{\|[R(V_{q'})]_{\mathcal{A}_1}\|_2}{\sqrt{\mu(V_{q'})}} \right)$. Then we apply the above inequality and the condition $\sqrt{H(V_q, V_{q'})} \gg \max_{V \in \{V_q, V_{q'}\}} \|R(V)\|_2 / \sqrt{\mu(V)}$ and establish that $\left|\mathcal{L}_n(V_q, V_{q'})\right| \xrightarrow{p} 0$. Together with the decomposition (60), and Lemma 6, we establish the asymptotic limiting distribution of $\widehat{\beta}(V_q) - \widehat{\beta}(V_{q'})$ in Theorem 2.

With the decomposition (60), we establish (45) in Theorem 5 by the following Lemma 6 and applying the condition $\mathcal{L}_n(V_q, V_{q'}) \xrightarrow{p} \mathcal{L}^*(V_q, V_{q'})$ and $\widehat{H}(V_q, V_{q'}) \xrightarrow{p} H(V_q, V_{q'})$.

B.5 Proof of Theorem 3

In the following, we shall prove

$$\liminf_{n \to \infty} \mathbb{P}\left[\beta \in \mathrm{CI}(V_{\widehat{q}})\right] \geq 1 - \alpha - 2\alpha_0 \quad \text{with} \quad \widehat{q} = \widehat{q}_c \text{ or } \widehat{q}_r, \tag{62}$$

Recall that the test $\mathcal{C}(V_q)$ is defined in (26) and note that

$$\{\widehat{q}_c = q^*\} = \{Q_{\max} \geq q^*\} \cap \left(\cap_{q=0}^{q^*-1} \{\mathcal{C}(V_q) = 1\} \right) \cap \{\mathcal{C}(V_{q^*}) = 0\}. \tag{63}$$

Define the events

$$\mathcal{B}_1 = \left\{ \max_{0 \leq q < q' \leq Q_{\max}} \left| \mathcal{G}_n(V_q, V_{q'}) \right| \leq \rho(\alpha_0) \right\} \quad \text{and} \quad \mathcal{B}_2 = \{|\widehat{\rho}/\rho(\alpha_0) - 1| \leq \tau_0\},$$

$$\mathcal{B}_3 = \left\{ \max_{0 \leq q < q' \leq Q_{\max}} \frac{1}{\sqrt{H(V_q, V_{q'})}} \left| \widetilde{\mathcal{E}}(V_q) - \widetilde{\mathcal{E}}(V_{q'}) \right| \leq \tau_1 \rho(\alpha_0) \right\}.$$

where $\rho(\alpha_0)$ is defined in (35), $\widehat{\rho}$ is defined in (38) with $\mathbf{M}_{\mathrm{RF}}$ replaced by $\mathbf{M}$, and $\mathcal{G}_n(V_q, V_{q'})$ is defined in (61). By the definition of $\rho(\alpha_0)$ in (35) and the following (81), we control the probability of the event $\lim_{n \to \infty} \mathbf{P}(\mathcal{B}_1) = 1 - 2\alpha_0$. By the following (6) and $\widehat{\rho}/\rho(\alpha_0) \xrightarrow{p} 1$,

we establish that, for any positive constants $\tau_0 > 0$ and $\tau_1 > 0$, $\liminf_{n\to\infty} \mathbf{P}(\mathcal{B}_2 \cap \mathcal{B}_3) = 1$. Combing the above two equalities, we have

$$\liminf_{n\to\infty} \mathbf{P}(\mathcal{B}_1 \cap \mathcal{B}_2 \cap \mathcal{B}_3) \geq 1 - 2\alpha_0. \tag{64}$$

For any $0 \leq q \leq q^* - 1$, the condition (R4) implies that there exists some $q+1 \leq q' \leq q^*$ such that $\left|\mathcal{L}_n(V_q, V_{q'})\right| \geq A\rho(\alpha_0)$. On the event $\mathcal{B}_1 \cap \mathcal{B}_2 \cap \mathcal{B}_3$, we apply the expression (60) and obtain

$$\left|[\widehat{\beta}(V_q) - \widehat{\beta}(V_{q'})]/\sqrt{H(V_q, V_{q'})}\right| \geq A\rho(\alpha_0) - \rho(\alpha_0) - \tau_1\rho(\alpha_0) \geq (A - 1 - \tau_1)(1 - \tau_0)\widehat{\rho}. \tag{65}$$

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 (60) 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 (65), the event $\mathcal{B}_1 \cap \mathcal{B}_2 \cap \mathcal{B}_3$ implies $\cap_{q=0}^{q^*-1} \{\mathcal{C}(V_q) = 1\} \cap \{\mathcal{C}(V_{q^*}) = 0\}$. By (64), we establish $\liminf_{n\to\infty} \mathbf{P}\left[\left(\cap_{q=0}^{q^*-1} \{\mathcal{C}(V_q) = 1\}\right) \cap \{\mathcal{C}(V_{q^*}) = 0\}\right] \geq 1 - 2\alpha_0$. Together with (63), we have $\liminf_{n\to\infty} \mathbf{P}(\widehat{q}_c \neq q^*) \leq 2\alpha_0$. To control the coverage probability, we decompose $\mathbf{P}\left(\frac{1}{\mathrm{SE}(V_{\widehat{q}_c})}\left|\widehat{\beta}(V_{\widehat{q}_c}) - \beta\right| \geq z_{\alpha/2}\right)$ as

$$\begin{aligned} &\mathbf{P}\left(\left\{\frac{1}{\mathrm{SE}(V_{\widehat{q}_c})}\left|\widehat{\beta}(V_{\widehat{q}_c}) - \beta\right| \geq z_{\alpha/2}\right\} \cap \{\widehat{q}_c = q^*\}\right) \\ &+ \mathbf{P}\left(\left\{\frac{1}{\mathrm{SE}(V_{\widehat{q}_c})}\left|\widehat{\beta}(V_{\widehat{q}_c}) - \beta\right| \geq z_{\alpha/2}\right\} \cap \{\widehat{q}_c \neq q^*\}\right). \end{aligned} \tag{66}$$

Note that

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

and $\mathbf{P}\left(\left\{\frac{1}{\mathrm{SE}(V_{\widehat{q}_c})}\left|\widehat{\beta}(V_{\widehat{q}_c}) - \beta\right| \geq z_{\alpha/2}\right\} \cap \{\widehat{q}_c \neq q^*\}\right) \leq \mathbf{P}(\widehat{q}_c \neq q^*) \leq 2\alpha_0$. By the decomposition (66), we combine the above two inequalities and establish

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

which implies (62) 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 (62) with $\widehat{q} = \widehat{q}_r$.

B.6 Proof of Theorem 4

If $\mathrm{Cov}(\epsilon_i, \delta_i \mid Z_i, X_i) = \mathrm{Cov}(\epsilon_i, \delta_i)$, then (50) further implies

$$\left|\epsilon_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)\delta_{\mathcal{A}_1} - \mathrm{Cov}(\epsilon_i, \delta_i) \cdot \mathrm{Tr}[\mathbf{M}_{\mathrm{RF}}(V)]\right| \leq t_0 K^2 \sqrt{\mathrm{Tr}([\mathbf{M}_{\mathrm{RF}}(V)]^2)}. \tag{67}$$

Proof of (42)**.** We decompose the error $\widehat{\mathrm{Cov}}(\delta_i, \epsilon_i) - \mathrm{Cov}(\delta_i, \epsilon_i)$ as,

$$\begin{aligned} &\frac{1}{n_1 - r}(f_{\mathcal{A}_1} - \widehat{f}_{\mathcal{A}_1} + \delta_{\mathcal{A}_1})^\intercal P^\perp_{V_{\mathcal{A}_1}}[\epsilon_{\mathcal{A}_1} + D_{\mathcal{A}_1}(\beta - \widehat{\beta}_{\mathrm{init}}(V)) + R_{\mathcal{A}_1}(V)] - \mathrm{Cov}(\delta_i, \epsilon_i) \\ &= T_1 + T_2 + T_3, \end{aligned} \tag{68}$$

where $T_1 = \frac{1}{n_1 - r}\left[(\delta_{\mathcal{A}_1})^\intercal P^\perp_{V^\intercal_{\mathcal{A}_1}}\, \epsilon_{\mathcal{A}_1} - (n_1 - r)\mathrm{Cov}(\delta_i, \epsilon_i)\right]$, and

$$\begin{aligned} T_2 &= \frac{1}{n_1 - r}(\delta_{\mathcal{A}_1})^\intercal P^\perp_{V^\intercal_{\mathcal{A}_1}}\left[D_{\mathcal{A}_1}(\beta - \widehat{\beta}_{\mathrm{init}}(V)) + R_{\mathcal{A}_1}(V)\right], \\ T_3 &= \frac{1}{n_1 - r}(f_{\mathcal{A}_1} - \widehat{f}_{\mathcal{A}_1})^\intercal P^\perp_{V^\intercal_{\mathcal{A}_1}}[\epsilon_{\mathcal{A}_1} + D_{\mathcal{A}_1}(\beta - \widehat{\beta}_{\mathrm{init}}(V)) + R_{\mathcal{A}_1}(V)]. \end{aligned}$$

