Confidence Intervals for Causal Effects with Invalid Instruments using Two-Stage Hard Thresholding

Zijian Guo
University of Pennsylvania, Philadelphia, USA
Hyunseung Kang
Stanford University, Stanford, USA
T. Tony Cai
University of Pennsylvania, Philadelphia, USA
Dylan S. Small
University of Pennsylvania, Philadelphia, USA

Summary.
The instrumental variable (IV) method is commonly used to estimate the causal effect of a treatment on an outcome by using IVs that satisfy the assumptions of association with treatment, no direct effect on the outcome and ignorability. A major challenge in IV analysis is to find said IVs, but typically one is unsure of whether all of the putative IVs are in fact valid (i.e. satisfy the assumptions). We propose a general inference procedure that provides honest inference in the presence of invalid IVs, even after controlling for a large number of covariates. The key step of our method is a novel selection procedure, which we call Two-Stage Hard Thresholding (TSHT), where we use hard thresholding to select the set of non-redundant instruments in the first stage and subsequently use hard thresholding to select the valid instruments in the second stage using the thresholding from the first stage. TSHT allows us to not only select invalid IVs, but also provides honest confidence intervals of the treatment effect at $\sqrt{n}$ rate. We establish asymptotic properties of our procedure and demonstrate that our procedure performs well in simulation studies compared to traditional IV methods, especially when the instruments are invalid.

Keywords: Causal effect; confidence interval; exclusion restriction; high-dimensional covariates; invalid instruments; treatment effect.

1. Introduction

1.1. Motivation: Invalid Instruments Even After Controlling for (Potentially Many) Confounders

Instrumental variables (IV) analysis is a popular method to deduce causal effects in the presence of unmeasured confounding. An IV analysis requires variables called instruments that (A1) are related to the exposure (A2) have no direct pathway to the outcome and (A3) are not related to unmeasured variables that affect the exposure and the outcome (see Section 2.2 for details). Variables that satisfy these

assumptions are referred to as valid instruments. A major challenge in IV analysis is to find valid instruments

In practice, it is often the case that potential candidate instruments become more plausible as valid instruments after controlling for some, possibly high dimensional, covariates (see Hernán and Robins [2006] and Baiocchi et al. [2014] for discussion on control for covariates for an instrument to be valid). For example, a long-standing interest in economics is the causal effect of education on earnings and often, IV analysis is used to deduce the effect [Angrist and Krueger, 1991, Card, 1993, 1999]. A popular instrument in this analysis is a person's proximity to a college when growing up [Card, 1999, 1993]. However, proximity to a college may be related to a person's socioeconomic status, characteristics of a person's high school and other covariates that may affect a person's earnings. Thus, these covariates need to be controlled for in order for proximity to college to be a valid IV and with the growing trend toward collecting large data sets with many variables, this approach of finding instrumental variables that are valid after conditioning on covariates has increasing promise [Hernán and Robins, 2006, Swanson and Hernán, 2013, Baiocchi et al., 2014, Varian, 2014, Imbens, 2014].

Yet, despite the promise that large data sets may bring in terms of finding valid instruments by conditioning on potentially many covariates, some IVs may still turn out to be invalid and subsequent analysis assuming that all the IVs are valid after conditioning can be misleading [Murray, 2006]. For example, suppose for studying the causal effect of education on earnings, we used proximity as an IV and to make sure the IV satisfies (A3), we control for confounders like high school test scores of the student, high school size, individual's genetic makeup, family education, and family's socioeconomic status. But, if living close to college had other benefits beyond getting more education, say by being exposed to many programs available to high school students for job preparation and employers who come to the area to discuss employment opportunities for college students, then the IV, proximity to college, can directly affect individual's earning potential and violate (A2) [Card, 1999]. This problem is also widely prevalent in other applications of instrumental variables, most notably in Mendelian randomization [Davey Smith and Ebrahim, 2003, 2004] where the instruments are genetic in nature and some instruments are likely to be invalid due to having pleiotropic effects [Lawlor et al., 2008, Burgess et al., 2015].

This paper tackles the problem of constructing confidence intervals for causal effects that are robust to invalid instruments even after controlling for possibly high dimensional covariates.

1.2. Prior Work

In non-IV settings with many high dimensional covariates, Zhang and Zhang [2014], Javanmard and Montanari [2014], van de Geer et al. [2014], Belloni et al. [2014] and Cai and Guo [2016a] provide honest confidence intervals for a treatment effect. In IV settings with high dimensional covariates (or IVs), Gautier and Tsybakov [2011], Belloni et al. [2012], Fan and Liao [2014] and Chernozhukov et al. [2015] provide

honest confidence intervals for a treatment effect, under the assumption that all the IVs are valid after controlling for said covariates. In invalid IV settings, Kolesár et al. [2015] and Bowden et al. [2015] provide inferential methods for treatment effects. However, the method requires that the effects of the IVs on the treatment be orthogonal to their direct effects on the outcome, a stringent assumption. Bowden et al. [2016], Burgess et al. [2016], Kang et al. [2016b] and Windmeijer et al. [2016] also work on the invalid IV setting, but without making the stringent orthogonality assumption. Unfortunately, all these papers focuses on the low dimensional setting and some only work in the case where the IVs are completely uncorrelated/orthogonal to each other unless modifications are made [Bowden et al., 2015, Burgess et al., 2016]. Furthermore, all the previous work only provides a consistent estimator of the treatment effect without any theoretical guarantees on inference; in fact, one of the simplest consistent estimators in this setting, the median estimator [Bowden et al., 2016, Burgess et al., 2016, Windmeijer et al., 2016] has been shown to be consistent, but not $\sqrt{n}$ consistent [Windmeijer et al., 2016].

There are two major challenges in obtaining valid confidence intervals in our problem: (i) potentially high-dimensional covariates and (ii) the invalid IVs. The problem related with high-dimensional covariates can be dealt with by applying recent debiasing methods [Zhang and Zhang, 2014, Javanmard and Montanari, 2014, van de Geer et al., 2014, Cai and Guo, 2016a]. However, the general idea behind debiasing does not inherently resolve the invalid IV problem as even a single IV that is improperly assumed as valid while it is truly invalid can make these debiased estimates useless. To put it another way, debiasing as a method is only meant to asymptotically remove the bias of regression coefficients from ℓ_1 shrinkage estimators and to conduct proper inference on these de-biased coefficients. This methodological goal is different than in the invalid IV problem where the goal is to properly estimate a set of valid IVs, as even a single error of declaring an IV that is invalid as valid can lead to dishonest inference. In fact, the methodological challenge is not only to correctly select IVs, but also once selected, to do robust inference using the selected IVs.

1.3. Our Contributions

Although there are existing methods for estimating the treatment effect in the presence of possibly invalid IVs, there is a paucity of procedures for selecting the set of valid instruments and forming confidence intervals for the treatment effects with theoretical coverage guarantees. In this paper, we propose a novel two-stage hard thresholding (TSHT) procedure to estimate the set of valid instruments and form confidence intervals with theoretical coverage guarantee. As the name suggests, a key component of TSHT is the two sequential steps of hard-thresholding procedures common in high dimensional inference [Donoho and Johnstone, 1994, Donoho, 1995] to simultaneously allow for invalid IVs and endogeneity of the treatment. Specifically, in the first thresholding stage, we select non-redundant IVs (see Definition 2 for details) and in the second thresholding stage, we use the thresholded estimates from the first thresholding step as pilot estimates to guide the selection of the set

of valid instruments; see Section 3.3 for details. Using our two-stage variant of thresholding properly accounts for the selection of IVs and leads to $1/\sqrt{n}$ rate confidence intervals with desired coverage in both low and high dimensional settings where invalid IVs are present and without knowing a priori which of these IVs are invalid. Also, for the low dimensional covariate setting, our procedure is the first to have theoretical guarantees that it performs as well asymptotically as the oracle procedure that knows which instruments are valid. For the high dimensional covariate setting, our procedure is the first available procedure for forming confidence intervals with desired coverage when there may be invalid IVs.

The outline of the paper is as follows. After describing the model setup in Section 2, we formulate our TSHT procedure in Section 3. In Section 4, we develop the theoretical properties of our procedure. In Section 5, we investigate the performance of our procedure in a large simulation study and find that our confidence interval performs very similarly with respect to the oracle even if some of the underlying theoretical assumptions made in Section 4 are violated (see Sections 4 and 5 for details). In Section 6, we present an empirical study where we revisit the question of the causal effect of years of schooling on income using data from the Wisconsin Longitudinal Study. We provide conclusions and discussion in Section 7.

2. Model

2.1. Notation

To define causal effects, the potential outcome approach Neyman [1923], Rubin [1974] for instruments laid out in Holland [1988] is used. For each individual $i \in \{1, \ldots, n\}$, let $Y_i^{(d,\mathbf{z})} \in \mathbb{R}$ be the potential outcome if the individual were to have exposure $d \in \mathbb{R}$ and instruments $\mathbf{z} \in \mathbb{R}^{p_z}$. Let $D_i^{(\mathbf{z})} \in \mathbb{R}$ be the potential exposure if the individual had instruments $\mathbf{z} \in \mathbb{R}^{p_z}$. For each individual, only one possible realization of $Y_i^{(d,\mathbf{z})}$ and $D_i^{(\mathbf{z})}$ is observed, denoted as Y_i and D_i, respectively, based on his observed instrument values $\mathbf{Z}_{i\cdot} \in \mathbb{R}^{p_z}$ and exposure D_i. We also denote pre-instrument covariates for each individual i as $\mathbf{X}_{i\cdot} \in \mathbb{R}^{p_x}$. In total, n sets of outcome, exposure, and instruments, denoted as $(Y_i, D_i, \mathbf{Z}_{i\cdot}, \mathbf{X}_{i\cdot})$, are observed in an i.i.d. fashion.

We denote $\mathbf{Y} = (Y_1, \ldots, Y_n)$ to be an n-dimensional vector of observed outcomes, $\mathbf{D} = (D_1, \ldots, D_n)$ to be an n-dimensional vector of observed exposures/treatment, $\mathbf{Z}$ to be a n by p_z matrix of instruments where row i consists of $\mathbf{Z}_{i\cdot}$, and $\mathbf{X}$ to be an n by p_x matrix of covariates where row i consists of $\mathbf{X}_{i\cdot}$. Let $\mathbf{W}$ be an n by $p = p_z + p_x$ matrix where $\mathbf{W}$ is a result of concatenating the matrices $\mathbf{Z}$ and $\mathbf{X}$ and $\mathbf{\Sigma}^* = \mathbf{E}(\mathbf{W}_{i\cdot}\mathbf{W}_{i\cdot}^\intercal)$ is positive definite. For any vector $\mathbf{v} \in \mathbb{R}^p$, let $\mathbf{v}_j$ denote the jth element of $\mathbf{v}$. Let $\|\mathbf{v}\|_1$, $\|\mathbf{v}\|_2$, and $\|\mathbf{v}\|_\infty$ denote the usual $1, 2$ and ∞-norms, respectively. Let $\|\mathbf{v}\|_0$ denote the number of non-zero elements in $\mathbf{v}$ and $\text{supp}(\mathbf{v}) \subseteq \{1, \ldots, p\}$, is defined as $\{j : \mathbf{v}_j \neq 0\}$.

For any n by p matrix $\mathbf{M} \in \mathbb{R}^{n \times p}$, we denote the (i, j) element of matrix $\mathbf{M}$ as M_{ij}, the ith row as $\mathbf{M}_{i\cdot}$, and the jth column as $\mathbf{M}_{\cdot j}$. Let $\mathbf{M}^\intercal$ be the transpose of $\mathbf{M}$ and $\|\mathbf{M}\|_\infty$ represent the element-wise matrix sup norm of matrix $\mathbf{M}$. For a

sequence of random variables X_n, we use $X_n \xrightarrow{p} X$ and $X_n \xrightarrow{d} X$ to represent that X_n converges to X in probability and in distribution, respectively. Finally, for any two sequences a_n and b_n, we will write $a_n \gg b_n$ if $\limsup \frac{b_n}{a_n} = 0$ and write $a_n \ll b_n$ if $b_n \gg a_n$. Also, for a set J, $|J|$ denotes its cardinality.

2.2. Model and Instrumental Variables Assumptions

We consider the Additive LInear, Constant Effects (ALICE) model of [Holland, 1988] and extend it to allow for multiple valid and possibly invalid instruments as in Small [2007] and Kang et al. [2016b]. For two possible values of the exposure d', d and instruments $\mathbf{z}', \mathbf{z}$, we assume the following potential outcomes model

$$Y_i^{(d',\mathbf{z}')} - Y_i^{(d,\mathbf{z})} = (\mathbf{z}' - \mathbf{z})^\intercal \boldsymbol{\kappa}^* + (d' - d)\beta^*, \quad \mathbf{E}(Y_i^{(0,\mathbf{0})} \mid \mathbf{Z}_{i.}, \mathbf{X}_{i.}) = \mathbf{Z}_{i.}^\intercal \boldsymbol{\eta}^* + \mathbf{X}_{i.}^\intercal \boldsymbol{\phi}^* \quad (1)$$

where $\boldsymbol{\kappa}^*, \beta^*, \boldsymbol{\eta}^*$, and $\boldsymbol{\phi}^*$ are unknown parameters. The parameter β^* represents the causal parameter of interest, the causal effect (divided by $d' - d$) of changing the exposure from d' to d on the outcome. The parameter $\boldsymbol{\phi}^*$ represents the impact of covariates on the baseline potential outcome $Y_i^{(0,\mathbf{0})}$. The parameter $\boldsymbol{\kappa}^*$ represents violation of (A2), the direct effect of the instruments on the outcome. If (A2) holds, then $\boldsymbol{\kappa}^* = 0$. The parameter $\boldsymbol{\eta}^*$ represents violation of (A3), the presence of unmeasured confounding between the instrument and the outcome. If (A3) holds, then $\boldsymbol{\eta}^* = 0$.

Let $\boldsymbol{\pi}^* = \boldsymbol{\kappa}^* + \boldsymbol{\eta}^*$ and $\epsilon_{i1} = Y_i^{(0,\mathbf{0})} - \mathbf{E}(Y_i^{(0,\mathbf{0})} \mid \mathbf{Z}_{i.}, \mathbf{X}_{i.})$. When we combine equation (1) along with the definition of ϵ_{i1}, the observed data model becomes

$$Y_i = \mathbf{Z}_{i.}^\intercal \boldsymbol{\pi}^* + D_i\beta^* + \mathbf{X}_{i.}^\intercal \boldsymbol{\phi}^* + \epsilon_{i1}, \quad \mathbf{E}(\epsilon_{i1} \mid \mathbf{Z}_{i.}, \mathbf{X}_{i.}) = 0 \quad (2)$$

and we denote $\sigma^2 = \mathrm{Var}(\epsilon_{i1} | \mathbf{Z}_{i.}, \mathbf{X}_{i.})$. The observed model is also known as the under-identified single-equation linear model in econometrics (page 83 of Wooldridge [2010]). This model is not a usual regression model because D_i might be correlated with ϵ_{i1}. In particular, the parameter β^* measures the causal effect of changing D on Y rather than an association. Also, the parameter $\boldsymbol{\pi}^*$ in model (2) combines both the violation of (A2), represented by $\boldsymbol{\kappa}^*$, and the violation of (A3), represented by $\boldsymbol{\eta}^*$. If both (A2) and (A3) are satisfied for IVs, typically referred to as valid IVs [Murray, 2006], then $\boldsymbol{\kappa}^* = \boldsymbol{\eta}^* = 0$ and $\boldsymbol{\pi}^* = 0$. Hence, $\boldsymbol{\pi}^*$ captures invalid IVs, i..e the violations of (A2) and (A3). We formalize this notion with the following definition.

Definition 1. *Suppose we have p_z candidate instruments along with the models* (1)–(2)*. We say that instrument $j = 1, \ldots, p_z$ is valid, i.e. satisfies* (A2) *and* (A3)*, if $\boldsymbol{\pi}_j^* = 0$.*

We also assume a linear association/observational model between the endogenous variable D_i, the instruments $\mathbf{Z}_{i.}$, and the covariates $\mathbf{X}_{i.}$.

$$D_i = \mathbf{Z}_{i.}^\intercal \boldsymbol{\gamma}^* + \mathbf{X}_{i.}^\intercal \boldsymbol{\psi}^* + \epsilon_{i2}, \quad \mathbf{E}(\epsilon_{i2} | \mathbf{Z}_{i.}, \mathbf{X}_{i.}) = 0 \quad (3)$$

Each element $\boldsymbol{\gamma}_j^*$ is the partial correlation between the jth instrument and D. The parameter $\boldsymbol{\psi}^*$ represents the association between the covariates and D_i. Also, unlike the models (1)-(2), we do not need a causal model between D_i, $\mathbf{Z}_{i\cdot}$, and $\mathbf{X}_{i\cdot}$; only the association model (3) is sufficient for our method. Finally, for notation, we let $s_{\text{z2}} = \|\boldsymbol{\pi}^*\|_0$, $s_{\text{x2}} = \|\boldsymbol{\phi}^*\|_0$, $s_{\text{z1}} = \|\boldsymbol{\gamma}^*\|_0$, $s_{\text{x1}} = \|\boldsymbol{\psi}^*\|_0$ and $s = \max\{s_{\text{z2}}, s_{\text{x2}}, s_{\text{z1}}, s_{\text{x1}}\}$. Finally, in both models, the instruments and the covariates are exogenous to the error terms; see Wooldridge [2010] for textbook discussion on exogeneity.

Based on model (3), we can define a set of instruments that satisfy (A1), or sometimes referred to as non-redundant instruments in the econometrics literature [Cheng and Liao, 2015].

DEFINITION 2. *Suppose we have p_{z} candidate instruments along with the model* (3)*. We say that instrument $j = 1, \ldots, p_z$ satisfies* (A1)*, or is a non-redundant IV, if $\boldsymbol{\gamma}_j^* \neq 0$ and denote $\mathcal{S}^*$ to be the set of these instruments.*

Typically, satisfying (A1) has been defined in a global sense where (A1) is satisfied if $\boldsymbol{\gamma}^* \neq 0$ [Wooldridge, 2010]. However, this global definition can be misleading in the presence of multiple candidate instruments. For example, it is possible that $\gamma_1^* \gg 0$ while $\gamma_j^* = 0$ for all $j \neq 1$ so that only the first instrument has an effect on the exposure while the rest do not. Using the global definition would imply that all the p_z instruments satisfy (A1) while Definition 2 makes it explicit and, perhaps less ambiguous, that it is only the first instrument $j = 1$ that satisfies (A1). Nevertheless, both the traditional global definition and Definition 2 are equivalent if $\gamma_j^* \neq 0$ for all j, that is where we only include relevant instruments, which is typically the case in practice and is the scenario studied by Kang et al. [2016b]. Finally, when there is only one candidate instrument so that $p_{\text{z}} = 1$, both definitions are equivalent to the definition presented in Holland [1988] and both become a special case of the definition presented in Angrist et al. [1996] under an additive, linear, constant effects model. In short, Definition 2 agrees with most definitions of satisfying (A1) in the literature.

