Endogenous treatment effect estimation with a large and mixed set of instruments and control variables*

Qingliang Fan † Yaqian Wu ‡

June 24, 2022

Abstract

Instrumental variables (IV) and control variables are frequently used to assist researchers in investigating the endogenous treatment effects. When used together, their identities are typically assumed to be known. However, in many practical situations, one is faced with a large and mixed set of covariates, some of which can serve as excluded IVs, some can serve as control variables, while others should be discarded from the model. It is often not possible to classify them based on economic theory alone. This paper proposes a data-driven method to classify a large (increasing with sample size) set of covariates into excluded IVs, controls, and noise to be discarded. The resulting IV estimator is shown to have the oracle property (to have the same first-order asymptotic distribution as the IV estimator, assuming the true classification is known).

JEL classification: C21, C26, C52, C55
Keywords: Endogeneity, model uncertainty, weak instruments, control variables, machine learning.

*Thanks to Xiaoxia Shi (the Editor) and three anonymous referees for many constructive comments. We are grateful to Mehmet Caner, Brigham Frandsen, Zijian Guo, Michal Kolesár, Wei Lin, Whitney Newey, Frank Windmeijer, Helen Zhang, Yichong Zhang, and the participants of CUHK econometrics group meetings, seminars at City University of Hong Kong, Korean Econometric Society, and Yonsei Economics Research Institute for their valuable comments. Fan acknowledges financial support from the Research Grants Council of Hong Kong, GRF-14617121 and CUHK direct grants.

†Department of Economics, The Chinese University of Hong Kong. E-mail: michaelqfan@gmail.com.
‡School of Economics, Xiamen University. E-mail: yaqian2018@stu.xmu.edu.cn.

1 Introduction

This paper considers the estimation of treatment effect with a potentially large number of covariates when the treatment is endogenous, even conditional on the covariates. The true identity of the high-dimensional covariates is often unknown to empirical researchers: some are excluded instruments, some are useful control variables (and among those, some are relevant for the treatment variable while some are not), and the rest are only noise. Thus, we are motivated to develop an IV estimator that utilizes the rich dataset while robust to weak instruments and unknown control variables.

We propose a three-step procedure for selecting the desired instruments and useful control variables using adaptive Lasso. First, we select the relevant variables in a reduced form model for the endogenous treatment variable. Second, we replace the treatment variable with its post-adaptive Lasso predicted value and select useful controls. Third, we take the selected controls and the predicted treatment variable to obtain the treatment effect estimator via OLS. Our estimator has the desired oracle property: it can consistently select the targeted instruments and controls in the first and second steps. Therefore, it is called the **R**obust **IV** **E**stimator for both the **I**rrelevant instrument and uncertain **I**ncluded controls (R2IVE). The "2" in R2IVE refers to both types (reduced form and structural equation) of model uncertainty. Monte Carlo simulations demonstrate that our estimator performs better than other existing IV estimators for endogenous treatment effects. User-friendly R code to implement our method is provided.

The most closely related study is Kang et al. (2016) and Windmeijer et al. (2019). Kang et al. (2016) propose a Lasso-type procedure ("sisVIVE") to identify the set of structural equation variables. They show that causal effects are identified and can be estimated as long as more than 50% of the covariates are excluded IVs (the "majority rule", which was first proposed by Han (2008)), without any prior knowledge about which variables are instruments or controls. Under the same identification condition, Windmeijer et al. (2019) propose a consistent median estimator that can be used for adaptive Lasso estimation, with the

resulting estimator ("ALasso") having oracle properties. However, these methods do not consider the commonly encountered weak IVs. When the majority rule is satisfied, some of the IVs might be irrelevant for the endogenous treatment variable. We extend the papers mentioned above by allowing some of the candidate instruments to be irrelevant for the treatment. Our method requires a modified sufficient condition for identification, which is the majority rule *among the relevant instruments* (i.e., more than half of the relevant IVs are excluded IVs). The majority rule allows some candidate instruments to have a direct effect on the dependent variable. As such, it imposes a weaker assumption than the ad hoc approach of choosing excluded instruments.

Furthermore, our paper extends the scope of Kang et al. (2016) and Windmeijer et al. (2019) by allowing both the number of candidate instruments and the number of relevant candidate instruments to grow with the sample size. In another related work, Guo et al. (2018) consider the presence of irrelevant instruments and propose two-stage hard thresholding (TSHT) with a voting procedure. Different from Guo et al. (2018), which select the excluded IVs from the first-step selection of relevant IVs, we consider *all candidate instruments* in both steps. Our procedure provides more robust finite sample performance in empirically relevant cases, especially when the magnitude of the first-step coefficient is relatively small.

This paper also relates to the study of IV selection with known validity and control variables, such as Donald and Newey (2001), Bai and Ng (2010), Okui (2011), Gautier and Tsybakov (2011), Belloni et al. (2012), Caner and Fan (2015), Lin et al. (2015), and Fan and Zhong (2018). The literature on selecting valid moment conditions (Cheng and Liao, 2015; Caner et al., 2018) is related in the sense that if the useful control variable is misclassified as an excluded IV, it leads to invalid moment conditions. The poor performances from 2SLS and LIML in Section 5 show the problem.

The following notations are used throughout the paper. For any $n \times L$ matrix $\mathbf{X}$, we denote the (i, j)-th element of matrix $\mathbf{X}$ as X_{ij}, the ith row as $\mathbf{X}_{i.}$, and the jth column as $\mathbf{X}_{.j}$.

$\mathbf{X}^\top$ is the transpose of $\mathbf{X}$. $\mathbf{X}_S = \mathbf{X}_{\cdot\{j:j\in S\}}$, where S is a subset of $\{1, \cdots, L\}$. $\mathcal{M}_{\mathbf{X}} = \mathbf{I}_n - \mathcal{P}_{\mathbf{X}}$, where $\mathcal{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is the projection matrix onto the column space of $\mathbf{X}$, and $\mathbf{I}_n$ is the n-dimensional identity matrix. Let $\boldsymbol{\iota}_s$ denote a $1 \times s$ vector of ones. The l_p-norm is denoted by $\|\cdot\|_p$, and the l_0-norm, $\|\cdot\|_0$, denotes the number of non-zero components of a vector. $\|\cdot\|_\infty$ denotes the maximal element of a vector. $\mathbf{1}(\cdot)$ is an indicator function that takes a value of 1 if the event is true and 0 otherwise. We denote A^c to be its complement for any set A, and $|A|$ is the cardinality of set A. For the order of magnitude symbols, $a_n \ll b_n$ ($a_n \gg b_n$) means that a_n is much less (greater) than b_n.

The rest of the article is organized as follows. After introducing the baseline model in Section 2, we describe the identification condition and estimation procedure in Section 3. Section 4 presents the theoretical results. Section 5 collects simulation results to evaluate the finite sample performance of the proposed estimator. In section 6, we illustrate the usefulness of our estimator by revisiting the trade and growth study. Section 7 concludes. Technical proofs are given in the online Appendix.

2 Model

The baseline structural model is given by

$$Y_i = D_i\beta^* + \mathbf{Z}_{i\cdot}^\top\boldsymbol{\alpha}^* + \varepsilon_i, \quad E\left(\varepsilon_i|\mathbf{Z}_{i\cdot}\right) = 0, \tag{2.1}$$

where Y_i is the outcome variable. D_i is an endogenous treatment variable, that is, $E(\varepsilon_i|D_i) \neq 0$, where ε_i is the unobserved random error. $\beta^* \in \mathbb{R}$ is the true treatment effect parameter of interest. $\mathbf{Z}_{i\cdot} \in \mathbb{R}^{L_n}$ is the L_n-dimensional vector of covariates, with $\boldsymbol{\alpha}^*$ being its true coefficients vector. We extend the scope of Kang et al. (2016) and Windmeijer et al. (2019) by allowing the dimensionality L_n to grow with n but not exceed the sample size (specifically, we require $L_n = o(n)$) and provide an inference procedure for the treatment effect. Denote $\mathbf{Y} = (Y_1, ..., Y_n)^\top$, $\mathbf{D} = (D_1, ..., D_n)^\top$, $\mathbf{Z} = (\mathbf{Z}_{1\cdot}, ..., \mathbf{Z}_{n\cdot})^\top$ and $\Sigma_n = n^{-1}\mathbf{Z}^\top\mathbf{Z}$. Note that the

model (2.1) can have known included exogenous variables $\mathbf{X}_i \in \mathbb{R}^p$ and if so, we can replace the variables Y_i, D_i and $\mathbf{Z}_{i\cdot}$ with the residuals after regressing them on $\mathbf{X}$ (e.g., replace $\mathbf{Y}$ by $\mathcal{M}_{\mathbf{X}}\mathbf{Y}$, and the same for $\mathbf{Z}$ and $\mathbf{D}$), following Zivot and Wang (1998). For simplicity, we also assume that $\mathbf{Y}$, $\mathbf{D}$, and the non-constant columns of $\mathbf{Z}$ are all centered, which can be obtained from a residual transformation with a vector of 1's, $\boldsymbol{\iota}_n$.

Definition 1. *Covariate Z_j is an excluded instrument if $\alpha_j^* = 0$, and it is a useful control variable if $\alpha_j^* \neq 0$, for $j \in \{1, \cdots, L_n\}$. Let $\mathcal{A}_C = \{j : \alpha_j^* \neq 0\}$ and $\mathcal{A}_C^c$ denote the set of useful control variables and excluded IVs, respectively, and $|\mathcal{A}_C| = s_C$. Denote $\alpha_{\min}^* = \min\left\{\left|\alpha_j^*\right| : j \in \mathcal{A}_C\right\}$, $\mathbf{Z}_C = \mathbf{Z}_{\cdot\{j:j\in\mathcal{A}_C\}}$ and $\Sigma_C = n^{-1}\mathbf{Z}_C^\top\mathbf{Z}_C$.*

Next, we consider the reduced form equation,

$$D_i = \mathbf{Z}_{i\cdot}^\top\boldsymbol{\gamma}^* + \xi_i, \quad E\left(\xi_i|\mathbf{Z}_{i\cdot}\right) = 0, \tag{2.2}$$

where ξ_i denotes i.i.d. random errors with mean 0 and finite variance, and $corr(\varepsilon_i, \xi_i) \neq 0$. This reduced form equation accommodates the optimal instrument estimation in high dimensions (Belloni et al., 2012). Since $\mathbf{Z}_{i\cdot}$ can include the nonlinear terms (such as B-splines, dummies, polynomials, and various interactions) of the original economic variables, it is without loss of generality to consider the dependence of $\mathbf{Z}_{i\cdot}$ and endogenous variables (D_i, Y_i) as potentially nonlinear[1].

Definition 2. *Instrument Z_j is a relevant IV if $\gamma_j^* \neq 0$, for $j \in \{1, \cdots, L_n\}$. Let $\mathcal{A}_R = \{j : \gamma_j^* \neq 0\}$ denote the set of these instruments that can approximate the conditional expectation of the endogenous variable and $s_R = |\mathcal{A}_R|$. Denote $\gamma_{\min}^* = \min\left\{\left|\gamma_j^*\right| : j \in \mathcal{A}_R\right\}$, $\mathbf{Z}_R = \mathbf{Z}_{\cdot\{j:j\in\mathcal{A}_R\}}$, and $\Sigma_R = n^{-1}\mathbf{Z}_R^\top\mathbf{Z}_R$.*

Remark 2.1. *In empirical studies, $\mathcal{A}_C \cap \mathcal{A}_R$ is the set of control variables that can cause omitted-variable bias, while $\mathcal{A}_C \cap \mathcal{A}_R^c$ is the set of exogenous control variables that do not cause*

[1]We thank the Editor for this point.

omitted-variable bias since it is not correlated to the treatment variable. $\mathcal{A}_C \cap \mathcal{A}_R^c$ *improves the efficiency of the IV estimator by reducing the variance in the "unobserved factors" and also makes the validity of* $\mathcal{A}_C^c \cap \mathcal{A}_R$ *more plausible in practice (Angrist and Pischke, 2009).*

Therefore, our goal is to select relevant excluded instruments in $\mathcal{A}_R$ consistently for model (2.2) and useful controls $\mathcal{A}_C$ consistently for model (2.1).

3 Identification and estimation

The models (2.1) and (2.2) imply the following moment conditions:

$$E\left(\mathbf{Z}_{i.}(Y_i - \mathbf{Z}_{i.}^\top \boldsymbol{\alpha}^* - D_i\beta^*)\right) = 0,\ E\left(\mathbf{Z}_{i.}(D_i - \mathbf{Z}_{i.}^\top \boldsymbol{\gamma}^*)\right) = 0.$$

Combining these two conditions, we obtain $\boldsymbol{\Gamma}^* = \boldsymbol{\alpha}^* + \beta^*\boldsymbol{\gamma}^*$, where $\boldsymbol{\Gamma}^* = E(\mathbf{Z}_{i.}\mathbf{Z}_{i.}^\top)^{-1}E\left(\mathbf{Z}_{i.}Y_i\right)$ and $\boldsymbol{\gamma}^* = E(\mathbf{Z}_{i.}\mathbf{Z}_{i.}^\top)^{-1}E\left(\mathbf{Z}_{i.}D_i\right)$. Since we can identify both $\boldsymbol{\Gamma}^*$ and $\boldsymbol{\gamma}^*$ based on observed data, β^* and $\boldsymbol{\alpha}^*$ are identified if and only if there is a unique solution to equation $\boldsymbol{\Gamma}^* = \boldsymbol{\alpha}^* + \beta^*\boldsymbol{\gamma}^*$ given $\boldsymbol{\Gamma}^*$ and $\boldsymbol{\gamma}^*$. When $\gamma_j^* \neq 0$ for all $j = 1, \cdots, L_n$, Kang et al. (2016) discuss a sufficient condition called the "majority rule"; that is, if $\|\boldsymbol{\alpha}^*\|_0 \leq L_n/2$, the parameters β^* and $\boldsymbol{\alpha}^*$ can always be identified. The "majority rule" was first proposed by Han (2008) and was also used in Windmeijer et al. (2019). Unlike Kang et al. (2016) and Windmeijer et al. (2019), we allow the presence of irrelevant instruments. Without loss of generality, we assume that the first s_R variables are relevant instruments. Let $\boldsymbol{\gamma}^* = (\boldsymbol{\gamma}_{(1)}^{*\top}, \mathbf{0}^\top)^\top$, $\boldsymbol{\Gamma}^* = (\boldsymbol{\Gamma}_{(1)}^{*\top}, \boldsymbol{\Gamma}_{(2)}^{*\top})^\top$ and $\boldsymbol{\alpha}^* = (\boldsymbol{\alpha}_{(1)}^{*\top}, \boldsymbol{\alpha}_{(2)}^{*\top})^\top$, where $\boldsymbol{\gamma}_{(1)}^*$, $\boldsymbol{\Gamma}_{(1)}^*$ and $\boldsymbol{\alpha}_{(1)}^*$ are all $s_R \times 1$ vectors. Then, we have

$$\boldsymbol{\Gamma}_{(1)}^* = \boldsymbol{\alpha}_{(1)}^* + \beta^*\boldsymbol{\gamma}_{(1)}^*, \tag{3.1}$$

$$\boldsymbol{\Gamma}_{(2)}^* = \boldsymbol{\alpha}_{(2)}^*. \tag{3.2}$$

The solution of $\boldsymbol{\alpha}_{(1)}^*$ and β^* is unique if the majority rule holds in the set of relevant instruments, and $\boldsymbol{\alpha}_{(2)}^*$ is unique due to the uniqueness of $\boldsymbol{\Gamma}_{(2)}^*$. Thus, we can obtain unique $\boldsymbol{\alpha}^*$ and

β^* if the majority rule holds among relevant instruments, that is, $\left\|\boldsymbol{\alpha}^*_{(1)}\right\|_0 \leq s_R/2$.

