Journal of the Royal Statistical Society Series B: Statistical Methodology, 2023, **85**, 959–985
https://doi.org/10.1093/jrsssb/qkad049
Advance access publication 20 May 2023
Original Article

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series B
Statistical Methodology
B

Causal inference with invalid instruments: post-selection problems and a solution using searching and sampling

Zijian Guo

Department of Statistics, Rutgers University, Piscataway, USA

Address for correspondence: Zijian Guo, Department of Statistics, Rutgers University, 110 Frelinghuysen Rd, Piscataway, NJ 08854, USA. Email: zijguo@stat.rutgers.edu

Abstract

Instrumental variable methods are among the most commonly used causal inference approaches to deal with unmeasured confounders in observational studies. The presence of invalid instruments is the primary concern for practical applications, and a fast-growing area of research is inference for the causal effect with possibly invalid instruments. This paper illustrates that the existing confidence intervals may undercover when the valid and invalid instruments are hard to separate in a data-dependent way. To address this, we construct uniformly valid confidence intervals that are robust to the mistakes in separating valid and invalid instruments. We propose to search for a range of treatment effect values that lead to sufficiently many valid instruments. We further devise a novel sampling method, which, together with searching, leads to a more precise confidence interval. Our proposed searching and sampling confidence intervals are uniformly valid and achieve the parametric length under the finite-sample majority and plurality rules. We apply our proposal to examine the effect of education on earnings. The proposed method is implemented in the R package `RobustIV` available from CRAN.

Keywords: majority rule, Mendelian randomization, plurality rule, uniform inference, unmeasured confounders

1 Introduction

Unmeasured confounders are a major concern for causal inference with observational studies. The instrumental variable (IV) method is one of the most commonly used causal inference approaches to deal with unmeasured confounders. The IVs are required to satisfy three identification conditions: conditioning on the baseline covariates,

(A1) the IVs are associated with the treatment;
(A2) the IVs are independent of the unmeasured confounders;
(A3) the IVs have no direct effect on the outcome.

The main challenge of IV-based methods is identifying instruments satisfying (A1), (A2), and (A3) simultaneously. Assumptions (A2) and (A3) are crucial for identifying the causal effect as they assume that the IVs can only affect the outcome through the treatment. However, assumptions (A2) and (A3) may be violated in applications and cannot even be tested in a data-dependent way. We define an IV as 'invalid' if it violates assumptions (A2) or (A3). If an invalid IV is mistakenly taken as valid, it generally leads to a biased estimator of the causal effect. A fast-growing literature is to conduct causal inference with possibly invalid IVs (e.g., Bowden et al., 2015, 2016; Fan & Wu, 2020; Guo et al., 2018; Kang et al., 2020, 2016; Kolesár et al., 2015; Tchetgen et al., 2021; Windmeijer et al., 2019). Many of these works are motivated by Mendelian Randomization studies using genetic variants as IVs (Burgess et al., 2017). The adopted genetic variants can be invalid

Received: August 19, 2021. **Revised:** February 21, 2023. **Accepted:** April 4, 2023

since they may affect both treatment and outcome due to the pleiotropy effect (Davey Smith & Ebrahim, 2003).

The current paper focuses on the linear outcome model with heteroscedastic errors and multiple possibly invalid IVs. Under this multiple IV framework, the effect identification requires extra conditions, such as the majority rule (Kang et al., 2016) and plurality rule (Guo et al., 2018), assuming that a sufficient proportion of IVs are valid, but the validity of any IV is unknown a priori. The existing works (e.g., Guo et al., 2018; Kang et al., 2016; Windmeijer et al., 2019, 2021) leveraged the majority and plurality rules to select valid IVs in a data-dependent way. The selected valid IVs were used for the following-up causal inference, with the invalid IVs being included as the baseline covariates. However, there are chances that we make mistakes in separating valid and invalid IVs. Certain invalid IVs can be hard to detect in applications with the given amount of data. We refer to such invalid IVs as 'locally invalid IVs' and provide a formal definition in Definition 1. In Section 3, we demonstrate that, when there exist locally invalid IVs, the existing inference methods `TSHT` (Guo et al., 2018) and `CIIV` (Windmeijer et al., 2021) may produce unreliable confidence intervals, where `TSHT` is shorthanded for two stage hard thresholding and `CIIV` is shorthanded for confidence interval method for selecting valid IVs.

The current paper proposes uniformly valid confidence intervals (CIs) robust to IV selection error. To better accommodate finite-sample inferential properties, we introduce the finite-sample majority and plurality rule in Conditions 3 and 4, respectively. We start with the finite-sample majority rule and explain the *searching* idea under this setting. For every value of the treatment effect, we implement a hard thresholding step to decide which candidate IVs are valid. We propose to search for a range of treatment effect values such that the majority of candidate IVs can be taken as valid. We further propose a novel *sampling* method to improve the precision of the searching CI. For the plurality rule setting, we first construct an initial estimator $\hat{\mathcal{V}}$ of the set of valid IVs and then apply the searching and sampling method over $\hat{\mathcal{V}}$. Our proposed searching CI works even if $\hat{\mathcal{V}}$ does not correctly recover the set of valid IVs.

Our proposed searching and sampling CIs are shown to achieve the desired coverage under the finite-sample majority or plurality rule. The CIs are uniformly valid in the sense that the coverage is guaranteed even in the presence of locally invalid IVs. We also establish that the searching and sampling CIs achieve the $1/\sqrt{n}$ length. The proposed CIs are computationally efficient as the searching method searches over one-dimension space, and we only resample the reduced-form estimators instead of the entire data.

We discuss other related works on invalid IVs in the following. Kang et al. (2020) proposed the `union` CI, which takes a union of intervals being constructed by a given number of valid IVs and passing the Sargan test (Sargan, 1958). The `union` CI requires an upper bound for the number of invalid IVs, while our proposed CI does not rely on such information. Our proposed searching and sampling CIs are typically much shorter and computationally more efficient than the `union` CI. Different identifiability conditions have been proposed to identify the causal effect when the IV assumptions (A2) and (A3) fail to hold. Bowden et al. (2015) and Kolesár et al. (2015) assumed that the IVs' direct effect on the outcome and the IVs' association with the treatment are nearly orthogonal. In addition, there has been progress in identifying the treatment effect when all IVs are invalid; for instance, Lewbel (2012), Tchetgen et al. (2021), and Liu et al. (2020) leveraged heteroscedastic covariance of regression errors, while Guo and Bühlmann (2022) relied on identifying nonlinear treatment models with machine learning methods. Goh and Yu (2022) emphasized the importance of accounting for the uncertainty of choosing valid IVs by the penalized methods (Kang et al., 2016) and proposed a Bayesian approach to construct a credible interval with possibly invalid IVs. However, no theoretical justification exists for this credible interval being a valid CI. In Mendelian Randomization studies, much progress has been made in inference with summary statistics, which is not the main focus of the current paper; see Bowden et al. (2015, 2016) and Zhao et al. (2020) for examples. In the Generalized Method of Moments setting, Liao (2013), Cheng and Liao (2015), and Caner et al. (2018) leveraged a set of prespecified valid moment conditions for the model identification and further tested the validity of another set of moment conditions. The current paper is entirely different in the sense that there is no prior knowledge of the validity of any given IV.

The construction of uniformly valid CIs after model selection is a major focus in statistics under the name of post-selection inference. Many methods (Berk et al., 2013; Cai & Guo, 2017; Chernozhukov et al., 2015; Javanmard & Montanari, 2014; Lee et al., 2016; Leeb & Pötscher,

2005; van de Geer et al., 2014; Xie & Wang, 2022; Zhang & Zhang, 2014) have been proposed, and the focus is on (but not limited to) inference for regression coefficients after some variables or submodels are selected. This paper considers a different problem, post-selection inference for causal effect with possibly invalid instruments. To our best knowledge, this problem has not been carefully investigated in the post-selection inference literature. Furthermore, our proposed sampling method differs from other existing post-selection inference methods.

Notations. For a set S and a vector $x \in \mathbb{R}^p$, S^c denotes the complement of S, $|S|$ denotes the cardinality of S, and x_S is the subvector of x with indices in S. For sets $B \subset A$, we define $A \backslash B = A \cap B^c$. The ℓ_q norm of a vector x is defined as $\|x\|_q = (\sum_{l=1}^p |x_l|^q)^{\frac{1}{q}}$ for $q \geq 0$ with $\|x\|_0 = |\{1 \leq l \leq p : x_l \neq 0\}|$ and $\|x\|_\infty = \max_{1 \leq l \leq p} |x_l|$. We use $\mathbf{0}_q$ and $\mathbf{1}_q$ to denote the q-dimension vector with all entries equal to 0 and 1, respectively. For a matrix X, $X_{i\cdot}$, $X_{\cdot j}$, and X_{ij} are used to denote its i-th row, j-th column, and (i, j) entry, respectively. For a sequence of random variables X_n indexed by n, we use $X_n \to^d X$ to denote that X_n converges to X in distribution. We use c and C to denote generic positive constants that may vary from place to place. 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 A, we use $\|A\|_2$ and $\|A\|_\infty$ to denote its spectral and element-wise maximum norm, respectively. For a symmetric matrix A, we use $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ to denote its maximum and minimum eigenvalues, respectively.

2 Models and reduced-form estimators

We consider the i.i.d. data $\{Y_i, D_i, X_{i\cdot}, Z_{i\cdot}\}_{1 \leq i \leq n}$, where $Y_i \in \mathbb{R}$, $D_i \in \mathbb{R}$, $X_{i\cdot} \in \mathbb{R}^{p_x}$, and $Z_{i\cdot} \in \mathbb{R}^{p_z}$ denote the outcome, the treatment, the baseline covariates, and candidate IVs, respectively. We consider the following outcome model with possibly invalid IVs (Kang et al., 2016; Small, 2007),

$$Y_i = D_i \beta^* + Z_{i\cdot}^\top \pi^* + X_{i\cdot}^\top \phi^* + e_i \quad \text{with} \quad \mathbf{E}(e_i Z_{i\cdot}) = \mathbf{0} \quad \text{and} \quad \mathbf{E}(e_i X_{i\cdot}) = \mathbf{0}, \tag{1}$$

where $\beta^* \in \mathbb{R}$ denotes the treatment effect, $\pi^* \in \mathbb{R}^{p_z}$, and $\phi^* \in \mathbb{R}^{p_x}$. We set the first element of $X_{i\cdot}$ as the constant 1. If the IVs satisfy (A2) and (A3), this leads to $\pi^* = \mathbf{0}$ in (1). A nonzero vector π^* indicates that the IVs violate the classical IV assumptions (A2) and (A3); see Figure 1 for an illustration.

We consider the association model for the treatment D_i,

$$D_i = Z_{i\cdot}^\top \gamma^* + X_{i\cdot}^\top \psi^* + \delta_i \quad \text{with} \quad \mathbf{E}(\delta_i Z_{i\cdot}) = \mathbf{0} \quad \text{and} \quad \mathbf{E}(\delta_i X_{i\cdot}) = \mathbf{0}. \tag{2}$$

The model (2) can be viewed as the best linear approximation of D_i by $Z_{i\cdot}$ and $X_{i\cdot}$ instead of a casual model. In (2), $\gamma_j^* \neq 0$ indicates that the j-th IV satisfies assumption (A1). Due to unmeasured confounders, e_i and δ_i can be correlated, and the treatment D_i is endogeneous with $\mathbf{E}(D_i e_i) \neq 0$.

Following Kang et al. (2016), we discuss the causal interpretation of the model (1) and explain why $\pi^* \neq 0$ represents the violation of assumptions (A2) and (A3). For two treatment values $d, d' \in \mathbb{R}$ and two realizations of IVs $\mathbf{z}, \mathbf{z}' \in \mathbb{R}^{p_z}$, define the following potential outcome model:

$$Y_i^{(d', \mathbf{z}')} - Y_i^{(d, \mathbf{z})} = (d' - d)\beta^* + (\mathbf{z}' - \mathbf{z})^\top \kappa^* \quad \text{and} \quad \mathbf{E}(Y_i^{(0,0)} \mid Z_{i\cdot}, X_{i\cdot}) = Z_{i\cdot}^\top \eta^* + X_{i\cdot}^\top \phi^*,$$

where $\beta^* \in \mathbb{R}$ is the treatment effect, $\kappa^*, \eta^* \in \mathbb{R}^{p_z}$, and $\phi^* \in \mathbb{R}^{p_x}$. As illustrated in Figure 1, $\kappa_j^* \neq 0$ indicates that the j-th candidate IV has a direct effect on the outcome, which violates the assumption (A3); $\eta_j \neq 0$ indicates that the j-th candidate IV is associated with the unmeasured confounder, which violates assumption (A2). Under the consistency condition $Y_i = Y_i^{(D_i, Z_{i\cdot})}$, the above potential outcome model implies (1) with the invalidity vector $\pi^* = \kappa^* + \eta^*$ and $e_i = Y_i^{(0,0)} - \mathbf{E}(Y_i^{(0,0)} \mid Z_{i\cdot}, X_{i\cdot})$.

We define the set $\mathcal{S}$ of relevant instruments and the set $\mathcal{V}$ of valid instruments as

$$\mathcal{S} = \{1 \leq j \leq p_z : \gamma_j^* \neq 0\} \quad \text{and} \quad \mathcal{V} = \{j \in \mathcal{S} : \pi_j^* = 0\}. \tag{3}$$

The IVs belonging to $\mathcal{S}$ satisfy the IV assumption (A1). The set $\mathcal{V}$ is a subset of $\mathcal{S}$ and the IVs belonging to $\mathcal{V}$ satisfy the classical IV assumptions (A1)–(A3) simultaneously.

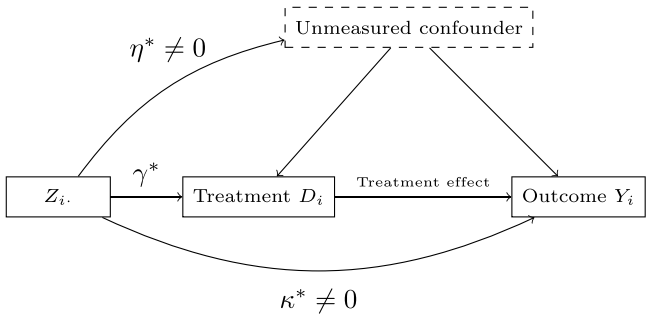

Figure 1. Illustration of (A2) and (A3) being violated in the model (1) with $\pi^* = \kappa^* + \eta^*$.

We now review the identification strategy under models (1) and (2). We plug in the treatment model (2) into the outcome model (1) and obtain the reduced-form model,

$$\begin{aligned} Y_i &= Z_{i\cdot}^\top \Gamma^* + X_{i\cdot}^\top \Psi^* + \epsilon_i \quad \text{with} \quad \mathbf{E}(Z_{i\cdot}\epsilon_i) = 0,\ \mathbf{E}(X_{i\cdot}\epsilon_i) = 0, \\ D_i &= Z_{i\cdot}^\top \gamma^* + X_{i\cdot}^\top \psi^* + \delta_i \quad \text{with} \quad \mathbf{E}(Z_{i\cdot}\delta_i) = 0,\ \mathbf{E}(X_{i\cdot}\delta_i) = 0, \end{aligned} \tag{4}$$

where $\Gamma^* = \beta^*\gamma^* + \pi^* \in \mathbb{R}^{p_z}$, $\Psi^* = \beta^*\psi^* + \phi^* \in \mathbb{R}^{p_x}$, and $\epsilon_i = \beta^*\delta_i + e_i$. We shall devise our methods under the model (4), which is induced by the models (1) and (2).

Since $Z_{i\cdot}$ and $X_{i\cdot}$ are uncorrelated with ϵ_i and δ_i, we can identify the reduced-form parameters Γ^* and γ^* in (4). However, since $\pi^* \neq \mathbf{0}$, the identification of β^* through solving the equation $\Gamma^* = \beta^*\gamma^* + \pi^* \in \mathbb{R}^{p_z}$ requires extra assumptions. Kang et al. (2016) and Bowden et al. (2016) proposed the following majority rule to identify β for $\pi^* \neq \mathbf{0}$.

Condition 1 (Population Majority Rule). More than half of the relevant IVs are valid; that is, $|\mathcal{V}| > |\mathcal{S}|/2$, where $\mathcal{V}$ and $\mathcal{S}$ are defined in (3).

Condition 1 requires that more than half of the relevant IVs are valid but does not directly require the knowledge of $\mathcal{V}$. For the j-th IV, we may identify the effect as $\beta^{[j]} = \Gamma_j^*/\gamma_j^*$. If the j-th IV is valid, $\beta^{[j]} = \beta^*$; otherwise, $\beta^{[j]} \neq \beta^*$. Under Condition 1, β^* can be identified using the majority of $\{\beta^{[j]}\}_{j\in\mathcal{S}}$. Guo et al. (2018) and Hartwig et al. (2017) proposed the following plurality rule as a weaker identification condition.