We apply (46) with $A = P^\perp_{V^\intercal_{\mathcal{A}_1}}$ and $t = t_0 K^2\sqrt{n_1 - r}$ for some $0 < t_0 \le \sqrt{n_1 - r}$, and establish

$$\mathbb{P}\left(\left|\epsilon^\intercal_{\mathcal{A}_1} P^\perp_{V^\intercal_{\mathcal{A}_1}}\, \delta_{\mathcal{A}_1} - \mathrm{Cov}(\epsilon_i, \delta_i) \cdot \mathrm{Tr}\left(P^\perp_{V^\intercal_{\mathcal{A}_1}}\right)\right| \ge t_0 K^2\sqrt{n_1 - r} \mid \mathcal{O}\right) \le 6\exp\left(-ct_0^2\right).$$

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

$$|T_1| \lesssim \sqrt{\log(n_1 - r)/(n_1 - r)}. \tag{69}$$

We apply (46) with $A = P^\perp_{V^\intercal_{\mathcal{A}_1}}$ and $t = t_0 K^2\sqrt{n_1 - r}$ for some $0 < t_0 \le \sqrt{n_1 - r}$,

$$\mathbb{P}\left(\left|\epsilon^\intercal_{\mathcal{A}_1} P^\perp_{V^\intercal_{\mathcal{A}_1}}\, D_{\mathcal{A}_1} - \mathrm{Cov}(\epsilon_i, D_i) \cdot \mathrm{Tr}\left(P^\perp_{V^\intercal_{\mathcal{A}_1}}\right)\right| \ge t_0 K^2\sqrt{n_1 - r} \mid \mathcal{O}\right) \le 6\exp\left(-ct_0^2\right).$$

The above concentration bound implies

$$\mathbb{P}\left(\frac{1}{n_1 - r}\left|\epsilon^\intercal_{\mathcal{A}_1} P^\perp_{V^\intercal_{\mathcal{A}_1}}\, D_{\mathcal{A}_1}\right| \ge C\left(1 + \sqrt{\frac{\log(n_1 - r)}{n_1 - r}}\right) \mid \mathcal{O}\right) \le 6(n_1 - r)^{-c}. \tag{70}$$

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

$$|T_2| \lesssim \left|\beta - \widehat{\beta}_{\mathrm{init}}(V)\right| + \|R(V)\|_2/\sqrt{n_1 - r}, \tag{71}$$

where the last inequality follows from (54). Regarding T_3, we have

$$\begin{aligned} |T_3| &= \left|\frac{1}{n_1 - r}(f_{\mathcal{A}_1} - \widehat{f}_{\mathcal{A}_1})^\intercal P^\perp_{V^\intercal_{\mathcal{A}_1}}[\epsilon_{\mathcal{A}_1} + D_{\mathcal{A}_1}(\beta - \widehat{\beta}_{\mathrm{init}}(V)) + R_{\mathcal{A}_1}(V)]\right| \\ &\lesssim \frac{\|P^\perp_{V^\intercal_{\mathcal{A}_1}}(f_{\mathcal{A}_1} - \widehat{f}_{\mathcal{A}_1})\|_2}{\sqrt{n_1 - r}} \frac{\|P^\perp_{V^\intercal_{\mathcal{A}_1}}\epsilon_{\mathcal{A}_1}\|_2 + \|P^\perp_{V^\intercal_{\mathcal{A}_1}} D_{\mathcal{A}_1}\|_2 \cdot |\beta - \widehat{\beta}_{\mathrm{init}}(V)| + \|R_{\mathcal{A}_1}(V)\|_2}{\sqrt{n_1 - r}}. \end{aligned} \tag{72}$$

By a similar argument as in (70), we establish that, with probability larger than $1 - (n_1 - r)^{-c}$, $\frac{1}{\sqrt{n_1}}\|P^\perp_{V^\intercal_{\mathcal{A}_1}}\epsilon_{\mathcal{A}_1}\|_2 + \frac{1}{\sqrt{n_1}}\|P^\perp_{V^\intercal_{\mathcal{A}_1}} D_{\mathcal{A}_1}\|_2 \le C$, for some positive constant $C > 0$. Together with (72), we establish that, with probability larger than $1 - (n_1 - r)^{-c}$,

$$|T_3| \lesssim \frac{\|P^\perp_{V^\intercal_{\mathcal{A}_1}}(f_{\mathcal{A}_1} - \widehat{f}_{\mathcal{A}_1})\|_2}{\sqrt{n_1 - r}} \cdot \left(1 + |\beta - \widehat{\beta}_{\mathrm{init}}(V)| + \frac{\|R(V)\|_2}{\sqrt{n_1 - r}}\right).$$

Together with (68) and the upper bounds (69) and (71), we establish (42).

Proof of (44)**.** By (49), we obtain the following decomposition for $\widetilde{\beta}_{\mathrm{RF}}(V)$ in (18),

$$\widehat{\beta}_{\mathrm{RF}}(V) - \beta = \frac{\epsilon_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)f_{\mathcal{A}_1} - [\widehat{\mathrm{Cov}}(\delta_i,\epsilon_i) - \mathrm{Cov}(\delta_i,\epsilon_i)]\mathrm{Tr}[\mathbf{M}(V)]}{D_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)D_{\mathcal{A}_1}} + \frac{\epsilon_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)\delta_{\mathcal{A}_1} - \mathrm{Cov}(\delta_i,\epsilon_i)\mathrm{Tr}[\mathbf{M}(V)]}{D_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)D_{\mathcal{A}_1}} + \frac{R_{\mathcal{A}_1}(V)^{\intercal}\mathbf{M}(V)D_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)D_{\mathcal{A}_1}}.$$

The above decomposition implies (55) with $\mathcal{G}(V) = \frac{1}{\mathrm{SE}(V)}\frac{\epsilon_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)f_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)D_{\mathcal{A}_1}}$, and

$$\mathrm{SE}(V)\cdot\mathcal{E}(V) = \frac{R_{\mathcal{A}_1}(V)^{\intercal}\mathbf{M}(V)D_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)D_{\mathcal{A}_1}} + \frac{[\mathrm{Cov}(\delta_i,\epsilon_i) - \widehat{\mathrm{Cov}}(\delta_i,\epsilon_i)]\mathrm{Tr}[\mathbf{M}(V)] + \epsilon_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)\delta_{\mathcal{A}_1} - \mathrm{Cov}(\delta_i,\epsilon_i)\mathrm{Tr}[\mathbf{M}(V)]}{D_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)D_{\mathcal{A}_1}}.$$

We establish $\mathcal{G}(V) \xrightarrow{d} N(0,1)$ by applying the same arguments as in Section B.3 in the main paper. We establish (44) by combining (42), (52), and (67).

Appendix C. Proof of Extra Lemmas

C.1 Proof of Lemma 4

Note that $\mathbf{E}\left[\epsilon_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)\delta_{\mathcal{A}_1} \mid \mathcal{O}\right] = \mathrm{Tr}\left(\mathbf{M}(V)\Lambda\right)$. We apply (46) with $A = \mathbf{M}(V)$ and $t = t_0K^2\|\mathbf{M}(V)\|_F$ for some $t_0 > 0$ and establish

$$\begin{aligned}&\mathbb{P}\left(\left|\epsilon_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)\delta_{\mathcal{A}_1} - \mathrm{Tr}\left(\mathbf{M}(V)\Lambda\right)\right| \geq t_0K^2\|\mathbf{M}(V)\|_F \mid \mathcal{O}\right)\\ &\leq 6\exp\left(-c\min\left\{t_0^2, t_0\frac{\|\mathbf{M}(V)\|_F}{\|\mathbf{M}(V)\|_2}\right\}\right) \leq 6\exp\left(-c\min\left\{t_0^2,t_0\right\}\right),\end{aligned} \tag{73}$$

where the last inequality follows from $\|\mathbf{M}(V)\|_F \geq \|\mathbf{M}(V)\|_2$. The above concentration bound implies (50) by taking an expectation with respect to $\mathcal{O}$.

Since $\mathbf{E}\left[\delta_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)\delta_{\mathcal{A}_1} \mid \mathcal{O}\right] = \mathrm{Tr}\left(\mathbf{M}(V)\Sigma^{\delta}\right)$, we apply a similar argument to (73) and establish

$$\mathbb{P}\left(\left|\delta_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)\delta_{\mathcal{A}_1} - \mathrm{Tr}\left(\mathbf{M}(V)\Sigma^{\delta}\right)\right| \geq t_0K^2\|\mathbf{M}(V)\|_F \mid \mathcal{O}\right) \leq 2\exp\left(-c\min\left\{t_0^2,t_0\right\}\right). \tag{74}$$

Note that $\mathbf{E}\left[\delta_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)f_{\mathcal{A}_1} \mid \mathcal{O}\right] = 0$, and conditioning on $\mathcal{O}$, $\{\delta_i\}_{i\in\mathcal{A}_1}$ are independent sub-gaussian random variables. We apply Proposition 5.16 of Vershynin (2010) and establish that, with probability larger than $1-\exp(-t_0^2)$ for some positive constant $c > 0$,

$$\mathbb{P}\left(\delta_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)f_{\mathcal{A}_1} \geq Ct_0K\sqrt{f_{\mathcal{A}_1}^{\intercal}[\mathbf{M}(V)]^2f_{\mathcal{A}_1}} \mid \mathcal{O}\right) \leq \exp(-ct_0^2). \tag{75}$$

The above concentration bound implies (50) by taking an expectation with respect to $\mathcal{O}$.

By the decomposition $D_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)D_{\mathcal{A}_1} - f_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)f_{\mathcal{A}_1} = \delta_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)f_{\mathcal{A}_1} + \delta_{\mathcal{A}_1}^{\intercal}\mathbf{M}(V)\delta_{\mathcal{A}_1}$, we establish (51) by applying the concentration bounds (74) and (75).

