Journal of the Royal Statistical Society Series B: Statistical Methodology, 2024, **00**, 1–21
https://doi.org/10.1093/jrsssb/qkae025

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

Original Article

On the instrumental variable estimation with many weak and invalid instruments

Yiqi Lin[1], Frank Windmeijer[2], Xinyuan Song[1] and Qingliang Fan[3]

[1]Department of Statistics, The Chinese University of Hong Kong, Hong Kong
[2]Department of Statistics and Nuffield College, University of Oxford, Oxford, UK
[3]Department of Economics, The Chinese University of Hong Kong, Hong Kong

Address for correspondence: Qingliang Fan, Department of Economics, The Chinese University of Hong Kong, 903, Esther Lee Building, Shatin, N.T., Hong Kong. Email: michaelqfan@gmail.com

Abstract

We discuss the fundamental issue of identification in linear instrumental variable (IV) models with unknown IV validity. With the assumption of the 'sparsest rule', which is equivalent to the plurality rule but becomes operational in computation algorithms, we investigate and prove the advantages of non-convex penalized approaches over other IV estimators based on two-step selections, in terms of selection consistency and accommodation for individually weak IVs. Furthermore, we propose a surrogate sparsest penalty that aligns with the identification condition and provides oracle sparse structure simultaneously. Desirable theoretical properties are derived for the proposed estimator with weaker IV strength conditions compared to the previous literature. Finite sample properties are demonstrated using simulations and the selection and estimation method is applied to an empirical study concerning the effect of body mass index on diastolic blood pressure.

Keywords: invalid instruments, model identification, non-convex penalty, treatment effect, weak instruments

1 Introduction

Recently, estimation of causal effects with high-dimensional observational data has drawn much attention in many research fields such as economics, epidemiology, and genomics. The instrumental variable (IV) method is widely used when the treatment variable of interest is endogenous. As shown in Figure 1, the ideal IV needs to be correlated with the endogenous treatment variable (C1), it should not have a direct effect on the outcome (C2) and should not be related to unobserved confounders that affect both outcome and treatment (C3).

Our research is motivated by the difficulty of finding IVs that satisfy all the above conditions. In applications, invalid IVs (violation of C2 or C3) (Davey Smith & Ebrahim, 2003; Kang et al., 2016; Windmeijer et al., 2019) and weak IVs (concerning the weak correlation in C1) (Bound et al., 1995; Staiger & Stock, 1997) are prevalent. A strand of literature studies the 'many weak IVs' problem (Andrews et al., 2018; Chao & Swanson, 2005). With the increasing availability of large datasets, IV models are often high-dimensional (Belloni et al., 2012; Fan & Zhong, 2018; Lin et al., 2015), and have potentially weak IVs (Andrews et al., 2018), and invalid IVs (Guo et al., 2018; Windmeijer et al., 2021). Among those problems, we mainly focus on the invalid IV problem, while allowing for potential high-dimensionality and weak signals.

1.1 Related works

Most related works fall into two main categories: robust estimation with invalid IVs and estimation which can select valid IVs without any prior knowledge of validity. The first strand of literature allows all IVs to be invalid. For example, Lewbel (2012), Tchetgen et al. (2021), and Guo and

Received: July 6, 2022. **Revised:** February 26, 2024. **Accepted:** February 26, 2024

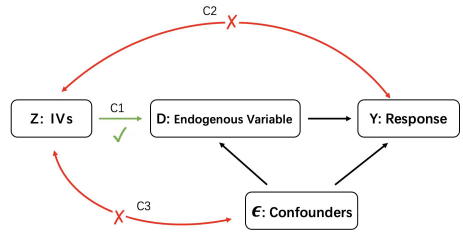

Figure 1. Relevance and validity of instrumental variables (IVs).

Bühlmann (2022) utilized conditional heteroskedasticity or heterogeneous curvatures to achieve robustness with potentially all IVs invalid. However, their performances are not satisfactory once the identification condition is not evident.

The second strand focused on unknown invalid IVs, while imposing certain identification conditions on the number of valid IVs. Kang et al. (2016) proposed a Lasso type estimator (sisVIVE). Windmeijer et al. (2019) pointed out the inconsistent variable selection of sisVIVE under a relative IV strength condition and proposed an adaptive Lasso estimator, which has asymptotic oracle properties under the assumption that more than half of the IVs are valid, also called the majority rule. Guo et al. (2018) and Windmeijer et al. (2021) further developed two-step (the first one for relevance, the second one for validity) selection approaches, two-stage hard thresholding (TSHT) and confidence intervals IV (CIIV), respectively, under the plurality rule conditional on the set of relevant IVs. The plurality rule states that the valid IVs form the largest group. However, the approaches mentioned above are not robust to many weak IVs due to the restriction of the majority/plurality rule amongst the strong IVs instead of all IVs. Our method closely follows this strand of literature. Instead of a two-step selection, we require the plurality rule for valid IVs for a one-step selection procedure, thus considerably relaxing the requirement of valid IVs in theory and most practical scenarios.

The study of many (weak) IVs originated from empirical motivations but often assumed known validity. For example, Staiger and Stock (1997), Hansen et al. (2008), Newey and Windmeijer (2009), and Hansen and Kozbur (2014) considered different estimators for situations with many (weak) valid IVs but fixed the number of known covariates. Kolesár (2018) allowed the number of covariates to grow with the sample size. Seng and Li (2022) introduced an alternative model averaging method to handle weak IVs in low- and high-dimensional settings. We consider the weak IV issues that are prevalent in empirical studies.

1.2 Main results and contributions

We propose a *W*eak and *I*nvalid IV robust *T*reatment effect (WIT) estimator. The sparsest rule is sufficient for identification and is operational in numerical optimization. The proposed procedure has a selection stage (regarding IV validity) and a post-selection estimation stage. The selection stage is a penalized IV-regression via minimax concave penalty (MCP, Zhang, 2010), a proper surrogate penalty aligned with the identification condition to achieve model selection consistency of valid IVs under much weaker technical conditions than existing methods (Guo et al., 2018; Windmeijer et al., 2021). In the estimation stage, we utilize the limited information maximum likelihood (LIML) estimator to handle the weak IVs (Staiger & Stock, 1997). An efficient computational algorithm for the optimal solution is provided in the online supplementary material. The computer codes for implementing the WIT estimator are available at https://github.com/QoifoQ/WIT.

The key contributions of this paper are summarized as follows:

(a) We provide a self-contained framework to investigate the fundamental problem in model identification for linear IV models with unknown validity of instruments. Specifically, we study the identification condition from the general data-generating process (DGP) framework. Furthermore, we discuss the alignment of model identification and variable selection regarding IV validity, which requires a non-convex penalty function.

(b) This study extends the IV estimation with unknown invalid IVs (namely, Guo et al., 2018; Kang et al., 2016; Windmeijer et al., 2019, 2021) to allow for many potentially weak IVs. We show that the sparsest rule, equivalent to the plurality rule on *the whole IV set*, could accommodate weak IVs in empirically relevant scenarios. Furthermore, we revisit the penalized approaches using the sparsest rule and propose a concept of proper surrogate sparsest penalty that targets identification conditions and provides sparse structure. We propose to deploy MCP as a surrogate sparsest penalty and ensure the targeted solution is the global minimizer. On the contrary, the existing methods (Kang et al., 2016; Windmeijer et al., 2019) do not fit the surrogate sparsest penalty and hence are mistargeting the model identification.

(c) Our method is a one-step valid IV selection instead of the previous sequential two-step selections (Guo et al., 2018; Windmeijer et al., 2021). This allows us to utilize individually weak IVs instead of discarding them. We provide theoretical foundations to ensure the compatibility of weak IVs under a mild minimal signal condition. Formally, we establish the selection consistency of the valid IV set, the consistency, and asymptotic normality of the proposed treatment effect estimator under many potentially invalid and weak IVs, where both the number of valid and invalid IVs are increasing with the sample size N. We also provide the theoretical results for the case of a fixed and finite number of IVs. Our model accommodates different rates for IV validity violations which are illustrated through representative low- and high-dimensional cases.

The article is organized as follows. In Section 2, we describe the model with some invalid IVs and analyze identification conditions in a general way. In Section 3, we present the methodology and the novel WIT estimator. We establish the theorems to identify the valid IVs, estimation consistency, and asymptotic normality. Section 4 shows the finite sample performance of our proposed estimator using comprehensive numerical experiments. Section 5 applies our methods to study the effect of body mass index (BMI) on diastolic blood pressure (DBP) using Mendelian Randomization. Section 6 concludes. An online supplementary material contains additional demonstrations and simulations as well as detailed proofs of the theoretical results.

2 Model and identification strategy

2.1 Potential outcome model with some invalid IVs

For $i = 1, 2, \ldots, n$, we have the random sample $(Y_i, D_i, \boldsymbol{Z}_{i\cdot})$, where $Y_i \in \mathbb{R}^1$ is the outcome variable, $D_i \in \mathbb{R}^1$ is the (endogenous) treatment variable and $\boldsymbol{Z}_i \in \mathbb{R}^p$ are the potential IVs. Following the same model setting as in Small (2007), Kang et al. (2016), Guo et al. (2018), and Windmeijer et al. (2019, 2021), we consider a linear functional form between treatments D_i and instruments $\boldsymbol{Z}_{i\cdot}$ as the first-stage specification; meanwhile, a linear exposure of Y_i and D_i and $\boldsymbol{Z}_i$ is assumed as follows:

$$\begin{aligned} Y_i &= D_i\beta^* + \boldsymbol{Z}_{i\cdot}^\top \boldsymbol{\alpha}^* + \epsilon_i, \\ D_i &= \boldsymbol{Z}_{i\cdot}^\top \boldsymbol{\gamma}^* + \eta_i, \end{aligned} \tag{1}$$

where ϵ_i, η_i are random errors.

Remark 1 Assuming a homogeneous treatment effect (denoted as β^*) among subjects simplifies the identification problem in instrumental variable analysis.

Following Kang et al. (2016), we define the valid instruments as follows,

Definition 1 For $j = 1, \ldots, p$, the jth instrument is valid if $\alpha_j^* = 0$.

The validity of the jth IV is quantified by α_j^*, which captures the direct effect of the potential IV $\boldsymbol{Z}_j$ on the outcome (C1) as well as its influence on unmeasured confounders (C2). More details can be found in Kang et al. (2016). Further, we define the valid IV set $\mathcal{V}^* = \{j : \alpha_j^* = 0\}$ and invalid IV set $\mathcal{V}^{c*} = \{j : \alpha_j^* \neq 0\}$. Let $p_{\mathcal{V}^*} = |\mathcal{V}^*|$, $p_{\mathcal{V}^{c*}} = |\mathcal{V}^{c*}|$ and $p = p_{\mathcal{V}^*} + p_{\mathcal{V}^{c*}}$. Notably, $p_{\mathcal{V}^*} \geq 1$ refers to the existence of an excluded IV, thus satisfying the order condition (Wooldridge, 2010). Let the $n \times p$

matrix of observations on the instruments be denoted by $\boldsymbol{Z}$, and the n-vectors of outcomes and treatments by $\boldsymbol{Y}$ and $\boldsymbol{D}$, respectively. We consider the cases of many and weak IVs in equation (1) and make the following model assumptions:

Assumption 1 (Many valid and invalid IVs). $p < n$, $p_{\mathcal{V}^c*}/n \to \upsilon_{p_{\mathcal{V}^c*}} + o(n^{-1/2})$, and $p_{\mathcal{V}^*}/n \to \upsilon_{p_{\mathcal{V}^*}} + o(n^{-1/2})$ for some non-negative constants $\upsilon_{p_{\mathcal{V}^c*}}$ and $\upsilon_{p_{\mathcal{V}^*}}$ such that $0 \leq \upsilon_{p_{\mathcal{V}^*}} + \upsilon_{p_{\mathcal{V}^c*}} < 1$.

Assumption 2 Assume $\boldsymbol{Z}$ is standardized. It then has full column rank and $\|\boldsymbol{Z}_j\|_2^2 \leq n$ for $j = 1, 2, \ldots, p$.

Assumption 3 Let $\boldsymbol{u}_i = (\epsilon_i, \eta_i)^\top$. $\boldsymbol{u}_i \mid \boldsymbol{Z}_i$ are i.i.d. and follow a multivariate normal distribution with mean zero and positive definite covariance matrix $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_\epsilon^2 & \sigma_{\epsilon,\eta} \\ \sigma_{\epsilon,\eta} & \sigma_\eta^2 \end{pmatrix}$. The elements of $\boldsymbol{\Sigma}$ are finite and $\sigma_{\epsilon,\eta} \neq 0$.

Assumption 4 (Strength of valid IVs). The concentration parameter μ_n grows at the same rate as n, i.e. $\mu_n := \boldsymbol{\gamma}_{\mathcal{V}^*}^{*\top} \boldsymbol{Z}_{\mathcal{V}^*}^\top M_{\boldsymbol{Z}_{\mathcal{V}^c*}} \boldsymbol{Z}_{\mathcal{V}^*} \boldsymbol{\gamma}_{\mathcal{V}^*}^* / \sigma_\eta^2 \to \mu_0 n$, for some $\mu_0 > 0$.

Assumption 1 is identical to the assumption of many instruments in Kolesár (2018). It relaxes the conventional many IVs assumptions (Bekker, 1994; Chao & Swanson, 2005) that only allow the dimension of valid IVs $p_{\mathcal{V}^*}$ to grow with n. Also, it has not been considered in the literature on selecting valid IVs (Guo et al., 2018; Kang et al., 2016; Windmeijer et al., 2019, 2021). Assumption 2 is standard for data pre-processing and scaling $\boldsymbol{Z}_i$. Assumption 3 follows Guo et al. (2018) and Windmeijer et al. (2021) to impose the homoskedasticity assumption and endogeneity of treatment D_i.