Condition 2 (Population Plurality Rule). The number of valid IVs is larger than the number of invalid IVs with any given invalidity level $v \neq 0$, that is,

$$|\mathcal{V}| > \max_{v\neq 0} |\mathcal{I}_v| \quad \text{with} \quad \mathcal{I}_v = \left\{ j \in \mathcal{S} : \pi_j^*/\gamma_j^* = v \right\},$$

where the set $\mathcal{V}$ of valid IVs is defined in (3).

The term π_j^*/γ_j^* represents the invalidity level: $\pi_j^*/\gamma_j^* \neq 0$ indicates that the j-th IV violates assumptions (A2) and (A3). $\mathcal{I}_v$ denotes the set of all IVs with the same invalidity level v. Note that $\mathcal{V} = \mathcal{I}_0$. Under the plurality rule, β^* can be identified using the largest cluster of $\{\beta^{[j]}\}_{j\in\mathcal{S}}$. We refer to Conditions 1 and 2 as *population* identification conditions since they are used to identify β^* with an infinite amount of data. We will propose finite-sample versions of Conditions 1 and 2 in Conditions 3 and 4, respectively.

We now present the data-dependent estimators of γ^* and Γ^* in the model (4). Define $p = p_x + p_z$, $W = (Z, X) \in \mathbb{R}^{n\times p}$, and $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n W_{i\cdot}(W_{i\cdot})^\top$. We estimate γ^* and Γ^* by the Ordinary Least Squares (OLS) estimators, defined as

$$(\hat{\Gamma}^\top, \hat{\Psi}^\top)^\top = (W^\top W)^{-1} W^\top Y \quad \text{and} \quad (\hat{\gamma}^\top, \hat{\psi}^\top)^\top = (W^\top W)^{-1} W^\top D. \tag{5}$$

Under regularity conditions, as $n \to \infty$, the OLS estimators satisfy

$$\sqrt{n}\begin{pmatrix} \hat{\Gamma} - \Gamma^* \\ \hat{\gamma} - \gamma^* \end{pmatrix} \xrightarrow{d} N(0, \text{Cov}) \quad \text{with} \quad \text{Cov} = \begin{pmatrix} \mathbf{V}^\Gamma & \mathbf{C} \\ \mathbf{C}^\top & \mathbf{V}^\gamma \end{pmatrix}, \tag{6}$$

where $\mathbf{V}^\Gamma \in \mathbb{R}^{p_z \times p_z}$, $\mathbf{V}^\gamma \in \mathbb{R}^{p_z \times p_z}$, and $\mathbf{C} \in \mathbb{R}^{p_z \times p_z}$ are explicitly defined in Online Supplementary Material, Lemma 1 in the supplement. Define the residues $\hat{\epsilon}_i = Y_i - Z_{i\cdot}^\top \hat{\Gamma} - X_{i\cdot}^\top \hat{\Psi}$ and $\hat{\delta}_i = D_i - Z_{i\cdot}^\top \hat{\gamma} - X_{i\cdot}^\top \hat{\psi}$ for $1 \le i \le n$. We estimate $\mathbf{V}^\Gamma$, $\mathbf{V}^\gamma$, and $\mathbf{C}$ in (6) by

$$\begin{aligned} \hat{\mathbf{V}}^\Gamma &= \left[\hat{\Sigma}^{-1}\left(\frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2 W_{i\cdot} W_{i\cdot}^\top\right)\hat{\Sigma}^{-1}\right]_{1:p_z,1:p_z}, \hat{\mathbf{V}}^\gamma = \left[\hat{\Sigma}^{-1}\left(\frac{1}{n}\sum_{i=1}^n \hat{\delta}_i^2 W_{i\cdot} W_{i\cdot}^\top\right)\hat{\Sigma}^{-1}\right]_{1:p_z,1:p_z}, \\ \hat{\mathbf{C}} &= \left[\hat{\Sigma}^{-1}\left(\frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i\hat{\delta}_i W_{i\cdot} W_{i\cdot}^\top\right)\hat{\Sigma}^{-1}\right]_{1:p_z,1:p_z}, \end{aligned} \tag{7}$$

where for a matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$, we use $\mathbf{A}_{1:p_z,1:p_z}$ to denote the $p_z \times p_z$ submatrix containing the first p_z rows and columns of $\mathbf{A}$. The variance estimators in (7) are robust to the heteroskedastic errors and referred to as the sandwich estimators (Eicker, 1967; Huber, 1967); see Chapter 4.2.3 of Wooldridge (2010) for details. We use the OLS estimators in (5) and the sandwich estimators in (7) as a prototype to discuss our proposed methods. Our proposed methods are effective for any reduced-form estimators satisfying (6). In Online Supplementary Material, Section A.3 in the supplement, we consider the high-dimensional setting with $p > n$ and construct the debiased Lasso estimators $\hat{\Gamma}$ and $\hat{\gamma}$ satisfying (6).

3 IV selection errors and nonuniform inference

In this section, we demonstrate that even if the majority rule (Condition 1) or plurality rule (Condition 2) holds, the CIs by `TSHT` (Guo et al., 2018) and `CIIV` (Windmeijer et al., 2021) may be unreliable. The main idea of Guo et al. (2018) and Windmeijer et al. (2021) is to estimate $\mathcal{V}$ by a set estimator $\hat{\mathcal{V}}$ and then identify β^* through the following expression or its weighted version:

$$\beta(\hat{\mathcal{V}}) = \sum_{j \in \hat{\mathcal{V}}} \Gamma_j^* \gamma_j^* / \sum_{j \in \hat{\mathcal{V}}} (\gamma_j^*)^2. \tag{8}$$

When there are no selection errors (i.e., $\hat{\mathcal{V}} = \mathcal{V}$), we have $\beta(\hat{\mathcal{V}}) = \beta^*$. The validity of the CIs by `TSHT` and `CIIV` requires $\hat{\mathcal{V}}$ to recover $\mathcal{V}$ correctly. However, there are chances to make mistakes in estimating $\mathcal{V}$ in finite samples, and the CIs by `TSHT` and `CIIV` are unreliable when invalid IVs are included in $\hat{\mathcal{V}}$. The IV selection error leads to a bias in identifying β^* with $\beta(\hat{\mathcal{V}})$. Even if $\hat{\mathcal{V}}$ used in (8) is selected in a data-dependent way, the target causal effect β^* is fixed, which is a main difference from the post-selection inference literature (e.g., Berk et al., 2013; Lee et al., 2016; Leeb & Pötscher, 2005). We provide more detailed discussions in Online Supplementary Material, Section A.1 in the supplement.

In the following, we define 'locally invalid IVs' as invalid IVs that are hard to be separated from valid IVs with a given amount of data. For $j, k \in \mathcal{S}$, define

$$\mathbf{T}_{j,k} := \min\{\mathbf{T}_{j,k}^0, \mathbf{T}_{k,j}^0\} \quad \text{with} \quad \mathbf{T}_{j,k}^0 = \sqrt{\frac{1}{n}\left(\mathbf{R}_{k,k}^{[j]}/[\gamma_k^*]^2 + \mathbf{R}_{j,j}^{[j]}/[\gamma_j^*]^2 - 2\mathbf{R}_{j,k}^{[j]}/[\gamma_k^*\gamma_j^*]\right)}, \tag{9}$$

where $\mathbf{R}^{[j]} = \mathbf{V}^\Gamma + (\beta^{[j]})^2 \mathbf{V}^\gamma - 2\beta^{[j]}\mathbf{C}$ and $\beta^{[j]} = \Gamma_j^*/\gamma_j^*$. As shown in Proposition 2, if the absolute difference between π_j^*/γ_j^* and π_k^*/γ_k^* for $j, k \in \mathcal{S}$ is above $2\sqrt{\log n} \cdot \mathbf{T}_{j,k}$, then we can tell that the j-th and k-th IVs have different invalidity levels. Particularly, the term $\gamma_k^* \mathbf{T}_{j,k}^0$ represents the standard

error of estimating π_k^* by $\hat{\Gamma}_k - \hat{\gamma}_k\hat{\beta}^{[j]}$, where $\hat{\beta}^{[j]} = \hat{\Gamma}_j/\hat{\gamma}_j$ denotes the causal effect estimator by assuming the j-th IV to be valid. By symmetry, $\gamma_j^*\mathbf{T}_{k,j}^0$ denotes the standard error of estimating π_j^* by $\hat{\Gamma}_j - \hat{\gamma}_j\hat{\beta}^{[k]}$. We now formally define locally invalid IVs.

Definition 1 (Locally invalid IV). For $j \in \mathcal{S}$, the j-th IV is locally invalid if

$$0 < |\pi_j^*/\gamma_j^*| < s_j(n) \quad \text{with} \quad s_j(n) := 2\sqrt{\log n} \cdot \max_{k\in\mathcal{V}} |\mathbf{T}_{j,k}|.$$

where $\mathbf{T}_{j,k}$ is defined in (9).

The definition of locally invalid IVs depends on the invalidity level π_j^*/γ_j^* and the separation level $s_j(n)$, which stands for the uncertainty level in separating the j-th IV and valid IVs. We show in Proposition 2 that the j-th IV, if invalid, can be separated from valid IVs if $|\pi_j^*/\gamma_j^*| \geq s_j(n)$. The separation level $s_j(n)$ is of the order $\sqrt{\log n/n}$ if all of $\{\gamma_j^*\}_{j\in\mathcal{S}}$ are constants. The large sample size enhances the power of detecting invalid IVs. For a sufficiently large n, the set of locally invalid IVs becomes empty since $s_j(n) \to 0$. Definition 1 is related to the local violation of IV exogeneity assumptions and valid moment conditions studied in Caner et al. (2018), Berkowitz et al. (2012), Guggenberger (2012), and Hahn and Hausman (2005).

The theoretical results of `TSHT` (Guo et al., 2018) or `CIIV` (Windmeijer et al., 2021) essentially assume that there are no locally invalid IVs; see Assumption 8 in Guo et al. (2018). However, the absence of locally invalid IVs may not properly accommodate real data analysis with a finite sample. For a given sample size n, there are chances that an invalid IV with $|\pi_j^*/\gamma_j^*| < s_j(n)$ is taken as valid by mistake. We now consider a numerical example and demonstrate that the coverage levels of CIs by `TSHT` and `CIIV` may be below the nominal level when locally invalid IVs exist.

Example 1 (Setting **S2** in Section 7). For the models (1) and (2), set $\gamma^* = 0.5 \cdot \mathbf{1}_{10}$ and $\pi^* = (\mathbf{0}_4^\top, \tau/2, \tau/2, -1/3, -2/3, -1, -4/3)^\top$. The plurality rule is satisfied with $|\mathcal{V}| = 4 > \max_{v\neq 0} |\mathcal{I}_v| = 2$. We vary τ across $\{0.025, 0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5\}$. $s_j(n)$ for $j = 5, 6$ in Definition 1 takes the value 0.96 for $n = 500$ and 0.53 for $n = 2{,}000$.

In this example, the fifth and sixth IVs (with a small τ) are locally invalid IVs, which are hard to detect for $n = 500$, 2,000. The empirical coverage is reported in Figure 2. For a small n and τ, the estimated sets $\hat{\mathcal{V}}$ by `TSHT` and `CIIV` often contain locally invalid IVs, and the coverage levels of `TSHT` and `CIIV` are below 95%. There are nearly no locally invalid IVs for $n = 2{,}000$ and $\tau = 0.5$, and the CIs by `TSHT` and `CIIV` achieve the nominal level 95%. For $\tau = 0.1$, with n increasing from 500 to 2,000, the fifth and sixth IVs may still be included in $\hat{\mathcal{V}}$. Consequently, the bias due to the locally invalid IVs remains unchanged, and the empirical coverage levels of `TSHT` and `CIIV` decrease since the standard errors get smaller with a larger n. In contrast, our proposed sampling CI in Algorithm 3 and the `Union` interval in Kang et al. (2020) achieve the desired coverage at the expense of wider intervals. Our proposed sampling CIs are significantly shorter than the `Union` intervals.

We introduce an oracle bias-aware CI as the benchmark when IV selection errors exist. The oracle bias-aware CI serves as a better benchmark than the oracle CI assuming the knowledge of $\mathcal{V}$ since it accounts for the IV selection error. For the `TSHT` estimator $\hat{\beta}$ and its standard error $\mathrm{SE}(\hat{\beta})$, we assume $(\hat{\beta} - \beta^*)/\mathrm{SE}(\hat{\beta}) \to^d N(b, 1)$ with b denoting the asymptotic bias. Following (7) in Armstrong et al. (2020), we leverage the oracle knowledge of $|\mathrm{E}\hat{\beta} - \beta^*|$ and form the oracle bias-aware CI as

$$(\hat{\beta} - \chi, \hat{\beta} + \chi) \quad \text{with} \quad \chi = \hat{\mathrm{SE}}(\hat{\beta}) \cdot \sqrt{\mathrm{cv}_\alpha(|\mathrm{E}\hat{\beta} - \beta^*|^2/\hat{\mathrm{SE}}^2(\hat{\beta}))}, \tag{10}$$

where $\hat{\mathrm{SE}}(\hat{\beta})$ is the empirical standard error computed over 500 simulations and $\mathrm{cv}_\alpha(B^2)$ is the $1-\alpha$ quantile of the χ^2 distribution with 1 degree of freedom and noncentrality parameter B^2. In

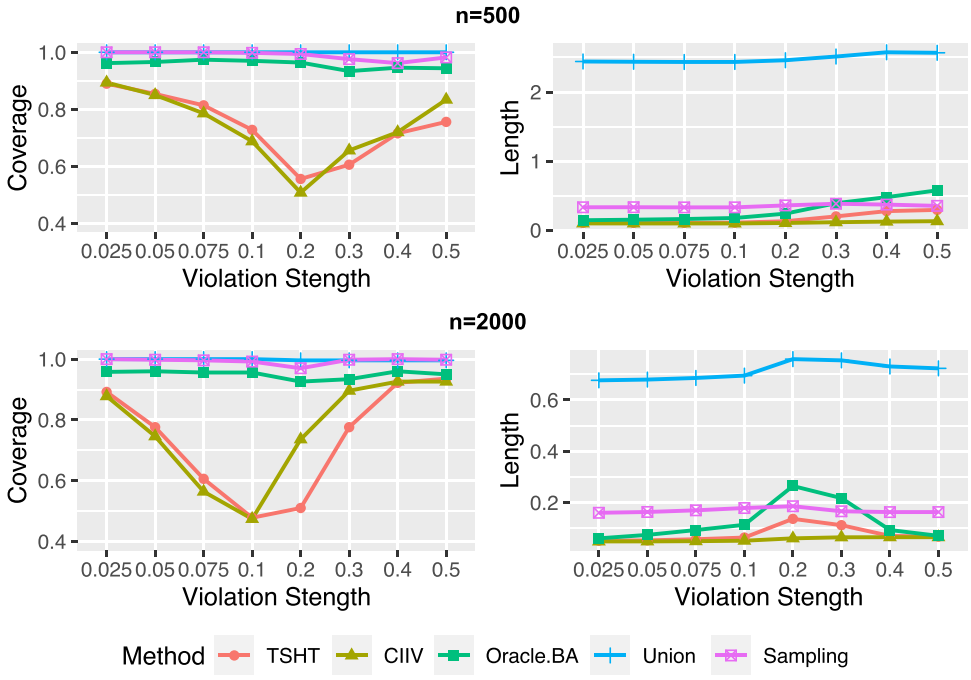

Figure 2. Empirical coverage and average lengths for Example 1 with τ varying across {0.025, 0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5}. `TSHT`, `CIIV`, and `Union` stand for the CIs by Guo et al. (2018), Windmeijer et al. (2021), and Kang et al. (2020), respectively. `OracleBA` stands for the oracle bias-aware CI in (10). `Sampling` represents our proposed sampling CI in Algorithm 3.

Figure 2, the oracle bias-aware CI achieves the desired coverage, and the sampling CI has a comparable length to the oracle bias-aware CI.

4 Searching and sampling: robust inference methods under majority rule

In this section, we focus on the majority rule setting and generalize the proposed methods to the plurality rule in Section 5. Our proposed procedure consists of two stages as an analogy to `TSLS`. We fit the treatment model in the first stage and select strong IVs; see Section 4.1. In the second stage, we propose novel searching and sampling CIs in Sections 4.2–4.4, which are robust to the mistakes in separating valid and invalid IVs.