3.1 Selection of relevant IVs

To select the relevant instruments, we consider the following objective function with an adaptive Lasso penalty (Zou, 2006):

$$\ddot{\boldsymbol{\gamma}}_n = \arg\min_{\boldsymbol{\gamma}} \left\{ \|\mathbf{D} - \mathbf{Z}\boldsymbol{\gamma}\|_2^2 + \lambda_n \sum_{j=1}^{L_n} \omega_j |\gamma_j| \right\}, \tag{3.3}$$

where the adaptive weights are defined by $\omega_j = |\widetilde{\gamma}_j|^{-1}$, and $\widetilde{\boldsymbol{\gamma}}_n = (\widetilde{\gamma}_1, \cdots, \widetilde{\gamma}_{L_n})^\top$ is obtained from the least squares estimator $\widetilde{\boldsymbol{\gamma}}_n(\text{ols})$ when the dimension is much smaller than the sample size ($L_n \ll n$) and the elastic-net estimator $\widetilde{\boldsymbol{\gamma}}_n(\text{enet})$ (to enhance the robustness of the initial estimator) when L_n is relatively large but no more than n. Specifically, the elastic-net estimator as initial $\widetilde{\boldsymbol{\gamma}}_n$ is defined as

$$\widetilde{\boldsymbol{\gamma}}_n(\text{enet}) = \left\{ \arg\min_{\boldsymbol{\gamma}} \|\mathbf{D} - \mathbf{Z}\boldsymbol{\gamma}\|_2^2 + \lambda_2 \|\boldsymbol{\gamma}\|_2^2 + \lambda_1 \|\boldsymbol{\gamma}\|_1 \right\}, \tag{3.4}$$

where $\lambda_1, \lambda_2 > 0$ are the tuning parameters. The initial estimator $\widetilde{\boldsymbol{\gamma}}_n(\text{enet})$ achieves $\sqrt{n/L_n}$-consistency under some mild conditions (Zou and Zhang, 2009), satisfying the adaptive Lasso estimator's oracle property requirements. The information from the initial estimator $\widetilde{\boldsymbol{\gamma}}_n$ can improve the variable selection performance. $\widehat{\mathcal{A}}_R = \{j : |\ddot{\gamma}_j| > 0\}$ is denoted as the adaptive Lasso selected instruments. We then run the least squares of D_i on the selected IVs, and we obtain the refitted estimator $\widehat{\boldsymbol{\gamma}}_n$ and the predicted D_i:

$$\widehat{D}_i = \sum_{j \in \widehat{\mathcal{A}}_R} \widehat{\gamma}_j Z_{ij}. \tag{3.5}$$

Denote $\widehat{\mathbf{D}} = (\widehat{D}_1, \cdots, \widehat{D}_n)^\top$. We suggest the BIC (Wang et al., 2009) method to choose the tuning parameter λ_n in (3.3) adaptively in practice for transparency[2] and model the selection consistency property.

3.2 Selection of useful controls

We then describe the procedure to select the useful controls in (2.1). By taking the conditional expectation of both sides of (2.1) given $\mathbf{Z}_{i.}$, we have

$$E(Y_i|\mathbf{Z}_{i.}) = D_i^*\beta^* + \mathbf{Z}_{i.}^\top \boldsymbol{\alpha}^*, \tag{3.6}$$

where $D_i^* = E(D_i|\mathbf{Z}_{i.})$. Denote $\nu_i = Y_i - E(Y_i|\mathbf{Z}_{i.})$. It is straightforward to show that $E(\nu_i) = E[E(\nu_i|\mathbf{Z}_{i.})] = 0$ and $\text{cov}(D_i^*\nu_i) = E[E(D_i^*\nu_i|\mathbf{Z}_{i.})] = E[D_i^*E(\nu_i|\mathbf{Z}_{i.})] = 0$. Adding ν_i to both sides of (3.6), we have

$$Y_i = D_i^*\beta^* + \mathbf{Z}_{i.}^\top \boldsymbol{\alpha}^* + \nu_i. \tag{3.7}$$

Thus, D_i^* is an exogenous variable in (3.7). The coefficient of the optimal instrument D_i^* in equation (3.7) remains the same β^* as in the structural equation (2.1). Since D_i^* is not observable in practice, we replace D_i^* with its estimate $\widehat{D}_i$ in (3.5). Substituting $\widehat{D}_i$ from (3.5) into (3.7) yields

$$\mathbf{Y} = \widehat{\mathbf{D}}\beta^* + \mathbf{Z}\boldsymbol{\alpha}^* + \ddot{\boldsymbol{\nu}}. \tag{3.8}$$

We then partial out the effect of $\widehat{\mathbf{D}}$ by multiplying by $\mathcal{M}_{\widehat{\mathbf{D}}}$ on both sides of (3.8),

$$\mathcal{M}_{\widehat{\mathbf{D}}}\mathbf{Y} = \mathcal{M}_{\widehat{\mathbf{D}}}\mathbf{Z}\boldsymbol{\alpha}^* + \mathcal{M}_{\widehat{\mathbf{D}}}\ddot{\boldsymbol{\nu}} \tag{3.9}$$

[2]We provide standard R packages for easy implementation by practitioners. In our baseline model, the BIC tuning parameter choice enjoys excellent variable selection performance and is easy to implement. For the initial elastic-net estimator, we use the BIC and a fast grid search.

. Denote $\widetilde{\mathbf{Y}} = \mathcal{M}_{\widehat{\mathbf{D}}}\mathbf{Y}$, $\widetilde{\mathbf{Z}} = \mathcal{M}_{\widehat{\mathbf{D}}}\mathbf{Z}$ and $\widetilde{\boldsymbol{\nu}} = \mathcal{M}_{\widehat{\mathbf{D}}}\ddot{\boldsymbol{\nu}}$. Then equation (3.9) can be written as $\widetilde{\mathbf{Y}} = \widetilde{\mathbf{Z}}\boldsymbol{\alpha}^* + \widetilde{\boldsymbol{\nu}}$, which is a linear model with a data matrix $(\widetilde{\mathbf{Y}}, \widetilde{\mathbf{Z}})$. The adaptive Lasso estimator for $\boldsymbol{\alpha}^*$ is

$$\widehat{\boldsymbol{\alpha}}_n = \left\{ \arg\min_{\boldsymbol{\alpha}} \left\| \widetilde{\mathbf{Y}} - \widetilde{\mathbf{Z}}\boldsymbol{\alpha} \right\|_2^2 + \lambda_n' \sum_{j=1}^{L_n} \omega_j' |\alpha_j| \right\}, \tag{3.10}$$

where the adaptive weights are defined by $\omega_j' = |\widetilde{\alpha}_j|^{-1}$ and the initial estimator $\widetilde{\boldsymbol{\alpha}}_n$ is constructed by elastic-net, with tuning parameters λ_1 and λ_2 that satisfy the same conditions as their counterparts in (3.4),

$$\widetilde{\boldsymbol{\alpha}}_n(\text{enet}) = \left\{ \arg\min_{\boldsymbol{\alpha}} \left\| \mathbf{Y} - \mathbf{D}\tilde{\beta} - \mathbf{Z}\boldsymbol{\alpha} \right\|_2^2 + \lambda_2 \|\boldsymbol{\alpha}\|_2^2 + \lambda_1 \|\boldsymbol{\alpha}\|_1 \right\}. \tag{3.11}$$

The $\tilde{\beta}$ in (3.11) is the median estimator proposed by Han (2008) and Windmeijer et al. (2019). Specifically, denote $\widetilde{\boldsymbol{\Gamma}}_n(\text{enet})$ as the elastic-net estimator of the reduced form equation of $\mathbf{Y}$ on $\mathbf{Z}$, and $\widetilde{\boldsymbol{\gamma}}_{nj}(\text{enet})$ is the elastic-net estimator from (3.4). Then,

$$\widetilde{\beta} = \text{median}\left(\left\{ \frac{\widetilde{\boldsymbol{\Gamma}}_{nj}(\text{enet})}{\widetilde{\boldsymbol{\gamma}}_{nj}(\text{enet})}, j \in \widehat{\mathcal{A}}_R \right\} \right). \tag{3.12}$$

Similar to the first step, we use the BIC to choose λ_n' in (3.10). Denote $\widehat{\mathcal{A}}_C = \{j : |\widehat{\alpha}_j| > 0\}$ as the selected set of useful controls in the structural equation.

Remark 3.1. *The multicollinearity in transformed $\widetilde{\mathbf{Z}}$ does not interfere with the consistent variable selection of the proposed adaptive Lasso procedure*[3]*. Theoretically, the initial estimator $\widetilde{\boldsymbol{\alpha}}_n$ is constructed based on an identified reduced form equation. The adaptive weight constructed with this initial estimator satisfies the adaptive irrepresentable condition, which is a sufficient consistent condition for variable selection (Huang et al., 2008). We formally show this in Lemma 4.2.*

[3]The Lasso-type estimation usually results in sparsity of $\widehat{\boldsymbol{\gamma}}_n$, implying the perfect multicollinearity of $\widetilde{\mathbf{Z}}$ only concentrates on the relevant set $\widetilde{\mathbf{Z}}_R$. Even in low-dimensional models, it does not affect the selection consistency. When $L_n \ll n$, it suffices to use the OLS estimator $\widehat{\boldsymbol{\Gamma}}_n(\text{ols})$.

The oracle property of step 2 indicates that the useful controls can be selected with probability approaching 1. This property is crucial for the consistency of the R2IVE. Additionally, since we consider all variables in the candidate set in both steps, we could select $\mathcal{A}_C \cap \mathcal{A}_R^c$ and $\mathcal{A}_C \cap \mathcal{A}_R$ as controls, in contrast to the TSHT, which selects the excluded IVs from the relevant IVs in a sequential way (so that $\mathcal{A}_C \cap \mathcal{A}_R^c$ cannot be selected as controls).

Remark 3.2. *The majority rule is not directly testable. Practically, the empirical researchers could first look at the first step variable selection result and only determine whether the majority rule would hold for the selected strong IVs. Furthermore, from the research design step, the covariates set should focus on those likely to be excluded IVs. We use an empirical example in Section 6 to demonstrate what to check in practice.*

3.3 Treatment effect estimation

In the final step, we install the selected controls in the structural equation. Formally, the proposed IV estimator for the treatment effect β^*, R2IVE, is the least squares solution

$$\widehat{\beta} = \left(\widehat{\mathbf{D}}^\top \mathcal{M}_{\widehat{\mathcal{A}}_C} \widehat{\mathbf{D}}\right)^{-1} \widehat{\mathbf{D}}^\top \mathcal{M}_{\widehat{\mathcal{A}}_C} \mathbf{Y}, \tag{3.13}$$

where $\mathcal{M}_{\widehat{\mathcal{A}}_C} = \mathbf{I}_n - \mathcal{P}_{\widehat{\mathcal{A}}_C}$, and $\mathcal{P}_{\widehat{\mathcal{A}}_C} = \mathbf{Z}_{\widehat{\mathcal{A}}_C}(\mathbf{Z}_{\widehat{\mathcal{A}}_C}^\top \mathbf{Z}_{\widehat{\mathcal{A}}_C})^{-1}\mathbf{Z}_{\widehat{\mathcal{A}}_C}^\top$ is the projection matrix of $\mathbf{Z}_{\widehat{\mathcal{A}}_C}$, where $\mathbf{Z}_{\widehat{\mathcal{A}}_C} = \mathbf{Z}_{\cdot\{j:j\in\widehat{\mathcal{A}}_C\}}$.

In summary, we present the following Algorithm 1 for the estimator.

4 Main theoretical results of R2IVE

We first define the desired convergence mode of an initial estimator. We will later verify that these conditions hold for our proposed initial estimators.

Algorithm 1 **R**obust **IV** **E**stimator to both the **I**rrelevant instrument and uncertain **I**ncluded controls (R2IVE)

Step 1. Obtain the penalized estimator $\ddot{\boldsymbol{\gamma}}_n$ in (3.3) using adaptive Lasso. The post-adaptive Lasso prediction of the conditional expectation of the endogenous treatment $\widehat{D}$ in (3.5) is used in the next step.

Step 2. Compute $\widetilde{\mathbf{Y}} = \mathcal{M}_{\widehat{\mathbf{D}}}\mathbf{Y}$ and $\widetilde{\mathbf{Z}} = \mathcal{M}_{\widehat{\mathbf{D}}}\mathbf{Z}$. Obtain the penalized estimator $\widehat{\boldsymbol{\alpha}}_n$ and the useful controls set $\widehat{\mathcal{A}}_C$ in (3.10) via adaptive Lasso.

Step 3. Take the selected controls from the previous step and the predicted $\widehat{D}$ from the first step to run a least squares regression for (3.8) to obtain the resulting IV estimator of $\widehat{\beta}$ in (3.13).