Remark 2 Under the setting of many valid and invalid IVs, i.e. $\upsilon_{p_{\mathcal{V}^c*}} \neq 0$ and $\upsilon_{p_{\mathcal{V}^c}} \neq 0$, the homoskedasticity assumption is necessary for the LIML estimator in the estimation stage (see details in Section 3) to be consistent. In the low-dimensional case, i.e. $\upsilon_{p_{\mathcal{V}^c*}} = \upsilon_{p_{\mathcal{V}^c}} = 0$, we can relax Assumption 3 and use a heteroskedasticity robust version of the TSLS estimator. The normality condition in Assumption 3 is not required for the selection stage; it is required only for the estimation stage (asymptotic results of the embedded LIML estimator). Relaxing the normality assumption could impact the formula of standard errors and its consistent variance estimator (Kolesár, 2018).

Assumption 4 implies a strong identification condition in terms of the concentration parameter (Bekker, 1994; Newey & Windmeijer, 2009). In the fixed p case, it indicates the presence of a constant coefficient γ_j, $\exists j$, and the rest of IVs could be weak (Staiger & Stock, 1997). Specifically, we model weakly correlated IVs as $\gamma = Cn^{-\tau}$, $0 < \tau \leq 1/2$, which is the 'local to zero' case (Staiger & Stock, 1997). Essentially this is a mixture of constant γ-type and asymptotically diminishing γ-type instruments for fixed p. For $p \to \infty$ in the same order as n, we allow all the IVs to be weak in the 'local to zero' case with specified rates. This IV strength assumption can be further weakened along the lines of Hansen et al. (2008) to have weak identification asymptotics. In this paper, we focus on the individually weak (diminishing to zero as in Staiger & Stock, 1997) signals model in high-dimensionality. Notice our model allows for much weaker individually weak IVs regardless of their validity (as long as the concentration parameter satisfies Assumption 4), unlike that of Guo et al. (2018). Nevertheless, the constant μ_0 can be a small number to accommodate empirically relevant finite samples with many individually weak IVs.

2.2 Identifiability of Model (1)

The following moment conditions can be derived from Model (1):

$$E\left(\boldsymbol{Z}^\top(\boldsymbol{D} - \boldsymbol{Z}\boldsymbol{\gamma}^*)\right) = 0, \quad E\left(\boldsymbol{Z}^\top(\boldsymbol{Y} - \boldsymbol{D}\beta^* - \boldsymbol{Z}\boldsymbol{\alpha}^*)\right) = 0 \Rightarrow \boldsymbol{\Gamma}^* = \boldsymbol{\alpha}^* + \beta^*\boldsymbol{\gamma}^*, \tag{2}$$

where $\mathbf{\Gamma}^* = E(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} E(\boldsymbol{Z}^\top \boldsymbol{Y})$ and $\boldsymbol{\gamma}^* = E(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} E(\boldsymbol{Z}^\top \boldsymbol{D})$, both are identified by the reduced-form models. Without the exact knowledge about which IVs are valid, Kang et al. (2016) considered the identification of $(\boldsymbol{\alpha}^*, \beta^*)$ via the unique mapping of

$$\beta_j^* = \mathbf{\Gamma}_j^* / \boldsymbol{\gamma}_j^* = \beta^* + \alpha_j^* / \gamma_j^*. \tag{3}$$

Notice that the moment conditions (2) consist of p equations, but $(\boldsymbol{\alpha}^*, \beta^*) \in \mathbb{R}^{p+1}$ need to be estimated and is hence under-identified without further restrictions. Kang et al. (2016) proposed a sufficient condition, called majority rule (first proposed by Han, 2008), such that $p_{\mathcal{V}^*} \geq \lceil p/2 \rceil$, to identify the model parameters without any prior knowledge of the validity of individual IVs. However, the majority rule could be restrictive in practice. Guo et al. (2018) further relaxed it to the plurality rule as follows:

$$\text{Plurality Rule}: \quad \left| \mathcal{V}^* = \left\{ j : \alpha_j^* / \gamma_j^* = 0 \right\} \right| > \max_{c \neq 0} \left| \left\{ j : \alpha_j^* / \gamma_j^* = c \right\} \right|, \tag{4}$$

which was stated as an 'if and only if' condition of identification of $(\boldsymbol{\alpha}^*, \beta^*)$.

We re-examine the identifiability problem from the model DGP perspective. Given first-stage information: $\{\boldsymbol{D}, \boldsymbol{Z}, \boldsymbol{\gamma}^*\}$, without loss of generality, we denote the true DGP with $\{\beta^*, \boldsymbol{\alpha}^*, \boldsymbol{\epsilon}\}$ in equation (1) as DGP $\mathcal{P}_0$ that generates $\boldsymbol{Y}$. Given this $\mathcal{P}_0$, for $j \in \mathcal{V}^{c*}$, we obtain an alternative representation: $\boldsymbol{Z}_j \alpha_j^* = \frac{\alpha_j^*}{\gamma_j^*} (\boldsymbol{D} - \sum_{l \neq j} \boldsymbol{Z}_l \gamma_l^* - \boldsymbol{\eta})$. Denote $\mathcal{I}_c = \{ j \in \mathcal{V}^{c*} : c = \alpha_j^* / \gamma_j^* \}$, where $c \neq 0$. For compatibility, we denote $\mathcal{I}_0 = \mathcal{V}^*$. Thus, we can reformulate $\boldsymbol{Y} = \boldsymbol{D}\beta^* + \boldsymbol{Z}\boldsymbol{\alpha}^* + \boldsymbol{\epsilon}$ in equation (1) to:

$$\boldsymbol{Y} = \boldsymbol{D}\tilde{\beta}^c + \boldsymbol{Z}\tilde{\boldsymbol{\alpha}}^c + \tilde{\boldsymbol{\epsilon}}^c, \tag{5}$$

where $\{\tilde{\beta}^c, \tilde{\boldsymbol{\alpha}}^c, \tilde{\boldsymbol{\epsilon}}^c\} = \{\beta^* + c, \boldsymbol{\alpha}^* - c\boldsymbol{\gamma}^*, \boldsymbol{\epsilon} - c\boldsymbol{\eta}\}$, for some $j \in \mathcal{V}^{c*}$. Evidently, for different $c \neq 0$, it forms different DGPs $\mathcal{P}_c = \{\tilde{\beta}^c, \tilde{\boldsymbol{\alpha}}^c, \tilde{\boldsymbol{\epsilon}}^c\}$ that can generate *the same* $\boldsymbol{Y}$ (given $\boldsymbol{\epsilon}$), which also satisfies the moment condition (2) as $\mathcal{P}_0$ since $E(\boldsymbol{Z}^\top \tilde{\boldsymbol{\epsilon}}^c) = \mathbf{0}$. Building on the argument of Guo et al. (2018), Theorem 1, the additional number of potential DGPs satisfying the moment condition (2) is the number of distinguished $c \neq 0$ for $j \in \mathcal{V}^{c*}$. We formally state this result in the following theorem.

Theorem 1 Suppose Assumptions 1–3 hold, given $\mathcal{P}_0$ and $\{\boldsymbol{D}, \boldsymbol{Z}, \boldsymbol{\gamma}^*, \boldsymbol{\eta}\}$, it can only produce additional $G = |\{c \neq 0 : \alpha_j^* / \gamma_j^* = c, j \in \mathcal{V}^{c*}\}|$ groups of different $\mathcal{P}_c$ such that $\mathcal{V}^* \cup \{ \cup_{c \neq 0} \mathcal{I}_c \} = \{1, 2 \ldots, p\}$, $\mathcal{V}^* \cap \mathcal{I}_c = \emptyset$ for any $c \neq 0$ and $\mathcal{I}_c \cap \mathcal{I}_{\tilde{c}} = \emptyset$ for $c \neq \tilde{c}$, and $E(\boldsymbol{Z}^\top \tilde{\boldsymbol{\epsilon}}^c) = \mathbf{0}$. The sparsity structure regarding $\boldsymbol{\alpha}$ is non-overlapping for different solutions.

Theorem 1 shows there is a collection of model DGPs (different parametrizations)

$$\mathcal{Q} = \{\mathcal{P} = \{\beta, \boldsymbol{\alpha}, \epsilon\} : \boldsymbol{\alpha} \text{ is sparse}, E(\boldsymbol{Z}^\top \boldsymbol{\epsilon}) = \mathbf{0}\} \tag{6}$$

corresponding to the same observation $\boldsymbol{Y}$ conditional on first-stage information. Given some $\mathcal{P}_0$, there are additional $1 \leq G \leq p_{\mathcal{V}^{c*}}$ equivalent DGPs. All members in $\mathcal{Q}$ are related through the transformation procedure (5) and $1 < |\mathcal{Q}| = G + 1 \leq p$. Notably, the non-overlapping sparse structure among all possible DGPs leads to the sparsest model regarding $\boldsymbol{\alpha}^*$ being equivalent to plurality rule $|\mathcal{V}^*| > \max_{c \neq 0} |\mathcal{I}_c|$ in the whole set of IVs.

2.3 The sparsest ($\boldsymbol{\alpha}$) rule

The sparsest rule is conceptually equivalent to the plurality rule on the whole IV set, considering the non-overlapping sparse solutions given by Theorem 1. To relax the majority rule, Guo et al. (2018) proposed to use the plurality rule based on the relevant IV set:

$$\left| \mathcal{V}_{\mathcal{S}^*}^* = \left\{ j \in \mathcal{S}^* : \alpha_j^* / \gamma_j^* = 0 \right\} \right| > \max_{c \neq 0} \left| \left\{ j \in \mathcal{S}^* : \alpha_j^* / \gamma_j^* = c \right\} \right|, \tag{7}$$

where $\mathcal{S}^*$ is the set of strong IVs estimated by $\hat{\mathcal{S}}$ via first-step hard thresholding. Thus, TSHT and CIIV explicitly leverage the $\hat{\mathcal{S}}$-based plurality rule to estimate $\mathcal{V}^*_{\hat{\mathcal{S}}}$ and β^*.

In contrast to earlier studies on invalid IVs, our approach leverages the information from weak IVs. First, weak IVs can be employed to estimate β^*. In situations where strong IVs are not available, weak IV robust estimators such as LIML prove useful (Andrews et al., 2018). Second, weak IVs can aid in the identification of the valid IVs set, as demonstrated in Theorem 2. When weak IVs are present, the plurality rule applied after the initial selection of strong instruments may yield unstable estimates of $\mathcal{V}^*$, as exemplified in the following scenario.

Example 1 (Weak and invalid IVs). Let $\boldsymbol{\gamma}^* = (\boldsymbol{0.04}_3, \boldsymbol{0.5}_2, 0.2, \boldsymbol{0.1}_4)^\top$ and $\boldsymbol{\alpha}^* = (0_5, 1, \boldsymbol{0.7}_4)^\top$. There are therefore three groups: $\mathcal{I}_0 = \mathcal{V}^* = \{1, 2, 3, 4, 5\}$, $\mathcal{I}_5 = \{6\}$, $\mathcal{I}_7 = \{7, 8, 9, 10\}$, and plurality rule $|\mathcal{I}_0| > \max_{c=5,7}|\mathcal{I}_c|$ holds in the whole IVs set. This set-up satisfies the individually weak IVs in fixed p (discussed later in Corollary 1). Figure 2 shows that the selection of valid IVs by CIIV and TSHT breaks down in finite samples, e.g. for $N \in [500, 2{,}000]$. This is because the solution of the plurality rule after first-stage selection in the finite sample (which may not hold in practice even though it is in theory) is quite sensitive to IV strength and sample sizes. On the other hand, weak IVs also deteriorate the performance of CIIV without first-step selection. Notably, the proposed WIT estimator significantly outperforms others. Figure 3 demonstrates the relevant set $\mathcal{S}^*$ selected by plurality rule-based TSHT and CIIV. It clearly shows $\hat{\mathcal{S}}$ is unstable and changes with sample size, even though the plurality rule holds in the whole IV set.

The mixture of weak and invalid IVs is ubiquitous in practice, especially in many IVs cases. For the sake of using all instruments' information for estimating β^* and identification of $\mathcal{V}^*$, we allow some individual IV strengths to be local to zero (Chao & Swanson, 2005), say $\gamma_j^* \to 0$, or a small constant that cannot pass the first-stage threshold (Guo et al., 2018) unless with a very large sample size. However, we can see that in equation (3), plurality rule-based methods that rely on first-stage selection are problematic, since $\mathcal{I}_0 = \{j : \alpha_j^*/\gamma_j^* = 0\}$ is ill-defined asymptotically due to the problem of '0/0' if γ_j^* is local to zero.

For using weak IVs information and improving finite sample performance, we turn to the sparsest rule that is also operational in computation algorithms. From the multiple DGPs $\mathcal{Q}$, recall $\mathcal{P}_c = \{\tilde{\beta}^c, \tilde{\boldsymbol{\alpha}}^c, \tilde{\boldsymbol{\epsilon}}^c\} = \{\beta^* + c, \boldsymbol{\alpha}^* - c\boldsymbol{\gamma}^*, \boldsymbol{\epsilon} - c\boldsymbol{\eta}\}$, where $\tilde{\boldsymbol{\alpha}}^c_{\mathcal{I}_c} = \mathbf{0}$. For other elements in $\tilde{\boldsymbol{\alpha}}^c$ (corresponding to a different DGP in $\mathcal{Q}$) and $j \in \{j : \alpha_j^*/\gamma_j^* = \tilde{c} \neq c\}$, we obtain

$$|\tilde{\alpha}_j^c| = |\alpha_j^* - c\gamma_j^*| = |\alpha_j^*/\gamma_j^* - c| \cdot |\gamma_j^*| = |\tilde{c} - c| \cdot |\gamma_j^*|. \tag{8}$$

The above $|\tilde{\alpha}_j^c|$ needs to be distinguished from 0 on the ground of the non-overlapping structure stated in Theorem 1. To facilitate the discovery of all solutions in $\mathcal{Q}$, we assume:

Assumption 5 (Separation condition). $|\boldsymbol{\alpha}^*_{\mathcal{V}^{c*}}|_{\min} > \kappa(n)$ and $|\tilde{\alpha}_j^c| > \kappa^c(n)$ for $j \in \{j : \alpha_j^*/\gamma_j^* = \tilde{c} \neq c\}$, where $\kappa(n)$ and $\kappa^c(n)$ are a generally vanishing rate specified by a particular estimator.