4.1 First-stage selection and the finite-sample majority rule

Following Guo et al. (2018), we estimate $\mathcal{S}$ by applying the first-stage hard thresholding (Donoho & Johnstone, 1994) to the reduced-form estimator $\hat{\gamma}$ defined in (5),

$$\hat{\mathcal{S}} = \left\{ 1 \le j \le p_z : |\hat{\gamma}_j| \ge \sqrt{\log n} \cdot \sqrt{\hat{\mathrm{V}}^{\gamma}_{jj}/n} \right\}, \tag{11}$$

with $\hat{\mathrm{V}}^{\gamma}$ defined in (7). The main purpose of (11) is to estimate the set of relevant IVs and screen out IVs weakly associated with the treatment. The term $\sqrt{\log n}$ is introduced to adjust for the multiplicity of testing p_z hypothesis; see more discussions in Remark 1. The screening step in (11) guarantees the robustness of our proposal when some IVs are weakly associated with the treatment. With a high probability, our estimated set $\hat{\mathcal{S}}$ in (11) belongs to $\mathcal{S}$ and contains the set $\mathcal{S}_{\mathrm{str}}$ of

strongly relevant IVs, defined as

$$\mathcal{S}_{\text{str}} = \left\{1 \le j \le p_z : \left|\gamma_j^*\right| \ge 2\sqrt{\log n} \cdot \sqrt{\mathbf{V}_{jj}^{\gamma}/n}\right\} \quad \text{with } \mathbf{V}^{\gamma} \text{ defined in (6).} \tag{12}$$

The set $\mathcal{S}_{\text{str}}$ contains the j-th IV if its individual strength $|\gamma_j^*|$ is well above its estimation accuracy $\sqrt{\mathbf{V}_{jj}^{\gamma}/n}$. The factor 2 in (12) ensures that the individual IV strength is sufficiently large such that the j-th IV can be included in the set $\hat{\mathcal{S}}$ defined in (11); see more detailed discussion in Online Supplementary Material, Section A.2 in the supplement.

We now modify Condition 1 to accommodate the finite-sample uncertainty.

Condition 3 (Finite-sample Majority Rule). More than half of the relevant IVs are strongly relevant and valid, that is, $|\mathcal{V} \cap \mathcal{S}_{\text{str}}| > |\mathcal{S}|/2$, where $\mathcal{S}$ and $\mathcal{V}$ are defined in (3) and $\mathcal{S}_{\text{str}}$ is defined in (12).

When the sample size n is small, Condition 3 is slightly stronger than Condition 1. When $n \to \infty$ and the IV strengths $\{\gamma_j^*\}_{1 \le j \le p_z}$ do not change with n, the set $\mathcal{S}_{\text{str}}$ converges to $\mathcal{S}$ and Condition 3 is asymptotically the same as Condition 1.

Remark 1 There are other choices of the threshold level in (11). We can replace $\sqrt{\log n}$ in (11) and (12) by any $f(n)$ satisfying $f(n) \to \infty$ with $n \to \infty$ and $f(n) > \sqrt{2 \log p_z}$. Guo et al. (2018) used $f(n) = \sqrt{2.01 \log \max\{n, p_z\}}$, which is applicable in both low and high dimensions. In Online Supplementary Material, Section B.1 in the supplement, we consider the settings with weak IVs and demonstrate that our proposed methods have nearly the same performance with both the thresholds in (11) and that in Guo et al. (2018).

4.2 The searching confidence interval

In the following, we restrict our attention to the estimated set $\hat{\mathcal{S}}$ of relevant IVs and leverage Condition 3 to devise the searching CI for β^*. The main idea is to search for β values that lead to the majority of the instruments being detected as valid. The searching idea is highly relevant to the Anderson–Rubin test (Anderson & Rubin, 1949) for the weak IV problem, which searches for a range of β values leading to a sufficiently small χ^2 test statistic; see Remark 2 for more discussions.

For any given $\beta \in \mathbb{R}$, we apply the relation $\Gamma_j^* = \beta^* \gamma_j^* + \pi_j^*$ and construct the initial estimator of π_j^* as $\hat{\Gamma}_j - \beta \hat{\gamma}_j$. The estimation error of $\hat{\Gamma}_j - \beta \hat{\gamma}_j$ is decomposed as

$$(\hat{\Gamma}_j - \beta \hat{\gamma}_j) - \pi_j^* = \hat{\Gamma}_j - \Gamma_j^* - \beta(\hat{\gamma}_j - \gamma_j^*) + (\beta^* - \beta)\gamma_j^*. \tag{13}$$

When $\beta = \beta^*$, the above estimation error is further simplified as $\hat{\Gamma}_j - \Gamma_j^* - \beta(\hat{\gamma}_j - \gamma_j^*)$. We quantify the uncertainty of $\{\hat{\Gamma}_j - \Gamma_j^* - \beta(\hat{\gamma}_j - \gamma_j^*)\}_{j \in \hat{\mathcal{S}}}$ by the following union bound,

$$\liminf_{n \to \infty} \mathbb{P}\left[\max_{j \in \hat{\mathcal{S}}} \frac{|\hat{\Gamma}_j - \Gamma_j^* - \beta(\hat{\gamma}_j - \gamma_j^*)|}{\sqrt{(\hat{\mathbf{V}}_{jj}^{\Gamma} + \beta^2 \hat{\mathbf{V}}_{jj}^{\gamma} - 2\beta \hat{\mathbf{C}}_{jj})/n}} \le \Phi^{-1}\left(1 - \frac{\alpha}{2|\hat{\mathcal{S}}|}\right)\right] \ge 1 - \alpha, \tag{14}$$

where $\alpha \in (0, 1)$ and Φ^{-1} is the inverse CDF of the standard normal distribution. By rescaling (14), we construct the following threshold for $\hat{\Gamma}_j - \Gamma_j^* - \beta(\hat{\gamma}_j - \gamma_j^*)$ with $j \in \hat{\mathcal{S}}$,

$$\hat{\rho}_j(\beta) := \Phi^{-1}\left(1 - \frac{\alpha}{2|\hat{\mathcal{S}}|}\right) \cdot \sqrt{(\hat{\mathbf{V}}_{jj}^{\Gamma} + \beta^2 \hat{\mathbf{V}}_{jj}^{\gamma} - 2\beta \hat{\mathbf{C}}_{jj})/n}. \tag{15}$$

We further apply the hard thresholding to $\{\hat{\Gamma}_j - \beta\hat{\gamma}_j\}_{j\in\hat{\mathcal{S}}}$ and estimate π_j^* by

$$\hat{\pi}_j(\beta) = \left(\hat{\Gamma}_j - \beta\hat{\gamma}_j\right) \cdot \mathbf{1}\left(\left|\hat{\Gamma}_j - \beta\hat{\gamma}_j\right| \geq \hat{\rho}_j(\beta)\right) \quad \text{for} \quad j \in \hat{\mathcal{S}}. \tag{16}$$

If $\beta = \beta^*$, the hard thresholding in (16) guarantees $\hat{\pi}_j(\beta^*) = 0$ for $j \in \mathcal{V}$ with probability larger than $1 - \alpha$. For any $\beta \in \mathbb{R}$, we can construct the vector $\hat{\pi}_{\hat{\mathcal{S}}}(\beta) = (\hat{\pi}_j(\beta))_{j\in\hat{\mathcal{S}}}$ and calculate the number of nonzero entries in $\hat{\pi}_{\hat{\mathcal{S}}}(\beta)$, denoted as $\|\hat{\pi}_{\hat{\mathcal{S}}}(\beta)\|_0$. Due to the hard thresholding, $\hat{\pi}_{\hat{\mathcal{S}}}(\beta)$ can be a sparse vector and $\|\hat{\pi}_{\hat{\mathcal{S}}}(\beta)\|_0$ stands for the number of invalid IVs corresponding to the given β value.

We search for a range of β such that the number of invalid IVs is below $|\hat{\mathcal{S}}|/2$,

$$\mathrm{CI}^{\mathrm{sear}} := \left\{\beta \in \mathbb{R} : \|\hat{\pi}_{\hat{\mathcal{S}}}(\beta)\|_0 < |\hat{\mathcal{S}}|/2\right\}. \tag{17}$$

The above searching CI is valid since Condition 3 implies that $\|\hat{\pi}_{\hat{\mathcal{S}}}(\beta^*)\|_0 < |\hat{\mathcal{S}}|/2$ with probability $1 - \alpha$; see Online Supplementary Material, equation (56) in the supplement. We consider the following example and illustrate the construction of $\mathrm{CI}^{\mathrm{sear}}$ in Figure 3.

Example 2 Generate the models (1) and (2) with no baseline covariates, set $\beta^* = 1$, $n = 2{,}000$, $\gamma_j^* = 0.5 \cdot \mathbf{1}_{10}$ and $\pi^* = (\mathbf{0}_6^\top, 0.05, 0.05, -0.5, -1)^\top$. In Figure 3, we plot $|\hat{\mathcal{S}}| - \|\hat{\pi}_{\hat{\mathcal{S}}}(\beta)\|_0$ over different β and the red interval (0.914, 1.112) covers $\beta^* = 1$.

Remark 2 The proposed searching idea is related to but different from the Anderson–Rubin test (Anderson & Rubin, 1949) for the weak IV problem. In (17), we use the sparsity as the test statistic and invert the sparsity level to determine the confidence region for β^*. Our inverted test statistics $\|\hat{\pi}_{\hat{\mathcal{S}}}(\beta)\|_0$ is a discrete function of β, and its limiting distribution is not standard, which is fundamentally different from the χ^2 statistics used in AR test; see Figure 3. Due to the discreteness of the test statistics, the searching interval in (17) can be a union of disjoint intervals.

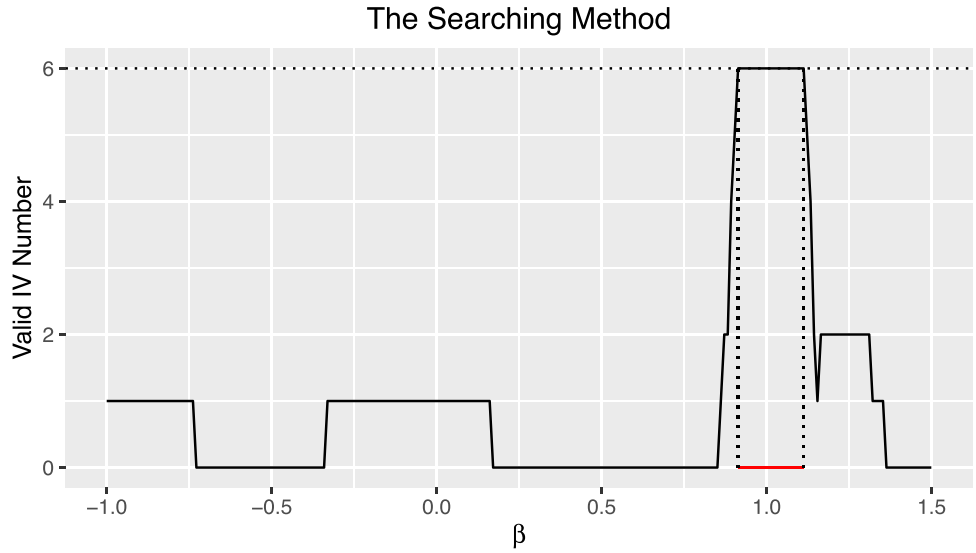

Figure 3. The x-axis plots a range of β values, and the y-axis plots the number of valid IVs (i.e., $|\hat{\mathcal{S}}| - \|\hat{\pi}_{\hat{\mathcal{S}}}(\beta)\|_0$) for every given β. The red interval (0.914, 1.112) denotes $\mathrm{CI}^{\mathrm{sear}}$ in (17).

4.3 Efficient implementation of the searching CI

We propose a computationally efficient implementation of CI^{sear} in (17). To implement (17), we can enumerate all values of β and calculate $\|\hat{\pi}_{\hat{\mathcal{S}}}(\beta)\|_0$. However, the enumeration method for constructing CI^{sear} can be time-consuming if there is a huge amount of β values. To improve the computational efficiency, we construct an initial set $[L, U]$ such that $\mathbb{P}(\beta^* \in [L, U]) \to 1$ and construct a grid set with the grid size n^{-a} for $a > 1/2$. Then, we discretize $[L, U]$ into the following grid set,

$$\mathcal{B} = \{\beta_1, \beta_2, \ldots, \beta_K\} \quad \text{with } \beta_1 = L,\ \beta_K = U,\ |\beta_{j+1} - \beta_j| = n^{-a} \quad \text{for} \quad 1 \le j \le K-2, \tag{18}$$

and $|\beta_K - \beta_{K-1}| \le n^{-a}$. The reason for requiring $a > 1/2$ is to ensure that the approximation error due to interval discretization is smaller than the parametric rate $n^{-1/2}$. Importantly, the constructed confidence interval is almost invariant to the choices of $[L, U]$ and the grid size n^{-a} as long as $[L, U]$ contains β^* and $a > 1/2$.

Throughout the paper, we choose the default grid size $n^{-0.6}$ and present the following default construction of the initial range $[L, U]$. For $j \in \hat{\mathcal{S}}$, we estimate β^* by the ratio $\hat{\Gamma}_j/\hat{\gamma}_j$ and estimate its variance by $\widehat{\text{Var}}(\hat{\Gamma}_j/\hat{\gamma}_j) = \frac{1}{n}(\hat{\mathbf{V}}^{\Gamma}_{jj}/\hat{\gamma}_j^2 + \hat{\mathbf{V}}^{\gamma}_{jj}\hat{\Gamma}_j^2/\hat{\gamma}_j^4 - 2\hat{\mathbf{C}}_{jj}\hat{\Gamma}_j/\hat{\gamma}_j^3)$. We then construct L and U as

$$L = \min_{j\in\hat{\mathcal{S}}}\left\{\hat{\Gamma}_j/\hat{\gamma}_j - \sqrt{\log n \cdot \widehat{\text{Var}}\left(\hat{\Gamma}_j/\hat{\gamma}_j\right)}\right\}, \quad U = \max_{j\in\hat{\mathcal{S}}}\left\{\hat{\Gamma}_j/\hat{\gamma}_j + \sqrt{\log n \cdot \widehat{\text{Var}}\left(\hat{\Gamma}_j/\hat{\gamma}_j\right)}\right\}, \tag{19}$$

where $\sqrt{\log n}$ is used to adjust for multiplicity. This initial range has been discussed in Section 3 of Windmeijer et al. (2021).

With the grid set defined in (18), we modify (17) and propose the following computationally efficient implementation,

$$\hat{\text{CI}}^{\text{sear}} = \left[\min_{\{\beta\in\mathcal{B}:\|\hat{\pi}_{\hat{\mathcal{S}}}(\beta)\|_0 < |\hat{\mathcal{S}}|/2\}} \beta,\ \max_{\{\beta\in\mathcal{B}:\|\hat{\pi}_{\hat{\mathcal{S}}}(\beta)\|_0 < |\hat{\mathcal{S}}|/2\}} \beta\right]. \tag{20}$$

Instead of searching over the real line as in (17), we restrict to the grid set $\mathcal{B}$ defined in (18) and (19) and search for the smallest and largest grid value such that more than half of the relevant IVs are valid. In contrast to CI^{sear} in (17), $\hat{\text{CI}}^{\text{sear}}$ is guaranteed to be an interval. In Online Supplementary Material, Section A.6 in the supplement, we compare $\hat{\text{CI}}^{\text{sear}}$ and CI^{sear} and find that both intervals guarantee the coverage properties and achieve similar lengths.

We summarize the construction of searching CI in Algorithm 1.

When the majority rule is violated, there is no β such that $\|\hat{\pi}_{\hat{\mathcal{S}}}(\beta)\|_0 < |\hat{\mathcal{S}}|/2$ and $\hat{\text{CI}}^{\text{sear}}$ in (20) are empty. This indicates that the majority rule is violated, which can be used as a partial check of the majority rule. Moreover, Algorithm 1 can be implemented with the summary statistics $\hat{\Gamma}$, $\hat{\gamma}$ and $\hat{\mathbf{V}}^{\Gamma}$, $\hat{\mathbf{V}}^{\gamma}$, $\hat{\mathbf{C}}$, which are the inputs for steps 2–6 of Algorithm 1.

Algorithm 1 Searching CI (Uniform Inference under Majority Rule)

Input: Outcome $Y \in \mathbb{R}^n$; Treatment $D \in \mathbb{R}^n$; IVs $Z \in \mathbb{R}^{n\times p_z}$; Covariates $X \in \mathbb{R}^{n\times p_x}$; Significance level $\alpha \in (0, 1)$.

Output: Confidence interval $\hat{\text{CI}}^{\text{sear}}$; Majority rule check $R \in \{0, 1\}$.

