On the instrumental variable estimation with potentially many (weak) and some invalid instruments

Yiqi Lin[a,b], Frank Windmeijer[b], Xinyuan Song[a], Qingliang Fan[c]†

[a] *Department of Statistics, The Chinese University of Hong Kong, Hong Kong.*

[b] *Department of Statistics, University of Oxford, Oxford, U.K.*

[c] *Department of Economics, The Chinese University of Hong Kong, Hong Kong.*

Abstract. This study extends the instrumental variable (IV) estimation with unknown invalid IVs (namely, Kang et al., 2016; Guo et al., 2018; Windmeijer et al., 2019, 2021) to allow for many (weak) IVs. We discuss the necessary condition of the plurality rule (Guo et al., 2018, Theorem 1) and show that no identification condition can be "if and only if" in general. With the assumption of the "sparsest rule", which is equivalent to the plurality rule and operational in computation algorithms, we investigate the advantage of non-convex penalized approaches over other IV estimators based on two-step selections, in terms of selection consistency and allowing for individually weak IVs. Furthermore, we propose a surrogate sparsest penalty that aligns with the identification condition and provides oracle sparse structure simultaneously. We develop a novel estimator, called WIT, to select valid IVs and improve estimation accuracy under many (weak) IVs with much relaxed individual IV strength conditions. Theoretical properties are derived for the proposed estimator. The finite sample property is demonstrated on simulated data and an empirical study concerning the effect of trade on economic growth.

Keywords: Model Identification, Invalid Instruments, Weak Instruments, Treatment Effect, Non-convex Penalty.

1. Introduction

Recently, estimation of causal effect 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), and it should not have a direct effect on the outcome (C2) or unobserved confounders for the outcome (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 and Ebrahim, 2003; Kang et al., 2016; Windmeijer et al., 2019) and weak IVs (violation of C1) (Bound et al., 1995; Staiger and Stock, 1997) are prevalent. A strand of literature studies the "many weak IVs" problem (Stock et al., 2002; Chao and Swanson, 2005). With the

†Correspondence: Qingliang Fan, Department of Economics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong. Email: `michaelqfan@gmail.com`

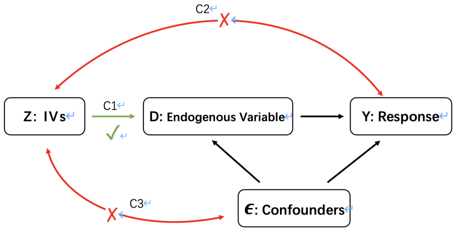

Figure 1. Relevance and Validity of IVs

increasing availability of large datasets, IV models are often high-dimensional (Belloni et al., 2012; Lin et al., 2015; Fan and Zhong, 2018), and have potentially weak IV (Andrews et al., 2018), and invalid IV (Guo et al., 2018; Windmeijer et al., 2021). Among those problems, we mainly focus on the invalid IV problem, while allowing for potential high-dimensional and weak IVs.

1.1. *Related Work*

Most prior works on IV's validity fall into two strands: robust estimation with all invalid IVs or estimation relying on valid IVs from a candidate IV set without prior knowledge of validity.

The first strand of literature allows all IVs to be invalid. For example, Kolesár et al. (2015) restrict the direct effects of IVs on treatment and response are random effects and independent. However, this assumption might be difficult to justify in practice. Lewbel (2012); Tchetgen et al. (2021); Guo and Bühlmann (2022) utilize the heterogeneity to achieve robustness with all invalid IVs, but their performances are not satisfactory once the heterogeneity (identification condition) is not entirely evident.

The second strand focuses on unknown invalid IVs, while imposing certain identification conditions on the number of valid IVs. Assuming more than half of IVs are valid, Kang et al. (2016) propose a Lasso type estimator (sisVIVE). Windmeijer et al. (2019) point out the inconsistent variable selection of sisVIVE under a relative IV strength condition and propose an adaptive Lasso estimator under the same conditions to bypass that problem. Guo et al. (2018); Windmeijer et al. (2021) further develop two selection-based two-step approaches, Two-Stage Hard Thresholding (TSHT) and Confidence Intervals IV (CIIV), respectively, under the plurality rule condition on relevant IVs set. However, the aforementioned approaches are not robust to many weak IVs due to the restriction of the majority/plurality rule in strong IVs instead of all IVs. Our method closely follows this strand of literature, requiring the plurality rule for valid IVs regardless of their strength, thus essentially relaxing the requirement in theory and for most practical cases.

The study of many (weak) IVs stems from empirical motivations but often assumes known validity. For example, Hansen et al. (2008); Chao and Swanson (2005); Hansen and Kozbur (2014); Newey and Windmeijer (2009) consider different estimators in many (weak) valid IVs but fix the number of known covariates. Kolesár et al. (2015); Kolesár (2018), on the other hand, allow the number of covariates to grow with sample size.

1.2. The 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 the identification and is operational in numerical optimization accommodating weak IVs. The proposed procedure has a selection stage and a post-selection estimation stage. The selection stage is a penalized IV-regression via minimax concave penalty (MCP, Zhang et al., 2010), a proper surrogate penalty aligned with the identification condition to achieve model selection consistency of valid IVs under weaker 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 estimate the treatment effect. An efficient computational algorithm for the optimal solution is provided. The computer codes for implementing the WIT estimator are available at `https://github.com/QoifoQ/WIT`.

The key contributions of this paper are summarized in the following.

(a) We provide a self-contained framework to investigate the fundamental problem in model identification for linear IV models with unknown validity. Specifically, we study the identification condition from the general data generating process (DGP) framework. It addresses an earlier caveat on the if and only if (iff) condition statement of the plurality rule (Guo et al., 2018, Theorem 1). Furthermore, we develop a theorem on the impossibility result of the existence of an iff condition on the model identification in the linear IV model framework.

(b) We show that the sparsest rule, equivalent to the plurality rule in 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, mistargeting the model identification.

(c) 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 IVs, where both the number of valid and invalid IVs are increasing with the sample size n. We also provide a simplified theoretical result in finite IVs case. Our method has a one-step valid IV selection instead of the previous two-step selections (Guo et al., 2018; Windmeijer et al., 2021). This allows us to utilize individually weak IVs, which are prevalent in empirical studies.

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 a 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 estimators using comprehensive numerical experiments. Section 5 applies our methods to an empirical international trade and growth study. Section 6 concludes. All the technical details and proofs are provided in the appendix.

2. Model and Identification Strategy

2.1. Potential Outcome Model with Some Invalid IVs

We consider the potential outcome model as in Small (2007); Kang et al. (2016). For $i = 1, 2, \ldots, n$, we have the random sample $(Y_i, D_i, \boldsymbol{Z}_{i.})$. Let $Y_i^{(D_i, \boldsymbol{Z}_{i.})}$ to be the potential outcome for the object i having exposure D_i and instrumental variables $\boldsymbol{Z}_{i.} \in \mathbb{R}^p$.

Given two different sets of treatment variables D_i^A, D_i^B, and corresponding instruments $\boldsymbol{Z}_{i.}^A$, $\boldsymbol{Z}_{i.}^B$, assume

$$Y_i^{(D_i^B, \boldsymbol{Z}_{i.}^B)} - Y_i^{(D_i^A, \boldsymbol{Z}_{i.}^A)} = \left(\boldsymbol{Z}_{i.}^B - \boldsymbol{Z}_{i.}^A\right)^\top \boldsymbol{\phi} + \left(D_i^B - D_i^A\right)\beta \text{ and } E\left(Y_i^{(0,0)} \mid \boldsymbol{Z}_{i.}\right) = \boldsymbol{Z}_{i.}^\top \boldsymbol{\theta}, \tag{1}$$

where $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ measure the direct effects of instruments on responses and the association between IVs and unmeasured confounders, respectively. Note that $\boldsymbol{Z}_{i.}$ could have nonlinear transformations of the original variables (such as polynomials and B-splines), so a high-dimensional model is plausible. A good instrument $\boldsymbol{Z}_j$ should not have a direct effect on the response and unmeasured confounders, i.e., $\phi_j = 0$ and $\theta_j = 0$, for $j = 1, 2, \ldots p$. Assuming the linear functional form between treatment effects $\boldsymbol{D}_i$ and instruments $\boldsymbol{Z}_{i.}$, the above potential outcome model (1) can be rewritten as follows,

$$\begin{aligned} Y_i &= D_i\beta + \boldsymbol{Z}_{i.}^\top \boldsymbol{\alpha} + \epsilon_i \\ D_i &= \boldsymbol{Z}_{i.}^\top \boldsymbol{\gamma} + \eta_i. \end{aligned} \tag{2}$$

where $\epsilon_i = Y_i^{(0,0)} - E\left(Y_i^{(0,0)} \mid \boldsymbol{Z}_{i.}\right)$, $\boldsymbol{\alpha} = \boldsymbol{\phi} + \boldsymbol{\theta}$. Let $\boldsymbol{\alpha}^*, \beta^*$ and $\boldsymbol{\gamma}^*$ represent the true coefficients in (2). Following Kang et al. (2016), we define the valid instruments as:

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

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 excluded IV, satisfying the order condition (Wooldridge, 2010). We consider many (weak) IVs cases in (2) 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 positive constant $\upsilon_{p_{\mathcal{V}^{c*}}}$ and $\upsilon_{p_{\mathcal{V}^*}}$ such that $\upsilon_{p_{\mathcal{V}^*}} + \upsilon_{p_{\mathcal{V}^{c*}}} < 1$.

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: $\boldsymbol{\gamma}^* = \left(E\left(\boldsymbol{Z}_{i.}\boldsymbol{Z}_{i.}^\top\right)\right)^{-1} E\left(\boldsymbol{Z}_{i.}\boldsymbol{D}_i\right)$, $\boldsymbol{\gamma}_j^* \neq 0$ for $j = 1, 2, \ldots, p$ and given n.

Assumption 4: Let $\boldsymbol{u}_i = (\epsilon_i, \eta_i)^\top$ and $\boldsymbol{u}_i \mid \boldsymbol{Z}$ follows i.i.d. multivariate normal distribution with mean zeros 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 5 (Strength of valid IVs): The concentration parameter μ_n grows at the same rate as n, i.e., $\mu_n := \boldsymbol{\gamma}_{\boldsymbol{Z}_{\mathcal{V}^*}}^{*\top} \boldsymbol{Z}_{\mathcal{V}^*}^\top M_{\boldsymbol{Z}_{\mathcal{V}^{c*}}} \boldsymbol{Z}_{\mathcal{V}^*} \boldsymbol{\gamma}_{\boldsymbol{Z}_{\mathcal{V}^*}}^* / \sigma_\eta^2 \to \mu_0 n$, for some $\mu_0 > 0$.

Assumption 1 is identical to the many instruments assumption in Kolesár et al. (2015); Kolesár (2018). It relaxes the conventional many IVs assumptions (Bekker, 1994; Chao and Swanson, 2005) that only allow the dimension of valid IVs $p_{\mathcal{V}^*}$ grows with n. Also, it has not been considered in the literature on selecting valid IVs (Kang et al., 2016; Guo et al., 2018; Windmeijer et al., 2021, 2019). Assumption 2 is standard for data preprocessing and scaling $\boldsymbol{Z}_j$. Assumptions 2 and 3 rule out redundant IVs. Assumption 4 follows Guo et al. (2018); Windmeijer et al. (2021) to impose the homoscedastic assumption and endogeneity issue of treatment D_i. The normality is not required for the selection stage, only for the estimation stage (asymptotics of embedded LIML estimator). Notice Assumption 5 requires a typically strong-identified moment condition in the literature (Bekker, 1994). This strength assumption can be further weakened along the lines of Chao and Swanson (2005); Hansen et al. (2008) to have many weak instruments asymptotics. Importantly, here we allow the individually weak (diminishing to zero in Staiger and Stock, 1997 style) signals for some but not all IVs regardless of their validity (as long as the concentration parameter satisfies Assumption 5), unlike that of Guo et al. (2018). Nevertheless, constant μ_0 can be a small number in a finite sample to accommodate empirically relevant many weak IVs.

2.2. Identifiability of Model (2)

The following moment conditions can be derived from Model (2):

$$E\Big(\boldsymbol{Z}^\top(\boldsymbol{D}-\boldsymbol{Z}\boldsymbol{\gamma}^*)\Big)=\boldsymbol{0},\quad E\Big(\boldsymbol{Z}^\top(\boldsymbol{Y}-\boldsymbol{D}\beta^*-\boldsymbol{Z}\boldsymbol{\alpha}^*)\Big)=\boldsymbol{0}\quad\Rightarrow\quad\boldsymbol{\Gamma}^*=\boldsymbol{\alpha}^*+\beta^*\boldsymbol{\gamma}^*,\tag{3}$$

where $\boldsymbol{\Gamma}^*=E(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}E(\boldsymbol{Z}^\top\boldsymbol{Y})$ 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) consider the identification of $(\boldsymbol{\alpha}^*,\beta^*)$ via the unique mapping of

$$\beta_j^*=\boldsymbol{\Gamma}_j^*/\boldsymbol{\gamma}_j^*=\beta^*+\alpha_j^*/\gamma_j^*.\tag{4}$$

Notice the moment conditions (3), consisting of p equations but $(\boldsymbol{\alpha}^*,\beta^*)\in\mathbb{R}^{p+1}$ need to be estimated, is under-identified without further restriction. Kang et al. (2016) propose the 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 IV. However, the majority rule could be restrictive in practice. Guo et al. (2018) further extend it to the plurality rule

$$\text{Plurality Rule:}\quad\Big|\mathcal{V}^*=\{j:\alpha_j^*/\gamma_j^*=0\}\Big|>\max_{c\neq0}\Big|\{j:\alpha_j^*/\gamma_j^*=c\}\Big|,\tag{5}$$

that is stated as an "if and only if" condition of identification of $(\boldsymbol{\alpha}^*,\beta^*)$. We re-examine the identifiability and show that the "only if" part is true only when all IVs are valid. In general, there is no iff condition in terms of model (3).

To illustrate the identification problem, we consider the model DGP. Given first stage information: $\{\boldsymbol{D},\boldsymbol{Z},\boldsymbol{\gamma}^*\}$, without loss of generality, we denote the DGP with some $\{\beta^*,\boldsymbol{\alpha}^*,\boldsymbol{\epsilon}\}$ in (2) as DGP $\mathcal{P}_0$ that generates $\boldsymbol{Y}$. Given this $\mathcal{P}_0$, for $j\in\mathcal{V}^{c*}$, we have $\boldsymbol{Z}_j\alpha_j^*=\frac{\alpha_j^*}{\gamma_j^*}(\boldsymbol{D}-\sum_{l\neq j}\boldsymbol{Z}_l\gamma_l^*-\boldsymbol{\eta})$. Denote $\mathcal{I}_c=\{j\in\mathcal{V}^{c*}:c=\alpha_j^*/\gamma_j^*\}$, where $c\neq0$ and c could have up to $p_{\mathcal{V}^{c*}}$ different values. 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 (2) to:

$$\boldsymbol{Y}=\boldsymbol{D}\tilde{\beta}^c+\boldsymbol{Z}\tilde{\boldsymbol{\alpha}}^c+\tilde{\boldsymbol{\epsilon}}^c,\tag{6}$$

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 (3) 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 (3) 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 1-4 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* $\mathcal{V}^* \cup \{\cup_{c\neq 0}\mathcal{I}_c\} = \{1, 2 \ldots, p\}$, $\mathcal{V}^* \cap \mathcal{I}_c = \varnothing$ *for any* $c \neq 0$ *and* $\mathcal{I}_c \cap \mathcal{I}_{\tilde{c}} = \varnothing$ *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

$$\mathcal{Q} = \{\mathcal{P} = \{\beta, \boldsymbol{\alpha}, \epsilon\} : \boldsymbol{\alpha} \text{ is sparse}, E(\boldsymbol{Z}^\top \boldsymbol{\epsilon}) = \mathbf{0}\} \tag{7}$$

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 (6) 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. In the following, we discuss why it is not a necessary condition for identification.

DEFINITION 2. *Let* $\mathcal{H}$ *be a collection of mappings* $h : \mathcal{Q} \to \mathcal{P} \in \mathcal{Q}$ *such that* h *maps a collection of DGPs* $\mathcal{Q}$ *to one specific DGP* $\mathcal{P} \in \mathcal{Q}$*. Moreover, two different mappings* h_i *and* h_j *are image equivalent, denoted as* $h_i \cong h_j$*, in the sense of sharing an identical image. Also, the size of* $\mathcal{H}$*,* $|\mathcal{H}|$ *is defined as the number of distinct images.*

Through the above definition, clearly, $|\mathcal{H}| \leq G+1$, and any mapping $h \in \mathcal{H}$ such that $h : \mathcal{Q} \to \mathcal{P}_0$ can be treated as a sufficient condition to identify the model (2) according to $\mathcal{P}_0$. We discuss the mappings contribute to $\mathcal{H}$ and $|\mathcal{H}|$. The following theorem shows, in general, no iff condition exists for identifying $(\boldsymbol{\alpha}^*, \beta^*)$ in model (2).

THEOREM 2. *Under the same conditions as Theorem 1,* $\exists i \in \{0, \ldots, G\}$*, let* $\mathcal{F} = \{f : \mathcal{P} \in \mathcal{Q} \to \mathbb{R}; f(\mathcal{P}_i) < f(\mathcal{P}_j),\ \forall j \neq i\}$ *and* $\mathcal{G} = \{g = \operatorname{argmin}_{\mathcal{P}\in\mathcal{Q}} f(\mathcal{P}); f \in \mathcal{F}\}$*, then we obtain:*
(a) $\mathcal{G} \subseteq \mathcal{H}$*.*
(b) There does not exist a necessary condition for identifying $(\boldsymbol{\alpha}^*, \beta^*)$ *unless* $\exists h \in \mathcal{H} : \mathcal{Q} \to \mathcal{P}_0$ *and* $|\mathcal{H}| = 1$*.*

REMARK 1. *In Theorem 2,* $f \in \mathcal{F}$ *maps* $\mathcal{P} \in \mathcal{Q}$ *on* $\mathbb{R}$ *is only for convenience; in fact, it could map to any ordinal object. Moreover,* $f \in \mathcal{F}$ *means there exists the unique minimum value of the mapping, for all* $\mathcal{P} \in \mathcal{Q}$*. The elements in* $\mathcal{G}$ *(*$g(\mathcal{Q})$*'s) typically correspond to some "most/least (fill in the blank)" criteria, and they can serve as the sufficient conditions for identification. For instance, the plurality rule is equivalent to the most sparse* $\boldsymbol{\alpha}$*:* $g(\mathcal{Q}) = \min \|\boldsymbol{\alpha}\|_0$*. One could consider other criteria such as the least standard deviation of the regression:* $g(\mathcal{Q}) = \min \operatorname{Var}(\boldsymbol{\epsilon})$*; the minimum* l_2 *norm of*

$\boldsymbol{\alpha}$: $g(\mathcal{Q}) = \min \|\boldsymbol{\alpha}\|_2$, etc. Notably, the majority rule (Kang et al., 2016; Windmeijer et al., 2019) does not belong to $\mathcal{G}$ in general since it can only hold for some specific $\mathcal{Q}$ instead.

COROLLARY 1. *Suppose the Assumptions 1-4 are satisfied. If all IVs are valid, then $|\mathcal{H}| = 1$. If there are invalid IVs, then $|\mathcal{H}| = G + 1 > 1$.*

Theorem 2 describes the properties of the constituents of the $\mathcal{H}$, while Corollary 1 states that none of the mappings in $\mathcal{H}$ can be treated as a necessary condition unless $|\mathcal{H}| = 1$ and $h \in \mathcal{H}$ maps to $\mathcal{P}_0$. The key implication of Theorem 2 (b) and Corollary 1 is that, unless all IVs are valid, the plurality rule (corresponds to $\min \|\boldsymbol{\alpha}\|_0$) is not the only criteria for identification; the information of $\text{Var}(\tilde{\boldsymbol{\epsilon}}^c)$ e.g., can also be used as one of many criteria for identification.

In light of Theorem 2, is there a guidance for researchers to choose a proper identification condition $h \in \mathcal{H}$? Generally, seeking a proper/optimal identification condition requires a clear-defined loss function to measure how good theidentification is and to perform an optimization in infinite functional space $\mathcal{H}$. That is infeasible to quantify. Even if we choose some common identification conditions, e.g. the ones in Remark 1, none of them can be uniformly optimal in all $\mathcal{Q}$. However, akin to the arguments in Guo et al. (2018); Windmeijer et al. (2021), it is often reasonable to impose the sparsest $\boldsymbol{\alpha}$ (plurality rule) in practice because it aligns with most of the research designs to have valid IVs forming the largest group (in ideal situations all IVs would be valid). However, in practice, valid IVs tend to be weak, which is exactly the motivation of our study. As we show that individually weak IVs would not affect the model identification as long as we have some strong and valid IVs (specifically, Assumption 5 is satisfied), the sparsest rule assumption is practical and operational. Besides, sparsest $\boldsymbol{\alpha}$ is only related to the number of valid IVs, while other criteria, e.g. $\min \|\boldsymbol{\alpha}\|_2$, $\min \|\boldsymbol{\epsilon}\|_2^2$, might involve more information that does not have a clear meaning in classical IV theory.

Basically, in the remaining content, our target is to find all possible DGPs in $\mathcal{Q}$ and identify the one based on the sparsest rule. Other criteria for a true DGP might need some attention, say the reasonable value of $\text{Var}(\boldsymbol{\epsilon})$.

2.3. The Sparsest (α) Rule

The sparsest rule is conceptually equivalent to the plurality rule on the whole IV set, given 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:

$$\left| \mathcal{V}^*_{\mathcal{S}^*} = \left\{ j \in \mathcal{S}^* : \alpha^*_j / \gamma^*_j = 0 \right\} \right| > \max_{c \neq 0} \left| \left\{ j \in \mathcal{S}^* : \alpha^*_j / \gamma^*_j = c \right\} \right|, \tag{8}$$

where $\mathcal{S}^*$ is the strong IVs estimated by $\hat{\mathcal{S}}$ via first-step hard thresholding, i.e., $\hat{\mathcal{S}} = \{j : \widehat{\gamma}_j > \sqrt{\hat{\text{Var}}\left(\widehat{\gamma}_j\right)} \cdot \sqrt{2.01 \log \max\{p, n\}}\}$. Thus, TSHT and CIIV explicitly leverage on $\hat{\mathcal{S}}$-based plurality rule to estimate $\mathcal{V}^*_{\hat{\mathcal{S}}}$ and β^*.

Unlike earlier literature on invalid IVs, our paper utilize the information on weak IVs also. For one, the weak IV can be used to estimate β^*. When we do not have

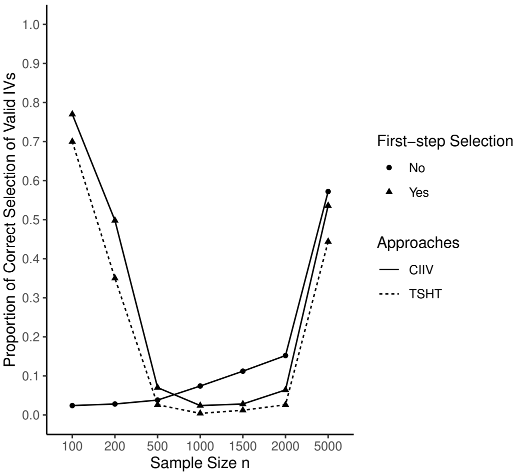

Figure 2. The proportion of correct selection of (subset) valid IVs based on 500 replications on each sample size.

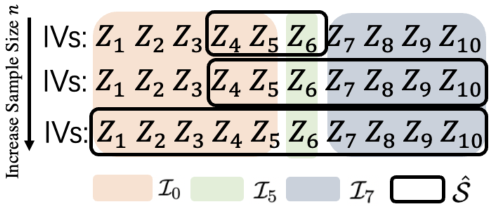
Increase Sample Size n
IVs: Z_1 Z_2 Z_3 Z_4 Z_5 Z_6 Z_7 Z_8 Z_9 Z_{10}
IVs: Z_1 Z_2 Z_3 Z_4 Z_5 Z_6 Z_7 Z_8 Z_9 Z_{10}
IVs: Z_1 Z_2 Z_3 Z_4 Z_5 Z_6 Z_7 Z_8 Z_9 Z_{10}
$\mathcal{I}_0$
$\mathcal{I}_5$
$\mathcal{I}_7$
$\hat{\mathcal{S}}$

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

strong IVs, the weak IV robust methods such as LIML are useful (Andrews et al., 2018); second, weak instruments could be used in the identification of valid IVs set as we show in Theorem 3. The plurality rule established on first stage selection may be unstable in estimating $\mathcal{V}^*$ as illustrated in the following toy example.

Example 1 (weak and invalid IVs). *Let* $\boldsymbol{\gamma}^* = (\mathbf{0.04}_3, \mathbf{0.5}_2, 0.2, \mathbf{0.1}_4)^\top$ *and* $\boldsymbol{\alpha}^* = (\mathbf{0}_5, 1, \mathbf{0.7}_4)^\top$. *Obviously it forms 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. Fig. 2 shows the selection of valid IVs by CIIV and TSHT breaks down in finite samples, e.g., for* $n \in [500, 2000]$. *This is because the solution of the plurality rule 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. Fig. 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 case. For the sake of using all IVs information for estimating β^* and identification of $\mathcal{V}^*$, we allow some individual IV strength to be local to zero (Chao and Swanson, 2005), say $\gamma_j^* \to 0$, or a small fixed constant that can not pass the first stage threshold (Guo et al., 2018) unless with very large sample size. However, we can see that in (4), 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.

To the end of using weak IVs information and improving finite sample performance, it motivates us to turn to the sparsest rule that is also operational in computation algorithms. Back to 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 $\tilde{\boldsymbol{\alpha}}^c$ (corresponds to a different DGP in $\mathcal{Q}$) and $j \in \{j : \alpha^*_j/\gamma^*_j = \tilde{c} \neq c\}$, we obtain

$$|\tilde{\alpha}^c_j| = |\alpha^*_j - c\gamma^*_j| = |\alpha^*_j/\gamma^*_j - c| \cdot |\gamma^*_j| = |\tilde{c} - c| \cdot |\gamma^*_j|. \tag{9}$$

The above $|\tilde{\alpha}^c_j|$ needs to be distinguished from 0 on the ground of non-overlapping structure stated in Theorem 1. To facilitate the discovery of all solutions in $\mathcal{Q}$, we assume: **Assumption** 6: $|\tilde{\alpha}^c_j| > \kappa(n)$ for $j \in \{j : \alpha^*_j/\gamma^*_j = \tilde{c} \neq c\}$ and $|\boldsymbol{\alpha}^*_{\mathcal{V}^{c*}}|_{\min} > \kappa(n)$, where $\kappa(n)$ is a generally vanishing rate specified by some estimator under consideration to separate zero and non-zero terms.

The above condition is known as the "beta-min" condition (van de Geer and Bühlmann, 2009; Loh and Wainwright, 2015). Notably, as shown in (9), $|\tilde{\alpha}^c_j| = |\tilde{c} - c| \cdot |\gamma^*_j| > \kappa(n)$ depends on the product of $|\tilde{c} - c|$ and $|\gamma^*_j|$. As discussed in Guo et al. (2018), $|\tilde{c} - c|$ can not be too small to separate different solutions in $\widehat{\boldsymbol{\alpha}}$, but the larger gap $|\tilde{c} - c|$ is helpful to mitigate the problem of small or local to zero $|\gamma^*_j|$ in favor of our model; see Appendix A1 for a more detailed discussion of Assumption 6.

Hence, the identification condition known as the sparsest rule is formally defined as **Assumption** 7: (The Sparsest Rule): $\boldsymbol{\alpha}^* = \operatorname{argmin}_{\mathcal{P}=\{\beta,\boldsymbol{\alpha},\boldsymbol{\epsilon}\}\in\mathcal{Q}} \|\boldsymbol{\alpha}\|_0$.