Definition 3 [Initial (Estimator) Consistency]**.** *The initial estimator $\widetilde{\gamma}_j$ is r_n-consistent if*

$$r_n \max_{j \leq L_n} \left| \widetilde{\gamma}_j - \gamma_j^* \right| = O_P(1), \quad r_n \to \infty. \tag{4.1}$$

Similarly, the initial estimator $\widetilde{\alpha}_j$ is r_n-consistent if it satisfies corresponding conditions as in (4.1)*.*

Then we invoke the following conditions for theoretical study.

Assumption 1. *(C1) $\sqrt{E(D_i^2)} < \infty$, for $i = 1, \cdots, n$.*

(C2) $|\beta^*| \leq C_1$ and $\|\boldsymbol{\alpha}^*\|_2/\sqrt{n} \leq C_2$ for some constants C_1 and C_2.

(C3) γ_j^* satisfies

$$\left\{ \sum_{j \in \mathcal{A}_R} \left(\frac{1}{\left|\gamma_j^*\right|} \right)^2 \right\}^{\frac{1}{2}} \leq M_R = o\left(r_n\right), \tag{4.2}$$

where M_R is a parameter which we specify in the following (C8). Similarly, α_j^* satisfies corresponding conditions (with counterpart constants M_C) as in (4.2)*.*

(C4) The number of useful controls relevant for D_i is less than 50% of the total relevant instruments, that is, $|\mathcal{A}_C \cap \mathcal{A}_R| \leq s_R/2$, where $s_R = |\mathcal{A}_R| \geq 1$.

(C5) The errors $\{\xi_i\}_{i=1,2,\cdots,n}$ are independent and identically distributed random variables with mean zero and finite variance σ_ξ^2, and for certain constants $1 \leq d \leq 2$, $C_3 > 0$

and K, the tail probability of ξ_i satisfies $P(|\xi_i| > x) \leq K \exp\left(-C_3 x^d\right)$ for all $x \geq 0$ and $i = 1, 2, \cdots, n$. The same conditions also hold for $\{\nu_i\}_{i=1,2,\cdots,n}$.

(C6) Let $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ denote the minimum and maximum eigenvalues of a positive definite matrix $\mathbf{M}$, respectively. Then, we assume

$$b_1 \leq \lambda_{\min}(\frac{1}{n}\mathbf{Z}^\top \mathbf{Z}) \leq \lambda_{\max}(\frac{1}{n}\mathbf{Z}^\top \mathbf{Z}) \leq B_1,$$

where b_1 and B_1 are positive constants.

(C7) $\lim_{n\to\infty} \dfrac{\log(L_n)}{\log(n)} = C_4$ *for some* $0 \leq C_4 < 1$.

(C8) The parameters $\{s_R, \lambda_n, M_R, \gamma^*_{\min}\}$ satisfy $\dfrac{M_R \lambda_n}{\gamma^*_{\min} n} \to 0$ and

(a) for $1 < d \leq 2$

$$\frac{(\log s_R)^{\frac{1}{d}}}{\sqrt{n}\gamma^*_{\min}} \to 0, \quad \frac{\sqrt{n}\,(\log(L_n - s_R))^{\frac{1}{d}}}{\lambda_n r_n} \to 0; \tag{4.3}$$

(b) for $d = 1$

$$\frac{(\log n)\,(\log s_R)}{\sqrt{n}\gamma^*_{\min}} \to 0, \quad \frac{\sqrt{n}(\log n)\,(\log(L_n - s_R))}{\lambda_n r_n} \to 0, \tag{4.4}$$

and $\{s_C, \lambda'_n, M_C, \alpha^*_{\min}\}$ satisfy similar conditions.

(C9) The tuning parameters $\{\lambda_1, \lambda_2\}$ satisfy

$$\frac{\lambda_1}{\sqrt{n}} \to 0, \quad \frac{\lambda_2}{n} \to 0, \quad \frac{n}{\lambda_1\sqrt{L_n}} \to \infty, \quad \frac{\lambda_2}{\sqrt{n}}\|\boldsymbol{\gamma}^*\|_2 \to 0, \quad \frac{\lambda_2}{\sqrt{n}}\|\boldsymbol{\alpha}^*\|_2 \to 0. \tag{4.5}$$

Condition (C1) imposes a mild restriction on the finite second moment of the endogenous treatment variable D_i. Condition (C2) requires the boundedness of the true treatment effect and the non-zero coefficients for $\boldsymbol{\alpha}^*$. Condition (C3) restricts the boundedness of non-zero γ^*_j and α^*_j. Condition (C4) formally present the majority rule. $|\mathcal{A}_C \cap \mathcal{A}_R| \leq s_R/2$ is the

identification condition, which is a sufficient condition here if the reduced form is linear. And we need at least one strong excluded instrument[4]. Condition (C5) requires that the distribution of the random errors should not be too heavy tailed. Condition (C6) assumes that the eigenvalues of Σ_n are bounded below and above to ensure that the gram matrix has good behavior. It then implies that the eigenvalues of Σ_R and Σ_C are also bounded. Condition (C7) restricts the growth rate of the number of parameters, following the literature on moderately high dimension case (Fan and Peng, 2004; Zou and Zhang, 2009). Condition (C8) puts restrictions on the numbers of covariates with non-zero coefficients s_R (s_C), the penalty parameter λ_n (λ_n'), and the minimum non-zero coefficient $\gamma^*_{\min}$ ($\alpha^*_{\min}$), which can imply that $\gamma^*_{\min}$ ($\alpha^*_{\min}$) are uniformly bounded away from zero over $j \in \mathcal{A}_R$ ($j \in \mathcal{A}_C$) and n. The maximum number of covariates permitted depends on the tail behavior of the error term. Heuristically, for $d = 2$ (sub-Gaussian tail), the model can include more covariates than the exponential tail case ($d = 1$). We can use some special cases to see the growth rate. For example, we consider $d = 2$, and $r_n = \sqrt{n/L_n}$ (the rate of our initial estimator). Assume that $1/\gamma^*_{\min} = O(1)$, $M_R = O\left(s_R^{1/2}\right)$ and $\lambda_n = n^a$ for some $0 < a < 1$, then the above conditions can be simplified as

$$\frac{\sqrt{s_R}}{n^{1-a}} \to 0, \quad \frac{\log s_R}{n} \to 0, \quad \frac{\sqrt{L_n \log(L_n - s_R)}}{n^a} \to 0.$$

Thus, we have $\lambda_n = o_p(n)$, $s_R = o_p(\min\{n^{2(1-a)}, n/L_n\})$ and $L_n - s_R = o_p(\exp(n^{2a}/L_n))$. Condition (C9) follows the assumptions (A5) and (A6) in Zou and Zhang (2009), which regulates the tuning parameters λ_1, λ_2. It allows the non-zero coefficients to vanish but at a rate that can be distinguished by the penalized least squares.

Remark 4.1. *We specify the convergence rate r_n of the initial estimator in the relevant IV*

[4]The high-dimensionality of the model provides a more plausible scenario that our assumption on the IV set could hold. Even if we have all weak IVs (or many weak controls), we show in the simulations that as long as there is model uncertainty in both structural and reduced form models, our method is still more robust than the method that could deal with only one type of undesired instrument (such as LIML for many weak instruments case).

selection. For the low dimension $n \gg L_n$ case, under conditions (C5) and (C6), the OLS initial estimator $\widetilde{\boldsymbol{\gamma}}_n(ols)$ is known to be $\sqrt{n}$-consistent. In high-dimensional models, under conditions (C5)-(C7), and (C9), the elastic-net estimator $\widetilde{\boldsymbol{\gamma}}_n(enet)$, proven by Zou and Zhang (2009) in Theorem 3.1, has the rate

$$E\left(\|\widetilde{\boldsymbol{\gamma}}_n\left(enet\right)-\boldsymbol{\gamma}^*\|_2^2\right) \leq 4\frac{\lambda_2^2\left\|\boldsymbol{\gamma}^*\right\|_2^2+BL_n n\sigma_\xi^2+\lambda_1^2 L_n}{\left(bn+\lambda_2\right)^2}=O\left(\frac{L_n}{n}\right).$$

Hence, $\widetilde{\boldsymbol{\gamma}}_n(enet)$ is a $\sqrt{n/L_n}$-consistent estimator.

Lemma 4.1. *Assume that condition (C3), (C8) and other conditions in Remark 4.1 hold, then*

$$P\left(\widehat{\mathcal{A}}_R=\mathcal{A}_R\right)\to 1, \text{ as } n\to\infty, \tag{4.6}$$

$$\left\|\mathbf{D}^*-\widehat{\mathbf{D}}\right\|_2=o_p(1). \tag{4.7}$$

The equation (4.6) shows the selection consistency of the adaptive Lasso for the high-dimensional reduced form model. Based on this selection consistency, we give the proof of equation (4.7) in the online Appendix.

Lemma 4.2. *Assume conditions (C2), (C4)-(C7) and (C9) hold, the median estimator $\widetilde{\beta}$ in* (3.12) *is consistent:*

$$\widetilde{\beta}\xrightarrow{p}\beta^*, \tag{4.8}$$

and the constructed initial estimator $\widetilde{\boldsymbol{\alpha}}_n(enet)$ in the second step of R2IVE is consistent:

$$\sqrt{n/L_n}\|\widetilde{\boldsymbol{\alpha}}_n(enet)-\boldsymbol{\alpha}^*\|_2=O_p(1). \tag{4.9}$$

Furthermore, based on this initial estimator, assume conditions (C1), (C2), (C3) and (C8) hold,

$$P\left(\widehat{\mathcal{A}}_C=\mathcal{A}_C\right)\to 1, \text{ as } n\to\infty. \tag{4.10}$$

Lemma 4.2 describes the behavior of the proposed median estimator $\widetilde{\beta}$ and initial estimator $\widetilde{\boldsymbol{\alpha}}_n$. The median estimator $\widetilde{\beta}$ is consistent, and the initial estimator $\widetilde{\boldsymbol{\alpha}}_n$ satisfies consistency in Definition 3. It suffices to provide control variable selection consistency of the adaptive Lasso estimator in the second step. Hence, our oracle property is more general than models that consider only $\mathcal{A}_R$ or $\mathcal{A}_C$.

Remark 4.2. *Note that despite its desired oracle property, the current procedure has a deficiency regarding uniform inference. When the coefficient of some variables is non-zero but small, the machine learning procedure may not select them and thus may lead to incorrect inference in finite samples. This is different from the double machine learning (DML) procedure of Belloni et al. (2014) and Chernozhukov et al. (2018), which gives uniform inference with an unknown functional form of the model. This paper focuses on the first-order problem of IV and control variable model uncertainty. In this regard, our study is rather complementary to the DML method rather than competing with it. The simulation results show that R2IVE is robust to the situation of many small coefficients; second, DML with the R2IVE-selected model works well. We leave the uniform inference problem for future studies.*

Theorem 4.1. *Assume Assumption 1 holds, the R2IVE in* (3.13) *is* $\sqrt{n}$*-consistent and asymptotically normal. That is*

$$\sigma_n^{-1}\sqrt{n}\left(\widehat{\beta}-\beta^*\right)\to N(0,1), \tag{4.11}$$

where $\sigma_n^2 = \left[E\left(\mathbf{D}^{*\top}\mathcal{M}_{\mathcal{A}_C}\mathbf{D}^*\right)\right]^{-1} E\left[\mathbf{D}^{*\top}\mathcal{M}_{\mathcal{A}_C}\mathbf{D}^*\nu_i^2\right]\left[E\left(\mathbf{D}^{*\top}\mathcal{M}_{\mathcal{A}_C}\mathbf{D}^*\right)\right]^{-1}$. *In the case in which the structural error is homoscedastic, that is,* $E(\nu_i^2|\mathbf{Z}_{i.}) = \sigma_\nu^2$, (4.11) *holds with* $\sigma_n^2 = \sigma_\nu^2\left[E\left(\mathbf{D}^{*\top}\mathcal{M}_{\mathcal{A}_C}\mathbf{D}^*\right)\right]^{-1}$.

Unlike the previous literature, adding controls $\mathcal{A}_C \cap \mathcal{A}_R$ in the first step and $\mathcal{A}_C \cap \mathcal{A}_R^c$ in the second step to our procedure improves efficiency by effectively reducing the space of the

error term. For statistical inference, the asymptotic variance can be estimated by

$$\widehat{\sigma}_n^2 = \left(\frac{1}{n}\sum_{i,j=1}^{n} \widehat{D}_i \mathcal{M}_{\widehat{\mathcal{A}}_C(i,j)} \widehat{D}_j\right)^{-1} \left(\frac{1}{n}\sum_{i,j=1}^{n} \widehat{D}_i \mathcal{M}_{\widehat{\mathcal{A}}_C(i,j)} \widehat{D}_j \widehat{\nu}_i^2\right) \left(\frac{1}{n}\sum_{i,j=1}^{n} \widehat{D}_i \mathcal{M}_{\widehat{\mathcal{A}}_C(i,j)} \widehat{D}_j\right)^{-1},$$

where $\widehat{\nu}_i = Y_i - \widehat{D}_i\widehat{\beta} - \mathbf{Z}_{i\cdot}^\top \widehat{\boldsymbol{\alpha}}$. In addition, $\widehat{\sigma}_n^2 = \left(\frac{1}{n}\sum_{i,j=1}^{n} \widehat{D}_i \mathcal{M}_{\widehat{\mathcal{A}}_C(i,j)} \widehat{D}_j\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \widehat{\nu}_i^2\right)$ under the homoscedastic structural errors case. Then the asymptotic confidence interval for the true treatment effect β^* can be obtained by $(\widehat{\beta} - z_{\tau/2}\widehat{\sigma}_n/\sqrt{n},\ \widehat{\beta} + z_{\tau/2}\widehat{\sigma}_n/\sqrt{n})$, where $z_{\tau/2}$ denotes the $\tau/2$ upper-tailed critical value of the standard normal distribution.