1: Construct $\hat{\Gamma} \in \mathbb{R}^{p_z}$, $\hat{\gamma} \in \mathbb{R}^{p_z}$ as in (5) and $\hat{\mathbf{V}}^{\Gamma}$, $\hat{\mathbf{V}}^{\gamma}$ and $\hat{\mathbf{C}}$ as in (7);

2: Construct $\hat{\mathcal{S}}$ as in (11); ▷ First-stage selection

3: Construct L and U as in (19);

4: Construct the grid set $\mathcal{B}$ as in (18) with $a = 0.6$;

5: Construct $\{\hat{\pi}_j(\beta)\}_{j\in\hat{\mathcal{S}},\beta\in\mathcal{B}}$ as in (16);

6: Construct $\hat{\text{CI}}^{\text{sear}}$ as in (20) and set $R = \mathbf{1}(\hat{\text{CI}}^{\text{sear}} \neq \emptyset)$. ▷ Searching CI

4.4 The sampling CI

In this subsection, we devise a novel sampling method to improve the precision of the searching CI. Conditioning on the observed data, we resample $\{\hat{\Gamma}^{[m]}, \hat{\gamma}[m]\}_{1\le m\le M}$ as

$$\begin{pmatrix}\hat{\Gamma}^{[m]}\\ \hat{\gamma}^{[m]}\end{pmatrix} \overset{\text{i.i.d.}}{\sim} N\left[\begin{pmatrix}\hat{\Gamma}\\ \hat{\gamma}\end{pmatrix}, \begin{pmatrix}\hat{\mathbf{V}}^{\Gamma}/n & \hat{\mathbf{C}}/n\\ \hat{\mathbf{C}}^{\top}/n & \hat{\mathbf{V}}^{\gamma}/n\end{pmatrix}\right] \quad \text{for} \quad 1\le m\le M, \tag{21}$$

where M denotes the resampling size (with the default value 1,000), $\hat{\gamma}$ and $\hat{\Gamma}$ are defined in (5), and $\hat{\mathbf{V}}^{\Gamma}$, $\hat{\mathbf{C}}$, and $\hat{\mathbf{V}}^{\gamma}$ are defined in (7).

For $1\le m\le M$, we use $\{\hat{\Gamma}^{[m]}, \hat{\gamma}^{[m]}\}$ to replace $\{\hat{\Gamma}, \hat{\gamma}\}$ and implement Algorithm 1 to construct a searching CI. We refer to this CI as the m-th sampled searching CI. In implementing Algorithm 1, we decrease the thresholding level used in the hard thresholding step of estimating π^*. Consequently, each sampled searching CI can be much shorter than the original searching CI; see Figure 4. Our proposal of decreasing the threshold relies on the following observation: there exists $1\le m^*\le M$ such that

$$\max_{j\in\hat{\mathcal{S}}}\frac{|\hat{\Gamma}_j^{[m^*]}-\Gamma_j^*-\beta(\hat{\gamma}_j^{[m^*]}-\gamma_j^*)|}{\sqrt{(\hat{\mathbf{V}}_{jj}^{\Gamma}+\beta^2\hat{\mathbf{V}}_{jj}^{\gamma}-2\beta\hat{\mathbf{C}}_{jj})/n}}\le\lambda\cdot\Phi^{-1}\left(1-\frac{\alpha}{2|\hat{\mathcal{S}}|}\right)\quad\text{with}\quad\lambda\asymp\left(\frac{\log n}{M}\right)^{\frac{1}{2|\hat{\mathcal{S}}|}}. \tag{22}$$

The rigorous statement is provided in Proposition 1. Compared with (14), the upper bound in (22) is shrunk by a factor λ, which is close to zero with a large M (e.g., $M=1{,}000$). If we had access to $\{\hat{\Gamma}^{[m^*]}, \hat{\gamma}^{[m^*]}\}$, we could use a much smaller threshold level in constructing the searching CI.

We now provide the complete details about constructing the sampling CI. We firstly consider that the tuning parameter λ in (22) is given and will present a data-dependent way of choosing λ in Remark 3. Motivated by (22), we multiply the threshold level by λ and implement Algorithm 1 for each of $\{\hat{\Gamma}^{[m]}, \hat{\gamma}^{[m]}\}_{1\le m\le M}$. That is, for $1\le m\le M$, we modify (16) and estimate $\pi^*_{\hat{\mathcal{S}}}$ by

$$\hat{\pi}_j^{[m]}(\beta,\lambda)=\left(\hat{\Gamma}_j^{[m]}-\beta\hat{\gamma}_j^{[m]}\right)\cdot\mathbf{1}\left(\left|\hat{\Gamma}_j^{[m]}-\beta\hat{\gamma}_j^{[m]}\right|\ge\lambda\cdot\hat{\rho}_j(\beta)\right)\quad\text{for}\quad j\in\hat{\mathcal{S}}, \tag{23}$$

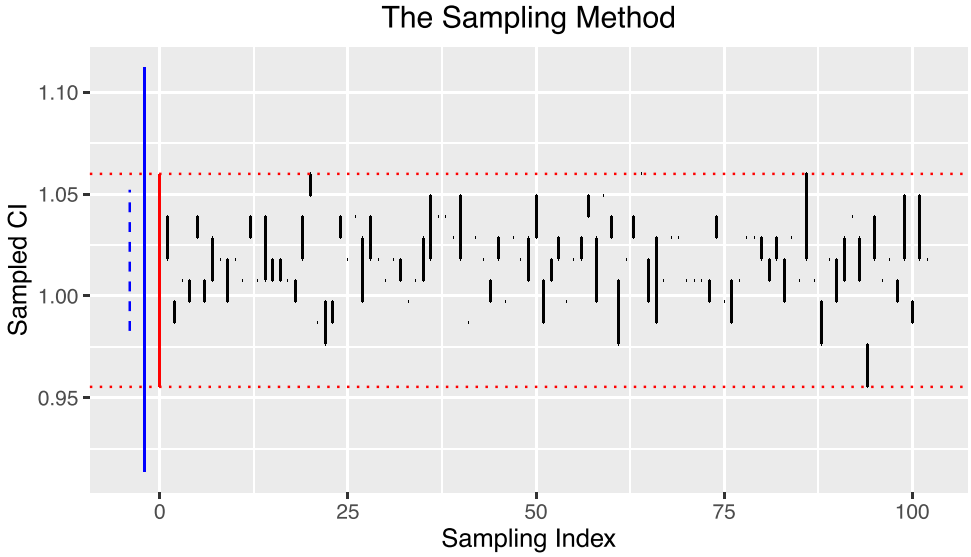

Figure 4. The axis corresponds to sampling indexes {1, 2, …, 102} (after re-ordering), and the y-axis reports the sampled CIs. The red interval is $\text{CI}^{\text{samp}} = (0.955, 1.060)$, the blue interval is $\hat{\text{CI}}^{\text{sear}} = (0.914, 1.112)$, and the blue dashed interval is the oracle CI (0.983, 1.052) with prior information on valid IVs.

with $\hat{\rho}_j(\beta)$ defined in (15). We apply (20) and construct the m-th sampled searching CI as $[\beta^{[m]}_{\min}(\lambda), \beta^{[m]}_{\max}(\lambda)]$ with

$$\beta^{[m]}_{\min}(\lambda) = \min_{\{\beta\in\mathcal{B}:\|\hat{\pi}^{[m]}_{\hat{S}}(\beta,\lambda)\|_0<|\hat{S}|/2\}} \beta \quad \text{and} \quad \beta^{[m]}_{\max}(\lambda) = \max_{\{\beta\in\mathcal{B}:\|\hat{\pi}^{[m]}_{\hat{S}}(\beta,\lambda)\|_0<|\hat{S}|/2\}} \beta. \tag{24}$$

Similarly to (20), we set $[\beta^{[m]}_{\min}(\lambda), \beta^{[m]}_{\max}(\lambda)] = \emptyset$ if there is no β such that $\|\hat{\pi}^{[m]}(\beta, \lambda)\|_0 < |\hat{S}|/2$. We aggregate the M searching CIs and propose the following sampling CI:

$$\text{CI}^{\text{samp}} = \left[\min_{m\in\mathcal{M}} \beta^{[m]}_{\min}(\lambda),\ \max_{m\in\mathcal{M}} \beta^{[m]}_{\max}(\lambda)\right], \tag{25}$$

where the index set $\mathcal{M} = \{1 \le m \le M : [\beta^{[m]}_{\min}(\lambda), \beta^{[m]}_{\max}(\lambda)] \neq \emptyset\}$ contains all indexes m corresponding to nonempty $[\beta^{[m]}_{\min}(\lambda), \beta^{[m]}_{\max}(\lambda)]$. In Figure 4, we demonstrate the sampling CI in (25) using Example 2. Many of the M sampled searching CIs are empty, and the nonempty ones can be much shorter than the searching CI. Consequently, the sampling CI (in red) is much shorter than the searching CI (in blue).

An important reason for the sampling method improving the precision is that the test statistics $\|\hat{\pi}_{\hat{S}}(\beta)\|_0$ used for constructing the searching CI in (20) is a discrete function of β, and the limiting distribution of the test statistics is not standard.

We summarize the sampling CI in Algorithm 2 and will provide two remarks about the implementation of Algorithm 2.

Algorithm 2 Sampling CI (Uniform Inference under Majority Rule)

Input: Outcome $Y \in \mathbb{R}^n$; Treatment $D \in \mathbb{R}^n$; IVs $Z \in \mathbb{R}^{n\times p_z}$; Covariates $X \in \mathbb{R}^{n\times p_x}$; Sampling number $M = 1{,}000$; $\lambda = c_*(\log n/M)^{1/(2|\hat{S}|)}$; Significance level $\alpha \in (0, 1)$

Output: Confidence interval CI^{samp}

1: Implement steps 1 to 4 as in Algorithm 1;
2: **for** $m \leftarrow 1$ to M **do**
3: Sample $\hat{\Gamma}^{[m]}$ and $\hat{\gamma}^{[m]}$ as in (21);
4: Compute $\{\hat{\pi}^{[m]}_j(\beta, \lambda)\}_{j\in\hat{S},\beta\in\mathcal{B}}$ as in (23);
5: Compute $\beta^{[m]}_{\min}(\lambda)$ and $\beta^{[m]}_{\max}(\lambda)$ in (24);
6: **end for**
7: Construct CI^{samp} as in (25) ▷ Sampling CI

Remark 3 (Tuning parameter selection for sampling). We demonstrate in Online Supplementary Material, Section B.2 in the supplement that the sampling CIs in Algorithm 2 do not materially change with different choices of L, U, a, and the resampling number M. Theorem 2 suggests the form of the tuning parameter λ as $c_*(\log n/M)^{1/(2|\hat{S}|)}$. We present a data-dependent way to specify the constant c_*. If the λ value is too small, very few of the resampled reduced-form estimators will pass the majority rule and most of the $M = 1{,}000$ sampled intervals will be empty. Hence, the proportion of the nonempty intervals indicates whether λ is large enough. We start with a small value $\lambda = 1/6 \cdot (\log n/M)^{1/(2|\hat{S}|)}$ and increase the value of λ by a factor of 1.25 until more than $\texttt{prop} = 10\%$ of the $M = 1{,}000$ intervals are nonempty. We choose the smallest λ value achieving this and use this λ value to implement Algorithm 2.

Remark 4 (Alternative aggregation). We may combine the sampled CIs as $\mathrm{CI}_0^{\mathrm{samp}} = \cup_{m=1}^{M} [\beta_{\min}^{[m]}(\lambda), \beta_{\max}^{[m]}(\lambda)]$. Since $\mathrm{CI}_0^{\mathrm{samp}}$ may not be an interval due to its definition, we focus on $\mathrm{CI}^{\mathrm{samp}}$ defined in (25). In addition, we may filter out $\{\hat{\gamma}^{[m]}, \hat{\Gamma}^{[m]}\}_{1\le m\le M}$ near the boundary of the sampling distribution in (21). Particularly, define

$$\mathcal{M}_0 = \left\{1 \le m \le M : \max_{j\in\hat{\mathcal{S}}} \max\left\{\left|\hat{\gamma}_j^{[m]} - \hat{\gamma}_j\right| / \sqrt{\hat{\mathbf{V}}_{jj}^{\gamma}/n}, \left|\hat{\Gamma}_j^{[m]} - \hat{\Gamma}_j\right| / \sqrt{\hat{\mathbf{V}}_{jj}^{\Gamma}/n}\right\} \le 1.1\Phi^{-1}\left(1 - \frac{\alpha_0}{4|\hat{\mathcal{S}}|}\right)\right\}$$

with $\alpha_0 = 0.05$. We modify (25) and construct the sampling CI as

$$\mathrm{CI}^{\mathrm{samp}}(\mathcal{M}_0) = \left[\min_{m\in\mathcal{M}'} \beta_{\min}^{[m]}(\lambda), \max_{m\in\mathcal{M}'} \beta_{\max}^{[m]}(\lambda)\right], \tag{26}$$

with the index set $\mathcal{M}' = \{m \in \mathcal{M}_0 : [\beta_{\min}^{[m]}(\lambda), \beta_{\max}^{[m]}(\lambda)] \neq \emptyset\}$. In Online Supplementary Material, Table B.6 in the supplement, we show that $\mathrm{CI}^{\mathrm{samp}}(\mathcal{M}_0)$ is nearly the same as $\mathrm{CI}^{\mathrm{samp}}$ in (25).

4.5 Robustness to the existence of weak IVs

The validity of our proposed CIs in Sections 4.2–4.4 relies on Condition 3, which requires more than half of the IVs to be strongly relevant and valid but allows the remaining IVs to be arbitrarily weak and invalid. In the following, we make two important remarks about weak IVs. Firstly, our focused setting is completely different from the classical weak IV framework (e.g., Staiger & Stock, 1997), where all IVs as a total are weak. The F test for the treatment model and the concentration parameters can provide evidence on whether the IVs are weak (e.g., Stock et al., 2002). Similarly, the estimated set $\hat{\mathcal{S}}$ in (11) can also be used to check whether the given data set falls into the classical weak IV framework. This paper focuses on the regime with nonempty $\hat{\mathcal{S}}$, indicating the existence of strong IVs. Secondly, the challenge due to weak IVs is fundamentally different from locally invalid IVs. In the first-stage selection, the set of relatively weak IVs $\mathcal{S}\backslash\mathcal{S}_{\mathrm{str}}$ is not guaranteed to be selected by $\hat{\mathcal{S}}$ in (11). The uncertainty of selecting IVs inside $\mathcal{S}\backslash\mathcal{S}_{\mathrm{str}}$ is of a different nature from the uncertainty of detecting locally invalid IVs. In Online Supplementary Material, Section B.1 in the supplement, we demonstrate that both `TSHT` and `CIIV` perform well even if there exists uncertainty of selecting IVs belonging to $\mathcal{S}\backslash\mathcal{S}_{\mathrm{str}}$.

5 Uniform inference methods under plurality rule

This section considers the more challenging setting with the plurality rule. We introduce the finite-sample plurality rule and then generalize the method proposed in Section 4. For a given invalidity level $\nu \in \mathbb{R}$, we define the set of IVs,

$$\mathcal{I}(\nu, \tau) = \left\{j \in \mathcal{S} : \left|\pi_j^*/\gamma_j^* - \nu\right| \le \tau\right\} \quad \text{with} \quad \tau \in \mathbb{R}. \tag{27}$$

When τ is small, $\mathcal{I}(\nu, \tau)$ denotes the set of IVs with the invalidity level around ν. With $\mathbf{T}_{j,k}$ defined in (9), we define the separation level as

$$\mathrm{sep}(n) := 2\sqrt{\log n} \max_{j,k\in\hat{\mathcal{S}}} \mathbf{T}_{j,k}. \tag{28}$$

Note that $\beta^{[j]} = \Gamma_j^*/\gamma_j^* = \beta^* + \pi_j^*/\gamma_j^*$. If the pairwise difference $\beta^{[j]} - \beta^{[k]}$ for $j, k \in \hat{\mathcal{S}}$ is above $\mathrm{sep}(n)$, the j-th and k-th IVs can be separated based on their invalidity levels. Particularly, the term $\max_{j,k\in\hat{\mathcal{S}}} \mathbf{T}_{j,k}$ denotes the largest error of estimating the pairwise difference $\{\beta^{[j]} - \beta^{[k]}\}_{j,k\in\hat{\mathcal{S}}}$ and $\sqrt{\log n}$ is used to adjust for the multiplicity of hypothesis testing; see more discussion after Proposition 2. When $\{\gamma_j^*\}_{j\in\hat{\mathcal{S}}}$ are nonzero constants, $\mathrm{sep}(n)$ is of order $\sqrt{\log n/n}$.