C.2 Proof of Lemma 5

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

$$Y_{\mathcal{A}_1} - \widehat{\beta}_{\text{init}}(V) D_{\mathcal{A}_1} = V_{\mathcal{A}_1}\pi + R_{\mathcal{A}_1} + D_{\mathcal{A}_1}(\beta - \widehat{\beta}_{\text{init}}(V)) + \epsilon_{\mathcal{A}_1}. \tag{76}$$

The least square estimator of π is expressed as, $\widehat{\pi} = (V^\intercal V)^{-1} V^\intercal \left(Y_{\mathcal{A}_1} - \widehat{\beta}_{\text{init}}(V) D_{\mathcal{A}_1}\right)$. Note that

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

Note that $\mathbf{E} V_i \epsilon_i = 0$ and conditioning on $\mathcal{O}$, $\{\epsilon_i\}_{i\in\mathcal{A}_1}$ are independent sub-gaussian random variables. By Proposition 5.16 of Vershynin (2010), there exist positive constants $C > 0$ and $c > 0$ such that $\mathbb{P}\left(\frac{1}{n_1}\left|\sum_{i=1}^{n_1} V_{ij}\epsilon_i\right| \geq C t_0 \sqrt{\sum_{i=1}^{n_1} V_{ij}^2}/n_1 \mid \mathcal{O}\right) \leq \exp(-ct_0^2)$. By Condition (R1), we have $\max_{i,j} V_{ij}^2 \leq C \log n$ and apply the union bound and establish

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n_1} V_i \epsilon_i / n_1\right\|_\infty \geq C \log n_1 / \sqrt{n_1}\right) \leq n_1^{-c}. \tag{77}$$

Similarly to (77), we establish $\mathbb{P}\left(\left\|\frac{1}{n_1}\sum_{i=1}^{n_1} V_i\delta_i\right\|_\infty \geq C\frac{\log n_1}{\sqrt{n_1}}\right) \leq n_1^{-c}$. Together with the expression $\frac{1}{n_1}\sum_{i=1}^{n_1} V_i D_i = \frac{1}{n_1}\sum_{i=1}^{n_1} V_i f_i + \frac{1}{n_1}\sum_{i=1}^{n_1} V_i\delta_i$, we apply Condition (R1) and establish $\mathbb{P}\left(\left|\frac{1}{n_1}\sum_{i=1}^{n_1} V_i D_i\right| \geq C\right) \leq n_1^{-c}$. Together with (77), and Condition (R1), we establish that, with probability larger than $1 - n_1^{-c}$,

$$\|\widehat{\pi} - \pi\|_2 \lesssim \|R(V)\|_\infty + \left|\beta - \widehat{\beta}_{\text{init}}(V)\right| + \log n/\sqrt{n}. \tag{78}$$

Our proposed estimator $\widehat{\epsilon}(V)$ defined in (20) has the following equivalent expression,

$$\begin{aligned}
\widehat{\epsilon}(V) &= P^\perp_{V_{\mathcal{A}_1}^\intercal}\left[Y_{\mathcal{A}_1} - D_{\mathcal{A}_1}\widehat{\beta}_{\text{init}}(V)\right] \\
&= Y_{\mathcal{A}_1} - D_{\mathcal{A}_1}\widehat{\beta}_{\text{init}}(V) - V(V^\intercal V)^{-1} V^\intercal \left(Y_{\mathcal{A}_1} - \widehat{\beta}_{\text{init}}(V) D_{\mathcal{A}_1}\right) \\
&= Y_{\mathcal{A}_1} - D_{\mathcal{A}_1}\widehat{\beta}_{\text{init}}(V) - V_{\mathcal{A}_1}\widehat{\pi}.
\end{aligned}$$

Then we apply (76) and obtain $[\widehat{\epsilon}(V)]_i - \epsilon_i = D_i(\beta - \widehat{\beta}_{\text{init}}(V)) + V_i^\intercal(\pi - \widehat{\pi}) + R_i(V)$. Together with Condition (R1) and (78), we establish Lemma 5.

C.3 Proof of Lemma 6

Similarly to (57), we decompose $\widetilde{\mathcal{E}}(V_q)$ as $\widetilde{\mathcal{E}}(V_q) = \frac{\text{Err}_1 + \text{Err}_2}{D_{\mathcal{A}_1}^\intercal \mathbf{M}(V_q) D_{\mathcal{A}_1}}$, where

$$\text{Err}_1 = \sum_{1\leq i\neq j\leq n_1} [\mathbf{M}(V_q)]_{ij}\delta_i\epsilon_j,$$

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

We apply the same analysis as that of (59) and establish

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

By Lemma 5 with $V = V_{Q_{\max}}$, we apply the similar argument as that of (56) and establish that

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

where $\eta_n(V)$ is defined in (29). 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,

$$\begin{aligned}
&\frac{f_{\mathcal{A}_1}^\intercal\mathbf{M}(V_{q'})\epsilon_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\intercal\mathbf{M}(V_{q'})D_{\mathcal{A}_1}} - \frac{f_{\mathcal{A}_1}^\intercal\mathbf{M}(V_q)\epsilon_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\intercal\mathbf{M}(V_q)D_{\mathcal{A}_1}} = \frac{f_{\mathcal{A}_1}^\intercal\mathbf{M}(V_{q'})\epsilon_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^\intercal\mathbf{M}(V_{q'})f_{\mathcal{A}_1}} - \frac{f_{\mathcal{A}_1}^\intercal\mathbf{M}(V_q)\epsilon_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^\intercal\mathbf{M}(V_q)f_{\mathcal{A}_1}} \\
&+ \frac{f_{\mathcal{A}_1}^\intercal\mathbf{M}(V_{q'})\epsilon_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^\intercal\mathbf{M}(V_{q'})f_{\mathcal{A}_1}}\left(\frac{f_{\mathcal{A}_1}^\intercal\mathbf{M}(V_{q'})f_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\intercal\mathbf{M}(V_{q'})D_{\mathcal{A}_1}} - 1\right) - \frac{f_{\mathcal{A}_1}^\intercal\mathbf{M}(V_q)\epsilon_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^\intercal\mathbf{M}(V_q)f_{\mathcal{A}_1}}\left(\frac{f_{\mathcal{A}_1}^\intercal\mathbf{M}(V_q)f_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\intercal\mathbf{M}(V_q)D_{\mathcal{A}_1}} - 1\right).
\end{aligned} \tag{79}$$

Since the vector S defined in (32) satisfies $\max_{i\in\mathcal{A}_1} S_i^2 / \sum_{i\in\mathcal{A}_1} S_i^2 \to 0$, we verify the Linderberg condition and establish

$$\frac{1}{\sqrt{H(V_q, V_{q'})}}\left(\frac{f_{\mathcal{A}_1}^\intercal\mathbf{M}(V_{q'})\epsilon_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^\intercal\mathbf{M}(V_{q'})f_{\mathcal{A}_1}} - \frac{f_{\mathcal{A}_1}^\intercal\mathbf{M}(V_q)\epsilon_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^\intercal\mathbf{M}(V_q)f_{\mathcal{A}_1}}\right) \xrightarrow{d} N(0,1). \tag{80}$$

We apply (50) and (51) and establish that with probability at least $1-\exp(-c\min\{t_0^2, t_0\})$ for some positive constants $c > 0$ and $t_0 > 0$,

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

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

$$\begin{aligned}
&\frac{1}{\sqrt{H(V_q, V_{q'})}}\left|\frac{f_{\mathcal{A}_1}^\intercal\mathbf{M}(V_{q'})\epsilon_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^\intercal\mathbf{M}(V_{q'})f_{\mathcal{A}_1}}\left(\frac{f_{\mathcal{A}_1}^\intercal\mathbf{M}(V_{q'})f_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\intercal\mathbf{M}(V_{q'})D_{\mathcal{A}_1}} - 1\right)\right| \\
&\lesssim \frac{1}{\sqrt{H(V_q, V_{q'})}}\frac{t_0}{\sqrt{\mu(V_q)}}\left(\frac{\mathrm{Tr}[\mathbf{M}(V_q)]}{\mu(V_q)} + \frac{t_0}{\sqrt{\mu(V_q)}}\right).
\end{aligned}$$

Under the condition (R3), we establish

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

Similarly, we have $\frac{1}{\sqrt{H(V_q, V_{q'})}} \left| \frac{f_{\mathcal{A}_1}^\intercal \mathbf{M}(V_{q'}) \epsilon_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^\intercal \mathbf{M}(V_{q'}) f_{\mathcal{A}_1}} \left(\frac{f_{\mathcal{A}_1}^\intercal \mathbf{M}(V_{q'}) f_{\mathcal{A}_1}}{D_{\mathcal{A}_1}^\intercal \mathbf{M}(V_{q'}) D_{\mathcal{A}_1}} - 1 \right) \right| \xrightarrow{p} 0$. The above inequalities and the decomposition (79) imply

$$\left| \mathcal{G}_n(V_q, V_{q'}) - \frac{1}{\sqrt{H(V_q, V_{q'})}} \left(\frac{f_{\mathcal{A}_1}^\intercal \mathbf{M}(V_{q'}) \epsilon_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^\intercal \mathbf{M}(V_{q'}) f_{\mathcal{A}_1}} - \frac{f_{\mathcal{A}_1}^\intercal \mathbf{M}(V_q) \epsilon_{\mathcal{A}_1}}{f_{\mathcal{A}_1}^\intercal \mathbf{M}(V_q) f_{\mathcal{A}_1}} \right) \right| \xrightarrow{p} 0. \tag{81}$$

Together with (80), we establish $\mathcal{G}_n(V_q, V_{q'}) \xrightarrow{d} N(0, 1)$.