5 Simulation

We conduct various simulation studies to evaluate the performance of the proposed method in finite samples. Specifically, the structural equation is

$$Y_i = D_i\beta^* + \mathbf{Z}_{i\cdot}^\top \boldsymbol{\alpha}^* + \varepsilon_i, \tag{5.1}$$

where $\beta^* = 0.75$. $\mathbf{Z}_{i\cdot} = (Z_{i1}, Z_{i2}, \cdots, Z_{iL_n})^\top$ is generated from a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$, and $\boldsymbol{\Sigma} = (\rho_{j_1 j_2})_{L_n \times L_n}$ with $\rho_{j_1 j_2} = 0.5^{|j_1 - j_2|}$, for $j_1, j_2 = 1, \cdots, L_n$, and $i = 1, \cdots, n$. We consider two different coefficients generating patterns for $\boldsymbol{\alpha}^*$. One is "cutoff at s_C" sparse design, that is, $\boldsymbol{\alpha}^* = c \times (\mathbf{0}_q, \boldsymbol{\iota}_{s_C}, \mathbf{0}_{L_n - q - s_C})^\top$, where $\boldsymbol{\iota}_{s_C}$ is a $1 \times s_C$ vector of 1's, which denotes that the first q and the last $L_n - q - s_C$ covariates are excluded IV, and s_C is the number of useful controls. The default value for the constant c is 1 until otherwise noted. The other is an "exponentially decaying" design, that is, $\boldsymbol{\alpha}^* = 0.5 * (\mathbf{0}_{L_n - s_C}, 1, 0.7, 0, 7^2, \cdots, 0.7^{s_C - 1})^\top$, which denotes that the first $L_n - s_C$ covariates are excluded IVs, the last s_C covariates are controls and the corresponding coefficients of the controls decrease to zero. In the decaying parameter case, we allow that most $\mathbf{Z}_i$s are controls (a violation of the majority rule), which are designed to test the robustness of the R2IVE in the more difficult situation to distinguish useful controls.

The endogenous variable D_i is generated based on the following reduced form model:

$$D_i = \mathbf{Z}_{i\cdot}^{\top}\boldsymbol{\gamma}^* + \xi_i. \tag{5.2}$$

For model (5.2), we consider three different coefficient patterns. The first is "cutoff at s_R" sparse design, that is, $\boldsymbol{\gamma}^* = (2, 0.75, 1.5, 1, \ldots, \mathbf{0}_{L_n - s_R})^{\top}$, where $\mathbf{0}_{L_n - s_R}$ is a $1 \times (L_n - s_R)$ vector of zeros and s_R is the number of relevant IVs. We fill in the values of non-zero elements in $\boldsymbol{\gamma}^*$ by replicating the non-zero elements of $(2, 0.75, 1.5, 1)$ until its length is s_R. For example, if $s_R = 6$, the non-zero elements of $\boldsymbol{\gamma}^*$ are $(2, 0.75, 1.5, 1, 2, 0.75)$. We also consider the fixed value of non-zero elements in $\boldsymbol{\gamma}^*$ and vary the magnitude of the coefficients in subsection 5.5. The second is the "many weak" design, and the corresponding coefficients are $\boldsymbol{\gamma}^* = (\frac{\tau}{\sqrt{n}}, \cdots, \frac{\tau}{\sqrt{n}})^{\top}$, where $\tau = 1.41$. The third is an "exponentially decaying" design, that is, $\boldsymbol{\gamma}^* = 0.5 * (1, 0.7, 0, 7^2, \cdots, 0.7^{L_n - 1})^{\top}$, which denotes that the reduced form model is non-sparse and has some weak instruments as the power series approaches 0.

In summary, we vary (i) the number of useful controls s_C, (ii) the number of relevant IVs s_R, (iii) the size of $|\mathcal{A}_C^c \cap \mathcal{A}_R|$ and $|\mathcal{A}_C \cap \mathcal{A}_R^c|$, which is controlled by the value of q in the strict cutoff designs, (iv) the sample size n and IV dimensionality L_n, (v) the strength of IVs and (vi) the coefficient patterns introduced above. The model settings are also summarized in Table 1.

[Insert Table 1 here]

The error terms in the structural model and reduced form models are generated by

$$\begin{pmatrix} \varepsilon_i \\ \xi_i \end{pmatrix} \overset{\text{i.i.d.}}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix} \right).$$

We repeat each simulation setting $M = 500$ times and compute the average of the estimation bias (denoted by "bias"), $M^{-1}\sum_{m=1}^{M}\left(\widehat{\beta}^m - \beta^*\right)$, with its empirical standard deviation and the mean squared errors (denoted by "MSE"), $M^{-1}\sum_{m=1}^{M}\left(\widehat{\beta}^m - \beta^*\right)^2$, where

$\widehat{\beta}^m$ denotes an estimator of β^* in the mth experiment. We compare our method with OLS, 2SLS, Oracle 2SLS (the linear 2SLS when the relevant set $\mathcal{A}_R$ and controls set $\mathcal{A}_C$ are all known), LIML, RJIVE, NAIVE, sisVIV, ALasso, TSHT, and median estimators. For 2SLS, LIML, RJIVE, and NAIVE, we report the estimation results using the endogenous treatment variable in the structural equation and all instrumental variables in the reduced form model. We also present results for the so-called post-sisVIVE estimator (to compare with other post-selection estimators), which is a 2SLS estimator that takes the set of controls selected by sisVIVE. Post-sisVIVE also provides standard error in an empirical study in Section 6. We report the model selection performance for the estimators that can perform variable selection. Specifically, we report the average number ("mean") of covariates selected as relevant IV (for NAIVE,TSHT, and R2IVE), excluded IV (for TSHT), or useful control (for sisVIVE, post-sisVIVE, ALasso, and R2IVE) together with the minimum, median and maximum numbers of selected covariates, and the rate of successfully recognized covariates ("freq")[5]. Finally, we report the result for a double machine learning (DML) estimator (Chernozhukov et al., 2015) using post Lasso with control variables selection by R2IVE. DML variable selection is not reported since the model selection result is not directly comparable.

All simulation studies are conducted using R. In particular, the R package `naivereg` (Fan and Zhong, 2018) and `gcdnet` (Yang and Zou, 2012) are used in R2IVE. The R package `sisVIVE` (Kang et al., 2016) and the R code provided by Guo et al. (2018) are used to obtain sisVIVE and TSHT, respectively. The computer codes for the implementation of the R2IVE are available at `https://github.com/microfan1/R2IVE`.

5.1 A brief description of sisVIVE, ALasso, and TSHT

Here, we briefly describe sisVIVE, ALasso and TSHT, the most relevant comparison estimators.

[5]The formula for this indicator is freq=$\frac{\sum_{m=1}^{M}\mathbf{1}(S_i)}{M}$, where S_i is the event in ith loop that the selected covariates include all the intended true set of instruments or controls.

sisVIVE is numerically equivalent to a two-step estimation procedure. The preparation step is to obtain the predicted value of $\widehat{D}$ using all IVs. Step 1 is the standard Lasso problem, which estimates the structural equation coefficients $\widehat{\boldsymbol{\alpha}}_\lambda$ based on a set of transformed variables similar to (3.9). Then, the causal effect parameter β is obtained directly using the estimates from step 1, $\widehat{\beta}_{\text{sisVIVE}} = \frac{\widehat{\mathbf{D}}^\top(\mathbf{Y}-\mathbf{Z}\widehat{\boldsymbol{\alpha}}_\lambda)}{\|\widehat{\mathbf{D}}\|_2^2}$.

ALasso uses OLS estimators to calculate median estimator $\widetilde{\beta}$ similar to (3.12); that is, $\widetilde{\boldsymbol{\Gamma}}$ and $\widetilde{\boldsymbol{\gamma}}$ are from OLS using all IVs. Then the initial estimator $\widetilde{\boldsymbol{\alpha}}$ is calculated by $\widetilde{\boldsymbol{\alpha}} = \widetilde{\boldsymbol{\Gamma}} - \widetilde{\boldsymbol{\gamma}}\widetilde{\beta}$ which is used in adaptive Lasso to obtain $\widehat{\boldsymbol{\alpha}}$. At last, β is obtained similar to sisVIVE, $\widehat{\beta}_{\text{ALasso}} = \frac{\widehat{\mathbf{D}}^\top(\mathbf{Y}-\mathbf{Z}\widehat{\boldsymbol{\alpha}})}{\|\widehat{\mathbf{D}}\|_2^2}$.

TSHT targets the set $\mathcal{A}_C^c \cap \mathcal{A}_R$. The first thresholding step selects $\mathcal{A}_R$, and the second thresholding step selects $\mathcal{A}_C^c$ based on the selected strong instruments. The voting procedure takes the candidate sets and uses majority and plurality rules to determine the true set of excluded IVs. Finally, a two-stage least squares estimator with the selected valid IV set gives a point estimate for the causal effect parameter.

5.2 Change the number of useful controls

In the first case, we fix the sample size $n = 200$, IV dimension $L_n = 100$, and the number of relevant IVs $s_R = 10$ and $c = 1$. We set $s_C = 0, 10, 30$ to check the influence of the number of useful controls on the estimation results. For $s_C = 0$, we have $|\mathcal{A}_C \cap \mathcal{A}_R| = |\mathcal{A}_C \cap \mathcal{A}_R^c| = 0$, $|\mathcal{A}_C^c \cap \mathcal{A}_R| = 10$ and $|\mathcal{A}_C^c \cap \mathcal{A}_R^c| = 90$. For other non-zero values of s_C, we set $q = 7$, which denotes $|\mathcal{A}_C^c \cap \mathcal{A}_R| = 7$, $|\mathcal{A}_C \cap \mathcal{A}_R| = 3$, $|\mathcal{A}_C^c \cap \mathcal{A}_R^c| = 100 - s_C - 7$ and $|\mathcal{A}_C \cap \mathcal{A}_R^c| = s_C - 3$. All settings here satisfy the majority rule. The results are shown in Table 2 and Figures 1(a) and 1(b).

[Insert Table 2, Figure 1 here]

Due to space considerations, Figures 1(a) and 1(b) only show the box plots of bias for these estimators for $s_C = 0, 30$. Table 2 reports the estimation results for $s_C = 0, 10, 30$.

In Figures 1(a) and 1(b), we do not include the results of OLS in both panels since they are always severely biased and have very large MSE, which inflates the scale of the figure and disturbs the visual demonstration of the main estimator. For the same reason, LIML is excluded from Figure 1(b) when it mistakenly uses the controls as excluded IVs (e.g., in Table 2, $s_C = 30$, the MSE of LIML is 7.618, which is much larger than the second-largest MSE, post-sisVIVE, which is 0.1493).

When there are no useful controls ($s_C = 0$), 2SLS, sisVIVE, and ALasso are outperformed by LIML, RJIVE, NAIVE, TSHT, and R2IVE due to the presence of many irrelevant instruments. This is shown in Figure 1(a). R2IVE is almost as good as LIML and RJIVE. We demonstrate the importance of control variable selection, e.g., when $s_C = 10, 30$, 2SLS, LIML, RJIVE, and NAIVE become severely biased when they confuse the true identities of controls and excluded IVs. The sisVIVE and post-sisVIVE are also substantially biased in models with irrelevant IVs. The reason is that sisVIVE tends to select too many controls (and sometimes missing the true controls), which causes bias. Post-sisVIVE does not mitigate this problem. The post-selection 2SLS using many wrong variables from sisVIVE (specifically, with many weak instruments that are regarded as excluded IV) aggravates the bias problem in our baseline simulation settings. The median estimator proposed in Windmeijer et al. (2019) is severely biased. Since the ALasso relies on this initial estimator, the performance of ALasso is hence negatively affected. As a result, the ALasso is also biased due to the biased initial estimator and not selecting all control variables correctly. TSHT tends to select too many excluded instruments, which means that some control variables were wrongly selected as excluded IVs. Moreover, the simulation results show the cost of ignoring the other useful covariates (namely, $\mathcal{A}_C \cap \mathcal{A}_R$ in step 1 and $\mathcal{A}_C \cap \mathcal{A}_R^c$ in step 2) in the finite sample. Specifically, in Table 2, Panel 2, where $|\mathcal{A}_C \cap \mathcal{A}_R^c| = 7$, the MSE of TSHT is much larger than that of R2IVE. The median estimator proposed in (3.12) is best in methods other than R2IVE and DML. R2IVE is very close to oracle 2SLS in linear reduced form models and performs the best among the non-oracle estimators. It can select

relevant and excluded IVs and utilize the useful IV sets in different estimation stages. Based on the R2IVE-selected controls, DML is the second-best estimator following R2IVE. DML has a notably larger bias that diminishes as the sample size increases, as shown in Figures 1(e)-1(g).

5.3 Change the number of relevant IVs

In the second case, we fix the sample size $n = 200$, IV dimension $L_n = 100$, and the number of useful controls $s_C = 30$ and $c = 1$ but change the number of relevant IVs $s_R = 4, 15, 20$. For each of the aforementioned s_R, we set the number of strong and excluded instruments $q = 3, 10, 14$, respectively. The majority rule is satisfied in this setting. The results are shown in Table 3 and Figures 1(c) and 1(d).

[Insert Table 3 here]

In Figures 1(c) and 1(d), we see that 2SLS, RJIVE, and NAIVE all have large biases and MSEs. LIML (very large MSE) is excluded from Figures 1(c) and 1(d). When the number of relevant IVs increases, these estimators perform better. With irrelevant instruments in the model, e.g., $s_R = 4$, sisVIVE is even outperformed by RJIVE and NAIVE. The post-sisVIVE does not help to reduce the bias. The sisVIVE is still biased when $s_R = 20$. This shows the importance of selecting relevant IVs. The median estimator used in ALasso is most biased other than LIML, and hence ALasso has a large bias. TSHT improves with s_R but has a larger MSE in finite samples than R2IVE. The simulation results show that R2IVE has better finite sample performance than DML in our baseline model setting in terms of bias and MSE. The median estimator in this paper has comparable performance to DML.

5.4 Change the sample size n, IV dimensionality L_n and the magnitude of control variable coefficients

In this case, we increase the sample size and IV dimensionality and simultaneously change the magnitude of $\boldsymbol{\alpha}^*$ (controlled by c), the control variable coefficients. We first fix the IV dimensionality at $L_n = 100$, and $s_R = 20$, $s_C = 20$ and $q = 14$ while changing the sample size (and the magnitude of $\boldsymbol{\alpha}^*$) to $n = 200$ ($c = 1$), 500 ($c = 0.75$), 1000 ($c = 0.5$). The results are shown in Table 4 and Figures 1(e)-1(g). The estimation performances of OLS, 2SLS, LIML, RJIVE, and NAIVE all improve with larger sample sizes, but they are always biased due to misclassifying the useful control as an excluded IV. sisVIVE and DML have diminishing bias and MSE when the sample size increases. sisVIVE tends to over select controls and has a relatively large bias compared to other selection-based estimators in different sample sizes. ALasso can not correctly select all control variables with a poorly performing median estimator. The R2IVE is quite robust to changes in c or n. TSHT's performance in sample size $n = 200$ is not very satisfactory. It improves significantly and is on par with R2IVE when the sample size $n = 500, 1000$. R2IVE is the best performer in different sample sizes.