With the separation level $\text{sep}(n)$, we introduce the finite-sample plurality rule.

Condition 4 (Finite-sample Plurality Rule). For $\tau_n = 3\text{sep}(n)$,

$$|\mathcal{V} \cap \mathcal{S}_{\text{str}}| > \max_{v \in \mathbb{R}} |\mathcal{I}(v, \tau_n) \backslash \mathcal{V}|, \tag{29}$$

where $\text{sep}(n)$, $\mathcal{V}$, $\mathcal{S}_{\text{str}}$, and $\mathcal{I}(v, \tau_n)$ are defined in (28), (3), (12), and (27), respectively.

The set $\mathcal{V} \cap \mathcal{S}_{\text{str}}$ consists of the strongly relevant and valid IVs, which we rely on to make inferences for β^*. Since $\tau_n \to 0$ as $n \to \infty$, the set $\mathcal{I}(v, \tau_n) \backslash \mathcal{V}$ contains all invalid IVs with invalidity levels $\pi_j^*/\gamma_j^* \approx v$. Condition 4 requires that the cardinality of $\mathcal{V} \cap \mathcal{S}_{\text{str}}$ is larger than that of invalid IVs with $\pi_j^*/\gamma_j^* \approx v$. When $v = 0$, the set $\mathcal{I}(0, \tau_n) \backslash \mathcal{V}$ is the set of invalid IVs with invalidity levels below τ_n. Condition 4 also requires that there are more IVs in $\mathcal{V} \cap \mathcal{S}_{\text{str}}$ than the locally invalid IVs with invalidity levels below τ_n.

In comparison to Condition 2, Condition 4 accommodates for the finite-sample approximation error by grouping together invalid IVs with similar invalidity levels. For a large sample size, Condition 4 is the same as Condition 2. Specifically, if $n \to \infty$ and $\{\pi_j^*\}_{1 \le j \le p_z}$ and $\{\gamma_j^*\}_{1 \le j \le p_z}$ do not change with n, then

$$\lim_{n \to \infty} |\mathcal{V} \cap \mathcal{S}_{\text{str}}| = |\mathcal{V}|, \quad \lim_{n \to \infty} |\mathcal{I}(v, \tau_n) \backslash \mathcal{V}| = |\mathcal{I}_v \backslash \mathcal{V}| = \begin{cases} 0 & \text{if } v = 0 \\ |\mathcal{I}_v| & \text{if } v \neq 0 \end{cases}$$

with $\mathcal{I}_v$ defined in Condition 2. We shall emphasize that $\text{sep}(n)$ defined in (28) is not required in the following confidence interval construction.

We propose a two-step inference procedure for β^* under Condition 4. In the first step, we construct an initial set $\hat{\mathcal{V}}$ satisfying

$$\mathcal{V} \cap \mathcal{S}_{\text{str}} \subset \hat{\mathcal{V}} \subset \mathcal{I}(0, \tau_n) \quad \text{with} \quad \tau_n = 3\text{sep}(n). \tag{30}$$

Since Condition 4 implies that $\mathcal{V} \cap \mathcal{S}_{\text{str}}$ is the majority of the set $\mathcal{I}(0, \tau_n)$, we apply (30) and establish that the set $\mathcal{V} \cap \mathcal{S}_{\text{str}}$ becomes the majority of the initial set $\hat{\mathcal{V}}$. In the second step, we restrict our attention to $\hat{\mathcal{V}}$ and generalize the methods in Section 4. Importantly, the initial set $\hat{\mathcal{V}}$ is allowed to contain locally invalid IVs.

We now construct an initial set $\hat{\mathcal{V}}$ satisfying (30) by modifying `TSHT` (Guo et al., 2018). Without loss of generality, we set $\hat{\mathcal{S}} = \{1, 2, \ldots, |\hat{\mathcal{S}}|\}$. For any $j \in \hat{\mathcal{S}}$, we construct an estimator of β^* and π^* as

$$\hat{\beta}^{[j]} = \hat{\Gamma}_j / \hat{\gamma}_j \quad \text{and} \quad \hat{\pi}_k^{[j]} = \hat{\Gamma}_k - \hat{\beta}^{[j]} \hat{\gamma}_k \quad \text{for} \quad k \in \hat{\mathcal{S}}, \tag{31}$$

where the super index j stands for the model identification by assuming the j-th IV to be valid. We define $\hat{\mathbf{R}}^{[j]} = \hat{\mathbf{V}}^{\Gamma} + (\hat{\beta}^{[j]})^2 \hat{\mathbf{V}}^{\gamma} - 2\hat{\beta}^{[j]} \hat{\mathbf{C}}$ and further estimate the standard error of $\hat{\pi}_k^{[j]}$ with $k \in \hat{\mathcal{S}}$ by

$$\hat{\text{SE}}(\hat{\pi}_k^{[j]}) = \sqrt{\left(\hat{\mathbf{R}}_{k,k}^{[j]} + \left(\hat{\gamma}_k / \hat{\gamma}_j \right)^2 \hat{\mathbf{R}}_{j,j}^{[j]} - 2\hat{\gamma}_k / \hat{\gamma}_j \hat{\mathbf{R}}_{k,j}^{[j]} \right) / n}. \tag{32}$$

For $1 \le k, j \le |\hat{\mathcal{S}}|$, we apply the following hard thresholding and construct the (k, j) entry of the voting matrix $\hat{\Pi} \in \mathbb{R}^{|\hat{\mathcal{S}}| \times |\hat{\mathcal{S}}|}$ as

$$\hat{\Pi}_{k,j} = \mathbf{1}\left(|\hat{\pi}_k^{[j]}| \le \hat{\text{SE}}(\hat{\pi}_k^{[j]}) \cdot \sqrt{\log n} \quad \text{and} \quad |\hat{\pi}_j^{[k]}| \le \hat{\text{SE}}(\hat{\pi}_j^{[k]}) \cdot \sqrt{\log n} \right), \tag{33}$$

with $\hat{\pi}_k^{[j]}$ and $\hat{\pi}_j^{[k]}$ defined in (31), $\hat{\text{SE}}(\hat{\pi}_k^{[j]})$ and $\hat{\text{SE}}(\hat{\pi}_j^{[k]})$ defined in (32), and $\sqrt{\log n}$ used to adjust for multiplicity. In (33), $\hat{\Pi}_{k,j} = 1$ represents that the k-th and j-th IVs support each other to be valid while

$\hat{\Pi}_{k,j} = 0$ represents that they do not. The voting matrix in (33) is a symmetric version of the voting matrix proposed in Guo et al. (2018).

We now construct the initial set by leveraging the voting matrix in (33). Define $\hat{\mathcal{W}} = \arg\max_{1 \le j \le |\hat{\mathcal{S}}|} \|\hat{\Pi}_{j\cdot}\|_0$ as the set of IVs receiving the largest number of votes. We construct the following initial set,

$$\hat{\mathcal{V}}^{\texttt{TSHT}} := \{1 \le l \le |\hat{\mathcal{S}}| : \text{there exist } 1 \le k \le |\hat{\mathcal{S}}| \text{ and } j \in \hat{\mathcal{W}} \text{ such that } \hat{\Pi}_{j,k}\hat{\Pi}_{k,l} = 1\}. \tag{34}$$

In words, if the l-th IV from $\hat{\mathcal{S}}$ and the j-th IV from $\hat{\mathcal{W}}$ are claimed to be valid by any IV from $\hat{\mathcal{S}}$, then the l-th IV is also included in $\hat{\mathcal{V}}^{\texttt{TSHT}}$. We show in Proposition 2 that $\hat{\mathcal{V}}^{\texttt{TSHT}}$ is guaranteed to satisfy (30). Together with Condition 4, $\mathcal{V} \cap \mathcal{S}_{\text{str}}$ becomes the majority of the initial set $\hat{\mathcal{V}}^{\texttt{TSHT}}$. Then, we restrict to the set $\hat{\mathcal{V}}^{\texttt{TSHT}}$ and apply Algorithms 1 and 2 with $\hat{\mathcal{S}}$ replaced by $\hat{\mathcal{V}}^{\texttt{TSHT}}$.

We summarize our proposed searching and sampling CIs in Algorithm 3, with the tuning parameters selected in the same way as that in Remark 3. Algorithm 3 can be implemented without requiring the raw data, but with $\hat{\Gamma}$, $\hat{\gamma}$ and $\hat{\mathbf{V}}^{\Gamma}$, $\hat{\mathbf{V}}^{\gamma}$, $\hat{\mathbf{C}}$ as the inputs. We have demonstrated our method by constructing $\hat{\mathcal{V}} = \hat{\mathcal{V}}^{\texttt{TSHT}}$ as in (34), but Algorithm 3 can be applied with any $\hat{\mathcal{V}}$ satisfying (30).

Comparison with the CIIV method. The idea of searching has been developed in Windmeijer et al. (2021) to select valid IVs. We now follow Windmeijer et al. (2021) and sketch the intuitive idea of the `CIIV` method. For any grid value $\delta_g \in [L, U]$, define the set

$$\hat{\mathcal{V}}(\delta_g) = \left\{ j \in \mathcal{S} : \Gamma_j^*/\gamma_j^* = \delta_g \text{ is not rejected} \right\}.$$

Here, $\hat{\mathcal{V}}(\delta_g)$ denotes a subset of IVs such that the corresponding hypothesis $\Gamma_j^*/\gamma_j^* = \delta_g$ is not rejected. As explained in Section 3 of Windmeijer et al. (2021), the `CIIV` method examines all values of δ_g and selects the largest set $\hat{\mathcal{V}}(\delta_g)$ as the set of valid IVs, that is,

$$\hat{\mathcal{V}}^{\texttt{CIIV}} = \hat{\mathcal{V}}(\hat{\delta}_g) \quad \text{with} \quad \hat{\delta}_g = \underset{\delta_g \in [L,U]}{\arg\max} |\hat{\mathcal{V}}(\delta_g)|. \tag{35}$$

Our proposed searching CI differs from Windmeijer et al. (2021) since we directly construct CIs by searching for a range of suitable β values, while the `CIIV` method applies the searching idea to select the set of valid IVs.

We provide some intuitions on why our proposal is more robust to the IV selection error. We first screen out the strongly invalid IVs and construct an initial set estimator $\hat{\mathcal{V}} = \hat{\mathcal{V}}^{\texttt{TSHT}}$; then, we restrict to the set $\hat{\mathcal{V}}$ and apply searching and sampling CIs developed under the majority rule. The majority rule in the second step explains the robustness: we compare the number of votes to $|\hat{\mathcal{V}}|/2$, which is fixed after computing $\hat{\mathcal{V}}$. However, the optimization in (35) chooses δ_g, giving the largest number of votes, which can be more vulnerable to selection/testing errors. The validity of the `CIIV` method requires that $\hat{\mathcal{V}}^{\texttt{CIIV}}$ does not contain any invalid IVs; in contrast, our method is still effective even if the initial set $\hat{\mathcal{V}}$ contains the invalid IVs but satisfies (30).

6 Theoretical justification

We focus on the low-dimensional setting with heteroscedastic errors and introduce the following regularity conditions. We always consider the asymptotic expressions in the limit with $n \to \infty$.

(C1) For $1 \le i \le n$, $W_{i\cdot} = (X_{i\cdot}^\top, Z_{i\cdot}^\top)^\top \in \mathbb{R}^p$ are i.i.d. Sub-Gaussian random vectors with $\Sigma = \mathbf{E}(W_{i\cdot}(W_{i\cdot})^\top)$ satisfying $c_0 \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le C_0$ for some positive constants $C_0 \ge c_0 > 0$.

(C2) For $1 \le i \le n$, the errors $(\epsilon_i, \delta_i)^\top$ in (4) are i.i.d. Sub-Gaussian random vectors; the conditional covariance matrix satisfying $c_1 \le \lambda_{\min}(\mathbf{E}[(\epsilon_i, \delta_i)^\top(\epsilon_i, \delta_i) \mid W_{i\cdot}]) \le \lambda_{\max}(\mathbf{E}[(\epsilon_i, \delta_i)^\top(\epsilon_i, \delta_i) \mid W_{i\cdot}]) \le C_1$ for some positive constants $C_1 \ge c_1 > 0$.

Algorithm 3 Uniform Inference with Searching and Sampling (Plurality Rule)

Input: Outcome $Y \in \mathbb{R}^n$; Treatment $D \in \mathbb{R}^n$; IVs $Z \in \mathbb{R}^{n\times p_z}$; Covariates $X \in \mathbb{R}^{n\times p_x}$; Sampling number $M = 1{,}000$; $\lambda = c_*(\log n/M)^{1/(2|\hat{\mathcal{S}}|)}$; Significance level $\alpha \in (0, 1)$.

Output: Confidence intervals $\hat{\text{CI}}^{\text{sear}}$ and CI^{samp}; Plurality rule check R.

1: Construct $\hat{\Gamma} \in \mathbb{R}^{p_z}$, $\hat{\gamma} \in \mathbb{R}^{p_z}$ as in (5) and $\hat{\mathbf{V}}^{\Gamma}$, $\hat{\mathbf{V}}^{\gamma}$ and $\hat{\mathbf{C}}$ as in (7);

2: Construct $\hat{\mathcal{S}}$ as in (11);

3: Construct the voting matrix $\hat{\Pi} \in \mathbb{R}^{|\hat{\mathcal{S}}|\times|\hat{\mathcal{S}}|}$ as in (33);

4: Construct $\hat{\mathcal{V}} = \hat{\mathcal{V}}^{\text{TSHT}}$ as in (34); ▷ Construction of $\hat{\mathcal{V}}$

5: Construct L and U as in (19) with $\hat{\mathcal{S}} = \hat{\mathcal{V}}$;

6: Construct the grid set $\mathcal{B} \subset [L, U]$ as in (18) with the grid size $n^{-0.6}$;

7: Compute $\{\hat{\rho}_j(\beta)\}_{j\in\hat{\mathcal{V}},\beta\in\mathcal{B}}$ as in (15) with $\hat{\mathcal{S}} = \hat{\mathcal{V}}$;

8: Compute $\{\hat{\pi}_j(\beta)\}_{j\in\hat{\mathcal{V}},\beta\in\mathcal{B}}$ as in (16) with $\hat{\mathcal{S}} = \hat{\mathcal{V}}$;

9: Construct $\hat{\text{CI}}^{\text{sear}}$ as in (20) with $\hat{\mathcal{S}} = \hat{\mathcal{V}}$ and set $R = \mathbf{1}(\hat{\text{CI}}^{\text{sear}} \neq \emptyset)$; ▷ Searching CI

10: **for** $m \leftarrow 1$ to M **do**

11: Sample $\hat{\Gamma}^{[m]}$ and $\hat{\gamma}^{[m]}$ as in (21);

12: Compute $\{\hat{\pi}_j^{[m]}(\beta, \lambda)\}_{j\in\hat{\mathcal{V}},\beta\in\mathcal{B}}$ as in (23) with $\hat{\mathcal{S}} = \hat{\mathcal{V}}$;

13: Construct $\beta_{\min}^{[m]}(\lambda)$ and $\beta_{\max}^{[m]}(\lambda)$ as in (24) with $\hat{\mathcal{S}} = \hat{\mathcal{V}}$;

14: **end for**

15: Construct CI^{samp} as in (25). ▷ Sampling CI

Conditions (C1) and (C2) are imposed on the reduced-form model (4), which includes the outcome model (1) and the treatment model (2) as a special case. We assume that the covariance matrix of $W_{i\cdot}$ is well conditioned and also the covariance matrix of the errors is well conditioned. Condition (C2) in general holds if e_i in (1) and δ_i in (2) are not perfectly correlated. Our setting allows for the heteroscedastic errors. If we further assume $(\epsilon_i, \delta_i)^\top$ to be independent of $W_{i\cdot}$, Condition (C2) assumes the covariance matrix of $(\epsilon_i, \delta_i)^\top$ to be well conditioned. As a remark, the Sub-Gaussian conditions on both $W_{i\cdot}$ and the errors might be relaxed to the moment conditions in low dimensions.

We start with the majority rule setting and will move to the plurality rule. The following theorem justifies the searching CI under the majority rule.