The conditions described above are comparable to the 'beta-min' condition (Loh & Wainwright, 2015; van de Geer & Bühlmann, 2009). The rates $\kappa(n)$ and $\kappa^c(n)$ will be further specified in Theorem 2. We aim to facilitate the understanding of these technical conditions by providing practical examples. In Section 3.3, we introduce two commonly encountered DGPs in both low and high dimensions. These examples demonstrate the validity of Assumption 5 without delving into the intricate specifics of the separation (beta-min) condition.

Remark 3 Notably, as shown in equation (8), $|\tilde{\alpha}_j^c| = |\tilde{c} - c| \cdot |\gamma_j^*| > \kappa^c(n)$ depends on the product of $|\tilde{c} - c|$ and $|\gamma_j^*|$. As discussed in Guo et al. (2018), $|\tilde{c} - c|$ cannot

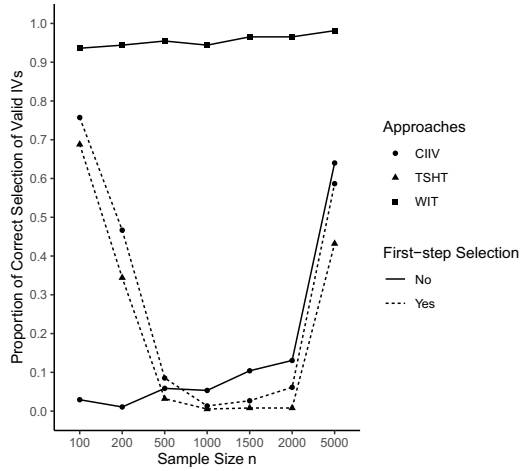

Figure 2. The proportion of correct selection of (subset) valid IVs based on 500 replications on each sample size.

be too small to separate different solutions in $\mathcal{Q}$, and a larger gap $|\tilde{c} - c|$ is helpful to mitigate the problem of small or local to zero $|\gamma_j^*|$ in favour of our model.

Hence, the identification condition known as the sparsest rule is formally defined as

Assumption 6 (The sparsest rule). $\boldsymbol{\alpha}^* = \text{argmin}_{\mathcal{P}=\{\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\epsilon}\}\in\mathcal{Q}} \|\boldsymbol{\alpha}\|_0$.

Example 1 (continued). Following the procedure (5), we are able to reformulate two additional solutions of equation (2) given the DGP of Example 1, $\boldsymbol{\alpha}^* = (0_5, 1, \boldsymbol{0.7}_4)^\top$: $\tilde{\boldsymbol{\alpha}}^5 = (\boldsymbol{-0.2}_3, \boldsymbol{-2.5}_2, 0, \boldsymbol{0.2}_4)^\top$ and $\tilde{\boldsymbol{\alpha}}^7 = (\boldsymbol{-0.28}_3, \boldsymbol{-3.5}_2, -0.4, 0_4)^\top$. Thus, the sparsest rule $\text{argmin}_{\boldsymbol{\alpha}\in\{\boldsymbol{\alpha}^*,\tilde{\boldsymbol{\alpha}}^5,\tilde{\boldsymbol{\alpha}}^7\}}\|\boldsymbol{\alpha}\|_0$ picks $\boldsymbol{\alpha}^*$ up, and Assumption 5 is easy to satisfy since fixed minimum absolute values except 0 are 0.7, 0.2, 0.28 in $\boldsymbol{\alpha}^*$, $\tilde{\boldsymbol{\alpha}}^5$, $\tilde{\boldsymbol{\alpha}}^7$, respectively. This example shows the first-stage signal should not interfere with the valid IV selection in the structural form equation in equation (1), as long as the first-stage has sufficient information (concentration parameter requirement in Assumption 4). Therefore, the most sparse rule using the whole IVs set is desirable. It is also shown to be stable in numerical studies. The detailed performance of the proposed method under this example refers to Case A1(II) in online supplementary Appendix A5.

In the next subsection, we re-examine the penalized approaches by Kang et al. (2016) and Windmeijer et al. (2019), and discuss a class of existing estimators concerning penalization, identification, and computation. We also explore the general penalization approach that aligns model identification with its objective function.

2.4 Penalization approaches with embedded surrogate sparsest rule

A Lasso penalization approach was first used in the unknown IV validity context by Kang et al. (2016). We extend this to a general formulation and discuss the properties of different classes of penalties.

Consider a general penalized estimator based on moment conditions (2),

$$(\widehat{\boldsymbol{\alpha}}^{\text{pen}}, \hat{\beta}^{\text{pen}}) = \underset{\boldsymbol{\alpha},\beta}{\text{argmin}} \underbrace{\frac{1}{2n}\|P_Z(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\alpha} - \boldsymbol{D}\beta)\|_2^2}_{\text{(I)}} + \underbrace{p_\lambda^{\text{pen}}(\boldsymbol{\alpha})}_{\text{(II)}}, \tag{9}$$

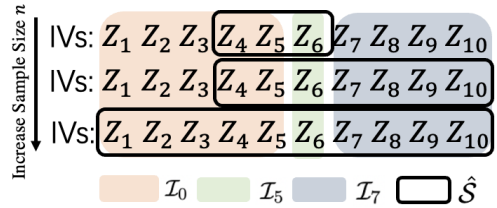

Figure 3. Illustration of plurality rule based on first-step selection.

where $p_\lambda^{\text{pen}}(\boldsymbol{\alpha}) = \sum_{j=1}^{p} p_\lambda^{\text{pen}}(\alpha_j)$ and $p_\lambda^{\text{pen}}(\cdot)$ is a general penalty function with tuning parameter $\lambda > 0$ and $p_\lambda^{\text{pen}'}(\cdot)$ is its derivative that satisfy: $\lim_{x \to 0^+} p_\lambda^{\text{pen}'}(x) = \lambda$, $p_\lambda^{\text{pen}}(0) = 0$, $p_\lambda^{\text{pen}}(x) = p_\lambda^{\text{pen}}(-x)$, $(x-y)(p_\lambda^{\text{pen}}(x) - p_\lambda^{\text{pen}}(y)) \geq 0$, and $p_\lambda^{\text{pen}'}(\cdot)$ is continuous on $(0, \infty)$.

In the RHS of equation (9), (I) and (II) correspond to two requirements for the collection of valid DGPs in $\mathcal{Q}$ defined in equation (6). Part (I) is a scaled finite sample version of $E((\boldsymbol{Z}^\top \boldsymbol{\epsilon})^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} (\boldsymbol{Z}^\top \boldsymbol{\epsilon}))$, which is a $(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}$ weighted quadratic term of condition $E(\boldsymbol{Z}^\top \boldsymbol{\epsilon}) = \boldsymbol{0}$, and (II) is imposed to ensure sparsity structure in $\widehat{\boldsymbol{\alpha}}$.

Further, regarding (I), one can reformulate (9) with respect to $\widehat{\boldsymbol{\alpha}}^{\text{pen}}$ as

$$\widehat{\boldsymbol{\alpha}}^{\text{pen}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \frac{1}{2n} \|\boldsymbol{Y} - \tilde{\boldsymbol{Z}}\boldsymbol{\alpha}\|_2^2 + p_\lambda^{\text{pen}}(\boldsymbol{\alpha}), \tag{10}$$

where $\tilde{\boldsymbol{Z}} = M_{\widehat{\boldsymbol{D}}} \boldsymbol{Z}$ and $\widehat{\boldsymbol{D}} = P_{\boldsymbol{Z}} \boldsymbol{D} = \boldsymbol{Z}\widehat{\boldsymbol{\gamma}}$, where $\widehat{\boldsymbol{\gamma}}$ is the least squares estimator of $\boldsymbol{\gamma}$, see Kang et al. (2016). The design matrix $\tilde{\boldsymbol{Z}}$ is rank-deficient with rank $p-1$ since $\tilde{\boldsymbol{Z}}\widehat{\boldsymbol{\gamma}} = 0$. However, we show that it does not affect the $\boldsymbol{\alpha}$ support recovery using a proper penalty function. On the other hand, $\tilde{\boldsymbol{Z}}$ is a function of $\boldsymbol{\eta}$, $\boldsymbol{\gamma}^*$ and $\boldsymbol{Z}$, hence is correlated with $\boldsymbol{\epsilon}$. This inherited endogeneity initially stems from $\widehat{\boldsymbol{D}}$, in which $E(\widehat{\boldsymbol{D}}^\top \boldsymbol{\epsilon}) = \sigma_{\epsilon,\eta}^2 p/n$ does not vanish in the many IVs model (Assumption 1). The following lemma implies that the level of endogeneity of each $\tilde{\boldsymbol{Z}}_j$ is limited.

Lemma 1 Suppose Assumptions 1–4 hold and denote average gram matrix $\boldsymbol{Q}_n = \boldsymbol{Z}^\top \boldsymbol{Z}/n$. The endogeneity level of the jth transformed IV $\tilde{\boldsymbol{Z}}_j$ follows:

$$\tilde{\boldsymbol{Z}}_j^\top \boldsymbol{\epsilon}/n = \underbrace{\sigma_{\epsilon,\eta}^2 p/n}_{E(\widehat{\boldsymbol{D}}^\top \boldsymbol{\epsilon}/n)} \cdot \underbrace{\frac{\boldsymbol{Q}_{nj}^\top \boldsymbol{\gamma}^*}{\boldsymbol{\gamma}^{*\top} \boldsymbol{Q}_n \boldsymbol{\gamma}^* + \sigma_\eta^2 p/n}}_{\text{dilutionweight}} + O_p(n^{-1/2}). \tag{11}$$

Remark 4 Under Assumption 1, $p/n \to \upsilon_{p_{\mathcal{V}^*}} + \upsilon_{p_{\mathcal{V}^{c*}}} < 1$ does not vanish as $n \to \infty$. The dilution weight is related to $\boldsymbol{Q}_n$ and first-stage signal $\boldsymbol{\gamma}^*$. In general the dilution weight is $o(1)$ and hence negligible except for the existence of dominated γ_j^*. However, in the fixed p case, since $p/n \to 0$, the endogeneity of $\tilde{\boldsymbol{Z}}$ disappears asymptotically.

Concerning (II) in equation (9), Theorem 1 shows that model (1) can be identified by different strategies with non-overlapping results. On the ground of the sparsest rule assumption, the role of the penalty on $\boldsymbol{\alpha}$, i.e. $p_\lambda^{\text{pen}}(\boldsymbol{\alpha})$, should not only impose a sparsity structure but also serve as an objective function corresponding to the identification condition we choose. For example, the penalty $\lambda\|\boldsymbol{\alpha}\|_0$ matches the sparsest rule.

To see the roles of a proper penalty function clearly, we rewrite equation (9) into an equivalent constrained objective function with the optimal penalty $\|\boldsymbol{\alpha}\|_0$ regarding the sparsest rule:

$$(\widehat{\boldsymbol{\alpha}}^{\text{opt}}, \hat{\beta}^{\text{opt}}) = \underset{\boldsymbol{\alpha},\beta}{\operatorname{argmin}} \|\boldsymbol{\alpha}\|_0 \quad \text{s.t. } \|P_{\boldsymbol{Z}}(\boldsymbol{Y} - \boldsymbol{D}\beta - \boldsymbol{Z}\boldsymbol{\alpha})\|_2^2 < \delta, \tag{12}$$

where δ is the tolerance level which we specify in Section 4. The constraint above narrows the feasible solutions into $\mathcal{Q}$ because it aligns with the Sargan test statistic, Sargan (1958), $\|P_Z(\boldsymbol{Y}-\boldsymbol{D}\beta-\boldsymbol{Z}\boldsymbol{\alpha})\|_2^2/\|(\boldsymbol{Y}-\boldsymbol{D}\beta-\boldsymbol{Z}\boldsymbol{\alpha})/\sqrt{n}\|_2^2 = O_p(1)$ under the null hypothesis $E(\boldsymbol{Z}^\top\boldsymbol{\epsilon})=\mathbf{0}$ as required in $\mathcal{Q}$; otherwise, the constraint becomes $O_p(n)$ that cannot be bounded by δ. Thus, a properly chosen δ in equation (12) leads to an equivalent optimization problem

$$(\widehat{\boldsymbol{\alpha}}^{\text{opt}}, \hat{\beta}^{\text{opt}}) = \underset{\mathcal{P}=\{\beta,\boldsymbol{\alpha},\boldsymbol{\epsilon}\}\in\mathcal{Q}}{\operatorname{argmin}} \|\boldsymbol{\alpha}\|_0$$

that matches the identification condition in Assumption 6. Therefore, the primary optimization object in equation (12) should also serve as an identification condition: the sparsest rule.

Remark 5 The constrained optimization perspective provides valuable insights into the issue of invalid IVs in equation (1). Specifically, the penalty term in equation (9) not only imposes a sparse structure but also serves as an identification rule differing from penalized linear regressions without an endogeneity issue.