Then, we fix $n = 500$, $s_R = 20$, $s_C = 20$ and $q = 14$ while letting $L_n = 100$ ($c = 1$), 250 ($c = 0.5$). The results are shown in Table 4 and Figures 1(f) and 1(h). As the IV dimensionality grows while the magnitude of non-zero α is smaller (e.g., Panel 4 of Table 4), the performance of TSHT becomes poor. R2IVE and DML are quite robust to the IV dimensionality when the sample size is large.

[Insert Table 4 here]

5.5 Change the IV strength

As discussed in Section 3, if a variable is irrelevant for D_i and is a useful control, the IV relevancy and control eligibility are evaluated in two separate steps of the procedure; therefore, the ratio of α_j and γ_j (in the cutoff design) does not directly affect the variable

selection results (in contrast to a sequential selection method such as TSHT). In this case, we change the magnitude of non-zero coefficients in $\boldsymbol{\gamma}^*$. We first fix the IV dimensionality at $n = 200$, $L_n = 100$, $s_R = 20$, $s_C = 20$, $q = 14$ and $c = 1$ while changing the magnitude of non-zero coefficients in $\boldsymbol{\gamma}^* = 1$, 0.75, and 0.5, respectively, e.g., $\boldsymbol{\gamma}^* = (1, \cdots, 1, 0, \cdots, 0)^\top$. The results are shown in Table 5 and Figures 2(a) and 2(b). The decrease in values of $\boldsymbol{\gamma}^*$ generally worsens the performance of all estimators, while R2IVE and DML always dominate other estimators. When $\gamma = 0.5$, the performance of DML is slightly better than R2IVE.

[Insert Table 5, Figure 2 here]

5.6 Change the IV model design patterns

Now, we consider some different coefficient patterns for IV models. Specifically, we use the conventional "many weak IVs" setting and the "exponentially decaying" design regarding 1) the IV strength for (5.2) only and 2) both IV strength for (5.1) and control variable coefficients in (5.2). Due to space limitations, we fix $n = 200$ and $L_n = 100$. The simulation results are shown in Tables 6-8 and Figures 2(c)-2(h). First, in the "many weak IVs" design, we consider two cases of $s_C = 0, 10$. When the true model has no useful controls ($s_C = 0$), LIML and RJIVE are the best estimators. R2IVE is the best after these two, as shown in Panel 1 of Table 6. However, there is a large bias to LIML and RJIVE when they do not distinguish useful controls from excluded IVs, as shown in Panel 2. sisVIVE and ALasso can correctly select all control variables but are still biased. TSHT always has a large bias and MSE due to undesirable variable selection performance. It tends to select fewer relevant instruments than NAIVE. R2IVE and DML have the smallest bias and MSE, and R2IVE is most robust to the many weak IVs and uncertainty of controls. Second, in the "exponentially decaying" design for the (5.2) case, which is shown in Table 7 and Figures 2(e) and 2(f), R2IVE consistently outperforms TSHT, and it is very close to LIML and RJIVE when there are no useful controls. LIML has good performance in the "many IVs" case (with the best MSE of 0.0072), as shown in Table 7. Similar to the previous case, when the model does

not select useful controls, the MSE of LIML is very large. In this design, R2IVE is again the most robust to both irrelevant IVs and structural model uncertainty. DML also has good performance when there are many weak instruments. Finally, in the "exponentially decaying" design (shrinking coefficients for both (5.1) and (5.2)), we allow more than 50% IVs that are useful controls (a violation of the majority rule, hence a difficult case by design) and consider $s_C = 90, 95$. When $s_C = 90$, DML and R2IVE perform best in estimators other than RJIVE, which performs relatively well since all useful controls are very weak in this setting. Ignoring these controls does not induce a large bias since they can be taken as nearly exogenous variables. When $s_C = 95$, RJIVE becomes seriously biased due to the presence of $\mathcal{A}_C \cap \mathcal{A}_R$, while R2IVE and DML are less affected by this change.

[Insert Tables 6-8 here]

6 Application to trade and economic growth

In this section, we illustrate the usefulness of R2IVE by revisiting the classic question of trade and growth, which remains a hot debate topic with important policy implications. One lingering issue in the empirical study of trade and growth is the endogeneity of trade due to the unobserved common driving forces that cause both trade and growth. Frankel and Romer (1999, *FR99* henceforth) construct an instrumental variable using the primary workhorse of empirical trade studies, the gravity model of trade (Anderson, 1979). The instrumental variable (called "proxy for trade" in *FR99*) is the sum of predicted bilateral trade shares for country i using geographical variables. They show that trade positively affects growth using cross-sectional data from 150 countries from the 1980s. Fan and Zhong (2018) extend the study of *FR99* by considering more potential instruments and a nonlinear reduced form equation. In addition to the instrument used in *FR99*, they also include total water area, coastline, arable land as a percentage of total land, land boundaries, forest area as a percentage of land area, the number of official and other commonly used languages in a

country, and the interaction terms of constructed trade proxy with all these variables. They provide a stronger result regarding trade on growth than *FR99*. However, *FR99* and Fan and Zhong (2018) did not consider that some instruments might actually be controls. We show the solutions to this structural model uncertainty using R2IVE. In this study, we use an additional air pollution variable (PM2.5) as a potential instrument.

Whether the air pollution variable can serve as an instrument or control is not clear from previous literature. Frankel and Rose (2005) find little evidence that trade is related to environmental pollution using cross-country data. In contrast, Kukla-Gryz (2009) finds that air pollution is related to international trade and per capita income. Many developing countries are gradually adopting new policies with more environment-friendly standards, hence, raising production costs, which means that air pollution can provide a direct path to growth. We try to get a clearer answer to this problem.

In addition to the aforementioned candidate IVs, we also include two randomly simulated variables from a standard normal distribution in the model to test the sensitivity of the proposed estimator[6]. The economic interpretation of the majority rule is that some candidate instruments may directly affect the outcome variable; however, the number of those unknown controls is less than valid IVs. Under the baseline model of Fan and Zhong (2018), the geographical variables form the majority group of the candidate IVs, therefore likely satisfying the majority rule. This point will be verified in a later subsection. Our study is a relaxation of Fan and Zhong (2018) regarding unknown exclusion restrictions.

6.1 Model and data

We use cross-sectional data from 158 countries (economies) and update the data to 2017 to investigate the contemporary effect of trade on growth. The summary statistics of the main data are presented in Table 9. Figure 3 is the scatter diagram of the actual and constructed share of international trade. Their correlation coefficient is 0.36.

[6]We thank an anonymous reviewer for this point.

[Insert Table 9, Figure 3 here]

We first standardize all data and consider a linear structural equation

$$Y_i = D_i\beta + \mathbf{Z}_{i\cdot}^\top \boldsymbol{\alpha} + \mathbf{S}_{i\cdot}^\top \boldsymbol{\delta} + \varepsilon_i \tag{6.1}$$

where Y_i is the log of GDP per worker in country i, D_i is the share of international trade to GDP, $\mathbf{S}_{i\cdot}$ is the size of the country, namely, population and land area, $\mathbf{Z}_{i\cdot}$ is the vector of other covariates that include all aforementioned candidate IVs summarized in Table 9 and the two randomly generated "noise" variables, and ε_i is unobserved random disturbances in the growth function. As discussed in Section 2, R2IVE considers a general nonlinear relationship between the covariates and endogenous variables. Here, we use the polynomial terms (up to order 3) of the original variables[7].

The reduced form model we consider is

$$D_i = \sum_j Z_{ij}\gamma_j + \xi_i \tag{6.2}$$

where Zs are the same variables in (6.1), and ξ_i is unobserved random disturbances, which is correlated with ε_i.

Note that we can replace the variables Y_i, D_i, and $\mathbf{Z}_{i\cdot}$ with the projections after regressing them on $\mathbf{S}_i$ (e.g., replace $\mathbf{Y}$ by $\ddot{\mathbf{Y}} = \mathcal{M}_{\mathbf{S}}\mathbf{Y}$). Then, equation (6.1) and (6.2) becomes

$$\begin{aligned} \ddot{Y}_i &= \ddot{D}_i\beta + \ddot{\mathbf{Z}}_{i\cdot}^\top \boldsymbol{\alpha} + \ddot{\varepsilon}_i \\ \ddot{D}_i &= \sum_j \ddot{Z}_{ij}\gamma_j + \ddot{\xi}_i \end{aligned} \tag{6.3}$$

[7]In an earlier version of the paper (Fan and Wu, 2020), we considered a nonparametric additive reduced form model. The new variable selection results (with data standardization to highlight variable selection performance across different methods) of R2IVE are comparable to the earlier version.

6.2 Empirical results

To investigate the influence of the uncertain IV on the estimation of β, we explore the empirical results with and without the air quality index (PM2.5) in $\mathbf{Z}$ and compare our estimated value with *FR99*, NAIVE, sisVIVE, post-sisVIVE, TSHT, and the median estimators defined in Windmeijer et al. (2019) and in (3.12), respectively. The results are summarized in Tables 10 and 11.

When the variable PM2.5 is not included in the candidate set, NAIVE and R2IVE select two relevant instruments in the reduced form: the proxy for trade and the interaction term of the proxy for trade and the number of official and other commonly used languages, which are likely to be excluded IVs (the J test p value is 0.40), while TSHT only selects the latter. In Table 10, OLS has severe bias because of the endogeneity issue. The t statistics for the NAIVE on trade is 2.76, compared to 4.66 for the 2SLS. Post-sisVIVE is not feasible here since sisVIVE does not select any excluded instrument.

[Insert Tables 10 and 11 here]

When the variable PM2.5 is considered, it is selected by R2IVE and NIAVE as a relevant variable for trade, in addition to the aforementioned two excluded IVs. The valid IV to control ratio is at worst 2:1. Therefore, the majority rule is likely satisfied here. Table 11 summarizes the estimation results. If we use NAIVE, under the operating assumption that all IVs are valid, the estimated causal effect is 0.88 (with a standard error of 0.18). For the identity of PM2.5, the sisVIVE, ALasso, and R2IVE all select it as a useful control. This result supports the conjecture that there may be a direct air pollution pathway to growth, as discussed in some theoretical models on the environmental Kuznets Curve (Dasgupta et al., 2002). The simulated completely random noise IVs are not selected by R2IVE as an excluded instrument or control. The causal effect is estimated to be 1.15, which is close to the results in Table 10. At last, the median estimators tend to be sensitive to the unknown control variables.

In summary, this empirical study shows that our method is most robust when facing "noisy" instruments and uncertain controls. Other methods, specifically, sisVIVE and ALasso can select the strong control variable (PM2.5). However, sisVIVE cannot select any excluded instrument without PM2.5 in the model, and ALasso does not discard the pure noise variables. TSHT does not perform well in the weak IV case and is inconsistent in the selection results (it selects the constructed trade and language interaction as a valid instrument without the PM2.5, and only PM2.5 as a control variable after it is added). In hindsight, the NAIVE in Fan and Zhong (2018) did not accurately estimate the effect of trade in the 2010s due to misclassifying instruments and controls. Our method can select useful control variables and discard weak IVs in the presence of a mixed set of covariates.

7 Conclusion

This paper develops an IV estimator (R2IVE) robust to both structural and reduced form model uncertainty when estimating endogenous treatment effects. The proposed method extends Kang et al. (2016) and Windmeijer et al. (2019) by considering a high-dimensional instrumental variable setting which allows for a more general (possibly nonlinear) relationship between the instruments and endogenous variables. The proposed R2IVE is shown to be root-n consistent and asymptotically normal. Monte Carlo simulations demonstrate that R2IVE performs better than the existing IV estimators (such as RJIVE, NAIVE, sisVIVE, ALasso, and TSHT) in many empirically relevant scenarios. The empirical study revisits the classic question of trade and growth. It is shown that the R2IVE can estimate the endogenous treatment effect with a large set of instruments without knowing which ones are relevant or valid and whether a variable is a useful control. We will pursue causal inference in the model of many weak and nearly valid instruments in the future.

References

Anderson, J., 1979. A theoretical foundation for gravity equation. American Economic Review 69, 106–16.

Angrist, J., Pischke, J., 2009. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press, Princeton NJ.

Bai, J., Ng, S., 2010. Instrumental variable estimation in a data rich environment. Econometric Theory 26, 1577–1606.

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

Belloni, A., Chernozhukov, V., Hansen, C., 2014. High-dimensional methods and inference on structural and treatment effects. Journal of Economic Perspectives 28, 29–50.

Caner, M., Fan, Q., 2015. Hybrid generalized empirical likelihood estimators: Instrument selection with adaptive lasso. Journal of Econometrics 187, 256–274.

Caner, M., Han, X., Lee, Y., 2018. Adaptive elastic net GMM estimation with many invalid moment conditions: Simultaneous model and moment selection. Journal of Business & Economic Statistics 36, 24–46.

Cheng, X., Liao, Z., 2015. Select the valid and relevant moments: An information-based lasso for GMM with many moments. Journal of Econometrics 186, 443–464.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J., 2018. Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal 21, C1–C68.

Chernozhukov, V., Hansen, C., Spindler, M., 2015. Post-selection and post-regularization inference in linear models with many controls and instruments. American Economic Review 105, 486–490.

Dasgupta, S., Laplante, B., Wang, H., Wheeler, D., 2002. Confronting the environmental Kuznets curve. Journal of Economic Perspectives 16, 147–168.

Donald, S.G., Newey, W.K., 2001. Choosing the number of instruments. Econometrica 69, 1161–1191.

Fan, J., Peng, H., 2004. Nonconcave penalized likelihood with a diverging number of parameters. Annals of Statistics 32, 928–961.

Fan, Q., Wu, Y., 2020. Endogenous treatment effect estimation with some invalid and irrelevant instruments. arXiv preprint arXiv:2006.14998 .

Fan, Q., Zhong, W., 2018. Nonparametric additive instrumental variable estimator: A group shrinkage estimation perspective. Journal of Business & Economic Statistics 36, 388–399.

Frankel, J., Romer, D., 1999. Does trade causes growth? American Economic Review 89, 379–399.