Theorem 1 Consider the model (4). Suppose that Condition 3, Conditions (C1) and (C2) hold, and $\alpha \in (0, 1/4)$ is the significance level. Then, CI^{sear} defined in (17) and $\hat{\text{CI}}^{\text{sear}}$ defined in (20) satisfy

$$\liminf_{n\to\infty} \mathbb{P}\left(\beta^* \in \text{CI}^{\text{sear}}\right) \geq 1 - \alpha \quad \text{and} \quad \liminf_{n\to\infty} \mathbb{P}\left(\beta^* \in \hat{\text{CI}}^{\text{sear}}\right) \geq 1 - \alpha.$$

There exists a positive constant $C > 0$ such that

$$\liminf_{n\to\infty} \mathbb{P}\left(\max\left\{\mathbf{L}(\text{CI}^{\text{sear}}), \mathbf{L}(\hat{\text{CI}}^{\text{sear}})\right\} \leq \frac{C}{\min_{j\in\hat{\mathcal{S}}\cap\mathcal{V}} |\gamma_j^*| \cdot \sqrt{n}}\right) \geq 1 - \alpha,$$

where $\mathbf{L}(\cdot)$ denotes the interval length.

If $\min_{j\in\hat{\mathcal{S}}\cap\mathcal{V}} |\gamma_j^*|$ is of a constant order, Theorem 1 implies that the length of the searching CI is of the parametric rate $1/\sqrt{n}$. The searching CI achieves the desired coverage level without relying on a perfection separation of valid and invalid IVs, which brings in a sharp contrast to the well-separation conditions required in TSHT (Guo et al., 2018) and CIIV (Windmeijer et al., 2021).

We now turn to the sampling CI. Before presenting the theory for the sampling CI, we justify in Proposition 1 why we are able to decrease the thresholding level in constructing the sampling CI. For $\alpha_0 \in (0, 1/4)$, define the positive constant

$$c^*(\alpha_0) = \frac{1}{[3\pi \cdot \lambda_{\min}(\mathrm{Cov})]^{|\hat{\mathcal{S}}|}} \exp\left(-\frac{|\hat{\mathcal{S}}| \cdot 3\lambda_{\max}(\mathrm{Cov})}{\lambda_{\min}(\mathrm{Cov})} \cdot \left[\Phi^{-1}\left(1 - \frac{\alpha_0}{4|\hat{\mathcal{S}}|}\right)\right]^2 \right), \tag{36}$$

with Cov defined in (6) and Φ^{-1} denoting the inverse CDF of the standard normal distribution. For a fixed p_z and $\alpha_0 \in (0, 1)$, we have $c_1/C_0 \le \lambda_{\min}(\mathrm{Cov}) \le \lambda_{\max}(\mathrm{Cov}) \le C_1/c_0$ and $c^*(\alpha_0)$ is a positive constant independent of n. With $c^*(\alpha_0)$ defined in (36), we introduce the following term quantifying the resampling property,

$$\mathrm{err}_n(M, \alpha_0) = \left[\frac{2 \log n}{c^*(\alpha_0) M} \right]^{\frac{1}{2|\hat{\mathcal{S}}|}}, \tag{37}$$

where $\mathrm{err}_n(M, \alpha_0)$ converges to 0 with $M \to \infty$ and M being much larger than $\log n$.

Proposition 1 Suppose Conditions (C1) and (C2) hold and $\alpha_0 \in (0, 1/4)$. If $\mathrm{err}_n(M, \alpha_0) \le c$ for a small positive constant $c > 0$, then there exists a positive constant $C > 0$ such that

$$\liminf_{n\to\infty} \mathbb{P}\left(\min_{1\le m\le M} \left[\max_{\beta\in\mathcal{U}(a)} \max_{j\in\hat{\mathcal{S}}} \frac{|\hat{\Gamma}_j^{[m]} - \Gamma_j^* - \beta(\hat{\gamma}_j^{[m]} - \gamma_j^*)|}{\sqrt{(\hat{\mathbf{V}}_{jj}^{\Gamma} + \beta^2 \hat{\mathbf{V}}_{jj}^{\gamma} - 2\beta \hat{\mathbf{C}}_{jj})/n}} \right] \le \mathrm{Cerr}_n(M, \alpha_0) \right) \ge 1 - \alpha_0, \tag{38}$$

with $\mathrm{err}_n(M, \alpha_0)$ defined in (37) and $\mathcal{U}(a) := \{\beta \in \mathbb{R} : |\beta - \beta^*| \le n^{-a}\}$ for any $a > 1/2$.

Since $c^*(\alpha_0)$ is of a constant order, the condition $\mathrm{err}_n(M, \alpha_0) \le c$ holds for a sufficiently large resampling size M. The above proposition states that, with a high probability, there exists $1 \le m^* \le M$ such that

$$\max_{\beta\in\mathcal{U}(a)} \max_{j\in\hat{\mathcal{S}}} \frac{|\hat{\Gamma}_j^{[m^*]} - \Gamma_j^* - \beta(\hat{\gamma}_j^{[m^*]} - \gamma_j^*)|}{\sqrt{(\hat{\mathbf{V}}_{jj}^{\Gamma} + \beta^2 \hat{\mathbf{V}}_{jj}^{\gamma} - 2\beta \hat{\mathbf{C}}_{jj})/n}} \le \mathrm{Cerr}_n(M, \alpha_0).$$

In comparison to (14), the threshold decreases from $\Phi^{-1}(1 - \frac{\alpha}{2|\hat{\mathcal{S}}|})$ to $\mathrm{Cerr}_n(M, \alpha_0)$. A related sampling property was established in Guo (2020) to address a different nonstandard inference problem.

We now apply Proposition 1 to justify the sampling CI under the majority rule.

Theorem 2 Suppose that the conditions of Proposition 1 hold, $\alpha_0 \in (0, 1/4)$, and λ used in (24) satisfies $\lambda \ge 2\mathrm{Cerr}_n(M, \alpha_0)/\Phi^{-1}[1 - \alpha/(2|\hat{\mathcal{S}}|)]$ and $\lambda \gg n^{1/2-a}$ with the constant C used in (38), $\alpha \in (0, 1/4)$ and $a > 1/2$. Then, $\mathrm{CI}^{\mathrm{samp}}$ defined in (25) and $\mathrm{CI}^{\mathrm{samp}}(\mathcal{M}_0)$ defined in (26) satisfy

$$\liminf_{n\to\infty} \mathbb{P}\big(\beta^* \in \mathrm{CI}^{\mathrm{samp}}\big) \ge \liminf_{n\to\infty} \mathbb{P}\big(\beta^* \in \mathrm{CI}^{\mathrm{samp}}(\mathcal{M}_0)\big) \ge 1 - \alpha_0.$$

There exists a positive constant $C > 0$ such that

$$\liminf_{n\to\infty} \mathbb{P}\left(\max\left\{\frac{\mathbf{L}(\mathrm{CI}^{\mathrm{samp}})}{\sqrt{\log|\mathcal{M}|}}, \mathbf{L}(\mathrm{CI}^{\mathrm{samp}}(\mathcal{M}_0))\right\} \le \frac{C}{\min_{j\in\hat{\mathcal{S}}\cap\mathcal{V}} |\gamma_j^*|\cdot\sqrt{n}}\right) \ge 1-\alpha_0.$$

Note that $2\mathrm{Cerr}_n(M,\alpha_0)/\Phi^{-1}[1-\alpha/(2|\hat{\mathcal{S}}|)] = c(\log n/M)^{1/(2|\hat{\mathcal{S}}|)}$ for some positive constant $c > 0$. Motivated by the condition $\lambda \ge 2\mathrm{Cerr}_n(M,\alpha_0)/\Phi^{-1}[1-\alpha/(2|\hat{\mathcal{S}}|)]$, we choose the tuning parameter λ in the form $\lambda = c_*(\log n/M)^{1/(2|\hat{\mathcal{S}}|)}$ in Remark 3. Similar to Theorem 1, Theorem 2 shows that our proposed searching CI does not require the well-separation condition on the invalidity levels. If the IV strengths $\{\gamma_j^*\}_{j\in\mathcal{S}}$ are assumed to be of a constant order, then the length of $\mathrm{CI}^{\mathrm{samp}}(\mathcal{M}_0)$ defined in (26) is $1/\sqrt{n}$. We can only establish the upper bound $\sqrt{\log|\mathcal{M}|/n}$ for the length of $\mathrm{CI}^{\mathrm{samp}}$ defined in (25). However, we believe that this is mainly a technical artifact since $\mathrm{CI}^{\mathrm{samp}}(\mathcal{M}_0)$ and $\mathrm{CI}^{\mathrm{samp}}$ are nearly the same in the numerical studies; see Online Supplementary Material, Section B.2 in the supplement.

We now switch to the more challenging setting only assuming the finite-sample plurality rule (Condition 4). The main extra step is to show that our constructed initial set $\hat{\mathcal{V}} = \hat{\mathcal{V}}^{\mathrm{TSHT}}$ satisfies (30). To establish this, we provide a careful finite-sample analysis of the voting scheme described in (33) and $\hat{\mathcal{V}}^{\mathrm{TSHT}}$ defined in (34).

Proposition 2 Suppose that Conditions (C1) and (C2) hold. Consider the indexes $j \in \hat{\mathcal{S}}$ and $k \in \hat{\mathcal{S}}$. (a) If $\pi_k^*/\gamma_k^* = \pi_j^*/\gamma_j^*$, then $\liminf_{n\to\infty}\mathbb{P}(\hat{\Pi}_{k,j} = \hat{\Pi}_{j,k} = 1) = 1$. (b)If $|\pi_k^*/\gamma_k^* - \pi_j^*/\gamma_j^*| \ge 2\sqrt{\log n}\cdot \mathbf{T}_{j,k}$, then $\liminf_{n\to\infty}\mathbb{P}(\hat{\Pi}_{k,j} = \hat{\Pi}_{j,k} = 0) = 1$, where $\mathbf{T}_{j,k}$ is defined in (9). Under the additional Condition 4, the constructed $\hat{\mathcal{V}} = \hat{\mathcal{V}}^{\mathrm{TSHT}}$ in (34) satisfies $\liminf_{n\to\infty}\mathbb{P}(\mathcal{V}\cap\mathcal{S}_{\mathrm{str}} \subset \hat{\mathcal{V}} \subset \mathcal{I}(0, 3\mathrm{sep}(n))) = 1$, with $\mathrm{sep}(n)$ defined in (28).

The above proposition shows that if two IVs are of the same invalidity level, then they vote for each other with a high probability. If two IVs are well-separated (i.e., $|\pi_k^*/\gamma_k^* - \pi_j^*/\gamma_j^*| \ge 2\sqrt{\log n}\cdot\mathbf{T}_{j,k}$), then they vote against each other with a high probability. If $0 < |\pi_k^*/\gamma_k^* - \pi_j^*/\gamma_j^*| < 2\sqrt{\log n}\cdot\mathbf{T}_{j,k}$, there is no theoretical guarantee on how the two IVs will vote. The above proposition explains why we are likely to make a mistake in detecting some locally invalid IVs defined in Definition 1.

With Proposition 2, we connect the finite-sample plurality rule to the finite-sample majority rule. We apply the voting algorithm to remove all strongly invalid IVs and the set $\hat{\mathcal{V}}^{\mathrm{TSHT}}$ only consists of valid IVs and the locally invalid IVs. Under Condition 4, $\mathcal{V}\cap\mathcal{S}_{\mathrm{str}}$ becomes the majority of $\hat{\mathcal{V}}^{\mathrm{TSHT}}$. We then establish the following theorem by applying the theoretical analysis of the majority rule by replacing $\hat{\mathcal{S}}$ with $\hat{\mathcal{V}}^{\mathrm{TSHT}}$.

Theorem 3 Consider the model (4). Suppose that Condition 4, Conditions (C1) and (C2) hold. Then with $\hat{\mathcal{S}}$ replaced by $\hat{\mathcal{V}}^{\mathrm{TSHT}}$, the coverage and precision properties in Theorem 1 hold for the searching interval in Algorithm 3 and the coverage and precision properties in Theorem 2 hold for the sampling interval in Algorithm 3.

7 Simulation studies

Throughout the numerical studies, we implement our proposed $\hat{\mathrm{CI}}^{\mathrm{sear}}$ and $\mathrm{CI}^{\mathrm{samp}}$ detailed in Algorithm 3. We illustrate the robustness of Algorithm 3 to different initial estimators of $\mathcal{V}$ by considering two choices of $\hat{\mathcal{V}}$: $\hat{\mathcal{V}}^{\mathrm{TSHT}}$ defined in (34) and the set of valid IVs $\hat{\mathcal{V}}^{\mathrm{CIIV}}$ outputted by `CIIV` (Windmeijer et al., 2021). We focus on the low-dimensional setting in the current section and will present the high-dimensional results in Online Supplementary Material, Section A.3 in the supplement. The code for replicating all numerical results in the current paper is available at https://github.com/zijguo/Searching-Sampling-Replication.

We compare with three existing CIs allowing for invalid IVs: `TSHT` (Guo et al., 2018), `CIIV` (Windmeijer et al., 2021), and the `Union` method (Kang et al., 2020). `TSHT` and `CIIV` are implemented with the codes on the Github websites[1] while the `Union` method is implemented with the code shared by the authors of Kang et al. (2020). The `Union` method takes a union of intervals that are constructed by a given number of candidate IVs and are not rejected by the Sargan test. An upper bound $\bar{s}$ for the number of invalid IVs is required for the implementation. We consider two specific upper bounds: $\bar{s} = p_z - 1$ corresponds to the existence of two valid IVs, and $\bar{s} = \lceil p_z/2 \rceil$ corresponds to the majority rule being satisfied. We conduct 500 replications of simulations and compare different CIs in terms of empirical coverage and average lengths.

We implement two oracle methods as the benchmark. Firstly, we implement the oracle TSLS assuming the prior knowledge of $\mathcal{V}$. This method serves as a benchmark when the set $\mathcal{V}$ of valid IVs is correctly recovered. Secondly, we implement the oracle bias-aware confidence interval in (10) assuming the oracle knowledge of the bias of `TSHT` estimator. We argue that the oracle bias-aware confidence interval serves as a better benchmark, especially when $\mathcal{V}$ might not be correctly recovered in finite samples.

We generate the i.i.d. data $\{Y_i, D_i, Z_{i\cdot}, X_{i\cdot}\}_{1 \le i \le n}$ using the outcome model (1) and the treatment model (2). We generate $\gamma^* \in \mathbb{R}^{p_z}$ and $\pi^* \in \mathbb{R}^{p_z}$ as follows:

S1 (Majority rule): set $\gamma^* = \gamma_0 \cdot \mathbf{1}_{10}$ and $\pi^* = (\mathbf{0}_6, \tau \cdot \gamma_0, \tau \cdot \gamma_0, -0.5, -1)^\top$;
S2 (Plurality rule): set $\gamma^* = \gamma_0 \cdot \mathbf{1}_{10}$ and $\pi^* = (\mathbf{0}_4, \tau \cdot \gamma_0, \tau \cdot \gamma_0, -\frac{1}{3}, -\frac{2}{3}, -1, -\frac{4}{3})^\top$;
S3 (Plurality rule): set $\gamma^* = \gamma_0 \cdot \mathbf{1}_{10}$ and $\pi^* = (\mathbf{0}_4, \tau \cdot \gamma_0, \tau \cdot \gamma_0, -\frac{1}{6}, -\frac{1}{3}, -\frac{1}{2}, -\frac{2}{3})^\top$;
S4 (Plurality rule): set $\gamma^* = \gamma_0 \cdot \mathbf{1}_6$ and $\pi^* = (\mathbf{0}_2, -0.8, -0.4, \tau \cdot \gamma_0, 0.6)^\top$;
S5 (Plurality rule): set $\gamma^* = \gamma_0 \cdot \mathbf{1}_6$ and $\pi^* = (\mathbf{0}_2, -0.8, -0.4, \tau \cdot \gamma_0, \tau \cdot \gamma_0 + 0.1)^\top$.

The parameter γ_0 denotes the IV strength and is set as 0.5. The parameter τ denotes the invalidity level and is varied across $\{0.2, 0.4\}$. The setting **S1** satisfies the population majority rule while the settings **S2–S5** only satisfy the population plurality rule. Settings **S4** and **S5** represent the challenging settings where there are only two valid IVs. We introduce settings **S3** and **S5** to test the robustness of our proposed method when the finite-sample plurality rule might be violated. For example, for the setting **S5** with small n (e.g., $n = 500$), the invalid IVs with π_j^* values $\tau \cdot \gamma_0, \tau \cdot \gamma_0 + 0.1$ have similar invalidity levels and may violate Condition 4.

We now specify the remaining details for the generating models (1) and (2). Set $p_x = 10$, $\phi^* = (0.6, 0.7, \ldots, 1.5)^\top \in \mathbb{R}^{10}$ in (2) and $\psi^* = (1.1, 1.2, \ldots, 2)^\top \in \mathbb{R}^{10}$ in (1). We vary n across $\{500, 1{,}000, 2{,}000, 5{,}000\}$. For $1 \le i \le n$, generate the covariates $W_{i\cdot} = (Z_{i\cdot}^\top, X_{i\cdot}^\top)^\top \in \mathbb{R}^p$ following a multivariate normal distribution with zero mean and covariance $\Sigma \in \mathbb{R}^{p \times p}$ where $\Sigma_{jl} = 0.5^{|j-l|}$ for $1 \le j, l \le p$; generate the errors $(e_i, \delta_i)^\top$ following bivariate normal with zero mean, unit variance and $\mathrm{Cov}(e_i, \delta_i) = 0.8$.