Due to computational NP-hardness for $\|\boldsymbol{\alpha}\|_0$ in equation (12), a surrogate penalty function is needed. Kang et al. (2016) proposed to replace the optimal l_0-norm with Lasso in equation (9), denoting their estimator sisVIVE as $(\widehat{\boldsymbol{\alpha}}^{\text{sis}}, \hat{\beta}^{\text{sis}})$. And $\mathcal{V}^*$ is estimated as $\hat{\mathcal{V}}^{\text{sis}} = \{j : \widehat{\alpha}_j^{\text{sis}} = 0\}$. However, the surrogate ℓ_1 penalty brings the following issues: (a) Failure in consistent variable selection under some deterministic conditions, namely the sign-aware invalid IV strength (SAIS) condition (Windmeijer et al., 2019, Proposition 2):

$$\left|\widehat{\boldsymbol{\gamma}}_{\mathcal{V}^{c*}}^\top \operatorname{sgn}\left(\boldsymbol{\alpha}_{\mathcal{V}^{c*}}^*\right)\right| > \left\|\widehat{\boldsymbol{\gamma}}_{\mathcal{V}^*}\right\|_1. \tag{13}$$

The sisVIVE cannot achieve $\mathcal{P}_0$ under the SAIS condition, which is likely to hold when the invalid IVs are relatively stronger than valid ones. (b) Kang et al. (2016, Theorem 2) proposed a non-asymptotic error bound $|\hat{\beta}^{\text{sis}} - \beta^*|$ for sisVIVE. The dependence of the restricted isometry property (RIP) constant of $P_{\widehat{\boldsymbol{D}}}\boldsymbol{Z}$ and the error bound is not clear due to the random nature of $\widehat{\boldsymbol{D}}$. (c) The objective function deviates from the original sparsest rule: $g_1(\mathcal{P}) = \|\boldsymbol{\alpha}\|_0$ and $g_2(\mathcal{P}) = \|\boldsymbol{\alpha}\|_1$ correspond to incompatible identification conditions unless satisfying an additional strong requirement

$$\boldsymbol{\alpha}^* = \underset{\mathcal{P}=\{\beta,\boldsymbol{\alpha},\boldsymbol{\epsilon}\}\in\mathcal{Q}}{\operatorname{argmin}} g_j(\mathcal{P}), \quad \forall\, j=1,2 \Leftrightarrow \|\boldsymbol{\alpha}^* - c\boldsymbol{\gamma}^*\|_1 > \|\boldsymbol{\alpha}^*\|_1, \quad \forall\, c\neq 0, \tag{14}$$

which further impedes sisVIVE in estimating $\beta^* \in \mathcal{P}_0$.

Problems (a) and (b) relate to the Lasso problem within invalid IVs, while (c) exposes a deeper issue beyond Lasso: a proper surrogate penalty in equation (10) must align with the identification condition.

Windmeijer et al. (2019) suggested the Adaptive Lasso (Zou, 2006), using as an initial estimator the median estimator of Han (2008) to tackle problem (a)'s SAIS issue. This solution also addresses (c) concurrently. However, it necessitates a more demanding majority rule and requires all IVs to be strong in the fixed p case. Like TSHT and CIIV, it is sensitive to weak IVs.

The following proposition outlines the appropriate surrogate sparsest penalty.

Proposition 1 (The proper surrogate sparsest penalty). Suppose Assumptions 1–6 are satisfied. If $p_\lambda^{\text{pen}}(\boldsymbol{\alpha})$ is the surrogate sparsest rule in the sense that it gives sparse solutions and

$$\boldsymbol{\alpha}^* = \underset{\mathcal{P}=\{\beta,\boldsymbol{\alpha},\boldsymbol{\epsilon}\}\in\mathcal{Q}}{\operatorname{argmin}} \|\boldsymbol{\alpha}\|_0 = \underset{\mathcal{P}=\{\beta,\boldsymbol{\alpha},\boldsymbol{\epsilon}\}\in\mathcal{Q}}{\operatorname{argmin}} p_\lambda^{\text{pen}}(\boldsymbol{\alpha}), \tag{15}$$

then $p_\lambda^{\text{pen}}(\cdot)$ must be concave and $p_\lambda^{\text{pen}\prime}(t) = O(\lambda\kappa(n))$ for any $t > \kappa(n)$.

The surrogate sparsest penalty requirement aligns with the folded-concave penalized method (Fan & Li, 2001; Zhang, 2010). We adopt MCP penalty for our proposed approach in the

following. In standard sparse linear regression, MCP necessitates $\lambda = \lambda(n) = O(\sqrt{\log p/n})$ and $p_\lambda^{\mathrm{MCP}'}(t) = 0$ for $t > C\lambda$, with C a constant, fulfilling Proposition 1. We further elaborate that this property is applicable in invalid IVs scenarios in the following section, and demonstrate its ability to bypass sisVIVE's shortcomings shown in (a) and (b).

Remark 6 Proposition 1 mandates concavity for the surrogate penalty, precluding the use of Adaptive Lasso. However, by imposing additional conditions like the majority rule, which identifies $\boldsymbol{\alpha}^*$, the Adaptive Lasso penalty can satisfy equation (15). It is crucial to differentiate the motivation for using concave penalties in the surrogate sparsest penalty from the one for debiasing techniques such as the debiased Lasso (Javanmard & Montanari, 2018). The latter may not fulfil the identification condition, thus potentially altering the objective function (12).

3 WIT estimator

3.1 Estimation procedure

We implement the penalized regression framework (9) and specifically employ a concave penalty in equation (10), the MCP, which provides nearly unbiased estimation. The MCP is favoured numerically for its superior convexity in penalized loss and its consistent variable selection property without imposing incoherence conditions on the design matrix (Feng & Zhang, 2019; Loh & Wainwright, 2017). This makes MCP more suited to two-stage estimation problems than Lasso.

The selection stage is formally defined as

$$\widehat{\boldsymbol{\alpha}}^{\mathrm{MCP}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \frac{1}{2n} \|Y - \widetilde{\boldsymbol{Z}}\boldsymbol{\alpha}\|_2^2 + p_\lambda^{\mathrm{MCP}}(\boldsymbol{\alpha}), \tag{16}$$

where $\widetilde{\boldsymbol{Z}} = M_{\widehat{\boldsymbol{D}}}\boldsymbol{Z}$, and $\widehat{\boldsymbol{D}} = P_{\boldsymbol{Z}}\boldsymbol{D} = \boldsymbol{Z}\widehat{\boldsymbol{\gamma}}$ are the same as in equation (10). $p_\lambda^{\mathrm{MCP}}(\boldsymbol{\alpha}) = \sum_{j=1}^{p} p_\lambda^{\mathrm{MCP}}(\alpha_j) = \sum_{j=1}^{p} \int_0^{|\alpha_j|} (\lambda - t/\rho)_+ \, \mathrm{d}t$ is the MCP penalty and $\rho > 1$ is the tuning parameter, which also controls convexity level $1/\rho$, and its corresponding derivative is $p_\lambda^{\mathrm{MCP}'}(t) = (\lambda - |t|/\rho)_+$. Unlike Lasso, MCP imposes no penalty when $|\alpha_j| > \lambda\rho$, resulting in nearly unbiased coefficient estimation and exact support recovery without the SAIS condition (see online supplementary Appendix A2 for further discussion). Consequently, a consistent estimation of the valid IVs set, i.e. $\hat{\mathcal{V}} = \{j : \widehat{\alpha}_j^{\mathrm{MCP}} = 0\}$ and $\Pr(\hat{\mathcal{V}} = \mathcal{V}^*) \xrightarrow{p} 1$, is expected to hold under more relaxed conditions. We next illustrate how the WIT blends the benefits of penalized TSLS and LIML estimators at various stages.

The LIML estimator is consistent not only in classic many (weak) IVs (Bekker, 1994; Hansen et al., 2008) but also in many IVs and many included covariates (Kolesár, 2018). However, simultaneous estimation of $\hat{\kappa}_{\mathrm{liml}}$ and $\hat{\mathcal{V}}^c$ in equation (18) below is difficult. In the selection stage (9), we use the moment-based objective function. If we do not consider the penalty term (II), the moment-based part (I) of equation (9) coincides with TSLS. Furthermore, the bias in TSLS has a limited effect on consistent variable selections (see Theorem 2). In the estimation step, however, due to LIML's superior finite sample performance and the issue of TSLS in the presence of many (or weak) IVs even when $\mathcal{V}^*$ is known (Chao & Swanson, 2005), we embed the LIML estimator to estimate β^* on the basis of estimated valid IVs set via equation (16). The performance of oracle-TSLS shown in simulations verifies this choice.

Consequently, we proposed the WIT estimator in the estimation stage,

$$\left(\hat{\beta}^{\mathrm{WIT}}, \hat{\boldsymbol{\alpha}}_{Z_{\hat{\mathcal{V}}^c}}^{\mathrm{WIT}}\right)^\top = \left([\boldsymbol{D}, \boldsymbol{Z}_{\hat{\mathcal{V}}^c}]^\top (I - \hat{\kappa}_{\mathrm{liml}} M_{\boldsymbol{Z}})[\boldsymbol{D}, \boldsymbol{Z}_{\hat{\mathcal{V}}^c}]\right)^{-1} \left([\boldsymbol{D}, \boldsymbol{Z}_{\hat{\mathcal{V}}^c}]^\top (I - \hat{\kappa}_{\mathrm{liml}} M_{\boldsymbol{Z}})\boldsymbol{Y}\right), \tag{17}$$

$$\hat{\kappa}_{\mathrm{liml}} = \min_{\beta} \left\{ G(\beta) = \left((\boldsymbol{Y} - \boldsymbol{D}\beta)^\top M_{\boldsymbol{Z}} (\boldsymbol{Y} - \boldsymbol{D}\beta)\right)^{-1} \left((\boldsymbol{Y} - \boldsymbol{D}\beta)^\top M_{\boldsymbol{Z}_{\hat{\mathcal{V}}^c}} (\boldsymbol{Y} - \boldsymbol{D}\beta)\right) \right\}, \tag{18}$$

Note, equation (17) belongs to the general k-class estimators (Nagar, 1959), whose properties vary upon the choice of $\hat{\kappa}$, i.e. $\hat{\kappa}=0$ refers to the ordinary least squares (OLS) and $\hat{\kappa}=1$ reduces to the TSLS estimator. Equation (18) has a closed-form solution: $\hat{\kappa}_{\text{liml}} = \lambda_{\min}(\{[\boldsymbol{Y},\boldsymbol{D}]^{\top}M_{\boldsymbol{Z}}[\boldsymbol{Y},\boldsymbol{D}]\}^{-1}\{[\boldsymbol{Y},\boldsymbol{D}]^{\top}\ M_{\boldsymbol{Z}_{\hat{\mathcal{V}}^c}}[\boldsymbol{Y},\boldsymbol{D}]\})$, where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue. Focusing on $\hat{\beta}^{\text{WIT}}$ as the primary interest, we reformulate equations (17) and (18) based on the residuals of $\boldsymbol{Y},\boldsymbol{D},\boldsymbol{Z}_{\hat{\mathcal{V}}}$ on $\boldsymbol{Z}_{\hat{\mathcal{V}}^c}$. Denote $\boldsymbol{Y}_{\perp}=M_{\boldsymbol{Z}_{\hat{\mathcal{V}}^c}}\boldsymbol{Y}$, $\boldsymbol{D}_{\perp}=M_{\boldsymbol{Z}_{\hat{\mathcal{V}}^c}}\boldsymbol{D}$ and $\boldsymbol{Z}_{\hat{\mathcal{V}}\perp}=M_{\boldsymbol{Z}_{\hat{\mathcal{V}}^c}}\boldsymbol{Z}_{\hat{\mathcal{V}}}$ and notice $M_{\boldsymbol{Z}_{\hat{\mathcal{V}}^c}}M_{\boldsymbol{Z}_{\hat{\mathcal{V}}\perp}}=M_{\boldsymbol{Z}}$, thus it is equivalent to derive asymptotic results on the following model (19).

$$\hat{\beta}^{\text{WIT}}=\left(\boldsymbol{D}_{\perp}^{\top}\left(\boldsymbol{I}-\hat{\kappa}_{\text{liml}}M_{\boldsymbol{Z}_{\hat{\mathcal{V}}\perp}}\right)\boldsymbol{D}_{\perp}\right)^{-1}\left(\boldsymbol{D}_{\perp}^{\top}\left(\boldsymbol{I}-\hat{\kappa}_{\text{liml}}M_{\boldsymbol{Z}_{\hat{\mathcal{V}}\perp}}\right)\boldsymbol{Y}_{\perp}\right), \tag{19}$$

$$\hat{\kappa}_{\text{liml}}=\lambda_{\min}\left(\{[\boldsymbol{Y}_{\perp},\boldsymbol{D}_{\perp}]^{\top}M_{\boldsymbol{Z}_{\hat{\mathcal{V}}\perp}}[\boldsymbol{Y}_{\perp},\boldsymbol{D}_{\perp}]\}^{-1}\{[\boldsymbol{Y}_{\perp},\boldsymbol{D}_{\perp}]^{\top}[\boldsymbol{Y}_{\perp},\boldsymbol{D}_{\perp}]\}\right). \tag{20}$$

3.2 Asymptotic behaviour of WIT estimator

Throughout this section, we aim to recover the one specific element in $\mathcal{Q}$, denoted as $(\beta^*,\boldsymbol{\alpha}^*,\boldsymbol{\epsilon})$. Though a slight abuse of notation, we use $\widehat{\boldsymbol{\alpha}}$ to denote a local solution of equation (16) with MCP.