Frankel, J., Rose, A., 2005. Is trade good or bad for the environment? sorting out the causality. The Review of Economics and Statistics 87, 85–91.

Gautier, E., Tsybakov, A., 2011. High-dimensional instrumental variables regression and confidence sets , working paper.

Guo, Z., Kang, H., Cai, T., Small, D., 2018. Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. Journal of the Royal Statistical Society: Series B 80, 793–815.

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

Huang, J., Ma, S., Zhang, C.H., 2008. Adaptive lasso for sparse high-dimensional regression models. Statistica Sinica 18, 1603–1618.

Kang, H., Zhang, A., Cai, T.T., Small, D.S., 2016. Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. Journal of the American Statistical Association 111, 132–144.

Kukla-Gryz, A., 2009. Economic growth, international trade and air pollution: A decomposition analysis. Ecological Economics 68, 1329–1339.

Lin, W., Feng, R., Li, H., 2015. Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. Journal of the American Statistical Association 110, 270–288.

Okui, R., 2011. Instrumental variable estimation in the presence of many moment conditions. Journal of Econometrics 165, 70–86.

Wang, H., Li, B., Leng, C., 2009. Shrinkage tuning parameter selection with a diverging number of parameters. Journal of the Royal Statistical Society: Series B 71, 671–683.

Windmeijer, F., Farbmacher, H., Davies, N., Smith, G., 2019. On the use of the lasso for instrumental variables estimation with some invalid instruments. Journal of the American Statistical Association 114, 1139–1350.

Yang, Y., Zou, H., 2012. An efficient algorithm for computing the HHSVM and its generalizations. Journal of Computational and Graphical Statistics 22, 396–415.

Zivot, E., Wang, J., 1998. Inference on structural parameters in instrumental variables regression with weak instruments. Econometrica 66, 1389–1404.

Zou, H., 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101, 1418–1429.

Zou, H., Zhang, H., 2009. On the adaptive elastic-net with a diverging number of parameters. Annals of Statistics 37, 1733–1751.

Table 1: The summary of all simulation model settings

(5.1)	(5.2)	n	L_n	s_R	s_C	q	$\lvert M1\rvert$	$\lvert M2\rvert$	$\lvert M3\rvert$	$\lvert M4\rvert$
Cutoff	Cutoff	200	100	10	**0**	**10**	10	0	90	0
		200	100	10	**10**	**7**	7	3	83	7
		200	100	10	**30**	7	7	3	63	27
		200	100	**4**	30	**3**	3	1	67	29
		200	100	**15**	30	**10**	10	5	60	25
		200	100	**20**	30	**14**	14	6	56	24
		200/500/1000	**100**	20	20	14	14	6	66	14
		500	**250**	20	20	14	14	6	216	14
Cutoff	Many weak	200	100	-	**0**	-	-	-	-	-
		200	100	-	**10**	7	-	-	-	-
Cutoff	Exp. Decay	200	100	-	**0**	-	-	-	-	-
		200	100	-	**10**	7	-	-	-	-
Exp. Decay	Exp. Decay	200	100	-	**90**	-	-	-	-	-
		200	100	-	**95**	-	-	-	-	-

NOTE: The sets $M1 - M4$ are $\mathcal{A}_C^c \cap \mathcal{A}_R$, $\mathcal{A}_C \cap \mathcal{A}_R$, $\mathcal{A}_C^c \cap \mathcal{A}_R^c$, and $\mathcal{A}_C \cap \mathcal{A}_R^c$. (5.1) and (5.2) are structural and reduced form equations stated in Section 5. L_n is the dimension of $\mathbf{Z}$; s_R and s_C are the number of relevant IV and useful controls, respectively; $q \leq s_R$ determines the distribution of 2×2 contingent sets $M1 - M4$. The cardinality of the IV set is shown for strict cutoff designs. The default coefficient values are $\boldsymbol{\gamma}^* = (2, 0.75, 1.5, 1, \ldots, \mathbf{0}_{L_n - s_R})^\top$ and $\boldsymbol{\alpha}^* = (\mathbf{0}_q, \boldsymbol{\iota}_{s_C}, \mathbf{0}_{L_n - q - s_C})^\top$ for the cutoff designs until further noted such as in Section 5.5. To distinguish from the strong IV case, "-" is shown in many weak IVs case (including exponential decaying non-sparse cases with some weak IVs). In the following, Section 5.2 uses row 1-3 of panel 1, Section 5.3 uses row 4-6 of panel 1, Section 5.4 uses row 7-8 of panel 1, Section 5.5 uses row 7 ($n = 200$) of panel 1, and Section 5.6 uses panels 2-4, which are the approximate sparse and nearly valid (panel 4) situations.

Table 2: Change the number of controls s_C, fix $n = 200, L_n = 100, s_R = 10, c = 1$

s_C		Bias	Std Dev	MSE	Oracle	Mean	Median	Max	Min	Freq
0	OLS	0.0171	0.0102	0.0004						
	2SLS	0.0079	0.0141	0.0002						
	Oracle 2SLS	0.0002	0.0169	0.0001						
	LIML	-0.0005	0.0102	0.0001						
	RJIVE	0.0001	0.0102	0.0001						
	NAIVE	0.0037	0.0157	0.0001	R(10)	12.21	12	17	10	1
	sisVIVE	0.0081	0.0107	0.0002	C(0)	0.16	0	63	0	-
	post-sisVIVE	0.0081	0.0142	0.0002						
	median	0.7030	0.1262	0.5102						
	ALasso	0.0078	0.0103	0.0002	C(0)	0	0	0	0	-
	TSHT	0.0024	0.0406	0.0002	$C^c \cap R(10)$	10.22	10	13	9	0.99
					R(10)	10.22	10	13	9	0.99
	median	-0.0023	0.0195	0.0004						
	R2IVE	0.0001	0.0168	0.0001	R(10)	10.14	10	15	10	1
					C(0)	0	0	1	0	-
10	OLS	0.2754	0.0497	0.0781						
	2SLS	0.2698	0.0511	0.0751						
	Oracle 2SLS	0.0008	0.0228	0.0002						
	LIML	-0.3085	0.1318	0.1126						
	RJIVE	0.2663	0.0531	0.0737						
	NAIVE	0.2679	0.0517	0.0741	R(10)	12.19	12	16	10	1
	sisVIVE	0.2620	0.1130	0.0814	C(10)	46.20	48	69	11	0.99
	post-sisVIVE	0.4824	0.1893	0.2722						
	median	0.7138	0.1200	0.5238						
	ALasso	0.2489	0.0368	0.0633	C(10)	5.93	6	10	0	0
	TSHT	0.1704	0.0650	0.0305	$C^c \cap R(7)$	9.06	9	12	7	0.99
					R(10)	10.23	10	13	9	0.99
	median	0.0468	0.0336	0.0033						
	R2IVE	0.0007	0.0226	0.0002	R(10)	10.16	10	14	10	1
					C(10)	10.01	11	11	10	1
	DML	-0.0037	0.0214	0.0003						
30	OLS	0.2731	0.0941	0.0838						
	2SLS	0.2674	0.0952	0.0809						
	Oracle 2SLS	0.0003	0.0241	0.0002						
	LIML	-2.6342	0.8242	7.6180						
	RJIVE	0.2658	0.0948	0.0796						
	NAIVE	0.2652	0.0957	0.0801	R(10)	12.10	12	17	10	1
	sisVIVE	0.1848	0.1144	0.0472	C(30)	60.88	65	87	31	1
	post-sisVIVE	0.3217	0.1901	0.1493						
	median	0.6904	0.1418	0.4968						
	ALasso	0.2620	0.0558	0.0718	C(30)	26.14	27	32	14	0
	TSHT	0.1864	0.0880	0.0353	$C^c \cap R(7)$	9.30	9	12	8	0.99
					R(10)	10.19	10	12	8	0.98
	median	0.0559	0.0361	0.0044						
	R2IVE	0.0015	0.0240	0.0003	R(10)	10.15	10	14	10	1
					C(30)	30.07	30	32	29	1
	DML	-0.0543	0.0248	0.0040						

NOTE: This table summarizes the averages of estimated bias, standard deviations, MSE, and model selection performance. The sixth column "Oracle" indicates the type and number of variables that are intended to be selected by various methods, where R, $C^c \cap R$, C represent relevant covariates in $\mathcal{A}_R$, excluded IV and control variables (and in the parenthesis, we put the true number of each type of variable that is intended to be selected by respective methods, e.g., for TSHT, the intended IV set is $\mathcal{A}_C^c \cap \mathcal{A}_R$). The average number of variables selected, together with the median, minimum and maximum numbers, and the proportion of times that selected variables include all relevant IVs, excluded IVs, or controls, are reported for respective estimators. Notice that sisVIVE and post-sisVIVE share the same selection results. The first median estimator in the same panel of ALasso is the median estimator defined in Windmeijer et al. (2019). The second median estimator in the same panel of R2IVE is defined in (3.12). DML uses post Lasso as the machine learning method. We show the results of DML for models with included controls, such as the cases in panels 2 and 3 (s_C=10, 30). DML variable selection is not reported due to the transformed model structure.

Table 3: Change the number of relevant instruments s_R, fix $n = 200, L_n = 100, s_C = 30, c = 1$

s_R		Bias	Std Dev	MSE	Oracle	Mean	Median	Max	Min	Freq
	OLS	0.3843	0.1694	0.1754						
	2SLS	0.3697	0.1735	0.1664						
	Oracle 2SLS	-0.0002	0.0453	0.0007						
	LIML	-12.9344	96.6007	9.50E+03						
	RJIVE	0.3577	0.1924	0.1650						
	NAIVE	0.3638	0.1756	0.1650	R(4)	5.42	5	9	4	1
4	sisVIVE	0.5236	0.0898	0.2822	C(0)	46.45	46	68	34	0.80
	post-sisVIVE	0.6771	0.1792	0.4660						
	median	0.7430	0.1409	0.5720						
	ALasso	0.2931	0.0694	0.0908	C(30)	29.79	30	32	26	0.01
	TSHT	0.1870	0.2916	0.0388	$C^c \cap R(4)$	4.11	4	6	3	0.97
					R(4)	4.27	4	7	3	0.99
	median	0.0553	0.0740	0.0085						
	R2IVE	0.0065	0.0454	0.0038	R(4)	4.15	4	8	4	1
					C(30)	30.30	30	36	29	0.99
	DML	-0.0510	0.0414	0.0068						
	OLS	0.2679	0.0749	0.0773						
	2SLS	0.2642	0.0756	0.0755						
	Oracle 2SLS	0.0019	0.0186	0.0001						
	LIML	-1.3929	96.6007	2.0783						
	RJIVE	0.2621	0.0774	0.0747						
	NAIVE	0.2629	0.0759	0.0750	R(15)	16.84	16	22	15	1
15	sisVIVE	0.0633	0.0526	0.0068	C(30)	46.14	42	92	30	1
	post-sisVIVE	0.0852	0.0512	0.0167						
	median	0.6746	0.1353	0.4734						
	ALasso	0.2682	0.0514	0.0746	C(30)	19.58	20	31	5	0
	TSHT	0.1826	0.0598	0.0344	$C^c \cap R(9)$	13.81	14	18	11	0.98
					R(15)	15.24	15	18	14	0.97
	median	0.0631	0.0348	0.0052						
	R2IVE	0.0016	0.0186	0.0002	R(15)	15.19	15	19	15	1
					C(30)	31.93	30	31	29	1
	DML	-0.0464	0.0205	0.0028						
	OLS	0.2378	0.0640	0.0605						
	2SLS	0.2350	0.0646	0.0592						
	Oracle 2SLS	0.0005	0.0152	0.0001						
	LIML	-0.8341	0.2274	0.7475						
	RJIVE	0.2393	0.0651	0.0615						
	NAIVE	0.2340	0.0648	0.0589	R(20)	21.81	21	29	20	1
20	sisVIVE	0.0329	0.0139	0.0013	C(30)	40.81	40	85	31	1
	post-sisVIVE	0.0352	0.0216	0.0018						
	median	0.6047	0.1448	0.3867						
	ALasso	0.2371	0.0529	0.0590	C(30)	13.52	14	29	0	0
	TSHT	0.1655	0.0474	0.0281	$C^c \cap R(14)$	18.33	18	22	15	0.97
					R(20)	20.14	20	22	19	0.96
	median	0.0517	0.0255	0.0033						
	R2IVE	0.0010	0.0155	0.0001	R(20)	20.25	20	24	20	1
					C(30)	31.92	31	31	30	1
	DML	-0.0404	0.0179	0.0021						

Please see the table notes in Table 2.