Example 2 (continued). *Following the procedure (6), we are able to reformulate two additional solutions of* (3) *given the DGP of Example 1,* $\boldsymbol{\alpha}^* = (\mathbf{0}_5, 1, \mathbf{0.7}_4)^\top$*:* $\tilde{\boldsymbol{\alpha}}^5 = (-\mathbf{0.2}_3, -\mathbf{2.5}_2, 0, \mathbf{0.2}_4)^\top$ *and* $\tilde{\boldsymbol{\alpha}}^7 = (-\mathbf{0.28}_3, -\mathbf{3.5}_2, -0.4, \mathbf{0}_4)^\top$*. Thus, the sparsest rule* $\operatorname{argmin}_{\boldsymbol{\alpha}\in\{\boldsymbol{\alpha}^*,\tilde{\boldsymbol{\alpha}}^5,\tilde{\boldsymbol{\alpha}}^7\}}\|\boldsymbol{\alpha}\|_0$ *picks* $\boldsymbol{\alpha}^*$ *up, and Assumption 6 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* (2)*, as long as the first stage has sufficient information (concentration parameter requirement in Assumption 5). 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 1(II) in Section 4.1.*

In the following subsection we revisit the penalized approaches by Kang et al. (2016) and Windmeijer et al. (2021) and discuss a class of existing estimators including Guo et al. (2018); Windmeijer et al. (2021) in terms of penalization, identification, and computation. We also discuss the general penalization approach which aligns model identification with its objective function.

2.4. Penalization Approaches with Embedded Surrogate Sparsest Rule

Penalized approach (Lasso) is firstly used in unknown IV validity context by Kang et al. (2016). We further extend the literature to a general formulation and discuss the properties of different classes of penalties.

For a general penalized estimator based on moment conditions (3),

$$(\widehat{\boldsymbol{\alpha}}^{\text{pen}}, \hat{\beta}^{\text{pen}}) = \underset{\boldsymbol{\alpha},\beta}{\operatorname{argmin}} \ \underbrace{\frac{1}{2n}\|P_{\boldsymbol{Z}}(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\alpha} - \boldsymbol{D}\beta)\|_2^2}_{(I)} + \underbrace{p_\lambda^{\text{pen}}(\boldsymbol{\alpha})}_{(II)}. \tag{10}$$

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 $p_\lambda^{\text{pen}\prime}(\cdot)$ is its derivative that satisfy: $\lim_{x \to 0^+} p_\lambda^{\text{pen}\prime}(x) = \lambda$, $p_\lambda^{\text{pen}}(0) = 0$, $p_\lambda^{\text{pen}}(x) = p_\lambda^{\text{pen}}(-x)$, $(x-y)(p_\lambda^{\text{pen}}(x) - p_\lambda^{\text{pen}}(y)) \geq 0$, and $p_\lambda^{\text{pen}\prime}(\cdot)$ is continuous on $(0, \infty)$.

In RHS of (10), (I) and (II) correspond to two requirements for the collection of valid DGPs in $\mathcal{Q}$ defined in (7). Between them, (I) is a scaled finite sample version of $E\Big((\boldsymbol{Z}^\top \boldsymbol{\epsilon})^\top (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} (\boldsymbol{Z}^\top \boldsymbol{\epsilon})\Big)$, which is a $(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}$ weighted quadratic term of condition $E(\boldsymbol{Z}^\top \boldsymbol{\epsilon}) = \mathbf{0}$, and (II) is imposed to ensure sparsity structure in $\widehat{\boldsymbol{\alpha}}$.

Further, regarding (I), one can reformulate (10) with respect to $\widehat{\boldsymbol{\alpha}}^{\text{pen}}$ as

$$\widehat{\boldsymbol{\alpha}}^{\text{pen}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \ \frac{1}{2n} \|\boldsymbol{Y} - \tilde{\boldsymbol{Z}} \boldsymbol{\alpha}\|_2^2 + p_\lambda^{\text{pen}}(\boldsymbol{\alpha}), \tag{11}$$

where $\tilde{\boldsymbol{Z}} = M_{\widehat{\boldsymbol{D}}} \boldsymbol{Z}$ and $\widehat{\boldsymbol{D}} = P_{\boldsymbol{Z}} \boldsymbol{D} = \boldsymbol{Z} \widehat{\boldsymbol{\gamma}}$, with similar arguments in Theorem 3 of Kang et al. (2016). Notably, 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 many IVs model (Assumption 1). The following lemma implies the issue of endogeneity of each $\tilde{\boldsymbol{Z}}_j$ is limited.

LEMMA 1. *Suppose Assumptions 1-5 hold and denote average gram matrix $\boldsymbol{Q_n} = \boldsymbol{Z}^\top \boldsymbol{Z}/n$, the endogeneity level of j-th transformed IV $\tilde{\boldsymbol{Z}}_j$ follows*

$$\tilde{\boldsymbol{Z}}_j^\top \boldsymbol{\epsilon}/n = \underbrace{\sigma_{\epsilon,\eta}^2 p/n}_{E(\widehat{\boldsymbol{D}}^\top \boldsymbol{\epsilon}/n)} \cdot \underbrace{\frac{\boldsymbol{Q}_{nj}^\top \boldsymbol{\gamma}^*}{\boldsymbol{\gamma}^{*\top} \boldsymbol{Q}_n \boldsymbol{\gamma}^* + \sigma_\eta^2 p/n}}_{\textit{dilution weight}} + O_p(n^{-1/2}). \tag{12}$$

REMARK 2. *Under Assumption 1, $p/n \to \mu_{p_{\mathcal{V}^*}} + \mu_{p_{\mathcal{V}^{c*}}} < 1$ does not vanish as $n \to \infty$. This 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 (10), Theorem 1 shows that model (2) can be identified by different strategies with non-overlapping results. On the ground of the sparsest rule assumption, the role of penalty on $\boldsymbol{\alpha}$, i.e. $p_\lambda^{\text{pen}}(\boldsymbol{\alpha})$, should not only impose 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 sparest rule.

To see the roles of a proper penalty function clearly, we rewrite (10) into equivalent constrained objective function with the optimal penalty $\|\boldsymbol{\alpha}\|_0$ regarding the sparest rule:

$$(\widehat{\boldsymbol{\alpha}}^{\text{opt}}, \hat{\beta}^{\text{opt}}) = \underset{\boldsymbol{\alpha}, \beta}{\operatorname{argmin}} \ \|\boldsymbol{\alpha}\|_0 \quad \text{s.t.} \ \|P_{\boldsymbol{Z}}(\boldsymbol{Y} - \boldsymbol{D}\beta - \boldsymbol{Z}\boldsymbol{\alpha})\|_2^2 < \delta, \tag{13}$$

where δ is the tolerance level we specify in Section 4. The constraint above narrows the feasible solutions into $\mathcal{Q}$ because it aligns with the Sargan test (Sargan, 1958) statistics

$\|P_{\boldsymbol{Z}}(\boldsymbol{Y}-\boldsymbol{D}\beta-\boldsymbol{Z}\boldsymbol{\alpha})\|_2^2/\|(\boldsymbol{Y}-\boldsymbol{D}\beta-\boldsymbol{Z}\boldsymbol{\alpha})/\sqrt{n}\|_2^2 = O_p(1)$ under 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 (13) leads to an equivalent optimization problem that

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

matches the identification condition. Therefore, the primary optimization object in (13) should also serve as an identification condition: the sparsest rule.

Due to computational NP-hardness for $\|\boldsymbol{\alpha}\|_0$ in (13), a surrogate penalty function is needed. Kang et al. (2016) proposed to replace the optimal l_0-norm with Lasso in (10), 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):

$$\left|\widehat{\boldsymbol{\gamma}}_{\mathcal{V}^{c*}}^\top \operatorname{sgn}\left(\boldsymbol{\alpha}^*_{\mathcal{V}^{c*}}\right)\right| > \|\widehat{\boldsymbol{\gamma}}_{\mathcal{V}^*}\|_1 \tag{14}$$

The SAIS is a rather common situation in practice, under which sisVIVE cannot achieve $\mathcal{P}_0$.

(b) Unclear dependency of regularization condition of $\tilde{\boldsymbol{Z}}$: Kang et al. (2016), Theorem 2, proposed a non-asymptotic error bound $|\hat{\beta}^{\text{sis}} - \beta^*|$ for sisVIVE. Under some regularity of restricted isometry property (RIP) constants of $\boldsymbol{Z}$ and $P_{\widehat{\boldsymbol{D}}}\boldsymbol{Z}$,

$$\|\hat{\beta}^{\text{sis}} - \beta^*\|_2 \le \frac{\left|\widehat{\boldsymbol{D}}^\top \boldsymbol{\epsilon}\right|}{\|\widehat{\boldsymbol{D}}\|_2^2} + \frac{1}{\|\widehat{\boldsymbol{D}}\|_2}\left(\frac{(4/3\sqrt{5})\lambda\sqrt{p_{\mathcal{V}^*}\delta^+_{2p_{\mathcal{V}^*}}\left(P_{\widehat{\boldsymbol{D}}}\boldsymbol{Z}\right)}}{2\delta^-_{2p_{\mathcal{V}^*}}(\boldsymbol{Z}) - \delta^+_{2p_{\mathcal{V}^*}}(\boldsymbol{Z}) - 2\delta^+_{2p_{\mathcal{V}^*}}\left(P_{\widehat{\boldsymbol{D}}}\boldsymbol{Z}\right)}\right),$$

where $\delta_k^{+/-}(\boldsymbol{H})$ refers to the upper and lower RIP constant of matrix $\boldsymbol{H}$. The dependence of RIP constant of $P_{\widehat{\boldsymbol{D}}}\boldsymbol{Z}$ is not clear due to the randomness nature of $\widehat{\boldsymbol{D}}$. Moreover, it is obscure what the impact of the potential failure of RIP in selecting valid IVs.

(c) The objective function deviates from the original sparsest rule: Multiple non-overlapping sparse solutions in (11) differentiate from standard Lasso problem, whereas the unique sparse solution satisfying (11) should share the same optimization target of l_1 and l_0 penalty. As shown in Theorem 2 and Remark 1, $g_1(\mathcal{P}) = \|\boldsymbol{\alpha}\|_0$ and $g_2(\mathcal{P}) = \|\boldsymbol{\alpha}\|_1$ correspond to incompatible identification conditions unless satisfying an additional strong requirement

$$\boldsymbol{\alpha}^* = \operatorname{argmin}_{\mathcal{P}=\{\beta,\boldsymbol{\alpha},\boldsymbol{\epsilon}\}\in\mathcal{Q}}\, g_j(\mathcal{P}), \forall j = 1, 2 \iff \|\boldsymbol{\alpha}^* - c\boldsymbol{\gamma}^*\|_1 > \|\boldsymbol{\alpha}^*\|_1, \forall c \neq 0, \tag{15}$$

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

Among the above problems, (a) and (b) are directly linked to the Lasso problem within the framework of invalid IVs, while (c) reveals the root of the problem beyond Lasso: a proper surrogate penalty in (11) should align with identification condition.

Windmeijer et al. (2019) proposed to use Adaptive Lasso (Zou, 2006) with a properly constructed initial estimator through median estimator to overcome the SAIS problem in (a). It also addresses (c) simultaneously. However, it requires the more stringent majority rule (see Remark 1), and all IVs are strong in fixed p case. Furthermore, it suffers from the same sensitivity issue on weak IVs as TSHT and CIIV.

The following proposition explains what should be a proper surrogate sparsest penalty.

Proposition 1. **(The proper Surrogate Sparsest penalty)** *Suppose Assumptions 1-7 are satisfied. If $p_\lambda^{pen}(\boldsymbol{\alpha})$ is the surrogate sparsest rule in the sense that it gives sparse solutions and*

$$\boldsymbol{\alpha}^* = \underset{\mathcal{P}=\{\beta,\boldsymbol{\alpha},\boldsymbol{\epsilon}\}\in\mathcal{Q}}{\operatorname{argmin}} \|\boldsymbol{\alpha}\|_0 = \underset{\mathcal{P}=\{\beta,\boldsymbol{\alpha},\boldsymbol{\epsilon}\}\in\mathcal{Q}}{\operatorname{argmin}} p_\lambda^{pen}(\boldsymbol{\alpha}), \tag{16}$$

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

Such a requirement of surrogate sparest penalty coincides with the folded-concave penalized method (Fan and Li, 2001; Zhang et al., 2010). Take MCP for example. In standard sparse linear regression, MCP requires $\lambda = \lambda(n) = O(\sqrt{\log p/n})$ and $p_\lambda^{\mathrm{MCP}\prime}(t) = 0$ when $t > C\lambda$, for some constant C, which satisfies Proposition 1. We specify that such property holds in invalid IVs cases in the next section and demonstrate it also circumvents the flaws of sisVIVE shown in (a) and (b).

Remark 3. *Proposition 1 shows the proper surrogate penalty must be concave and thus excludes Adaptive Lasso. Notice Adaptive Lasso penalty with adaptive weight constructed by a consistent estimate of $\boldsymbol{\alpha}^*$ could still satisfy (16) by adding more restrictive conditions, such as the majority rule, to identify $\boldsymbol{\alpha}^*$. The motivation of the surrogate sparsest penalty deploying concave penalties differentiates from debiasing purpose. Some other debias-oriented techniques, like the debiased Lasso (Javanmard and Montanari, 2018), cannot fit the identification condition and hence would deflect the objective function (13).*

In a nutshell, the proper surrogate sparsest penalty for (10) is to align the targeted identification condition and the global solution for $\mathcal{P}_0 \in \mathcal{Q}$.

3. WIT Estimator

3.1. Estimation Procedure

We adopt the penalized regression framework (10) and deploy a concave penalty in (11), the MCP in particular, which is a nearly unbiased estimator. Numerically, MCP penalty is shown to be the best choice in terms of the convexity of the penalized loss. Besides, it has consistent variable selection property without adding incoherence conditions on the design matrix (Loh and Wainwright, 2017; Feng and Zhang, 2019), which suits the two-stage type of estimation problem better than Lasso. Formally, the selection stage is

$$\widehat{\boldsymbol{\alpha}}^{\mathrm{MCP}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \frac{1}{2n}\|\boldsymbol{Y} - \widetilde{\boldsymbol{Z}}\boldsymbol{\alpha}\|_2^2 + p_\lambda^{\mathrm{MCP}}(\boldsymbol{\alpha}), \tag{17}$$

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 (11). $p_\lambda^{\text{MCP}}(\boldsymbol{\alpha}) = \sum_{j=1}^p p_\lambda^{\text{MCP}}(\alpha_j) = \sum_{j=1}^p \int_0^{|\alpha_j|} \Big(\lambda - t/\rho\Big)_+ dt$ is the MCP penalty and $\rho > 1$ is the tuning parameter, which also controls convexity level $1/\rho$, and its corresponding derivative is $p_\lambda^{\text{MCP}\prime}(t) = (\lambda - |t|/\rho)_+$. Unlike Lasso, MCP has no penalty once $|\alpha_j| > \lambda\rho$. It has nearly unbiased coefficients estimation, and achieves exact support recovery regardless of SAIS condition (see more intuitions in Appendix A2). Therefore, a consistent estimation of valid IVs set, i.e. $\hat{\mathcal{V}} = \{j : \widehat{\alpha}_j^{\text{MCP}} = 0\}$ and $\Pr(\hat{\mathcal{V}} = \mathcal{V}^*) \xrightarrow{p} 1$, is expected to hold under weaker conditions. Next, we show WIT combines the advantages of penalized TSLS and LIML estimators in different 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 et al., 2015; Kolesár, 2018). However, simultaneous estimations in $\hat{\kappa}_{\text{liml}}$ and $\hat{\mathcal{V}}^c$ in (19) are difficult to analyze. In the selection stage (10), we use the moment-based objective function. If we do not consider the penalty term (II), the moment-based part (I) of (10) coincides with TSLS. Furthermore, the bias in TSLS has a limited effect on consistent variable selections (see Theorem 3). 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 (Sawa, 1969; Chao and Swanson, 2005), we embed the LIML estimator to estimate β^* on the basis of estimated valid IVs set via (17). The performance of oracle-TSLS shown in simulations verifies this choice.

Consequently, we proposed the **W**eak and some **I**nvalid instruments robust **T**reatment effect (WIT) estimator in the estimation stage,

$$\Big(\hat{\beta}^{\text{WIT}}, \hat{\boldsymbol{\alpha}}^{\text{WIT}}_{\boldsymbol{Z}_{\hat{\mathcal{V}}^c}}\Big)^\top = \Big(\left[\boldsymbol{D}, \boldsymbol{Z}_{\hat{\mathcal{V}}^c}\right]^\top (\boldsymbol{I} - \hat{\kappa}_{\text{liml}} M_{\boldsymbol{Z}}) \left[\boldsymbol{D}, \boldsymbol{Z}_{\hat{\mathcal{V}}^c}\right]\Big)^{-1} \Big(\left[\boldsymbol{D}, \boldsymbol{Z}_{\hat{\mathcal{V}}^c}\right]^\top (\boldsymbol{I} - \hat{\kappa}_{\text{liml}} M_{\boldsymbol{Z}}) \boldsymbol{Y}\Big), \tag{18}$$

$$\hat{\kappa}_{\text{liml}} = \min_\beta \Big\{G(\beta) = \Big((\boldsymbol{Y} - \boldsymbol{D}\beta)^\top M_{\boldsymbol{Z}} (\boldsymbol{Y} - \boldsymbol{D}\beta)\Big)^{-1} \Big((\boldsymbol{Y} - \boldsymbol{D}\beta)^\top M_{\boldsymbol{Z}_{\hat{\mathcal{V}}^c}} (\boldsymbol{Y} - \boldsymbol{D}\beta)\Big)\Big\}, \tag{19}$$

Note, (18) 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 OLS and $\hat{\kappa} = 1$ reduced to the TSLS estimator. (19) has a closed-form solution: $\hat{\kappa}_{\text{liml}} = \lambda_{\min}\Big(\{[\boldsymbol{Y}, \boldsymbol{D}]^\top M_{\boldsymbol{Z}} [\boldsymbol{Y}, \boldsymbol{D}]\}^{-1} \{[\boldsymbol{Y}, \boldsymbol{D}]^\top M_{\boldsymbol{Z}_{\hat{\mathcal{V}}^c}} [\boldsymbol{Y}, \boldsymbol{D}]\}\Big)$, where $\lambda_{\min}(\cdot)$ means the smallest eigenvalue. Focusing on $\hat{\beta}^{\text{WIT}}$ as primary interest, we reformulate (18) and (19) 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 (20).

$$\begin{aligned} \hat{\beta}^{\text{WIT}} &= \Big(\boldsymbol{D}_\perp^\top \left(\boldsymbol{I} - \hat{\kappa}_{\text{liml}} M_{\boldsymbol{Z}_{\hat{\mathcal{V}}\perp}}\right) \boldsymbol{D}_\perp\Big)^{-1} \Big(\boldsymbol{D}_\perp^\top \left(\boldsymbol{I} - \hat{\kappa}_{\text{liml}} M_{\boldsymbol{Z}_{\hat{\mathcal{V}}\perp}}\right) \boldsymbol{Y}_\perp\Big), &(20) \\ \hat{\kappa}_{\text{liml}} &= \lambda_{\min}\Big(\{[\boldsymbol{Y}_\perp, \boldsymbol{D}_\perp]^\top M_{\boldsymbol{Z}_{\hat{\mathcal{V}}\perp}} [\boldsymbol{Y}_\perp, \boldsymbol{D}_\perp]\}^{-1} \{[\boldsymbol{Y}_\perp, \boldsymbol{D}_\perp]^\top [\boldsymbol{Y}_\perp, \boldsymbol{D}_\perp]\}\Big). &(21) \end{aligned}$$

3.2. Asymptotic Behavior of WIT Estimator

Throughout this subsection, we aim to recover the one specific element in $\mathcal{Q}$, denoted as $(\beta^*, \boldsymbol{\alpha}^*, \boldsymbol{\epsilon})$ temporally. Though a slight abuse of notation, we use $\widehat{\boldsymbol{\alpha}}$ to denote a local solution of (17) with MCP.

All local solutions of (17) we consider are characterized by the Karush–Kuhn–Tucker (KKT) or first-order condition, i.e.

$$\widetilde{\boldsymbol{Z}}^{\top}(\boldsymbol{Y}-\widetilde{\boldsymbol{Z}}\widehat{\boldsymbol{\alpha}})/n=\frac{\partial}{\partial \boldsymbol{t}}\sum_{j=1}^{p}p_{\lambda}'(t_j)\Big|_{\boldsymbol{t}=\widehat{\boldsymbol{\alpha}}}. \tag{22}$$

Explicitly, to the end of finding valid IVs via comparing with true signal $\boldsymbol{\alpha}^*$, we rewrite (22) as

$$\begin{cases}\left(\lambda-\frac{1}{\rho}\left|\widehat{\alpha}_j\right|\right)_+\leqslant \operatorname{sgn}\left(\widehat{\alpha}_j\right)\tilde{\boldsymbol{Z}}_j^{\top}(\boldsymbol{Y}-\tilde{\boldsymbol{Z}}\widehat{\boldsymbol{\alpha}})/n\leqslant\lambda, & j\in\hat{\mathcal{V}}^c\\ \left|\tilde{\boldsymbol{Z}}_j^{\top}(\boldsymbol{Y}-\tilde{\boldsymbol{Z}}\widehat{\boldsymbol{\alpha}})/n\right|\leqslant\lambda, & j\in\hat{\mathcal{V}}\end{cases} \tag{23}$$

where the inequalities in the first line stem from the convexity of MCP penalty and last 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 (PLSE) (Feng and Zhang, 2019; Loh and Wainwright, 2017; Zhang and Zhang, 2012).

The sisVIVE imposed RIP condition directly on $\tilde{\boldsymbol{Z}}$ to establish error bound. However, the restricted eigenvalue (RE) condition (Bickel et al., 2009) is the weakest condition (van de Geer and 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 design matrix $\tilde{\boldsymbol{Z}}$ of (17) in the following. Define restricted cone $\mathscr{C}(\mathcal{V}^*;\xi)=\{\boldsymbol{u}:\|\boldsymbol{u}_{\mathcal{V}^*}\|_1\le\xi\,\|\boldsymbol{u}_{\mathcal{V}^{c*}}\|_1\}$ for some $\xi>0$ that estimation error $\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}^*$ belongs to. The restricted eigenvalue $K_{\mathscr{C}}$ for $\tilde{\boldsymbol{Z}}$ is defined as $K_{\mathscr{C}}=K_{\mathscr{C}}(\mathcal{V}^*,\xi):=\inf\limits_{\boldsymbol{u}}\{\|\tilde{\boldsymbol{Z}}\boldsymbol{u}\|_2/(\|\boldsymbol{u}\|_2n^{1/2}):\boldsymbol{u}\in\mathscr{C}(\mathcal{V}^*;\xi)\}$ and RE condition refers to that $K_{\mathscr{C}}$ for $\tilde{\boldsymbol{Z}}$ should be bounded away from zero.

Lemma 2. *(RE condition of $\tilde{\boldsymbol{Z}}$) Under assumptions A1-4, for any given $\boldsymbol{\gamma}^*\neq\mathbf{0}$, there exists a constant $\xi\in(0,\|\widehat{\boldsymbol{\gamma}}_{\mathcal{V}^*}\|_1/\|\widehat{\boldsymbol{\gamma}}_{\mathcal{V}^{c*}}\|_1)$ and further, a restricted cone $\mathscr{C}(\mathcal{V}^*;\xi)$ defined by chosen ξ such that $K_{\mathscr{C}}^2>0$ holds strictly.*

Lemma 2 elaborates that RE condition on $\tilde{\boldsymbol{Z}}$ holds without any additional assumptions on $\tilde{\boldsymbol{Z}}$, unlike the extra RIP condition in sisVIVE. Moreover, this restricted cone is invariant of scaling, thus, indicating accommodation of many weak IVs. These two features suggest the theoretical advantages of penalized methods (11) over existing methods.

Next, we discuss the selection of valid IVs by comparing the local solution of (23) with the oracle (moment-based) counterpart. Define $\widehat{\boldsymbol{\alpha}}^{\text{or}}_{\mathcal{V}^*}=\mathbf{0}$ and

$$\widehat{\boldsymbol{\alpha}}^{\text{or}}_{\mathcal{V}^{c*}}=(\tilde{\boldsymbol{Z}}_{\mathcal{V}^{c*}}^{\top}\tilde{\boldsymbol{Z}}_{\mathcal{V}^{c*}})^{-1}\tilde{\boldsymbol{Z}}_{\mathcal{V}^{c*}}^{\top}\boldsymbol{Y}\quad\text{or}\quad\widehat{\boldsymbol{\alpha}}^{\text{or}}_{\mathcal{V}^{c*}}=(\boldsymbol{Z}_{\mathcal{V}^{c*}}^{\top}\boldsymbol{Z}_{\mathcal{V}^{c*}})^{-1}[\boldsymbol{Z}_{\mathcal{V}^{c*}}^{\top}(\boldsymbol{Y}-\widehat{\boldsymbol{D}}\hat{\beta}^{\text{TSLS}}_{\text{or}})], \tag{24}$$

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 above defined $\widehat{\boldsymbol{\alpha}}^{\text{or}}_{\mathcal{V}^{c*}}$'s 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 $\boldsymbol{R}^{\text{or}} = \widetilde{\boldsymbol{Z}}^\top(\boldsymbol{Y} - \widetilde{\boldsymbol{Z}}\widehat{\boldsymbol{\alpha}}^{\text{or}})/n$ are 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 $\tilde{\tilde{\boldsymbol{Q}}}_n = \boldsymbol{Z}_{\mathcal{V}^*}^\top(P_{\boldsymbol{Z}} - P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{Z}_{\mathcal{V}^*}/n$, we derive the following lemma.

Lemma 3. *Suppose Assumptions 1-5 hold and let*

$$\zeta \asymp \frac{p_{\mathcal{V}^*}}{n} \cdot \frac{\|\tilde{\tilde{\boldsymbol{Q}}}_n \gamma^*_{\mathcal{V}^*}\|_\infty}{\gamma^{*\top}_{\mathcal{V}^*}\tilde{\tilde{\boldsymbol{Q}}}_n \gamma^*_{\mathcal{V}^*}} + \sqrt{\frac{\log p_{\mathcal{V}^*}}{n}}. \tag{25}$$

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

$$\|\boldsymbol{R}^{or}\|_\infty \quad \leq \quad \left\|\frac{\boldsymbol{Z}^\top_{\mathcal{V}^*}\tilde{\tilde{\boldsymbol{\epsilon}}}}{n}\right\|_\infty + \left\|\frac{\frac{\boldsymbol{Z}^\top_{\mathcal{V}^*}\tilde{\tilde{\boldsymbol{D}}}}{n} \cdot \frac{\tilde{\tilde{\boldsymbol{D}}}^\top \boldsymbol{\epsilon}}{n}}{\frac{\tilde{\tilde{\boldsymbol{D}}}^\top\tilde{\tilde{\boldsymbol{D}}}}{n}}\right\|_\infty \leq \zeta \tag{26}$$

holds with probability approaching 1.

Based on Lemmas 2 and 3, we consider the set $\mathscr{B}(\lambda, \rho) = \{\widehat{\boldsymbol{\alpha}} \text{ in } (23) : \lambda \geq \zeta, \rho > K_{\mathscr{C}}^{-2}(\mathcal{V}^*, \xi) \vee 1\}$, in which ζ is defined in (25) and ξ is guaranteed by Lemma 2, as a collection of all local solutions $\widehat{\boldsymbol{\alpha}}$ computed in (23) through a broad class of MCP under certain penalty level λ and convexity $1/\rho$. Given the computed local solutions in practice are through a discrete path given some starting point (see Section 3.3), we further consider the computable solution set $\mathscr{B}_0(\lambda, \rho)$, introduced by Feng and Zhang (2019), i.e.,

$$\mathscr{B}_0(\lambda, \rho) = \{\widehat{\boldsymbol{\alpha}} : \widehat{\boldsymbol{\alpha}} \text{ and starting point } \widehat{\boldsymbol{\alpha}}^{(0)} \text{ are connected in } \mathscr{B}(\lambda, \rho)\}. \tag{27}$$

The connection between the above two sets is that $\exists\ \widehat{\boldsymbol{\alpha}}^{(l)} \in \mathscr{B}(\lambda, \rho)$ with penalty level $\lambda^{(l)}$ increasing with the index $l = 1, 2, \ldots$, such that $\widehat{\boldsymbol{\alpha}}^{(0)} - \boldsymbol{\alpha}^* \in \mathscr{C}(\mathcal{V}^*, \xi)$, $\widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\alpha}}^{(l)}$ 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 Lemma S3 of Appendix B9. Thus, $\mathscr{B}_0(\lambda, \rho)$ is a collection of approximations of $\boldsymbol{\alpha}$ in all DGPs.