All local solutions of equation (16) we consider are characterized by the Karush–Kuhn–Tucker or first-order condition, i.e.

$$\widetilde{\boldsymbol{Z}}^{\top}(\boldsymbol{Y}-\widetilde{\boldsymbol{Z}}\widehat{\boldsymbol{\alpha}})/n=\frac{\partial}{\partial \boldsymbol{t}}\sum_{j=1}^{p}p_{\lambda}'(t_j)\Big|_{\boldsymbol{t}=\widehat{\boldsymbol{\alpha}}}. \tag{21}$$

Explicitly, to the end of finding valid IVs via comparing with true signal $\boldsymbol{\alpha}^*$, we rewrite equation (21) as

$$\begin{cases}\left(\lambda-\frac{1}{\rho}|\widehat{\alpha}_j|\right)_{+}\leqslant \operatorname{sgn}(\widehat{\alpha}_j)\tilde{\boldsymbol{Z}}_j^{\top}(\boldsymbol{Y}-\tilde{\boldsymbol{Z}}\widehat{\boldsymbol{\alpha}})/n\leqslant\lambda, & j\in\hat{\mathcal{V}}^c\\ \left|\tilde{\boldsymbol{Z}}_j^{\top}(\boldsymbol{Y}-\tilde{\boldsymbol{Z}}\widehat{\boldsymbol{\alpha}})/n\right|\leqslant\lambda, & j\in\hat{\mathcal{V}},\end{cases} \tag{22}$$

where the inequalities in the first line stem from the convexity of the MCP penalty and in the last line originate in the sub-derivative of the MCP penalty at the origin.

As discussed in Section 2, $\tilde{\boldsymbol{Z}}$ is a function of $(\boldsymbol{Z},\boldsymbol{\gamma}^*,\boldsymbol{\eta})$ and thus endogenous with $\boldsymbol{\epsilon}$. The fact that $\tilde{\boldsymbol{Z}}$ inherits the randomness of $\boldsymbol{\eta}$ distinguishes itself from the general assumptions put on the design matrix, and obscures the feasibility of conditions required to achieve exact support recovery in the literature of penalized least squares estimator (Feng & Zhang, 2019; Loh & Wainwright, 2017; Zhang & Zhang, 2012). The sisVIVE method imposed the RIP condition directly on $\tilde{\boldsymbol{Z}}$ to establish an error bound. However, the restricted eigenvalue (RE) condition (Bickel et al., 2009) is the weakest condition (van de Geer & Bühlmann, 2009) available to guarantee rate minimax performance in prediction and coefficient estimation, as well as to establish variable selection consistency for Lasso penalty. Feng and Zhang (2019) further adopted the RE condition for non-convex penalty analysis. We then state the conditions on the design matrix $\tilde{\boldsymbol{Z}}$ of equation (16) in the following. Define restricted cone $C(\mathcal{V}^*;\xi)=\{\boldsymbol{u}:\|\boldsymbol{u}_{\mathcal{V}^*}\|_1\leq\xi\|\boldsymbol{u}_{\mathcal{V}^{c*}}\|_1\}$ for some $\xi>0$ that estimation error $\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*$ belongs to. The restricted eigenvalue K_C for $\tilde{\boldsymbol{Z}}$ is defined as $K_C=K_C(\mathcal{V}^*,\xi):=\inf_{\boldsymbol{u}}\{\|\tilde{\boldsymbol{Z}}\boldsymbol{u}\|_2/(\|\boldsymbol{u}\|_2 n^{1/2}):\boldsymbol{u}\in C(\mathcal{V}^*;\xi)\}$ and the RE condition refers to the condition that K_C for $\tilde{\boldsymbol{Z}}$ should be bounded away from zero.

Lemma 2 (RE condition of $\tilde{\boldsymbol{Z}}$). Under Assumptions 1–3, there exists a constant $\xi\in(0,\|\widehat{\boldsymbol{\gamma}}_{\mathcal{V}^*}\|_1/\|\widehat{\boldsymbol{\gamma}}_{\mathcal{V}^{c*}}\|_1)$ and a restricted cone $C(\mathcal{V}^*;\xi)$ defined by chosen ξ such that $K_C^2>0$ holds strictly.

Lemma 2 elaborates that the RE condition on $\tilde{\boldsymbol{Z}}$ holds without any additional assumptions on $\tilde{\boldsymbol{Z}}$, unlike the extra RIP condition for sisVIVE. Moreover, this restricted cone is invariant to scaling,

thus indicating the accommodation of many weak IVs. These two features suggest the theoretical advantages of penalized methods (10) over existing methods.

Next, we discuss the selection of valid IVs by comparing the local solution of equation (22) with the oracle (moment-based) counterpart. Define $\widehat{\boldsymbol{\alpha}}^{\text{or}}_{\mathcal{V}^*} = \mathbf{0}$ and

$$\widehat{\boldsymbol{\alpha}}^{\text{or}}_{\mathcal{V}^{c*}} = (\tilde{\boldsymbol{Z}}^{\top}_{\mathcal{V}^{c*}}\tilde{\boldsymbol{Z}}_{\mathcal{V}^{c*}})^{-1}\tilde{\boldsymbol{Z}}^{\top}_{\mathcal{V}^{c*}}\boldsymbol{Y} \quad \text{or} \quad \widehat{\boldsymbol{\alpha}}^{\text{or}}_{\mathcal{V}^{c*}} = (\boldsymbol{Z}^{\top}_{\mathcal{V}^{c*}}\boldsymbol{Z}_{\mathcal{V}^{c*}})^{-1}[\boldsymbol{Z}^{\top}_{\mathcal{V}^{c*}}(\boldsymbol{Y} - \widehat{\boldsymbol{D}}\hat{\beta}^{\text{TSLS}}_{\text{or}})], \tag{23}$$

where $\hat{\beta}^{\text{TSLS}}_{\text{or}} = [\boldsymbol{D}^{\top}(P_{\boldsymbol{Z}} - P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{D}]^{-1}[\boldsymbol{D}^{\top}(P_{\boldsymbol{Z}} - P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{Y}]$ and the two versions of $\widehat{\boldsymbol{\alpha}}^{\text{or}}_{\mathcal{V}^{c*}}$ are equivalent. Notice this $\hat{\beta}^{\text{TSLS}}_{\text{or}}$ is not for the final treatment effect estimation, but to illustrate the selection stage consistency only. To this end, we show the supremum norm of $\boldsymbol{R}^{\text{or}} = \widetilde{\boldsymbol{Z}}^{\top}(\boldsymbol{Y} - \widetilde{\boldsymbol{Z}}\widehat{\boldsymbol{\alpha}}^{\text{or}})/n$ is bounded by an inflated order of $O(\sqrt{\log p_{\mathcal{V}^*}/n})$. Denote $\tilde{\tilde{\boldsymbol{D}}} = (P_{\boldsymbol{Z}} - P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{D}$ and $\tilde{\tilde{\boldsymbol{Q}}}_n = \boldsymbol{Z}^{\top}_{\mathcal{V}^*}(P_{\boldsymbol{Z}} - P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{Z}_{\mathcal{V}^*}/n$, we derive the following lemma.

Lemma 3 Suppose Assumptions 1–4 hold and let

$$\zeta \asymp \frac{p_{\mathcal{V}^*}}{n} \cdot \frac{\|\tilde{\tilde{\boldsymbol{Q}}}_n \boldsymbol{\gamma}^*_{\mathcal{V}^*}\|_{\infty}}{\boldsymbol{\gamma}^{*\top}_{\mathcal{V}^*}\tilde{\tilde{\boldsymbol{Q}}}_n\boldsymbol{\gamma}^*_{\mathcal{V}^*}} + \sqrt{\frac{\log p_{\mathcal{V}^*}}{n}}. \tag{24}$$

Then, the supremum norms of residual $\boldsymbol{R}^{\text{or}}$ are bounded by ζ, i.e.

$$\|\boldsymbol{R}^{\text{or}}\|_{\infty} \le \left\| \frac{\boldsymbol{Z}^{\top}_{\mathcal{V}^*}\tilde{\tilde{\boldsymbol{\epsilon}}}}{n} \right\|_{\infty} + \left\| \frac{\dfrac{\boldsymbol{Z}^{\top}_{\mathcal{V}^*}\tilde{\tilde{\boldsymbol{D}}}}{n} \cdot \dfrac{\tilde{\tilde{\boldsymbol{D}}}^{\top}\boldsymbol{\epsilon}}{n}}{\dfrac{\tilde{\tilde{\boldsymbol{D}}}^{\top}\tilde{\tilde{\boldsymbol{D}}}}{n}} \right\|_{\infty} \le \zeta \tag{25}$$

holds with probability approaching 1.

Based on Lemmas 2 and 3, we consider the set $B(\lambda, \rho) = \{\widehat{\boldsymbol{\alpha}}$ in equation (22) $: \lambda \ge \zeta, \rho > K_C^{-2}(\mathcal{V}^*, \xi) \vee 1\}$, in which ζ is defined in equation (24) and ξ is guaranteed by Lemma 2, as a collection of all local solutions $\widehat{\boldsymbol{\alpha}}$ computed in equation (22) through a broad class of MCP under a certain penalty level λ and convexity $1/\rho$. Given that the computed local solutions in practice are through a discrete path with some starting point (see online supplementary Appendix A3), we further consider the computable solution set $B_0(\lambda, \rho)$, introduced by Feng and Zhang (2019), i.e.

$$B_0(\lambda, \rho) = \{\widehat{\boldsymbol{\alpha}} : \widehat{\boldsymbol{\alpha}} \text{ and starting point } \widehat{\boldsymbol{\alpha}}^{(0)} \text{ are connected in } B(\lambda, \rho)\}. \tag{26}$$

The connection between $B_0(\lambda, \rho)$ and $B(\lambda, \rho)$ is that $\exists\, \widehat{\boldsymbol{\alpha}}^{(l)} \in B(\lambda, \rho)$ with penalty level $\lambda^{(l)}$ increasing with the index $l = 1, 2, \ldots$, such that $\widehat{\boldsymbol{\alpha}}^{(0)} - \boldsymbol{\alpha}^* \in C(\mathcal{V}^*, \xi)$, $\widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\alpha}}^{(l)}$ for large enough l and $\|\widehat{\boldsymbol{\alpha}}^{(l)} - \widehat{\boldsymbol{\alpha}}^{(l-1)}\|_1 < a_0\lambda^{(l)}$, where a_0 is specified in online supplementary Lemma A3 of Appendix B7. Thus, $B_0(\lambda, \rho)$ is a collection of approximations of $\boldsymbol{\alpha}$ in all DGPs.

Denote

$$\text{Bias}(\hat{\beta}^{\text{TSLS}}_{\text{or}}) = \frac{\boldsymbol{D}^{\top}(P_{\boldsymbol{Z}} - P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{\epsilon}}{\boldsymbol{D}^{\top}(P_{\boldsymbol{Z}} - P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{D}} = \frac{\boldsymbol{D}^{\top}_{\perp}P_{\boldsymbol{Z}_{\perp}}\boldsymbol{\epsilon}_{\perp}}{\boldsymbol{D}^{\top}_{\perp}P_{\boldsymbol{Z}_{\perp}}\boldsymbol{D}_{\perp}};$$

$$\bar{\boldsymbol{\gamma}}^*_{\mathcal{V}^{c*}} = \boldsymbol{\gamma}^*_{\mathcal{V}^{c*}} + (\boldsymbol{Z}^{\top}_{\mathcal{V}^{c*}}\boldsymbol{Z}_{\mathcal{V}^{c*}})^{-1}\boldsymbol{Z}^{\top}_{\mathcal{V}^{c*}}\boldsymbol{Z}_{\mathcal{V}^*}\boldsymbol{\gamma}^*_{\mathcal{V}^*}.$$

Also, let $\tilde{\tilde{\boldsymbol{Q}}}^{c}_n$ and $\text{Bias}(\hat{\beta}^{c\text{TSLS}}_{\text{or}})$ be defined as $\mathcal{P}_c$ version of $\tilde{\tilde{\boldsymbol{Q}}}_n$ and $\text{Bias}(\hat{\beta}^{\text{TSLS}}_{\text{or}})$. Then, we provide the asymptotic result of selection consistency of WIT.

Theorem 2 (Selection consistency). Specify separation conditions $\kappa(n)$ and $\kappa^c(n)$ in Assumption 5 as

$$\kappa(n) \asymp \underbrace{\sqrt{\frac{\log p_{\mathcal{V}^*}}{n}}}_{T_1} + \underbrace{\frac{p_{\mathcal{V}^*}}{n} \cdot \frac{\|\tilde{\tilde{\boldsymbol{Q}}}_n \boldsymbol{\gamma}^*_{\mathcal{V}^*}\|_\infty}{\boldsymbol{\gamma}^{*\top}_{\mathcal{V}^*} \tilde{\tilde{\boldsymbol{Q}}}_n \boldsymbol{\gamma}^*_{\mathcal{V}^*}}}_{T_2} + \underbrace{|\text{Bias}(\hat{\beta}^{\text{TSLS}}_{\text{or}})| \|\bar{\boldsymbol{\gamma}}^*_{\mathcal{V}^{c*}}\|_\infty}_{T_3}, \tag{27}$$

$$\kappa^c(n) \asymp (1+c)\left\{\sqrt{\frac{\log|\mathcal{I}_c|}{n}} + \frac{|\mathcal{I}_c|}{n} \cdot \frac{\|\tilde{\tilde{\boldsymbol{Q}}}^c_n \boldsymbol{\gamma}^*_{\mathcal{I}_c}\|_\infty}{\boldsymbol{\gamma}^{*\top}_{\mathcal{I}_c} \tilde{\tilde{\boldsymbol{Q}}}^c_n \boldsymbol{\gamma}^*_{\mathcal{I}_c}}\right\} + |\text{Bias}(\hat{\beta}^{c\,\text{TSLS}}_{\text{or}})| \|\bar{\boldsymbol{\gamma}}^*_{\mathcal{I}^c_c}\|_\infty, \tag{28}$$

where $T_1 \to 0$ as $n \to \infty$. Suppose Assumptions 1–6 hold and consider computable local solutions specified in equation (26). Then

$$\widehat{\boldsymbol{\alpha}}^{\text{MCP}} = \underset{\hat{\boldsymbol{\alpha}} \in B_0(\lambda,\rho)}{\text{argmin}} \|\widehat{\boldsymbol{\alpha}}\|_0, \quad \Pr(\hat{\mathcal{V}} = \mathcal{V}^*, \widehat{\boldsymbol{\alpha}}^{\text{MCP}} = \widehat{\boldsymbol{\alpha}}^{\text{or}}) \xrightarrow{p} 1. \tag{29}$$

In Theorem 2, T_1 is similar to a standard rate $\sqrt{\log p/n}$ in penalized linear regression, while T_2 and T_3 are the additional terms that only play a role in the many IVs context and vanish fast in the finite strong IVs case (see Corollary 1). This result is new to the literature.