Combining Definitions 1 and 2, we can formally define the usual three core conditions, i.e. (A1)-(A3), that define instruments.

DEFINITION 3. *Suppose we have p_{z} candidate instruments along with the models* (1)–(3)*. We say that the $Z_{\cdot j}$, $j = 1, \ldots, p_z$, is an instrument if* (A1) − (A3) *are satisfied, i.e. if $\pi_j^* = 0$ and $\gamma_j^* \neq 0$. Let $\mathcal{V}^*$ be the set of instruments.*

When there is only one instrument, $p_z = 1$, Definition 3 of an instrument is identical to the definition of an instrument in Holland [1988]. Specifically, Definition 2 satisfies assumption (A1) that the instrument is related to the exposure. Also, assumption (A2), the exclusion restriction, which means $Y_i^{(d,\mathbf{z})} = Y_i^{(d,\mathbf{z}')}$ for all $d, \mathbf{z}, \mathbf{z}'$, is equivalent to $\boldsymbol{\kappa}^* = 0$ and assumption (A3), no unmeasured confounding, which means $Y_i^{(d,\mathbf{z})}$ and $D_i^{(\mathbf{z})}$ are independent of $Z_{i\cdot}$ for all d and $\mathbf{z}$, is equivalent to $\boldsymbol{\eta}^* = 0$, implying $\boldsymbol{\pi}^* = \boldsymbol{\kappa}^* + \boldsymbol{\eta}^* = \mathbf{0}$. Definition 3 is also a special case of the definition of an instrument in Angrist et al. [1996] where here we assume the model is additive, linear, and has a constant treatment effect β^*. Hence, when multiple

instruments, $p_z > 1$, are present, our models (1)–(3) and Definition 3 can be viewed as a generalization of the definition of instruments in Holland [1988].

Note that the models presented above are commonly used in applications of IVs in econometrics [Wooldridge, 2010] and applications of IVs in genetic epidemiology and Mendelian randomization [Didelez and Sheehan, 2007]. However, we generalize these widely used models in two important ways: (i) the model in (2) allows for possibly invalid instruments and (ii) we allow the number of covariates p_x (and even the number of instruments p_z) to be larger than the sample size n.

3. Confidence Interval Estimation via Two-Stage Hard Thresholding

3.1. General Approach

The construction of our confidence interval can be broken down into two parts. The first part, detailed in Section 3.2, is estimating ITT effects based on the models (2) and (3). As we will see, the first part primarily deals with the problem posed by potentially high dimensional covariates, specifically the bias that comes from penalized estimators for high dimensional regression. The second part, which is elaborated in Section 3.3, tackles the heart of the problem in this paper, the presence of invalid IVs even after conditioning on high dimensional controls. Here, we take a novel two-stage hard thresholding approach to correctly select the valid IVs. Specifically, in the first step, we estimate the set of IVs that satisfy (A1) and in the second step, we use these IVs as initial guides to find IVs that satisfy (A2) and (A3) using the estimated set in the first step. Combining the two parts gives our confidence interval estimation procedure and is summarized in Procedure 1.

Procedure 1 Two-Stage Hard Thresholding (TSHT) for Confidence Interval for β^* under Invalid IVs with High-Dimensional Covariates

---

Input: Outcome $\mathbf{Y}$, treatment $\mathbf{D}$, instrument $\mathbf{Z}$, covariates $\mathbf{X}$, significance level α

STEP 1: Estimate ITT effects (i.e. $\widetilde{\mathbf{\Gamma}}$, $\widetilde{\boldsymbol{\gamma}}$) via debiased scaled Lasso in (6)-(9)

STEP 2: Select valid IVs (i.e. $\widetilde{V}$) via two-stage hard thresholding

STEP 2a: Estimate IVs satisfying (A1) (i.e. $\widetilde{\mathcal{S}}$) via hard thresholding $\widetilde{\boldsymbol{\gamma}}$ in (10)

STEP 2b: For each IV satisfying (A1) (i.e. $j \in \widetilde{\mathcal{S}}$), estimate IVs satisfying (A2) and (A3) via hard-thresholding $\widetilde{\boldsymbol{\pi}}^{[j]}$ in (11)-(12)

STEP 3: Combine STEP 1 and STEP 2 via (13)-(15) to obtain confidence interval

Output: $1 - \alpha$ Confidence interval for β^*

Procedure 1 provides a general recipe to construct confidence intervals in the presence of invalid IVs and high dimensional covariates. The key step in the procedure is STEP 2, where we utilize two-stage hard thresholding, to deal with the problem posed by invalid IVs; as such, we call our procedure TSHT procedure, akin to the acronym for two-stage least squares (TSLS) procedure in IV, arguably the most popular IV estimator in the literature. We also note that the Procedure 1 as stated can handle (i) low dimensional covariates, (ii) high dimensional covariates, and (iii) settings with all IVs having no direct effect and no unmeasured confound-

ing, which is unrealistic in practice, and can still obtain valid confidence intervals. However, depending on particular data sets one may have, the procedure can be modified for simplicity and, in some cases, efficiency; Sections 3.5 and 3.6 discusses these cases in detail.

3.2. Estimating ITT Effects

The first part of the confidence interval procedure involves estimation of ITT effects. Specifically, given the observed models (2) and (3), we can write reduced-forms models where both models are only functions of $\mathbf{Z}_{i.}$ and $\mathbf{X}_{i.}$

$$Y_i = \mathbf{Z}_{i.}^\intercal \mathbf{\Gamma}^* + \mathbf{X}_{i.}^\intercal \Psi^* + e_{i1} \tag{4}$$
$$D_i = \mathbf{Z}_{i.}^\intercal \boldsymbol{\gamma}^* + \mathbf{X}_{i.}^\intercal \boldsymbol{\psi}^* + \epsilon_{i2} \tag{5}$$

Here, $\mathbf{\Gamma}^* = \beta^* \boldsymbol{\gamma}^* + \boldsymbol{\pi}^*$ and $\Psi^* = \boldsymbol{\phi}^* + \beta^* \boldsymbol{\psi}^*$ are the parameters of the reduced-form model with $\mathbf{\Gamma}^*$ representing the ITT effect of the instruments on the outcome and $\boldsymbol{\gamma}^*$ representing the ITT effect of the instruments on the exposure. The term $e_{i1} = \beta^* \epsilon_{i2} + \epsilon_{i1}$ is the reduced-form error term in (4). The errors have the property that $\mathbf{E}(e_{i1}|\mathbf{Z}_{i.}, \mathbf{X}_{i.}) = 0$ and $\mathbf{E}(\epsilon_{i2}|\mathbf{Z}_{i.}, \mathbf{X}_{i.}) = 0$ with the variances $\Theta_{11}^* = \mathrm{Var}(e_{i1}|\mathbf{Z}_{i.}, \mathbf{X}_{i.})$, $\Theta_{22}^* = \mathrm{Var}(\epsilon_{i2}|\mathbf{Z}_{i.}, \mathbf{X}_{i.})$, and $\Theta_{12}^* = \mathrm{Cov}(e_{i1}, \epsilon_{i2}|\mathbf{Z}_{i.}, \mathbf{X}_{i.})$. Thus, each equation in the reduced-form model is a usual (high dimensional) regression model with (high dimensional) covariates $\mathbf{Z}_{i.}$ and $\mathbf{X}_{i.}$ and outcomes Y_i and D_i, respectively.

There are many methods in the literature to estimate the parameters of high dimensional regression models like the reduced-form models in (4) and (5). One approach is the scaled Lasso estimator proposed by Sun and Zhang [2012],

$$\{\widehat{\mathbf{\Gamma}}, \widehat{\mathbf{\Psi}}, \widehat{\Theta}_{11}\}$$
$$= \underset{\mathbf{\Gamma} \in \mathbb{R}^{p_z}, \mathbf{\Psi} \in \mathbb{R}^{p_x}, \Theta_{11} \in \mathbb{R}^+}{\operatorname{argmin}} \frac{\|\mathbf{Y} - \mathbf{Z}\mathbf{\Gamma} - \mathbf{X}\mathbf{\Psi}\|_2^2}{2n\sqrt{\Theta_{11}}} + \frac{\sqrt{\Theta_{11}}}{2} + \frac{\lambda_0}{\sqrt{n}} \left(\sum_{j=1}^{p_z} \|\mathbf{Z}_{.j}\|_2 |\Gamma_j| + \sum_{j=1}^{p_x} \|\mathbf{X}_{.j}\|_2 |\Psi_j| \right) \tag{6}$$

for the reduced model in (4) and

$$\{\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\psi}}, \widehat{\Theta}_{22}\}$$
$$= \underset{\mathbf{\Gamma} \in \mathbb{R}^{p_z}, \mathbf{\Psi} \in \mathbb{R}^{p_x}, \Theta_{22} \in \mathbb{R}^+}{\operatorname{argmin}} \frac{\|\mathbf{D} - \mathbf{Z}\boldsymbol{\gamma} - \mathbf{X}\boldsymbol{\psi}\|_2^2}{2n\sqrt{\Theta_{22}}} + \frac{\sqrt{\Theta_{22}}}{2} + \frac{\lambda_0}{\sqrt{n}} \left(\sum_{j=1}^{p_z} \|\mathbf{Z}_{.j}\|_2 |\gamma_j| + \sum_{j=1}^{p_x} \|\mathbf{X}_{.j}\|_2 |\psi_j| \right) \tag{7}$$

for the reduced model in (5). The term λ_0 in both estimation problems (6) and (7) represents the penalty term in the scaled Lasso estimator and we choose $\lambda_0 = \sqrt{a_0 \log p / n}$ for some constant $a_0 > 2$; in practice, we find that setting $a_0 = 2$ or 2.05 works well. Also, we can estimate Θ_{12}^* from the estimation problems (6) and (7) by $\widehat{\Theta}_{12} = 1/n \left(\mathbf{Y} - \mathbf{Z}\widehat{\mathbf{\Gamma}} - \mathbf{X}\widehat{\mathbf{\Psi}} \right)^\intercal \left(\mathbf{D} - \mathbf{Z}\widehat{\boldsymbol{\gamma}} - \mathbf{X}\widehat{\boldsymbol{\psi}} \right)$.

Unfortunately, most penalized estimators for high dimensional regression problems are biased and the scaled Lasso estimators are no exception. In our case, using the estimates, say $\widehat{\mathbf{\Gamma}}$ and $\widehat{\boldsymbol{\gamma}}$, are biased for the parameters that they estimate $\mathbf{\Gamma}^*$ and

$\boldsymbol{\gamma}^*$. Thankfully, recent works by Zhang and Zhang [2014], Javanmard and Montanari [2014], van de Geer et al. [2014] and Cai and Guo [2016a] allow us to debias these biased estimates. Specifically, let $\mathbf{W}$ be the concatenated matrix of the instruments $\mathbf{Z}$ and the covariates $\mathbf{X}$. Suppose we solve p_z optimization problems where the solution to each p_z optimization problem, denoted as $\widehat{\mathbf{u}}^{[j]} \in \mathbb{R}^p$, $j = 1, \ldots, p_z$, is

$$\widehat{\mathbf{u}}^{[j]} = \underset{\mathbf{u}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n}\|\mathbf{W}\mathbf{u}\|_2^2 \quad \text{s.t.} \quad \|\frac{1}{n}\mathbf{W}^\intercal\mathbf{W}\mathbf{u} - \mathbf{I}_{\cdot j}\|_\infty \leq \lambda_n, \tag{8}$$

with $\mathbf{I}_{\cdot j}$ denoting the j-th column of the identity matrix $\mathbf{I}$. The tuning parameter λ_n is chosen to be $12M_1^2\sqrt{\log p/n}$ with M_1 defined as the largest eigenvalue of $\boldsymbol{\Sigma}^*$. Let $\widehat{\mathbf{U}}$ denote the concatenation of the p_z solutions to the optimization problem, i.e. $\widehat{\mathbf{U}} = (\widehat{\mathbf{u}}^{[1]}, \ldots, \widehat{\mathbf{u}}^{[p_z]})^\intercal$. Then, the debiased estimates of $\widehat{\boldsymbol{\Gamma}}$ and $\widehat{\boldsymbol{\gamma}}$, denoted as $\widetilde{\boldsymbol{\Gamma}}$ and $\widetilde{\boldsymbol{\gamma}}$, are

$$\widetilde{\boldsymbol{\Gamma}} = \widehat{\boldsymbol{\Gamma}} + \frac{1}{n}\widehat{\mathbf{U}}\mathbf{W}^\intercal\left(\mathbf{Y} - \mathbf{Z}\widehat{\boldsymbol{\Gamma}} - \mathbf{X}\widehat{\boldsymbol{\Psi}}\right), \quad \widetilde{\boldsymbol{\gamma}} = \widehat{\boldsymbol{\gamma}} + \frac{1}{n}\widehat{\mathbf{U}}\mathbf{W}^\intercal\left(\mathbf{D} - \mathbf{Z}\widehat{\boldsymbol{\gamma}} - \mathbf{X}\widehat{\boldsymbol{\psi}}\right). \tag{9}$$

In short, we used scaled Lasso along with de-biasing methods on the reduced-form models to obtain de-biased estimates $\widetilde{\boldsymbol{\Gamma}}$ and $\widetilde{\boldsymbol{\gamma}}$ of the intent-to-treat effects of the instruments on the outcome and the exposure, respectively.

3.3. Two-Stage Hard Thresholding

The second part of the confidence interval procedure deals with the problem posed by invalid IVs. Specifically, we need to select valid IVs among p_z candidate IVs that satisfy all (A1)-(A3) assumptions, that is the set $\mathcal{V}^*$ in Definition 3. As discussed before, we do this by first, finding IVs that satisfy (A1), that is the set $\mathcal{S}^*$ in Definition 2 consisting of js where $\gamma_j^* \neq 0$, by thresholding the de-biased estimate $\widetilde{\boldsymbol{\gamma}}$

$$\widetilde{\mathcal{S}} = \left\{ j : |\widetilde{\gamma}_j| \geq \frac{\sqrt{\widehat{\Theta}_{22}}\|\mathbf{W}\widehat{\mathbf{u}}^{[j]}\|_2}{\sqrt{n}}\sqrt{\frac{a_0 \log p_\text{z}}{n}} \right\}. \tag{10}$$

where $\widetilde{\mathcal{S}}$ denotes an estimate of $\mathcal{S}^*$. The threshold is based on the noise level of $\widetilde{\gamma}_j$ in (9) (represented by $\sqrt{\widehat{\Theta}_{22}\|\mathbf{W}\widehat{\mathbf{u}}^{[j]}\|_2/n}$), adjusted by dimensionality of the instrument size (represented by $\sqrt{a_0 \log p_\text{z}}$).

The second thresholding step involves selecting IVs that satisfy (A2) and (A3). Specifically, by Definition 1, the set of instruments that satisfy (A2) and (A3) are those js where $\pi_j^* = 0$. Consequently, to estimate $\boldsymbol{\pi}^*$, we take each instrument j in $\widetilde{\mathcal{S}}$ that satisfy (A1) and we define $\widehat{\beta}^{[j]}$ to be a "pilot" estimate of β^* by using this IV and dividing the reduced-form estimates, i.e. $\widehat{\beta}^{[j]} = \widetilde{\Gamma}_j/\widetilde{\gamma}_j$, and $\widehat{\boldsymbol{\pi}}^{[j]}$ to be the estimate of $\boldsymbol{\pi}^*$ using this jth instrument's estimate of β^*, i.e. $\widehat{\boldsymbol{\pi}}^{[j]} = \widetilde{\boldsymbol{\Gamma}} - \widehat{\beta}^{[j]}\widetilde{\boldsymbol{\gamma}}$; we also construct corresponding pilot estimates of σ^2, i.e. $\widehat{\sigma^2}^{[j]} = \widehat{\Theta}_{11} + (\widehat{\beta}^{[j]})^2\widehat{\Theta}_{22} - 2\widehat{\beta}^{[j]}\widehat{\Theta}_{12}$. Then, for each $\widehat{\boldsymbol{\pi}}^{[j]}$ in $j \in \widetilde{\mathcal{S}}$, we threshold each element of $\widehat{\boldsymbol{\pi}}^{[j]}$ to create the thresholded

estimate $\widetilde{\boldsymbol{\pi}}^{[j]}$,

$$\widetilde{\pi}_k^{[j]} = \widehat{\pi}_k^{[j]} \mathbf{1}\left(k \in \widetilde{\mathcal{S}} \ \cap \ |\widehat{\pi}_k^{[j]}| \geq a_0 \sqrt{\widehat{\sigma^2}^{[j]}} \frac{\|\mathbf{W}(\widehat{\mathbf{u}}^{[k]} - \frac{\widetilde{\gamma}_k}{\widetilde{\gamma}_j}\widehat{\mathbf{u}}^{[j]})\|_2}{\sqrt{n}} \sqrt{\frac{\log p_{\mathrm{z}}}{n}} \right) \tag{11}$$

for all $1 \leq k \leq p_{\mathrm{z}}$. Each thresholded estimate $\widetilde{\boldsymbol{\pi}}^{[j]}$ is obtained by looking at the elements of the un-thresholded estimate, $\widehat{\boldsymbol{\pi}}^{[j]}$, and examining whether each element of it exceeds the noise threshold, denoted by the term $\sqrt{\widehat{\sigma^2}^{[j]}}\|\mathbf{W}(\widehat{\mathbf{u}}^{[k]} - \frac{\widetilde{\gamma}_k}{\widetilde{\gamma}_j}\widehat{\mathbf{u}}^{[j]})\|_2/n$, adjusting for the multiplicity of the selection procedure by the term $a_0\sqrt{\log p_{\mathrm{z}}}$. Among the $|\widetilde{\mathcal{S}}|$ candidate estimates of $\boldsymbol{\pi}^*$ based on each instrument in $\widetilde{\mathcal{S}}$, i.e. $\widetilde{\boldsymbol{\pi}}^{[j]}$, and we choose $\widetilde{\boldsymbol{\pi}}^{[j]}$ with most valid instruments, or equivalently choose $j^* \in \widetilde{\mathcal{S}}$ where $j^* = \operatorname{argmin} \|\widetilde{\boldsymbol{\pi}}^{[j]}\|_0$; if there is a non-unique solution, we choose $\widetilde{\boldsymbol{\pi}}^{[j]}$ with the smallest ℓ_1 norm, the closest convex norm of ℓ_0.