Denote $\text{Bias}(\hat{\beta}^{\text{TSLS}}_{\text{or}}) = \frac{\boldsymbol{D}^\top(P_{\boldsymbol{Z}} - P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{\epsilon}}{\boldsymbol{D}^\top(P_{\boldsymbol{Z}} - P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{D}} = \frac{\boldsymbol{D}_\perp^\top P_{\boldsymbol{Z}_\perp}\boldsymbol{\epsilon}_\perp}{\boldsymbol{D}_\perp^\top P_{\boldsymbol{Z}_\perp}\boldsymbol{D}_\perp}$ and $\bar{\gamma}^*_{\mathcal{V}^{c*}} = \gamma^*_{\mathcal{V}^{c*}} + (\boldsymbol{Z}^\top_{\mathcal{V}^{c*}}\boldsymbol{Z}_{\mathcal{V}^{c*}})^{-1}\boldsymbol{Z}^\top_{\mathcal{V}^{c*}}\boldsymbol{Z}_{\mathcal{V}^*}\gamma^*_{\mathcal{V}^*}$. Then, we provide the asymptotic result of selection consistency of WIT.

Theorem 3. **(Selection Consistency)** *Specify $\kappa(n)$ in Assumption 6 as*

$$\kappa(n) \asymp \underbrace{\sqrt{\frac{\log p_{\mathcal{V}^*}}{n}}}_{T_1} + \underbrace{\frac{p_{\mathcal{V}^*}}{n} \cdot \frac{\|\tilde{\tilde{\boldsymbol{Q}}}_n \gamma^*_{\mathcal{V}^*}\|_\infty}{\gamma^{*\top}_{\mathcal{V}^*}\tilde{\tilde{\boldsymbol{Q}}}_n \gamma^*_{\mathcal{V}^*}}}_{T_2} + \underbrace{|\text{Bias}(\hat{\beta}^{TSLS}_{or})| \cdot \|\bar{\gamma}^*_{\mathcal{V}^{c*}}\|_\infty}_{T_3}, \tag{28}$$

where $T_1 \to 0$ as $n \to \infty$. Suppose Assumptions 1-7 hold and consider computable local solutions specified in (27), we have

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

In Theorem 3, 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 happen to many IV content and varnish fast in finite strong IV (see Colorrary 2). This result is new to the literature.

PROPOSITION 2. *Under the same assumptions of Theorem 3, if there does not dominant scaled γ_j^*, i.e. $\|\tilde{\tilde{\boldsymbol{Q}}}_n^{1/2}\boldsymbol{\gamma}_{\mathcal{V}^*}^*\|_\infty/\|\tilde{\tilde{\boldsymbol{Q}}}_n^{1/2}\boldsymbol{\gamma}_{\mathcal{V}^*}^*\|_1 = o\Big(\|\tilde{\tilde{\boldsymbol{Q}}}_n^{1/2}\boldsymbol{\gamma}_{\mathcal{V}^*}^*\|_1/(p_{\mathcal{V}^*}\|\tilde{\tilde{\boldsymbol{Q}}}_n^{1/2}\|_\infty)\Big)$, then $T_2 \to 0$.*

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

PROPOSITION 3. **(Approximation of** $\text{Bias}(\hat{\beta}_{or}^{TSLS})$**)** *Let $s = \max(\mu_n, p_{\mathcal{V}^*})$, under the Assumptions 1-5, we obtain*

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

REMARK 4. *The rate of concentration parameter μ_n will affect T_3 through $|\text{Bias}(\hat{\beta}_{or}^{TSLS})|$ under many IVs setting. Suppose Assumption 5 holds, that $\mu_n \xrightarrow{p} \mu_0 n$, the leading term in (30) is $\dfrac{\sigma_{\epsilon\eta}}{\sigma_\eta^2}\dfrac{\nu_{p_{\mathcal{V}^*}}}{\mu_0 + \nu_{p_{\mathcal{V}^*}}} \ll \dfrac{\sigma_{\epsilon\eta}}{\sigma_\eta^2}$ for moderate μ_0 since $0 < \nu_{p_{\mathcal{V}^*}} < 1$ while $\mu_0 > 0$. While under many weak IVs setting (Chao and Swanson, 2005; Hansen et al., 2008; Newey and Windmeijer, 2009), $\mu_n/n \xrightarrow{p} 0$ and the leading term in (30) becomes $\sigma_{\epsilon\eta}/\sigma_\eta^2$. Thus, many weak IVs setting imposes some difficulty (a higher T_3) for selecting valid IVs in Theorem 3. We further discuss the eligibility of many weak IVs in Remark 5.*

The following theorem describes the asymptotic behavior of 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

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

as the estimates of the covariance matrix of reduced-form error $\boldsymbol{\Omega} = \text{Cov}(\boldsymbol{\epsilon} + \beta^*\boldsymbol{\eta}, \boldsymbol{\eta})$ 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 $\hat{\boldsymbol{a}} = (\hat{\beta}^{\text{WIT}}, 1)$ and $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$.

THEOREM 4. *Under the same conditions as in Theorem 3, we obtain:*

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

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

(c) (Consistent variance estimator):

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

where $\hat{\boldsymbol{b}} = (1, -\hat{\beta}^{WIT})$ 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 typical many IVs case. Hansen et al. (2008) show that it is still valid for many weak IVs asymptotics.

Remark 5. **(Many Weak IVs Asymptotics)** *Regarding the many weak instruments asymptotic sequence considered in Chao and Swanson (2005); Hansen et al. (2008); Newey and Windmeijer (2009), we need to modify the number of invalid IVs $p_{\mathcal{V}^{c*}}$ as fixed. Thus, Assumption 5 can be relaxed to $\mu_n/n \to 0$ but $\mu_n/\sqrt{n} \to \infty$ and Theorem 4 holds according to Hansen et al. (2008), Theorems 1 and 3.*

In the following, we show WIT is a more powerful tool than many existing methods (Kang et al., 2016; Windmeijer et al., 2019) even under the strong and finite IVs settings. Typically, under strong and 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 $p_{\mathcal{V}^{c*}} + p_{\mathcal{V}^*} = p < n$.

Also, in the finite and strong IVs case, the threshold to distinguish the zero and nonzero parameters in Theorem 3 for selecting valid IVs goes to zero and thus helps relaxing such a requirement of unvanished margin. We now present the asymptotic theorem for WIT estimator in the finite IVs.

Corollary 2. **(Strong and Finite IVs Case)** *Suppose Assumptions 1′, 2-5 and 7 hold, for any fixed $\min_{\mathrm{j}\in\mathcal{V}^{c*}} \alpha_j^* > 0$, we have*

(a) (Selection consistency): $\widehat{\boldsymbol{\alpha}}^{MCP} = \underset{\widehat{\boldsymbol{\alpha}}\in\mathscr{B}_0(\lambda,\rho)}{\operatorname{argmin}} \|\widehat{\boldsymbol{\alpha}}\|_0, \Pr(\hat{\mathcal{V}} = \mathcal{V}^*, \widehat{\boldsymbol{\alpha}}^{MCP} = \widehat{\boldsymbol{\alpha}}^{or}) \xrightarrow{p} 1.$

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

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

(d) (Consistent variance estimator):

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

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

3.3. Computational Implementation of WIT Estimator

Through Proposition 1, we clearly know that MCP belongs to the surrogate sparsest penalty under Assumptions 1-7, and it ensures the global solution in (17) matches the sparsest rule. However, each element in $\mathcal{Q}$ is the local solution of (17) and we can only obtain one local solution from one initial value practically. A multiple starting points strategy is needed to achieve the global solution. Enumerating the whole $\boldsymbol{\alpha}^* \in \mathbb{R}^p$ is impossible. Therefore, it is important to develop an efficient algorithm for MCP penalty.

In light of practical use, we adopt the iterative local adaptive majorize-minimization (I-LAMM) algorithm (Fan et al., 2018), which satisfies (27) as shown in Feng and Zhang (2019) Section 2.1, with different initial values to achieve the local solution of $\widehat{\boldsymbol{\alpha}}$ in (17). See more technical details and derivation in Appendix A3.

Motivated by the individual IV estimator (Windmeijer et al., 2021) such that $\hat{\beta}_j = \hat{\Gamma}_j/\hat{\gamma}_j \xrightarrow{p} \beta^* + \alpha_j^*/\gamma_j^*$, where $\hat{\boldsymbol{\Gamma}} = (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{Y}$ and $\hat{\boldsymbol{\Gamma}} = (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{D}$, we can construct

$$\check{\boldsymbol{\alpha}}(\check{\beta}) = \hat{\boldsymbol{\Gamma}} - \check{\beta}\hat{\gamma} = \hat{\boldsymbol{\Gamma}} - \beta^*\hat{\boldsymbol{\gamma}} + (\beta^* - \check{\beta})\hat{\boldsymbol{\gamma}}, \tag{32}$$

with any initial value $\check{\beta}$. Thus $\|\check{\boldsymbol{\alpha}}(\check{\beta}) - \boldsymbol{\alpha}^*\|_1 \xrightarrow{p} |\beta^* - \check{\beta}| \cdot \|\hat{\boldsymbol{\gamma}}\|_1$. It is valid to replace $(\beta^*, \boldsymbol{\alpha}^*)$ by $(\tilde{\beta}^c, \tilde{\boldsymbol{\alpha}}^c)$, we know that $\|\check{\boldsymbol{\alpha}}(\check{\beta}) - \tilde{\boldsymbol{\alpha}}^c\|_1$ is asymptotically controlled by $|\tilde{\beta}^c - \check{\beta}| \cdot \|\hat{\boldsymbol{\gamma}}\|_1$. Thus, varying $\check{\beta} \in \mathbb{R}^1$ is equivalent to varying $\check{\boldsymbol{\alpha}}(\check{\beta}) \in \mathbb{R}^p$ of a close DGP $\mathcal{Q}$.

Here we provide a simple heuristic procedure to efficiently choose a proper $\check{\beta}$. Let $\hat{\beta}_{[j]}$ be the sorted $\hat{\beta}_j$. A fuzzy MCP regression is conducted to clustering $\hat{\beta}_j$:

$$\bar{\boldsymbol{\beta}} = \underset{\grave{\boldsymbol{\beta}}}{\operatorname{argmin}} \sum_{j=1}^{p} \|\grave{\beta}_j - \hat{\beta}_{[j]}\|_2^2 + \sum_{j=2}^{p} p_{\bar{\lambda}}^{\mathrm{MCP}}(|\grave{\beta}_j - \grave{\beta}_{j-1}|), \tag{33}$$

where $\bar{\boldsymbol{\beta}} \in \mathbb{R}^p$, $\bar{\lambda}$ is a prespecified penalty level and $\grave{\boldsymbol{\beta}}$ is initialized by $\grave{\beta}_j = \hat{\beta}_{[j]}$. Therefore, $\bar{\boldsymbol{\beta}}$ consist of less than p distinct values. Thus we can choose the initial $\widehat{\boldsymbol{\alpha}}$ in (27) such that

$$\widehat{\boldsymbol{\alpha}}^{(0)}(\check{\beta}) = \check{\boldsymbol{\alpha}}(\check{\beta}), \tag{34}$$

where $\check{\beta}$'s are chosen in $\bar{\boldsymbol{\beta}}$ with the priority given to the largest (remaining) cluster and $\widehat{\alpha}_j^{(0)}(\check{\beta}) = 0$ for j to be the unsorted index of $\hat{\beta}_j$ such that $\check{\beta} = \bar{\beta}_j$.

Intrigued by (27), which provides the theoretical guideline of tuning parameter, we look for a data-driven tuning procedure that has good performance in practice. From numerical studies, ρ is not sensitive so it is fixed as 2 for most applications. But λ is required to be tuned. Cross-validation is implemented in sisVIVE, but is known to be time-consuming and select too many invalid IVs. Windmeijer et al. (2019, 2021) propose to use Sargan test under low dimensions to choose tuning parameter consistently and obtain a good finite sample performance. But Sargan test is designed for fixed p and cannot handle many IVs. Motivated by this, we propose the modified Cragg-Donald (MCD, Kolesár, 2018) test-based tuning procedure, which extends the Sargan test to allow high-dimensional covariates and instruments.

Specifically, consider a local solution $\widehat{\boldsymbol{\alpha}}$ in (27), Denote $p_{\widehat{\mathcal{V}}} = |\{j : \widehat{\alpha}_j = 0\}|$ and $p_{\widehat{\mathcal{V}}^c} = |\{j : \widehat{\alpha}_j \neq 0\}|$. Let $m_{\min}$ be the minimum eigenvalue of $\boldsymbol{S}^{-1}\boldsymbol{T}$, where $\boldsymbol{S}$ and $\boldsymbol{T}$ are defined as a version of $\widehat{\boldsymbol{\alpha}}$ version in (31). Then, the MCD test is given by $nm_{\min}$. According to Kolesár (2018) Proposition 4, MCD test with asymptotic size ϱ_n would reject the null of $\boldsymbol{\alpha}_{\widehat{\mathcal{V}}} = \mathbf{0}$, when

$$nm_{\min} > \chi^2_{p_{\widehat{\mathcal{V}}^c}-1}\left\{\Phi\left(\sqrt{\frac{n - p_{\widehat{\mathcal{V}}^c}}{n - p_{\widehat{\mathcal{V}}^c} - p_{\widehat{\mathcal{V}}}}}\Phi^{-1}(\varrho_n)\right)\right\}, \tag{35}$$

where $\Phi(\cdot)$ denotes the CDF of standard normal distribution and $\chi^2_{p_{\widehat{\mathcal{V}}^c}-1}(\varrho_n)$ is $1 - \varrho_n$ quantile of $\chi^2_{p_{\widehat{\mathcal{V}}^c}-1}$ distribution. This property holds regardless of $p_{\mathcal{V}^{c*}}$ is fixed or grows with n. For the sake of the model selection consistency, the size of MCD test needs to be $o(1)$. Following Belloni et al. (2012); Windmeijer et al. (2019, 2021), we adopt a scaled rate $\varrho_n = 0.5/\log n$ that works well in simulation studies. Thus, with $\widehat{\boldsymbol{\alpha}}^{(0)}$ in (34) and

Algorithm 1 WIT estimator with MCD test tuning strategy

```
Input: $\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{D}, \boldsymbol{\lambda}^{\mathrm{seq}}, \varrho_n,$ and $J$
Calculate $\bar{\boldsymbol{\beta}}$ in (33), initialize $\widehat{\boldsymbol{\alpha}}^{\mathrm{MCP}} = \mathbf{1}$ and $\boldsymbol{I} = \mathbf{1}$
for $\widehat{\boldsymbol{\alpha}}^{(0)} = \mathbf{0}, \widehat{\boldsymbol{\alpha}}^{(0)}(\check{\beta}_j)$ (34) do    ▷ $\check{\beta}_j \in \bar{\boldsymbol{\beta}}$ in priority of largest cluster, for $j = 1, \ldots, J$
    for $\lambda$ in $\boldsymbol{\lambda}^{\mathrm{seq}}$ do
        Calculate $\widehat{\boldsymbol{\alpha}}$ through I-LAMM in Algorithm 2
        if $\widehat{\boldsymbol{\alpha}}$ is not rejected by MCD test (35) with size $\varrho_n$ then
            $\boldsymbol{I}[l] = 0$ for $l \in \{l : \widehat{\alpha}_l = 0\}$
            if $|\{j : \widehat{\alpha}_j = 0\}| > |\{j : \widehat{\alpha}_j^{\mathrm{MCP}} = 0\}|$ then
                $\widehat{\boldsymbol{\alpha}}^{\mathrm{MCP}} = \widehat{\boldsymbol{\alpha}}$
            end if
        end if
    end for
    if $\|\boldsymbol{I}\|_1 \leq |\{j : \{j : \widehat{\alpha}_j^{\mathrm{MCP}} = 0\}|$ then    ▷ Impossible for the existence of more sparse $\widehat{\boldsymbol{\alpha}}$
        Break
    end if
end for
Output: $\widehat{\boldsymbol{\alpha}}^{MCP}$
```

a sequence of $\boldsymbol{\lambda}^{\mathrm{seq}} = \boldsymbol{C}\sqrt{\log p/n}$ where $\boldsymbol{C} = 0.1 \times (1, \ldots, 20)^\top$, we propose to use MCD test to select the proper $\lambda \in \boldsymbol{\lambda}^{\mathrm{seq}}$ that is not rejected by (35) with $\varrho_n = 0.5/\log n$ and meanwhile with largest $p_{\widehat{\mathcal{V}}}$.

To sum up, we provide Algorithm 1 to demonstrate the detailed implementation steps.

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 $\boldsymbol{Z}_{\boldsymbol{i}\cdot} \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^{\boldsymbol{Z}})$, 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 $(\epsilon_i, \eta_i)^\top \overset{\text{iid}}{\sim} \mathcal{N}\left(\mathbf{0}, \left(\begin{smallmatrix} \sigma_\epsilon^2 & \sigma_{\epsilon,\eta} \\ \sigma_{\epsilon,\eta} & \sigma_\eta^2 \end{smallmatrix}\right)\right)$. We let $\sigma_\epsilon^2 = 1$ and $\mathrm{corr}(\epsilon_i, \eta_i) = 0.6$ 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, 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, MCD tuning strategy is implemented to determine λ, and we fix $\rho = 2$. In the I-LAMM

Table 1. Simulation results in low dimension

Case	Approaches	$n=200$				$n=500$			
		MAD	CP	FPR	FNR	MAD	CP	FPR	FNR
1(I)	TSLS	0.532	0	-	-	0.530	0	-	-
	LIML	0.952	0.004	-	-	0.978	0	-	-
	oracle-TSLS	0.038	0.938	-	-	0.023	0.958	-	-
	oracle-LIML	0.038	0.938	-	-	0.023	0.958	-	-
	TSHT	0.060	0.828	0.158	0.018	0.023	0.950	0	0.002
	CIIV	0.045	0.810	0.072	0.009	0.023	0.946	0.004	0.003
	sisVIVE	0.539	-	0.428	0.957	0.589	-	0.479	1
	Post-Alasso	0.532	0	1	0	0.530	0	0.996	0
	WIT	0.046	0.818	0.068	0.065	0.024	0.948	0.004	0.020
1(II)	TSLS	1.098	0	-	-	1.111	0	-	-
	LIML	7.437	0.292	-	-	7.798	0.030	-	-
	oracle-TSLS	0.072	0.938	-	-	0.046	0.948	-	-
	oracle-LIML	0.072	0.950	-	-	0.046	0.958	-	-
	TSHT	0.110	0.914	0.122	0.585	0.742	0.598	0.423	0.712
	CIIV	0.099	0.724	0.088	0.642	4.321	0.334	0.360	0.824
	sisVIVE	0.259	-	0.018	0.234	0.154	-	0	0.168
	Post-Alasso	1.831	0.016	0.429	0.258	3.533	0	0.560	0.387
	WIT	0.079	0.914	0.016	0.034	0.049	0.920	0.016	0.016

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 (STD), and coverage probability (CP) based on 95% confidence interval. 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 selection of valid and invalid IV as FP, and FN, respectively, and the number of correct selection of valid and invalid as TP, TN, respectively. Thus, FPR = FP/(FP +TN) and FNR = FN/(FN + TP).

4.1. Case 1: Low dimension

We first consider the low dimension scenario:

Case 1(I): $\boldsymbol{\gamma}^* = (\mathbf{0.5}_4, \mathbf{0.6}_6)^\top$ and $\boldsymbol{\alpha}^* = (\mathbf{0}_5, \mathbf{0.4}_3, \mathbf{0.8}_2)^\top$.

Case 1(II): $\boldsymbol{\gamma}^* = (\mathbf{0.04}_3, \mathbf{0.5}_2, 0.2, \mathbf{0.1}_4)^\top$ and $\boldsymbol{\alpha}^* = (\mathbf{0}_5, 1, \mathbf{0.7}_4)^\top$.

In above two cases, we maintain $\sigma_\eta^2 = 1$ and vary sample size $n = 200$ to 500. Case 1(I) refers to all strong IVs case, but SAIS (14) condition holds. Case 1(II) refers to Example 1 with mixed strong and weak IVs.

Table 1 presents the detailed results. In Case 1(I), high FPR or FNP indicates that sisVIVE and Post-Alasso mistarget $\boldsymbol{\alpha}^*$ because of lack of majority rule and SIAS holds. Their performances do not improve much with sample size n. Due to finite strong IVs, WIT performs similarly to TSHT, CIIV, and oracle-LIML. In low dimension settings, the oracle-TSLS is very close to oracle-LIML. Case 1(II) shows how weak IVs break strong

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

Case	Approaches	$n = 500$				$n = 1000$			
		MAD	CP	FPR	FNR	MAD	CP	FPR	FNR
1(III)	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
1(IV)	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

IVs based plurality rule. As shown in Example 1, TSHT and CIIV worsen in terms of all measures though the sample size increase from $n = 200$ to 500. Post-Alasso also fails due to the failure of majority rule. As the analysis in Example 2, such a mixture of weak IVs does not impede penalized methods. WIT estimator outperforms even when $n = 200$ and approaches oracle-LIML when n goes to 500. Interestingly, the comparably low FPR and MAD imply sisVIVE correctly target true $\boldsymbol{\alpha}^*$ since the additional requirement of matching objective function (15) happens to hold in this example. However, its FNR and MAD are worse than WIT due to the conservative cross-validation tuning strategy and the non-ignorable bias of Lasso, respectively.

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

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

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

We now vary sample size $n = 500$ to 1000 and fix $\sigma_\eta^2 = 1$ to strictly follow their design. Between them, Case 1(III) corresponds to the exact setting, while Case 1(IV) scales down the magnitude of $\boldsymbol{\gamma}^*$ to introduce small coefficients in the first stage.

Table 2 shows the results. In Case 1(III), CIIV outperforms TSHT because CIIV can utilize available information better (Windmeijer et al., 2021, Section 7). 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 1(IV), scaling down first stage coefficients causes some troubles for CIIV and TSHT, since 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 in $n = 500$ and improve in $n = 1000$ by mitigating such

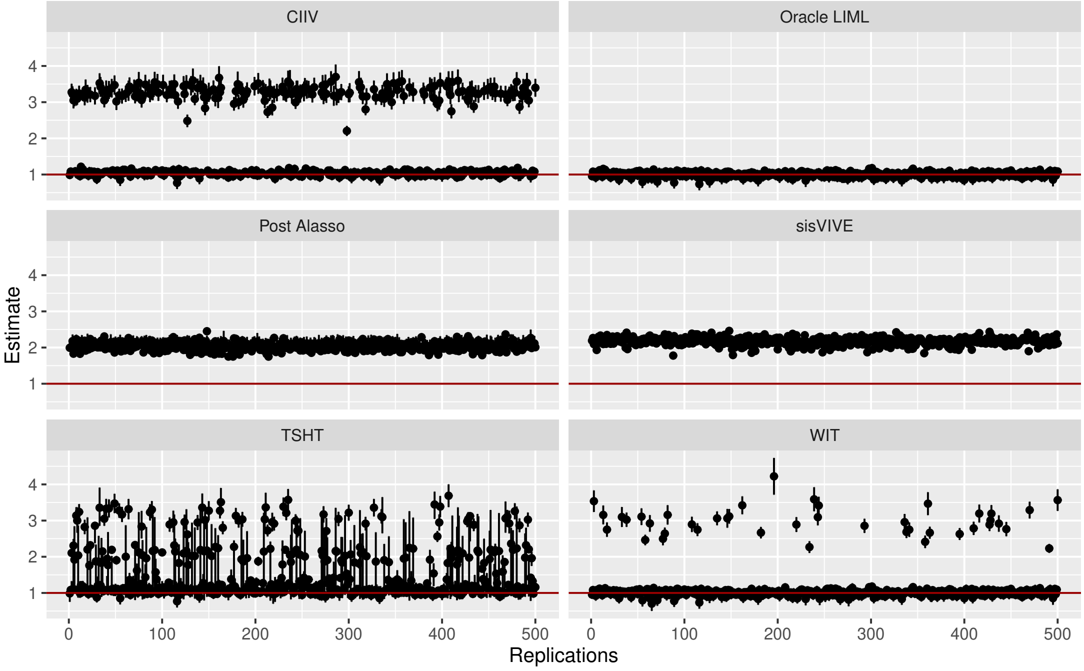

Figure 4. Scatter plot of estimations of β^* with confidence intervals of Case 1(IV) for $n = 1000$. Red line means true value $\beta^* = 1$.

coincidental issue of violating the plurality rule. Among penalized methods, sisVIVE and Post-Alasso mistarget and perform like TSLS because an additional requirement for sisVIVE (15) and majority rule fail simultaneously. Distinguished from them, the WIT estimator far outperforms with acceptable MAD in $n = 200$. The FPR and FNR are largely improved when the sample size increases. Fig. 4 presents all replication in Case 1(IV) when $n = 1000$. It shows that WIT is nearly oracle besides the mild number of incorrect selections. Nevertheless, CIIV and TSHT fail in selections more frequently.

4.2. Case 2: High dimension (many IVs)

To assess performance in many IVs, we consider the following examples:
Case 2(I) (increasing $p = 0.5n$): $\boldsymbol{\gamma}^* = (\mathbf{1.5}/\sqrt{\boldsymbol{n}})_p^\top$ and $\boldsymbol{\alpha}^* = (\mathbf{0}_{0.6p}, \mathbf{0.5}_{0.4p})^\top$.
Case 2(II) (increasing $p = 0.6n$): $\boldsymbol{\gamma}^* = (\mathbf{1.5}/\sqrt{\boldsymbol{n}})_p^\top$ and $\boldsymbol{\alpha}^* = (\mathbf{0}_{0.4p}, \mathbf{-0.5}_{0.2p}, \mathbf{1}_{0.3p}, \mathbf{-1}_{0.1p})^\top$.
The number of valid and invalid IVs is growing with the sample size. To verify the theoretical result, we maintain the ratio of concentration parameter to sample size n in a low constant level, i.e. $\mu_n/n = 0.5$, by adjusting σ_η^2. We vary sample size n from 500 to 1000, and let the first stage coefficient goes to 0. Due to computational burden in CIIV for many IVs case, we omit its results.

Table 3 provides the detailed estimation results. Case 2(I) satisfies the majority rule, and only two groups are present. All weak IVs narrow available choices in TSHT. When $n = 1000$, Low FPR but high FNR indicates that TSHT only selected the limited num-

Table 3. Simulation Results in High dimension (many IVs)

Case	Approaches	$n = 500$				$n = 1000$			
		MAD	CP	FPR	FNR	MAD	CP	FPR	FNR
2(I)	TSLS	2.170	0	-	-	3.015	0	-	-
	LIML	8.266	0	-	-	11.195	0	-	-
	oracle-TSLS	0.177	0	-	-	0.176	0	-	-
	oracle-LIML	0.044	0.950	-	-	0.027	0.956	-	-
	TSHT	0.972	0.108	0.221	0.698	0.376	0.112	0.002	0.649
	sisVIVE	0.552	-	0	0.214	0.551	-	0	0.123
	Post-Alasso	1.801	0	0.540	0	0.403	0	0.083	0
	WIT	0.047	0.936	0.003	0.013	0.029	0.946	0	0.001
2(II)	TSLS	1.300	0.096	-	-	1.733	0.010	-	-
	LIML	102.843	0.586	-	-	132.231	0.212	-	-
	oracle-TSLS	0.191	0.008	-	-	0.196	0	-	-
	oracle-LIML	0.054	0.958	-	-	0.035	0.936	-	-
	TSHT	1.592	0.254	0.293	0.937	2.577	0.174	0.296	0.955
	sisVIVE	0.280	-	0	0.201	0.284	-	0	0.118
	Post-Alasso	0.195	0	0.030	0	0.198	0	0.002	0
	WIT	0.085	0.760	0.005	0.010	0.044	0.904	0	0.010

ber of valid IVs. Both low FPR and FNR implies sisVIVE targets the correct solutions. Nevertheless, high MAD shows TSLS based method is biased in many IVs. The majority rule holds to ensure Post-Alasso is consistent. However, when $n = 500$, Post-Alasso is severely biased with comparable higher FPR. It may result from the sensitivity problem of initial estimator of Post-Alasso in weak IVs. Among these approaches, WIT estimator behaves highly similar to oracle-LIML and achieves the best performance in every measure, even when $n = 500$. On the contrary, oracle-TSLS has much larger bias than LIML in many IVs cases, and the coverage rate collapses.