Proposition 2 Under the same assumptions of Theorem 2, if there does not exist a dominant scaled γ^*_j, i.e. $\|\tilde{\tilde{\boldsymbol{Q}}}^{1/2}_n \boldsymbol{\gamma}^*_{\mathcal{V}^*}\|_\infty / \|\tilde{\tilde{\boldsymbol{Q}}}^{1/2}_n \boldsymbol{\gamma}^*_{\mathcal{V}^*}\|_1 = o(\|\tilde{\tilde{\boldsymbol{Q}}}^{1/2}_n \boldsymbol{\gamma}^*_{\mathcal{V}^*}\|_1 / (p_{\mathcal{V}^*} \|\tilde{\tilde{\boldsymbol{Q}}}^{1/2}_n\|_\infty))$, then $T_2 \to 0$.

Proposition 2 shows that T_2 is limited in the general case where dominant scaled γ^*_j does not exist. For example, if we assume $\boldsymbol{Q}_n = \boldsymbol{I}$ and $\boldsymbol{\gamma}^*_{\mathcal{V}^*} = C\mathbf{1}_{p_{\mathcal{V}^*}}$, where C is a constant or diminishing to zero, then $\|\boldsymbol{\gamma}^*_{\mathcal{V}^*}\|_\infty / \|\boldsymbol{\gamma}^*_{\mathcal{V}^*}\|_1 = o(\|\boldsymbol{\gamma}^*_{\mathcal{V}^*}\|_1 / p_{\mathcal{V}^*}) = o(Cp_{\mathcal{V}^*}/p_{\mathcal{V}^*}) = o(C)$ holds and it follows that $T_2 \to 0$.

Proposition 3 (Approximation of $\text{Bias}(\hat{\beta}^{\text{TSLS}}_{or})$). Let $s = \max(\mu_n, p_{\mathcal{V}^*})$. Under Assumptions 1–4, we obtain

$$E\big[\text{Bias}(\hat{\beta}^{\text{TSLS}}_{\text{or}})\big] = \frac{\sigma_{\epsilon\eta}}{\sigma^2_\eta}\left(\frac{p_{\mathcal{V}^*}}{(\mu_n + p_{\mathcal{V}^*})} - \frac{2\mu_n^2}{(\mu_n + p_{\mathcal{V}^*})^3}\right) + o\big(s^{-1}\big). \tag{30}$$

Remark 7 The rate of concentration parameter μ_n will affect T_3 through $|\text{Bias}(\hat{\beta}^{\text{TSLS}}_{\text{or}})|$ in the many IVs setting. Suppose Assumption 4 holds, that $\mu_n \xrightarrow{p} \mu_0 n$, the leading term in equation (30) is $\frac{\sigma_{\epsilon\eta}}{\sigma^2_\eta} \frac{v_{p_{\mathcal{V}^*}}}{\mu_0 + v_{p_{\mathcal{V}^*}}} \ll \frac{\sigma_{\epsilon\eta}}{\sigma^2_\eta}$ for moderate μ_0 since $0 < v_{p_{\mathcal{V}^*}} < 1$ while μ_0 could be larger than 1. While in the many weak IVs setting (Chao & Swanson, 2005; Hansen et al., 2008; Newey & Windmeijer, 2009), $\mu_n/n \xrightarrow{p} 0$ and the leading term in equation (30) becomes $\sigma_{\epsilon\eta}/\sigma^2_\eta$. Thus, the many weak IVs setting imposes some difficulty (a higher T_3) for selecting valid IVs in Theorem 2.

The following theorem describes the asymptotic behaviour of the WIT estimator for many valid and invalid IVs cases by combining Theorem 1 and invariant likelihood arguments in Kolesár (2018). We further denote two statistics

$$\boldsymbol{S} = \frac{1}{n-p}(\boldsymbol{Y}, \boldsymbol{D})^\top M_Z(\boldsymbol{Y}, \boldsymbol{D}), \quad \boldsymbol{T} = \frac{1}{n}(\boldsymbol{Y}_\perp, \boldsymbol{D}_\perp)^\top M_{Z_\perp}(\boldsymbol{Y}_\perp, \boldsymbol{D}_\perp) \tag{31}$$

as the estimates of the covariance matrix of reduced-form error $\boldsymbol{\Omega} = \mathrm{Cov}(\boldsymbol{\epsilon} + \beta^* \boldsymbol{\eta}, \boldsymbol{\eta})$ and a variant of concentration parameter, respectively. Also, let $m_{\max} = \lambda_{\max}(\boldsymbol{S}^{-1}\boldsymbol{T})$, $\hat{\mu}_n = \max(m_{\max} - p_{\mathcal{V}^*}/n, 0)$ and $\widehat{\boldsymbol{\Omega}} = \frac{n-p}{n-p_{\mathcal{V}^{c*}}/n}\boldsymbol{S} + \frac{n}{n-p_{\mathcal{V}^{c*}}/n}(\boldsymbol{T} - \frac{\hat{\mu}_n}{\hat{\boldsymbol{a}}^\top \boldsymbol{S}^{-1}\hat{\boldsymbol{a}}}\hat{\boldsymbol{a}}\hat{\boldsymbol{a}}^\top)$, where $\hat{\boldsymbol{a}} = (\hat{\beta}^{\mathrm{WIT}}, 1)$ and $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$.

Theorem 3 Under the same conditions as in Theorem 2, we obtain:

(a) (Consistency): $\hat{\beta}^{\mathrm{WIT}} \xrightarrow{p} \beta^*$ with $\hat{\kappa}_{\mathrm{liml}} = \frac{1-v_{p_{\mathcal{V}^*}}}{1-v_{p_{\mathcal{V}^{c*}}}-v_{p_{\mathcal{V}^*}}} + o_p(1)$.

(b) (Asymptotic normality): $\sqrt{n}(\hat{\beta}^{\mathrm{WIT}} - \beta^*) \xrightarrow{d} \mathcal{N}(0, \mu_0^{-2}[\sigma_\epsilon^2 \mu_0 + \frac{v_{p_{\mathcal{V}^{c*}}}(1-v_{p_{\mathcal{V}^*}})}{1-v_{p_{\mathcal{V}^{c*}}}-v_{p_{\mathcal{V}^*}}}|\boldsymbol{\Sigma}|])$.

(c) (Consistent variance estimator):

$$\widehat{\mathrm{Var}}(\hat{\beta}^{\mathrm{WIT}}) = \frac{\hat{\boldsymbol{b}}^\top \widehat{\boldsymbol{\Omega}}\hat{\boldsymbol{b}}(\hat{\mu}_n + p_{\mathcal{V}^*}/n)}{-\hat{\mu}_n}\left(\hat{Q}_S \widehat{\boldsymbol{\Omega}}_{22} - \boldsymbol{T}_{22} + \frac{\hat{c}}{1-\hat{c}}\frac{\hat{Q}_S}{\hat{\boldsymbol{a}}^\top \widehat{\boldsymbol{\Omega}}^{-1}\hat{\boldsymbol{a}}}\right)^{-1}$$
$$\xrightarrow{p} \mu_0^{-2}\left[\sigma_\epsilon^2 \mu_0 + \frac{v_{p_{\mathcal{V}^{c*}}}(1-v_{p_{\mathcal{V}^*}})}{1-v_{p_{\mathcal{V}^{c*}}}-v_{p_{\mathcal{V}^*}}}|\boldsymbol{\Sigma}|\right],$$

where $\hat{\boldsymbol{b}} = (1, -\hat{\beta}^{\mathrm{WIT}})$ and $\hat{Q}_S = \frac{\hat{\boldsymbol{b}}^\top \boldsymbol{T}\hat{\boldsymbol{b}}}{\hat{\boldsymbol{b}}^\top \widehat{\boldsymbol{\Omega}}\hat{\boldsymbol{b}}}$.

Notably, when the number of invalid IVs $p_{\mathcal{V}^*}$ is a constant, the variance estimator above is reduced to the one that Bekker (1994) derived for the typical many IVs case. Hansen et al. (2008) showed that it is still valid under many weak IVs asymptotics.

3.3 Special cases in low and high dimensions

The values of $\kappa(n)$ and $\kappa^c(n)$ outlined in Theorem 2 provide the general formulae to confirm Assumption 5. In this subsection, we discuss some representative cases in low and high dimensions and verify that Assumption 5 holds.

3.3.1 Finite p *case*

First, we show that the WIT estimator is a more powerful tool than the existing methods requiring the majority rule (Kang et al., 2016; Windmeijer et al., 2019) also under the *finite number of IVs with a mixture of strong and weak IVs* settings. WIT achieves the same asymptotic results as Windmeijer et al. (2019) under more relaxed conditions. Specifically, under finite IVs, Assumption 1 can be reduced to Assumption 1′ as follows.

Assumption 1′ (Finite number of IVs). $p_{\mathcal{V}^{c*}} \geq 1$ and $p_{\mathcal{V}^*} \geq 1$ are fixed constants, and $p_{\mathcal{V}^{c*}} + p_{\mathcal{V}^*} = p < n$.

In the finite IVs case, the T_2 and T_3 terms in Theorem 2 for selecting valid IVs go to zero fast and the required separation rate reduces to $\kappa(n) \asymp n^{-1/2}$. We present the asymptotic properties for the WIT estimator here. Consider the following mixed IV strengths case. Let $\gamma_j^* = Cn^{-\tau_j}$ for $j = 1, 2, \ldots, p$, $\tau_{\mathcal{V}^*} = \mathrm{argmax}_{\tau_j}\{\tau_j : j \in \mathcal{V}^*\}$, $\tau_{\mathcal{I}_c} = \mathrm{argmax}_{\tau_j}\{\tau_j : j \in \mathcal{I}_c\}$, and $\tau_{\mathcal{I}_{\tilde{c}}} = \mathrm{argmax}_{\tau_j}\{\tau_j : j \in \mathcal{I}_{\tilde{c}}\}$, where $c \neq \tilde{c}$.

Corollary 1 (Finite p with Mixture of Strong and Weak IVs). Suppose Assumptions 1′, 2–4, and 6 hold. Additionally, we assume each IV is at least a weak one such that $\gamma_j^* = O(n^{-\tau_j})$ and $0 \leq \tau_j \leq 1/2$ for $j = 1, 2, \ldots, p$. For any fixed $\min_{j \in \mathcal{V}^{c*}} |\alpha_j^*| > 0$, if $\tau_{\mathcal{V}^*} + 2\tau_{\mathcal{I}_c} < 1$, $\tau_{\mathcal{V}^*} + 2\tau_{\mathcal{I}_{\tilde{c}}} < 1$ and $\tau_{\mathcal{I}_c} + \tau_{\mathcal{I}_{\tilde{c}}} < 2/3$, then we have

(a) (Selection consistency): $\widehat{\boldsymbol{\alpha}}^{\mathrm{MCP}} = \underset{\widehat{\boldsymbol{\alpha}} \in B_0(\lambda,\rho)}{\mathrm{argmin}} \|\widehat{\boldsymbol{\alpha}}\|_0$, $\Pr(\hat{\mathcal{V}} = \mathcal{V}^*, \widehat{\boldsymbol{\alpha}}^{\mathrm{MCP}} = \widehat{\boldsymbol{\alpha}}^{\mathrm{or}})$

$\xrightarrow{p} 1$.

(b) (Consistency & equivalence of WIT and TSLS): $\hat{\beta}^{\text{WIT}} \xrightarrow{p} \beta^*$ with $\hat{\kappa}_{\text{liml}} = 1 + o_p(1)$.

(c) (Asymptotic normality): $\sqrt{n}(\hat{\beta}^{\text{WIT}} - \beta^*) \xrightarrow{d} \mathcal{N}(0, \mu_0^{-1}\sigma_\epsilon^2)$.

(d) (Consistent variance estimator):

$$\widehat{\text{Var}}(\hat{\beta}^{\text{WIT}}) = \frac{\hat{\boldsymbol{b}}^\top \widehat{\boldsymbol{\Omega}} \hat{\boldsymbol{b}} (\hat{\mu}_n + p_{\mathcal{V}^*}/n)}{-\hat{\mu}_n} \left(\hat{Q}_S \widehat{\boldsymbol{\Omega}}_{22} - \boldsymbol{T}_{22} + \frac{\hat{c}}{1-\hat{c}} \frac{\hat{Q}_S}{\hat{\boldsymbol{a}}^\top \widehat{\boldsymbol{\Omega}}^{-1} \hat{\boldsymbol{a}}} \right)^{-1} \xrightarrow{p} \mu_0^{-1}\sigma_\epsilon^2,$$

where $\hat{\boldsymbol{b}} = (1, -\hat{\beta}^{\text{WIT}})$ and $\hat{Q}_S = \frac{\hat{\boldsymbol{b}}^\top \boldsymbol{T} \hat{\boldsymbol{b}}}{\hat{\boldsymbol{b}}^\top \widehat{\boldsymbol{\Omega}} \hat{\boldsymbol{b}}}$.