Intuitively, the second-stage thresholding selects the invalid IVs and valid IVs as follows. Among the $|\widetilde{\mathcal{S}}|$ pilot estimates $\widetilde{\boldsymbol{\pi}}^{[j]}$, the best estimate of π^* is the one that uses a valid IV from the set $\widetilde{\mathcal{S}}$. In particular, if the jth pilot estimate is actually based on a valid IV, then all the invalid IVs will be included in the support of the thresholded estimate $\widetilde{\boldsymbol{\pi}}^{[j]}$ because their π^* will be away from zero and all the valid instruments will be excluded from the support because their π^* are zero. On the other hand, if the jth pilot estimate is based on an invalid IV, the pilot estimate $\widetilde{\boldsymbol{\pi}}^{[j]}$ will be biased in the sense that the valid IVs will no longer have $\widetilde{\boldsymbol{\pi}}^{[j]}$ that will be thresholded to zero and most of the elements of $\widetilde{\boldsymbol{\pi}}^{[j]}$ will be away from zero. Consequently, many IVs will be declared invalid based on $\widetilde{\boldsymbol{\pi}}^{[j]}$ and when we minimize with respect to the number of non-zero elements of the vector, i.e. $\min \|\widetilde{\boldsymbol{\pi}}^{[j]}\|_0$ among all pilot estimates, we should be able to select the best estimate of π^*. We remark that the latter ℓ_0 minimization is reminiscent of Theorem 1 in Kang et al. [2016b] where a necessary and sufficient condition for identification of β^* under invalid instruments is by looking at the largest subset of valid instruments that converge on a unique (i.e. identified) value; the search for the largest subset of valid IVs is essentially a minimization of ℓ_0 norm, which counts the number of invalid IVs, and hence, there is some sense that our procedure is both sufficient and necessary way to estimate β^*.

Finally, we note that it is crucial to construct pilot estimates of π^* from the IVs in the first thresholding step, that is $\widetilde{\mathcal{S}}$, as each of these IVs represent strong IVs and have non-zero effects on the exposure; using IVs that are not in $\widetilde{\mathcal{S}}$ may lead to poor estimates of the direct effect on the outcome since a redundant instrument j, whose true γ_j^* is zero, can lead to a large, unstable value of $|\widetilde{\gamma}_k/\widetilde{\gamma}_j|$ and the threshold value in (11), which will make it difficult to distinguish truly invalid IVs from noise.

3.4. Confidence Interval Estimation

After the two thresholding steps, we estimate the set of valid instruments $\widetilde{\mathcal{V}} \subseteq \{1, \ldots, p_z\}$ as those elements of $\widetilde{\boldsymbol{\pi}}^{[j^*]}$ that are zero,

$$\widetilde{\mathcal{V}} = \widetilde{\mathcal{S}} \setminus \operatorname{supp}\left(\widetilde{\boldsymbol{\pi}}^{[j^*]}\right) \tag{12}$$

Then, using the estimated $\widetilde{\mathcal{V}}$, we obtain our estimate of β^*

$$\widehat{\beta} = \frac{\sum_{j\in\widetilde{\mathcal{V}}} \widetilde{\gamma}_j \widetilde{\Gamma}_j}{\sum_{j\in\widetilde{\mathcal{V}}} \widetilde{\gamma}_j^2}, \tag{13}$$

along with an estimate of its standard error

$$\widehat{\mathrm{V}} = \frac{\left\|\sum_{j\in\widetilde{\mathcal{V}}} \widetilde{\gamma}_j \frac{1}{\sqrt{n}} \mathbf{W}\widehat{\mathbf{u}}^{[j]}\right\|_2^2}{\left(\sum_{j\in\widetilde{\mathcal{V}}} \widetilde{\gamma}_j^2\right)^2} \widehat{\sigma^2} \quad \text{and} \quad \widehat{\sigma^2} = \widehat{\Theta}_{11} + \widehat{\beta}^2\widehat{\Theta}_{22} - 2\widehat{\beta}\widehat{\Theta}_{12}, \tag{14}$$

and the usual form for the confidence interval for β^*,

$$\left(\widehat{\beta} - z_{1-\alpha/2}\sqrt{\widehat{\mathrm{V}}/n}, \quad \widehat{\beta} + z_{1-\alpha/2}\sqrt{\widehat{\mathrm{V}}/n}\right), \tag{15}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

In the equations for $\widehat{\beta}$ and the standard error $\widehat{\mathrm{V}}$, we see some familiar expressions from the traditional IV literature. First, $\widehat{\beta}$ has the "correct" form in that if, by chance, we correctly estimated the set of valid instruments $\mathcal{V}^*$ and our debiased estimates of the reduced-form parameters, $\widetilde{\boldsymbol{\Gamma}}$ and $\widetilde{\boldsymbol{\gamma}}$, are perfect estimates of the reduced-form parameters, our estimate of β^* in (13) would become $\widehat{\beta} = \sum_{j\in\widetilde{\mathcal{V}}} \widetilde{\gamma}_j\widetilde{\Gamma}_j / \sum_{j\in\widetilde{\mathcal{V}}} \widetilde{\gamma}_j^2 = \sum_{j\in\mathcal{V}^*} \gamma_j^{*2}\beta^* / \sum_{j\in\mathcal{V}^*} \gamma_j^{*2} = \beta^*$. Hence, our estimator in (13) would identify β^*. Clearly, we would never have a perfect estimate of the set $\mathcal{V}^*$ or of the reduced-form parameters in finite sample and Section 4 describes the properties of our estimate $\widehat{\beta}$ under these uncertainties. Second, in the standard error formula in (14), the $\widehat{\Theta}_{11} + \widehat{\beta}^2\widehat{\Theta}_{22} - 2\widehat{\beta}\widehat{\Theta}_{12}$ is of the similar form to the usual IV estimator of σ^2, variance of the error term in our original model (2). But, our standard error estimator is scaled by terms that depend on the estimated set of valid instruments $\widetilde{\mathcal{V}}$.

3.5. Special Case of Procedure 1: Valid IVs After Controlling for High Dimensional Covariates

To better understand the components of our inference procedure 1, it is instructive to go through some specific cases of estimating β^* that is common in the literature as these special cases can greatly simplify the procedure and remove unnecessary components. The first case is when the instruments are assumed to be valid (i.e. no direct effect and no unmeasured confounding) after conditioning on high dimensional

covariates. This setup was considered in Gautier and Tsybakov [2011], Belloni et al. [2012], Fan and Liao [2014], Chernozhukov et al. [2015]. Under this case, our procedure doesn't have to go through STEP 2b, the estimation of π^*, illustrated in Section 3.3. Instead, we can simply replace STEP 2b with $\widetilde{\mathcal{V}} = \widetilde{\mathcal{S}}$ and the resulting estimator for β^* is

$$\widehat{\beta}_H = \frac{\sum_{j\in\widetilde{\mathcal{S}}} \widetilde{\boldsymbol{\gamma}}_j \widetilde{\boldsymbol{\Gamma}}_j}{\sum_{j\in\widetilde{\mathcal{S}}} \widetilde{\boldsymbol{\gamma}}_j^2}. \tag{16}$$

The corresponding confidence interval for β^* would be

$$\left(\widehat{\beta}_H - z_{1-\alpha/2}\sqrt{\widehat{\mathrm{V}}_H/n}, \quad \widehat{\beta}_H + z_{1-\alpha/2}\sqrt{\widehat{\mathrm{V}}_H/n}\right). \tag{17}$$

Here, $\widehat{\mathrm{V}}_H$ is $\widehat{\mathrm{V}}$ in (14) except we replace $\widetilde{\mathcal{V}} = \widetilde{\mathcal{S}}$ and $\widehat{\beta}$ with $\widehat{\beta}_H$.

3.6. Special Case of Procedure 1: Invalid Instruments After Controlling for Low Dimensional Covariates

The second special case worth examining is the problem of invalid instruments after controlling for low dimensional covariates. While the plausibility of candidate IVs adherence to assumptions (A1)-(A3), especially (A3), is higher with many covariates, the low dimensional setting has recently received much attention and is discussed in Bowden et al. [2015, 2016], Burgess et al. [2016], Kang et al. [2016a], and Windmeijer et al. [2016]. As we will see below, our procedure simplifies greatly and with a minor modification, our estimator, unlike the estimators proposed in said prior literature, achieves optimal performance.

Specifically, under the low-dimensional scenario, there is no need to use the debiased scaled lasso in STEP 1 of Procedure 1. Instead, we can replace STEP 1 with the simple ordinary least square (OLS) estimates of the reduced-forms, $(\widetilde{\boldsymbol{\Gamma}}, \widetilde{\boldsymbol{\Psi}})^\intercal = (\mathbf{W}^\intercal\mathbf{W})^{-1}\mathbf{W}^\intercal\mathbf{Y}$ and $(\widetilde{\boldsymbol{\gamma}}, \widetilde{\boldsymbol{\psi}})^\intercal = (\mathbf{W}^\intercal\mathbf{W})^{-1}\mathbf{W}^\intercal\mathbf{D}$, and of the covariance terms $\widehat{\Theta}_{11} = \left\|\mathbf{Y} - \mathbf{Z}\widehat{\boldsymbol{\Gamma}} - \mathbf{X}\widehat{\boldsymbol{\Psi}}\right\|_2^2/n$, $\widehat{\Theta}_{22} = \left\|\mathbf{D} - \mathbf{Z}\widehat{\boldsymbol{\gamma}} - \mathbf{X}\widehat{\boldsymbol{\psi}}\right\|_2^2/n$, and $\widehat{\Theta}_{12} = \left(\mathbf{Y} - \mathbf{Z}\widehat{\boldsymbol{\Gamma}} - \mathbf{X}\widehat{\boldsymbol{\Psi}}\right)^\intercal\left(\mathbf{D} - \mathbf{Z}\widehat{\boldsymbol{\gamma}} - \mathbf{X}\widehat{\boldsymbol{\psi}}\right)/n$. As a result of using OLS in STEP 1, we need to replace $\widehat{\mathbf{u}}^{[j]}$ from (8) with $\widehat{\mathbf{u}}^{[j]} = (\widehat{\boldsymbol{\Sigma}})^{-1}_{\cdot j}$, $\widehat{\boldsymbol{\Sigma}} = \mathbf{W}^\intercal\mathbf{W}/n$, and replace the log terms in our thresholds in (10) and (11) from $\sqrt{\log p_{\mathrm{z}}}$ to $\sqrt{\log n}$.

We can then proceed to use the estimator defined in (13). Alternatively, we can use a modified version of $\widehat{\beta}$, denoted as $\widehat{\beta}_E$, using the weighting matrix $\mathbf{A} = \widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathcal{V}},\widetilde{\mathcal{V}}} - \widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathcal{V}},\widetilde{\mathcal{V}}^c}\widehat{\boldsymbol{\Sigma}}^{-1}_{\widetilde{\mathcal{V}}^c,\widetilde{\mathcal{V}}^c}\widehat{\boldsymbol{\Sigma}}_{\widetilde{\mathcal{V}}^c,\widetilde{\mathcal{V}}}$

$$\widehat{\beta}_E = \frac{\widetilde{\boldsymbol{\gamma}}_{\widetilde{\mathcal{V}}}^\intercal A \widetilde{\boldsymbol{\Gamma}}_{\widetilde{\mathcal{V}}}}{\widetilde{\boldsymbol{\gamma}}_{\widetilde{\mathcal{V}}}^\intercal A \widetilde{\boldsymbol{\gamma}}_{\widetilde{\mathcal{V}}}} \tag{18}$$

along with the estimated standard error $\widehat{\mathrm{V}}_E = \widehat{\sigma}^2/\widetilde{\boldsymbol{\gamma}}_{\widetilde{\mathcal{V}}}^\intercal A\widetilde{\boldsymbol{\gamma}}_{\widetilde{\mathcal{V}}}$ where $\widehat{\sigma}^2 = \widehat{\Theta}_{11} + \widehat{\beta}_E^2\widehat{\Theta}_{22} - 2\widehat{\beta}_E\widehat{\Theta}_{12}$ and confidence interval

$$\left(\widehat{\beta}_E - z_{1-\alpha/2}\sqrt{\widehat{\mathrm{V}}_E/n}, \quad \widehat{\beta}_E + z_{1-\alpha/2}\sqrt{\widehat{\mathrm{V}}_E/n}\right). \tag{19}$$

Note that $\widehat{\beta}_E$ in (18) is reduced to $\widehat{\beta}$ in (13) by setting $A = \mathrm{I}$. As we will see in Section 4.1, our estimator $\widehat{\beta}_E$, compared to other estimators in prior work, achieves optimal performance in the sense that our performance is asymptotically identical to the TSLS estimator for β^* that knows which IVs are valid a priori, i.e. the set $\mathcal{V}^*$.

4. Theoretical Results

In this section, we investigate the properties of the confidence interval proposed in Procedure 1. We first consider in Section 4.1 the coverage property in the case of invalid IVs with low dimensional covariates where p_{x} and p_{z} are fixed. In Section 4.2, we establish the coverage property for the general case, invalid IVs even after controlling for many covariates.

4.1. Invalid IVs After Controlling for Low Dimensional Covariates

We state the following mild assumption commonly used in the invalid IV literature

(IN1) (50% Rule) The number of valid IVs is more than half of the number of non-redundant IVs, that is $|\mathcal{V}^*| > \frac{1}{2}|\mathcal{S}^*|$.

We denote the assumption as "IN" since the assumption is specific to the case of invalid IVs. Assumption (IN1) is the generalization of the 50% rule in Kang et al. [2016b] and Han [2008] in the presence of possibly redundant IVs. In a nutshell, (IN1) states that if the number of invalid instruments is not too large, then we can detect the invalid IVs from valid IVs, without knowing a priori which IVs are valid or invalid; see Kang et al. [2016b] for a detailed discussion of this assumption and how this type of proportion-based assumption is a necessary component for identification of model parameters under invalid instruments.

Under the 50% assumption alone, Theorem 1 states that we can show that our procedure produces confidence intervals with desired coverage and optimal length in low dimensional settings where p_{x} and p_{z} are fixed.

THEOREM 1. *Suppose that the assumption* (IN1) *holds. Then the following property holds for the estimator* $\widehat{\beta}_E$ *defined in* (18)*,*

$$\sqrt{n}\left(\widehat{\beta}_E - \beta^*\right) \xrightarrow{d} N\left(0, \frac{\sigma^2}{\boldsymbol{\gamma}_{\mathcal{V}^*}^{*\mathsf{T}}\left(\boldsymbol{\Sigma}^*_{\mathcal{V}^*\mathcal{V}^*} - \boldsymbol{\Sigma}^*_{\mathcal{V}^*(\mathcal{V}^*)^C}\boldsymbol{\Sigma}^{*-1}_{(\mathcal{V}^*)^C(\mathcal{V}^*)^C}\boldsymbol{\Sigma}^*_{(\mathcal{V}^*)^C\mathcal{V}^*}\right)\boldsymbol{\gamma}^*_{\mathcal{V}^*}}\right). \tag{20}$$

Consequently, the confidence interval given in (19) *has asymptotically coverage probability* $1-\alpha$*, i.e.,*

$$\mathbf{P}\left\{\beta \in \left(\widehat{\beta}_E - z_{1-\alpha/2}\sqrt{\widehat{\mathrm{V}}_E/n}, \quad \widehat{\beta}_E + z_{1-\alpha/2}\sqrt{\widehat{\mathrm{V}}_E/n}\right)\right\} \to 1-\alpha. \tag{21}$$

We note that the proposed estimator $\widehat{\beta}_E$ has the same asymptotic variance as the oracle TSLS estimator with the prior knowledge of $\mathcal{V}^*$, which is shown to be efficient under the homoskedastic variance assumption (Theorem 5.2 in Wooldridge [2010]); consequently, our confidence interval asymptotically performs like the oracle TSLS confidence interval and is of optimal length. But, unlike TSLS, we achieve this oracle performance without prior knowledge of $\mathcal{V}^*$. We also note the estimators proposed in prior work [Bowden et al., 2015, 2016, Burgess et al., 2016, Kang et al., 2016a, Windmeijer et al., 2016] do not achieve oracle performance and TSLS-like efficiency.

4.2. Invalid IVs After Controlling for High Dimensional Covariates

We now consider the coverage property for the general case, invalid IVs even after controlling for many confounders. We first introduce the usual regularity assumptions used in high-dimensional statistical inference [Bickel et al., 2009, Bühlmann and van de Geer, 2011, Cai and Guo, 2016a].

(R1) (Coherence): The matrix $\mathbf{\Sigma}^*$ satisfies $1/M_1 \leq \lambda_{\min}(\mathbf{\Sigma}^*) \leq \lambda_{\max}(\mathbf{\Sigma}^*) \leq M_1$ for some constant $M_1 > 1$ and has bounded sub-Gaussian norm.

(R2) (Normality): The error terms in (4) and (5) follow a bivariate normal distribution.

(R3) (Global IV Strength): The IVs are globally strong with $\|\boldsymbol{\gamma}^*_{\mathcal{V}^*}\|_2 = \sqrt{\sum_{j\in\mathcal{V}^*}\gamma_j^2} \geq \delta \gg s_{z1}\log p/\sqrt{n}$, where $\mathcal{V}^*$ is the set of valid IVs defined in Definition 3.

Assumption (R1) places a condition on the spectrum of the design matrix $\mathbf{W}$ and the tail distribution of $\mathbf{W}_{i,\cdot}$, which is related to the restricted eigenvalue condition in Bickel et al. [2009]. For simplicity, we also assume that the sub-Gaussian norm of $\mathbf{W}_{i\cdot}$ is upper bounded by M_1, that is, $\sup_{\boldsymbol{v}\in S^{p-1}}\sup_{q\geq 1}(\mathbf{E}|\boldsymbol{v}^\intercal\mathbf{W}_{i\cdot}|^q/q)^{\frac{1}{q}} \leq M_1$ where S^{p-1} is the unit sphere in $\mathbb{R}^p$; see Vershynin [2012] for details on sub-Gaussian random variables and bounds. Assumption (R2) states that the errors (e_{i1}, ϵ_{i2}) are bivariate normal. Here, we make the normality assumption out of simplicity, similar to the work on inference in weak IV literature where error terms are typically assumed to be normal (e.g. Section 2 of Moreira [2003] and Section 2.2.1 Andrews et al. [2007]). Finally, Assumption (R3) states that the global strength of instruments, measured by the ℓ_2 norm of $\boldsymbol{\gamma}^*$ among valid IVs $\mathcal{V}^*$, is bounded away from zero. This type of global strength assumption is commonly made in the IV literature under the guise as a concentration parameter, which is a measure of strength of the instrument (see Section 5 for details) and is the weighted ℓ_2 norm of $\boldsymbol{\gamma}^*_{\mathcal{V}^*}$, and is often referred to as "traditional/strong" asymptotics [Stock et al., 2002, Wooldridge, 2010]. Recent works by Belloni et al. [2012] and Chernozhukov et al. [2015], which considered the setting where all IVs were valid after conditioning on high dimensional covariates, also make this type of assumption, specifically condition SM in Belloni et al. [2012] and condition RF in the supplementary materials of Chernozhukov et al. [2015]. Essentially, both these works require $\|\boldsymbol{\gamma}^*\|_2$ to be

bounded away from zero by a constant and are actually stronger than our (R3). In practice, (R3) is satisfied so long as there is at least one IV that has a constant non-zero effect on the treatment, or a non-zero effect that doesn't diminish with sample size. However, if the IVs are arbitrary weak in the sense of Staiger and Stock [1997], then (R3), let alone the said assumptions in high dimensional valid IV literature [Belloni et al., 2012], do not hold, and we leave this as a future topic of research to deal with arbitrary weak IVs in invalid IV settings.