Table 4: Change the sample size n, IV dimensionality L_n, and α^* values, fix $s_R = 20, s_C = 20$

		Bias	Std Dev	MSE	Oracle	Mean	Median	Max	Min	Freq
	OLS	0.2453	0.0509	0.0626						
	2SLS	0.2424	0.0515	0.0612						
	Oracle 2SLS	0.0012	0.0149	0.0001						
	LIML	-0.3567	0.1289	0.1439						
	RJIVE	0.2409	0.0515	0.0607						
$n = 200$	NAIVE	0.2417	0.0518	0.0610	R(20)	21.70	21	29	20	1
$L_n = 100$	sisVIVE	0.0340	0.0127	0.0013	C(20)	28.01	27	62	20	1
$c = 1$	post-sisVIVE	0.0327	0.0192	0.0016						
	median	0.6331	0.1184	0.4148						
	ALasso	0.2350	0.0466	0.0574	C(20)	6.32	7	17	0	0
	TSHT	0.1459	0.0403	0.0220	$C^c \cap R(14)$	17.58	17	21	14	0.97
					R(20)	20.18	20	23	19	0.97
	median	0.0484	0.0228	0.0029						
	R2IVE	0.0005	0.0149	0.0001	R(20)	20.24	20	24	20	1
					C(20)	21.44	21	30	20	1
	DML	-0.0180	0.0156	0.0006						
	OLS	0.1834	0.0243	0.0342						
	2SLS	0.1783	0.0251	0.0324						
	Oracle 2SLS	0.0003	0.0092	0.0001						
	LIML	-0.1082	0.0868	0.0131						
	RJIVE	0.1776	0.0253	0.0322						
$n = 500$	NAIVE	0.1781	0.0251	0.0323	R(20)	25.64	26	31	21	1
$L_n = 100$	sisVIVE	0.0176	0.0065	0.0004	C(20)	24.31	23	45	20	1
$c = 0.75$	post-sisVIVE	0.0121	0.0111	0.0003						
	median	0.5996	0.1105	0.3718						
	ALasso	0.1748	0.0191	0.0309	C(20)	13.32 13	21	6	0	
	TSHT	0.0018	0.0276	0.0001	$C^c \cap R(14)$	14.14	14	17	14	1
					R(20)	20.18	20	23	20	1
	median	0.0277	0.0126	0.0009						
	R2IVE	0.0001	0.0092	0.0001	R(20)	20.09	20	23	20	1
					C(20)	20	20	20	20	1
	DML	-0.0062	0.0094	0.0001						
	OLS	0.1250	0.0117	0.0158						
	2SLS	0.1188	0.0125	0.0142						
	Oracle 2SLS	0.0001	0.0065	0.0001						
	LIML	0.0041	0.019	0.0002						
	RJIVE	0.1184	0.0118	0.0142						
$n = 1000$	NAIVE	0.1186	0.0125	0.0142	R(20)	25.76	26	30	22	1
$L_n = 100$	sisVIVE	0.0121	0.0042	0.0002	C(20)	22.33	22	48	20	1
$c = 0.5$	post-sisVIVE	0.0056	0.0074	0.0001						
	median	0.5444	0.1124	0.3090						
	ALasso	0.1129	0.0128	0.0129	C(20)	13.80	14	21	6	0
	TSHT	0.0005	0.0239	0.0001	$C^c \cap R(14)$	14.14	14	17	14	1
					R(20)	20.17	20	23	20	1
	median	0.0191	0.0085	0.0004						
	R2IVE	0.0003	0.0065	0.0001	R(20)	20.05	20	22	20	1
					C(20)	20	20	20	20	1
	DML	-0.0027	0.0065	0.0001						
	OLS	0.1259	0.0165	0.0161						
	2SLS	0.1218	0.0173	0.0151						
	Oracle 2SLS	0.0005	0.0092	0.0001						
	LIML	0.0047	0.0305	0.0004						
	RJIVE	0.1195	0.0168	0.0146						
$n = 500$	NAIVE	0.1206	0.0175	0.0148	R(20)	20.03	20	21	20	1
$L_n = 250$	sisVIVE	0.0226	0.0074	0.0006	C(20)	26.93	25	118	20	1
$c = 0.5$	post-sisVIVE	0.0188	0.0110	0.0006						
	median	0.7136	0.0789	0.5155						
	ALasso	0.1209	0.0170	0.0149	C(20)	0.40	0	14	0	0
	TSHT	0.0856	0.0271	0.0075	$C^c \cap R(14)$	18.74	19	22	16	1
					R(20)	20.19	20	22	20	1
	median	0.0270	0.0126	0.0009						
	R2IVE	0.0007	0.0092	0.0001	R(20)	20.09	20	22	20	1
					C(20)	20	20	20	20	1
	DML	-0.0043	0.0091	0.0001						

Please see the table notes in Table 2.

Table 5: Change the size of non-zero coefficients in $\boldsymbol{\gamma}^*$, fix $n = 200$, $L_n = 100$, $s_R = 20$, $s_C = 20$, $q = 14$, $c = 1$

γ		Bias	Std Dev	MSE	Oracle	Mean	Median	Max	Min	Freq
1	OLS	0.2484	0.0505	0.0643						
	2SLS	0.2434	0.0517	0.0619						
	Oracle 2SLS	0.0017	0.0200	0.0002						
	LIML	-0.3454	0.1239	0.1346						
	RJIVE	0.2387	0.0545	0.0600						
	NAIVE	0.2417	0.0521	0.0612	R(20)	34.83	35	43	26	1
	sisVIVE	0.0560	0.0244	0.0037	C(20)	31.17	29	83	20	1
	post-sisVIVE	0.0643	0.0314	0.0062						
	median	0.6815	0.1467	0.4859						
	ALasso	0.3061	0.0514	0.0963	C(20)	10.60	11	22	0	0
	TSHT	0.0792	0.0708	0.0111	$C^c \cap R(14)$	15.50	15	21	14	1
					R(20)	20.19	20	22	20	1
	median	0.0585	0.0260	0.0041						
	R2IVE	0.0017	0.0200	0.0002	R(20)	20.26	20	24	20	1
					C(20)	20.01	20	21	20	1
	DML	-0.0139	0.0194	0.0005						
0.75	OLS	0.3350	0.0667	0.1170						
	2SLS	0.3278	0.0685	0.1124						
	Oracle 2SLS	0.0020	0.0265	0.0002						
	LIML	-1.0585	0.3538	1.2455						
	RJIVE	0.3214	0.0749	0.1089						
	NAIVE	0.3256	0.0692	0.1111	R(20)	34.77	35	44	27	1
	sisVIVE	0.1147	0.0777	0.0192	C(20)	41.66	35	86	21	1
	post-sisVIVE	0.1610	0.0764	0.0440						
	median	0.6791	0.1518	0.4843						
	ALasso	0.3465	0.0808	0.1266	C(20)	15.14	15	24	6	0.03
	TSHT	0.1198	0.1270	0.0239	$C^c \cap R(14)$	15.47	15	21	13	0.94
					R(20)	20.11	20	23	18	0.92
	median	0.0793	0.0333	0.0074						
	R2IVE	0.0028	0.0265	0.0003	R(20)	20.38	20	24	20	1
					C(20)	20.03	20	21	20	1
	DML	-0.0109	0.0248	0.0006						
0.5	OLS	0.6529	0.1307	0.4417						
	2SLS	0.6477	0.1346	0.4360						
	Oracle 2SLS	0.0054	0.0399	0.0006						
	LIML	-7.3013	1.44E+5	8.22E+04						
	RJIVE	0.6457	0.1438	0.4376						
	NAIVE	0.6479	0.1366	0.4378	R(20)	27.06	27	34	21	1
	sisVIVE	0.2788	0.1147	0.0909	C(20)	53.62	56	83	23	1
	post-sisVIVE	0.4140	0.1664	0.2027						
	median	0.6845	0.1412	0.4885						
	ALasso	0.3416	0.0942	0.1256	C(20)	18.92	19	21	13	0.39
	TSHT	0.1944	0.4754	0.0573	$C^c \cap R(14)$	12.24	12	18	5	0.04
					R(20)	16.31	16	20	8	0.01
	median	0.1171	0.0535	0.0166						
	R2IVE	0.0073	0.0397	0.0008	R(20)	20.64	20	26	20	1
					C(20)	20.04	20	22	20	1
	DML	0.0016	0.0358	0.0008						

Please see the table notes in Table 2.

Table 6: Many weak setting, $n = 200$, $L_n = 100$, $c = 1$

s_C		Bias	Std Dev	MSE	Oracle	Mean	Median	Max	Min	Freq
0	OLS	0.2013	0.0327	0.0415						
	2SLS	0.1141	0.0517	0.0142						
	Oracle 2SLS	0.1141	0.0517	0.0142						
	LIML	-0.0023	0.0432	0.0019						
	RJIVE	-0.0032	0.0453	0.0021						
	NAIVE	0.0961	0.0554	0.0105	R(100)	60.20	60	72	47	-
	sisVIVE	0.1781	0.0859	0.0391	C(0)	20.07	13.5	80	0	-
	post-sisVIVE	0.2432	0.0712	0.0788						
	median	0.4206	0.0943	0.1858						
	ALasso	0.1153	0.0349	0.0145	C(0)	0	0	0	0	-
	TSHT	0.4496	3.5185	0.2758	$C^c \cap R(100)$	59.01	100	100	1	-
					R(100)	59.03	100	100	1	-
	median	0.1008	0.0825	0.0170						
	R2IVE	0.1069	0.0545	0.0128	R(100)	44.24	58	84	1	-
					C(0)	20.13	0	96	0	-
10	OLS	0.9637	0.1726	0.9596						
	2SLS	0.9872	0.1896	1.0154						
	Oracle 2SLS	0.1239	0.0574	0.0167						
	LIML	81.7957	943.7155	8.97E+05						
	RJIVE	1.0280	0.2641	1.1266						
	NAIVE	0.9936	0.1942	1.0348	R(100)	59.72	60	75	47	-
	sisVIVE	0.2706	0.0776	0.0792	C(10)	40.72	41	81	10	1
	post-sisVIVE	0.3540	0.0940	0.1404						
	median	0.4797	0.0998	0.2400						
	ALasso	0.3254	0.0755	0.1116	C(10)	9.99	10	10	9	0.99
	TSHT	0.8227	8.0574	0.8978	$C^c \cap R(90)$	56.93	93	99	1	-
					R(100)	59.80	100	100	1	-
	median	0.2230	0.2730	0.1242						
	R2IVE	0.1080	0.0929	0.0414	R(100)	44.58	58.5	86	1	-
					C(10)	10.58	10	23	9	0.99
	DML	0.1052	0.0572	0.0128						

Please see the table notes in Table 2.

Table 7: Exponentially decaying design for Eq. (5.2), $n = 200$, $L_n = 100$, $c = 1$

s_C		Bias	Std Dev	MSE	Oracle	Mean	Median	Max	Min	Freq
0	OLS	0.3979	0.0412	0.1601						
	2SLS	0.2657	0.0767	0.0730						
	Oracle 2SLS	0.2657	0.0767	0.0730						
	LIML	0.0007	0.0850	0.0072						
	RJIVE	-0.0160	0.0912	0.0086						
	NAIVE	0.1439	0.0978	0.0244	R(100)	18.39	18	28	9	-
	sisVIVE	0.5398	0.0749	0.2970	C(0)	8.67	8	26	2	-
	post-sisVIVE	0.7398	0.1281	0.5534						
	median	0.7642	0.1203	0.5985						
	ALasso	0.2620	0.0464	0.0708	C(0)	0	0	0	0	-
	TSHT	0.1970	6.1882	0.1090	$C^c \cap R(100)$	4.52	2	100	1	-
					R(100)	4.61	2	100	1	-
	median	0.0436	0.1189	0.0160						
	R2IVE	0.0409	0.1185	0.0099	R(100)	4.93	5	10	2	-
					C(0)	52.89	57	97	2	-
10	OLS	0.6553	0.2541	0.4927						
	2SLS	0.6050	0.3004	0.4805						
	Oracle 2SLS	0.2552	0.0819	0.0677						
	LIML	-14.9401	515.5947	2.66E+05						
	RJIVE	0.5998	0.6050	0.7258						
	NAIVE	0.5636	0.3367	0.5233	R(100)	17.98	18	28	8	-
	sisVIVE	0.5961	0.0782	0.3614	C(10)	20.53	20	36	13	1
	post-sisVIVE	0.7374	0.1445	0.5500						
	median	0.7605	0.1231	0.5935						
	ALasso	0.3128	0.0751	0.1035	C(10)	10.01	10	11	10	1
	TSHT	0.1785	7.3405	0.0988	$C^c \cap R(90)$	3.82	2	98	1	-
					R(100)	4.03	2	100	1	-
	median	0.0270	0.1426	0.0211						
	R2IVE	0.0249	0.1230	0.0098	R(100)	5.08	5	12	2	-
					C(10)	11.84	11	21	10	1
	DML	0.0915	0.1061	0.0172						

Please see the table notes in Table 2.

Table 8: Exponentially decaying design for both Eqs. (5.1) and (5.2), $n = 200$, $L_n = 100$

s_C		Bias	Std Dev	MSE	Oracle	Mean	Median	Max	Min	Freq
	OLS	0.4254	0.0641	0.1849						
	2SLS	0.3013	0.0955	0.0972						
	Oracle 2SLS	0.0651	0.1472	0.0104						
	LIML	-1.0370	0.3769	1.2174						
	RJIVE	0.0534	0.1476	0.0246						
90	NAIVE	0.1426	0.0977	0.0243	R(100)	18.28	18	29	9	-
	sisVIVE	0.5397	0.0700	0.2962	C(90)	8.78	8	37	3	-
	post-sisVIVE	0.7372	0.1282	0.5493						
	median	0.7529	0.1189	0.5810						
	ALasso	0.2646	0.0500	0.0725	C(90)	0	0	1	0	-
	TSHT	0.1784	6.5985	0.0930	$C^c \cap R(10)$	3.42	2	100	1	-
					R(100)	3.49	2	100	1	-
	median	0.0415	0.1132	0.0145						
	R2IVE	0.0377	0.1188	0.0104	R(100)	4.98	5	10	2	-
					C(90)	52.34	55	94	2	-
	DML	0.0390	0.1975	0.0265						
	OLS	0.5428	0.0597	0.2989						
	2SLS	0.4571	0.0931	0.2158						
	Oracle 2SLS	0.0237	0.1671	0.0114						
	LIML	-1.2486	1.4646	3.7042						
	RJIVE	0.3182	0.1391	0.1206						
95	NAIVE	0.1372	0.0980	0.0223	R(100)	18.09	18	28	9	-
	sisVIVE	0.6316	0.0767	0.4048	C(95)	13.65	13	45	7	-
	post-sisVIVE	0.7517	0.1354	0.5712						
	median	0.7461	0.1201	0.5711						
	ALasso	0.2636	0.0478	0.0718	C(95)	0	0	0	0	-
	TSHT	0.1853	6.3317	0.0996	$C^c \cap R(5)$	3.14	2	100	1	-
					R(100)	3.23	2	100	1	-
	median	0.0351	0.1186	0.0153						
	R2IVE	0.0375	0.1196	0.0097	R(100)	5.02	5	10	2	-
					C(95)	52.14	57	97	2	-
	DML	0.0371	0.1980	0.0254						

Please see the table notes in Table 2.

Table 9: Summary statistics

	mean	std.dev	median	minimum	maximum	sample size
Ln Income Per Capita	10.18	1.1	10.42	7.46	12.03	158
Real Trade Share	0.87	0.52	0.76	0.2	4.13	158
Constructed Trade Share	0.09	0.05	0.08	0.02	0.3	158
Ln Population	1.38	1.8	1.48	-3.04	6.67	158
Ln Area (Land)	11.73	2.26	12.02	5.7	16.61	158
Area (Water)	25,378	100,818.4	2,365	0	891,163	158
Coastline	4,268.6	17,451.71	523	0	202,080	158
Land Boundaries	2,837.8	3,407.8	1,899.5	0	22,147	158
% Forest	29.89	22.38	30.62	0	98.26	158
% Arable Land	40.95	21.55	42.06	0.56	82.56	158
PM2.5	25.05	19.43	22	5.9	100	158
Languages	1.87	2.13	1	1	16	158

NOTE: Income per capita is measured in dollars. Population is measured in millions. Land area and water area are measured in square kilometers. Coastline and land boundaries are measured in kilometers. PM2.5 is measured in micrograms per cubic meter. Source: *FR99*, Penn World Table (PWT 9.1), the World Bank, and State of Global Air.