C.4 Proof of Lemma 1

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

$$b^\intercal [\mathbf{M}_{\mathrm{RF}}(V)]^2 b = (P_{\widehat{V}_{\mathcal{A}_1}}^\perp \Omega b)^\intercal \Omega (\Omega)^\intercal P_{\widehat{V}_{\mathcal{A}_1}}^\perp \Omega b \leq \|P_{\widehat{V}_{\mathcal{A}_1}}^\perp \Omega b\|_2^2 \|\Omega\|_2^2. \tag{82}$$

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 (82), we establish (40). Note that $b^\intercal \mathbf{M}_{\mathrm{RF}}(V) b = b^\intercal (\Omega)^\intercal P_{\widehat{V}_{\mathcal{A}_1}}^\perp \Omega b \leq \|\Omega\|_2^2 \|b\|_2^2$, we establish that $\lambda_{\max}(\mathbf{M}_{\mathrm{RF}}(V)) \leq 1$.

We apply the minimax expression of eigenvalues and obtain that

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

where the inequality follows from (40). The above inequality leads to $\mathrm{Tr}\left([\mathbf{M}_{\mathrm{RF}}(V)]^2\right) \leq \mathrm{Tr}[\mathbf{M}(V)]$. Since $P_{BW}^\perp P_{V_{BW},W}^\perp = P_{BW}^\perp$, we have

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

Similar to the above proof, we establish $[\mathbf{M}_{\mathrm{DNN}}(V)]^2 = \mathbf{M}_{\mathrm{DNN}}(V)$. The proof of (41) is the same as that of (40) by replacing (82) with $b^\intercal [\mathbf{M}_{\mathrm{RF}}(V)]^2 b \leq b^\intercal \mathbf{M}_{\mathrm{RF}}(V) b \cdot \|\Omega\|_2^2$.

C.5 Proof of Lemma 2

By rewriting Lemma 5, we have that, with probability larger than $1 - n^{-c}$,

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

It is sufficient to show

$$\frac{(f_{\mathcal{A}_1}^\intercal \mathbf{M}(V) f_{\mathcal{A}_1})^2}{\sum_{i=1}^{n_1} \sigma_i^2 [\mathbf{M}(V) f_{\mathcal{A}_1}]_i^2} \frac{\sum_{i=1}^{n_1} [\widehat{\epsilon}(V)]_i^2 [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2}{(D_{\mathcal{A}_1}^\intercal \mathbf{M}(V) D_{\mathcal{A}_1})^2} \xrightarrow{p} 1. \tag{83}$$

Note that

$$\frac{\sum_{i=1}^{n_1} [\widehat{\epsilon}(V)]_i^2 [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2}{\sum_{i=1}^{n_1} \sigma_i^2 [\mathbf{M}(V) f_{\mathcal{A}_1}]_i^2} = \frac{\sum_{i=1}^{n_1} \sigma_i^2 [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2}{\sum_{i=1}^{n_1} \sigma_i^2 [\mathbf{M}(V) f_{\mathcal{A}_1}]_i^2} \cdot \frac{\sum_{i=1}^{n_1} [\widehat{\epsilon}(V)]_i^2 [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2}{\sum_{i=1}^{n_1} \sigma_i^2 [\mathbf{M}(V) D_{\mathcal{A}_1}]_i^2}. \tag{84}$$

We further decompose

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

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 Vershynin (2010), we establish that,

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

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

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

By (84) 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 $\frac{(f_{\mathcal{A}_1}^\intercal \mathbf{M}(V) f_{\mathcal{A}_1})^2}{(D_{\mathcal{A}_1}^\intercal \mathbf{M}(V) D_{\mathcal{A}_1})^2} \xrightarrow{p} 1$, we establish (83).

Appendix D. Additional Simulation Results

D.1 The Appropriate Threshold of IV Strength for Reliable Inference

In this section, we show the performance of `TSCI` in terms of RMSE, bias, and CI coverage with different IV strengths. We generate Z_i, X_i and errors $\{(\delta_i, \epsilon_i)^\intercal\}_{1 \leq i \leq n}$ following the same procedure as in Settings B1 and B2 in Section 6.2 but with $Z_i = \Psi(X_{i,p_x+1}^*) \in (0,1)$. We generate the treatment and outcome $\{D_i, Y_i\}_{1 \leq i \leq n}$ following $D_i = 1/2 \cdot Z_i + a \cdot (\sin(2\pi Z_i) + 3/2 \cdot \cos(2\pi Z_i)) - 3/10 \cdot \sum_{j=1}^{10} X_{i,j} + \delta_i$ and $Y_i = 1/2 \cdot D_i + 1/5 \cdot \sum_{j=1}^{10} X_{i,j} + \epsilon_i$.

With fixing the sample size $n = 3000$, we control the IV strength by varying a across $\{0.15, 0.17, \cdots, 0.35\}$. Since we consider the valid IV setting, we implement `TSCI` with random forests for 500 times and specify $\mathcal{V}_0 = \{\overrightarrow{\mathbf{w}}(x)\}$ with $\overrightarrow{\mathbf{w}}(x) = \{1, x_1, \cdots, x_{p_x}\}$. In Figure 6, we show that when IV strength is above 40, `TSCI` has a smaller RMSE and bias, and its confidence interval achieves the desired coverage.

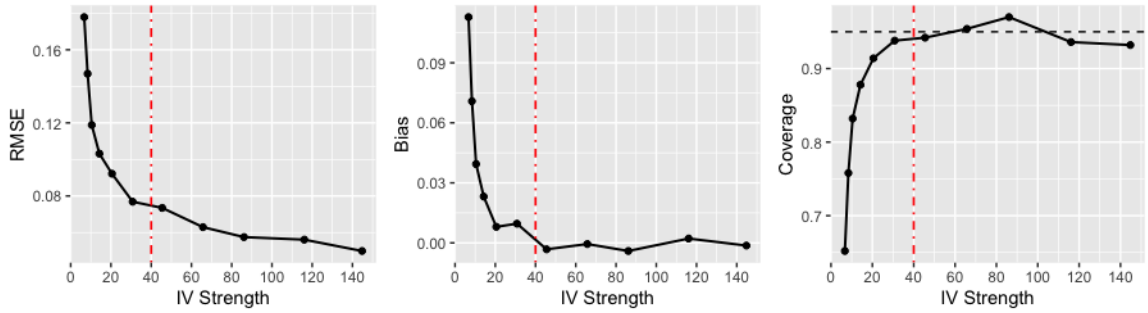

Figure 6: RMSE, bias and CI coverage of `TSCI` with different IV strengths. The red dashed line indicates the IV strength as 40. The black dashed line indicates the 95% coverage level.

We also use the setting with $a \in \{0.25, 0.3, 0.33\}$ to illustrate the bias of $\widehat{\beta}_{\text{init}}(V)$ and the effect of the proposed bias correction in Figure 1 in the main paper.

D.2 Comparison with Machine-Learning IV

In the following, we compare our `TSCI` estimator with the `MLIV` estimator described in (28) while setting the full set of observed covariates as adjusted forms in our method. We use the exactly same setting in Section D.1. In Table 7, we compare `TSCI` with `MLIV` for different $n \in \{1000, 3000, 5000\}$ and different $a \in \{0.2, 0.25, 0.3, 0.35\}$ with a larger value of a corresponding to a stronger IV. We implement 500 rounds of simulations and report the average measure.

In Table 7, we observe that the `MLIV` has a larger bias and standard error than `TSCI`, leading to an inflated MSE of `MLIV`. We have explained in Section 4.6 that this happens when the IV strength captured by RF is relatively weak. For a stronger IV (with $a = 0.35$), when the sample size is 3000 or 5000 and the self-prediction is excluded, the `MLIV` and `TSCI` have comparable MSE, but our proposed `TSCI` estimator generally exhibits a smaller bias. Moreover, by comparing `TSCI` and initial estimators, our proposed bias correction step effectively removes the bias due to the high complexity of RF algorithm. Another interesting observation is that the removal of self-prediction helps reducing the bias of both `TSCI` and `MLIV` estimators for the reason of reducing the correlation between the ML predicted value and the unmeasured confounders.