Section A in the supplementary materials shows that if the IVs are valid after conditioning on many covariates, then Asumptions (R1)-(R3) are sufficient for the confidence interval proposed in (17) to have correct coverage. However, when IVs are invalid after conditioning on said controls, we need to make two additional assumptions that are not in the usual high dimensional inference or instrumental variables literature and may be of theoretical interest in future work.

(IN2) (Individual IV Strength) For IVs in $\mathcal{S}^*$, $\min_{j\in\mathcal{S}^*}\left|\gamma_j^*\right| \geq \delta_{\min} \gg \sqrt{\log p/n}$.

(IN3) (Strong violation) Among IVs in the set $\mathcal{S}^* \setminus \mathcal{V}^*$, we have

$$\min_{j\in\mathcal{S}^*\setminus\mathcal{V}^*}\left|\frac{\pi_j^*}{\gamma_j^*}\right| \geq \frac{12(1+|\beta^*|)}{\delta_{\min}}\sqrt{\frac{M_1\log p_{\mathrm{z}}}{\lambda_{\min}(\mathbf{\Theta}^*)n}}. \tag{22}$$

Assumption (IN2) requires individual IV strength to be bounded away from zero so that all IVs in selected $\widetilde{\mathcal{S}}$ are strong. This assumption is needed primarily for cleaner technical exposition and our simulation studies in Section 5 and Section C in the supplementary materials demonstrate that (IN2) is largely unnecessary for our confidence interval to guarantee coverage. In the literature, (IN2) is similar to the "beta-min" condition assumption in high dimensional linear regression without IVs, with the exception that this condition is not imposed on our inferential quantity of interest, β^*. Also, (IN2) is different from Assumption (R3), where (R3) only requires the global IV strength to be bounded away from zero. Next, Assumption (IN3) requires the ratios π_j^*/γ_j^* for invalid IVs to be large and this assumption is needed to correctly identify IVs that violate (A2) and (A3). Specifically, for any IV with $|\pi_j^*/\gamma_j^*|$ being non-zero but small, it's difficult to distinguish such a weakly invalid IV from valid IVs where $\pi_j^*/\gamma_j^* = 0$. If a weakly invalid IV is mistakenly declared as valid, the bias from this mistake is of the order $\sqrt{\log p_{\mathrm{z}}/n}$, which has consequences, not for consistency of the point estimate, but for a $\sqrt{n}$ confidence interval; see Theorem 2 and Section 7 for more discussions.

With (R1)-(R3) and (IN1)-(IN3), our general Procedure 1 produces a consistent and asymptotic normal estimate of β^* even if IVs are invalid after conditioning on high dimensional controls.

THEOREM 2. *Suppose the assumptions* (R1) − (R3) *and* (IN1) − (IN2) *hold. As* $\sqrt{s_{\mathrm{z}1}}s\log p/\sqrt{n} \to 0$*, with probability larger than* $1 - c\left(p^{-c} + \exp\left(-cn\right)\right)$*,*

$$\left|\widehat{\beta} - \beta^*\right| \leq C\frac{1}{\delta_{\min}}\sqrt{\frac{\log p_{\mathrm{z}}}{n}}, \tag{23}$$

where $c, C > 0$ *are constants independent of* n *and* p. *In addition, if* (IN3) *holds, we have*

$$\sqrt{n}\left(\widehat{\beta} - \beta^*\right) = T^{\beta^*} + \Delta^{\beta^*} \tag{24}$$

where $T^{\beta^*} \mid \mathbf{W} \sim N(0, \mathrm{V})$, $\mathrm{V} = \sigma^2 / \left(\sum_{j\in\mathcal{V}^*}(\gamma_j^*)^2\right)^2 \left\|\sum_{j\in\mathcal{V}^*}\gamma_j^* \mathbf{W}\widehat{\mathbf{u}}^{[j]}/\sqrt{n}\right\|_2^2$, *and* $\Delta^{\beta^*}/\sqrt{\mathrm{V}} \xrightarrow{p} 0$ *as* $\sqrt{s_{\mathrm{z}1}}\, s \log p/\sqrt{n} \to 0$. *Consequently, the confidence interval given in* (15) *has asymptotically coverage probability* $1-\alpha$, *i.e.,*

$$\mathbf{P}\left\{\beta^* \in \left(\widehat{\beta} - z_{1-\alpha/2}\sqrt{\widehat{\mathrm{V}}/n}, \quad \widehat{\beta} + z_{1-\alpha/2}\sqrt{\widehat{\mathrm{V}}/n}\right)\right\} \to 1-\alpha. \tag{25}$$

In Theorem 2, the consistency of our estimator in (23) is established without (IN3) because the bias term $\sqrt{\log p_{\mathrm{z}}/n}$ discussed above is still going to zero. Section C.2 of the supplementary materials numerically demonstrates this by showing that the estimator $\widehat{\beta}$ is still consistent even when (IN3) is violated. However, for $\sqrt{n}$ asymptotic normality, Theorem 2 requires Assumption (IN3) to eliminate said bias so that our confidence interval (15) has correct coverage even with invalid IVs and high dimensional controls.

5. Simulation

5.1. Setup

In addition to the theoretical analysis of our method in Section 4, we also conduct a simulation study to investigate (i) the performance of our method and other comparators and (ii) sensitivity of our method to violations of the regularity assumptions mentioned above, most notably (IN2) and (IN3). The data generating process for the simulation follows the models (2) and (3) in Section 2.2 with $p_{\mathrm{z}} = 100$ instruments and $p_{\mathrm{x}} = 150$ covariates where $\mathbf{W}_{i\cdot}$ is a multivariate normal with mean zero and covariance $\Sigma^*_{ij} = 0.5^{|i-j|}$ for $1 \le i, j \le 250$. The parameters for the models are: $\beta^* = 1$, $\phi^* = (0.6, 0.7, 0.8, \cdots, 1.5, 0, 0, \cdots, 0) \in \mathbb{R}^{150}$ so that $s_{\mathrm{x}1} = 10$, $\psi^* = (1.1, 1.2, 1.3, \cdots, 2.0, 0, 0, \cdots, 0) \in \mathbb{R}^{150}$ so that $s_{\mathrm{x}2} = 10$, and variance-covariance of the error terms are $\mathrm{Var}(\epsilon_{i1}) = \mathrm{Var}(\epsilon_{i2}) = 1.5$, and $\mathrm{Cov}(\epsilon_{i1}, \epsilon_{i2}) = 0.75$. Instruments that satisfy Assumption (A1) are $\mathcal{S}^* = \{1, \ldots, 7\}$ and instruments that satisfy all three IV assumptions (A1)-(A3) are $\mathcal{V}^* = \{1, 2, 3, 4, 5\}$; thus instruments 6 and 7 only satisfy (A1), but do not satisfy (A2) and (A3). We fix these values throughout the entire simulation study.

The parameters we vary in the simulation study are: the sample size n, the strength of IVs via $\boldsymbol{\gamma}^*$, and violations of (A2) and (A3) via $\boldsymbol{\pi}^*$. For sample size, we let $n = (100, 200, 300, 1000, 3000)$. For IV strength, we set $\boldsymbol{\gamma}^*_{\mathcal{V}^*} = K(1, 1, 1, 1, \rho_1)$ and $\boldsymbol{\gamma}^*_{\mathcal{S}^*\setminus\mathcal{V}^*} = K(1, 1)$ and $\boldsymbol{\gamma}_{(\mathcal{S}^*)^C} = \mathbf{0}$, where we vary K (to be discussed later) and $\rho_1 = (0, 0.1, 0.2)$ across simulations. The value K controls the global strength of instruments, with higher $|K|$ indicating strong instruments in a global sense. The value ρ_1 controls the relative individual strength of instruments, specifically between the first four instruments in $\mathcal{V}^*$ and the fifth instrument. For example, $\rho_1 = 0.2$ implies that the fifth IV's individual strength is only 20% of the other four

valid instruments, i.e IVs 1 to 4. Also, varying ρ_1 would simulate the adherence of regularity assumption (IN2).

To specify K across simulations, we introduce a quantity we call the oracle concentration parameter (OCP) denoted as $C(\boldsymbol{\gamma}^*, \mathcal{V}^*, n)$

$$C(\boldsymbol{\gamma}^*, \mathcal{V}^*, n) = n\frac{\boldsymbol{\gamma}_{\mathcal{V}^*}^{*\intercal}\left(\boldsymbol{\Sigma}_{\mathcal{V}^*\mathcal{V}^*}^* - \boldsymbol{\Sigma}_{\mathcal{V}^*(\mathcal{V}^*)^C}^*\boldsymbol{\Sigma}_{(\mathcal{V}^*)^C(\mathcal{V}^*)^C}^{*-1}\boldsymbol{\Sigma}_{(\mathcal{V}^*)^C\mathcal{V}^*}^*\right)\boldsymbol{\gamma}_{\mathcal{V}^*}^*}{|\mathcal{V}^*|\Theta_{22}^*}, \tag{26}$$

where $\boldsymbol{\Sigma}_{IJ}^*$ denotes the submatrix containing Σ_{ij}^* for $i \in I$ and $j \in J$ and $\boldsymbol{\gamma}_{\mathcal{V}^*}^*$ denotes the subvector containing γ_j^* for $j \in \mathcal{V}^*$. In Section C.1 of the Supplementary Materials, we discuss the OCP and its relation to the usual concentration parameter in the IV literature [Stock and Wright, 2000]. In short, we define the OCP because the usual concentration parameter can be misleading when there are unknown redundant and invalid instruments and the OCP serves as a proxy for the usual concentration parameter.

Having defined the OCP, we can specify K as a function of n and $C(\boldsymbol{\gamma}^*, \mathcal{V}^*, n)$. Specifically, if n is set at a baseline of 100 and the simulation parameters $\mathcal{V}^*, \rho_1, \boldsymbol{\Sigma}^*$ and Θ_{22}^* are specified as above, we can find K for a particular value of the expected oracle concentration parameter $C(\boldsymbol{\gamma}^*, \mathcal{V}^*, 100)$. Thus, by varying $C(\boldsymbol{\gamma}^*, \mathcal{V}^*, 100) = (50, 100, 150, 200, 250, 500, 1000)$, we vary K.

Finally, we vary $\boldsymbol{\pi}^*$, which controls the validity of the IVs by defining $\pi_j^* = \rho_2\gamma_j^*$ for $j = 6, 7$ and $\pi_j^* = 0$ for all other j so that ρ_2 controls the magnitude of the violation of IV assumptions (A2) and (A3) from the 6th and 7th instruments. In the ideal case, we would have $\rho_2 = 0$ so that $\mathcal{S}^* = \mathcal{V}^* = \{1, 2, 3, 4, 5, 6, 7\}$. But, $\rho_2 \neq 0$ implies that the last two instruments do not satisfy (A2) and (A3). As such, we vary $\boldsymbol{\pi}^*$ by varying $\rho_2 = (0, 1, 2)$. Also, varying ρ_2 would simulate the adherence of regularity assumption (IN3).

In summary, we vary n, the strength of IVs via $\boldsymbol{\gamma}^*$, and violations of (A2) and (A3) via $\boldsymbol{\pi}^*$ in our simulation study, with ρ_1 and ρ_2 simulating the adherence to the new regularity assumptions in the paper, (IN2), and (IN3), respectively. For the setting $n \leq p$, we compare our procedure to $\widehat{\beta}_H$, which assumes IVs are valid. For the setting $n \geq p$, we add two additional comparators, the two-stage least squares (TSLS) and OLS. TSLS is the most popular IV method where one regresses $\mathbf{D}$ on $\mathbf{Z}$ and $\mathbf{X}$, and uses the predicted value of $\mathbf{D}$ in the regression of $\mathbf{Y}$ on $\mathbf{X}$ and $\mathbf{D}$. Note that the way we implement TSLS mimics most practitioners' use of TSLS by simply assuming all the instruments $\mathbf{Z}$ are valid. OLS is defined as where one regresses $\mathbf{Y}$ on $\mathbf{D}$ and $\mathbf{X}$. OLS will be biased because of confounding on $\mathbf{D}$. Finally, for both low and high dimensional settings, we have the oracle TSLS where an oracle provides us with the true set of valid IVs, which will not occur in practice. Our simulations are repeated 500 times.

5.2. Results

We present the most representative results from our simulation study; all the simulations are in Sections C.2, C.3 and C.4 of the supplementary materials. First,

Figure 1 considers the high dimensional setting with $n = 200$ and three comparators, our procedure $\widehat{\beta}$ that is robust to invalid IVs, our procedure $\widehat{\beta}_H$ that assumes all valid IVs, and the oracle TSLS. Columns "Weak" and "Strong" in the figure represent cases where $\rho_1 = 0.2$ and $\rho_1 = 0$, respectively. Columns "Valid" and "Invalid" represent cases where $\rho_2 = 0$ and $\rho_2 = 2$, respectively. The row "MAE" in the figure represents the median absolute error of the estimators, which measures the performance of the point estimators. The row "Coverage" represents the coverage performance of the confidence intervals. Finally, the row "Length" represents the average length of confidence intervals across simulations.

Both estimators $\widehat{\beta}$ and $\widehat{\beta}_H$ perform well in terms of estimation accuracy, coverage and length of confidence intervals and have similar performance to the benchmark, $\widehat{\beta}_{\text{oracle}}$, when all the instruments are valid (i.e. first and second columns of Figure 1). For example, in the MAE and length plots, the solid lines, which represent our estimator, the dashed lines, which represent our estimator assuming all valid IVs after conditioning on covariates, and the dotted lines, which represent the oracle, overlap with each other. However, if the instruments are invalid (i.e. the third and fourth columns of Figure 1), $\widehat{\beta}_H$ is not consistent and loses coverage, which makes sense since $\widehat{\beta}_H$ assumes all the IVs are valid after conditioning. However, our proposed estimator $\widehat{\beta}$ allows for possibly invalid instruments and performs as well as the oracle in terms of estimation accuracy and coverage. The average length of our robust confidence interval is only slightly larger than that of the oracle.

Figure 2 represents the same setting as Figure 1 except we now consider a larger sample size $n = 1000$. Even though n is larger than p, we still consider this to be in the many controls/high dimensional setting because the ratio of p to n is away from zero at 1/4. As expected, the estimators $\widehat{\beta}$ and $\widehat{\beta}_H$ along with the traditional TSLS estimator perform similarly to the oracle benchmark in terms of estimation accuracy, coverage and the length of confidence intervals when all the instruments are actually valid. For example, in the MAE plot of Figure 2, the solid, dashed, green and dotted lines, representing $\widehat{\beta}$, $\widehat{\beta}_H$, TSLS and the oracle, respectively, overlap with each other. Note that OLS cannot deal with confounding and hence, produces a biased estimate. However, when the instruments are invalid, the traditional TSLS estimator and $\widehat{\beta}_H$ are biased and fail to have the correct coverage. In contrast, the proposed estimator $\widehat{\beta}$ performs as well as the oracle estimator in terms of estimation accuracy and coverage, with the length of the proposed estimator being slightly longer than that for the oracle.

Finally, Figure 3 represents the setting where invalid instruments are present after conditioning on low dimensional covariates where $p_{\text{z}} = 9$ and $p_{\text{x}} = 10$ so that no coefficients for ϕ^* and ψ^* are zero and the sample size is $n = 1000$. If we use the estimator $\widehat{\beta}_E$ defined in (18) and the confidence interval (19), the proposed procedure performs almost the same as the oracle in terms of accuracy, coverage property and length, which supports the theory established in Theorem 1. Note that the performance of our procedure under the low dimensional setting with invalid IVs does not rely on assumptions (R1)-(R3) and, more importantly, (IN2)-(IN3).

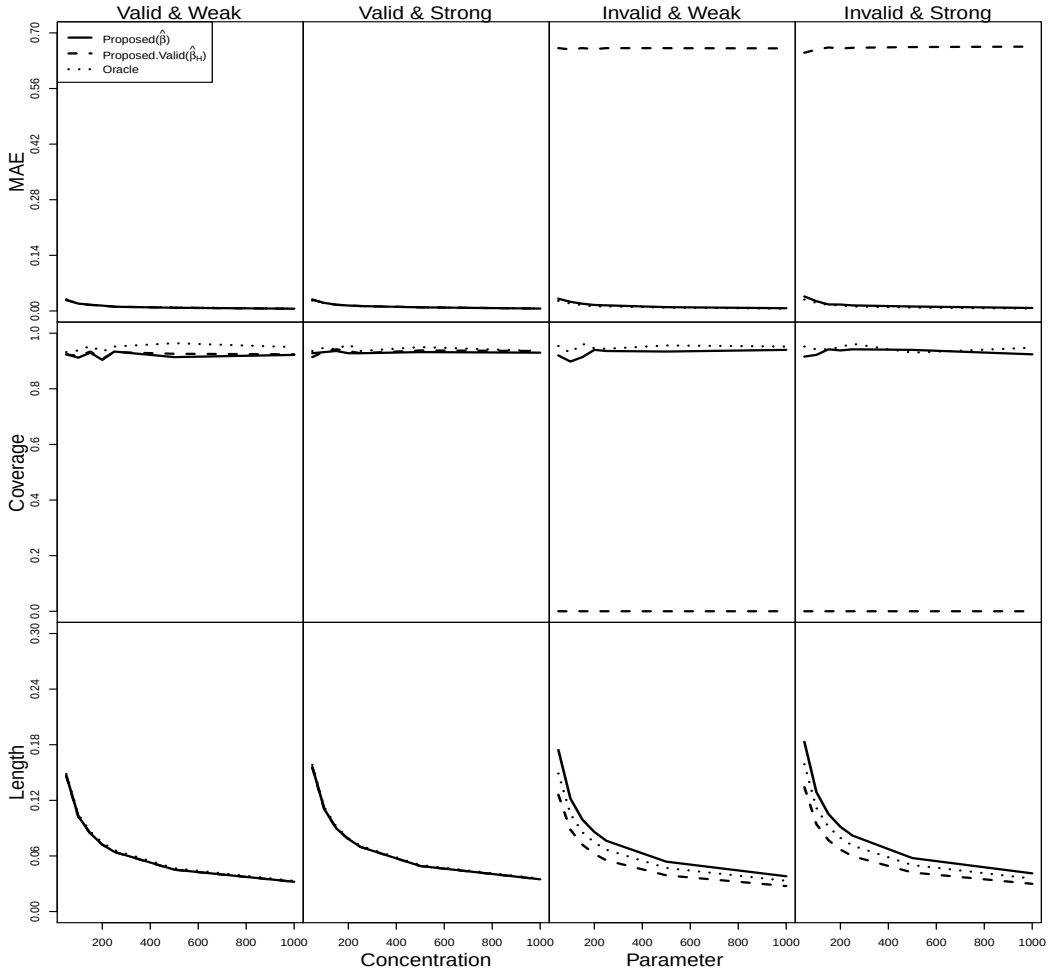

Fig. 1: Comparison of different methods when $p_z = 100$, $p_x = 150$ and $n = 200$. The x-axis represents the concentration parameter. On the y-axis, MAE represents Median Absolute Error of the estimators, Coverage represents coverage of the confidence intervals and Length represents the average length of confidence intervals. Proposed is our method allowing for invalid IVs and is represented by the solid line. Proposed.valid is our method that assumes all the IVs are valid and is represented by the dashed line. Oracle is the method that knows exactly which instruments are valid and is represented by the dotted line. The column labeled with Valid & Weak represents the case $\rho_1 = 0.2$ and $\rho_2 = 0$. The column labeled with Valid &Strong represents the case $\rho_1 = 0$ and $\rho_2 = 0$. The column labeled with Invalid &Weak represents the case $\rho_1 = 0.2$ and $\rho_2 = 2$. Finally, the column labeled with Invalid & Strong represents the case $\rho_1 = 0$ and $\rho_2 = 2$.