Case 2(II) has more invalid IVs with the sparsest rule. TSHT breaks down in this case since strong IVs-based plurality rule is unlikely to hold. sisVIVE correctly identifies solutions by chance in this example. But its estimate suffers from the bias of Lasso and TSLS-based bias in many IVs. Without majority rule, Post-Alasso performs comparably to WIT estimator in terms of FPR and FNR. But it is only a coincidence that the initial estimator in Post-Alasso is consistent in this example since $E(\text{median}(\beta_j^*)) = \text{median}(\{\mathbf{1}_{0.4p}, \mathbf{-9.540}_{0.2p}, \mathbf{22.081}_{0.3p}, \mathbf{-20.081}_{0.1p}\}) = 1$. Compared with sisVIVE, Post-Alasso removes the bias of Lasso. Compared with Post-Alasso, WIT circumvents the majority rule and bias from TSLS. Thus, WIT outperforms in many IVs cases and approaches oracle-LIML when the sample size increases.

5. Application to Trade and Economic Growth

In this section, we revisit the classic empirical study in trade and growth (Frankel and Romer, 1999, FR99 henceforth). This remains a frontier and intensely debated issue in the field of international economics and also has strong guidance for policy-making. We investigate the causal effect of trade on income using more comprehensive and updated data, taking into account that trade is an endogenous variable (it correlates with unob-

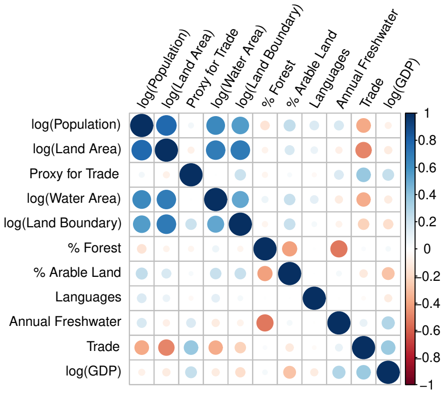

Figure 5. Correlation of all variables

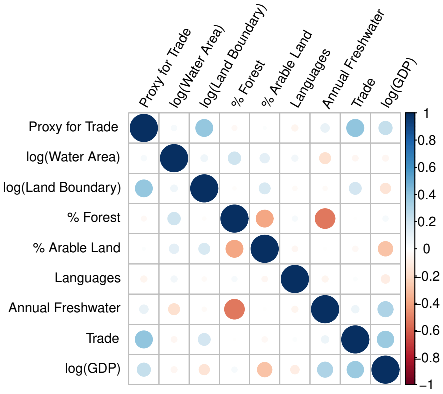

Figure 6. Correlation of transformed variables

served common factors driving both trade and growth), and some instruments might be invalid.

The structural equation considered in FR99 is,

$$\log(Y_i) = \alpha + \beta T_i + \psi S_i + \epsilon_i,$$

where for each country i, Y_i is the GDP per worker, T_i is the share of international trade to GDP, S_i is the size of the country, such as area, population, and ϵ_i is the error term. FR99 proposed to construct an IV (called a proxy for trade) based on the celebrated gravity theory of trade (Anderson, 1979). The logic of IV validity in aggregate variables is that the geographical variables, such as common border and distance between countries, indirectly affect growth through the channel of convenience for trade.

Following the same logic, Fan and Zhong (2018) extended the IV set to include more geographic and meteorological variables. The reduced form equation is

$$T_i = \boldsymbol{\gamma}^\top \boldsymbol{Z_i} + \nu_i, \tag{36}$$

where $\boldsymbol{Z_i}$ is a vector of instruments that we elaborate in Section 5.1. In this study, we expand the candidate IV set even further. On the one hand, more information contained in newly introduced IVs could increase the accuracy of estimating the causal effect of trade. On the other hand, some of the newly introduced IVs might be invalid. Also, with a large IV set with potentially invalid IVs, it is desirable to have a robust estimate of the treatment effect when the signal in the first stage is weak. This problem will be addressed by the proposed WIT estimator as discussed in previous sections.

5.1. Data Description and Empirical Model

We collect cross-sectional data from 158 countries in the year 2017. Table S1 (in Appendix A4) presents the summary statistics of the main variables.

We first standardize all the variables, then we formulate the structural equation as:

$$\log(Y_i) = T_i\beta + \boldsymbol{Z}_{i.}^\top \boldsymbol{\alpha} + \boldsymbol{S}_{i.}^\top \boldsymbol{\psi} + \epsilon_i \quad \text{for } i = 1, 2, \ldots, 158, \tag{37}$$

Table 4. Empirical Results of Various Estimators

	$\hat{\beta}\left(\widehat{\text{Var}}^{1/2}(\hat{\beta})\right)$	95% CI	Valid IVs $\widehat{\mathcal{V}}$	Relevant IVs $\hat{\mathcal{S}}$	Sargan Test
OLS	0.413(0.084)	(0.246, 0.581)	-	-	-
FR99	0.673(0.220)	(0.228, 1.117)	-	-	0.999
LIML	2.969(1.503)	(0.023, 5.916)	-	-	0.001
TSHT	0.861(0.245)	(0.380, 1.342)	{1}	{1}	0.999
CIIV*	2.635(1.974)	(-1.233, 6.504)	{2,4,5,6,7}	-	0.385
sisVIVE	0.819(-)	-	{1,2,4}	-	0.418
Post-Alasso	0.964(0.251)	(0.471, 1.457)	{1,2,4,5,6}	-	0.086
WIT	0.974(0.323)	(0.340, 1.609)	{1,2,4,6}	-	0.275

Note: CIIV* stands for CIIV method without first stage IVs selection because it reports that "Less than two IVs are individually relevant, treat all IVs as strong". Sargan test p-value is shown in the last column. The selection of relevant IVs $\hat{\mathcal{S}}$ is only implemented in TSHT and CIIV.

where $\boldsymbol{S}_{i.}$ consists of log(Population) and log(Land Area) for country i serve as control variables following FR99 and $\boldsymbol{Z}_{i.}$ are the potential IVs including all from Fan and Zhong (2018) and more geo-economic variables from the World Bank database. $\boldsymbol{\alpha}$ indicates whether the IV is invalid. In the first stage, we consider the linear reduced form equation (36). Fig. 5 shows the plot of correlations between the variables.

We partial out the effect from control variables $\boldsymbol{S}$ in (37), which does not affect the estimation of $\boldsymbol{\alpha}$ and β. Denote $\boldsymbol{M_S}$ as the projection matrix on orthogonal space with respect to column space of $\boldsymbol{S}$, we transform $\{\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{Z}\}$ to $\{\boldsymbol{M_S Y}, \boldsymbol{M_S T}, \boldsymbol{M_S Z}\}$ and denote the transformed observations as $\{\ddot{\boldsymbol{Y}}, \ddot{\boldsymbol{T}}, \ddot{\boldsymbol{Z}}\}$. So as the error terms $\ddot{\boldsymbol{\epsilon}}$ and $\ddot{\boldsymbol{\nu}}$. Hence, we work on the transformed model as follows:

$$\ddot{Y}_i = \ddot{T}_i\beta + \ddot{\boldsymbol{Z}}_{i.}^\top \boldsymbol{\alpha} + \ddot{\epsilon}_i, \quad \ddot{T}_i = \ddot{\boldsymbol{Z}}_{i.}^\top \boldsymbol{\phi} + \ddot{\nu}_i.$$

The correlation matrix of transformed variables is plotted in Fig 6.

5.2. Empirical Result

We explore the causal effect of trade using the proposed WIT estimator and also provide the estimation results from other popular estimators, including sisVIVE, TSHT, CIIV, OLS, LIML, and FR99 (TSLS using one IV $\widehat{T}$) for comparison.

Table 4 provides the detailed results of estimation and inference. The p-value of the Hausman test for endogeneity is 1.81e-3 using the proxy for trade as IV. The OLS estimate is likely biased due to the endogeneity of trade. The FR99 result using $\boldsymbol{Z}_1 = \widehat{\boldsymbol{T}}$ as instruments gives a smaller treatment effect estimate compared to WIT. LIML using all potential IVs (without distinguishing the invalid ones) likely overestimates the treatment effect. The 0.001 p-value of Sargan test strongly rejects the null of all potential IVs are valid.

For the penalized IV regression approaches: sisVIVE, Post-Alasso and WIT estimator jointly select 3 valid IVs: $\boldsymbol{Z}_1$, $\boldsymbol{Z}_2$ and $\boldsymbol{Z}_4$. $\boldsymbol{Z}_6$: Languages is selected by WIT as valid but not by sisVIVE. Regarding p-value of Sargan test for WIT and sisVIVE, 0.275, and 0.418, it supports that $\boldsymbol{Z}_6$ is valid. Compared with Post-Alasso, $\boldsymbol{Z}_5$: % Arable Land is chosen by Post-Alasso while not by WIT estimator. In view of the marginal correlation in Fig. 6, $\boldsymbol{Z}_5$ is nearly uncorrelated to trade but significantly correlated with log(GDP).

Concerning the Sargen test, p-value of $0.086 < 0.1$ in Post-Alasso indicates $\boldsymbol{Z}_5$ is not very credible to be valid. From an economic perspective, more arable land generates higher crop yields and maintains a higher agriculture sector labor force, which directly affects GDP. Thus it has a channel to affect GDP directly. Put together, we conclude $\boldsymbol{Z}_5$ should be a invalid IV.

Regarding individual IV estimator-based approaches, their selection results might not be accurate compared with WIT estimator. These methods impose a stronger condition of relevant IVs than WIT, which prohibits their selection in weaker IVs, narrows the available choices, and cannot handle weak IVs problems. Specifically, the first stage threshold of TSHT only selects proxy for trade, a consensus valid instrument. While CIIV*, which treats all IVs as strong, apparently chooses the incorrect set since it excludes proxy of trade. It shows the sensitivity problem of individual IV estimator-based approaches when weak IVs are present.

Four out of seven IVs are estimated as valid by WIT implies the majority rule holds. That supports the selection result of Post-Alasso is closed to WIT. Even though $\hat{\beta} = 0.964$ in Post-Alasso is almost identical to WIT estimator, Post-Alasso utilizes TSLS as second-stage estimator instead of LIML in WIT estimator. Nevertheless, a closer check finds that the first stage F-value is $3.682 < 5$, indicating the weak IVs problem in Post-Alasso estimated valid IVs. Using LIML as the second stage in their estimated valid IVs provides a much different estimate: 1.441(0.421) that is far away from the result of WIT estimator.

To sum up, the proposed WIT estimator overcomes potential weak IV problem and provides the most accurate estimation with reasonable explanation among the existing methods. This empirical problem is further evidence of the usefulness of our approach.

6. Conclusion

In this study, 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 model and showed the impossibility of the existence of if and only if identification condition. Sticking to the sparsest rule, we proposed the surrogate sparsest penalty that fits the identification condition. Further, we proposed a novel WIT estimator that addresses the issues in sisVIVE, Post-Alasso, and outperforms the plurality rule based TSHT and CIIV. Simulations and real data analysis support the theoretical findings and advantages over existing approaches.

Acknowledgements

References

Anderson, J. E. (1979) A theoretical foundation for the gravity equation. *Am. Econ. Rev.*, **69**, 106–116.

Andrews, I., Stock, J. and Sun, L. (2018) Weak instruments in IV regression: Theory and practice. *Annu. Rev. Econ.*

Bekker, P. A. (1994) Alternative approximations to the distributions of instrumental variable estimators. *Econometrica*, 657–681.

Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C. (2012) Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, **80**, 2369–2429.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009) Simultaneous analysis of lasso and dantzig selector. *The Ann. Statist.*, **37**, 1705–1732.

Bound, J., Jaeger, D. A. and Baker, R. M. (1995) Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *J. Am. Statist. Ass.*, **90**, 443–450.

Bun, M. J. and Windmeijer, F. (2011) A comparison of bias approximations for the two-stage least squares (2sls) estimator. *Econ. Lett.*, **113**, 76–79.

Chao, J. C. and Swanson, N. R. (2005) Consistent estimation with a large number of weak instruments. *Econometrica*, **73**, 1673–1692.

Daubechies, I., Defrise, M. and De Mol, C. (2004) An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, **57**, 1413–1457.

Davey Smith, G. and Ebrahim, S. (2003) 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International journal of epidemiology*, **32**, 1–22.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass.*, **96**, 1348–1360.

Fan, J., Liu, H., Sun, Q. and Zhang, T. (2018) I-lamm for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *Ann. Statist.*, **46**, 814.

Fan, Q. and Zhong, W. (2018) Nonparametric additive instrumental variable estimator: A group shrinkage estimation perspective. *J. Bus. Econ. Statist.*, **36**, 388–399.

Feng, L. and Zhang, C.-H. (2019) Sorted concave penalized regression. *The Ann. Statist.*, **47**, 3069–3098.

Frankel, J. A. and Romer, D. H. (1999) Does trade cause growth? *Am. Econ. Rev.*, **89**, 379–399.

van de Geer, S. A. and Bühlmann, P. (2009) On the conditions used to prove oracle results for the lasso. *Electron. J. Statist.*, **3**, 1360–1392.

Guo, Z. and Bühlmann, P. (2022) Two stage curvature identification with machine learning: Causal inference with possibly invalid instrumental variables. *arXiv preprint arxiv.2203.12808*.

Guo, Z., Kang, H., Tony Cai, T. and Small, D. S. (2018) Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *J. R. Statist. Soc. B*, **80**, 793–815.

Han, C. (2008) Detecting invalid instruments using l_1-GMM. *Economics Letters*, **3**, 285–287.

Hansen, C., Hausman, J. and Newey, W. (2008) Estimation with many instrumental variables. *J. Bus. Econ. Statist.*, **26**, 398–422.

Hansen, C. and Kozbur, D. (2014) Instrumental variables estimation with many weak instruments using regularized JIVE. *J. Econometrics*, **182**, 290–308.

Javanmard, A. and Montanari, A. (2018) Debiasing the lasso: Optimal sample size for gaussian designs. *The Ann. Statist.*, **46**, 2593–2622.

Kang, H., Zhang, A., Cai, T. T. and Small, D. S. (2016) Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *J. Am. Statist. Ass.*, **111**, 132–144.

Kolesár, M. (2018) Minimum distance approach to inference with many instruments. *J. Econometrics*, **204**, 86–100.

Kolesár, M., Chetty, R., Friedman, J., Glaeser, E. and Imbens, G. W. (2015) Identification and inference with many invalid instruments. *J. Bus. Econ. Statist.*, **33**, 474–484.

Lewbel, A. (2012) Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models. *J. Bus. Econ. Statist.*, **30**, 67–80.

Lin, W., Feng, R. and Li, H. (2015) Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. *J. Am. Statist. Ass.*, **110**, 270–288.

Loh, P.-L. and Wainwright, M. J. (2015) Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *J. Mach. Learn. Res.*, **16**, 559–616.

— (2017) Support recovery without incoherence: A case for nonconvex regularization. *The Ann. Statist.*, **45**, 2455–2482.

Nagar, A. L. (1959) The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica*, 575–595.

Newey, W. K. and Windmeijer, F. (2009) Generalized method of moments with many weak moment conditions. *Econometrica*, **77**, 687–719.

Sargan, J. D. (1958) The estimation of economic relationships using instrumental variables. *Econometrica*, 393–415.

Sawa, T. (1969) The exact sampling distribution of ordinary least squares and two-stage least squares estimators. *J. Am. Statist. Ass.*, **64**, 923–937.

Small, D. S. (2007) Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *J. Am. Statist. Ass.*, **102**, 1049–1058.

Staiger, D. and Stock, J. H. (1997) Instrumental variables regression with weak instruments. *Econometrica*, 557–586.

Stock, J., Yogo, M. and Wright, J. (2002) A survey of weak instruments and weak identification in generalized method of moments. *J. Bus. Econ. Statist.*, **20**, 518–529.

Tchetgen, E. T., Sun, B. and Walter, S. (2021) The GENIUS approach to robust mendelian randomization inference. *Statistical Science*, **36**, 443–464.

Windmeijer, F., Farbmacher, H., Davies, N. and Davey Smith, G. (2019) On the use of the lasso for instrumental variables estimation with some invalid instruments. *J. Am. Statist. Ass.*, **114**, 1339–1350.

Windmeijer, F., Liang, X., Hartwig, F. P. and Bowden, J. (2021) The confidence interval method for selecting valid instrumental variables. *J. R. Statist. Soc. B*, **83**, 752–776.

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

Zhang, C.-H. and Zhang, T. (2012) A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, **27**, 576–593.

Zhang, C.-H. et al. (2010) Nearly unbiased variable selection under minimax concave penalty. *The Ann. Statist.*, **38**, 894–942.

Zhao, P. and Yu, B. (2006) On model selection consistency of lasso. *J. Mach. Learn. Res.*, **7**, 2541–2563.

Zou, H. (2006) The adaptive lasso and its oracle properties. *J. Am. Statist. Ass.*, **101**, 1418–1429.

Zou, H. and Li, R. (2008) One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.*, **36**, 1509.

Web-based supporting materials for "On the instrumental variable estimation with potentially many (weak) and some invalid instruments"

Yiqi Lin[a,b], Frank Windmeijer[b], Xinyuan Song[a], Qingliang Fan[c]

[a]Department of Statistics, The Chinese University of Hong Kong, Hong Kong.
[b]Department of Statistics, University of Oxford, Oxford, U.K.
[c]Department of Economics, The Chinese University of Hong Kong, Hong Kong.

This supplementary material mainly includes the following parts: Section Appendix A provides additional details of the main paper. Section Appendix B contains all technical proofs. Throughout the online supplementary material, we allow constant C to be a generic positive constant that may differ in different cases.

Appendix A. Additional Demonstrations

Appendix A1. Discussion of Assumption 6

Recall that Assumption 6: $|\tilde{\alpha}_j^c| = |\alpha_j^* - c\gamma_j^*| = |\tilde{c} - c| \cdot |\gamma_j^*| > \kappa(n)$ and $|\boldsymbol{\alpha}^*_{\mathcal{V}^{c*}}|_{\min} > \kappa(n)$, is a "beta-min" condition for penalized method while $|\tilde{c} - c| \neq 0$ required for individual IV estimator based approaches: TSHT and CIIV. We consider finite p case where above two mentioned methods primary works on and, thus, $\kappa(n)$ is specified as $\sqrt{\log p_{\mathcal{V}^*}/n} \asymp n^{-1/2}$.

For individual IV estimator based approaches, they rely on $\hat{\beta}_j = \hat{\Gamma}_j/\hat{\gamma}_j \xrightarrow{p} \beta^* + \alpha_j^*/\gamma_j^*$, where $\hat{\boldsymbol{\Gamma}} = \left(\boldsymbol{Z}^\top \boldsymbol{Z}\right)^{-1} \boldsymbol{Z}^\top \boldsymbol{Y}$ and $\hat{\boldsymbol{\Gamma}} = \left(\boldsymbol{Z}^\top \boldsymbol{Z}\right)^{-1} \boldsymbol{Z}^\top \boldsymbol{D}$. To grouping different value of α_j^*/γ_j^*, the assumption $|\tilde{c} - c| \neq 0$ is naturally to be required. However, the penalized methods direct target to $\boldsymbol{\alpha} \in \mathcal{Q}$ though (11) instead of α_j^*/γ_j^*. Hence it leads to a different requirement: $|\tilde{c} - c| \cdot |\gamma_j^*| \neq 0$.

First-stage thresholding to construct relevant IVs set $\hat{\mathcal{S}}$ in Guo et al. (2018) rule out the weak IVs because the small value in $\hat{\gamma}_j$ appearing in denominator of $\hat{\beta}_j$ can cause trouble to the grouping $\hat{\beta}_j$ subsequently. However, the weak IVs will also cause a large $\tilde{c} = \alpha_j^*/\gamma_j^*$ for a fixed α_j^*. Then let $c = \alpha_i^*/\gamma_i^*$, $\tilde{c} = \alpha_j^*/\gamma_j^*$ and $\gamma_i = C_1/\sqrt{n}$, $\gamma_j = C_2/\sqrt{n}$. Therefore, even though $|\tilde{c} - c| = \sqrt{n}|\alpha_i^*/C_1 - \alpha_j^*/C_2| \neq 0$ grows with n and is easy to distinguish, TSHT or CIIV cannot use that because weakness in γ_i and γ_j. Whereas, $|\tilde{c} - c| \cdot |\gamma_j^*| = |C_2| \cdot |\alpha_i^*/C_1 - \alpha_j^*/C_2| \neq 0$ are able to be utilized for penalized methods.

This phenomenon of accommodation of weak IVs for penalized methods is numerically demonstrated by simulation Case1(II), 1(IV), Case 2 and application in trade economic growth.

Appendix A2. SAIS condition and intuition of why MCP can circumvent it?

Correct selection of valid IVs is a more subtle and important issue in IV content. Recall Windmeijer et al. (2019) indicated that failure of consistent variable selection of sisVIVE is guaranteed if the SAIS condition holds. The SAIS condition was first proposed in Windmeijer et al. (2019) derived from Irrepresentable Condition (IRC) directly. SAIS condition originally comes from IRC known as (almost) necessary and sufficient condition for variable selection consistency of Lasso (Zhao and Yu, 2006) for $n^{-1}\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}$, i.e.,

$$\max_{j \in \mathcal{V}^*} \left\| \left(\tilde{\boldsymbol{Z}}_{\mathcal{V}^{c*}}^\top \tilde{\boldsymbol{Z}}_{\mathcal{V}^{c*}} \right)^{-1} \tilde{\boldsymbol{Z}}_{\mathcal{V}^{c*}}^\top \tilde{\boldsymbol{Z}}_j \right\|_1 \leq \xi < 1, \text{ for some } \xi \in [0,1). \tag{A1}$$

In the standard Lasso regression problem, the IRC only relates to the design matrix and holds for many standard designs (see corollaries in Zhao and Yu (2006)). However, the IRC on $\tilde{\boldsymbol{Z}}$ (instead of $\boldsymbol{Z}$) involves the first stage signal estimate $\widehat{\boldsymbol{\gamma}}$, which further complicates the verifiability of the SAIS condition. Typically, two-stage IV regression modeling exacerbates the difficulty of detecting valid IVs through penalized methods than support recovery problem in a simple linear model. While, among penalty choices, Lasso penalty even worsens it if first stage coefficients related SAIS condition hold.

The MCP penalty inherits a much weaker condition for oracle property than IRC that Lasso required(Zhao and Yu, 2006). (Zhang and Zhang, 2012, Theorem 6) has generalized the (Irrepresentable Condition) IRC in Lasso to a concave penalty in a linear regression problem. We briefly state the key result by defining two quantities:

$$\theta_{\text{select}} = \inf \left\{ \theta : \left\| \left(\frac{\tilde{\boldsymbol{Z}}_{\mathcal{V}^{c*}}^\top \tilde{\boldsymbol{Z}}_{\mathcal{V}^{c*}}}{n} \right)^{-1} p'_\lambda(\boldsymbol{\varphi}_{\mathcal{V}^{c*}} + \widehat{\boldsymbol{\alpha}}_{\mathcal{V}^{c*}}^{\text{or}}) \right\|_\infty \leq \theta\lambda, \forall \left\| \boldsymbol{\varphi}_{\mathcal{V}^{c*}} \right\|_\infty \leq \theta\lambda \right\}, \tag{A2}$$

$$\kappa_{\text{select}} = \sup \left\{ \| \tilde{\boldsymbol{Z}}_{\mathcal{V}^*}^\top \tilde{\boldsymbol{Z}}_{\mathcal{V}^{c*}} (\tilde{\boldsymbol{Z}}_{\mathcal{V}^{c*}}^\top \tilde{\boldsymbol{Z}}_{\mathcal{V}^{c*}})^{-1} p'_\lambda(\boldsymbol{\varphi}_{\mathcal{V}^{c*}} + \widehat{\boldsymbol{\alpha}}_{\mathcal{V}^{c*}}^{\text{or}}) \|_\infty / \lambda : \| \boldsymbol{\varphi}_{\mathcal{V}^{c*}} \|_\infty \leq \theta_{\text{select}} \lambda \right\}, \tag{A3}$$

where $\boldsymbol{\varphi}_{\mathcal{V}^{c*}}$ is a $|\mathcal{V}^{c*}|$-vector, and let the $\widehat{\boldsymbol{\alpha}}^{\text{or}}$ to be the oracle estimate, i.e., $\widehat{\boldsymbol{\alpha}}_{\mathcal{V}^{c*}}^{\text{or}} = (\tilde{\boldsymbol{Z}}_{\mathcal{V}^{c*}}^\top \tilde{\boldsymbol{Z}}_{\mathcal{V}^{c*}})^{-1} \tilde{\boldsymbol{Z}}_{\mathcal{V}^{c*}} \boldsymbol{Y}$ and $\widehat{\boldsymbol{\alpha}}_{\mathcal{V}^*}^{\text{or}} = \boldsymbol{0}$. The extended IRC for concave penalty required $\kappa_{\text{select}} < 1$ as the most crucial one to achieve selection consistency. When replace MCP penalty $p_\lambda^{\text{MCP}}(\boldsymbol{\alpha})$ with Lasso penalty $\lambda\|\boldsymbol{\alpha}\|_1$ whose coordinate sub-derivertive lies in $[-\lambda, \lambda]$, extended condition will be reduced to $\kappa_{\text{select}}(\boldsymbol{\alpha}) = \|\tilde{\boldsymbol{Z}}_{\mathcal{V}^*}^\top \tilde{\boldsymbol{Z}}_{\mathcal{V}^{c*}} (\tilde{\boldsymbol{Z}}_{\mathcal{V}^{c*}}^\top \tilde{\boldsymbol{Z}}_{\mathcal{V}^{c*}})^{-1}\|_\infty < 1$ as identical to IRC of Lasso (A1). By the nature of nearly unbiasedness characteristic of MCP, $p'_n(t) = 0 \; \forall |t| > \lambda\rho$. Once the mild condition $\min_{j \in \mathcal{V}^{c*}} |\widehat{\alpha}_j^{\text{or}}| > \lambda\rho$ holds, then it implies $\theta_{\text{select}} = 0$ with $\boldsymbol{\varphi}_{\mathcal{V}^{c*}} = \boldsymbol{0}$ and $\kappa_{\text{select}} = 0$ sequentially. Thus the extended IRC $\kappa_{\text{select}} < 1$ holds automatically for MCP but does not for Lasso. Consequently, this property is desirable that achieving exact support recovery regardless constraint of SAIS condition.

Appendix A3. I-LAMM Algorithm for MCP penalty

Theoretically, the proposed WIT estimator enjoys weaker conditions and better performance for low and high-dimension cases with weak IVs. Fan et al. (2018) proposed the iterative local adaptive majorize-minimization (I-LAMM) algorithm for non-convex regularized model. I-LAMM first contracts the initial in the neighborhood of the optimum solutions to serve as a better sparse coarse initial $\widehat{\boldsymbol{\alpha}}^{(1)}$ then tighten it to the solution under precision tolerance and compute in polynomial time.