In Table 1, we report the empirical coverage and interval length for $\tau = 0.2$. The CIs by `TSHT` and `CIIV` undercover for $n = 500$, 1,000, and 2,000 and only achieve the 95% coverage level for a large sample size $n = 5{,}000$. Our proposed searching and sampling CIs achieve the desired coverage levels in most settings. For settings **S1–S4**, both initial estimates of set of valid IVs $\hat{\mathcal{V}}^{\texttt{TSHT}}$ and $\hat{\mathcal{V}}^{\texttt{CIIV}}$ lead to CIs achieving the 95% coverage level. For the more challenging settings **S5**, the empirical coverage level of our proposed searching and sampling CIs achieve the desired coverage with sample sizes above 2,000. For $n = 500$ and $n = 1{,}000$, our proposed searching and sampling methods improve the coverage of `TSHT` and `CIIV`. The undercoverage happens mainly due to the fact that the finite-sample plurality rule might fail for setting **S5** with a relatively small sample size. The CIs by the `Union` method (Kang et al., 2020) with $\bar{s} = p_z - 1$ (assuming there are two valid IVs) achieve the desired coverage level while those with $\bar{s} = \lceil p_z/2 \rceil$ (assuming the majority rule) do not achieve the desired coverage level for settings **S2–S5**.

We now compare the CI lengths. When the CIs by `TSHT` and `CIIV` are valid, their lengths are similar to the length of the CI by oracle TSLS, which match with the theory in Guo et al. (2018) and Windmeijer et al. (2021). For the CIs achieving valid coverage, our proposed sampling

[1] The code for `TSHT` is obtained from https://github.com/hyunseungkang/invalidIV and for `CIIV` is obtained from https://github.com/xlbristol/CIIV.

Table 1. Empirical coverage and average lengths for **S1** to **S5** with $\tau = 0.2$

		Oracle				Searching		Sampling			Union	
Set	n	TSLS	BA	TSHT	CIIV	$\hat{\mathcal{V}}^{\text{TSHT}}$	$\hat{\mathcal{V}}^{\text{CIIV}}$	$\hat{\mathcal{V}}^{\text{TSHT}}$	$\hat{\mathcal{V}}^{\text{CIIV}}$	Check	$p_z - 1$	$\lceil p_z/2 \rceil$
Empirical coverage of confidence intervals for $\tau = 0.2$												
	500	0.95	0.97	0.53	0.61	1.00	1.00	1.00	1.00	1.00	1.00	0.99
	1,000	0.94	0.97	0.45	0.69	1.00	1.00	1.00	1.00	1.00	1.00	0.98
	2,000	0.96	0.95	0.63	0.79	1.00	1.00	1.00	1.00	1.00	1.00	0.96
S1	5,000	0.94	0.95	0.89	0.92	1.00	1.00	1.00	1.00	1.00	1.00	0.96
	500	0.95	0.96	0.56	0.51	1.00	1.00	1.00	0.99	1.00	1.00	0.27
	1,000	0.94	0.95	0.45	0.58	0.99	0.99	1.00	0.97	1.00	1.00	0.03
	2,000	0.94	0.93	0.51	0.74	0.98	0.98	0.98	0.97	1.00	1.00	0.00
S2	5,000	0.96	0.93	0.85	0.95	0.99	1.00	1.00	1.00	1.00	1.00	0.00
	500	0.95	0.95	0.63	0.63	0.99	0.99	0.99	0.99	1.00	1.00	0.62
	1,000	0.94	0.97	0.62	0.60	0.99	0.99	0.99	0.97	1.00	1.00	0.18
	2,000	0.94	0.93	0.62	0.73	0.97	0.98	0.97	0.96	1.00	1.00	0.01
S3	5,000	0.96	0.93	0.85	0.95	0.99	1.00	0.99	1.00	1.00	1.00	0.00
	500	0.95	0.96	0.72	0.66	0.94	0.95	0.94	0.93	0.98	0.98	0.00
	1,000	0.94	0.95	0.65	0.58	1.00	0.97	0.99	0.96	0.95	0.98	0.00
	2,000	0.93	0.99	0.68	0.58	0.98	0.97	0.97	0.93	0.99	0.95	0.00
S4	5,000	0.95	0.95	0.91	0.88	0.98	0.95	0.98	0.95	1.00	0.94	0.00
	500	0.95	0.97	0.49	0.51	0.81	0.89	0.88	0.86	0.98	0.97	0.14
	1,000	0.94	0.96	0.31	0.50	0.68	0.89	0.76	0.86	0.91	0.98	0.00
	2,000	0.93	0.98	0.46	0.57	0.86	0.96	0.86	0.92	0.84	0.95	0.00
S5	5,000	0.95	0.95	0.90	0.88	0.98	0.95	0.97	0.94	0.98	0.94	0.00
Average lengths of confidence intervals for $\tau = 0.2$												
	500	0.10	0.16	0.08	0.08	0.59	0.61	0.34	0.34	–	1.16	0.25
	1,000	0.07	0.14	0.06	0.06	0.39	0.43	0.24	0.24	–	0.63	0.16
	2,000	0.05	0.10	0.05	0.05	0.27	0.30	0.17	0.17	–	0.42	0.09
S1	5,000	0.03	0.04	0.03	0.03	0.17	0.18	0.10	0.10	–	0.27	0.04
	500	0.13	0.24	0.13	0.10	0.58	0.59	0.37	0.36	–	2.46	0.07
	1,000	0.09	0.24	0.13	0.08	0.37	0.41	0.26	0.26	–	1.45	0.02
	2,000	0.06	0.26	0.14	0.06	0.25	0.29	0.19	0.18	–	0.76	0.00
S2	5,000	0.04	0.13	0.08	0.04	0.16	0.17	0.10	0.10	–	0.28	0.00
	500	0.13	0.22	0.10	0.10	0.62	0.62	0.45	0.38	–	1.77	0.13
	1,000	0.09	0.21	0.10	0.08	0.38	0.41	0.29	0.27	–	1.36	0.03
	2,000	0.06	0.25	0.13	0.06	0.26	0.29	0.19	0.18	–	0.86	0.00
S3	5,000	0.04	0.13	0.08	0.04	0.16	0.17	0.10	0.10	–	0.35	0.00
	500	0.23	0.62	0.24	0.18	0.56	0.56	0.48	0.44	–	0.87	0.00
	1,000	0.16	0.56	0.17	0.13	0.44	0.36	0.38	0.29	–	0.42	0.00
	2,000	0.11	0.32	0.14	0.10	0.27	0.24	0.22	0.18	–	0.20	0.00
S4	5,000	0.07	0.13	0.08	0.07	0.14	0.13	0.11	0.10	–	0.09	0.00
	500	0.23	0.63	0.27	0.17	0.42	0.51	0.41	0.42	–	1.01	0.05
	1,000	0.16	0.61	0.18	0.13	0.32	0.36	0.30	0.29	–	0.50	0.00
	2,000	0.11	0.38	0.12	0.10	0.28	0.24	0.25	0.18	–	0.22	0.00

(*continued*)

Table 1. Continued

		Oracle				Searching		Sampling			Union	
Set	n	TSLS	BA	TSHT	CIIV	$\hat{\mathcal{V}}^{\text{TSHT}}$	$\hat{\mathcal{V}}^{\text{CIIV}}$	$\hat{\mathcal{V}}^{\text{TSHT}}$	$\hat{\mathcal{V}}^{\text{CIIV}}$	Check	$p_z - 1$	$\lceil p_z/2 \rceil$
S5	5,000	0.07	0.15	0.08	0.07	0.15	0.13	0.12	0.10	–	0.09	0.00

Note. The columns indexed with TSLS, BA, TSHT, and CIIV represent the oracle TSLS CI with the knowledge of $\mathcal{V}$, the oracle bias-aware CI in (10), the CI by Guo et al. (2018), and the CI by Windmeijer et al. (2021), respectively. Under the columns indexed with 'Searching' (or 'Sampling'), the columns indexed with $\hat{\mathcal{V}}^{\text{TSHT}}$ and $\hat{\mathcal{V}}^{\text{CIIV}}$ represent our proposed searching (or sampling) CI in Algorithm 3 with $\hat{\mathcal{V}}^{\text{TSHT}}$ and $\hat{\mathcal{V}}^{\text{CIIV}}$, respectively. The column indexed with 'Check reports the proportion of simulations passing the plurality rule check in Algorithm 3. The columns indexed with Union represent the method by Kang et al. (2020). The columns indexed with $p_z - 1$ and $\lceil p_z/2 \rceil$ correspond to the Union methods assuming two valid IVs and the majority rule, respectively.

CI is, in general, shorter than the searching CI and the Union CI with $p_z - 1$. We shall point out that our proposed sampling CI can be even shorter than the oracle bias-aware CI (the benchmark). As an important remark, the oracle bias-aware CI, the sampling CI, the searching CI, and the Union CI are, in general, longer than the CI by the oracle TSLS, which is a price to pay for constructing uniformly valid CIs. In Online Supplementary Material, Section E.1 in the supplement, we consider the settings with $\tau = 0.4$ and heteroscedastic errors. In Online Supplementary Material, Section E.3 in the supplement, we further explore the settings considered in Windmeijer et al. (2021). The results for these settings are similar to those in Table 1.

Varying violation strength. In Figure 5, we focus on the setting **S2** and vary τ across $\{0.025, 0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5\}$. We follow Bekker and Crudu (2015) and generate heteroscedastic errors as follows: for $1 \le i \le n$, generate $\delta_i \sim N(0, 1)$ and $e_i = 0.3\delta_i + \sqrt{[1 - 0.3^2]/[0.86^4 + 1.38^2]}(1.38 \cdot \tau_{1,i} + 0.86^2 \cdot \tau_{2,i})$, where conditioning on Z_i, $\tau_{1,i} \sim N(0, [0.5 \cdot Z_{i,1}^2 + 0.25]^2)$, $\tau_{2,i} \sim N(0, 1)$, and $\tau_{1,i}$ and $\tau_{2,i}$ are independent of δ_i.

As reported in Figure 5, our proposed searching and sampling CIs achieve the desired coverage (95%) while TSHT and CIIV only achieve the desire coverage for $\tau = 0.5$ with $n = 2{,}000$ and $\tau \ge 0.3$ with $n = 5{,}000$. In terms of length, we observe that the sampling CI is shorter than the searching CI. The sampling CI can be even shorter than the oracle bias-aware CI (the benchmark). We do not plot the length of the Union CI, which is three to six times longer than our proposed sampling CI; see Figure 2 for details. An interesting observation is that the empirical coverage of TSHT and CIIV is around 90% for $\tau = 0.025$. This happens since the invalidity levels of the IVs are small, and even the inclusion of such invalid IVs does not significantly worsen the empirical coverage.

We present the results for homoscedastic errors in Online Supplementary Material, Figure E.1 in the supplement. In Online Supplementary Material, Section E.2 in the supplement, we explore the performance of different methods for settings with the locally invalid IVs where the violation levels are scaled to $\sqrt{\log n/n}$.

Tuning parameter selection. We investigate the robustness of Algorithm 3 to different choices of tuning parameters. For the searching CI, we observe in Table 2 that the empirical coverage and the average lengths are almost invariant to different choices of L, U, and a. For two intervals (a_1, b_1) and (a_2, b_2), we define its difference as $|a_1 - a_2| + |b_1 - b_2|$. In Table 2, we report the difference between the searching CI constructed with the default choice of L, U, a in Algorithm 3 and searching CIs with other choices of L, U, a. The average interval difference is smaller than twice the default grid size $n^{-0.6}$.

In Online Supplementary Material, Section B.2 in the supplement, we demonstrate that the sampling CIs have nearly the same empirical coverage and length for different choices of L, U, a, and the resampling size M. The choice of the shrinkage parameter $\lambda > 0$ has a more obvious effect on the sampling CI. As explained in Remark 3, we choose the smallest λ such that more than a prespecified proportion (denoted as prop) of the $M = 1{,}000$ intervals are nonempty. In Figure 6, we compare the coverage and length properties of the sampling CIs by varying prop across {1%, 5%, 10%, 20%, 30%}. If prop $\ge$ 5%, the empirical coverage reaches the nominal level and vary slightly with prop. In terms of length, the intervals get slightly longer with a larger value of prop. The empirical coverage and average lengths of the sampling CIs are robust to a wide range of λ values, as long as the corresponding λ guarantees a sufficient proportion of nonempty sampled searching CIs.

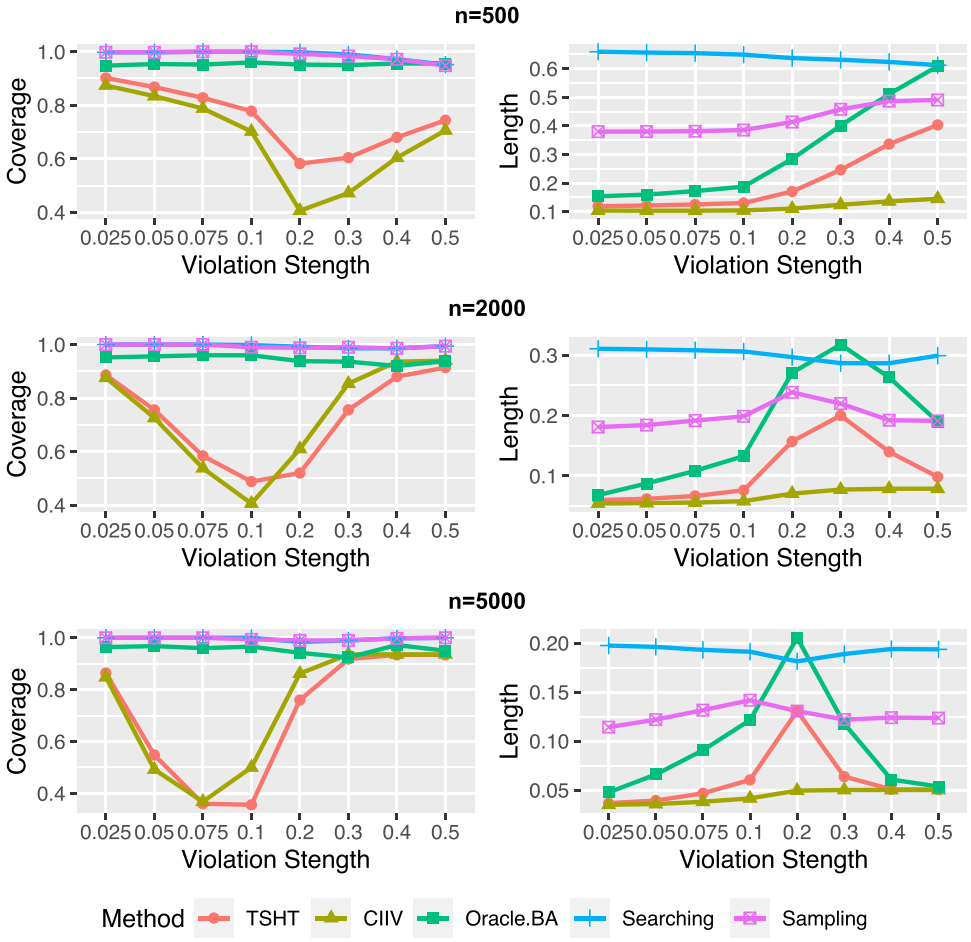

Figure 5. Empirical coverage and average lengths for the setting **S2** with $\tau \in$ {0.025, 0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5} and heteroscedastic errors. `Oracle` – `BA`, `TSHT`, and `CIIV` represent the oracle bias-aware CI in (10), the CI by Guo et al. (2018), and the CI by Windmeijer et al. (2021), respectively. The searching and sampling CIs are implemented as in Algorithm 3.

8 Real data analysis

We study the effect of education on earnings by analyzing the 2018 survey of China Family Panel Studies (Xie, 2012). The outcome is the logarithm transformation of salary in the past 12 months, and the treatment is the years of education. We include three binary baseline covariates: gender, urban (whether the subject lives in the urban area), and hukou (whether the subject's hukou is agricultural or nonagricultural). Following the literature, we have induced the following nine IVs:

(a) Family background variables (e.g., Behrman et al., 2012; Blackburn & Neumark, 1993; Trostel et al., 2002): the father's education level, the mother's education level, the spouse's education level, the family size, and the log transformation of the education expenditure in the past 12 months;
(b) The group-level years of education where the groups are formulated by age, gender, and region;
(c) Personal traits (Behrman et al., 2012): the statement on the fair competition, the statement on the talent pay-off, and whether the subject reads some books or not in the past 12 months.