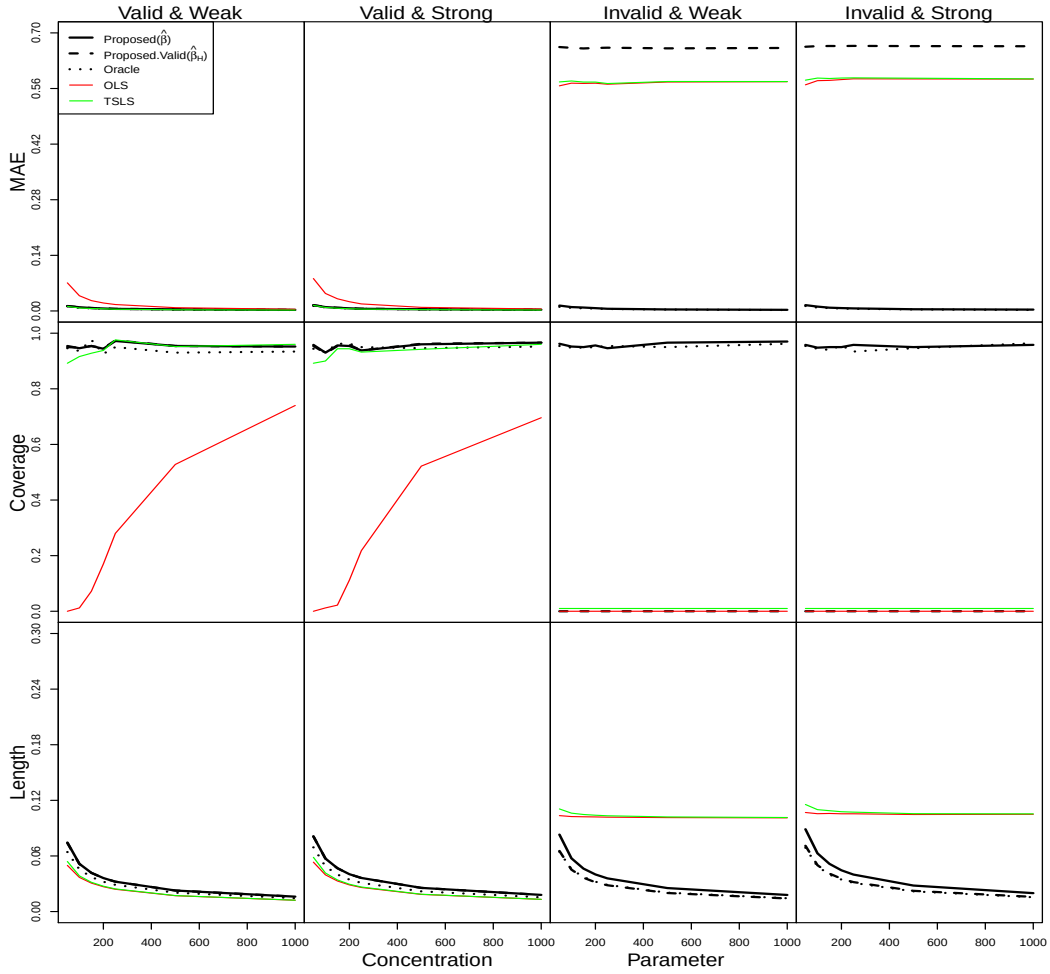

Fig. 2: Comparison of different methods when $p_z = 100$, $p_x = 150$ and $n = 1000$. The x-axis represents the concentration parameter. On the y-axis, MAE represents Median Absolute Error of the estimators, Coverage represents coverage of confidence intervals and Length represents the average length of confidence intervals. Proposed is our method allowing for invalid IVs and is represented by the solid line. Proposed.valid is our method that assumes all the IVs are valid and is represented by the dashed line. Oracle is the method that knows exactly which instruments are valid and is represented by the dotted line. The column labeled with Valid & Weak represents the case $\rho_1 = 0.2$ and $\rho_2 = 0$. The column labeled with Valid &Strong represents the case $\rho_1 = 0$ and $\rho_2 = 0$. The column labeled with Invalid &Weak represents the case $\rho_1 = 0.2$ and $\rho_2 = 2$. Finally, the column labeled with Invalid & Strong represents the case $\rho_1 = 0$ and $\rho_2 = 2$.

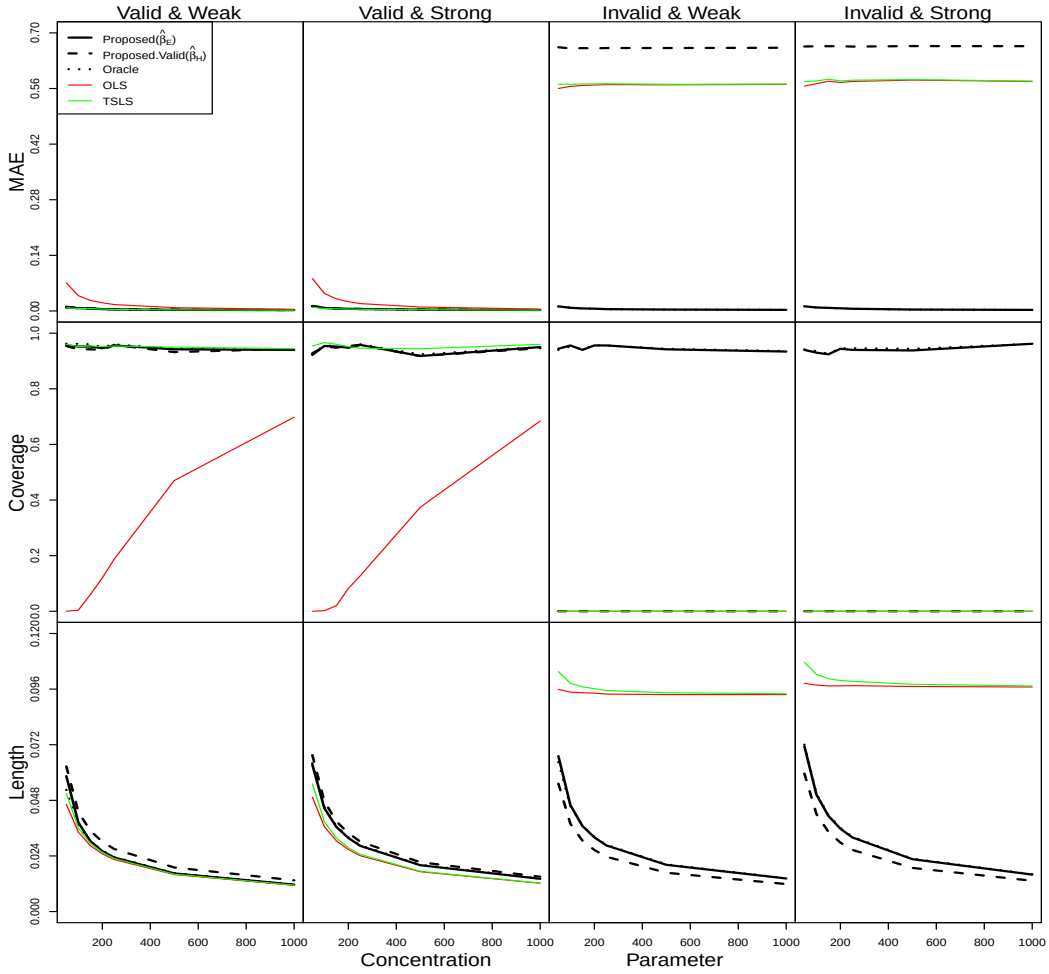

Fig. 3: Comparison of different methods when $p_z = 9$, $p_x = 10$ and $n = 1000$. The x-axis represents the concentration parameter. On the y-axis, MAE represents Median Absolute Error of the estimators, Coverage represents coverage of confidence intervals and Length represents the average length of confidence intervals. Proposed is our method allowing for invalid IVs and is represented by the solid line. Proposed.valid is our method that assumes all the IVs are valid and is represented by the dashed line. Oracle is the method that knows exactly which instruments are valid and is represented by the dotted line. The column labeled with Valid & Weak represents the case $\rho_1 = 0.2$ and $\rho_2 = 0$. The column labeled with Valid &Strong represents the case $\rho_1 = 0$ and $\rho_2 = 0$. The column labeled with Invalid &Weak represents the case $\rho_1 = 0.2$ and $\rho_2 = 2$. Finally, the column labeled with Invalid & Strong represents the case $\rho_1 = 0$ and $\rho_2 = 2$.

5.3. Violation of (IN2) and (IN3)

This section summarizes the simulation study mostly in Section C in the supplementary materials that assess the violation of regularity assumptions (IN2) and (IN3), which were used to prove properties of our procedure in high dimensions. First, in Figures 1 and 2 of the main text, our proposed estimator $\widehat{\beta}$ performs similarly for individually weak and strong IVs (i.e. for different ρ_1s). Even with $\rho_1 = 0.1$ (shown in supplementary materials), our procedure still retains the desired level of coverage, reiterating our comment in Section 4.2 that the proposed estimator is not sensitive to the individual instrument strength assumption (IN2) required in the theoretical analysis. Also, when $\rho_2 = 2$ so that the invalid IVs violate the assumptions (A2) and (A3) in a strong sense and subsequently, (IN3) is satisfied, our method is similar to the oracle; this is in contrast to the estimator $\widehat{\beta}_H$ and TSLS which are biased and lose coverage when any invalid instruments are present, even if the instruments strongly violate (A2) and (A3).

However, as discussed in Section 4.2, in high dimensions, if the instruments violate (A2) and (A3) weakly and consequently, violate (IN3), our procedure tends to suffer. In particular, in Section C.2 of the supplementary materials, when the invalid IVs weakly violated instrument with $\rho_2 = 1$, our method does not achieve 95 percent coverage even though the point estimate is still consistent, whcih was expected from Theorem 2. Fortunately, the under-coverage is at most 5 percent and is still much higher than TSLS or other methods assuming valid IVs after conditioning; see figures A2, A5, A8 and A11 of the supplementary materials for details. Also, in Section C.4 of the supplementary materials, when the number of candidate instruments are small and we have high dimensional covariates, which we think is the most typical setting in practice, the simulation results match closely with the simulation results presented in Figures 1 and 2. Specifically, in Figure A17 of the supplementary materials, the proposed confidence intervals have 95% coverage even though (IN3) is violated. Overall, the simulation study demonstrates that our procedure of selecting valid instruments can still improve the coverage property significantly compared to naive and popular methods, even if assumption (IN3) is violated.

6. Application: Causal Effect of Years of Education on Annual Earnings

To demonstrate our procedure 1 in real settings, we analyze the causal effect of years of education on yearly earnings, which has been studied extensively in economics using IV methods [Angrist and Krueger, 1991, Card, 1993, 1999]. The data comes from the Wisconsin Longitudinal Study (WLS), a longitudinal study that has kept track of American high school graduates from Wisconsin since 1957, and we examine the relationship between graduates' earnings and education from the 1974 survey Hauser [2005], roughly 20 years after they graduated from high school. Our analysis includes $N = 3772$ individuals, 1784 males and 1988 females. For our outcome, we use imputed log total yearly earnings prepared by WLS (see WLS documentation and Hauser [2005] for details) and for the treatment, we use the total years of education, all from the 1974 survey. The median total earnings is $\$9,200$ with a

25% quartile of $\$1,000$ and a 75% quartile of $\$15,320$ in 1974 dollars. The mean years of total education is 13.7 years with a standard deviation of 2.3 years.

We incorporate many covariates, including sex, graduate's hometown population, educational attainment of graduates' parents, graduates' family income, relative income in graduates' hometown, graduates' high school denomination, high school class size, all measured in 1957 when the participants were high school seniors. We also include 81 genetic covariates, specifically single nucleotide polymorphisms (SNPs), that were part of WLS to further control for potential variations between graduates; see Section B in Supplementary Materials for details on the non-genetic and genetic covariates. In summary, our data analysis includes 7 non-genetic covariates and 81 genetic covariates.

We used five instruments in our analysis, all derived from past studies of education on earnings [Card, 1993, Blundell et al., 2005, Gary-Bobo et al., 2006]. They are (i) total number of sisters, (ii) total number of brothers, (iii) individual's birth order in the family, all from Gary-Bobo et al. [2006], (iv) proximity to college from Card [1993], and (v) teacher's interest in individual's college education from Blundell et al. [2005], all measured in 1957. Although all these IVs have been suggested to be valid with varying explanations as to why they satisfy (A2) and (A3) after controlling for the aforementioned covariates, in practice, we are always uncertain due to the lack of complete socioeconomic knowledge about the effect of these IVs. Our method should provide some protection against this uncertainty compared to traditional methods where they simply assume that all five IVs are valid. Also, the first-stage F-test produces an F-statistic of 90.3 with a p-value less than 10^{-16}, which indicates very strong set of instruments. For more details on the instruments, see Section B of the Supplementary Materials.

Table 1 summarizes the results of our data analysis. OLS refers to running a regression of the treatment and the covariates on the outcome and looking at the slope coefficient of the treatment variable. TSLS refers to running two-stage least squares as described in Section 5 under the operating assumption that all the five instruments are valid; this is the usual and most popular analysis in the IV literature. Finally, we run the Pocedure 1.

Method	Point Estimate	95% Confidence Interval
OLS	0.097	(0.051, 0.143)
TSLS	0.169	(0.029, 0.301)
TSHT	0.062	(0.046, 0.077)

Table 1: Estimates of the Effect of Years of Education on Log Earnings. OLS is ordinary least squares, TSLS is two-stage least squares, and TSHT is Procedure 1.

The OLS estimate suggests a positive association between education and earnings, with statistically significant result at $\alpha = 0.05$ level. This agrees with previous literature which suggests a statistically significant positive association between years of education and log earnings Card [1999]. However, OLS does not completely control for confounding even after controlling for covariates. TSLS provides an al-

ternative method of controlling for confounding by using instruments so long as all the instruments satisfy the three core assumptions and the inclusion of covariates helps make these assumptions more plausible. Unfortunately, we notice that the TSLS estimate in Table 1 is inconsistent with previous studies' estimates among individuals from the U.S. between 1950s to 1970s, which range from 0.06 to 0.13 (see Table 4 in Card [1999]). Our method, which addresses the concern for invalid instruments with TSLS, provides an estimate of 0.062, which is more consistent with previous studies' estimates of the effect of years of education on earnings.

The data analysis suggests that our method can be a useful tool in IV analysis when there is concern for invalid instruments, even after attempting to mitigate this problem via covariates. Our method provides much more accurate estimates of the returns on education than TSLS, which naively assumes all the instruments are valid.

7. Conclusion and Discussion

We present a method to estimate the effect of the treatment on the outcome using instrumental variables where we do not make the assumption that all the instruments are valid. Our approach is based on the novel TSHT procedure, which is shown to succeed in selecting valid IVs in the presence of possibly invalid IVs. Our approach provides robust confidence intervals in the presence of invalid IVs even after controlling for many covariates. In simulation and in real data settings, our approach provides a more robust analysis than the traditional IV approaches, most notably TSLS, by providing some protection against possibly invalid instruments.

As discussed in Section 4.2, our theoretical analysis for the case of invalid IVs even after controlling for high-dimensional covariates require Assumptions (IN2) and (IN3). While (IN2) is not crucial in practice as our simulation study demonstrates and is made for a cleaner technical exposition, we believe (IN3) is most likely necessary for invalid IV problems and this is echoed in the model selection literature by Leeb and Pötscher [2005] who pointed out that "in general no model selector can be uniformly consistent for the most parsimonious true model" and hence the post-model-selection inference is generally non-uniform. Consequently, the set of competing models has to be "well separated" such that we can consistently select a correct model and Assumption (IN3) serves as this "well separated" condition in our invalid IV problem. While some recent work in high dimensional inference [Zhang and Zhang, 2014, Javanmard and Montanari, 2014, van de Geer et al., 2014, Chernozhukov et al., 2015, Cai and Guo, 2016a] do not make this "well separated" assumption, as we stressed before, our invalid IV problem is of different nature than the prior work because a single invalid IV declared as valid can ruin inference while said prior works assume covariates are exogenous and moments are known perfectly.

Finally, in practice, we believe that violation of (IN3) in high dimensions will not drastically harm inference and our CI will still have coverage around $1 - \alpha$, which is much better than TSLS and prior work assuming valid IVs after conditioning on many covariates, which have no coverage. In particular, our empirical investigations generally show that the under-coverage is no more than 5% and we think this is

partly due to the fact that (i) our procedure will still pick up the strongly invalid IVs and (ii) if the instruments are weakly invalid, the bias from them via π^* will be relatively small. It is certainly possible that advanced methods can weaken (IN3) and we leave this as a direction for further research.

Acknowledgments

The research of Hyunseung Kang was supported in part by NSF Grant DMS-1502437. The research of T. Tony Cai was supported in part by NSF Grants DMS-1208982 and DMS-1403708, and NIH Grant R01 CA127334. The research of Dylan S. Small was supported in part by NSF Grant SES-1260782.