To concretely, the I-LAMM algorithm combines the adaptive Local Linear Approximation (LLA) method and proximal gradient method or iterative shrinkage-thresholding (ISTA) algorithm (Daubechies et al., 2004). We adopt their method to solve a sequence of optimization problems through LLA,

$$\widehat{\boldsymbol{\alpha}}^{(t)} = \operatorname*{argmin}_{\boldsymbol{\alpha}} \frac{1}{2n}\|\boldsymbol{Y} - \widetilde{\boldsymbol{Z}}\boldsymbol{\alpha}\|_2^2 + \sum_{j=1}^{p} \left[p'_{\lambda}(|\widehat{\alpha}_j^{(t-1)}|)|\alpha_j|\right]. \tag{A4}$$

For $t = 1$ refers to contraction stage to obtain the better initial estimators falls in contraction region and $t = 2, 3, \ldots$, saying tightening stage, until it converges to stationary point. It is worthy to note Zou and Li (2008) use one-step LLA iteration to achieve the oracle estimator based on ols estimate as initial. Nevertheless, in our case, the ordinal ols estimate is not achievable under stringent conditions, Windmeijer et al. (2019) use the adaptive Lasso with embedded information in the median estimator as the one-step LLA to achieve the oracle property. Compared with the one-step method, iterations of $\boldsymbol{\lambda}^{(t-1)}$ (defined as below) in I-LAMM circumvent the stringent condition to obtain the root n initial estimator as the proper adaptive weight for ALasso.

But for each iteration (A4), the ISTA method is implemented to achieve the closed-form updating formula, which is the reduced model in Fan et al. (2018)'s derivation. For a given iteration t, let $k = 0, 1, 2, \ldots$ denotes the iteration in the proximal gradient updating, thus,

$$\begin{aligned}
\widehat{\boldsymbol{\alpha}}^{(t,k)} &= \operatorname*{argmin}_{\boldsymbol{\alpha}} \left\{ \frac{1}{2n}\left\|\boldsymbol{Y} - \widetilde{\boldsymbol{Z}}\widehat{\boldsymbol{\alpha}}^{(t,k-1)}\right\|_2^2 - \frac{1}{n}\left[\widetilde{\boldsymbol{Z}}(\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}^{(t,k-1)})\right]^{\top}(\boldsymbol{Y} - \widetilde{\boldsymbol{Z}}\widehat{\boldsymbol{\alpha}}^{(t,k-1)}) \right.\\
&\qquad \left. + \frac{\phi}{2}\|\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}}^{(t,k-1)}\|_2^2 + \sum_{j=1}^{p}\left[p'_{\lambda}(|\widehat{\alpha}_j^{(t-1)}|)|\alpha_j|\right] \right\}\\
&= \operatorname*{argmin}_{\boldsymbol{\alpha}} \left\{ \frac{\phi}{2}\left\|\boldsymbol{\alpha} - \left(\widehat{\boldsymbol{\alpha}}^{(t,k-1)} + \frac{1}{\phi n}\widetilde{\boldsymbol{Z}}^{\top}(\boldsymbol{Y} - \widetilde{\boldsymbol{Z}}\widehat{\boldsymbol{\alpha}}^{(t,k-1)})\right)\right\|_2^2 + \sum_{j=1}^{p}\left[p'_{\lambda}(|\widehat{\alpha}_j^{(t-1)}|)|\alpha_j|\right] \right\}\\
&= S\left(\widehat{\boldsymbol{\alpha}}^{(t,k-1)} + \frac{1}{\phi n}\widetilde{\boldsymbol{Z}}^{\top}(\boldsymbol{Y} - \widetilde{\boldsymbol{Z}}\widehat{\boldsymbol{\alpha}}^{(t,k-1)}), \frac{1}{\phi}\boldsymbol{\lambda}^{(t-1)}\right),
\end{aligned} \tag{A5}$$

where ϕ should be no smaller than the largest eigenvalue of $\widetilde{\boldsymbol{Z}}^{\top}\widetilde{\boldsymbol{Z}}$, or simply take it, to ensure the majorization and $S(\boldsymbol{x}, \boldsymbol{a})$ denotes the component-wise soft-thresholding operator, i.e. $S(\boldsymbol{x}, \boldsymbol{a})_j = \operatorname{sgn}(x_j)(|x_j| - a_j)_+$, with $\boldsymbol{\lambda}^{(t-1)} = \left(p'_n(|\widehat{\alpha}_1^{(t-1)}|), \ldots, p'_n(|\widehat{\alpha}_p^{(t-1)}|)\right)^{\top}$.

We adopt the first order optimality condition as a stopping criterion in the sub-problem. Let

$$\omega_{\boldsymbol{\lambda}^{(t-1)}}(\boldsymbol{\alpha}) = \min_{\boldsymbol{\xi} \in \partial|\boldsymbol{\alpha}|}\left\{\| - \widetilde{\boldsymbol{Z}}^{\top}(\boldsymbol{Y} - \widetilde{\boldsymbol{Z}}\boldsymbol{\alpha}) + \boldsymbol{\lambda}^{(t-1)} \odot \boldsymbol{\xi}\|_{\infty}\right\}$$

as a natural measure of suboptimality of $\boldsymbol{\alpha}$, where $\odot$ is the Hadamard product. Once $\omega_{\widehat{\boldsymbol{\lambda}}^{(t-1)}}(\boldsymbol{\alpha}^{(t,k)}) \leq \delta$, where δ is a pre-determined tolerance and is assigned δ_c and δ_t for contraction and tightening stage respectively, we stop the inner iteration and say $\widehat{\boldsymbol{\alpha}}^{(t)} = \widehat{\boldsymbol{\alpha}}^{(t,k)}$ as the δ-optimal solution in the sub-problem. Notably, it is an early stopped variant of ISTA method in each sub-problem to obtain $\widehat{\boldsymbol{\alpha}}^{(t)}$ from $\widehat{\boldsymbol{\alpha}}^{(t-1)}$. The following pseudo-code Algorithm 2 demonstrate the details of I-LAMM algorithm for WIT.

Algorithm 2 I-LAMM algorithm for $\widehat{\boldsymbol{\alpha}}$ with MCP penalty

```
Input: $\boldsymbol{Y}, \widetilde{\boldsymbol{Z}}, \widehat{\boldsymbol{\alpha}}^{(0)}, \lambda, \phi = \lambda_{\max}(\widetilde{\boldsymbol{Z}}^\top \widetilde{\boldsymbol{Z}}), \delta_c = 10^{-3}, \delta_t = 10^{-5}$
for $t = 1, 2, \ldots$ do
    $\boldsymbol{\lambda}^{(t-1)} = (p'_\lambda(|\widehat{\alpha}_1^{(t-1)}|), \ldots, p'_\lambda(|\widehat{\alpha}_p^{(t-1)}|))^\top$        ▷ $p_n^\top$ is the derivative of MCP penalty
    $\widehat{\boldsymbol{\alpha}}^{(t,0)} = \widehat{\boldsymbol{\alpha}}^{(t-1)}$
    for $k = 1, 2, \ldots$ do
        $\widehat{\boldsymbol{\alpha}}^{(t,k)} = S(\widehat{\boldsymbol{\alpha}}^{(t,k-1)} + \frac{1}{n\phi}\widetilde{\boldsymbol{Z}}^\top(\boldsymbol{Y} - \widetilde{\boldsymbol{Z}}\widehat{\boldsymbol{\alpha}}^{(t,k-1)}), \frac{1}{\phi}\boldsymbol{\lambda}^{(t-1)}))$        ▷ LAMM updating
        $\omega_{\boldsymbol{\lambda}^{(t-1)}}(\widehat{\boldsymbol{\alpha}}^{(t,k)}) = \min_{\boldsymbol{\xi} \in \partial|\widehat{\boldsymbol{\alpha}}^{(t,k)}|}\left\{\| - \widetilde{\boldsymbol{Z}}^\top(\boldsymbol{Y} - \widetilde{\boldsymbol{Z}}\widehat{\boldsymbol{\alpha}}^{(t,k)}) + \boldsymbol{\lambda}^{(t-1)} \odot \boldsymbol{\xi}\|_\infty\right\}$
        if $\omega_{\boldsymbol{\lambda}^{(t-1)}}(\widehat{\boldsymbol{\alpha}}^{(t,k)}) \leq \mathbf{1}(t=1)\delta_c + \mathbf{1}(t \neq 1)\delta_t$ then
            $\widehat{\boldsymbol{\alpha}}^{(t)} = \widehat{\boldsymbol{\alpha}}^{(t,k)}$
            break
        end if
    end for
    if $\|\widehat{\boldsymbol{\alpha}}^{(t)} - \widehat{\boldsymbol{\alpha}}^{(t-1)}\|_\infty \leq \delta_t$ then
        $\widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\alpha}}^{(t)}$
        break
    end if
end for
Output: $\widehat{\boldsymbol{\alpha}}$
```

Appendix A4. Additional Information on Real Data Analysis

The following table S1 provide the detailed summary statistics of variables used in our analysis.

Appendix B. Proofs

We follow the sequence of appearing to show the proofs. To proceed, some lemmas from literature are needed. We first restate their results.

Appendix B1. Ancillary lemma

Lemma S1. *(Lemma 1 and 2 in (Bekker, 1994) and Lemma A.1 in (Kolesár et al., 2015)). Consider the quadratic form $Q = (M + U)^\top C(M + U)$, where $M \in \mathbb{R}^{n\times S}, C \in \mathbb{R}^{n\times n}$ are non-stochastic, C is symmetric and idempotent with rank J_n which may depend on n, and $U = (u_1, \ldots, u_n)^\top$, with $u_i \sim [0, \Omega]$ iid. Let $a \in \mathbb{R}^S$ be a non-stochastic vector. Then:*

Table S1. Summary statistics of main variables

	Notation	Type	Mean	Std	Median	Min	Max
log(GDP)	$\log(Y)$	Response	10.177	1.0102	10.416	7.463	12.026
Trade	T	Endogenous Variable	0.866	0.520	0.758	0.198	4.128
log(Population)	$\boldsymbol{S}_1$	Control Variable	1.382	1.803	1.480	−3.037	6.674
log(Land Area)	$\boldsymbol{S}_2$	Control Variable	11.726	2.260	12.015	5.680	16.611
$\widehat{T}$ (proxy for trade)	$\boldsymbol{Z}_1$	IV	0.093	0.052	0.079	0.015	0.297
log(Water Area)	$\boldsymbol{Z}_2$	IV	6.756	3.654	7.768	0	13.700
log(Land Boundaries)	$\boldsymbol{Z}_3$	IV	6.507	2.920	7.549	0	10.005
% Forest	$\boldsymbol{Z}_4$	IV	29.89	22.380	30.62	0	98.26
% Arable Land	$\boldsymbol{Z}_5$	IV	40.947	21.549	42.062	0.558	82.560
Languages	$\boldsymbol{Z}_6$	IV	1.873	2.129	1	1	16
Annual Freshwater	$\boldsymbol{Z}_7$	IV	2.190	2.129	2.155	-2.968	8.767

Source: FR99, the World Bank, and CIA world Factbook.

(a) If u_i has finite fourth moments:

$$
\begin{aligned}
\mathbb{E}[Q \mid C] =& M^\top CM + J_N\Omega \\
\operatorname{var}(Qa \mid C) =& a^\top \Omega a M^\top CM + a^\top M^\top CMa\Omega + \Omega aa^\top M^\top CM + MCMaa^\top\Omega + J_N\left(a^\top\Omega a\Omega + \Omega aa^\top\Omega\right) \\
&+ d_C^\top d_C\left[\mathbb{E}\left(a^\top u\right)^2 uu^\top - a^\top\Omega aa^\top\Omega - a^\top\Omega a\Omega\right] + 2d_C^\top CMa\mathbb{E}\left[\left(a^\top u\right)uu^\top\right] \\
&+ M^\top Cd_C\mathbb{E}\left[\left(a^\top u\right)^2 u^\top\right] + \mathbb{E}\left[\left(a^\top u\right)^2 u\right] d_C^\top CM
\end{aligned}
$$

where $d_C = \operatorname{diag}(C)$. If the distribution of u_i is normal, the last two lines of the variance expression equals zero.

(b) Suppose that the distribution of u_i is normal, and that, as $n \to \infty$:

$$M^\top CM/N \to Q_{CM}, J_n/n \to \alpha_r$$

where the elements c_{is} of C may depend on N. Then:

$$\sqrt{n}(Qa/n - \mathbb{E}Qa/n) \xrightarrow{d} \mathcal{N}(0, V),$$

where $V = a^\top\Omega aQ_{CM} + a^\top Q_{CM}a\Omega + \Omega aa^\top Q_{CM} + Q_{CM}aa^\top\Omega + \alpha_r\left(a^\top\Omega a\Omega + \Omega aa^\top\Omega\right)$.

Appendix B2. Proof of Theorem 1

Proof. Firstly, we prove the procedure (6) can generate the G different groups of $\mathcal{P}_c$ satisfying the requirements. With direct calculation, $\mathcal{P}_c = \{\tilde{\beta}^c, \tilde{\boldsymbol{\alpha}}^c, \tilde{\boldsymbol{\epsilon}}^c\} = \{\beta^* + c, \boldsymbol{\alpha}^* - c\boldsymbol{\gamma}^*, \boldsymbol{\epsilon} - c\boldsymbol{\eta}\}$ and $E(\tilde{\boldsymbol{\epsilon}}^c) = E(\boldsymbol{\epsilon}) - cE(\boldsymbol{\eta}) = \mathbf{0}$ for $c = \alpha_j^*/\gamma_j^*, j \in \mathcal{I}_c$. Therefore, $\tilde{\boldsymbol{\alpha}}_{\mathcal{I}_c}^c = \boldsymbol{\alpha}_{\mathcal{I}_c}^* - c\boldsymbol{\gamma}_{\mathcal{I}_c}^* = \boldsymbol{\alpha}_{\mathcal{I}_c}^* - \boldsymbol{\alpha}_{\mathcal{I}_c}^* = \mathbf{0}$ and $\tilde{\alpha}_j^c = \alpha_j^* - c\gamma_j^* \neq 0$ for $j \notin \mathcal{I}_c$. Thought out all possible $c = \{\alpha_j^*/\gamma_j^* : j \notin \mathcal{V}^*\}$, we conclude the procedure (6) has generate G groups additional DGP$_c$. The exhaustive and mutual exclusive property of $\mathcal{I}_c$ and $\mathcal{V}^*$ is also guaranteed by construction.

Second, we prove there is no more possible DGP with sparse structure and zero mean structural error by contradiction.

Assume there is a additional DGP $\{\breve{\beta}, \breve{\boldsymbol{\alpha}}, \breve{\epsilon}\}$ differentiating with $\mathcal{P}_c$ and $\mathcal{P}_0$ but still have $E(\breve{\epsilon}) = \mathbf{0}$ and zero(s) component in $\breve{\alpha}$, i.e., $\breve{\mathcal{I}} = \{j : \breve{\boldsymbol{\alpha}}_j = 0\} \neq \varnothing$. By the property of exhaustive and mutually exclusive of $\mathcal{V}^*$ and $\{\mathcal{I}_c\}_{c \neq 0}$, WLOG, we assume $\mathcal{I}_g \cap \breve{\mathcal{I}} \neq \varnothing$ for some g and $\text{DGP}_g = \{\tilde{\beta}^g, \tilde{\boldsymbol{\alpha}}^g, \tilde{\boldsymbol{\epsilon}}^g\}$. Since $E(\breve{\boldsymbol{\epsilon}}) = E(\tilde{\boldsymbol{\epsilon}}^g) = \mathbf{0}$, it suffices to show the conflict in terms of moment condition (3). Hence, $\{\tilde{\beta}^g, \tilde{\boldsymbol{\alpha}}^g\}$ and $\{\breve{\beta}, \breve{\boldsymbol{\alpha}}\}$ all are solutions of

$$\boldsymbol{\Gamma}^* = \boldsymbol{\alpha} + \beta \boldsymbol{\gamma}^*.$$

For $j \in \mathcal{I}_g \cap \breve{\mathcal{I}}$, $\Gamma_j^* = \tilde{\beta}^g \gamma_j^* = \breve{\beta} \gamma_j^*$ forces $\tilde{\beta}^g = \breve{\beta}$. In turn, $\forall j \notin \mathcal{I}_g \cap \breve{\mathcal{I}}$,

$$\Gamma_j^* - \breve{\beta} \gamma_j^* = \tilde{\alpha}_j = \breve{\alpha}_j.$$

Thus, $\{\tilde{\beta}^g, \tilde{\boldsymbol{\alpha}}^g\}$ and $\{\breve{\beta}, \breve{\boldsymbol{\alpha}}\}$ are equivalent. So as $\tilde{\boldsymbol{\epsilon}}^g$ and $\breve{\boldsymbol{\epsilon}}$ since

$$\tilde{\boldsymbol{\epsilon}}^c = \boldsymbol{Y} - \boldsymbol{D}\tilde{\beta}^g - \boldsymbol{Z}\tilde{\boldsymbol{\alpha}}^g = \boldsymbol{Y} - \boldsymbol{D}\breve{\beta} - \boldsymbol{Z}\breve{\boldsymbol{\alpha}} = \breve{\boldsymbol{\epsilon}}.$$

Hence DGP $\{\breve{\beta}, \breve{\boldsymbol{\alpha}}, \breve{\epsilon}\}$ are equivalent to DGP_g and forms a contradiction. It concludes the procedure (6) can produce all possible DGPs. □

Appendix B3. Proof of Theorem 2

Proof. Recall the construction that $\exists i \in \{0, \ldots, G\}$, let $\mathcal{F} = \{f : \mathcal{P} \in \mathcal{Q} \to \mathbb{R}; f(\mathcal{P}_i) < f(\mathcal{P}_j), \forall j \neq i\}$ and $\mathcal{G} = \{g = \text{argmin}_{\mathcal{P} \in \mathcal{Q}} f(\mathcal{P}); f \in \mathcal{F}\}$ To show any elements in $\mathcal{G}$ can are sufficient to identify one specify DGP in $\mathcal{Q}$, we assume $\mathcal{P}_m$ and $\mathcal{P}_n$ both are elements in $\mathcal{Q}$ but with minimum value under same specific $f \in \mathcal{F}$ and its corresponding $g \in \mathcal{G}$. According to Theorem 1, β must be different in $\mathcal{P}_m$ and $\mathcal{P}_n$ if $\mathcal{P}_m \neq \mathcal{P}_n$. Hence,

(1) $\mathcal{P}_m = \mathcal{P}_n$, it leads to $f(\mathcal{P}_m) = f(\mathcal{P}_n) < f(\mathcal{P}_j)$ for any $j \neq m, n$ and sharing the same β in $\mathcal{P}_m$ and $\mathcal{P}_n$.

(2) $\mathcal{P}_m \neq \mathcal{P}_n$, apply the minimum value assumption of f on $\mathcal{P}_m \neq \mathcal{P}_n$ respectively, we obtain

$$f(\mathcal{P}_m) < f(\mathcal{P}_i), \quad \forall i \neq m \quad \text{and} \quad f(\mathcal{P}_n) < f(\mathcal{P}_j), \quad \forall j \neq n. \tag{B6}$$

Taking $i = m, j = n$, it forms a contradiction that

$$f(\mathcal{P}_m) < f(\mathcal{P}_n) < f(\mathcal{P}_m) \tag{B7}$$

holds strictly. Therefore, $\mathcal{P}_m$ and $\mathcal{P}_n$ must be equivalent and identifiable under f.

Due to the arbitrage of $f \in \mathcal{F}$, it concludes (a) $\mathcal{G} \subseteq \mathcal{H}$.

Move to the second part (b): it states that if $\exists h \in \mathcal{H} : Q \to \mathcal{P}_0$ and such h is the necessary condition for identifying β^*, then it must have $|\mathcal{H}| = 1$ and $\exists h \in \mathcal{H} : Q \to \mathcal{P}_0$. We prove by contradiction. Consider two case (1) $\forall h \in \mathcal{H} : \mathcal{Q} \to \mathcal{P}_k$, where $\mathcal{P}_k \neq \mathcal{P}_0$ for some k, and (2) $1 < |\mathcal{H}| \leq G + 1$ with $\exists h \in \mathcal{H} : \mathcal{Q} \to \mathcal{P}_0$:

(1) $\forall h \in \mathcal{H} : \mathcal{Q} \to \mathcal{P}_k$, where $\mathcal{P}_k \neq \mathcal{P}_0$ for some k: It directly forms contradiction of $\exists h \in \mathcal{H} : \mathcal{Q} \to \mathcal{P}_0$.

(2) $1 < |\mathcal{H}| \leq G+1$ with $\exists h \in \mathcal{H} : \mathcal{Q} \to \mathcal{P}_0$: It at least exist two distinct mappings $h_m(\mathcal{Q}) \cong P_m$, $h_n(\mathcal{Q}) \cong P_n$ and $\mathcal{P}_m \neq \mathcal{P}_n$. By assumption that there is necessary condition $h_i \in \mathcal{H}$ of identifying β^*. Thus, h_i must be image equivalent to h_m or h_n. WLOG, we let $h_0 = h_m$ and $\mathcal{P}_m = \mathcal{P}_0 \neq \mathcal{P}_n$. Since necessity of h_m, the contrapositive arguments must hold. That is supposing $h_m(\mathcal{Q}) \neq P_0$ leads to there is not other mappings such that maps $\mathcal{Q}$ to $\mathcal{P}_0$. However, the distinct image equivalent mapping h_n forms contradiction that it is possible to pick $h_n(\mathcal{Q}) = \mathcal{P}_0$ since it only requires h_m has different image with h_n.

Together with above arguments, it conclude (b) that there is no necessary condition of identification of β^* unless $\exists h \in \mathcal{H} : \mathcal{Q} \to \mathcal{P}_0$ and $|\mathcal{H}| = 1$. □

Appendix B4. Proof of Corollary 1

Proof. To show that, we only need to consider some specific constructions of $g \in \mathcal{G} \subseteq \mathcal{H}$. Let $f_i = \mathbf{1}(\alpha_i = 0)$ for $i = 1, 2 \ldots, p$ to be the individual valid IV indicators. According to Theorem 1, non-empty set $\mathcal{I}_c$ and $\mathcal{V}^*$ are are exhaustive and mutual exclusive zeros structure in $\tilde{\boldsymbol{\alpha}}^c$ and $\boldsymbol{\alpha}^*$ correspondingly. Thus, $\{f_i : i = 1, 2, \ldots, p\}$ must enumerate all solutions. Thus,

$$G + 1 = |\{f_i : i = 1, 2, \ldots, p\}| \leq |\mathcal{G}| \leq |\mathcal{H}| <= G + 1 \tag{B8}$$

holds. It leads to $|\mathcal{H}| = G + 1$. According to Theorem 2, there is not a necessary and sufficient condition to identify β^*, unless $G = 0$, i.e. all potential IVs are valid. □

Appendix B5. Proof of Proposition 1

Proof.

$$\begin{aligned} \boldsymbol{\alpha}^* = & \underset{\mathcal{P}=\{\beta,\boldsymbol{\alpha},\boldsymbol{\epsilon}\}\in\mathcal{Q}}{\operatorname{argmin}} \ p_\lambda^{\text{pen}}(\boldsymbol{\alpha}), \\ \iff & \sum_{j=1}^{p} p_\lambda^{\text{pen}}(\alpha_j^*) < \sum_{j=1}^{p} p_\lambda^{\text{pen}}(\tilde{\alpha}_j^c) \\ \iff & \sum_{j\in\mathcal{V}^{c*}} p_\lambda^{\text{pen}}(\alpha_j^*) < \sum_{j\in\mathcal{I}_c^c} p_\lambda^{\text{pen}}(\tilde{\alpha}_j^c), \end{aligned} \tag{B9}$$

where $\mathcal{I}_c = \{j : \alpha_j^*/\gamma_j^* = c, c \neq 0\}$ and $\mathcal{I}_c^c$ means the complement of $\mathcal{I}_c$. By the Assumption 7: $\boldsymbol{\alpha}^* = \operatorname{argmin}_{\mathcal{P}=\{\beta,\boldsymbol{\alpha},\boldsymbol{\epsilon}\}\in\mathcal{Q}} \|\boldsymbol{\alpha}\|_0$, we have $|\mathcal{V}^{c*}| < |\mathcal{I}_c^c|$.

Define $\pi(\mathcal{I}_c^c, \mathcal{V}^{c*})$ as $|\mathcal{V}^{c*}|$-combination of $\mathcal{I}_c^c$. Now we rewrite (B9) as

$$\sum_{j\in\mathcal{V}^{c*}} p_\lambda^{\text{pen}}(\alpha_j^*) - \sum_{k\in\pi(\mathcal{I}_c^c,\mathcal{V}^{c*})} p_\lambda^{\text{pen}}(\tilde{\alpha}_k^c) < \sum_{l\in\mathcal{I}_c^c/\pi(\mathcal{I}_c^c,\mathcal{V}^{c*})} p_\lambda^{\text{pen}}(\tilde{\alpha}_l^c), \tag{B10}$$

and it should holds for any $|\mathcal{V}^{c*}|$-combination and any $\mathcal{P}_0$. Note that RHS in (B10) are non-negative by definition of penalty and $|\mathcal{I}_c^c/\pi(\mathcal{I}_c^c, \mathcal{V}^{c*})| > 0$.

By the arbitrage of $\pi(\mathcal{I}_c^c, \mathcal{V}^{c*})$ and $\mathcal{P}_0$, we consider the worst case where

$$|\tilde{\alpha}_l^c| < |\tilde{\alpha}_k^c| \text{ and } |\tilde{\alpha}_k^c| < |\alpha_j^*|,$$

for $\forall k \in \pi(\mathcal{I}_c^c, \mathcal{V}^{c*}), l \in \mathcal{I}_c^c/\pi(\mathcal{I}_c^c, \mathcal{V}^{c*})$ and for each pair $(j,k) \in (\mathcal{V}^{c*}, \pi(\mathcal{I}_c^c, \mathcal{V}^{c*}))$. Using Taylor expansion and mean value theorem, we obtain

$$\begin{aligned}
&\Big\{ \sum_{(j,k)\in(\mathcal{V}^{c*},\pi(\mathcal{I}_c^c,\mathcal{V}^{c*}))} |\alpha_j^*| - |\tilde{\alpha}_k^c| \Big\} \min_{(j,k)\in(\mathcal{V}^{c*},\pi(\mathcal{I}_c^c,\mathcal{V}^{c*}))} [p_\lambda^{\text{pen}\prime}(\bar{\xi}_{(j,k)})] \\
&< \sum_{j\in\mathcal{V}^{c*}} p_\lambda^{\text{pen}}(\alpha_j^*) - \sum_{k\in\pi(\mathcal{I}_c^c,\mathcal{V}^{c*})} p_\lambda^{\text{pen}}(\tilde{\alpha}_k^c) \\
&< \sum_{l\in\mathcal{I}_c^c/\pi(\mathcal{I}_c^c,\mathcal{V}^{c*})} p_\lambda^{\text{pen}}(\tilde{\alpha}_l^c) \\
&< \|\tilde{\boldsymbol{\alpha}}^c_{\mathcal{I}_c^c/\pi(\mathcal{I}_c^c,\mathcal{V}^{c*})}\|_1 \max_{l\in\mathcal{I}_c^c/\pi(\mathcal{I}_c^c,\mathcal{V}^{c*})} [p_\lambda^{\text{pen}\prime}(\bar{\xi}_l)],
\end{aligned}$$

where $\bar{\xi}_{(j,k)} \in (|\tilde{\alpha}_k^c|, |\alpha_j^*|)$ and $\bar{\xi}_l \in (0, |\tilde{\alpha}_l^c|)$. Thus, It leads to

$$\frac{\min_{(j,k)\in(\mathcal{V}^{c*},\pi(\mathcal{I}_c^c,\mathcal{V}^{c*}))} [p_\lambda^{\text{pen}\prime}(\bar{\xi}_{(j,k)})]}{\max_{l\in\mathcal{I}_c^c/\pi(\mathcal{I}_c^c,\mathcal{V}^{c*})} [p_\lambda^{\text{pen}\prime}(\bar{\xi}_l)]} < \frac{\|\tilde{\boldsymbol{\alpha}}^c_{\mathcal{I}_c^c/\pi(\mathcal{I}_c^c,\mathcal{V}^{c*})}\|_1}{\Big\{ \sum_{(j,k)\in(\mathcal{V}^{c*},\pi(\mathcal{I}_c^c,\mathcal{V}^{c*}))} |\alpha_j^*| - |\tilde{\alpha}_k^c| \Big\}}. \tag{B11}$$

Because $p_\lambda^{\text{pen}\prime}$ is free of $\mathcal{P}_0$, that RHS in (B11) could be much smaller than 1 in some extreme case leads to $p_\lambda^{\text{pen}\prime}$ is a bounded monotone deceasing function. That is to say, p_λ^{pen} should be concave penalty.