Table 2. Searching CIs with different choices of L, U, a for settings **S1**–**S4** with $\tau = 0.4$ and $n = 2{,}000$

		S1 ($n = 2{,}000$)			S2 ($n = 2{,}000$)			S3 ($n = 2{,}000$)			S4 ($n = 2{,}000$)		
[L,U]	a	Cov	Len	Diff	Cov	Len	Diff	Cov	Len	Diff	Cov	Len	Diff
As in (19)	0.6	1	0.287	0.000	1	0.268	0.000	0.994	0.275	0.000	0.980	0.274	0.000
	0.8	1	0.295	0.008	1	0.276	0.008	0.996	0.283	0.008	0.982	0.282	0.008
	1.0	1	0.297	0.010	1	0.278	0.010	0.996	0.284	0.010	0.984	0.284	0.010
[−10,10]	0.6	1	0.287	0.007	1	0.268	0.008	0.994	0.275	0.008	0.978	0.274	0.007
	0.8	1	0.295	0.008	1	0.277	0.009	0.996	0.283	0.009	0.984	0.282	0.009
	1.0	1	0.297	0.010	1	0.278	0.010	0.996	0.285	0.010	0.984	0.284	0.010
[−20,20]	0.6	1	0.287	0.007	1	0.268	0.007	0.992	0.275	0.007	0.982	0.274	0.007
	0.8	1	0.295	0.008	1	0.277	0.009	0.996	0.283	0.009	0.982	0.282	0.009
	1.0	1	0.297	0.010	1	0.278	0.010	0.996	0.285	0.010	0.984	0.284	0.010

Note. The columns indexed with 'Cov and 'Len' denote empirical coverage and average lengths, respectively. The column indexed with 'Diff' represents the average length difference between the searching CI in Algorithm 3 and the searching CIs with other choices of L, U, and n^{-a}.

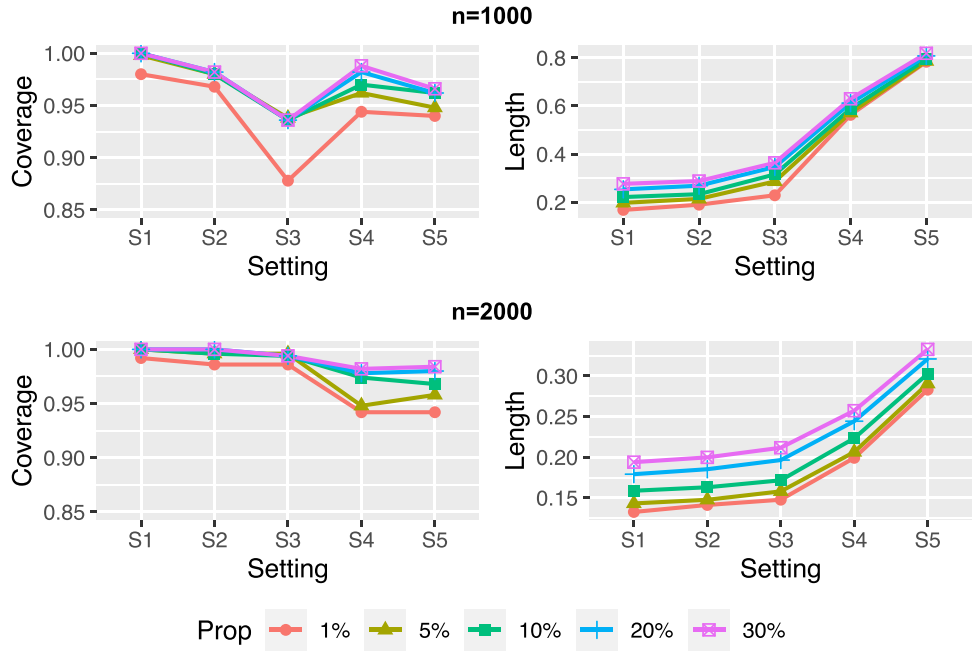

Figure 6. Comparison of the sampling CIs with different λ for $\tau = 0.4$ and $M = 1{,}000$. We choose the smallest value of λ such that $1{,}000 \cdot \texttt{prop}$ sampled intervals are nonempty. For example, for $\texttt{prop} = 10\%$, we choose the smallest λ leading to 100 nonempty sampled intervals.

The statement on fair competition measures an individual's willingness to compete through ability. The statement on the talent pay-off is about an individual's viewpoint on whether their educational endeavors will pay off. After removing the missing values, the data consists of 3,758 observations. In the supplement, we report the summary statistics of all IVs and baseline covariates in Online Supplementary Material, Table E.8.

In Table 3, we compare our proposed searching and sampling CIs with existing methods. By applying `TSHT`, we identified six relevant instruments: the father's education level, the mother's education level, the spouse's education level, the group-level years of education, the family size, whether to read some books or not in the past 12 months. Out of these IVs, the family size is detected as the invalid IV. `CIIV` outputs the same set of valid IVs. When we include all nine IVs, the concentration parameter is 2,906.36. If we only include the five IVs selected by `TSHT` and `CIIV`, the concentration parameter is 2,850.57. This indicates that the whole set of IVs are strongly associated with the treatment, but some IVs (e.g., the family size) are possibly invalid.

As reported in Table 3, CIs by `TSHT` and `CIIV` are relatively short, but they may undercover due to the IV selection error. We compare the lengths of `Union` method and our proposed searching and sampling CIs, which are all robust to the IV selection error. The sampling CI is the shortest among these CIs. We further plot the searching and sampling CIs in Online Supplementary Material, Figure E.4 in the supplement. The validity of the `Union` CI with $\bar{s} = \lceil p_z/2 \rceil$ requires half of the candidate IVs to be valid; if we can only assume that two of candidate IVs are valid, then the CI by `Union` with $\bar{s} = \lceil p_z/2 \rceil$ may not be valid but the CI by `Union` with $\bar{s} = p_z - 1$ is valid.

9 Conclusion and discussion

Causal inference from observational studies is a challenging task. Typically, stringent identification conditions are required to facilitate various causal inference approaches. The valid IV assumption is one of such assumptions to handle unmeasured confounders. In the current paper, we devise uniformly valid confidence intervals for the causal effect when the candidate IVs are possibly invalid. Our proposed searching and sampling confidence intervals add to the fast-growing literature on robust inference with possibly invalid IVs. The proposed method has the advantage of

Table 3. Confidence intervals for the effect of education on earnings

Method	CI	Method	CI
OLS	(0.0305, 0.0503)	Searching CI	(0.0409, 0.1698)
TSLS	(0.0959, 0.1190)	Sampling CI	(0.0552, 0.1268)
TSHT	(0.0946, 0.1178)	Union ($\bar{s} = p_z - 1$)	(−0.4915, 1.6043)
CIIV	(0.0948, 0.1175)	Union($\bar{s} = \lceil p_z/2 \rceil$)	(0.0409, 0.1342)

being more robust to the mistakes in separating the valid and invalid IVs at the expense of a wider confidence interval. The proposed intervals are computationally efficient and less conservative than existing uniformly valid confidence intervals.

Acknowledgments

The research of Z. Guo was partly supported by the NSF grants DMS 1811857 and 2015373 and NIH grants R01GM140463 and R01LM013614. Z. Guo is grateful to the participants at Penn Causal Reading Group and CUHK econometrics seminar for their helpful discussion, to Dr. Frank Windmeijer for pointing out the connection to the CIIV method, and to Dr. Hyunseung Kang for sharing the code for replicating the `Union` method. Z. Guo thanks the editors, two anonymous referees, and Drs. Hyunseung Kang, Zhonghua Liu, and Molei Liu for constructive comments on a previous draft. Z. Guo thanks Mr. Zhenyu Wang for the help with the numerical implementation and Dr. Junhui Yang and Mr. Zhenyu Wang for cleaning the CFHS data set.

Data availability

The data underlying this article is the 2018 survey of China Family Panel Studies (Xie, 2012), available at https://www.isss.pku.edu.cn/cfps/en/data/public/index.htm.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series B*.

References

Anderson T. W., & Rubin H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, *20*(1), 46–63. https://doi.org/10.1214/aoms/1177730090

Armstrong T. B., Kolesár M., & Kwon S. (2020). *Bias-aware inference in regularized regression models*. arXiv preprint arXiv:2012.14823.

Behrman J. R., Mitchell O. S., Soo C. K., & Bravo D. (2012). How financial literacy affects household wealth accumulation. *American Economic Review*, *102*(3), 300–304. https://doi.org/10.1257/aer.102.3.300

Bekker P. A., & Crudu F. (2015). Jackknife instrumental variable estimation with heteroskedasticity. *Journal of Econometrics*, *185*(2), 332–342. https://doi.org/10.1016/j.jeconom.2014.08.012

Berk R., Brown L., Buja A., Zhang K., & Zhao L. (2013). Valid post-selection inference. *The Annals of Statistics*, *41*(2), 802–837. https://doi.org/10.1214/12-AOS1077

Berkowitz D., Caner M., & Fang Y. (2012). The validity of instruments revisited. *Journal of Econometrics*, *166*(2), 255–266. https://doi.org/10.1016/j.jeconom.2011.09.038

Blackburn M. L., & Neumark D. (1993). Omitted-ability bias and the increase in the return to schooling. *Journal of Labor Economics*, *11*(3), 521–544. https://doi.org/10.1086/298306

Bowden J., Davey Smith G., & Burgess S. (2015). Mendelian randomization with invalid instruments: Effect estimation and bias detection through egger regression. *International Journal of Epidemiology*, *44*(2), 512–525. https://doi.org/10.1093/ije/dyv080

Bowden J., Davey Smith G., Haycock P. C., & Burgess S. (2016). Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology*, *40*(4), 304–314. https://doi.org/10.1002/gepi.21965

Burgess S., Small D. S., & Thompson S. G. (2017). A review of instrumental variable estimators for Mendelian randomization. *Statistical Methods in Medical Research*, *26*(5), 2333–2355. https://doi.org/10.1177/0962280215597579

Cai T. T., & Guo Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Annals of Statistics*, *45*(2), 615–626. https://doi.org/10.1080/02331888.2016.1265969
Caner M., Han X., & Lee Y. (2018). Adaptive elastic net GMM estimation with many invalid moment conditions: Simultaneous model and moment selection. *Journal of Business & Economic Statistics*, *36*(1), 24–46. https://doi.org/10.1080/07350015.2015.1129344
Cheng X., & Liao Z. (2015). Select the valid and relevant moments: An information-based Lasso for GMM with many moments. *Journal of Econometrics*, *186*(2), 443–464. https://doi.org/10.1016/j.jeconom.2015.02.019
Chernozhukov V., Hansen C., & Spindler M. (2015). Post-selection and post-regularization inference in linear models with many controls and instruments. *American Economic Review*, *105*(5), 486–490. http://dx.doi.org/10.1257/aer.p20151022
Davey Smith G., & Ebrahim S. (2003). Mendelian randomization: Can genetic epidemiology contribute to understanding environmental determinants of disease?. *International Journal of Epidemiology*, *32*(1), 1–22. https://doi.org/10.1093/ije/dyg070
Donoho D. L., & Johnstone J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, *81*(3), 425–455. https://doi.org/10.1093/biomet/81.3.425
Eicker F. (1967). Limit theorems for regression with unequal and dependent errors. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 59–82).
Fan Q., & Wu Y. (2020). *Endogenous treatment effect estimation with some invalid and irrelevant instruments*. arXiv preprint arXiv:2006.14998.
Goh G., & Yu J. (2022). Causal inference with some invalid instrumental variables: A quasi-Bayesian approach. *Oxford Bulletin of Economics and Statistics*, *84*(6), 1432–1451. https://doi.org/10.1111/obes.12513
Guggenberger P. (2012). On the asymptotic size distortion of tests when instruments locally violate the exogeneity assumption. *Econometric Theory*, *28*(2), 387–421. https://doi.org/10.1017/S0266466611000375
Guo Z. (2020). *Statistical inference for maximin effects: Identifying stable associations across multiple studies*. arXiv preprint arXiv:2011.07568.
Guo Z., & Bühlmann P. (2022). *Two stage curvature identification with machine learning: Causal inference with possibly invalid instrumental variables*. arXiv preprint arXiv:2203.12808.
Guo Z., Kang H., Tony Cai T., & Small D. S. (2018). 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. https://doi.org/10.1111/rssb.12275
Hahn J., & Hausman J. (2005). Estimation with valid and invalid instruments. *Annals of Economics and Statistics*, *79/80*, 25–57. https://doi.org/10.2307/20777569
Hartwig F. P., Davey Smith G., & Bowden J. (2017). Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *International Journal of Epidemiology*, *46*(6), 1985–1998. https://doi.org/10.1093/ije/dyx102
Huber P. J. (1967). Under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (p. 221).
Javanmard A., & Montanari A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, *15*(1), 2869–2909. https://dl.acm.org/doi/10.5555/2627435.2697057
Kang H., Lee Y., Cai T. T., & Small D. S. (2020). Two robust tools for inference about causal effects with invalid instruments. *Biometrics*, *78*(1), 24-34. https://doi.org/10.1111/biom.13415
Kang H., Zhang A., Cai T. T., & Small D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *Journal of the American Statistical Association*, *111* (513), 132–144. https://doi.org/10.1080/01621459.2014.994705
Kolesár M., Chetty R., Friedman J., Glaeser E., & Imbens G. W. (2015). Identification and inference with many invalid instruments. *Journal of Business & Economic Statistics*, *33*(4), 474–484. https://doi.org/10.1080/07350015.2014.978175
Lee J. D., Sun D. L., Sun Y., & Taylor J. E. (2016). Exact post-selection inference, with application to the Lasso. *The Annals of Statistics*, *44*(3), 907–927. https://doi.org/10.1214/15-AOS1371
Leeb H., & Pötscher B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, *21*(1), 21–59. https://doi.org/10.1017/S0266466605050036
Lewbel A. (2012). Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models. *Journal of Business & Economic Statistics*, *30*(1), 67–80. https://doi.org/10.1080/07350015.2012.643126
Liao Z. (2013). Adaptive GMM shrinkage estimation with consistent moment selection. *Econometric Theory*, *29*(5), 857–904. https://doi.org/10.1017/S0266466612000783
Liu Z., Ye T., Sun B., Schooling M., & Tchetgen E. T. (2020). *On Mendelian randomization mixed-scale treatment effect robust identification (MR MiSTERI) and estimation for causal inference*. arXiv preprint arXiv:2009.14484.

Sargan J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica*, *26*(3), 393–415. https://doi.org/10.2307/1907619

Small D. S. (2007). Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association*, *102*(479), 1049–1058. https://doi.org/10.1198/016214507000000608

Staiger D., & Stock J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, *65*(3), 557–586. https://doi.org/10.2307/2171753

Stock J. H., Wright J. H., & Yogo M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, *20*(4), 518–529. https://doi.org/10.1198/073500102288618658

Tchetgen E., Sun B., & Walter S. (2021). The genius approach to robust Mendelian randomization inference. *Statistical Science*, *36*(3), 443–464. https://doi.org/10.1214/20-STS802

Trostel P., Walker I., & Woolley P. (2002). Estimates of the economic return to schooling for 28 countries. *Labour Economics*, *9*(1), 1–16. https://doi.org/10.1016/S0927-5371(01)00052-5

van de Geer S., Bühlmann P., Ritov Y., & Dezeure R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, *42*(3), 1166–1202. https://doi.org/10.1214/14-AOS1221

Windmeijer F., Farbmacher H., Davies N., & Davey Smith G. (2019). On the use of the Lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association*, *114*(527), 1339–1350. https://doi.org/10.1080/01621459.2018.1498346

Windmeijer F., Liang X., Hartwig F. P., & Bowden J. (2021). The confidence interval method for selecting valid instrumental variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *83*(4), 752–776. https://doi.org/10.1111/rssb.12449

Wooldridge J. M. (2010). *Econometric analysis of cross section and panel data*. MIT Press.

Xie M.-g., & Wang P. (2022). *Repro samples method for finite-and large-sample inferences*. arXiv preprint arXiv:2206.06421.

Xie Y. (2012). *China family panel studies. (2010) User's manuals*. Institute of Social Science Survey, Peking University.

Zhang C.-H., & Zhang S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *76*(1), 217–242. https://doi.org/10.1111/rssb.12026

Zhao Q., Wang J., Hemani G., Bowden J., & Small D. S. (2020). Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *The Annals of Statistics*, *48*(3), 1742–1769. https://doi.org/10.1214/19-AOS1866