Table 10: Estimation results for the trade and income data (PM2.5 not included)

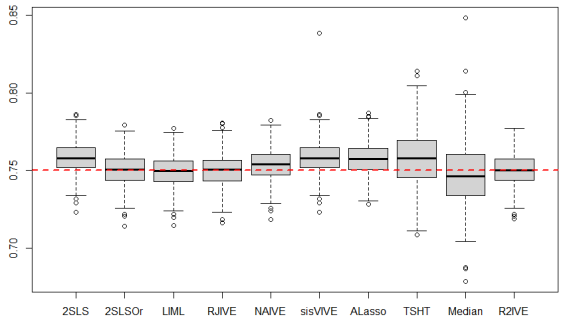
	OLS	*FR99*	2SLS	NAIVE	sisVIVE
trade share	0.41***	0.67***	0.82***	0.61***	0.83
	(0.08)	(0.23)	(0.17)	(0.22)	-
R^2	0.13	0.05	0.12	0.05	0.15
Sample Size	158	158	158	158	158
	median-W	ALasso	TSHT	median	R2IVE
trade share	0.79	0.76***	0.76**	0.93	1.11***
	-	(0.16)	(0.40)	-	(0.24)
R^2	-	0.22	0.02	-	0.33
Sample Size	158	158	158	158	158

The sisVIVE method does not report standard deviation. All variables are standardized. ***, ** stand for significance levels of 1%, 5%, respectively. The first "median" column (median-W) is the median estimator in Windmeijer et al. (2019). The second "median" column is proposed in (3.12).

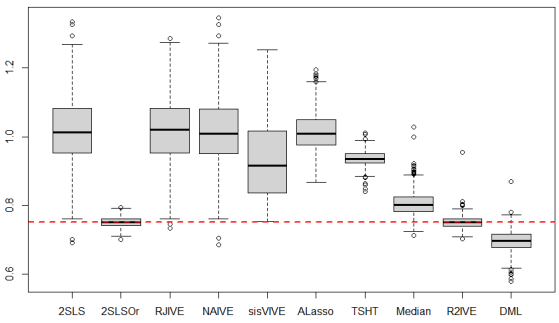
Table 11: Estimation results for the trade and income data (PM2.5 included)

	OLS	*FR99*	2SLS	NAIVE	sisVIVE	post-sisVIVE
trade share	0.41***	0.67***	0.94***	0.88***	0.94	0.89***
	(0.08)	(0.23)	(0.15)	(0.18)	-	(0.14)
R^2	0.13	0.05	0.19	0.12	0.19	0.31
Sample Size	158	158	158	158	158	158
	median-W	ALasso	TSHT	median	R2IVE	
trade share	1.16	0.72***	1.48***	0.96	1.15***	
	-	(0.15)	(0.41)	-	(0.19)	
R^2	-	0.30	0.12	-	0.42	
Sample Size	158	158	158	158	158	

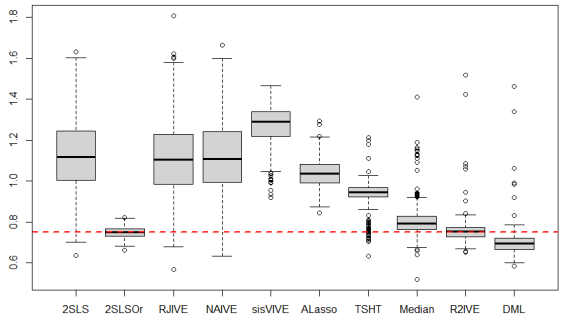
Please see the table notes in Table 10. The post-sisVIVE is a 2SLS estimator that takes the set of controls selected by sisVIVE. All variables are standardized. *** and ** denote significance levels of 1% and 5%, respectively.



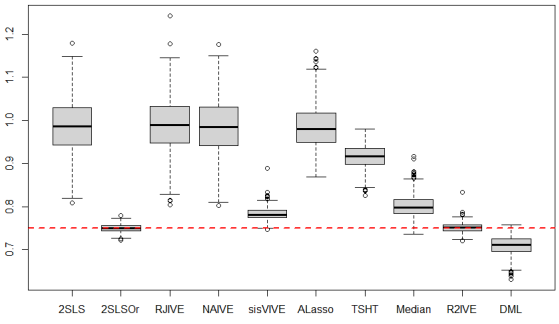
(a) $s_C = 0, q = 10$



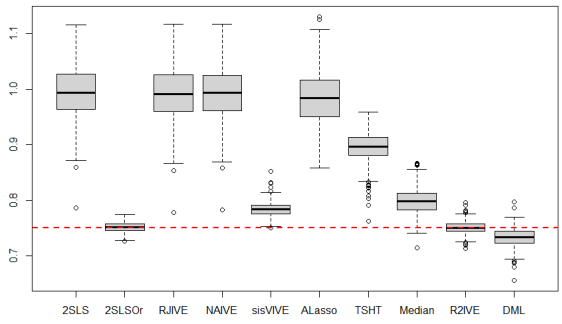
(b) $s_C = 30, q = 7$



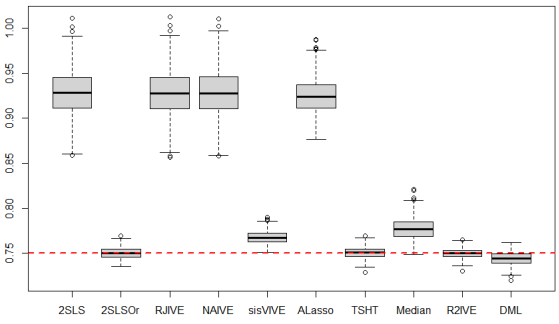
(c) $s_R = 4, q = 3$



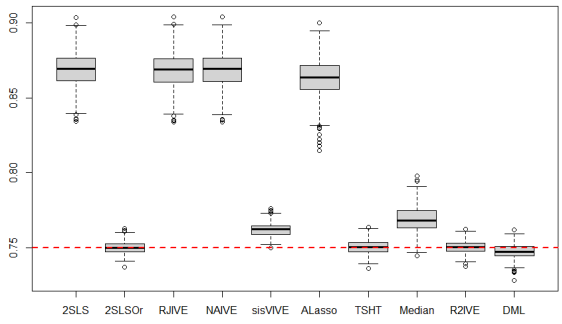
(d) $s_R = 20, q = 14$



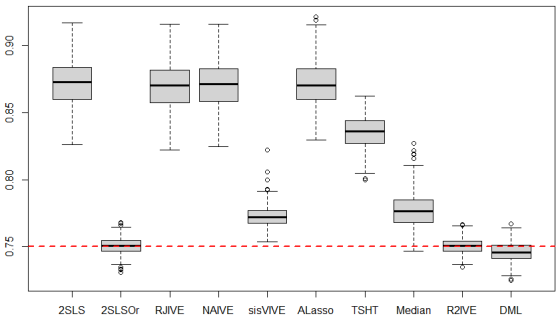
(e) $n = 200, L_n = 100, c = 1$



(f) $n = 500, L_n = 100, c = 0.75$

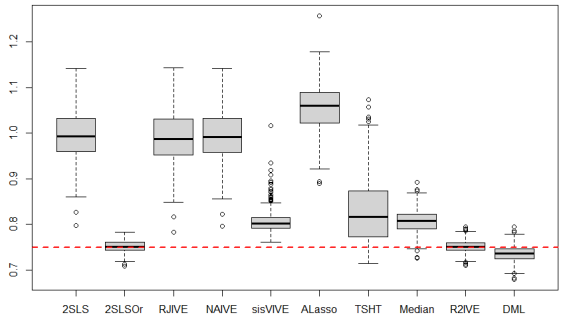


(g) $n = 1000, L_n = 100, c = 0.5$

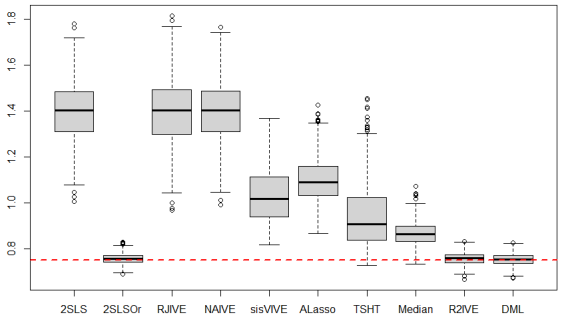


(h) $n = 500, L_n = 250, c = 0.5$

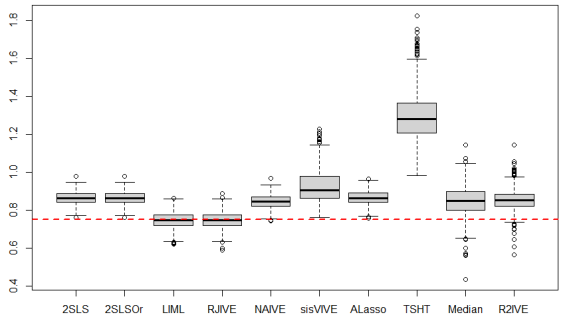
Figure 1: Boxplots of simulation settings in Sections 5.2-5.4



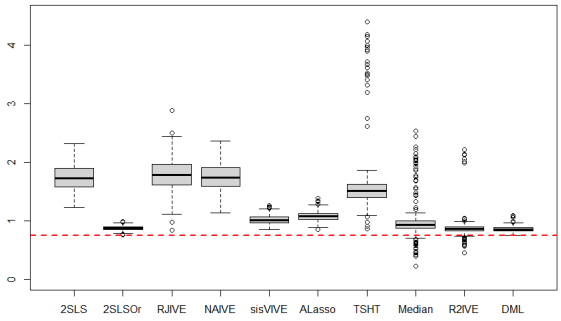
(a) $\gamma = 1$



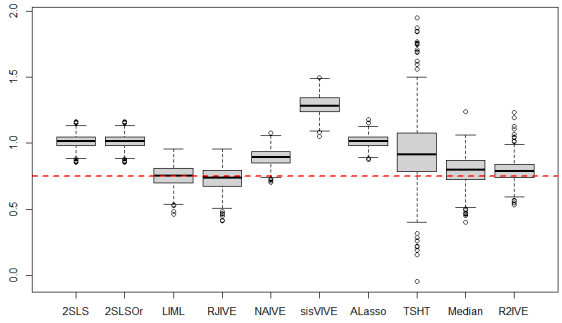
(b) $\gamma = 0.5$



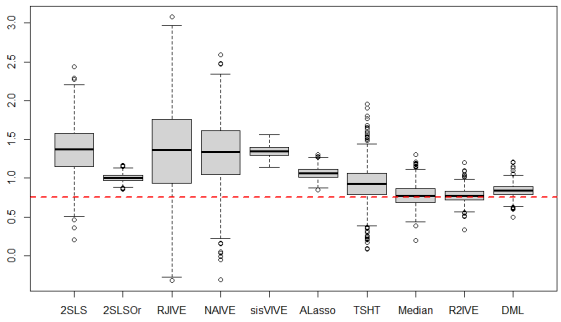
(c) "Many weak" setting, $s_C = 0$



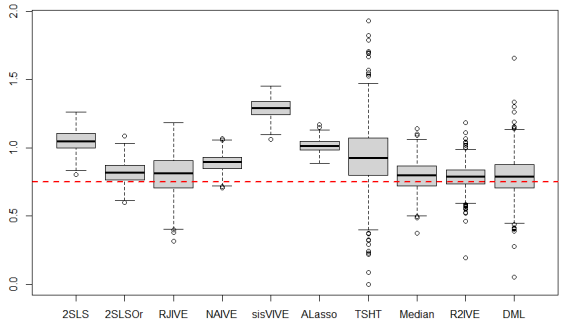
(d) "Many weak" setting, $s_C = 10$



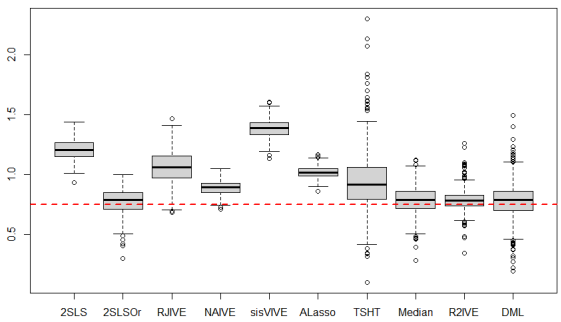
(e) Exp. decaying design for Eq. (5.2), $s_C = 0$



(f) Exp. decaying design for Eq. (5.2), $s_C = 10$



(g) Exp. decaying design for both Eqs. (5.1) and (5.2), $s_C = 90$



(h) Exp. decaying design for both Eqs. (5.1) and (5.2), $s_C = 95$

Figure 2: Boxplots of simulation settings in Sections 5.5-5.6

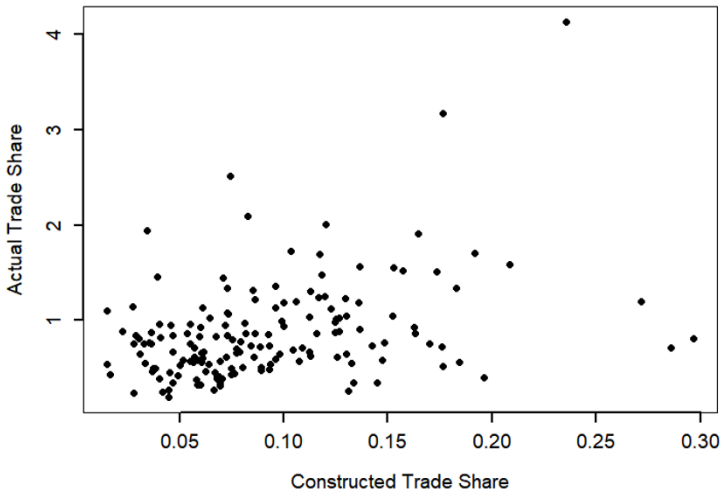


Figure 3: Scatter plot of real and constructed trade share