References

Donald W. K. Andrews, Marcelo J. Moreira, and James H. Stock. Performance of conditional wald tests in {IV} regression with weak instruments. *Journal of Econometrics*, 139(1):116–132, 2007.

Joshua D Angrist and Alan B Krueger. Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, 106(4):979–1014, 1991.

Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

Michael Baiocchi, Jing Cheng, and Dylan S. Small. Instrumental variable methods for causal inference. *Statistics in Medicine*, 33(13):2297–2340, 2014.

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

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Richard Blundell, Lorraine Dearden, and Barbara Sianesi. Evaluating the effect of education on earnings: models, methods and results from the national child development survey. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(3):473–512, 2005.

J. Bowden, G. Davey Smith, and S. Burgess. Mendelian randomization with invalid instruments: effect estimation and bias detection through egger regression. *International Journal of Epidemiology*, 44(2):512–525, 2015.

Jack Bowden, George Davey Smith, Philip C. Haycock, and Stephen Burgess. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology*, 40(4):304–314, 2016.

Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer, 2011.

Stephen Burgess, Nicholas J. Timpson, Shah Ebrahim, and George Davey Smith. Mendelian randomization: where are we now and where are we going? *International Journal of Epidemiology*, 44(2):379–388, 2015.

Stephen Burgess, Jack Bowden, Frank Dudbridge, and Simon G. Thompson. Robust instrumental variable methods using multiple candidate instruments with application to mendelian randomization. *arXiv*, 2016.

T Tony Cai and Zijian Guo. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of statistics*, To appear, 2016a.

T Tony Cai and Zijian Guo. Supplement to "confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity". *The Annals of statistics*, To appear, 2016b.

David Card. Using geographic variation in college proximity to estimate the return to schooling. Working Paper 4483, National Bureau of Economic Research, October 1993.

David Card. Chapter 30 - the causal effect of education on earnings. In Orley C. Ashenfelter and David Card, editors, *Handbook of Labor Economics*, volume 3, Part A, pages 1801 – 1863. Elsevier, 1999.

Xu Cheng and Zhipeng Liao. Select the valid and relevant moments: An information-based lasso for gmm with many moments. *Journal of Econometrics*, 186(2):443–464, 2015.

Victor Chernozhukov, Christian Hansen, and Martin Spindler. Post-selection and post-regularization inference in linear models with many controls and instruments. 2015.

George Davey Smith and Shah Ebrahim. Mendelian randomization: can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, 32(1):1–22, 2003.

George Davey Smith and Shah Ebrahim. Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology*, 33(1):30–42, 2004.

Vanessa Didelez and Nuala Sheehan. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical methods in medical research*, 16 (4):309–330, 2007.

David L Donoho. De-noising by soft-thresholding. *IEEE transactions on information theory*, 41(3):613–627, 1995.

David L Donoho and Jain M Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

Jianqing Fan and Yuan Liao. Endogeneity in high dimensions. *Annals of statistics*, 42(3):872, 2014.

Robert Gary-Bobo, Nathalie Picard, and Ana Prieto. Birth order and sibship sex composition as instruments in the study of education and earnings. CEPR Discussion Papers 5514, C.E.P.R. Discussion Papers, 2006.

Eric Gautier and Alexandre B. Tsybakov. High-dimensional instrumental variables regression and confidence sets. *arXiv preprint arXiv:1105.2454*, 2011.

Chirok Han. Detecting invalid instruments using l 1-gmm. *Economics Letters*, 101 (3):285–287, 2008.

Robert M. Hauser. Survey response in the long run: The wisconsin longitudinal study. *Field Methods*, 17(1):3–29, 2005.

Miguel A. Hernán and James M. Robins. Instruments for causal inference: An epidemiologist's dream? *Epidemiology*, 17(4):360–372, 2006.

Paul W. Holland. Causal inference, path analysis, and recursive structural equations models. *Sociological Methodology*, 18(1):449–484, 1988.

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

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15 (1):2869–2909, 2014.

Hyunseung Kang, T. Tony Cai, and D. S. Small. A simple and robust confidence interval for causal effects with possibly invalid instruments. *arXiv preprint arXiv:1504.03718*, 2016a.

Hyunseung Kang, Anru Zhang, T Tony Cai, and Dylan S Small. Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association*, 111:132–144, 2016b.

Michal Kolesár, Raj Chetty, John N. Friedman, Edward L. Glaeser, and Guido W. Imbens. Identification and inference with many invalid instruments. *Journal of Business & Economic Statistics*, 33(4):474–484, 2015.

Debbie A. Lawlor, Roger M. Harbord, Jonathan A. C. Sterne, Nic Timpson, and George Davey Smith. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27(8):1133–1163, 2008.

Hannes Leeb and Benedikt M Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, 21(01):21–59, 2005.

Marcelo J. Moreira. A conditional likelihood ratio test for structural models. *Econometrica*, 71(4):1027–1048, 2003.

Michael P. Murray. Avoiding invalid instruments and coping with weak instruments. *The Journal of Economic Perspectives*, 20(4):111–132, 2006.

Jerzy Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472, 1923.

Zhao Ren, Tingni Sun, Cun-Hui Zhang, and Harrison H Zhou. Asymptotic normality and optimalities in estimation of large gaussian graphical model. *arXiv preprint arXiv:1309.6024*, 2013.

Nicholas S. Roetker, James A. Yonker, Chee Lee, Vicky Chang, Jacob J. Basson, Carol L. Roan, Taissa S. Hauser, Robert M. Hauser, and Craig S. Atwood. Multigene interactions and the prediction of depression in the wisconsin longitudinal study. *British Medical Journal Open*, 2(4):e000944, 2012.

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

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

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

James H Stock and Jonathan H Wright. Gmm with weak identification. *Econometrica*, pages 1055–1096, 2000.

James H Stock and Motohiro Yogo. Asymptotic distributions of instrumental variables statistics with many instruments. *Identification and inference for econometric models: essays in honor of Thomas Rothenberg*, 2005.

James H. Stock, Jonathan H. Wright, and Motohiro Yogo. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 20(4), 2002.

Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 101 (2):269–284, 2012.

Sonja A. Swanson and Miguel A. Hernán. Commentary: How to report instrumental variables analyses (suggestions welcome). *Epidemiology*, 24:370–374, 2013.

Nicholas J. Timpson, Debbie A. Lawlor, Roger M. Harbord, Tom R Gaunt, Ian NM Day, Lyle J. Palmer, Andrew T. Hattersley, Shah Ebrahim, Gordon Lowe, Ann Rumley, and George Davey Smith. C-reactive protein and its role in metabolic

syndrome: Mendelian randomisation study. *The Lancet*, 366(9501):1954–1959, 2005.

Sara van de Geer, Peter Bühlmann, Yaacov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

Hal R Varian. Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, 28(2):3–27, 2014.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina Eldar and Gitta Kutyniok, editors, *Compressed Sensing: Theory and Applications*, pages 210–268. Cambridge University Press, 2012.

Frank Windmeijer, Helmut Farbmacher, Neil Davies, and George Davey Smith. On the use of the lasso for instrumental variables estimation with some invalid instruments. Technical report, Department of Economics, University of Bristol, UK, 2016.

Jeffrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT press, 2nd ed. edition, 2010.

Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

Shuheng Zhou. Restricted eigenvalue conditions on subgaussian random matrices. *arXiv preprint arXiv:0912.4045*, 2009.

A. Supplementary Materials: Theory for Valid IVs After Controlling for High Dimensional Covariates

In this section, we state the theoretical results for valid IVs after controlling for High Dimensional Covariates. Under the assumptions (R1)-(R3), Theorem 3 shows if the instruments are valid after conditioning on many covariates, then the estimator $\widehat{\beta}_H$ in our procedure is consistent and asymptotically normal.

THEOREM 3. *Suppose we have valid IVs, that is* $\boldsymbol{\pi}^* = 0$ *in* (2)*, and the assumptions* (R1) − (R3) *hold. The following property holds for the estimator* $\widehat{\beta}_H$*,*

$$\sqrt{n}\left(\widehat{\beta}_H - \beta^*\right) = T^{\beta^*} + \Delta^{\beta^*}, \tag{27}$$

where $T^{\beta^*} \mid \mathbf{W} \sim N\left(0, \mathrm{V}_H\right)$, $\mathrm{V}_H = \sigma^2/\|\boldsymbol{\gamma}^*\|_2^4 \left\|\sum_{j\in\mathcal{S}^*} \boldsymbol{\gamma}_j^* \mathbf{W}\widehat{\mathbf{u}}^{[j]}/\sqrt{n}\right\|_2^2$ *and* $\Delta^{\beta^*}/\sqrt{\mathrm{V}_H} \xrightarrow{p} 0$ *as* $\sqrt{s_{z1}}s\log p/\sqrt{n} \to 0$*.*

Theorem 3 states that if the IVs satisfy the exclusion restriction and no unmeasured confounding after conditioning on many covariates, $\widehat{\beta}_H$ defined in (16) is a consistent and the dominating part of the scaled difference $\sqrt{n}(\widehat{\beta}_H - \beta)$ is normal. Based on the asymptotic normality established in (27), the following theorem justifies the coverage property of the confidence interval proposed in (17) under the assumption that all instruments have no direct effect and are unconfounded after conditioning on many covariates.

THEOREM 4. *Suppose we have valid IVs, that is* $\boldsymbol{\pi}^* = 0$ *in* (2) *and the assumptions* (R1)−(R3) *hold. Assuming* $\sqrt{s_{z1}}s\log p/\sqrt{n} \to 0$*, the confidence interval given in* (17) *has asymptotically coverage probability* $1-\alpha$*, i.e.,*

$$\mathbf{P}\left\{\beta^* \in \left(\widehat{\beta}_H - z_{1-\alpha/2}\sqrt{\widehat{\mathrm{V}}_H/n}, \quad \widehat{\beta}_H + z_{1-\alpha/2}\sqrt{\widehat{\mathrm{V}}_H/n}\right)\right\} \to 1-\alpha, \tag{28}$$

Theorem 4 is similar to a result given in Chernozhukov et al. [2015], who studied IV estimators in high dimensional regime where all the instruments are valid after conditioning. However, there are some notable differences between our results and those in Chernozhukov et al. [2015] in terms of sparsity and instrument-covariate modeling assumptions that are required to achieve $1-\alpha$ coverage. A simulation study is carried out in Section 5 to compare our procedure to that of the oracle.

B. Supplementary Materials: Details on Data Analysis

B.1. Genetic Data

This is the list of genetic covariates, i.e. 81 SNPs, used in the data analysis: rs1018381, rs1042838, rs11178997, rs1137100, rs11564774, rs11902591, rs12152850, rs12313279, rs12602084, rs12664989, rs12913832, rs1421085, rs1424954, rs1435252, rs144848, rs1501299, rs1535_chr11, rs17070145, rs174575, rs17529477, rs17561, rs17571, rs1799913, rs1799945, rs1799966, rs1799998, rs1800046, rs1800497, rs1800795, rs1800955,

Covariate	Statistics
Sex	1784 males (47.3%), 1988 females (52.7%)
Hometown Population	
Under 1,000	585 (15.5%)
1,000 - 2,499	584 (15.5%)
2,500 - 9,999	771 (20.4 %)
10,000 - 24,999	375 (9.9%)
25,000 - 49,999	602 (16.0 %)
50,000 - 99,999	226 (7.1%)
100,000 - 150,000	112 (3.0%)
Over 150,000	477 (12.6%)
Father's Total Years of Education	10.4 years (SD: 3.2)
Mother's Total Years of Education	10.8 years (SD: 3.0)
Family Income Relative to Community	
Considerably below average	15 (0.4%)
Somewhat below average	246 (6.5 %)
Average	2670 (70.8%)
Somewhat above average	783 (20.8%)
Considerably above average	58 (1.5%)
School Type	479 catholic/private (12.7%) , 3293 public (87.3%)
Graduating Class Size	166.8 students (SD: 131.5)

Table 2: Covariates used in the data analysis. All the covariates were measured in 1957. SD stands for standard deviation.

rs1805420, rs1937_chr10, rs2059693, rs2061174, rs2071219, rs2237436, rs2241766, rs2242592, rs2254298, rs2306604, rs25533, rs2760118, rs2779562, rs2963238, rs35118453, rs363050, rs3749386, rs3788862, rs3797297, rs3802657, rs3853248, rs3990403, rs4073366, rs4245147, rs429358, rs4502885, rs45537037, rs4680, rs4986852, rs5031016, rs592389, rs6152_chrX, rs6166, rs6169, rs6265, rs6277, rs6312, rs6314 rs6318 rs669, rs707555, rs7412, rs760761, rs7761133, rs7997012, rs8039957, rs8076005, rs8191992, rs821616, rs878567, rs908867.

Similar to another analysis with the WLS genetic data, we remove individuals with more than 10% missing genotype data Roetker et al. [2012]. We also remove SNPs with no variation in our data set or SNPs with more than 20% missing. The other missing values were imputed using the most frequent genotype for that SNP. Following a Mendelian randomization study by Timpson et al. [2005], we code the SNPs using an additive model.

B.2. Covariates in the Data Analysis

We include the following nine covariates for our dat analysis: sex, graduate's hometown population, educational attainment of graduates' parents, graduates' family income, relative income in graduates' hometown, graduates' high school denomination, high school class size, all measured in 1957 when the participants were high school seniors. The details of these covariates are in Table 2

Instrument	Statistics
Total number of sisters	Median: 1 (25Q: 1, 75Q: 2)
Total number of brothers	Median: 1 (25Q: 1, 75Q: 2)
Birth order	Median: 1 (25Q: 1, 75Q: 3)
Distance from High School (HS) to College	
HS more than 15 miles from any college	1,487 (39.4%)
HS 15 miles or less from extension center	168 (4.4 %)
HS 15 miles or less from state college	196 (5.2%)
HS less than 15 miles from private college	159 (4.2 %)
HS in city with extension center	446 (1.2 %)
HS in city with stage college	283 (7.5 %)
HS in city with private college	205 (5.4 %)
HS 15 miles or less from state university	239 (6.3 %)
HS in city with state university (Milwaukee or Madison)	589 (15.6%)
Teacher's Interest in Individual's College Education	
Discouraged to attend college	47 (1.2%)
Had no effect or no response	1825 (48.4%)
Encouraged to attend college	1900 (50.4%)

Table 3: Instruments used in the data analysis. All the instruments were measured in 1957. 25Q and 75Q stand for 25% and 75% quartile, respectively. There were 119 individuals who had no response to teacher's interest in individual's college education.

B.3. Instruments in the Data Analysis

They are five instruments in the analysis: (i) total number of sisters, (ii) total number of brothers, (iii) individual's birth order in the family, (iv) proximity to college, and (v) teacher's interest in individuals' college education, all measured in 1957.

A small proportion of individuals had missing teacher's interest in individual's college education. To accommodate this, we created a binary covariate to indicate missingness and imputed missing observations as the average value for that covariate. In total, there were 119 individuals with missing values for teacher's interest in education out of 3772 individuals.

C. Supplementary Materials: Details of the Simulation Study

C.1. Oracle Concentration Parameter

In the IV literature, the concentration parameter is a measure of instruments' global strength [Stock and Wright, 2000] where a high value of the concentration parameter denotes strong instruments; the concentration parameter is also the expected value of the partial F-test where we regress $\mathbf{D}$ on the instruments $\mathbf{Z}$ and covariates $\mathbf{X}$ and conduct the usual F-test on the coefficients of $\mathbf{Z}$. We can also define the population version of the concentration parameter by taking the expected value of the concentration parameter over many samples [Stock and Yogo, 2005]. Unfortunately, the traditional notions of concentration parameter cannot deal with invalid

instruments because the traditional concentration parameter assumes, a priori, that we know exactly which instruments are valid and invalid. The oracle concentration parameter resolves this problem by making a simple modification to the traditional concentration parameter where an oracle provides us with the true set of valid instruments $\mathcal{V}^*$ and we examine the strength, specifically the elements of $\boldsymbol{\gamma}^*$ in $\mathcal{V}^*$ only among the valid instruments. In fact, much like the usual concentration parameter, the oracle concentration parameter grows with sample size n. For example, when sample size increases from 100 to 200, the concentration parameter will double, that is, $C\left(\boldsymbol{\gamma}^*, \mathcal{V}^*, 200\right) = 2C\left(\boldsymbol{\gamma}^*, \mathcal{V}^*, 100\right)$.

We call it the oracle expected concentration parameter because $C(\boldsymbol{\gamma}^*, \mathcal{V}^*, n)$ is usually not computable in the real application when there is no prior information about which instruments are valid or invalid, unless, of course, an oracle provides such information to the data analyst.

C.2. Extended simulations for high-dimensional instruments and covariates

In this section, we present the extended simulation analysis by varying more parameters from the main text. In particular, we keep the data generating model and the parameter settings identical, except we vary ρ_1, ρ_2, and n more broadly than what is presented in the main text. These results are in Figures A1 to A15.

Specifically, across the figures, the column titled "WeakIV1" represents the case $\rho_1 = 0.1$, the column titled "WeakIV2" represents the case $\rho_1 = 0.2$ and the column titled with "Strong" represents the case $\rho_1 = 0$. The row indexed with "MAE" represents the Median Absolute Error of the proposed estimators, which measures the performance of the point estimator. The row indexed with "Coverage" represents the coverage performance of the confidence intervals proposed. The row indexed with "Length" represents the length performance of corresponding confidence intervals. Finally, different figures vary with sample size n with $n = (100, 200, 300, 1000, 3000)$ and violation of instrumental variable, where $\rho_2 = 0$ represents the case of all valid instruments, $\rho_2 = 1$ represents the case of weakly violated instruments and $\rho_2 = 2$ represents the case of strongly violated instruments.

The extended simulation results presented here match closely with the main simulation results presented in the main paper. Our proposed procedure performs as well as the oracle in terms of accuracy and coverage property. In addition, when there are invalid instruments, the estimator $\widehat{\beta}_H$ and TSLS are biased and lose the coverage while our procedure incorporating invalid instruments performs as well as the oracle.

The case that our method fails to do as well as advertised is in the coverage of the confidence interval when $\rho_2 = 1$ where the violation of instrumental variable is not very strong. Here, our method does not achieve 95 percent coverage even though it is still consistent. The under-coverage is up to 5 percent. We believe this is due to the failure of selecting valid instruments. Specifically, based on the theoretical results in Theorem 2 of the paper, the constructed confidence intervals might not achieve the guaranteed coverage if Assumption (IN3) in the main paper is not satisfied, that is, the violation of instruments is weak. We leave it as a direction

of future research to construct honest confidence intervals in the case of weakly violated instruments.