We then show how the coefficient c_f in (28) in the main paper inflates the estimation error of `MLIV` when IV is weak. In Figure 7, we display the distribution of `TSCI` and `MLIV` estimators in box plots and the frequency of c_f in the histogram in the setting with $n = 1000$

		without self-prediction					with self-prediction				
		Bias			RMSE		Bias			RMSE	
a	n	TSCI	Init	MLIV	Ratio	IV Str	TSCI	Init	MLIV	Ratio	IV Str
0.20	1000	0.17	0.30	-4.31	374.06	2.18	0.28	0.38	0.48	1.51	11.90
	3000	0.03	0.14	-0.15	2.94	12.11	0.17	0.29	0.45	2.17	42.06
	5000	0.01	0.10	-0.07	1.67	27.97	0.12	0.23	0.42	2.69	76.41
0.25	1000	0.06	0.18	-0.38	23.14	3.84	0.17	0.28	0.41	1.88	15.79
	3000	0.00	0.06	-0.05	1.37	30.35	0.08	0.17	0.35	2.97	74.95
	5000	0.00	0.04	-0.02	1.11	77.97	0.04	0.12	0.32	3.84	156.57
0.30	1000	0.02	0.09	-0.11	2.01	7.49	0.09	0.18	0.32	2.22	23.92
	3000	-0.00	0.02	-0.02	1.09	74.82	0.03	0.09	0.25	3.74	151.09
	5000	0.00	0.02	-0.01	1.06	194.74	0.01	0.06	0.21	4.51	316.40
0.35	1000	0.00	0.05	-0.05	1.29	14.32	0.04	0.11	0.24	2.57	41.62
	3000	-0.01	0.01	-0.01	0.99	144.72	0.01	0.04	0.17	3.55	255.65
	5000	-0.00	0.01	-0.01	1.00	337.34	0.01	0.04	0.16	4.44	526.98

Table 7: Comparison of TSCI and MLIV with the treatment model fitted by RF. The columns indexed with "without self-prediction" stands for the split RF being implemented without self-prediction, while "with self-prediction" stands for the split RF being implemented with self-prediction. The columns indexed by "TSCI", "Init", "MLIV" correspond to our proposed TSCI estimator in (18), the initial estimator in (15), and the MLIV estimator in (28). The column indexed with "RMSE Ratio" represents the MSE ratio of the MLLV estimator to TSCI estimator; the column indexed with "IV Str" stands for our proposed IV strength in (21).

and $a = 0.2$. We can see that there are certain proportion of the coefficient c_f close to 0, which would inflate the estimator to extremely large values and thus increase the MLIV estimator variance when IV is weak.

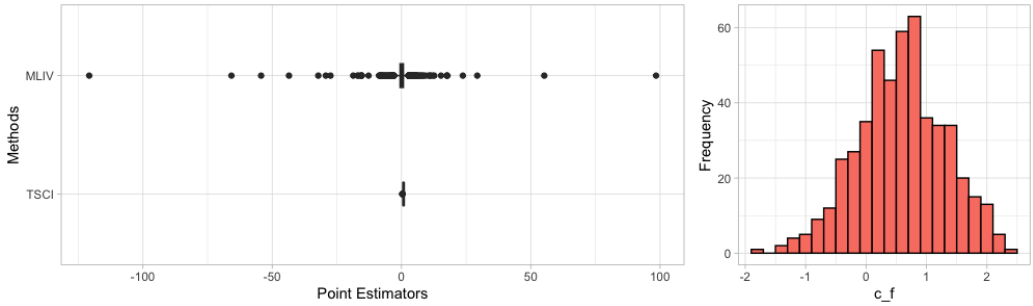

Figure 7: The left panel shows the boxplot of TSCI and MLIV estimators and the right panel shows the histogram of the coefficient c_f of $\widehat{f}$ in the first stage regression $D \sim \widehat{f} + X$ in the MLIV method.

D.3 Multiple IV (With Non-Linearity)

In this section, we consider multiple IVs and compare `TSCI` with existing methods dealing with invalid IVs, including `TSHT` (Guo et al., 2018) and `CIIV` (Windmeijer et al., 2021). We consider the setting with 10 IVs. With fixing the sample size $n = 3000$, we generate $X_i^* \in \mathbb{R}^{p_x+p_z}$ following the multivariate normal distribution with zero mean and covariance Σ where $\Sigma_{i,j} = 0.5^{|i-j|}$ for $1 \leq i, j \leq p_x + p_z$. We define the first p_z columns of X^* as IVs and denote it by $Z_i = X^*_{i,1:p_z}$. The remaining columns are defined as observed covariates, that is $X_i = X^*_{i,(p_z+1):(p_z+p_x)}$. We generate errors $\{(\delta_i, \epsilon_i)^\intercal\}_{1\leq i\leq n}$ following the bivariate normal distribution with zero mean, unit variance and covariance as 0.5. We generate $\{D_i, Y_i\}_{1\leq i\leq n}$ following $D_i = \sum_{j=1}^{p_z} |Z_{i,j}| + \sum_{j=1}^{p_z} \tanh(Z_{i,j}) + 1/2 \cdot \sum_{j=1}^{p_x} X_{i,j} + \delta_i$ and $Y_i = D_i + Z_i\pi + 1/2 \cdot \sum_{j=1}^{p_x} X_{i,j} + \epsilon_i$ where the vector $\pi \in \mathbb{R}^{p_z}$ indicates the invalidity level of each IV. The j-th IV is valid if $\pi_j = 0$; otherwise, it is invalid. A larger $|\pi_j|$ value indicates that the j-th IV is more severely invalid. We consider the following two settings,

- Setting E1: $\pi = a \cdot (0, 0, 0, 0, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5)^\intercal$,
- Setting E2: $\pi = a \cdot (0, 0, 0, 0, 0.5, 0.5, 0.5, 1, 1, 1)^\intercal$.

For Setting E1, neither the majority nor the plurality rule is satisfied. For Setting E2, the plurality rule is satisfied. We vary the value of $a \in \{0.1, 0.2, \cdots, 1\}$ to simulate the different levels of invalidity. We specify the sets of basis as $\mathcal{V}_0 = \{\overrightarrow{\mathbf{w}}(x)\}$ with $\overrightarrow{\mathbf{w}}(x) = \{1, x_1, \cdots, x_{p_x}\}$, $\mathcal{V}_1 = \{\mathbf{z}, \overrightarrow{\mathbf{w}}(x)\}$ and $\mathcal{V}_2 = \{\mathbf{z}, \mathbf{z}^2, \overrightarrow{\mathbf{w}}(x)\}$ with $\mathbf{z}^2 = \{z_1^2, \cdots, z_{10}^2\}$. We run 500 rounds of simulation and report the results in Figure 8.

In Setting E1, neither the majority rule nor the plurality rule is satisfied, so `TSHT` and `CIIV` cannot select valid IVs and do not achieve a desired coverage level of 95%. In contrast, `TSCI` is able to detect the existing invalidity and obtain valid inference. When the invalidity level is low (with $a = 0.1$), `TSCI` may not be able to identify the invalidity and thus lead to coverage below 95%.

In Setting E2, `TSCI` achieves the desired coverage level while `TSHT` and `CIIV` achieve the desired coverage for a relatively large invalidity level. When the invalidity level is low, say $a = 0.1$ and $a = 0.2$, `TSHT` and `CIIV` are significantly impacted by locally invalid IVs and generate poor coverage. In comparison to `TSHT` and `CIIV`, `TSCI` provides much more robust performance to the low invalidity levels because it evaluates the total invalidity levels accumulated by all ten invalid IVs.

D.4 Additional Results in Section 6.2

In this section, we further report the additional results for Setting B1 and results for Setting B2, and introduce a setting with binary IV denoted as Setting B3.

For Setting B1, we report the coverage for Vio=2 in Table 8 and we further report the mean absolute bias and the confidence interval length in for both violation forms in Tables 9 and 10, respectively. In Table 9, `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. Compared with RF-Init, `TSCI` corrects the bias effectively and thus have the desired 95% coverage. In Table 10, the CI of `TSCI` is shorter than the oracle method in settings with small sample size or weak interaction strength, because it fails to select the correct violation forms.

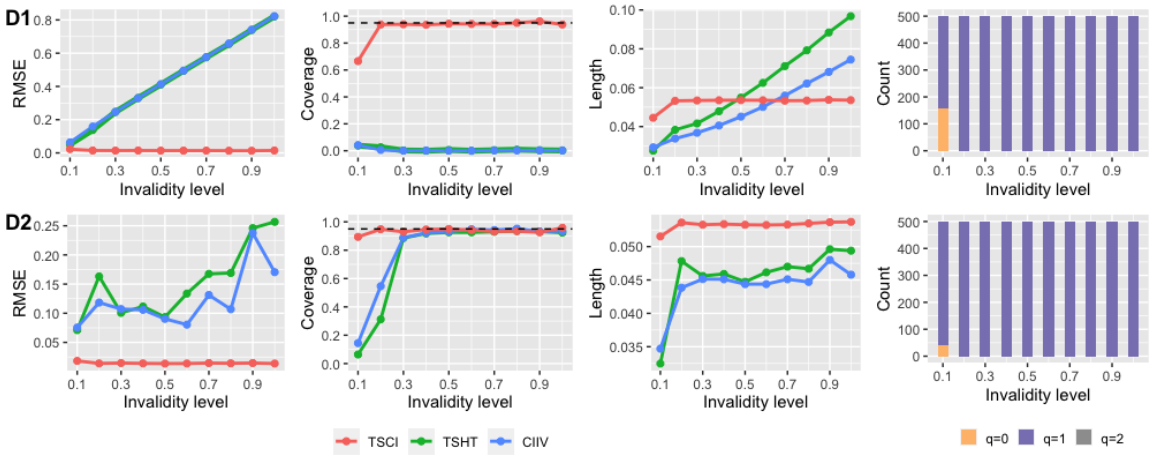

Figure 8: Comparison of `TSCI` with `TSHT` and `CIIV` in terms of RMSE, CI coverage and length under Settings E1 and E2. Setting E1 corresponds to a setting that the plurality rule fails while Setting E2 corresponds to a setting that the plurality holds. The stacked bar charts in the last column show the basis selection of `TSCI` where $\mathcal{V}_{q=1}$ is the oracle basis set. "`TSCI`" is our proposed method detailed in Algorithm 1. "`TSHT`" and "`CIIV`" are the methods proposed in Guo et al. (2018) and Windmeijer et al. (2021).