Consider another case where that keep $\boldsymbol{\alpha}^*$ fixed, but vary γ^* to make $\tilde{\boldsymbol{\alpha}}^c$ be in the same order with $\kappa(n)$ defined in Assumption 6. Again by (B11),

$$\begin{aligned}
\min_{(j,k)\in(\mathcal{V}^{c*},\pi(\mathcal{I}_c^c,\mathcal{V}^{c*}))} [p_\lambda^{\text{pen}\prime}(\bar{\xi}_{(j,k)})] &< \frac{\|\tilde{\boldsymbol{\alpha}}^c_{\mathcal{I}_c^c/\pi(\mathcal{I}_c^c,\mathcal{V}^{c*})}\|_1}{\Big\{ \sum_{(j,k)\in(\mathcal{V}^{c*},\pi(\mathcal{I}_c^c,\mathcal{V}^{c*}))} |\alpha_j^*| - |\tilde{\alpha}_k^c| \Big\}} \cdot \max_{l\in\mathcal{I}_c^c/\pi(\mathcal{I}_c^c,\mathcal{V}^{c*})} [p_\lambda^{\text{pen}\prime}(\bar{\xi}_l)] \\
&< \frac{\lambda\|\tilde{\boldsymbol{\alpha}}^c_{\mathcal{I}_c^c/\pi(\mathcal{I}_c^c,\mathcal{V}^{c*})}\|_1}{\Big\{ \sum_{(j,k)\in(\mathcal{V}^{c*},\pi(\mathcal{I}_c^c,\mathcal{V}^{c*}))} |\alpha_j^*| - |\tilde{\alpha}_k^c| \Big\}} \\
&\asymp \frac{\lambda\kappa(n)(|\mathcal{I}_c^c| - |\mathcal{V}^{c*}|)}{\Big\{ \sum_{(j,k)\in(\mathcal{V}^{c*},\pi(\mathcal{I}_c^c,\mathcal{V}^{c*}))} |\alpha_j^*| - |\tilde{\alpha}_k^c| \Big\}} \asymp C\lambda\kappa(n).
\end{aligned}$$

Thus, it concludes that $p_\lambda^{\text{pen}\prime}(t) = O(\lambda\kappa(n))$ for any $t > \kappa(n)$. □

Appendix B6. Proof of Lemma 1

Proof. $\tilde{\boldsymbol{Z}}_j = M_{\widehat{\boldsymbol{D}}}\boldsymbol{Z}_j$, $\boldsymbol{Z}_j$ is the j-th instrument. The target is to analyze the rate of $\tilde{\boldsymbol{Z}}_j^\top \boldsymbol{\epsilon}/n$ is not inflated. To proceed, $|\tilde{\boldsymbol{Z}}_j^\top \boldsymbol{\epsilon}/n| \le |\boldsymbol{Z}_j^\top \boldsymbol{\epsilon}/n| + |\boldsymbol{Z}_j^\top P_{\widehat{\boldsymbol{D}}}\boldsymbol{\epsilon}/n|$. The first term is known to be $O_p(n^{-1/2})$. For the second term, we have,

$$|\boldsymbol{Z}_j^\top P_{\widehat{\boldsymbol{D}}}\boldsymbol{\epsilon}/n| = \underbrace{\frac{|\boldsymbol{Z}_j^\top \widehat{\boldsymbol{D}}|}{n}}_{(I)} \cdot \underbrace{\frac{|\boldsymbol{\epsilon}^\top \widehat{\boldsymbol{D}}|}{n}}_{(II)} \cdot \left(\underbrace{\frac{|\widehat{\boldsymbol{D}}^\top \widehat{\boldsymbol{D}}|}{n}}_{(III)} \right)^{-1}. \tag{B12}$$

Notably, $\widehat{\boldsymbol{D}} = \boldsymbol{Z}\boldsymbol{\gamma}^* + P_{\boldsymbol{Z}}\boldsymbol{\eta}$ is a projected endogenous variable and actually consists of random part and non-random part. Then we analyze size of the above three terms sequentially. For (I), we have

$$\frac{\boldsymbol{Z}_j^\top \widehat{\boldsymbol{D}}}{n} = \frac{\boldsymbol{Z}_j^\top(\boldsymbol{Z}\boldsymbol{\gamma}^* + \boldsymbol{\eta})}{n} = (\boldsymbol{Q}_{nj})^\top \gamma^* + \frac{\boldsymbol{Z}_j^\top \boldsymbol{\epsilon}}{n} = \boldsymbol{Q}_{nj}^\top \gamma^* + O_p(n^{-1/2}). \tag{B13}$$

With regard to (II),

$$\frac{\boldsymbol{\epsilon}^\top \widehat{\boldsymbol{D}}}{n} = \frac{\boldsymbol{\epsilon}^\top P_{\boldsymbol{Z}}(\boldsymbol{Z}\boldsymbol{\gamma}^* + \boldsymbol{\eta})}{n} = \frac{\boldsymbol{\epsilon}^\top \boldsymbol{Z}\boldsymbol{\gamma}^*}{n} + \frac{\boldsymbol{\epsilon} P_{\boldsymbol{Z}}\boldsymbol{\eta}}{n}. \tag{B14}$$

Within this decomposition, we first have $E(\boldsymbol{\epsilon}^\top \boldsymbol{Z}\boldsymbol{\gamma}^*/n) = E(E(\boldsymbol{\epsilon}^\top|\boldsymbol{Z})\boldsymbol{Z}\boldsymbol{\gamma}^*) = 0$ and

$$\mathrm{var}(\boldsymbol{\epsilon}^\top \boldsymbol{Z}\boldsymbol{\gamma}^*/n) = E(\mathrm{var}(\boldsymbol{\epsilon}^\top \boldsymbol{Z}\boldsymbol{\gamma}^*/n|\boldsymbol{Z})) = E(\sigma_\epsilon^2 \boldsymbol{\gamma}^{*\top}\boldsymbol{Z}^\top \boldsymbol{Z}\boldsymbol{\gamma}^*/n^2) = O(\boldsymbol{\gamma}^{*\top}\boldsymbol{Q}_n\boldsymbol{\gamma}^*/n).$$

Thus we obtain $\boldsymbol{\epsilon}^\top \boldsymbol{Z}\boldsymbol{\gamma}^*/n = O_P(\sqrt{\boldsymbol{\gamma}^{*\top}\boldsymbol{Q}_n\boldsymbol{\gamma}^*/n})$. Regarding the second term, we have $E(\boldsymbol{\epsilon}^\top P_{\boldsymbol{Z}}\boldsymbol{\eta}/n) = E(\mathrm{tr}(P_{\boldsymbol{Z}}\boldsymbol{\eta}\boldsymbol{\epsilon}^\top)/n) = \mathrm{tr}(E(P_{\boldsymbol{Z}}E(\boldsymbol{\eta}\boldsymbol{\epsilon}^\top|\boldsymbol{Z}))/n) = \sigma_{\epsilon,\eta}^2 p/n$ and

$$\begin{aligned}\mathrm{var}(\boldsymbol{\epsilon}^\top P_{\boldsymbol{Z}}\boldsymbol{\eta}/n) = E(\mathrm{var}(\boldsymbol{\epsilon}^\top P_{\boldsymbol{Z}}\boldsymbol{\eta}/n|\boldsymbol{Z})) &= E(2\sigma_\epsilon^2\sigma_\eta^2[\mathrm{tr}(P_{\boldsymbol{Z}})/n^2]) + E([n^{-2}\sum_i (P_{\boldsymbol{Z}})_{ii}^2])([\sigma_{\epsilon,\eta}^2]^2 - 2\sigma_\epsilon^2\sigma_\eta^2) \\ &\le O(p/n^2) + O(p/n^2) = O(p/n^2),\end{aligned}$$

where the inequality holds since $\mathrm{tr}(P_{\boldsymbol{Z}}) = p$, $(P_{\boldsymbol{Z}})_{ii} \in (0,1)$ and $\sum_i (P_{\boldsymbol{Z}})_{ii}^2 \le \sum_i (P_{\boldsymbol{Z}})_{ii} = p$. Thus, we conclude the size of (II) is $O_P(\sqrt{\boldsymbol{\gamma}^{*\top}\boldsymbol{Q}\boldsymbol{\gamma}^*/n} \vee \sqrt{p/n^2}) + \sigma_{\epsilon,\eta}^2 p/n$.

Now we turn to (III), with similar argument,

$$\begin{aligned}\frac{\widehat{\boldsymbol{D}}^\top \widehat{\boldsymbol{D}}}{n} = \frac{\boldsymbol{D}^\top P_{\boldsymbol{Z}}\boldsymbol{D}}{n} &= \frac{\boldsymbol{\gamma}^{*\top}\boldsymbol{Z}^\top \boldsymbol{Z}\boldsymbol{\gamma}^*}{n} + 2\frac{\boldsymbol{\eta}^\top \boldsymbol{Z}\boldsymbol{\gamma}^*}{n} + \frac{\boldsymbol{\eta}^\top P_{\boldsymbol{Z}}\boldsymbol{\eta}}{n} \\ &= \boldsymbol{\gamma}^{*\top}\boldsymbol{Q}_n\boldsymbol{\gamma}^* + o_p(1) + O_P(\sqrt{\boldsymbol{\gamma}^{*\top}\boldsymbol{Q}_n\boldsymbol{\gamma}^*/n} \vee \sqrt{p/n^2}) + \sigma_\eta^2 p/n \\ &= \boldsymbol{\gamma}^{*\top}\boldsymbol{Q}_n\boldsymbol{\gamma}^* + \sigma_\eta^2 p/n + O_P(\sqrt{\boldsymbol{\gamma}^{*\top}\boldsymbol{Q}_n\boldsymbol{\gamma}^*/n} \vee \sqrt{p/n^2}).\end{aligned}$$

Therefore, together with above three terms, we are able to derive the size of $|\boldsymbol{Z}_j^\top P_{\widehat{\boldsymbol{D}}}\boldsymbol{\epsilon}/n|$,

$$\begin{aligned}\boldsymbol{Z}_j^\top P_{\widehat{\boldsymbol{D}}}\boldsymbol{\epsilon}/n &= \frac{\left[\boldsymbol{Q}_{nj}^\top \gamma^* + O_p(n^{-1/2})\right] \cdot \left[\sigma_{\epsilon,\eta}^2 p/n + O_P(\sqrt{\boldsymbol{\gamma}^{*\top}\boldsymbol{Q}_n\boldsymbol{\gamma}^*/n} \vee \sqrt{p/n^2})\right]}{\boldsymbol{\gamma}^{*\top}\boldsymbol{Q}_n\boldsymbol{\gamma}^* + \sigma_\eta^2 p/n + O_P(\sqrt{\boldsymbol{\gamma}^{*\top}\boldsymbol{Q}_n\boldsymbol{\gamma}^*/n} \vee \sqrt{p/n^2})} \\ &= \sigma_{\epsilon,\eta}^2 p/n \cdot \frac{\boldsymbol{Q}_{nj}^\top \boldsymbol{\gamma}^*}{\boldsymbol{\gamma}^{*\top}\boldsymbol{Q}_n\boldsymbol{\gamma}^* + \sigma_\eta^2 p/n} + O_p(n^{-1/2})\end{aligned} \tag{B15}$$

□

Appendix B7. Proof of Lemma 2

The proof technique of Lemma 2 is simple but constructive.

Proof. For any given $\boldsymbol{\gamma}^* \neq \mathbf{0}$, the transformed $\tilde{\boldsymbol{Z}} = M_{\widehat{\boldsymbol{D}}}\boldsymbol{Z}$, where $\widehat{\boldsymbol{D}} = P_{\boldsymbol{Z}}\boldsymbol{D} = \boldsymbol{Z}\widehat{\boldsymbol{\gamma}}$. Thus,

$$\begin{aligned} \tilde{\boldsymbol{Z}} &= \boldsymbol{Z} - \boldsymbol{Z}\widehat{\boldsymbol{\gamma}}\left(\widehat{\boldsymbol{\gamma}}^\top \boldsymbol{Z}^\top \boldsymbol{Z}\widehat{\boldsymbol{\gamma}}\right)^{-1}\widehat{\boldsymbol{\gamma}}^\top \boldsymbol{Z}^\top \boldsymbol{Z}, \\ \boldsymbol{C}_n = \frac{\tilde{\boldsymbol{Z}}^\top \tilde{\boldsymbol{Z}}}{n} &= \left(\frac{\boldsymbol{Z}^\top \boldsymbol{Z}}{n}\right) - \left(\frac{\boldsymbol{Z}^\top \boldsymbol{Z}}{n}\right)\widehat{\boldsymbol{\gamma}}\left(\widehat{\boldsymbol{\gamma}}^\top\left(\frac{\boldsymbol{Z}^\top \boldsymbol{Z}}{n}\right)\widehat{\boldsymbol{\gamma}}\right)^{-1}\widehat{\boldsymbol{\gamma}}^\top\left(\frac{\boldsymbol{Z}^\top \boldsymbol{Z}}{n}\right). \end{aligned} \tag{B16}$$

Denote $\boldsymbol{Q}_n = \boldsymbol{Z}^\top \boldsymbol{Z}/n$, consider the square of restricted eigenvalue of $\tilde{\boldsymbol{Z}}$, $K^2_{\mathscr{C}}$, we have

$$\begin{aligned} K^2_{\mathscr{C}}(\mathcal{V}^*, \xi) &= \inf_{\boldsymbol{u}}\{\|\boldsymbol{u}^\top(n^{-1}\tilde{\boldsymbol{Z}}^\top\tilde{\boldsymbol{Z}})\boldsymbol{u}/\|\boldsymbol{u}\|_2^2; \boldsymbol{u} \in \mathscr{C}(\mathcal{V}^*;\xi)\} \\ &= \inf_{\boldsymbol{u}\in\mathscr{C}(\mathcal{V}^*;\xi)}\{\|\boldsymbol{u}^\top\boldsymbol{C}_n\boldsymbol{u}/\|\boldsymbol{u}\|_2^2\} \\ &= \inf_{\boldsymbol{u}\in\mathscr{C}(\mathcal{V}^*;\xi)}\frac{\boldsymbol{u}^\top(\boldsymbol{Q}_n - \boldsymbol{Q}_n\widehat{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\gamma}}^\top\boldsymbol{Q}_n\widehat{\boldsymbol{\gamma}})^{-1}\widehat{\boldsymbol{\gamma}}^\top\boldsymbol{Q}_n)\boldsymbol{u}}{\|\boldsymbol{u}\|_2^2} \\ &= \inf_{\boldsymbol{u}\in\mathscr{C}(\mathcal{V}^*;\xi)}\frac{\boldsymbol{u}^\top\boldsymbol{Q}_n\boldsymbol{u}\widehat{\boldsymbol{\gamma}}^\top\boldsymbol{Q}_n\widehat{\boldsymbol{\gamma}}^\top - (\widehat{\boldsymbol{\gamma}}^\top\boldsymbol{Q}_n\boldsymbol{u})^2}{\|\boldsymbol{u}\|_2^2\widehat{\boldsymbol{\gamma}}^\top\boldsymbol{Q}_n\widehat{\boldsymbol{\gamma}}} \end{aligned} \tag{B17}$$

Notice the denominator $\boldsymbol{u}^\top\boldsymbol{Q}_n\boldsymbol{u}\widehat{\boldsymbol{\gamma}}^\top\boldsymbol{Q}_n\widehat{\boldsymbol{\gamma}}^\top - (\widehat{\boldsymbol{\gamma}}^\top\boldsymbol{Q}_n\boldsymbol{u})^2 \geq 0$ by Cauchy–Schwarz inequality and equality holds if and only if $\boldsymbol{u} = k\widehat{\boldsymbol{\gamma}}$ for any $k \neq 0$. Furthermore, the exact difference arising from this equation can be determined by the Lagrange's identity. One can always take $\xi \in (0, \|\widehat{\boldsymbol{\gamma}}_{\mathcal{V}^*}\|_1/\|\widehat{\boldsymbol{\gamma}}_{\mathcal{V}^{c*}}\|_1)$ such that the cone $\mathscr{C}(\mathcal{V}^*;\xi) = \{\boldsymbol{u} : \|\boldsymbol{u}_{\mathcal{V}^*}\|_1 \leq \xi\,\|\boldsymbol{u}_{\mathcal{V}^{c*}}\|_1\}$ exclude the membership of $k\widehat{\boldsymbol{\gamma}}$ because cone $\mathscr{C}(\mathcal{V}^*;\xi)$ is invariant of scale. Therefore, the denominator $\boldsymbol{u}^\top\boldsymbol{Q}_n\boldsymbol{u}\widehat{\boldsymbol{\gamma}}^\top\boldsymbol{Q}_n\widehat{\boldsymbol{\gamma}}^\top - (\widehat{\boldsymbol{\gamma}}^\top\boldsymbol{Q}_n\boldsymbol{u})^2 > 0$ and $K^2_{\mathscr{C}}(\mathcal{V}^*, \xi) > 0$ holds strictly as desire. □

Appendix B8. Proof of Lemma 3

Proof. Let $\boldsymbol{R}^* = \widetilde{\boldsymbol{Z}}^\top(\boldsymbol{Y} - \widetilde{\boldsymbol{Z}}\boldsymbol{\alpha}^*)/n$ and $\tilde{\boldsymbol{Z}} = M_{\widehat{\boldsymbol{D}}}\boldsymbol{Z}$,$\boldsymbol{D} = \boldsymbol{Z}\boldsymbol{\gamma}^* + \boldsymbol{\eta}$, we have

$$\begin{aligned} \boldsymbol{R}^* &= \widetilde{\boldsymbol{Z}}^\top(\boldsymbol{Y} - \widetilde{\boldsymbol{Z}}\boldsymbol{\alpha}^*)/n = \boldsymbol{Z}^\top M_{\widehat{\boldsymbol{D}}}(\boldsymbol{D}\beta^* + \boldsymbol{\epsilon})/n = \boldsymbol{Z}^\top\left(\boldsymbol{I} - \frac{\widehat{\boldsymbol{D}}\widehat{\boldsymbol{D}}^\top}{\widehat{\boldsymbol{D}}^\top\widehat{\boldsymbol{D}}}\right)(\boldsymbol{D}\beta^* + \boldsymbol{\epsilon})/n \\ &= \beta^*\frac{\boldsymbol{Z}^\top\boldsymbol{D}}{n} + \frac{\boldsymbol{Z}^\top\boldsymbol{\epsilon}}{n} - \frac{\frac{\boldsymbol{Z}^\top\widehat{\boldsymbol{D}}}{n}\cdot\frac{\widehat{\boldsymbol{D}}^\top(\boldsymbol{D}\beta^*+\boldsymbol{\epsilon})}{n}}{\frac{\widehat{\boldsymbol{D}}^\top\widehat{\boldsymbol{D}}}{n}} \\ &= \beta^*\frac{\boldsymbol{Z}^\top\boldsymbol{D}}{n} + \frac{\boldsymbol{Z}^\top\boldsymbol{\epsilon}}{n} - \beta^*\frac{\boldsymbol{Z}^\top\widehat{\boldsymbol{D}}}{n} - \frac{\frac{\boldsymbol{Z}^\top\widehat{\boldsymbol{D}}}{n}\cdot\frac{\widehat{\boldsymbol{D}}^\top\boldsymbol{\epsilon}}{n}}{\frac{\widehat{\boldsymbol{D}}^\top\widehat{\boldsymbol{D}}}{n}} \\ &= \frac{\boldsymbol{Z}^\top\boldsymbol{\epsilon}}{n} - \frac{\frac{\boldsymbol{Z}^\top\widehat{\boldsymbol{D}}}{n}\cdot\frac{\widehat{\boldsymbol{D}}^\top\boldsymbol{\epsilon}}{n}}{\frac{\widehat{\boldsymbol{D}}^\top\widehat{\boldsymbol{D}}}{n}}. \end{aligned} \tag{B18}$$

By direct algebra, using $\widehat{\boldsymbol{D}} = \boldsymbol{Z}\boldsymbol{\gamma}^* + P_{\boldsymbol{Z}}\boldsymbol{\eta}$ and $\widehat{\boldsymbol{D}}^\top\widehat{\boldsymbol{D}}/n = \boldsymbol{\gamma}^{*\top}\boldsymbol{Q}_n\boldsymbol{\gamma}^* + 2\boldsymbol{\eta}^\top\boldsymbol{Z}\boldsymbol{\gamma}^*/n + \boldsymbol{\eta}^\top P_{\boldsymbol{Z}}\boldsymbol{\eta}/n$, we obtain

$$\boldsymbol{R}^* = \frac{\boldsymbol{Z}^\top\boldsymbol{\epsilon}}{n} - \frac{\frac{\boldsymbol{Z}^\top\widehat{\boldsymbol{D}}}{n}\cdot\frac{\widehat{\boldsymbol{D}}^\top\boldsymbol{\epsilon}}{n}}{\frac{\widehat{\boldsymbol{D}}^\top\widehat{\boldsymbol{D}}}{n}} = \frac{\boldsymbol{Z}^\top\boldsymbol{\epsilon}}{n} - \frac{\Big(\boldsymbol{Q}_n\boldsymbol{\gamma}^* + \boldsymbol{Z}^\top\boldsymbol{\eta}/n\Big)\Big(\boldsymbol{\epsilon}^\top\boldsymbol{Z}\boldsymbol{\gamma}^*/n + \boldsymbol{\epsilon}^\top P_{\boldsymbol{Z}}\boldsymbol{\eta}/n\Big)}{\boldsymbol{\gamma}^{*\top}\boldsymbol{Q}_n\boldsymbol{\gamma}^* + 2\boldsymbol{\eta}^\top\boldsymbol{Z}\boldsymbol{\gamma}^*/n + \boldsymbol{\eta}^\top P_{\boldsymbol{Z}}\boldsymbol{\eta}/n},$$

in which the expression is free of β^*. Thus, by triangle inequality, we attain the bound

$$\|\boldsymbol{R}\|_\infty \le \left\|\frac{\boldsymbol{Z}^\top\boldsymbol{\epsilon}}{n}\right\|_\infty + \left\|\frac{\Big(\boldsymbol{Q}_n\boldsymbol{\gamma}^* + \boldsymbol{Z}^\top\boldsymbol{\eta}/n\Big)\Big(\boldsymbol{\epsilon}^\top\boldsymbol{Z}\boldsymbol{\gamma}^*/n + \boldsymbol{\epsilon}^\top P_{\boldsymbol{Z}}\boldsymbol{\eta}/n\Big)}{\boldsymbol{\gamma}^{*\top}\boldsymbol{Q}_n\boldsymbol{\gamma}^* + 2\boldsymbol{\eta}^\top\boldsymbol{Z}\boldsymbol{\gamma}^*/n + \boldsymbol{\eta}^\top P_{\boldsymbol{Z}}\boldsymbol{\eta}/n}\right\|_\infty. \tag{B19}$$

Under standard argument through concentration inequality,

$$Pr\left(\left\|\frac{\boldsymbol{Z}^\top\boldsymbol{\epsilon}}{n}\right\|_\infty \ge t\right) = Pr\Big(\max_{1\le j\le p}|\boldsymbol{Z}_j^\top\boldsymbol{\epsilon}| \ge nt\Big) \le \sum_{1\le j\le p} Pr(|\boldsymbol{Z}_j^\top\boldsymbol{\epsilon}| \ge nt) \le 2p\exp(-\frac{nt^2}{2\sigma_\epsilon^2}).$$

Let $t = \sigma_\epsilon\sqrt{\frac{2}{n}\log(2p)}$, we obtain $\|\boldsymbol{Z}^\top\boldsymbol{\epsilon}/n\|_\infty = O_p\Big(\sigma_\epsilon\sqrt{\frac{2}{n}\log(2p)}\Big)$ holds.

For the second term,

$$\left\|\frac{\Big(\boldsymbol{Q}_n\boldsymbol{\gamma}^* + \boldsymbol{Z}^\top\boldsymbol{\eta}/n\Big)\Big(\boldsymbol{\epsilon}^\top\boldsymbol{Z}\boldsymbol{\gamma}^*/n + \boldsymbol{\epsilon}^\top P_{\boldsymbol{Z}}\boldsymbol{\eta}/n\Big)}{\boldsymbol{\gamma}^{*\top}\boldsymbol{Q}_n\boldsymbol{\gamma}^* + 2\boldsymbol{\eta}^\top\boldsymbol{Z}\boldsymbol{\gamma}^*/n + \boldsymbol{\eta}^\top P_{\boldsymbol{Z}}\boldsymbol{\eta}/n}\right\|_\infty = \frac{\Big|\Big(\boldsymbol{\epsilon}^\top\boldsymbol{Z}\boldsymbol{\gamma}^*/n + \boldsymbol{\epsilon}^\top P_{\boldsymbol{Z}}\boldsymbol{\eta}/n\Big)\Big|\Big\|\boldsymbol{Q}_n\boldsymbol{\gamma}^* + \boldsymbol{Z}^\top\boldsymbol{\eta}/n\Big\|_\infty}{\Big|\boldsymbol{\gamma}^{*\top}\boldsymbol{Q}_n\boldsymbol{\gamma}^* + 2\boldsymbol{\eta}^\top\boldsymbol{Z}\boldsymbol{\gamma}^*/n + \boldsymbol{\eta}^\top P_{\boldsymbol{Z}}\boldsymbol{\eta}/n\Big|}$$

Similarly, we attain $\|\boldsymbol{Q}_n\boldsymbol{\gamma}^* + \boldsymbol{Z}^\top\boldsymbol{\eta}/n\|_\infty \le \|\boldsymbol{Q}_n\boldsymbol{\gamma}^*\|_\infty + O_p\Big(\sigma_\eta\sqrt{\frac{2}{n}\log(2p)}\Big)$.

Also repeatedly using ancillary lemmas and analysis in proof of lemma 1, we immediately have second term in (B19) is upper bounded by:

$$\begin{aligned}&\frac{\Big(\|\boldsymbol{Q}_n\boldsymbol{\gamma}^*\|_\infty + O_p\Big(\sigma_\eta\sqrt{\frac{2}{n}\log(2p)}\Big)\Big)\Big(\sigma_{\epsilon,\eta}^2 p/n + O_P(\sqrt{\boldsymbol{\gamma}^{*\top}\boldsymbol{Q}_n\boldsymbol{\gamma}^*/n} \vee \sqrt{p/n^2})\Big)}{\boldsymbol{\gamma}^{*\top}\boldsymbol{Q}_n\boldsymbol{\gamma}^* + \sigma_\eta^2 p/n + O_P(\sqrt{\boldsymbol{\gamma}^{*\top}\boldsymbol{Q}_n\boldsymbol{\gamma}^*/n} \vee \sqrt{p/n^2})}\\ =&\sigma_{\epsilon,\eta}^2 p/n\frac{\|\boldsymbol{Q}_n\boldsymbol{\gamma}^*\|_\infty}{\boldsymbol{\gamma}^{*\top}\boldsymbol{Q}_n\boldsymbol{\gamma}^* + \sigma_\eta^2 p/n} + O_p\Big(\sigma_\eta\sqrt{\frac{2}{n}\log(2p)}\Big).\end{aligned} \tag{B20}$$

Hence, together with above result, we obtain the final upper bound of $\|\boldsymbol{R}^*\|_\infty$:

$$\begin{aligned}\|\boldsymbol{R}^*\|_\infty &\le \sigma_{\epsilon,\eta}^2 p/n\frac{\|\boldsymbol{Q}_n\boldsymbol{\gamma}^*\|_\infty}{\boldsymbol{\gamma}^{*\top}\boldsymbol{Q}_n\boldsymbol{\gamma}^* + \sigma_\eta^2 p/n} + O_p\Big((\sigma_\eta + \sigma_\epsilon)\sqrt{\frac{2}{n}\log(2p)}\Big)\\ &= O_p\Big(\frac{p}{n}\cdot\frac{\|\boldsymbol{Q}_n\boldsymbol{\gamma}^*\|_\infty}{\boldsymbol{\gamma}^{*\top}\boldsymbol{Q}_n\boldsymbol{\gamma}^*} + \sqrt{\frac{\log p}{n}}\Big),\end{aligned} \tag{B21}$$

where the second line holds provided $\boldsymbol{\gamma}^{*\top}\boldsymbol{Q}_n\boldsymbol{\gamma}^*$ dominates $\sigma_{\epsilon,\eta}^2 p/n$.