C.3. Low-dimensional instruments and covariates

The set up is similar to Section 5.1 except for the number of instruments and covariates. The data generating process for the simulation follows the models (2) and (3) in Section 2.2 with $p_{\rm z} = 9$ instruments and $p_{\rm x} = 10$ covariates where $\mathbf{W}_{i\cdot}$ is a multivariate normal with mean zero and covariance $\mathbf{\Sigma}^*_{ij} = 0.5^{|i-j|}$ for $1 \leq i, j \leq 19$. The parameters for the models are: $\beta^* = 1$, $\phi^* = (0.6, 0.7, 0.8, \cdots, 1.5) \in \mathbb{R}^{10}$ so that $s_{\rm x1} = 10$, $\psi^* = (1.1, 1.2, 1.3, \cdots, 2.0) \in \mathbb{R}^{10}$ so that $s_{\rm x2} = 10$, Instruments that satisfy Assumption (A1) are $\mathcal{S}^* = \{1, \ldots, 7\}$ and instruments that satisfy all three IV assumptions (A1)-(A3) are $\mathcal{V}^* = \{1, 2, 3, 4, 5\}$; thus instruments 6 and 7 only satisfy (A1), but do not satisfy (A2) and (A3). The sample size is 1000.

The extended simulation results presented here match closely with the main simulation results presented in the main paper and the simulation results presented in Section C.2. Our proposed procedure performs as well as the oracle in terms of accuracy and coverage property. In addition, when there are invalid instruments, the estimator $\widehat{\beta}_H$ and TSLS are biased and lose the coverage while our procedure incorporating invalid instruments performs as well as the oracle.

If we use the estimator $\widehat{\beta}_E$ defined in (18) and the confidence interval (19), the proposed procedure performs almost the same as the oracle in terms of accuracy, coverage property and length, which supports the theory established in Theorem 1. Specifically, in Figure A14, the proposed confidence intervals have 95% coverage even though (IN3) is violated.

C.4. Low-dimensional instruments and high-dimensional covariates

The set up is similar to Section 5.1 except for the number of instruments. The data generating process for the simulation follows the models (2) and (3) in Section 2.2 with $p_{\rm z} = 9$ instruments and $p_{\rm x} = 150$ covariates where $\mathbf{W}_{i\cdot}$ is a multivariate normal with mean zero and covariance $\mathbf{\Sigma}^*_{ij} = 0.5^{|i-j|}$ for $1 \leq i, j \leq 159$. The parameters for the models are: $\beta^* = 1$, $\phi^* = (0.6, 0.7, 0.8, \cdots, 1.5, 0, 0, \cdots, 0) \in \mathbb{R}^{150}$ so that $s_{\rm x1} = 10$, $\psi^* = (1.1, 1.2, 1.3, \cdots, 2.0, 0, 0, \cdots, 0) \in \mathbb{R}^{150}$ so that $s_{\rm x2} = 10$. Instruments that satisfy Assumption (A1) are $\mathcal{S}^* = \{1, \ldots, 7\}$ and instruments that satisfy all three IV assumptions (A1)-(A3) are $\mathcal{V}^* = \{1, 2, 3, 4, 5\}$; thus instruments 6 and 7 only satisfy (A1), but do not satisfy (A2) and (A3). The sample size is 1000.

The extended simulation results presented here match closely with the main simulation results presented in the main paper and the simulation results presented in Section C.2. Our proposed procedure performs as well as the oracle in terms of accuracy and coverage property. In addition, when there are invalid instruments, the estimator $\widehat{\beta}_H$ and TSLS are biased and lose the coverage while our procedure incorporating invalid instruments performs as well as the oracle. Specifically, in Figure A17, the proposed confidence intervals have 95% coverage even though (IN3) is violated.

D. Supplementary Materials: Proofs of Theorems

In this section, we provide detailed proofs of Theorem 1 3, 4 and 2. Proof of extra lemmas are presented in next section. Before presenting the proof, we will introduce the notations used throughout the proof.

D.1. Notations

For any vector $\mathbf{v} \in \mathbb{R}^p$, let $\mathbf{v}_j$ denote the jth element of $\mathbf{v}$. Let $\|\mathbf{v}\|_1$, $\|\mathbf{v}\|_2$, and $\|\mathbf{v}\|_\infty$ be the usual $1, 2$ and ∞-norms, respectively. Let $\|\mathbf{v}\|_0$ denote the 0-norm, i.e. the number of non-zero elements in $\mathbf{v}$. The support of $\mathbf{v}$, denoted as $\text{supp}(\mathbf{v}) \subseteq \{1, \ldots, p\}$, is defined as the set containing the non-zero elements of the vector $\mathbf{v}$, i.e. $j \in \text{supp}(\mathbf{v})$ if and only if $\mathbf{v}_j \neq 0$. Also, for a vector $\mathbf{v} \in \mathbb{R}^p$ and set $J \subseteq \{1, \ldots, p\}$, we denote $\mathbf{v}_J \in \mathbb{R}^p$ to be the vector where all the elements except whose indices are in J are zero. For a set J, $|J|$ denotes its cardinality.

For any n by p matrix $\mathbf{M} \in \mathbb{R}^{n\times p}$, we denote the (i, j) element of matrix $\mathbf{M}$ as $\mathbf{M}_{ij}$, the ith row as $\mathbf{M}_{i\cdot}$, and the jth column as $\mathbf{M}_{\cdot j}$. Let $\mathbf{M}^\intercal$ be the transpose of $\mathbf{M}$. Finally, $\|\mathbf{M}\|_\infty$ represents the element-wise matrix sup norm of matrix $\mathbf{M}$.

For a sequence of random variables X_n, we use $X_n \xrightarrow{p} X$ and $X_n \xrightarrow{d} X$ to represent that X_n converges to X in probability and in distribution, respectively. For any two sequences a_n and b_n, we will write $a_n \gg b_n$ if $\limsup \frac{b_n}{a_n} = 0$ and write $a_n \ll b_n$ if $b_n \gg a_n$. We use c and C to denote generic positive constants that may vary from place to place.

Throughout the whole proof section, we will use $\beta, \boldsymbol{\gamma}, \boldsymbol{\Gamma}, \boldsymbol{\psi}, \boldsymbol{\Psi}, \boldsymbol{\pi}, \boldsymbol{\Theta}_{11}, \boldsymbol{\Theta}_{22}, \boldsymbol{\Theta}_{12}, \boldsymbol{\Sigma}, T^\beta, \Delta^\beta$ to stand for $\beta^*, \boldsymbol{\gamma}^*, \boldsymbol{\Gamma}^*, \boldsymbol{\psi}^*, \boldsymbol{\Psi}^*, \boldsymbol{\pi}^*, \boldsymbol{\Theta}^*_{11}, \boldsymbol{\Theta}^*_{22}, \boldsymbol{\Theta}^*_{12}, \boldsymbol{\Sigma}^*, T^{\beta^*}, \Delta^{\beta^*}$ respectively and define

$$\widehat{\mathbf{v}}^{[j]} = \mathbf{W}^\intercal \widehat{\mathbf{u}}^{[j]} \quad \text{for} \quad 1 \leq j \leq p_{\mathrm{z}}.$$

We also introduce the notation $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, $\sigma_1 = \sqrt{\boldsymbol{\Theta}_{11}}$, $\sigma_2 = \sqrt{\boldsymbol{\Theta}_{22}}$ and $\boldsymbol{\Pi}_{i\cdot} = (e_{i1}, \epsilon_{i2})$. Let $M_2 = \max\{1/\lambda_{\min}(\boldsymbol{\Theta}), \lambda_{\max}(\boldsymbol{\Theta})\}$ and hence $1/M_2 \leq \lambda_{\min}(\boldsymbol{\Theta}) \leq \lambda_{\max}(\boldsymbol{\Theta}) \leq M_2$. We normalize the columns of $\mathbf{W}$ as $\mathbf{H}_{\cdot j} = \sqrt{n}\mathbf{W}_{\cdot j}/\|\mathbf{W}_{\cdot j}\|_2$ for $j \in [p]$. Let $\text{Diag} = \text{diag}\left(\|\mathbf{W}_{\cdot j}\|_2/\sqrt{n}\right)_{1\leq j\leq p}$ denote the $p \times p$ diagonal matrix with (j, j) entry to be $\|\mathbf{W}_{\cdot j}\|_2/\sqrt{n}$. We set $\lambda_0 = \sqrt{2.05 \log p/n} = (1 + \gamma_0)\sqrt{2\delta_0 \log p/n}$, where $\delta_0 = \sqrt{1.025} > 1$ and $\gamma_0 = (1.025)^{\frac{1}{4}} - 1 > 0$. Take $\epsilon_0 = 2.01/\gamma_0 + 1$, $\nu_0 = 0.01$, $\tau_0 = 0.01$, $C_1 = 2.25$, $c_0 = 1/6$ and $C_0 = 3$. We also assume that $\log p/n \to 0$ and $\delta_0 \log p > 2$. Rather than use the constants directly in the following discussion, we use $\delta_0, \pi_0, \epsilon_0, \nu_0, C_1, C_0$ and c_0 to represent the above fixed constants in the following discussion. We review the following definition of restricted eigenvalue introduced in Bickel et al. [2009],

$$\kappa(X, k, \alpha_0) = \min_{\substack{J_0\subset\{1,\cdots,p\},\\ |J_0|\leq k}} \min_{\substack{\delta\neq 0,\\ \|\delta_{J_0^c}\|_1\leq\alpha_0\|\delta_{J_0}\|_1}} \frac{\|X\delta\|_2}{\sqrt{n}\|\delta_{J_0}\|_2}. \tag{29}$$

Define the oracle estimator of σ_1 and σ_2 as

$$\sigma_1^{ora} = \frac{1}{\sqrt{n}}\|Y - \mathbf{Z}\boldsymbol{\Gamma} - \mathbf{X}\boldsymbol{\Psi}\|_2 \text{ and } \sigma_2^{ora} = \frac{1}{\sqrt{n}}\|D - \mathbf{Z}\boldsymbol{\gamma} - \mathbf{X}\boldsymbol{\psi}\|_2,$$

and

$$\tau = \sqrt{1+\epsilon_0}\frac{2\sqrt{s}\lambda_0}{\kappa(\mathbf{H}, 4s, 1+2\epsilon_0)}. \tag{30}$$

D.2. Proof of Theorem 1

Define

$$A(\mathcal{V}) = \widehat{\mathbf{\Sigma}}_{\mathcal{V},\mathcal{V}} - \widehat{\mathbf{\Sigma}}_{\mathcal{V},\mathcal{V}^c}\widehat{\mathbf{\Sigma}}^{-1}_{\mathcal{V}^c,\mathcal{V}^c}\widehat{\mathbf{\Sigma}}_{\mathcal{V}^c,\mathcal{V}} \quad \text{and} \quad A^*(\mathcal{V}) = \mathbf{\Sigma}_{\mathcal{V},\mathcal{V}} - \mathbf{\Sigma}_{\mathcal{V},\mathcal{V}^c}\mathbf{\Sigma}^{-1}_{\mathcal{V}^c,\mathcal{V}^c}\mathbf{\Sigma}_{\mathcal{V}^c,\mathcal{V}}.$$

We introduce the following lemmas to facilitate the proof.

LEMMA 1. *Under the assumptions of Theorem 1, we have*

$$\lim_{n\to\infty} \mathbf{P}\left(\widetilde{\mathcal{V}} = \mathcal{V}^*\right) = 1 \tag{31}$$

LEMMA 2. *Under the assumptions of Theorem 1, we have*

$$\sqrt{n}\left(\frac{\widetilde{\gamma}^\intercal_{\mathcal{V}^*}A(\mathcal{V}^*)\widetilde{\mathbf{\Gamma}}_{\mathcal{V}^*}}{\widetilde{\gamma}^\intercal_{\mathcal{V}^*}A(\mathcal{V}^*)\widetilde{\gamma}_{\mathcal{V}^*}} - \beta\right) \xrightarrow{d} N\left(0, \frac{\mathbf{\Theta}_{11}+\beta^2\mathbf{\Theta}_{22}-2\beta\mathbf{\Theta}_{12}}{\gamma^\intercal_{\mathcal{V}^*}A^*(\mathcal{V})\gamma_{\mathcal{V}^*}}\right). \tag{32}$$

The estimator defined in (18) can be expressed as $\widehat{\beta}_E = \frac{\widetilde{\gamma}^\intercal_{\widetilde{\mathcal{V}}}A(\widetilde{\mathcal{V}})\widetilde{\mathbf{\Gamma}}_{\widetilde{\mathcal{V}}}}{\widetilde{\gamma}^\intercal_{\widetilde{\mathcal{V}}}A(\widetilde{\mathcal{V}})\widetilde{\gamma}_{\widetilde{\mathcal{V}}}}$, and hence the difference $\sqrt{n}\left(\widehat{\beta}_E - \beta\right)$ can be expressed as

$$\sqrt{n}\left(\widehat{\beta}_E - \beta\right) = \sqrt{n}\left(\frac{\widetilde{\gamma}^\intercal_{\mathcal{V}^*}A(\mathcal{V}^*)\widetilde{\mathbf{\Gamma}}_{\mathcal{V}^*}}{\widetilde{\gamma}^\intercal_{\mathcal{V}^*}A(\mathcal{V}^*)\widetilde{\gamma}_{\mathcal{V}^*}} - \beta\right)\mathbf{1}_{\widetilde{\mathcal{V}}=\mathcal{V}^*} + \sum_{\mathcal{V}\neq\mathcal{V}^*}\sqrt{n}\left(\frac{\widetilde{\gamma}^\intercal_{\mathcal{V}}A(\mathcal{V})\widetilde{\mathbf{\Gamma}}_{\mathcal{V}}}{\widetilde{\gamma}^\intercal_{\mathcal{V}}A(\mathcal{V})\widetilde{\gamma}_{\mathcal{V}}} - \beta\right)\mathbf{1}_{\widetilde{\mathcal{V}}=\mathcal{V}} \tag{33}$$

By Lemma 1, we have $\mathbf{1}_{\widetilde{\mathcal{V}}=\mathcal{V}^*} \xrightarrow{p} 1$ and $\mathbf{1}_{\widetilde{\mathcal{V}}=\mathcal{V}} \xrightarrow{p} 0$ if $\mathcal{V} \neq \mathcal{V}^*$. Combined with Lemma 2 and Slutsky's theorem, we establish

$$\sqrt{n}\left(\frac{\widetilde{\gamma}^\intercal_{\mathcal{V}^*}A(\mathcal{V}^*)\widetilde{\mathbf{\Gamma}}_{\mathcal{V}^*}}{\widetilde{\gamma}^\intercal_{\mathcal{V}^*}A(\mathcal{V}^*)\widetilde{\gamma}_{\mathcal{V}^*}} - \beta\right)\mathbf{1}_{\widetilde{\mathcal{V}}=\mathcal{V}^*} \xrightarrow{d} N\left(0, \frac{\mathbf{\Theta}_{11}+\beta^2\mathbf{\Theta}_{22}-2\beta\mathbf{\Theta}_{12}}{\gamma^\intercal_{\mathcal{V}^*}A^*(\mathcal{V}^*)\gamma_{\mathcal{V}^*}}\right).$$

Note that for any $\epsilon_0 > 0$,

$$\mathbf{P}\left(\left|\sqrt{n}\left(\widehat{\beta}_E - \beta\right) - \sqrt{n}\left(\frac{\widetilde{\gamma}^\intercal_{\mathcal{V}^*}A(\mathcal{V}^*)\widetilde{\mathbf{\Gamma}}_{\mathcal{V}^*}}{\widetilde{\gamma}^\intercal_{\mathcal{V}^*}A(\mathcal{V}^*)\widetilde{\gamma}_{\mathcal{V}^*}} - \beta\right)\mathbf{1}_{\widetilde{\mathcal{V}}=\mathcal{V}^*}\right| \geq \epsilon_0\right) \leq \mathbf{P}\left(\widetilde{\mathcal{V}} \neq \mathcal{V}^*\right) \tag{34}$$

and it follows from Lemma 1 that

$$\sqrt{n}\left(\widehat{\beta}_E - \beta\right) - \sqrt{n}\left(\frac{\widetilde{\gamma}^\intercal_{\mathcal{V}^*}A(\mathcal{V}^*)\widetilde{\mathbf{\Gamma}}_{\mathcal{V}^*}}{\widetilde{\gamma}^\intercal_{\mathcal{V}^*}A(\mathcal{V}^*)\widetilde{\gamma}_{\mathcal{V}^*}} - \beta\right)\mathbf{1}_{\widetilde{\mathcal{V}}=\mathcal{V}^*} \xrightarrow{p} 0. \tag{35}$$

By Lemma 3.7 in Wooldridge [2010], we establish (20).