			`TSCI-RF`				Proportions of selection				`TSLS`	Other RF(oracle)		
vio	a	n	Oracle	Comp	Robust	Invalidity	$\widehat{q}_c = 0$	$\widehat{q}_c = 1$	$\widehat{q}_c = 2$	$\widehat{q}_c = 3$		Init	Plug	Full
2	0.0	1000	0.92	0.01	0.01	0.00	1.00	0.00	0.00	0.00	0.00	0.82	0.30	0.01
		3000	0.94	0.94	0.94	1.00	0.00	0.00	1.00	0.00	0.00	0.91	0.63	0.00
		5000	0.94	0.94	0.94	1.00	0.00	0.00	1.00	0.00	0.00	0.88	0.77	0.00
	0.5	1000	0.92	0.11	0.12	0.21	0.79	0.08	0.13	0.00	0.00	0.86	0.49	0.01
		3000	0.94	0.93	0.92	1.00	0.00	0.00	0.97	0.03	0.00	0.90	0.37	0.00
		5000	0.94	0.94	0.94	1.00	0.00	0.00	0.99	0.01	0.00	0.88	0.03	0.00
	1.0	1000	0.95	0.89	0.89	0.96	0.04	0.02	0.93	0.01	0.00	0.89	0.40	0.01
		3000	0.93	0.93	0.93	1.00	0.00	0.00	0.99	0.01	0.00	0.92	0.00	0.00
		5000	0.93	0.93	0.93	1.00	0.00	0.00	1.00	0.00	0.00	0.90	0.00	0.00

Table 8: Empirical Coverage for Setting B1 with Vio=2. "`TSCI-RF`" is our proposed `TSCI` estimator with random forests. "Oracle", "Comp" and "Robust" correspond to the estimators with $\mathcal{V}_q$ selected by the oracle knowledge, the comparison method, and the robust method. "Invalidity" reports the proportion of detecting the proposed IV as invalid. "Proportions of selection" reports the selection frequencies of $\mathcal{V}_q$, $0 \leq q \leq 3$. 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				Proportions of selection				TSLS	Other RF(oracle)		
vio	a	n	Oracle	Comp	Robust	Invalidity	$\widehat{q}_c = 0$	$\widehat{q}_c = 1$	$\widehat{q}_c = 2$	$\widehat{q}_c = 3$		Init	Plug	Full
1	0.0	1000	0.02	0.53	0.53	0.01	0.99	0.01	0.00	0.00	0.56	0.13	0.48	0.30
		3000	0.01	0.01	0.01	1.00	0.00	0.84	0.16	0.00	0.56	0.04	0.14	0.25
		5000	0.00	0.00	0.00	1.00	0.00	0.85	0.15	0.00	0.56	0.03	0.05	0.23
	0.5	1000	0.01	0.26	0.25	0.24	0.76	0.24	0.01	0.00	0.33	0.08	0.24	0.22
		3000	0.00	0.00	0.00	1.00	0.00	0.97	0.02	0.01	0.33	0.03	0.16	0.19
		5000	0.00	0.00	0.00	1.00	0.00	0.98	0.01	0.01	0.33	0.02	0.22	0.18
	1.0	1000	0.00	0.01	0.01	0.95	0.05	0.93	0.01	0.00	0.23	0.04	0.15	0.13
		3000	0.00	0.00	0.00	1.00	0.00	0.99	0.01	0.00	0.23	0.01	0.38	0.11
		5000	0.00	0.00	0.00	1.00	0.00	0.98	0.02	0.01	0.23	0.01	0.37	0.10
2	0.0	1000	0.00	0.54	0.54	0.00	1.00	0.00	0.00	0.00	0.56	0.11	0.53	0.29
		3000	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.56	0.05	0.13	0.25
		5000	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.56	0.03	0.05	0.23
	0.5	1000	0.01	0.38	0.38	0.21	0.79	0.08	0.13	0.00	0.33	0.08	0.34	0.23
		3000	0.00	0.00	0.01	1.00	0.00	0.00	0.97	0.03	0.33	0.03	0.19	0.20
		5000	0.00	0.00	0.01	1.00	0.00	0.00	0.99	0.01	0.33	0.03	0.29	0.19
	1.0	1000	0.01	0.04	0.04	0.96	0.04	0.02	0.93	0.01	0.23	0.04	0.25	0.15
		3000	0.00	0.00	0.00	1.00	0.00	0.00	0.99	0.01	0.23	0.02	0.52	0.12
		5000	0.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.23	0.01	0.48	0.11

Table 9: Absolute Bias for Setting B1. "TSCI-RF" is our proposed TSCI estimator with random forests. "Oracle", "Comp" and "Robust" correspond to the estimators with $\mathcal{V}_q$ selected by the oracle knowledge, the comparison method, and the robust method. "Invalidity" reports the proportion of detecting the proposed IV as invalid. "Proportions of selection" reports the selection frequencies of $\mathcal{V}_q$, $0 \leq q \leq 3$. 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 $\mathcal{V}_q$.

For Setting B2, we report the empirical coverage in Table 11 which are generally similar to those for Setting B1.

To approximate the real data analysis in Section 7.1, 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 $g(Z_i, X_i)$,

- Setting B3 (binary IV): $f(Z_i, X_i) = Z_i \cdot (1 + a \sum_{j=1}^{4} X_{i,j}(1 + X_{i,j+1})) - 3/10 \cdot \sum_{i=1}^{5} X_{i,j}$ and $g(Z_i, X_i) = Z_i + 1/2 \cdot Z_i \cdot (\sum_{i=1}^{3} X_{i,j}))$.

Compared to Settings B1 and B2, the outcome model in Setting B3 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. We specify $\mathcal{V}_0 = \{\overrightarrow{\mathbf{w}}(x)\}$ with $\overrightarrow{\mathbf{w}}(x) = \{1, x_1, \cdots, x_5\}$ and $\mathcal{V}_1 = \mathcal{V}_0 \cup \{z, z \cdot x_1, \cdots, z \cdot x_5\}$. With the specified basis sets, we implement TSCI with random forests as detailed in Algorithm 1 in the main paper. In Table 12, we demonstrate our proposed TSCI method for Setting B3. The observations are coherent with those for Settings B1 and B2. The main difference between the binary IV (Setting B3) and the continuous IV (Settings B1 and B2) 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.

			TSCI-RF				Proportions of selection				TSLS	Other RF(oracle)		
vio	a	n	Oracle	Comp	Robust	Invalidity	$\widehat{q}_c = 0$	$\widehat{q}_c = 1$	$\widehat{q}_c = 2$	$\widehat{q}_c = 3$		Init	Plug	Full
1	0.0	1000	0.49	0.11	0.11	0.01	0.99	0.01	0.00	0.00	0.08	0.49	0.82	0.22
		3000	0.32	0.32	0.32	1.00	0.00	0.84	0.16	0.00	0.05	0.32	0.38	0.14
		5000	0.23	0.23	0.23	1.00	0.00	0.85	0.15	0.00	0.04	0.23	0.27	0.11
	0.5	1000	0.38	0.13	0.14	0.24	0.76	0.24	0.01	0.00	0.05	0.38	0.60	0.19
		3000	0.22	0.22	0.23	1.00	0.00	0.97	0.02	0.01	0.03	0.22	0.26	0.11
		5000	0.17	0.17	0.17	1.00	0.00	0.98	0.01	0.01	0.02	0.17	0.19	0.09
	1.0	1000	0.25	0.24	0.24	0.95	0.05	0.93	0.01	0.00	0.04	0.25	0.33	0.14
		3000	0.13	0.13	0.13	1.00	0.00	0.99	0.01	0.00	0.02	0.13	0.18	0.08
		5000	0.10	0.10	0.10	1.00	0.00	0.98	0.02	0.01	0.02	0.10	0.13	0.06
2	0.0	1000	0.49	0.17	0.17	0.00	1.00	0.00	0.00	0.00	0.12	0.49	0.85	0.21
		3000	0.31	0.31	0.31	1.00	0.00	0.00	1.00	0.00	0.07	0.31	0.38	0.14
		5000	0.23	0.23	0.23	1.00	0.00	0.00	1.00	0.00	0.06	0.23	0.27	0.11
	0.5	1000	0.38	0.16	0.16	0.21	0.79	0.08	0.13	0.00	0.07	0.38	0.70	0.18
		3000	0.23	0.23	0.26	1.00	0.00	0.00	0.97	0.03	0.04	0.23	0.28	0.11
		5000	0.17	0.17	0.21	1.00	0.00	0.00	0.99	0.01	0.03	0.17	0.20	0.09
	1.0	1000	0.24	0.24	0.24	0.96	0.04	0.02	0.93	0.01	0.05	0.24	0.40	0.14
		3000	0.13	0.13	0.13	1.00	0.00	0.00	0.99	0.01	0.03	0.13	0.22	0.08
		5000	0.10	0.10	0.11	1.00	0.00	0.00	1.00	0.00	0.02	0.10	0.15	0.06

Table 10: Average CI length for Setting B1. "TSCI-RF" is our proposed TSCI estimator with random forests. "Oracle", "Comp" and "Robust" correspond to the estimators with $\mathcal{V}_q$ selected by the oracle knowledge, the comparison method, and the robust method. "Invalidity" reports the proportion of detecting the proposed IV as invalid. "Proportions of selection" reports the selection frequencies of $\mathcal{V}_q$, $0 \leq q \leq 3$. 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 $\mathcal{V}_q$.

We also observe that the bias correction is effective and improves the coverage when the interaction a is relatively small.