Similarly, $\boldsymbol{R}^{\text{or}} = \widetilde{\boldsymbol{Z}}^\top(\boldsymbol{Y} - \widetilde{\boldsymbol{Z}}\widehat{\boldsymbol{\alpha}}^{\text{or}})/n = \widetilde{\boldsymbol{Z}}^\top\big[\boldsymbol{Y} - \tilde{\boldsymbol{Z}}_{\mathcal{V}^{c*}}(\tilde{\boldsymbol{Z}}_{\mathcal{V}^{c*}}^\top\tilde{\boldsymbol{Z}}_{\mathcal{V}^{c*}})^{-1}\tilde{\boldsymbol{Z}}_{\mathcal{V}^{c*}}^\top\boldsymbol{Y}\big]/n$. Therefore, for the sake of controlling the supremum norm of $\boldsymbol{R}^{\text{or}}$, we only need to consider the valid IV $\mathcal{V}^*$ since $\boldsymbol{R}_j^{\text{or}} = 0$ for $j \in \mathcal{V}^{c*}$. Recall $\hat{\beta}_{\text{or}}^{\text{TSLS}} = [\boldsymbol{D}^\top (P_{\boldsymbol{Z}} - P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{D}]^{-1}[\boldsymbol{D}^\top (P_{\boldsymbol{Z}} - P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{Y}]$, we have

$$\begin{aligned}
\widehat{\boldsymbol{\alpha}}_{\mathcal{V}^{c*}}^{\text{or}} &= (\boldsymbol{Z}_{\mathcal{V}^{c*}}^\top \boldsymbol{Z}_{\mathcal{V}^{c*}})^{-1}[\boldsymbol{Z}_{\mathcal{V}^{c*}}^\top (\boldsymbol{Y} - \widehat{\boldsymbol{D}}\hat{\beta}_{\text{or}}^{\text{TSLS}})] \\
&= (\boldsymbol{Z}_{\mathcal{V}^{c*}}^\top \boldsymbol{Z}_{\mathcal{V}^{c*}})^{-1}\boldsymbol{Z}_{\mathcal{V}^{c*}}^\top \boldsymbol{Y} - (\boldsymbol{Z}_{\mathcal{V}^{c*}}^\top \boldsymbol{Z}_{\mathcal{V}^{c*}})^{-1}\boldsymbol{Z}_{\mathcal{V}^{c*}}^\top \boldsymbol{D}[\boldsymbol{D}^\top (P_{\boldsymbol{Z}} - P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{D}]^{-1}[\boldsymbol{D}^\top (P_{\boldsymbol{Z}} - P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{Y}] \\
&= \boldsymbol{\alpha}_{\mathcal{V}^{c*}}^* + (\boldsymbol{Z}_{\mathcal{V}^{c*}}^\top \boldsymbol{Z}_{\mathcal{V}^{c*}})^{-1}\boldsymbol{Z}_{\mathcal{V}^{c*}}^\top \Big[\boldsymbol{I} - \boldsymbol{D}[\boldsymbol{D}^\top (P_{\boldsymbol{Z}} - P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{D}]^{-1}\boldsymbol{D}^\top (P_{\boldsymbol{Z}} - P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\Big]\boldsymbol{\epsilon}.
\end{aligned} \tag{B22}$$

Thus, for any $j \in \mathcal{V}^*$, we obtain

$$\begin{aligned}
\boldsymbol{R}_{\mathcal{V}^*}^{\text{or}} &= \tilde{\boldsymbol{Z}}_{\mathcal{V}^*}^\top \Big\{\boldsymbol{Y} - \tilde{\boldsymbol{Z}}(\widehat{\boldsymbol{\alpha}}^{\text{or}} - \boldsymbol{\alpha}^* + \boldsymbol{\alpha}^*)\Big\}/n \\
&= \tilde{\boldsymbol{Z}}_{\mathcal{V}^*}^\top (\boldsymbol{Y} - \tilde{\boldsymbol{Z}}\boldsymbol{\alpha}^*)/n + \tilde{\boldsymbol{Z}}_{\mathcal{V}^*}^\top \tilde{\boldsymbol{Z}}_{\mathcal{V}^{c*}}(\boldsymbol{\alpha}_{\mathcal{V}^{c*}}^* - \widehat{\boldsymbol{\alpha}}_{\mathcal{V}^{c*}}^{\text{or}})/n \\
&= \boldsymbol{R}_{\mathcal{V}^*}^* + \boldsymbol{Z}_{\mathcal{V}^*}^\top M_{\widehat{\boldsymbol{D}}} P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}}\Big(\boldsymbol{D} \cdot \text{Bias}(\hat{\beta}_{\text{or}}^{\text{TSLS}}) - \boldsymbol{\epsilon}\Big)/n,
\end{aligned} \tag{B23}$$

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

To explore the second term, we denote $\bar{\boldsymbol{D}} = P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}}\boldsymbol{D}$, $\bar{\boldsymbol{\epsilon}} = P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}}\boldsymbol{\epsilon}$. Due to blockwise formula for projection matrix, we have

$$P_{\boldsymbol{Z}} - P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}} = P_{M_{\boldsymbol{Z}_{\mathcal{V}^{c*}}}\boldsymbol{Z}_{\mathcal{V}^*}},$$

which is still a projection matrix. Thus, we denote $\tilde{\tilde{\boldsymbol{D}}} = P_{M_{\boldsymbol{Z}_{\mathcal{V}^{c*}}}\boldsymbol{Z}_{\mathcal{V}^*}}\boldsymbol{D}$, $\tilde{\tilde{\boldsymbol{\epsilon}}} = P_{M_{\boldsymbol{Z}_{\mathcal{V}^{c*}}}\boldsymbol{Z}_{\mathcal{V}^*}}\boldsymbol{\epsilon}$ and $\text{Bias}(\hat{\beta}_{\text{or}}^{\text{TSLS}}) = \frac{\boldsymbol{D}^\top (P_{\boldsymbol{Z}} - P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{\epsilon}}{\boldsymbol{D}^\top (P_{\boldsymbol{Z}} - P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{D}} = \frac{\tilde{\tilde{\boldsymbol{D}}}^\top \tilde{\tilde{\boldsymbol{\epsilon}}}}{\tilde{\tilde{\boldsymbol{D}}}^\top \tilde{\tilde{\boldsymbol{D}}}}$. Thus we have

$$\widehat{\boldsymbol{D}} - \bar{\boldsymbol{D}} = (P_{\boldsymbol{Z}} - P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{D} = \tilde{\tilde{\boldsymbol{D}}}. \tag{B24}$$

Therefore, the second term insides in RHS of (B23) can be reformulated as:

$$\begin{aligned}
&\boldsymbol{Z}_{\mathcal{V}^*}^\top M_{\widehat{\boldsymbol{D}}} P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}}\Big(\boldsymbol{D} \cdot \text{Bias}(\hat{\beta}_{\text{or}}^{\text{TSLS}}) - \boldsymbol{\epsilon}\Big)/n \\
=&\boldsymbol{Z}_{\mathcal{V}^*}^\top \left(\boldsymbol{I} - \frac{\widehat{\boldsymbol{D}}\widehat{\boldsymbol{D}}^\top}{\widehat{\boldsymbol{D}}^\top \widehat{\boldsymbol{D}}}\right)(\bar{\boldsymbol{D}}\,\text{Bias}(\hat{\beta}_{\text{or}}^{\text{TSLS}}) - \bar{\boldsymbol{\epsilon}})/n \\
=&\frac{(I) + (II) + (III) + (IV)}{\widehat{\boldsymbol{D}}^\top \widehat{\boldsymbol{D}}/n},
\end{aligned} \tag{B25}$$

where

$$\begin{aligned}
(I) =& \text{Bias}(\hat{\beta}_{\text{or}}^{\text{TSLS}}) \cdot \frac{\boldsymbol{Z}_{\mathcal{V}^*}^\top \widehat{\boldsymbol{D}}^\top}{n} \cdot \frac{\widehat{\boldsymbol{D}}^\top \bar{\boldsymbol{D}}}{n} = \frac{\tilde{\tilde{\boldsymbol{D}}}^\top \tilde{\tilde{\boldsymbol{\epsilon}}}}{\tilde{\tilde{\boldsymbol{D}}}^\top \tilde{\tilde{\boldsymbol{D}}}} \cdot \frac{\widehat{\boldsymbol{D}}^\top \widehat{\boldsymbol{D}}}{n} \cdot \frac{\boldsymbol{Z}_{\mathcal{V}^*}^\top \bar{\boldsymbol{D}}}{n}, \\
(II) =& -\frac{\boldsymbol{Z}_{\mathcal{V}^*}^\top \widehat{\boldsymbol{D}}^\top}{n} \cdot \frac{\widehat{\boldsymbol{D}}\bar{\boldsymbol{\epsilon}}}{n} = -\frac{\widehat{\boldsymbol{D}}^\top \widehat{\boldsymbol{D}}}{n} \cdot \frac{\boldsymbol{Z}_{\mathcal{V}^*}^\top \bar{\boldsymbol{\epsilon}}}{n}, \\
(III) =& -\text{Bias}(\hat{\beta}_{\text{or}}^{\text{TSLS}}) \cdot \frac{\boldsymbol{Z}_{\mathcal{V}^*}^\top \widehat{\boldsymbol{D}}}{n} \cdot \frac{\widehat{\boldsymbol{D}}^\top \bar{\boldsymbol{D}}}{n} = -\frac{\tilde{\tilde{\boldsymbol{D}}}^\top \tilde{\tilde{\boldsymbol{\epsilon}}}}{\tilde{\tilde{\boldsymbol{D}}}^\top \tilde{\tilde{\boldsymbol{D}}}} \cdot \frac{\boldsymbol{Z}_{\mathcal{V}^*}^\top \widehat{\boldsymbol{D}}}{n} \cdot \frac{\widehat{\boldsymbol{D}}^\top \bar{\boldsymbol{D}}}{n}, \\
(IV) =& \frac{\boldsymbol{Z}_{\mathcal{V}^*}^\top \widehat{\boldsymbol{D}}}{n} \cdot \frac{\widehat{\boldsymbol{D}}^\top \bar{\boldsymbol{\epsilon}}}{n}.
\end{aligned}$$

Now, using identity (B24), we replace $\frac{\widehat{\boldsymbol{D}}^\top \bar{\boldsymbol{D}}}{n}$ in (III) as $\frac{\widehat{\boldsymbol{D}}^\top (\widehat{\boldsymbol{D}} - \tilde{\bar{\boldsymbol{D}}})}{n}$. Therefor we obtain

$$\begin{aligned}(I)+(III) =& \frac{\tilde{\bar{\boldsymbol{D}}}^\top \tilde{\bar{\boldsymbol{\epsilon}}}}{\tilde{\bar{\boldsymbol{D}}}^\top \tilde{\bar{\boldsymbol{D}}}} \cdot \frac{\widehat{\boldsymbol{D}}^\top \widehat{\boldsymbol{D}}}{n} \cdot \frac{\boldsymbol{Z}_{\mathcal{V}^*}^\top (\bar{\boldsymbol{D}} - \widehat{\boldsymbol{D}})}{n} + \frac{\tilde{\bar{\boldsymbol{D}}}^\top \tilde{\bar{\boldsymbol{\epsilon}}}}{\tilde{\bar{\boldsymbol{D}}}^\top \tilde{\bar{\boldsymbol{D}}}} \cdot \frac{\widehat{\boldsymbol{D}}^\top \tilde{\bar{\boldsymbol{D}}}}{n} \cdot \frac{\boldsymbol{Z}_{\mathcal{V}^*}^\top \widehat{\boldsymbol{D}}}{n} \\ =& \frac{\tilde{\bar{\boldsymbol{D}}}^\top \tilde{\bar{\boldsymbol{\epsilon}}}}{\tilde{\bar{\boldsymbol{D}}}^\top \tilde{\bar{\boldsymbol{D}}}} \cdot \frac{\widehat{\boldsymbol{D}}^\top \widehat{\boldsymbol{D}}}{n} \cdot \frac{-\boldsymbol{Z}_{\mathcal{V}^*}^\top \tilde{\bar{\boldsymbol{D}}}}{n} + \frac{\tilde{\bar{\boldsymbol{D}}}^\top \tilde{\bar{\boldsymbol{\epsilon}}}}{\tilde{\bar{\boldsymbol{D}}}^\top \tilde{\bar{\boldsymbol{D}}}} \cdot \frac{\widehat{\boldsymbol{D}}^\top \tilde{\bar{\boldsymbol{D}}}}{n} \cdot \frac{\boldsymbol{Z}_{\mathcal{V}^*}^\top \widehat{\boldsymbol{D}}}{n}.\end{aligned} \tag{B26}$$

Hence, (B25) becomes

$$\begin{aligned}& \frac{(I)+(II)+(III)+(IV)}{\widehat{\boldsymbol{D}}^\top \widehat{\boldsymbol{D}}/n} \\ =& \frac{-\frac{\widehat{\boldsymbol{D}}^\top \widehat{\boldsymbol{D}}}{n}\left(\frac{\boldsymbol{Z}_{\mathcal{V}^*}^\top \tilde{\bar{\boldsymbol{D}}}}{n} \cdot \frac{\tilde{\bar{\boldsymbol{D}}}^\top \tilde{\bar{\boldsymbol{\epsilon}}}}{\tilde{\bar{\boldsymbol{D}}}^\top \tilde{\bar{\boldsymbol{D}}}} + \frac{\boldsymbol{Z}_{\mathcal{V}^*}^\top \bar{\boldsymbol{\epsilon}}}{n}\right) + \frac{\boldsymbol{Z}_{\mathcal{V}^*}^\top \widehat{\boldsymbol{D}}}{n}\left(\frac{\widehat{\boldsymbol{D}}^\top \bar{\boldsymbol{\epsilon}}}{n} + \frac{\widehat{\boldsymbol{D}}^\top \tilde{\bar{\boldsymbol{D}}}}{n} \cdot \frac{\tilde{\bar{\boldsymbol{D}}}^\top \tilde{\bar{\boldsymbol{\epsilon}}}}{\tilde{\bar{\boldsymbol{D}}}^\top \tilde{\bar{\boldsymbol{D}}}}\right)}{\frac{\widehat{\boldsymbol{D}}^\top \widehat{\boldsymbol{D}}}{n}} \\ =& -\frac{\boldsymbol{Z}_{\mathcal{V}^*}^\top}{n}(P_{\tilde{\bar{\boldsymbol{D}}}} \tilde{\bar{\boldsymbol{\epsilon}}} + \bar{\boldsymbol{\epsilon}}) + \frac{\frac{\boldsymbol{Z}_{\mathcal{V}^*}^\top \widehat{\boldsymbol{D}}}{n} \cdot \frac{\widehat{\boldsymbol{D}}^\top}{n}(P_{\tilde{\bar{\boldsymbol{D}}}} \tilde{\bar{\boldsymbol{\epsilon}}} + \bar{\boldsymbol{\epsilon}})}{\frac{\widehat{\boldsymbol{D}}^\top \widehat{\boldsymbol{D}}}{n}},\end{aligned} \tag{B27}$$

where $P_{\tilde{\bar{\boldsymbol{D}}}} = \tilde{\bar{\boldsymbol{D}}} \tilde{\bar{\boldsymbol{D}}}^\top / \tilde{\bar{\boldsymbol{D}}}^\top \tilde{\bar{\boldsymbol{D}}}$ is a projection matrix of $\tilde{\bar{\boldsymbol{D}}}$. Notice

$$\begin{aligned} P_{\tilde{\bar{\boldsymbol{D}}}} \tilde{\bar{\boldsymbol{\epsilon}}} + \bar{\boldsymbol{\epsilon}} &= \{P_{\tilde{\bar{\boldsymbol{D}}}}(P_{\boldsymbol{Z}} - P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}}) + P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}}\}\boldsymbol{\epsilon} = (P_{\tilde{\bar{\boldsymbol{D}}}} + P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{\epsilon}, \\ \widehat{\boldsymbol{D}}^\top \tilde{\bar{\boldsymbol{D}}} &= \boldsymbol{D}^\top P_{\boldsymbol{Z}}(P_{\boldsymbol{Z}} - P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{D} = \boldsymbol{D}^\top (P_{\boldsymbol{Z}} - P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{D} = \tilde{\bar{\boldsymbol{D}}}^\top \tilde{\bar{\boldsymbol{D}}}.\end{aligned} \tag{B28}$$

Thus, we are able to further simplify (B27) as

$$\begin{aligned} & -\frac{\boldsymbol{Z}_{\mathcal{V}^*}^\top}{n}(P_{\tilde{\bar{\boldsymbol{D}}}} \tilde{\bar{\boldsymbol{\epsilon}}} + \bar{\boldsymbol{\epsilon}}) + \frac{\frac{\boldsymbol{Z}_{\mathcal{V}^*}^\top \widehat{\boldsymbol{D}}}{n} \cdot \frac{\widehat{\boldsymbol{D}}^\top}{n}(P_{\tilde{\bar{\boldsymbol{D}}}} \tilde{\bar{\boldsymbol{\epsilon}}} + \bar{\boldsymbol{\epsilon}})}{\frac{\widehat{\boldsymbol{D}}^\top \widehat{\boldsymbol{D}}}{n}} \\ =& \frac{-\boldsymbol{Z}_{\mathcal{V}^*}^\top (P_{\tilde{\bar{\boldsymbol{D}}}} + P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{\epsilon}}{n} + \frac{\frac{\boldsymbol{Z}_{\mathcal{V}^*}^\top \widehat{\boldsymbol{D}}}{n} \cdot \frac{\widehat{\boldsymbol{D}}^\top}{n}(P_{\tilde{\bar{\boldsymbol{D}}}} + P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{\epsilon}}{\frac{\widehat{\boldsymbol{D}}^\top \widehat{\boldsymbol{D}}}{n}} \\ =& \frac{-\frac{\boldsymbol{Z}_{\mathcal{V}^*}^\top \tilde{\bar{\boldsymbol{D}}}}{n} \cdot \frac{\tilde{\bar{\boldsymbol{D}}}^\top \boldsymbol{\epsilon}}{n}}{\frac{\tilde{\bar{\boldsymbol{D}}}^\top \tilde{\bar{\boldsymbol{D}}}}{n}} + \frac{-\boldsymbol{Z}_{\mathcal{V}^*}^\top P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}} \boldsymbol{\epsilon}}{n} + \frac{\frac{\boldsymbol{Z}_{\mathcal{V}^*}^\top \widehat{\boldsymbol{D}}}{n}}{\frac{\widehat{\boldsymbol{D}}^\top \widehat{\boldsymbol{D}}}{n}} \cdot \left\{ \frac{\frac{\widehat{\boldsymbol{D}}^\top \tilde{\bar{\boldsymbol{D}}}}{n} \cdot \frac{\tilde{\bar{\boldsymbol{D}}}^\top \boldsymbol{\epsilon}}{n}}{\frac{\tilde{\bar{\boldsymbol{D}}}^\top \tilde{\bar{\boldsymbol{D}}}}{n}} + \frac{\widehat{\boldsymbol{D}}^\top P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}} \boldsymbol{\epsilon}}{n} \right\} \\ =& \frac{-\frac{\boldsymbol{Z}_{\mathcal{V}^*}^\top \tilde{\bar{\boldsymbol{D}}}}{n} \cdot \frac{\tilde{\bar{\boldsymbol{D}}}^\top \boldsymbol{\epsilon}}{n}}{\frac{\tilde{\bar{\boldsymbol{D}}}^\top \tilde{\bar{\boldsymbol{D}}}}{n}} + \frac{-\boldsymbol{Z}_{\mathcal{V}^*}^\top P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}} \boldsymbol{\epsilon}}{n} + \frac{\frac{\boldsymbol{Z}_{\mathcal{V}^*}^\top \widehat{\boldsymbol{D}}}{n}}{\frac{\widehat{\boldsymbol{D}}^\top \widehat{\boldsymbol{D}}}{n}} \cdot \frac{(\tilde{\bar{\boldsymbol{D}}}^\top + \hat{\boldsymbol{D}}^\top P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{\epsilon}}{n} \\ =& \frac{-\frac{\boldsymbol{Z}_{\mathcal{V}^*}^\top \tilde{\bar{\boldsymbol{D}}}}{n} \cdot \frac{\tilde{\bar{\boldsymbol{D}}}^\top \boldsymbol{\epsilon}}{n}}{\frac{\tilde{\bar{\boldsymbol{D}}}^\top \tilde{\bar{\boldsymbol{D}}}}{n}} + \frac{-\boldsymbol{Z}_{\mathcal{V}^*}^\top P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}} \boldsymbol{\epsilon}}{n} + \frac{\frac{\boldsymbol{Z}_{\mathcal{V}^*}^\top \widehat{\boldsymbol{D}}}{n} \cdot \frac{\widehat{\boldsymbol{D}}^\top \boldsymbol{\epsilon}}{n}}{\frac{\widehat{\boldsymbol{D}}^\top \widehat{\boldsymbol{D}}}{n}}\end{aligned} \tag{B29}$$

Thus, combining (B29), (B18) and (B23), we obtain

$$R^{\text{or}}_{\mathcal{V}^*} = \frac{\boldsymbol{Z}^\top_{\mathcal{V}^*}\tilde{\tilde{\boldsymbol{\epsilon}}}}{n} - \frac{\frac{\boldsymbol{Z}^\top_{\mathcal{V}^*}\tilde{\tilde{\boldsymbol{D}}}}{n} \cdot \frac{\tilde{\tilde{\boldsymbol{D}}}^\top \boldsymbol{\epsilon}}{n}}{\frac{\tilde{\tilde{\boldsymbol{D}}}^\top \tilde{\tilde{\boldsymbol{D}}}}{n}}, \tag{B30}$$

which is similar to (B18). Consequently, with analogous argument, we derive

$$\|\boldsymbol{R}^{\text{or}}\|_\infty = \|\boldsymbol{R}^{\text{or}}_{\mathcal{V}^*}\|_\infty = O_p\Big(\frac{p_{\mathcal{V}^*}}{n} \cdot \frac{\|\tilde{\tilde{\boldsymbol{Q}}}_n \boldsymbol{\gamma}^*_{\mathcal{V}^*}\|_\infty}{\boldsymbol{\gamma}^{*\top}_{\mathcal{V}^*}\tilde{\tilde{\boldsymbol{Q}}}_n \boldsymbol{\gamma}^*_{\mathcal{V}^*}} + \sqrt{\frac{\log p_{\mathcal{V}^*}}{n}}\Big), \tag{B31}$$

where $\tilde{\tilde{\boldsymbol{Q}}}_n = \boldsymbol{Z}^\top_{\mathcal{V}^*}(P_{\boldsymbol{Z}} - P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{Z}_{\mathcal{V}^*}/n$. □

Appendix B9. Proof of Theorem 3

The proof of Theorem 3 is similar to arguments in (Feng and Zhang, 2019). However, we pay more carefulness to the issue of endogeneity of $\tilde{\boldsymbol{Z}}$ and ξ, which is chosen in Lemma 2 and may be significant less than 1.

To proceed, we first prove a helpful inequality, known as basic inequality in sparse regression literature. Denote $\boldsymbol{\alpha}^0$ as $\boldsymbol{\alpha}^*$ or $\widehat{\boldsymbol{\alpha}}^{\text{or}}$, sharing the same support on $\mathcal{V}^{c*}$, and $\boldsymbol{R}^0$ is $\boldsymbol{R}^*$ or $\boldsymbol{R}^{\text{or}}$ defined in Lemma 2 upon the choice of $\boldsymbol{\alpha}^0$.