Corollary 1 addresses a scenario where the violation of validity, α_j^*, is held constant and thus is of the same order when $\tau_j = 0$, or exceeds its strength parameter, γ_j^*, for $j \in \mathcal{V}_c^*$. This scenario frequently arises when conducting robust estimations, particularly in circumstances that inadvertently incorporate weakly relevant but strongly invalid IVs, or when there is no prior information on the candidate IVs set. Dealing with these situations effectively is a crucial component of robust statistical analysis.

Conversely, as noted by an anonymous reviewer, experienced researchers often select IVs that are argued to be both relevant and valid. A violation of IV validity could happen among relatively strong IVs. This corresponds to a scenario where the ratio of α_j^*/γ_j^* is small. In the following proposition, we demonstrate that the WIT estimator can effectively manage cases where $\alpha_j^*/\gamma_j^* = o(1)$.

Proposition 4 (Finite p with $\alpha_j^*/\gamma_j^* = o(1)$). Suppose Assumptions 1′, 2, 3, and 6 hold, and Assumption 4 holds for each DGP in $\mathcal{Q}$. Each γ_j^* is at least a weak IV such that $\gamma_j^* = O(n^{-\tau_j})$ and $0 \le \tau_j \le 1/2$ for $j = 1, 2, \ldots, p$ and satisfy $\gamma_l^* > \max_{j \in \mathcal{V}^{c*}} O(n^{-1/2}\gamma_j^*/\alpha_j^*)$ for valid $l \in \mathcal{V}^*$. Additionally, we suppose $min_{j \in \mathcal{V}^{c*}} |\alpha_j^*| > O(n^{-1/2})$, $\alpha_j^*/\gamma_j^* = o(1)$. Under these conditions, conclusions (a) to (d) in Corollary 1 remain valid.

Proposition 4 gives the desired asymptotic results for the WIT estimator for DGPs in the $\alpha_j^*/\gamma_j^* = o(1)$ scenario, which aligns with the aforementioned case that some researchers may pick relatively strong IVs in the design phase, and true α_j^* to satisfy the separation condition in Assumption 5. It only requires that $|\gamma_l^*| > \inf_{j \in \mathcal{V}^{c*}} O(n^{-1/2}\gamma_j^*/\alpha_j^*)$, which means the valid IVs' strength signals are stronger than the rate of $n^{-1/2}$ over α_j^*/γ_j^*. It can be easily implied from this condition that a smaller value in ratio of α_j^* to γ_j^* helps to reduce the needed strength signal γ_l^* of the valid IVs. Therefore, $\gamma_l^* = o(\gamma_j^*)$ for $l \in \mathcal{V}^*$ and $j \in \mathcal{V}^{c*}$ is a sufficient condition to verify it. For the second part in Assumption 5: $|\tilde{\alpha}_j^c| > \kappa^c(n)$ for $j \in \{j : \alpha_j^*/\gamma_j^* = \tilde{c} \neq c\}$ will hold automatically under $\alpha_j^*/\gamma_j^* = o(1)$. Table 1 subsequently illustrates a simplified representative example featuring only 4 IVs, two of which are valid, and demonstrates the rates of $\kappa(n)$ and $\kappa^c(n)$ in true and transformed DGPs.

As a result, Corollary 1 and Proposition 4 provide easily interpretable examples of common DGPs across two distinct scenarios. These examples affirm the validity of Assumption 5, facilitating a more straightforward understanding for practitioners.

3.3.2 High-dimensional (many IVs) case

Next, we consider a common scenario (Andrews et al., 2018; Chao & Swanson, 2005) with a large number (can grow with n) of individually weak instruments, where $\gamma_j^* = O(\sqrt{\log p/n^{1-\delta}})$ for $j = 1, 2, \ldots, p$, and some small enough $\delta > 0$ such that $\log n/n^{-\delta} < \infty$ to avoid an exploded variance in the outcome variable Y_i. Regarding invalid IVs, we assume the same rate of $|\boldsymbol{\alpha}_{\mathcal{V}^{c*}}^*|_{\min} = O(\sqrt{\log p/n^{1-\delta}})$.

Table 1. Illustration of separation condition in case of finite p with $\alpha_j^*/\gamma_j^* = o(1)$

	Coefficients				Thresholds	
$\boldsymbol{\alpha}^*$	0	0	α_3^*	α_4^*	$\kappa(n)$	$n^{-1/2}$
$\boldsymbol{\gamma}^*$	γ_1^*	γ_2^*	γ_3^*	γ_4^*	–	–
$\tilde{\boldsymbol{\alpha}}^{c_3}$	$\gamma_1^* \cdot c_3$	$\gamma_2^* \cdot c_3$	0	$\alpha_4^* - \gamma_4^* \cdot c_3$	$\kappa^{c_3}(n)$	$n^{-1/2}$
$\tilde{\boldsymbol{\alpha}}^{c_4}$	$\gamma_1^* \cdot c_4$	$\gamma_2^* \cdot c_4$	$\alpha_3^* - \gamma_3^* \cdot c_4$	0	$\kappa^{c_4}(n)$	$n^{-1/2}$

Note. This example illustrates a scenario with four IVs, where the first two are valid, $c_3 = \alpha_3^*/\gamma_3^*$, $c_4 = \alpha_4^*/\gamma_4^*$, and $c_1 \neq c_2$. The first two rows represent the true DGPs. The last two rows, $\tilde{\boldsymbol{\alpha}}^c$ and $\kappa^c(n)$, correspond to the transformed DGPs according to equation (8) and the threshold defined in the separation condition: Assumption 5. The corresponding $\kappa^c(n)$ has been simplified through the proof presented in Proposition 4. IVs = instrumental variables; DGPs = data-generating processes.

Consequently, the separation condition $\kappa(n)$ is reduced to:

$$\kappa(n) \asymp \sqrt{\frac{\log p_{\mathcal{V}^*}}{n}} + \frac{p_{\mathcal{V}^*}}{n} \cdot \frac{\|\tilde{\tilde{\mathbf{Q}}}_n \boldsymbol{\gamma}^*_{\mathcal{V}^*}\|_\infty}{\boldsymbol{\gamma}^{*\top}_{\mathcal{V}^*} \tilde{\tilde{\mathbf{Q}}}_n \boldsymbol{\gamma}^*_{\mathcal{V}^*}} + |\text{Bias}(\hat{\beta}^{\text{TSLS}}_{\text{or}})| \|\bar{\boldsymbol{\gamma}}^*_{\mathcal{V}^{c*}}\|_\infty = O\left(\sqrt{\frac{\log p_{\mathcal{V}^*}}{n}}\right)$$

as per Proposition (2). This satisfies the first part of Assumption 5, that is, $|\boldsymbol{\alpha}^*_{\mathcal{V}^{c*}}|_{\min} > \kappa(n)$. The second part of Assumption 5, $|\tilde{\alpha}_j^c| > \kappa^c(n)$ for $j \in \{j : \alpha_j^*/\gamma_j^* = \tilde{c} \neq c\}$, can be confirmed according to equation (8),

$$|\tilde{\alpha}_j^c| = |\alpha_j^* - c\gamma_j^*| = |\tilde{c} - c| \cdot |\gamma_j^*| = O\left(\sqrt{\log p / n^{1-\delta}}\right), \tag{32}$$

where $|\tilde{c} - c|$ has the same order of $\alpha_j^*/\gamma_j^* \asymp O(1)$ for $j \in \mathcal{V}_c^*$. The threshold of $\kappa^c(n)$ in equation (28) is now specified as

$$\begin{aligned} \kappa^c(n) &\asymp (1+c)\left\{\sqrt{\frac{\log|\mathcal{I}_c|}{n}} + \frac{|\mathcal{I}_c|}{n} \cdot \frac{\|\tilde{\tilde{\mathbf{Q}}}^c_n \boldsymbol{\gamma}^*_{\mathcal{I}_c}\|_\infty}{\boldsymbol{\gamma}^{*\top}_{\mathcal{I}_c} \tilde{\tilde{\mathbf{Q}}}^c_n \boldsymbol{\gamma}^*_{\mathcal{I}_c}}\right\} + |\text{Bias}(\hat{\beta}^c_{\text{or}}\text{TSLS})| \|\bar{\boldsymbol{\gamma}}^*_{\mathcal{I}_c^c}\|_\infty \\ &\asymp \sqrt{\frac{\log|\mathcal{I}_c|}{n}}. \end{aligned} \tag{33}$$

Therefore, the second part of Assumption 5 holds as well.

We summarize the above discussions as the following proposition.

Proposition 5 Suppose Assumptions 1–4 and 6 hold. If we assume the uniform rate of $\gamma_j^* = \alpha_l^* = O(\sqrt{\log p/n^{1-\delta}})$ for $j = 1, 2, \ldots, p$, $l \in \mathcal{V}_c^*$ and small enough $\delta > 0$ such that $\log n/n^{-\delta} < \infty$, then Assumption 5 is satisfied and the results of Theorems 3 and 4 hold consequently.

4 Numerical simulations

In this section, we conduct numerical studies to evaluate the finite sample performance of the proposed WIT estimator. In the design of the simulation experiments, we consider scenarios corresponding to different empirically relevant problems.

We consider the same model in Section 2,

$$\boldsymbol{Y} = \boldsymbol{D}\beta^* + \boldsymbol{Z}\boldsymbol{\alpha}^* + \boldsymbol{\epsilon}, \quad \boldsymbol{D} = \boldsymbol{Z}\boldsymbol{\gamma}^* + \boldsymbol{\eta}.$$

Throughout all settings, we fix true treatment effect $\beta^* = 1$. $\boldsymbol{Z}$ is the $n \times p$ potential IV matrix and $\boldsymbol{Z}_{i\cdot} \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}^{\boldsymbol{Z}})$, where $\boldsymbol{\Sigma}^{\boldsymbol{Z}}_{jj} = 0.3$ and $\boldsymbol{\Sigma}^{\boldsymbol{Z}}_{jk} = 0.3|j-k|^{0.8}$ for $i = 1, \ldots, n$ and $k, j = 1, \ldots, p$. Denote $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^\top$ and $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_n)^\top$ and generate $(\epsilon_i, \eta_i)^\top \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \begin{pmatrix} \sigma_\epsilon^2 & \sigma_{\epsilon,\eta} \\ \sigma_{\epsilon,\eta} & \sigma_\eta^2 \end{pmatrix})$. We let $\sigma_\epsilon = 0.5$ and $\text{corr}(\epsilon_i, \eta_i) = 0.6$ in all settings but vary σ_η^2 to get different concentration parameters concerning strong or weak IVs cases.

We compare the WIT estimator with other popular estimators in the literature. Specifically, sisVIVE is computed by R package `sisVIVE`; Post-Alasso (Windmeijer et al., 2019), TSHT, and CIIV are implemented using codes on Github (Guo et al., 2018; Windmeijer et al., 2021). TSLS, LIML, oracle-TSLS, and oracle-LIML (the truly valid set $\mathcal{V}^*$ is known a priori) are also included. Regarding our proposed WIT estimator, the modified Cragg–Donald (MCD) tuning strategy is implemented to determine λ, and we fix $\rho = 2$. In the iterative local adaptive majorize-minimization (I-LAMM) algorithm, we take $\delta_c = 10^{-3}$ and $\delta_t = 10^{-5}$ as the tolerance levels. We report results based on 500 simulations.

We measure the performance of all estimators in terms of median absolute deviation (MAD), standard deviation, and coverage probability (CP) based on 95% confidence intervals. Moreover, we provide measurements on the estimation of $\boldsymbol{\alpha}^*$ and IV model selection. Specifically, we measure the performance of invalid IVs selection by false positive rate (FPR) and false negative rate (FNR). To be concrete, denote the number of incorrect selections of valid and invalid IVs as FP and FN, respectively, and the number of correct selections of valid and invalid as TP and TN, respectively. Thus, FPR =FP/(FP +TN) and FNR = FN/(FN + TP).

Here, we present a replication of the simulation design considered in Windmeijer et al. (2021) and its weak IVs variant:

$$\text{Case 1}: \boldsymbol{\gamma}^* = (\boldsymbol{0.4}_{21})^\top \text{ and } \boldsymbol{\alpha}^* = (0_9, \boldsymbol{0.4}_6, \boldsymbol{0.2}_6)^\top.$$
$$\text{Case 2}: \boldsymbol{\gamma}^* = (\boldsymbol{0.15}_{21})^\top \text{ and } \boldsymbol{\alpha}^* = (0_9, \boldsymbol{0.4}_6, \boldsymbol{0.2}_6)^\top.$$

We now vary sample size $N = 500$ to 1,000 and fix $\sigma_\eta = 1$ to strictly follow their design. Between them, Case 1 corresponds to the exact setting, while Case 2 scales down the magnitude of $\boldsymbol{\gamma}^*$ to introduce small coefficients in the first-stage.

Table 2 shows the results. In Case 1, CIIV outperforms TSHT because CIIV can utilize available information better (Windmeijer et al., 2021, Section 7). The WIT estimator performs similar to CIIV and approaches oracle-LIML. sisVIVE and Post-Alasso fail again due to a lack of majority rule. In Case 2, scaling down the first-stage coefficients causes some problems for CIIV and TSHT, since the first-stage selection thresholding $\sigma_\eta\sqrt{2.01 \log p/n} = 0.111 < 0.15$, which might break the plurality rule numerically. TSHT and CIIV perform poorly when $N = 500$ and improve when $N = 1,000$ when the issue of violating the plurality rule is mitigated. Among penalized methods, sisVIVE and Post-Alasso mistarget and perform like TSLS because an additional requirement for sisVIVE (14) and majority rule fail simultaneously. Distinguished from them, the WIT estimator outperforms with acceptable MAD when $N = 500$. The FPR and FNR improve when the sample size increases.