D.5 Binary Treatment

We consider the binary treatment setting and explore the finite-sample performance of our proposed TSCI method. We consider the outcome model $Y_i = D_i\beta + Z_i + 0.2 \cdot \sum_{j=1}^{p} X_{i,j} + \epsilon_i$ with $\beta = 1$ and $p_x = 20$. We generate ϵ_i and δ_i following bivariate normal distribution with zero mean, unit variance and covariance as 0.5. We generate the binary treatment D_i with the conditional mean as

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

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

- Setting 1 (continuous IV): generate Z_i, X_i following Section 6.2. $f(Z_i, X_i) = -25/12 + Z_i + Z_i^2 + 1/8 \cdot Z_i^4 + Z_i \cdot (a \cdot \sum_{j=1}^{5} X_{i,j}) - 3/10 \cdot \sum_{j=1}^{p_x} X_{i,j}$.
- Setting 2 (binary IV): generate $f(Z_i, X_i) = Z_i \cdot (1 + a \cdot \sum_{j=1}^{5} X_{i,j}) - \sum_{j=1}^{p} 0.3 X_{i,j}$.

			TSCI-RF				Proportions of selection				TSLS	Other RF(oracle)		
vio	a	n	Oracle	Comp	Robust	Invalidity	$\widehat{q}_c = 0$	$\widehat{q}_c = 1$	$\widehat{q}_c = 2$	$\widehat{q}_c = 3$		Init	Plug	Full
1	0	1000	0.84	0.84	0.84	1.00	0.00	1.00	0.00	0.00	0.58	0.81	0.16	0.00
		3000	0.93	0.93	0.94	1.00	0.00	0.96	0.04	0.00	0.11	0.91	0.00	0.00
		5000	0.93	0.94	0.94	1.00	0.00	0.95	0.05	0.00	0.01	0.93	0.00	0.00
	0.5	1000	0.94	0.94	0.95	1.00	0.00	0.95	0.04	0.01	0.00	0.88	0.00	0.01
		3000	0.93	0.93	0.92	1.00	0.00	0.95	0.05	0.00	0.00	0.92	0.00	0.01
		5000	0.93	0.93	0.93	1.00	0.00	0.99	0.01	0.00	0.00	0.92	0.00	0.00
	1	1000	0.94	0.94	0.95	1.00	0.00	0.98	0.01	0.00	0.00	0.92	0.00	0.01
		3000	0.95	0.96	0.96	1.00	0.00	0.98	0.02	0.00	0.00	0.93	0.00	0.00
		5000	0.93	0.93	0.93	1.00	0.00	0.99	0.01	0.00	0.00	0.91	0.00	0.00
2	0	1000	0.84	0.17	0.17	0.99	0.01	0.97	0.02	0.00	0.55	0.16	0.13	0.00
		3000	0.96	0.96	0.94	1.00	0.00	0.00	0.99	0.01	0.14	0.93	0.00	0.00
		5000	0.95	0.95	0.94	1.00	0.00	0.00	0.99	0.01	0.02	0.92	0.00	0.01
	0.5	1000	0.95	0.73	0.72	0.92	0.08	0.17	0.70	0.05	0.00	0.68	0.00	0.03
		3000	0.94	0.94	0.94	1.00	0.00	0.00	0.95	0.05	0.00	0.93	0.00	0.00
		5000	0.95	0.95	0.95	1.00	0.00	0.00	0.98	0.02	0.00	0.92	0.00	0.00
	1	1000	0.96	0.95	0.96	1.00	0.00	0.01	0.93	0.06	0.00	0.94	0.00	0.01
		3000	0.96	0.95	0.94	1.00	0.00	0.00	0.95	0.05	0.00	0.93	0.00	0.00
		5000	0.96	0.96	0.95	1.00	0.00	0.00	0.97	0.03	0.00	0.95	0.00	0.00

Table 11: CI coverage for Setting B2. "TSCI-RF" is our proposed TSCI estimator with random forests. "Oracle", "Comp" and "Robust" correspond to the estimators with $\mathcal{V}_q$ selected by the oracle knowledge, the comparison method, and the robust method. "Invalidity" reports the proportion of detecting the proposed IV as invalid. "Proportions of selection" reports the selection frequencies of $\mathcal{V}_q$, $0 \leq q \leq 3$. 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 $\mathcal{V}_q$.

The binary treatment result is summarized in Table 13. 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 $Q_{\max} = 0$. 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.

Appendix E. Additional Results for Real Data Analysis

This section contains the additional results for the real data analysis. In Figure 9, we first show the importance score of all variables in random forests, based on which the six most important covariates are selected to construct the basis set $\mathcal{V}_1$ in Section 7.1.

		TSCI-RF										RF-Init	
		Bias			Length			Coverage			Invalidity	Bias	Coverage
a	n	Oracle	Comp	Robust	Oracle	Comp	Robust	Oracle	Comp	Robust		Oracle	Oracle
0.25	1000	0.01	0.56	0.56	0.42	0.23	0.23	0.90	0.28	0.28	0.31	0.10	0.82
	3000	0.00	0.01	0.01	0.23	0.23	0.23	0.95	0.95	0.95	0.99	0.05	0.84
	5000	0.00	0.00	0.00	0.18	0.18	0.18	0.94	0.94	0.94	1.00	0.03	0.86
0.50	1000	0.00	0.02	0.02	0.22	0.22	0.22	0.93	0.90	0.90	0.97	0.04	0.90
	3000	-0.00	-0.00	-0.00	0.12	0.12	0.12	0.93	0.93	0.93	1.00	0.02	0.89
	5000	0.00	0.00	0.00	0.09	0.09	0.09	0.95	0.95	0.95	1.00	0.02	0.89
0.75	1000	0.00	0.00	0.00	0.15	0.15	0.15	0.92	0.90	0.90	0.99	0.02	0.91
	3000	0.00	0.00	0.00	0.08	0.08	0.08	0.94	0.94	0.94	1.00	0.01	0.89
	5000	-0.00	-0.00	-0.00	0.06	0.06	0.06	0.93	0.93	0.93	1.00	0.01	0.91

Table 12: Bias, length, and coverage for Setting B3. The columns indexed with "TSCI-RF" corresponds to our proposed TSCI with random forests, where the columns indexed with "Bias", "Length", and "Coverage" correspond to the bias of the point estimator, the length and empirical coverage of the constructed CI respectively. The columns indexed with "Oracle", "Comp" and "Robust" correspond to 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 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 $\mathcal{V}_q$.

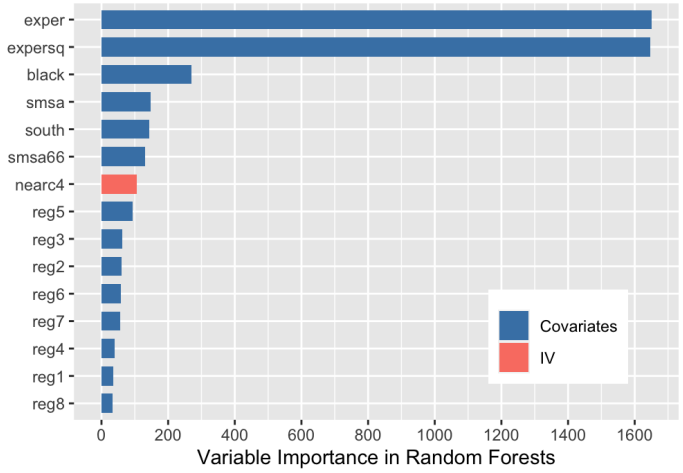

Figure 9: The importance score of each variable in random forest fitting.

E.1 Other Specifications of $\mathcal{V}_2$

In the following, in addition to the basis set $\mathcal{V}_2$ considered in Section 7.1, we consider three more specifications of $\mathcal{V}_2$, as detailed in Table 14, and test the robustness of TSCI's selection process.

We implement TSCI with random forests as detailed in Algorithm 1 with specifying $\mathcal{V}_0, \mathcal{V}_1$ as in Section 7.1 and different $\mathcal{V}_2$ in Table 14. In Table 15, we observe that the point

			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
2	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 13: 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 $Q_{\max} < 1$.

estimators and confidence intervals are relatively stable even with different specifications of $\mathcal{V}_2$.

$\mathcal{V}_2$	$\mathcal{V}_1 \cup$ {nearc4·reg1, nearc4·reg2, nearc4·reg3, nearc4·reg4,nearc4·reg5, nearc4·reg6, nearc4·reg7, nearc4·reg8}
$\mathcal{V}_2^{(1)}$	$\mathcal{V}_1 \cup \{\texttt{nearc4} \cdot \texttt{exper}^3\}$
$\mathcal{V}_2^{(2)}$	$\mathcal{V}_1 \cup \{\texttt{nearc4} \cdot \texttt{exper} \cdot \texttt{black}, \texttt{nearc4} \cdot \texttt{exper} \cdot \texttt{south}, \texttt{nearc4} \cdot \texttt{exper} \cdot \texttt{smsa}, \texttt{nearc4} \cdot \texttt{exper} \cdot \texttt{smsa66},$ $\texttt{nearc4} \cdot \texttt{expersq} \cdot \texttt{black}, \texttt{nearc4} \cdot \texttt{expersq} \cdot \texttt{south}, \texttt{nearc4} \cdot \texttt{expersq} \cdot \texttt{smsa}, \texttt{nearc4} \cdot \texttt{expersq} \cdot \texttt{smsa66},$ $\texttt{nearc4} \cdot \texttt{black} \cdot \texttt{south}, \texttt{nearc4} \cdot \texttt{black} \cdot \texttt{smsa}, \texttt{nearc4} \cdot \texttt{black} \cdot \texttt{smsa66},$ $\texttt{nearc4} \cdot \texttt{south} \cdot \texttt{smsa}, \texttt{nearc4} \cdot \texttt{south} \cdot \texttt{smsa66}, \texttt{nearc4} \cdot \texttt{smsa} \cdot \texttt{smsa66}\}$
$\mathcal{V}_2^{(3)}$	$\mathcal{V}_1 \cup \{\texttt{nearc4} \cdot \texttt{exper}^3, \texttt{nearc4} \cdot \texttt{exper} \cdot \texttt{black}, \texttt{nearc4} \cdot \texttt{exper} \cdot \texttt{south}, \texttt{nearc4} \cdot \texttt{exper} \cdot \texttt{smsa}, \texttt{nearc4} \cdot \texttt{exper} \cdot \texttt{smsa66},$ $\texttt{nearc4} \cdot \texttt{expersq} \cdot \texttt{black}, \texttt{nearc4} \cdot \texttt{expersq} \cdot \texttt{south}, \texttt{nearc4} \cdot \texttt{expersq} \cdot \texttt{smsa}, \texttt{nearc4} \cdot \texttt{expersq} \cdot \texttt{smsa66},$ $\texttt{nearc4} \cdot \texttt{black} \cdot \texttt{south}, \texttt{nearc4} \cdot \texttt{black} \cdot \texttt{smsa}, \texttt{nearc4} \cdot \texttt{black} \cdot \texttt{smsa66},$ $\texttt{nearc4} \cdot \texttt{south} \cdot \texttt{smsa}, \texttt{nearc4} \cdot \texttt{south} \cdot \texttt{smsa66}, \texttt{nearc4} \cdot \texttt{smsa} \cdot \texttt{smsa66}\}$