Lemma S2. *Suppose $\widehat{\boldsymbol{\alpha}}$ is a solution of (23) and denote $\boldsymbol{\Delta} = \widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^0$ and let*

$$\boldsymbol{\omega}(\boldsymbol{\alpha}) = \Big[\frac{\partial}{\partial \boldsymbol{t}} p^{MCP}_\lambda(\boldsymbol{t})|_{\boldsymbol{t}=\boldsymbol{\alpha}} - \tilde{\boldsymbol{Z}}^\top(\boldsymbol{Y} - \tilde{\boldsymbol{Z}}\boldsymbol{\alpha})/n\Big]/\lambda \tag{B32}$$

to measure the scaled violation of first order condition in (22), then

$$\boldsymbol{\Delta}^\top \boldsymbol{C}_n \boldsymbol{\Delta} \leq -\lambda \boldsymbol{\Delta}^\top \boldsymbol{\omega}(\boldsymbol{\alpha}^0) + 1/\rho\|\boldsymbol{\Delta}\|_2^2. \tag{B33}$$

Further, choosing a proper sub-derivative at origin: $\frac{\partial}{\partial t}p^{MCP}_\lambda(t)|_{t=\alpha^0_j} = \lambda sgn(\Delta_j), \forall j \in \mathcal{V}^*$,

$$\boldsymbol{\Delta}^\top \boldsymbol{C}_n \boldsymbol{\Delta} + (\lambda - \|\boldsymbol{R}_{\mathcal{V}^*}\|_\infty)\|\boldsymbol{\Delta}_{\mathcal{V}^*}\|_1 \leq -\lambda \boldsymbol{\Delta}^\top_{\mathcal{V}^{c*}}\boldsymbol{\omega}_{\mathcal{V}^{c*}}(\boldsymbol{\alpha}^0) + 1/\rho\|\boldsymbol{\Delta}\|_2^2. \tag{B34}$$

Proof. Because $\boldsymbol{\omega}(\widehat{\boldsymbol{\alpha}}) = \boldsymbol{0}$ in (22), we have $\tilde{\boldsymbol{Z}}^\top \boldsymbol{Y}/n = \frac{\partial}{\partial \boldsymbol{t}} p^{\text{MCP}}_\lambda(\boldsymbol{t})|_{\boldsymbol{t}=\widehat{\boldsymbol{\alpha}}} + \tilde{\boldsymbol{Z}}^\top \tilde{\boldsymbol{Z}}\widehat{\boldsymbol{\alpha}}/n$. Recall $\boldsymbol{C}_n = \tilde{\boldsymbol{Z}}^\top \tilde{\boldsymbol{Z}}/n$. Thus, we replace $\tilde{\boldsymbol{Z}}^\top \boldsymbol{Y}/n$ in $\boldsymbol{\omega}(\boldsymbol{\alpha}^*)$ and obtain

$$\boldsymbol{C}_n \boldsymbol{\Delta} = -\lambda \boldsymbol{\omega}(\boldsymbol{\alpha}^0) + \frac{\partial}{\partial \boldsymbol{t}} p^{\text{MCP}}_\lambda(\boldsymbol{t})|_{\boldsymbol{t}=\boldsymbol{\alpha}^0} - \frac{\partial}{\partial \boldsymbol{t}} p^{\text{MCP}}_\lambda(\boldsymbol{t})|_{\boldsymbol{t}=\widehat{\boldsymbol{\alpha}}}.$$

Multiply $\boldsymbol{\Delta}^\top$ on both sides, we have

$$\begin{aligned} \boldsymbol{\Delta}^\top \boldsymbol{C}_n \boldsymbol{\Delta} &= -\lambda \boldsymbol{\Delta}^\top \boldsymbol{\omega}(\boldsymbol{\alpha}^0) + \boldsymbol{\Delta}^\top \Big(\frac{\partial}{\partial \boldsymbol{t}} p^{\text{MCP}}_\lambda(\boldsymbol{t})|_{\boldsymbol{t}=\boldsymbol{\alpha}^0} - \frac{\partial}{\partial \boldsymbol{t}} p^{\text{MCP}}_\lambda(\boldsymbol{t})|_{\boldsymbol{t}=\widehat{\boldsymbol{\alpha}}}\Big) \\ &\leq -\lambda \boldsymbol{\Delta}^\top \boldsymbol{\omega}(\boldsymbol{\alpha}^0) + 1/\rho\|\boldsymbol{\Delta}\|_2^2, \end{aligned} \tag{B35}$$

where second line follows convexity level of MCP is up to $1/\rho$ and conclude (B33). Moreover, we further examine the terms in $\boldsymbol{\omega}(\boldsymbol{\alpha}^0)$ with respect to $\mathcal{V}^*$. We obtain,

$$\begin{aligned}\boldsymbol{\omega}_{\mathcal{V}^*}(\boldsymbol{\alpha}^0) &= \Big[\frac{\partial}{\partial \boldsymbol{t}} p_\lambda^{\mathrm{MCP}}(\boldsymbol{t})|_{\boldsymbol{t}=\boldsymbol{\alpha}^0_{\mathcal{V}^*}} - \tilde{\boldsymbol{Z}}^\top_{\mathcal{V}^*}(\boldsymbol{Y} - \tilde{\boldsymbol{Z}}\boldsymbol{\alpha}^0)/n\Big]/\lambda \\ &= \Big[\frac{\partial}{\partial \boldsymbol{t}} p_\lambda^{\mathrm{MCP}}(\boldsymbol{t})|_{\boldsymbol{t}=\boldsymbol{0}_{\mathcal{V}^*}} - \boldsymbol{R}^0_{\mathcal{V}^*}\Big]/\lambda.\end{aligned} \tag{B36}$$

Thus, we rewrite (B33) as

$$\begin{aligned}-\lambda\boldsymbol{\Delta}^\top_{\mathcal{V}^{c*}}\boldsymbol{\omega}_{\mathcal{V}^{c*}}(\boldsymbol{\alpha}^0) + 1/\rho\|\boldsymbol{\Delta}\|_2^2 &\geq \boldsymbol{\Delta}^\top\boldsymbol{C}_n\boldsymbol{\Delta} + \lambda\boldsymbol{\Delta}^\top_{\mathcal{V}^*}\boldsymbol{\omega}_{\mathcal{V}^*}(\boldsymbol{\alpha}^0) \\ &= \boldsymbol{\Delta}^\top\boldsymbol{C}_n\boldsymbol{\Delta} + \lambda\boldsymbol{\Delta}^\top_{\mathcal{V}^*}\Big[\frac{\partial}{\partial \boldsymbol{t}} p_\lambda^{\mathrm{MCP}}(\boldsymbol{t})|_{\boldsymbol{t}=\boldsymbol{0}_{\mathcal{V}^*}} - \boldsymbol{R}^0_{\mathcal{V}^*}\Big]/\lambda \\ &\geq \boldsymbol{\Delta}^\top\boldsymbol{C}_n\boldsymbol{\Delta} + \boldsymbol{\Delta}^\top_{\mathcal{V}^*}\Big[\frac{\partial}{\partial \boldsymbol{t}} p_\lambda^{\mathrm{MCP}}(\boldsymbol{t})|_{\boldsymbol{t}=\boldsymbol{0}_{\mathcal{V}^*}}\Big] - \|\boldsymbol{\Delta}_{\mathcal{V}^*}\|_1\|\boldsymbol{R}^0_{\mathcal{V}^*}\|_\infty \\ &= \boldsymbol{\Delta}^\top\boldsymbol{C}_n\boldsymbol{\Delta} + (\lambda - \|\boldsymbol{R}^0_{\mathcal{V}^*}\|_\infty)\|\boldsymbol{\Delta}_{\mathcal{V}^*}\|_1,\end{aligned} \tag{B37}$$

where the last equality holds for a proper sub-derivative at origin, i.e. for $j \in \mathcal{V}^*$, $\frac{\partial}{\partial t}p_\lambda^{\mathrm{MCP}}(t)|_{t=\alpha_j^0} = \lambda sgn(\Delta_j)$. □

Under the event $\Omega = \{\boldsymbol{\omega}_{\mathcal{V}^{c*}}(\widehat{\boldsymbol{\alpha}}^{\mathrm{or}}) = \boldsymbol{0}\}$, we move to prove the estimation error $\boldsymbol{\Delta}$ belongs to the cone $\mathscr{C}(\mathcal{V}^*;\xi)$, where ξ is chosen in Lemma 2. Recall $\mathscr{B}(\lambda,\rho) = \{\widehat{\boldsymbol{\alpha}}$ in (23) : $\lambda \geq \zeta, \rho > K_{\mathscr{C}}^{-2}(\mathcal{V}^*,\xi) \vee 1\}$ as a collection of $\widehat{\boldsymbol{\alpha}}$ computed in (23) through a broad class of MCP, in which ζ is define in (25).

LEMMA S3. *Under the event* $\Omega = \{\boldsymbol{\omega}_{\mathcal{V}^{c*}}(\widehat{\boldsymbol{\alpha}}^{or}) = \boldsymbol{0}\}$, *consider* $\widehat{\boldsymbol{\alpha}}_{\lambda_1}, \widehat{\boldsymbol{\alpha}}_{\lambda_2} \in \mathscr{B}(\lambda,\rho)$ *with different penalty level* λ_1 *and* λ_2, *and denote their estimation error as* $\boldsymbol{\Delta}_1 = \widehat{\boldsymbol{\alpha}}_{\lambda_1} - \widehat{\boldsymbol{\alpha}}^{or}$ *and* $\boldsymbol{\Delta}_2 = \widehat{\boldsymbol{\alpha}}_{\lambda_2} - \widehat{\boldsymbol{\alpha}}^{or}$, *respectively. Define* $a_1 = 1 - \|\boldsymbol{R}^{or}{}_{\mathcal{V}^*}\|_\infty/\lambda, a_2 = a_1\xi\rho/[2(\xi+1)], a_3 = a_1\xi/\,(\xi+1+a_1)$ *and* $a_0 = a_2 \wedge \{a_2a_3/(1\vee\xi)\}$. *Then, once* $\boldsymbol{\Delta}_1 \in \mathscr{C}(\mathcal{V}^*;\xi)$ *and* $\|\boldsymbol{\Delta}_1 - \boldsymbol{\Delta}_2\|_1 \leq a_0\lambda$ *hold, we conclude* $\boldsymbol{\Delta}_2 \in \mathscr{C}(\mathcal{V}^*;\xi)$.

PROOF. Let $a_1 = 1 - \|\boldsymbol{R}^{\mathrm{or}}{}_{\mathcal{V}^*}\|_\infty/\lambda > 0$. Applying Lemma S2 to $\boldsymbol{\Delta_2}$ and under event Φ, we have

$$\begin{aligned}\boldsymbol{\Delta}_2^\top\boldsymbol{C}_n\boldsymbol{\Delta}_2 + (\lambda - \|\boldsymbol{R}^{\mathrm{or}}_{\mathcal{V}^*}\|_\infty)\|\boldsymbol{\Delta}_{2\mathcal{V}^*}\|_1 \leq 1/\rho\|\boldsymbol{\Delta}_2\|_2^2 \\ \iff \boldsymbol{\Delta}_2^\top\boldsymbol{C}_n\boldsymbol{\Delta}_2 + a_1\lambda\|\boldsymbol{\Delta}_{2\mathcal{V}^*}\|_1 \leq 1/\rho\|\boldsymbol{\Delta}_2\|_2^2.\end{aligned} \tag{B38}$$

Consider the first case $\|\boldsymbol{\Delta}_1\|_1 \vee \|\boldsymbol{\Delta}_1 - \boldsymbol{\Delta}_2\|_1 \leqslant a_2\lambda$, where $a_2 = a_1\xi\rho/[2(\xi+1)]$. We have

$$\|\boldsymbol{\Delta}_2\|_2^2 \leq \|\boldsymbol{\Delta}_2\|_\infty \cdot \|\boldsymbol{\Delta}_2\|_1 \leq (\|\boldsymbol{\Delta}_2 - \boldsymbol{\Delta}_1\|_1 + \|\boldsymbol{\Delta}_1\|_1) \cdot \|\boldsymbol{\Delta}_2\|_1 \leq 2a_2\lambda\|\boldsymbol{\Delta}_2\|_1.$$

The above inequalities yield

$$a_1\lambda\|\boldsymbol{\Delta}_{2\mathcal{V}^*}\|_1 \leq 1/\rho\|\boldsymbol{\Delta}_2\|_2^2 \leq \{a_1\xi/(\xi+1)\}\left(\|\boldsymbol{\Delta}_{2\mathcal{V}^{c*}}\|_1 + \|\boldsymbol{\Delta}_{2\mathcal{V}^*}\|_1\right) \tag{B39}$$

and is equivalent to $\boldsymbol{\Delta}_2 \in \mathscr{C}(\mathcal{V}^*,\xi)$ by algebra in the first case.

Consider the second case that $\|\boldsymbol{\Delta}_1\|_1 \geq a_2\lambda$ and $\|\boldsymbol{\Delta}_1 - \boldsymbol{\Delta}_2\|_1 \leq \lambda a_2a_3/(1\vee\xi)$, where $a_3 = a_1\xi/(\xi+1+a_1)$. Similarly, applying Lemma S2 to $\boldsymbol{\Delta}_1$ and using $\boldsymbol{\Delta}_1 \in \mathscr{C}\,(\mathcal{V}^*;\xi)$, we obtain

$$a_1\lambda\|\boldsymbol{\Delta}_{1\mathcal{V}^*}\|_1 \leq 0 \quad \Rightarrow \quad \boldsymbol{\Delta}_{1\mathcal{V}^*} = \boldsymbol{0}.$$

Triangle inequalities

$$\|\boldsymbol{\Delta}_{1\mathcal{V}^*}-\boldsymbol{\Delta}_{2\mathcal{V}^*}\|_1+\|\boldsymbol{\Delta}_{2\mathcal{V}^*}\|_1\geq\|\boldsymbol{\Delta}_{1\mathcal{V}^*}\|_1,\quad \|\boldsymbol{\Delta}_{2\mathcal{V}^{c*}}-\boldsymbol{\Delta}_{1\mathcal{V}^{c*}}\|_1+\|\boldsymbol{\Delta}_{1\mathcal{V}^{c*}}\|_1\geq\|\boldsymbol{\Delta}_{2\mathcal{V}^{c*}}\|_1$$

give rise to

$$\begin{aligned}&\|\boldsymbol{\Delta}_{2\mathcal{V}^*}\|_1-\xi\|\boldsymbol{\Delta}_{2\mathcal{V}^{c*}}\|_1\\ \leq&\|\boldsymbol{\Delta}_{1\mathcal{V}^*}\|_1-\xi\|\boldsymbol{\Delta}_{1\mathcal{V}^{c*}}\|_1+(1\vee\xi)\|\boldsymbol{\Delta}_2-\boldsymbol{\Delta}_1\|_1\\ \leq&\|\boldsymbol{\Delta}_{1\mathcal{V}^*}\|_1-\xi\|\boldsymbol{\Delta}_{1\mathcal{V}^{c*}}\|_1+a_3\|\boldsymbol{\Delta}_1\|_1\\ =&(a_3-\xi)\|\boldsymbol{\Delta}_{1\mathcal{V}^{c*}}\|_1=\frac{-\xi(\xi+1)}{\xi+1+a_1}\|\boldsymbol{\Delta}_{1\mathcal{V}^{c*}}\|_1\leq 0,\end{aligned}\tag{B40}$$

where the second inequality follows the assumptions of $\|\boldsymbol{\Delta}_1\|_1$ and $\|\boldsymbol{\Delta}_1-\boldsymbol{\Delta}_2\|_1$.

Thus, together with above two cases, it conclude the $\boldsymbol{\Delta}_2\in\mathscr{C}(\mathcal{V}^*,\xi)$ when $\|\boldsymbol{\Delta}_1-\boldsymbol{\Delta}_2\|_1<a_0\lambda$, where $a_0=a_2\wedge\{a_2a_3/(1\vee\xi)\}$ □

Based on the above lemmas, we are able to derive the theoretical results stated in Theorem 1.

Proof (of Theorem 1). Consider the local solution $\widehat{\boldsymbol{\alpha}}$ in $\mathscr{B}_0(\lambda,\rho)$ and denote $\hat{\mathcal{V}}=\{j:\widehat{\alpha}_j=0\}$ and event $\Phi=\{\hat{\mathcal{V}}=\mathcal{V}^*\}$ of most interest. Thus,

$$\Pr(\Phi)=\Pr(\Phi,\Omega)+\Pr(\Phi,\Omega^c)\geq\Pr(\Phi|\Omega)\Pr(\Omega).\tag{B41}$$

Firstly, conditional on the event Ω, we denote $\boldsymbol{\Delta}=\widehat{\boldsymbol{\alpha}}-\widehat{\boldsymbol{\alpha}}^{\text{or}}$ and immediately have $\boldsymbol{\Delta}\in\mathscr{C}(\mathcal{V}^*;\xi)$ by Lemma S3. Applying Lemma S2 to $\boldsymbol{\Delta}$, we have

$$\boldsymbol{\Delta}^\top\boldsymbol{C}_n\boldsymbol{\Delta}+(\lambda-\|\boldsymbol{R}^{\text{or}}_{\mathcal{V}^*}\|_\infty)\|\boldsymbol{\Delta}_{\mathcal{V}^*}\|_1\leq-\lambda\boldsymbol{\Delta}^\top_{\mathcal{V}^{c*}}\boldsymbol{\omega}_{\mathcal{V}^{c*}}(\widehat{\boldsymbol{\alpha}}^{\text{or}})+1/\rho\|\boldsymbol{\Delta}\|_2^2.\tag{B42}$$

By Cauchy-Schwarz inequality,

$$-\lambda\boldsymbol{\Delta}^\top_{\mathcal{V}^{c*}}\boldsymbol{\omega}_{\mathcal{V}^{c*}}(\boldsymbol{\alpha}^{\text{or}})=\lambda\boldsymbol{\Delta}^\top_{\mathcal{V}^{c*}}[-\boldsymbol{\omega}_{\mathcal{V}^{c*}}(\boldsymbol{\alpha}^{\text{or}})]\leq\lambda\|\boldsymbol{\Delta}_{\mathcal{V}^{c*}}\|_2\|\boldsymbol{\omega}_{\mathcal{V}^{c*}}(\boldsymbol{\alpha}^{\text{or}})\|_2=0$$

follows the definition of Ω. Rearranging (B42) yields,

$$\begin{aligned}0&\geq\boldsymbol{\Delta}^\top\boldsymbol{Q}_n\boldsymbol{\Delta}-1/\rho\|\boldsymbol{\Delta}\|_2^2+(\lambda-\|\boldsymbol{R}^{\text{or}}_{\mathcal{V}^*}\|_\infty)\|\boldsymbol{\Delta}_{\mathcal{V}^*}\|_1\\ &\geq(K^2_{\mathscr{C}}(\mathcal{V}^*,\xi)-1/\rho)\|\boldsymbol{\Delta}\|_2^2+(\lambda-\|\boldsymbol{R}^{\text{or}}_{\mathcal{V}^*}\|_\infty)\|\boldsymbol{\Delta}_{\mathcal{V}^*}\|_1\geq 0,\end{aligned}\tag{B43}$$

where the second line follows membership of cone $\mathscr{C}(\mathcal{V}^*;\xi)$ of $\boldsymbol{\Delta}$ and the RE condition of $\tilde{\boldsymbol{Z}}$ in Lemma 2. Inequality (B43) force $\|\boldsymbol{\Delta}_{\mathcal{V}^*}\|_1=0$ and $\|\boldsymbol{\Delta}\|_2^2=0$, i.e., $\mathcal{V}^*=\hat{\mathcal{V}}$, in probability because $\|\boldsymbol{R}^{\text{or}}_{\mathcal{V}^*}\|_\infty<\lambda$ holds with probability approaching 1 in Lemma 3.

Therefore, it remains to investigate event Φ.

$$\boldsymbol{\omega}_{\mathcal{V}^{c*}}(\boldsymbol{\alpha}^{\text{or}})=\Big[\frac{\partial}{\partial\boldsymbol{t}}p^{\text{MCP}}_\lambda(\boldsymbol{t})|_{\boldsymbol{t}=\boldsymbol{\alpha}^{\text{or}}_{\mathcal{V}^{c*}}}-\tilde{\boldsymbol{Z}}^\top_{\mathcal{V}^{c*}}(\boldsymbol{Y}-\tilde{\boldsymbol{Z}}\widehat{\boldsymbol{\alpha}}^{\text{or}})\Big]=\frac{\partial}{\partial\boldsymbol{t}}p^{\text{MCP}}_\lambda(\boldsymbol{t})|_{\boldsymbol{t}=\widehat{\boldsymbol{\alpha}}^{\text{or}}_{\mathcal{V}^{c*}}}\tag{B44}$$

follows definition of $\widehat{\boldsymbol{\alpha}}^{\text{or}}$. By the definition of Φ and MCP,

$$\Phi=\{\boldsymbol{\omega}_{\mathcal{V}^{c*}}(\widehat{\boldsymbol{\alpha}}^{\text{or}})=\frac{\partial}{\partial\boldsymbol{t}}p^{\text{MCP}}_\lambda(\boldsymbol{t})|_{\boldsymbol{t}=\widehat{\boldsymbol{\alpha}}^{\text{or}}_{\mathcal{V}^{c*}}}=\boldsymbol{0}\}=\{|\widehat{\boldsymbol{\alpha}}^{\text{or}}_{\mathcal{V}^{c*}}|_{\min}>\lambda/\rho\}.$$

Rearrange the (B22) and yields

$$\left\{|\boldsymbol{\alpha}^*_{\mathcal{V}^{c*}}|_{\min} > \lambda/\rho + \|(\boldsymbol{Z}^\top_{\mathcal{V}^{c*}}\boldsymbol{Z}_{\mathcal{V}^{c*}})^{-1}\boldsymbol{Z}^\top_{\mathcal{V}^{c*}}\boldsymbol{\epsilon}\|_\infty + \left\|(\boldsymbol{Z}^\top_{\mathcal{V}^{c*}}\boldsymbol{Z}_{\mathcal{V}^{c*}})^{-1}\boldsymbol{Z}^\top_{\mathcal{V}^{c*}}\boldsymbol{D}\frac{\boldsymbol{D}^\top(P_{\boldsymbol{Z}} - P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{\epsilon}}{\boldsymbol{D}^\top(P_{\boldsymbol{Z}} - P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{D}}\right\|_\infty\right\} \subseteq \Phi. \tag{B45}$$

Thus, it suffices to examine

$$\left\|(\boldsymbol{Z}^\top_{\mathcal{V}^{c*}}\boldsymbol{Z}_{\mathcal{V}^{c*}})^{-1}\boldsymbol{Z}^\top_{\mathcal{V}^{c*}}\boldsymbol{D}\frac{\boldsymbol{D}^\top(P_{\boldsymbol{Z}} - P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{\epsilon}}{\boldsymbol{D}^\top(P_{\boldsymbol{Z}} - P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{D}}\right\|_\infty = \left|\frac{\boldsymbol{D}^\top(P_{\boldsymbol{Z}} - P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{\epsilon}}{\boldsymbol{D}^\top(P_{\boldsymbol{Z}} - P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{D}}\right| \cdot \|(\boldsymbol{Z}^\top_{\mathcal{V}^{c*}}\boldsymbol{Z}_{\mathcal{V}^{c*}})^{-1}\boldsymbol{Z}^\top_{\mathcal{V}^{c*}}\boldsymbol{D}\|_\infty.$$

The first term in RHS measure the estimation error of TSLS estimator, i.e.

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

is unvanished term under many (weak) IVs setting. While for the second term,

$$\begin{aligned}
&\|(\boldsymbol{Z}^\top_{\mathcal{V}^{c*}}\boldsymbol{Z}_{\mathcal{V}^{c*}})^{-1}\boldsymbol{Z}^\top_{\mathcal{V}^{c*}}\boldsymbol{D}\|_\infty \\
=&\|\boldsymbol{\gamma}^*_{\mathcal{V}^{c*}} + (\boldsymbol{Z}^\top_{\mathcal{V}^{c*}}\boldsymbol{Z}_{\mathcal{V}^{c*}})^{-1}\boldsymbol{Z}^\top_{\mathcal{V}^{c*}}\boldsymbol{Z}_{\mathcal{V}^*}\boldsymbol{\gamma}^*_{\mathcal{V}^*}\|_\infty + \|(\boldsymbol{Z}^\top_{\mathcal{V}^{c*}}\boldsymbol{Z}_{\mathcal{V}^{c*}})^{-1}\boldsymbol{Z}^\top_{\mathcal{V}^{c*}}\boldsymbol{\eta}\|_\infty \\
=&\|\bar{\boldsymbol{\gamma}}^*_{\mathcal{V}^{c*}}\|_\infty + O_p\Big(\sigma_\eta\sqrt{\frac{2\log(2p_{\mathcal{V}^{c*}})}{n}}\Big),
\end{aligned}$$

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

Thereby, (B45) reduces to

$$\left\{|\boldsymbol{\alpha}^*_{\mathcal{V}^{c*}}|_{\min} > \lambda/\rho + O_p\Big(\sigma_\epsilon\sqrt{\frac{2\log(2p_{\mathcal{V}^{c*}})}{n}}\Big) + |\,\text{Bias}(\hat{\beta}^{TSLS}_{or})|\cdot\Big[\|\bar{\boldsymbol{\gamma}}^*_{\mathcal{V}^{c*}}\|_\infty + O_p\Big(\sigma_\eta\sqrt{\frac{2\log(2p_{\mathcal{V}^{c*}})}{n}}\Big)\Big]\right\} \subseteq \Phi.$$

Combining with $\lambda > \zeta$, we now specify (B45) as

$$\left\{|\boldsymbol{\alpha}^*_{\mathcal{V}^{c*}}|_{\min} > C\Big[\sqrt{\frac{\log p_{\mathcal{V}^*}}{n}} + \frac{p_{\mathcal{V}^*}}{n}\cdot\frac{\|\tilde{\tilde{\boldsymbol{Q}}}_n\boldsymbol{\gamma}^*_{\mathcal{V}^*}\|_\infty}{\boldsymbol{\gamma}^{*\top}_{\mathcal{V}^*}\tilde{\tilde{\boldsymbol{Q}}}_n\boldsymbol{\gamma}^*_{\mathcal{V}^*}} + |\,\text{Bias}(\hat{\beta}^{\text{TSLS}}_{\text{or}})|\cdot\|\bar{\boldsymbol{\gamma}}^*_{\mathcal{V}^{c*}}\|_\infty\Big]\right\} \subseteq \Phi$$

Thus, under the condition event Φ holds in finite sample or in probability, we achieve consistency of identification of valid IVs. □

Appendix B10. Proof of Proposition 2

Proof.

$$T_2 = \frac{p_{\mathcal{V}^*}}{n}\cdot\frac{\|\tilde{\tilde{\boldsymbol{Q}}}_n\boldsymbol{\gamma}^*_{\mathcal{V}^*}\|_\infty}{\boldsymbol{\gamma}^{*\top}_{\mathcal{V}^*}\tilde{\tilde{\boldsymbol{Q}}}_n\boldsymbol{\gamma}^*_{\mathcal{V}^*}} \le \frac{p_{\mathcal{V}^*}}{n}\cdot\frac{\|\tilde{\tilde{\boldsymbol{Q}}}_n^{1/2}\|_\infty\|\tilde{\tilde{\boldsymbol{Q}}}_n^{1/2}\boldsymbol{\gamma}^*_{\mathcal{V}^*}\|_\infty}{\|\tilde{\tilde{\boldsymbol{Q}}}_n^{1/2}\boldsymbol{\gamma}^*_{\mathcal{V}^*}\|_2^2} \le \frac{p_{\mathcal{V}}}{n}\cdot\frac{p_{\mathcal{V}^*}\|\tilde{\tilde{\boldsymbol{Q}}}_n^{1/2}\|_\infty\|\tilde{\tilde{\boldsymbol{Q}}}_n^{1/2}\boldsymbol{\gamma}^*_{\mathcal{V}^*}\|_\infty}{\|\tilde{\tilde{\boldsymbol{Q}}}_n^{1/2}\boldsymbol{\gamma}^*_{\mathcal{V}^*}\|_1^2} \to 0$$

□

Appendix B11. Proof of Proposition 3

This proof is extended from Bun and Windmeijer (2011)'s higher order approximation arguments and with same notation.

PROOF. Recall

$$\text{Bias}(\hat{\beta}_{or}^{TSLS}) = \frac{\boldsymbol{D}^\top(P_{\boldsymbol{Z}} - P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{\epsilon}}{\boldsymbol{D}^\top(P_{\boldsymbol{Z}} - P_{\boldsymbol{Z}_{\mathcal{V}^{c*}}})\boldsymbol{D}} = \frac{\tilde{\tilde{\boldsymbol{D}}}^\top\tilde{\tilde{\boldsymbol{\epsilon}}}}{\tilde{\tilde{\boldsymbol{D}}}^\top\tilde{\tilde{\boldsymbol{D}}}} := \frac{c}{d}, \tag{B46}$$

we have $\bar{c} := E(c) = \sigma^2_{\epsilon,\eta}(p - p_{\mathcal{V}^{c*}}) = \sigma^2_{\epsilon,\eta}p_{\mathcal{V}^*}$ and $\bar{d} := E(d) = \sigma^2_\eta(\mu_n + L)$. That is free of number of invalid IVs $p_{\mathcal{V}^{c*}}$. Thereby, let $s = \max(\mu_n, p_{\mathcal{V}^*})$,

$$\begin{aligned}
\text{Bias}(\hat{\beta}_{\text{or}}^{\text{TSLS}}) &= \frac{\bar{c}}{\bar{d}} + \frac{c-\bar{c}}{\bar{d}} - \frac{\bar{c}(d-\bar{d})}{\bar{d}^2} - \frac{(c-\bar{c})(d-\bar{d})}{\bar{d}^2} + \frac{\bar{c}(d-\bar{d})^2}{\bar{d}^3} + O_p\left(s^{-\frac{3}{2}}\right) \\
E[\text{Bias}(\hat{\beta}_{\text{or}}^{\text{TSLS}})] &= \frac{\sigma^2_{\epsilon\eta}}{\sigma^2_\eta}\left(\frac{p_{\mathcal{V}^*}}{(\mu_n + p_{\mathcal{V}^*})} - \frac{2\mu_n^2}{(\mu_n + p_{\mathcal{V}^*})^3}\right) + o\left(s^{-1}\right)
\end{aligned} \tag{B47}$$

follows from Section 3 in Bun and Windmeijer (2011). □

Appendix B12. Proof of Theorem 4

PROOF. Under the conditions of Theorem 3, we have $\Pr(\hat{\mathcal{V}} = \mathcal{V}^*) \xrightarrow{p} 1$. Thus, $\hat{\beta}^{\text{WIT}} \xrightarrow{p} \hat{\beta}_{\text{or}}^{\text{liml}}$, where $\hat{\beta}_{\text{or}}^{\text{liml}}$ stands for LIML estimator estimator with known $\mathcal{V}^*$ in prior. Thus, (a) follows Corollary 1(iv) in (Kolesár et al., 2015) with $\min\text{eig}\left(\Sigma^{-1}\Lambda\right) = 0$ in their content. (b) and (c) follow (Kolesár, 2018, proposition 1). □

Appendix B13. Proof of Corollary 2

PROOF. Notice $p_{\mathcal{V}^*}/n < p/n \to 0$ and $\mu_n/n \xrightarrow{p} \mu_0$, thereby threshold $T_2 \to 0$ in Theorem 3. Likewise $T_3 \to 0$ follows $\text{Bias}(\hat{\beta}_{\text{or}}^{\text{TSLS}}) \xrightarrow{p} \frac{\sigma_{\epsilon\eta}}{\sigma^2_\eta}\left(\frac{p_{\mathcal{V}^*}}{(\mu_n+p_{\mathcal{V}^*})} - \frac{2\mu_n^2}{(\mu_n+p_{\mathcal{V}^*})^3}\right) = o(1)$. Thus, $\kappa(n)$ in (28) diminishes to 0 and Assumption 6 holds automatically. Then it follows Theorem 4. □