More simulation results, including comprehensive comparisons in various situations, are provided in online supplementary Appendix A5.

5 The effect of BMI on DBP

This section illustrates the usefulness of the proposed WIT estimator for the method of Mendelian Randomization. We implement the WIT estimator to obtain an estimate of the causal effect of BMI on DBP and compare it to the other estimators, OLS, sisVIVE, Post-Alasso, TSHT, and CIIV. This comparative analysis is designed to exemplify the efficiency and robustness of the WIT estimator in various research scenarios.

Table 2. Simulation results in low dimension: a replication of experiment (Windmeijer et al., 2021)

Case	Approaches	$N = 500$				$N = 1{,}000$			
		MAD	CP	FPR	FNR	MAD	CP	FPR	FNR
1	TSLS	0.436	0	–	–	0.435	0	–	–
	LIML	0.729	0	–	–	0.739	0	–	–
	oracle-TSLS	0.021	0.936	–	–	0.014	0.942	–	–
	oracle-LIML	0.021	0.932	–	–	0.014	0.944	–	–
	TSHT	0.142	0.404	0.398	0.150	0.016	0.924	0.023	0.004
	CIIV	0.037	0.710	0.125	0.032	0.017	0.894	0.031	0.002
	sisVIVE	0.445	–	0.463	0.972	0.465	–	0.482	0.999
	Post-Alasso	0.436	0	1	0	0.435	0	0.999	0
	WIT	0.036	0.708	0.121	0.099	0.016	0.910	0.020	0.027
2	TSLS	1.124	0	–	–	1.144	0	–	–
	LIML	1.952	0	–	–	1.976	0	–	–
	oracle-TSLS	0.060	0.936	–	–	0.044	0.942	–	–
	oracle-LIML	0.056	0.948	–	–	0.042	0.962	–	–
	TSHT	0.532	0.058	0.342	0.457	0.155	0.660	0.310	0.208
	CIIV	1.213	0.224	0.337	0.670	0.100	0.574	0.300	0.426
	sisVIVE	1.101	–	0.392	0.936	1.175	–	0.428	0.996
	Post-Alasso	1.112	0	0.945	0.010	1.029	0	0.652	0.205
	WIT	0.102	0.634	0.198	0.220	0.047	0.898	0.051	0.064

Note. CP = coverage probability; FPR = false positive rate; FNR = false negative rate; MAD = mean absolute deviation.

5.1 Data description

In Mendelian Randomization genetic markers, called single nucleotide polymorphisms (SNPs), function as IVs for the identification and estimation of causal effects of modifiable phenotypes on outcomes (Von Hinke et al., 2016). This research design utilizes the random distribution of alleles at conception (Locke et al., 2015).

Nonetheless, SNPs can be invalid IVs. This is primarily due to pleiotropy, which refers to the potential of genetic variants to be associated with multiple phenotypes, but can also be due to influences such as linkage disequilibrium and population stratification. Unfortunately, these violations are typically unidentified prior to their selection. Furthermore, the correlation between SNPs and treatments is often weak, and so we are dealing with the problem of potentially many weak instruments.

We analyzed data from 105,239 individuals from the UK Biobank (http://www.ukbiobank.ac.uk/) to investigate the effect of BMI on DBP. We use the data as in Windmeijer et al. (2021), with a slightly smaller number of individuals due to withdrawals from the study. Following Windmeijer et al. (2019) and Windmeijer et al. (2021), we used 96 SNPs as potential IVs for BMI. To account for skewness, we applied log-transformations to both BMI and DBP. We consider the same model specification detailed in Section 8 of Windmeijer et al. (2021). The model also included age, age squared, and sex as explanatory variables, along with 15 principal components of the genetic relatedness matrix. The corresponding data pre-processing code can be found in the online supplementary Appendix A6.

5.2 Result and analysis

Table 3 provides the estimation results of the effect of log(BMI) on log(DBP).

The OLS estimate of 0.206 is potentially severely biased due to endogeneity issues, such as reverse causality or latent confounders.

Table 3. Empirical results, the effect of log(BMI) on log(DBP)

	$\hat{\beta}$	SE($\hat{\beta}$)	95% CI	# Valid IVs $\widehat{\mathcal{V}}$	# Relevant IVs selected	*p*-value Sargan Test
OLS	0.206	0.002	(0.202, 0.210)	N.A.	–	N.A.
TSLS	0.087	0.016	(0.055, 0.119)	96	–	2.05e–19
TSHT*	0.087	0.016	(0.055, 0.119)	96	–	2.05e–19
TSHT	0.098	0.016	(0.066, 0.130)	61	62	5.29e–14
CIIV*	0.140	0.019	(0.103, 0.177)	83	–	0.011
CIIV	0.174	0.020	(0.135, 0.213)	49	62	0.014
sisVIVE	0.111	N.A.	N.A.	76	–	0.064
Post-Alasso	0.163	0.018	(0.128, 0.198)	85	–	0.013
WIT	0.123	0.020	(0.083,0.163)	81	–	0.140

Note. Sample size $N = 105{,}276$; $p = 96$ potential SNPs (IVs). Two-stage methods were used for both TSHT and CIIV: these methods first select strong IVs and then pick valid IVs from this subset. TSHT* and CIIV* represent the methods without employing a first-stage thresholding process. The sisVIVE does not report standard error (SE) or confidence interval (CI). The symbol '–' denotes that no first-stage selection was performed, and all the original 96 potential IVs were directly used. BMI = body mass index; DBP = diastolic blood pressure; IVs = instrumental variables; SNPs = single nucleotide polymorphisms.

Regarding the IV methods, we first denote TSHT* and CIIV* to indicate the estimation results without first-stage thresholding. For TSLS and TSHT*, both without first-stage thresholding, they yield identical estimates of 0.087. However, the near-zero Sargan test *p*-value strongly rejects the model, implying TSHT's inability to detect potentially invalid IVs among weak ones. Despite the slight improvement of TSHT in discerning invalid IVs through first-stage thresholding, it still leaves some possibly invalid instruments in the model, as indicated by the slightly increased but still very small Sargan test *p*-values.

Compared to TSHT, CIIV identified 13 strong and invalid IVs, leading to an estimate of 0.174 and a Sargan *p*-value of 0.011. CIIV* detected the same (as CIIV) strong invalid IVs with a similar Sargan *p*-value. CIIV* adjusted the estimate to 0.140.

The sisVIVE approach yielded an estimate of 0.111 which is lower than that obtained with the CIIV method. First, we observe that sisVIVE picks 20 invalid IVs (the highest count among comparison methods). Second, a Sargan *p*-value of 0.064 indicates that it overly penalizes invalid IVs while missing some true targets. In contrast, the Post-Alasso estimate is equal to 0.163, with the minimum number of 11 identified invalid IVs, and a Sargan *p*-value of 0.013.

The WIT estimator produced an estimate of 0.123, markedly lower than that of CIIV (and a little lower than CIIV*), accompanied by the highest Sargan test *p*-value of 0.140. The invalid IVs detected by WIT encompassed all the relevant and invalid IVs identified by CIIV without first-stage thresholding. Moreover, compared to CIIV*, WIT effectively identified two additional individually weak and invalid IVs. It thus significantly improved the Sargan *p*-value and estimation while capturing all valid yet weak IVs' information. In hindsight, sisVIVE penalized too many weak and valid IVs, leading to a loss in efficiency. Notably, the minimum validity violation $|\hat{\boldsymbol{\alpha}}^{MCP}|$ is 0.0014, aligned with the average magnitude of first-stage coefficients. In summary, the WIT method minimizes the risk of including invalid IVs while fully utilizing all valid IVs.

To summarize, the WIT estimator serves as a powerful tool for estimating treatment effects in biomedical research using SNPs as potential IVs. Its robustness to invalid and weak IVs makes it highly suitable for Mendelian Randomization applications, where there are potentially many weak IVs with uncertain validity.

6 Conclusion

We extended the study of IV models with unknown invalid IVs to allow for many weak IVs. We provided a complete framework to investigate the identification issue of such models. Sticking to the sparsest rule, we proposed the surrogate sparsest penalty that fits the identification condition. We proposed a novel WIT estimator that addresses the issues that can lead to poor performance of

sisVIVE and Post-Alasso, and can outperform the plurality rule-based TSHT and CIIV. Simulations and real data analysis support the theoretical findings and the advantages of the proposed method over existing approaches.

Acknowledgments

The authors thank two referees, an associate editor and the editors, Aurore Delaigle and Daniela Witten for their useful comments, which helped to improve the paper. Qingliang Fan acknowledges support from the Research Grants Council of Hong Kong, GRF-14500822, and Xinyuan Song from the Research Grants Council of Hong Kong, GRF-14303622.

Conflict of interest: None declared.

Data availability

The data that support the findings of this study are openly available at https://github.com/QoifoQ/WIT.

Supplementary material

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

References

Andrews I., Stock J., & Sun L. (2018). Weak instruments in IV regression: Theory and practice. *Annual Review of Economics*, *11*(1), 727–753. https://doi.org/10.1146/annurev-economics-080218-025643

Bekker P. A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica*, *62*(3), 657–681. https://doi.org/10.2307/2951662

Belloni A., Chen D., Chernozhukov V., & Hansen C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, *80*(6), 2369–2429. https://doi.org/10.3982/ECTA9626

Bickel P. J., Ritov Y., & Tsybakov A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, *37*(4), 1705–1732. https://doi.org/10.1214/08-AOS620

Bound J., Jaeger D. A., & Baker R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of The American Statistical Association*, *90*, 443–450. https://doi.org/10.1080/01621459.1995.10476536

Chao J. C., & Swanson N. R. (2005). Consistent estimation with a large number of weak instruments. *Econometrica*, *73*(5), 1673–1692. https://doi.org/10.1111/ecta.2005.73.issue-5

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

Fan J., & Li R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360. https://doi.org/10.1198/016214501753382273

Fan Q., & Zhong W. (2018). Nonparametric additive instrumental variable estimator: A group shrinkage estimation perspective. *Journal of Business & Economic Statistics*, *36*(3), 388–399. https://doi.org/10.1080/07350015.2016.1180991

Feng L., & Zhang C.-H. (2019). Sorted concave penalized regression. *The Annals of Statistics*, *47*(6), 3069–3098. https://doi.org/10.1214/18-AOS1759

Guo Z., & Bühlmann P. (2022). 'Causal inference with invalid instruments: Exploring nonlinear treatment models with machine learning', arXiv, arXiv:2203.12808, preprint: not peer reviewed.

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

Han C. (2008). Detecting invalid instruments using l_1-GMM. *Economics Letters*, *3*(3), 285–287. https://doi.org/10.1016/j.econlet.2008.09.004

Hansen C., Hausman J., & Newey W. (2008). Estimation with many instrumental variables. *Journal of Business & Economic Statistics*, *26*(4), 398–422. https://doi.org/10.1198/073500108000000024

Hansen C., & Kozbur D. (2014). Instrumental variables estimation with many weak instruments using regularized JIVE. *Journal of Econometrics*, *182*(2), 290–308. https://doi.org/10.1016/j.jeconom.2014.04.022

Javanmard A., & Montanari A. (2018). Debiasing the Lasso: Optimal sample size for Gaussian designs. *Annals of Statistics*, *46*, 2593–2622. https://doi.org/10.1214/17-AOS1630

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. (2018). Minimum distance approach to inference with many instruments. *Journal of Econometrics*, *204*(1), 86–100. https://doi.org/10.1016/j.jeconom.2018.01.004

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

Lin W., Feng R., & Li H. (2015). Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. *Journal of the American Statistical Association*, *110*(509), 270–288. https://doi.org/10.1080/01621459.2014.908125

Locke A. E., Kahali B., Berndt S. I., Justice A. E., Pers T. H., Day F. R., Powell C., Vedantam S., Buchkovich M. L., Yang J., Croteau-Chonka D. C., Esko T., Fall T., Ferreira T., Gustafsson S., Kutalik Z., Luan J., Mägi R., Randall J. C., … Wood A. R . (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, *518*(7538), 197–206. https://doi.org/10.1038/nature14177

Loh P.-L., & Wainwright M. J. (2015). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, *16*, 559–616.

Loh P.-L., & Wainwright M. J. (2017). Support recovery without incoherence: A case for nonconvex regularization. *Annals of Statistics*, *45*, 2455–2482. https://doi.org/10.1214/16-AOS1530

Nagar A. L. (1959). The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica*, *27*(4), 575–595. https://doi.org/10.2307/1909352

Newey W. K., & Windmeijer F. (2009). Generalized method of moments with many weak moment conditions. *Econometrica*, *77*(3), 687–719. https://doi.org/10.3982/ECTA6224

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

Seng L., & Li J. (2022). Structural equation model averaging: Methodology and application. *Journal of Business & Economic Statistics*, *40*(2), 815–828. https://doi.org/10.1080/07350015.2020.1870479

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

Tchetgen E. T., 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

van de Geer S. A., & Bühlmann P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, *3*, 1360–1392. https://doi.org/10.1214/09-EJS506

Von Hinke S., Smith G. D., Lawlor D. A., Propper C., & Windmeijer F. (2016). Genetic markers as instrumental variables. *Journal of Health Economics*, *45*, 131–148. https://doi.org/10.1016/j.jhealeco.2015.10.007

Windmeijer F., Farbmacher H., Davies N., & Smith G. D. (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.

Zhang C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*(2), 894–942. https://doi.org/10.1214/09-AOS729

Zhang C.-H., & Zhang T. (2012). A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, *27*(4), 576–593. https://doi.org/10.1214/12-STS399

Zou H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429. https://doi.org/10.1198/016214506000000735