Table 14: Different specifications of the second basis set $\mathcal{V}_2$. $\mathcal{V}_2$ is the specification we used in Section 7.1. $\mathcal{V}_2^{(1)}$ contains the two-way interaction of the IV with covariates `exper` and `expersq` and $\mathcal{V}_1$; $\mathcal{V}_2^{(2)}$ includes all the two-way interactions of the IV with 6 most important covariates excluding the interaction of the IV with `exper`, `expersq`; $\mathcal{V}_2^{(3)}$ includes all the two-way interactions of the IV with 6 most important covariates.

Identifications	TSCI		Proportions of selection			Prob($Q_{\max} > \widehat{q}_c$)	IV str.
	Estimate	CI	$\widehat{q}_c = 0$	$\widehat{q}_c = 1$	$\widehat{q}_c = 2$		
TSCI-$\mathcal{V}_2$	0.0604	(0.0294, 0.0914)	59.2%	38.2%	2.6%	93.6%	112.7950
TSCI-$\mathcal{V}_2^{(1)}$	0.0575	(0.0263, 0.0886)	40.0%	58.6%	1.4%	2.3%	111.8835
TSCI-$\mathcal{V}_2^{(2)}$	0.0614	(0.0303, 0.0916)	55.0%	40.0%	5.0%	72.9%	111.6233
TSCI-$\mathcal{V}_2^{(3)}$	0.0575	(0.0268, 0.0891)	39.2%	60.6%	0.2%	0%	113.0134

Table 15: Inference and basis selection of `TSCI` with different identifications of $\mathcal{V}_2$. The columns indexed with "Inference" correspond to the point estimators and the CI reported by `TSCI` in Algorithm 1 with $\mathcal{V}_{\widehat{q}_c}$. The columns indexed with "Proportions of selection" correspond to proportions of different selections over 500 rounds of simulations. The column indexed with "Prob($Q_{\max} > \widehat{q}_c$)" indicates the proportion of $Q_{\max} > \widehat{q}_c$. This case means that `TSCI` is able to selects $\widehat{q}_c$ without any weak IV constraints. The last column shows the average IV strength for `TSCI` estimators.

References

Takeshi Amemiya. The nonlinear two-stage least-squares estimator. *Journal of Econometrics*, 2(2):105–110, 1974.

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

Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2009.

Joshua D. Angrist, Guido W. Imbens, and Alan B. Krueger. Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, 14(1):57–67, 1999.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *Annals of Statistics*, 47(2):1148–1178, 2019.

Philipp Bach, Malte S. Kurz, Victor Chernozhukov, Martin Spindler, and Sven Klaassen. DoubleML: An object-oriented implementation of double machine learning in R. *Journal of Statistical Software*, 108(3):1–56, 2024. doi: 10.18637/jss.v108.i03.

Paul A. Bekker and Federico Crudu. Jackknife instrumental variable estimation with heteroskedasticity. *Journal of Econometrics*, 185(2):332–342, 2015.

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.

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.

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.

Peter Bühlmann and Torsten Hothorn. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, 22(4):477–505, 2007.

Peter Bühlmann and Bin Yu. Sparse boosting. *Journal of Machine Learning Research*, 7 (6), 2006.

David Card. Using geographic variation in college proximity to estimate the return to schooling. *National Bureau of Economic Research*, 1993.

David Carl, Corinne Emmenegger, Peter Bühlmann, and Zijian Guo. Tsci: Two stage curvature identification for causal inference with invalid instruments in r. *Journal of Statistical Software*, 114:1–21, 2025.

Jiafeng Chen, Daniel L. Chen, and Greg Lewis. Mostly harmless machine learning: learning optimal instruments in linear iv models. *Journal of Machine Learning Research*, page forthcoming, 2023.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *Journal of Econometrics*, 21(1), 2018.

Corinne Emmenegger. *dmlalg: Double machine learning algorithms*, 2021.

Corinne Emmenegger and Peter Bühlmann. Regularizing double machine learning in partially linear endogenous models. *Electronic Journal of Statistics*, 15(2):6461–6543, 2021.

Peter J. Green and Bernard W. Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach*. Crc Press, 1993.

Zijian Guo. Causal inference with invalid instruments: post-selection problems and a solution using searching and sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 85(3):959–985, 2023. ISSN 1369-7412.

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.

Chirok Han. Detecting invalid instruments using l1-gmm. *Economics Letters*, 101(3):285–287, 2008.

Christian Hansen, Jerry Hausman, and Whitney Newey. Estimation with many instrumental variables. *Journal of Business & Economic Statistics*, 26(4):398–422, 2008.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, pages 1029–1054, 1982.

Jerry A. Hausman, Whitney K. Newey, Tiemen Woutersen, John C. Chao, and Norman R. Swanson. Instrumental variable estimation with heteroskedasticity and many instruments. *Quantitative Economics*, 3(2):211–255, 2012.

Thomas R. Ten Have, Marshall M. Joffe, Kevin G. Lynch, Gregory K. Brown, Stephen A. Maisto, and Aaron T. Beck. Causal mediation analyses with rank preserving models. *Biometrics*, 63(3):926–934, 2007.

James J. Heckman. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, 5(4):475–492, 1976.

Paul W. Holland. Causal inference, path analysis and recursive structural equations models. *ETS Research Report Series*, 1988(1):i–50, 1988.

Guido W. Imbens. Instrumental variables: An econometrician's perspective. *Statistical Science*, 29(3):323–358, 2014.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

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.

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.

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.

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.

Arthur Lewbel. Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models. *Journal of Business & Economic Statistics*, 30(1):67–80, 2012.

Arthur Lewbel. The identification zoo: Meanings of identification in econometrics. *Journal of Economic Literature*, 57(4):835–903, 2019.

Yi Lin and Yongho Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590, 2006.

Ruiqi Liu, Zuofeng Shang, and Guang Cheng. On deep instrumental variables estimate. *arXiv preprint arXiv:2004.14954*, 2020.

Clement J. McDonald, Siu L. Hui, and William M. Tierney. Effects of computer reminders for influenza vaccination on morbidity during influenza epidemics. *MD Computing: Computers in Medical Practice*, 9(5):304–312, 1992.

Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999, 2006.

Nicolai Meinshausen, Lukas Meier, and Peter Bühlmann. P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 2009.

Whitney K. Newey. Efficient instrumental variables estimation of nonlinear models. *Econometrica*, pages 809–837, 1990.

Patrick Puhani. The heckman correction for sample selection and its critique. *Journal of Economic Surveys*, 14(1):53–68, 2000.

Thomas J. Rothenberg. Approximating the distributions of econometric estimators and test statistics. *Handbook of Econometrics*, 2:881–935, 1984.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

John D. Sargan. The estimation of economic relationships using instrumental variables. *Econometrica*, pages 393–415, 1958.

Michelle Shardell and Luigi Ferrucci. Instrumental variable analysis of multiplicative models with potentially invalid instruments. *Statistics in Medicine*, 35(29):5430–5447, 2016.

Dylan S. Small. Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association*, 102(479):1049–1058, 2007.

Jerzy Splawa-Neyman, Dorota M. Dabrowska, and Terence P. Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472, 1990.

Douglas Staiger and James H. Stock. Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586, 1997.

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.

Baoluo Sun, Zhonghua Liu, and E. J. Tchetgen Tchetgen. Semiparametric efficient G-estimation with invalid instrumental variables. *Biometrika*, 110(4):953–971, 02 2023.

Eric Tchetgen Tchetgen, BaoLuo Sun, and Stefan Walter. The genius approach to robust mendelian randomization inference. *Statistical Science*, 36(3):443–464, 2021.

Stijn Vansteelandt and Vanessa Didelez. Improving the robustness and efficiency of covariate-adjusted linear instrumental variable estimators. *Scandinavian Journal of Statistics*, 45(4):941–961, 2018.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

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.

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.

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.

Tiemen Woutersen and Jerry A. Hausman. Increasing the power of specification tests. *Journal of Econometrics*, 211(1):166–175, 2019.