D.3. Preliminary lemmas for high dimension case

We first define the following events for the random design $\mathbf{W}$ (the normalized $\mathbf{H}$) and the error $\mathbf{\Pi}$,

$$
\begin{aligned}
G_1 &= \left\{ \frac{2}{5}\frac{1}{\sqrt{M_1}} < \frac{\|\mathbf{W}_{\cdot j}\|_2}{\sqrt{n}} < \frac{7}{5}\sqrt{M_1} \text{ for } 1 \le j \le p \right\}, \\
G_2 &= \left\{ \left| \frac{(\sigma_i^{ora})^2}{\sigma_i^2} - 1 \right| \le 2\sqrt{\frac{\log p}{n}} + 2\frac{\log p}{n} \text{ for } i = 1, 2 \right\}, \\
G_3 &= \left\{ \left| \frac{\boldsymbol{\gamma}^\intercal \widehat{\mathbf{\Sigma}} \boldsymbol{\gamma}}{\boldsymbol{\gamma}^\intercal \mathbf{\Sigma} \boldsymbol{\gamma}} - 1 \right| \le 12\sqrt{\frac{\log p}{n}} \text{ and } \left| \frac{\mathbf{\Omega}_{j\cdot}{}^\intercal \widehat{\mathbf{\Sigma}} \mathbf{\Omega}_{j\cdot}}{\mathbf{\Omega}_{jj}} - 1 \right| \le 12\sqrt{\frac{\log p}{n}}, 1 \le j \le p_{\mathrm{z}} \right\}, \\
G_4 &= \left\{ \kappa(\mathbf{H}, 4s, 1 + 2\epsilon_0) \ge \frac{1}{2\sqrt{M_1}} \right\}, \\
G_5 &= \left\{ \frac{\|\mathbf{H}^\intercal \mathbf{\Pi}_{i\cdot}\|_\infty}{n} \le \sigma_i \sqrt{\frac{2\delta_0 \log p}{n}} \text{ for } i = 1, 2 \right\}, \\
S_1 &= \left\{ \frac{\|\mathbf{H}^\intercal \mathbf{\Pi}_{i\cdot}\|_\infty}{n} \le \sigma_i^{ora} \lambda_0 \frac{\epsilon_0 - 1}{\epsilon_0 + 1} (1 - \tau) \text{ for } i = 1, 2 \right\}, \\
S_2 &= \left\{ (1 - \nu_0)\, \hat{\sigma}_i \le \sigma_i \le (1 + \nu_0) \hat{\sigma}_i \text{ for } i = 1, 2 \right\},
\end{aligned}
\tag{36}
$$

and

$$
\begin{aligned}
A_1 &= \left\{ \|e_j^\intercal \mathbf{\Omega} \widehat{\mathbf{\Sigma}} - e_j^\intercal\|_\infty \le \lambda_n,\ j = 1, 2, \cdots, p_{\mathrm{z}} \right\}, \text{ where } \lambda_n = 2eC_0 M_1^2 \sqrt{\frac{\log p}{n}}, \\
A_2 &= \left\{ |\widetilde{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j| \le \frac{\|\widehat{\mathbf{v}}^{[j]}\|_2 \sigma_2}{\sqrt{n}} \sqrt{2.05 \log p_{\mathrm{z}}} \text{ for } 1 \le j \le p_{\mathrm{z}} \right\}, \\
A_3 &= \left\{ \max_{1 \le j \le p_{\mathrm{z}}} \left\| \frac{1}{n} (\widehat{v}^{[j]})^\intercal \mathbf{\Pi}_{\cdot i} \right\|_\infty \le \left(1 + 12\sqrt{\frac{\log p}{n}} \right) M_1 \sqrt{\frac{2.05 \log p_{\mathrm{z}}}{n}} \sigma_i,\ \text{for } i = 1, 2 \right\}, \\
A_4 &= \left\{ \frac{2}{\sqrt{n}} \sum_{j \in \mathcal{S}^*} \boldsymbol{\gamma}_j (\widehat{\mathbf{v}}^{[j]})^\intercal \mathbf{\Pi}_{\cdot 2} \le \frac{2\sqrt{\log p}}{\sqrt{n}} \left\| \sum_{j \in \mathcal{S}^*} \boldsymbol{\gamma}_j \widehat{\mathbf{v}}^{[j]} \right\|_2 \sqrt{\mathbf{\Theta}_{22}} \right\}, \\
A_5 &= \left\{ \frac{1}{\sqrt{n}} \sum_{j \in \mathcal{S}^*} \boldsymbol{\gamma}_j (\widehat{\mathbf{v}}^{[j]})^\intercal \left(\mathbf{\Pi}_{\cdot 1} + \beta \mathbf{\Pi}_{\cdot 2} \right) \le \frac{\sqrt{\log p}}{\sqrt{n}} \left\| \sum_{j \in \mathcal{S}^*} \boldsymbol{\gamma}_j \widehat{\mathbf{v}}^{[j]} \right\|_2 \sqrt{\mathbf{\Theta}_{11} + \beta^2 \mathbf{\Theta}_{22} + 2\beta \mathbf{\Theta}_{12}} \right\},
\end{aligned}
\tag{37}
$$

where $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{W}^\intercal \mathbf{W}$ and $\widehat{\mathbf{v}}^{[j]} = \mathbf{W}^\intercal \widehat{\mathbf{u}}^{[j]}$. Define

$$
G = \cap_{i=1}^5 G_i \quad \text{and} \quad S = \cap_{i=1}^2 S_i \quad \text{and} \quad A = \cap_{i=1}^5 A_i.
$$

We introduce the following lemmas to control the probability of events G, S and A. The detailed proofs of the following lemmas are presented in Section E.4 and E.5.

LEMMA 3. *If $s \le cn/\log p$, then*

$$\mathbf{P}\left(G\right) \ge 1 - \frac{6}{p} - 2p^{1-C_1} - \frac{1}{2\sqrt{\pi\delta_0 \log p}} p^{1-\delta_0} - 2\exp\left(-\frac{c'n}{M_1^3}\right), \tag{38}$$

and

$$\mathbf{P}\left(G \cap S\right) \ge \mathbf{P}\left(G\right) - 2\exp\left(-\left(\frac{g_0 + 1 - \sqrt{2g_0+1}}{2}\right)n\right) - c''\frac{1}{\sqrt{\log p}} p^{1-\delta_0}, \tag{39}$$

where $g_0 = \nu_0/(2+3\nu_0)$ and c, c', c_* and c'' are universal positive constants, not depending on n and p. We also have

$$\mathbf{P}\left(A_1\right) \ge 1 - 2p_{\mathrm{z}} p^{1-c_0 C_0^2}, \quad \textit{and} \quad \mathbf{P}(A_4 \cap A_5) \ge 1 - p^{-c}, \tag{40}$$

$$\min\{\mathbf{P}(A_2), \mathbf{P}(A_3)\} \ge \mathbf{P}\left((A_1 \cap G_1 \cap G_3)\right) - \frac{1}{2\sqrt{\pi \log p_{\mathrm{z}}}} p_{\mathrm{z}}^{-0.02}. \tag{41}$$

LEMMA 4. *On the event $A_1 \cap G_1 \cap G_3$, we have*

$$\frac{\left(1-\lambda_n\right)^2}{2M_1} \le \frac{\|\widehat{\mathbf{v}}^{[j]}\|_2^2}{n} \le \left(1 + 12\sqrt{\frac{\log p}{n}}\right) M_1, \quad \textit{for } 1 \le j \le p_{\mathrm{z}}. \tag{42}$$

If $s_{\mathrm{z}1}\sqrt{\log p/n} \to 0$, on the event G_3, we have

$$\frac{1}{n}\left\|\sum_{j\in\mathcal{S}^*} \boldsymbol{\gamma}_j \widehat{\mathbf{v}}^{[j]}\right\|_2^2 \ge \frac{M_1\|\boldsymbol{\gamma}\|_2^2\left(1 - s_{\mathrm{z}1}\lambda_n\right)^2}{1 - 12\sqrt{\frac{\log p}{n}}} \quad \textit{and} \quad \frac{1}{n}\left\|\sum_{j\in\mathcal{V}^*} \boldsymbol{\gamma}_j \widehat{\mathbf{v}}^{[j]}\right\|_2^2 \ge \frac{M_1\|\boldsymbol{\gamma}_{\mathcal{V}^*}\|_2^2\left(1 - s_{\mathrm{z}1}\lambda_n\right)^2}{1 - 12\sqrt{\frac{\log p}{n}}}. \tag{43}$$

Furthermore, we have

$$\frac{M_1\left(1 - s_{\mathrm{z}1}\lambda_n\right)^2}{\|\boldsymbol{\gamma}\|_2^2\left(1 - 12\sqrt{\frac{\log p}{n}}\right)} \frac{1}{M_2} \le \mathrm{V}_H \le \frac{4s_{\mathrm{z}1}M_1^2 M_2(1+\beta^2)}{\|\boldsymbol{\gamma}\|_2^2}, \tag{44}$$

and

$$\frac{M_1\left(1 - s_{\mathrm{z}1}\lambda_n\right)^2}{\|\boldsymbol{\gamma}_{\mathcal{V}^*}\|_2^2\left(1 - 12\sqrt{\frac{\log p}{n}}\right)} \frac{1}{M_2} \le \mathrm{V} \le \frac{4s_{\mathrm{z}1}M_1^2 M_2(1+\beta^2)}{\|\boldsymbol{\gamma}_{\mathcal{V}^*}\|_2^2}. \tag{45}$$

D.4. Proof of Theorem 3

The proof of Theorem 3 is based on Lemma 5 and the following expression for the estimator $\widehat{\beta}_H$, $\widehat{\beta}_H = \widehat{\boldsymbol{\gamma}^{\intercal}\boldsymbol{\Gamma}}/\widehat{\|\boldsymbol{\gamma}\|_2^2}$, where $\widehat{\|\boldsymbol{\gamma}\|_2^2} = \sum_{j\in\widetilde{\mathcal{S}}} \widetilde{\boldsymbol{\gamma}}_j^2$and $\widehat{\boldsymbol{\gamma}^{\intercal}\boldsymbol{\Gamma}} = \sum_{j\in\widetilde{\mathcal{S}}} \widetilde{\boldsymbol{\gamma}}_j \widetilde{\boldsymbol{\Gamma}}_j$.

LEMMA 5. *Suppose that* $\sqrt{s_{z1}}s\log p/\sqrt{n} \to 0$, $\boldsymbol{\pi}^* = 0$ *and the assumptions* (R1)−(R3) *hold. Then we have the following decompositions,*

$$\sqrt{n}\left(\widehat{\|\boldsymbol{\gamma}\|_2^2} - \|\boldsymbol{\gamma}\|_2^2\right) = \frac{2}{\sqrt{n}}\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j(\widehat{\mathbf{v}}^{[j]})^\intercal\boldsymbol{\Pi}_{\cdot 2} + R^\gamma, \tag{46}$$

and

$$\sqrt{n}\left(\widehat{\boldsymbol{\gamma}^\intercal\boldsymbol{\Gamma}} - \boldsymbol{\gamma}^\intercal\boldsymbol{\Gamma}\right) = \frac{1}{\sqrt{n}}\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j(\widehat{\mathbf{v}}^{[j]})^\intercal\left(\boldsymbol{\Pi}_{\cdot 1} + \beta\boldsymbol{\Pi}_{\cdot 2}\right) + R^{\text{inter}}, \tag{47}$$

where

$$\frac{2}{\sqrt{n}}\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j(\widehat{\mathbf{v}}^{[j]})^\intercal\boldsymbol{\Pi}_{\cdot 2} \sim N\left(0, \frac{4}{n}\left\|\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j\widehat{\mathbf{v}}^{[j]}\right\|_2^2\boldsymbol{\Theta}_{22}\right), \tag{48}$$

$$\frac{1}{\sqrt{n}}\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j(\widehat{\mathbf{v}}^{[j]})^\intercal\left(\boldsymbol{\Pi}_{\cdot 1} + \beta\boldsymbol{\Pi}_{\cdot 2}\right) \sim N\left(0, \frac{1}{n}\left\|\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j\widehat{\mathbf{v}}^{[j]}\right\|_2^2\left(\boldsymbol{\Theta}_{11} + \beta^2\boldsymbol{\Theta}_{22} + 2\beta\boldsymbol{\Theta}_{12}\right)\right), \tag{49}$$

and on the event $A\cap S\cap G$*, we have*

$$\max\left\{\left|R^\gamma\right|, \left|R^{\text{inter}}\right|\right\} \le C\left(|\beta|+1\right)\|\boldsymbol{\gamma}\|_2\sqrt{s_{z1}}s\frac{\log p}{\sqrt{n}} + Cs_{z1}\frac{\log p_z}{\sqrt{n}}. \tag{50}$$

Then on the event $A\cap S\cap G$*, we have*

$$\max\left\{\left|\widehat{\|\boldsymbol{\gamma}\|_2^2} - \|\boldsymbol{\gamma}\|_2^2\right|, \left|\widehat{\boldsymbol{\gamma}^\intercal\boldsymbol{\Gamma}} - \boldsymbol{\gamma}^\intercal\boldsymbol{\Gamma}\right|\right\} \le C\|\boldsymbol{\gamma}\|_2 s_{z1}\sqrt{\frac{\log p}{n}} + Cs_{z1}\frac{\log p_z}{n} \le C\|\boldsymbol{\gamma}\|_2 s_{z1}\sqrt{\frac{\log p}{n}}. \tag{51}$$

In the following, we will prove (27) in the main paper. Note that

$$\widetilde{\beta}-\beta = -\frac{\beta}{\|\boldsymbol{\gamma}\|_2^2}\left(\widehat{\|\boldsymbol{\gamma}\|_2^2} - \|\boldsymbol{\gamma}\|_2^2\right) + \frac{1}{\|\boldsymbol{\gamma}\|_2^2}\left(\widehat{\boldsymbol{\gamma}^\intercal\boldsymbol{\Gamma}} - \boldsymbol{\gamma}^\intercal\boldsymbol{\Gamma}\right) + \frac{\|\boldsymbol{\gamma}\|_2^2 - \widehat{\|\boldsymbol{\gamma}\|_2^2}}{\|\boldsymbol{\gamma}\|_2^2}\left(\frac{\widehat{\boldsymbol{\gamma}^\intercal\boldsymbol{\Gamma}}}{\widehat{\|\boldsymbol{\gamma}\|_2^2}} - \frac{\boldsymbol{\gamma}^\intercal\boldsymbol{\Gamma}}{\|\boldsymbol{\gamma}\|_2^2}\right). \tag{52}$$

By Lemma 5, we have the following decomposition,

$$\sqrt{n}\left(\widetilde{\beta} - \beta\right) = T^\beta + \Delta^\beta, \tag{53}$$

where

$$\begin{aligned} T^\beta &= -\frac{\beta}{\|\boldsymbol{\gamma}\|_2^2}\frac{2}{\sqrt{n}}\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j(\widehat{\mathbf{v}}^{[j]})^\intercal\boldsymbol{\Pi}_{\cdot 2} + \frac{1}{\|\boldsymbol{\gamma}\|_2^2}\frac{1}{\sqrt{n}}\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j(\widehat{\mathbf{v}}^{[j]})^\intercal\left(\boldsymbol{\Pi}_{\cdot 1} + \beta\boldsymbol{\Pi}_{\cdot 2}\right) \\ &= \frac{1}{\|\boldsymbol{\gamma}\|_2^2}\frac{1}{\sqrt{n}}\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j(\widehat{\mathbf{v}}^{[j]})^\intercal\left(\boldsymbol{\Pi}_{\cdot 1} - \beta\boldsymbol{\Pi}_{\cdot 2}\right), \end{aligned}$$

and $\Delta^\beta = \mathrm{Res}_1 + \mathrm{Res}_2$ with

$$\mathrm{Res}_1 = \frac{1}{\|\boldsymbol{\gamma}\|_2^2}\left(-\beta R^{\boldsymbol{\gamma}} + R^{\mathrm{inter}}\right) \text{ and } \mathrm{Res}_2 = \sqrt{n}\frac{\|\boldsymbol{\gamma}\|_2^2 - \widehat{\|\boldsymbol{\gamma}\|_2^2}}{\|\boldsymbol{\gamma}\|_2^2}\left(\frac{\widehat{\boldsymbol{\gamma}^\intercal\boldsymbol{\Gamma}}}{\widehat{\|\boldsymbol{\gamma}\|_2^2}} - \frac{\boldsymbol{\gamma}^\intercal\boldsymbol{\Gamma}}{\|\boldsymbol{\gamma}\|_2^2}\right).$$

By the distribution of $\boldsymbol{\Pi}$, we establish that

$$T^\beta \mid \mathbf{W} \sim N\left(0, \frac{1}{n\|\boldsymbol{\gamma}\|_2^4}\left\|\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j\widehat{\mathbf{v}}^{[j]}\right\|_2^2\left(\boldsymbol{\Theta}_{11} + \beta^2\boldsymbol{\Theta}_{22} - 2\beta\boldsymbol{\Theta}_{12}\right)\right). \tag{54}$$

By Lemma 5, on the event $G \cap S \cap A$, we have

$$\frac{1}{\sqrt{\mathrm{V}_H}}|\mathrm{Res}_1| \le C\frac{1}{\|\boldsymbol{\gamma}\|_2}\left(|\beta|\,|R^{\boldsymbol{\gamma}}| + \left|R^{\mathrm{inter}}\right|\right) \le C\left(|\beta|+1\right)\sqrt{s_{\mathrm{z}1}}s\frac{\log p}{\sqrt{n}} + C\frac{1}{\|\boldsymbol{\gamma}\|_2}\frac{s_{\mathrm{z}1}\log p}{\sqrt{n}}. \tag{55}$$

Note that on the event $G \cap S \cap A$,

$$\frac{1}{\sqrt{\mathrm{V}_H}}\mathrm{Res}_2 \le C\sqrt{n}\frac{\|\boldsymbol{\gamma}\|_2^2 - \widehat{\|\boldsymbol{\gamma}\|_2^2}}{\|\boldsymbol{\gamma}\|_2} \times \frac{\left(\widehat{\boldsymbol{\gamma}^\intercal\boldsymbol{\Gamma}} - \boldsymbol{\gamma}^\intercal\boldsymbol{\Gamma}\right) + \beta\left(\|\boldsymbol{\gamma}\|_2^2 - \widehat{\|\boldsymbol{\gamma}\|_2^2}\right)}{\|\boldsymbol{\gamma}\|_2^2 + \left(\widehat{\|\boldsymbol{\gamma}\|_2^2} - \|\boldsymbol{\gamma}\|_2^2\right)} \le C\frac{s_{\mathrm{z}1}^3(\log p)^{\frac{3}{2}}}{n}, \tag{56}$$

where the last inequality follows from (51). Combined with (55), by $\sqrt{s_{\mathrm{z}1}}s\log p/\sqrt{n} \to 0$, we can establish that on the event $G \cap S \cap A$,

$$\left|\Delta^\beta/\sqrt{\mathrm{V}_H}\right| \le C\sqrt{s_{\mathrm{z}1}}s\frac{\log p}{\sqrt{n}} + C\frac{1}{\|\boldsymbol{\gamma}\|_2}\frac{s_{\mathrm{z}1}\log p}{\sqrt{n}}. \tag{57}$$

Since $\sqrt{s_{\mathrm{z}1}}s\log p/\sqrt{n} \to 0$, we establish $\Delta^\beta/\sqrt{\mathrm{V}_H} \xrightarrow{p} 0$. Combined with (54), we establish (27).

D.5. Proof of Theorem 4

We first introduce the following lemma to establish the coverage property.

LEMMA 6. *Suppose that* $\boldsymbol{\pi}^* = 0$ *and the assumptions* (R1) − (R3) *hold. As* $\sqrt{s_{\mathrm{z}1}}s\log p/\sqrt{n} \to 0$*, then we have*

$$\frac{\widehat{\mathrm{V}}_H}{\mathrm{V}_H} \xrightarrow{p} 1. \tag{58}$$

By (54), we have $\frac{T^\beta}{\sqrt{\mathrm{V}_H}} \sim N(0,1)$. Combined with (57) and Lemma 6, we have

$$\sqrt{n}\frac{\widehat{\beta}_H - \beta}{\sqrt{\widehat{\mathrm{V}}_H}} = \frac{T^\beta + \Delta^\beta}{\sqrt{\mathrm{V}_H}} \times \frac{\sqrt{\mathrm{V}_H}}{\sqrt{\widehat{\mathrm{V}}_H}} \xrightarrow{d} N(0,1). \tag{59}$$

and hence the coverage property (28) follows.

D.6. Proof of Theorem 2

The proof of the theorem follows from the following lemma, which characterizes the behavior of the selection process (10) and (12) in the main paper.

LEMMA 7. *Suppose that* $\sqrt{s_{z1}}s\log p/\sqrt{n}\to 0$ *and the assumptions* (R1) − (R3) *and* (IN1) − (IN2) *are satisfied. With probability larger than* $1-c\left(p^{-c}+\exp\left(-cn\right)\right)$, *we have*

$$\widetilde{\mathcal{V}}\subset\left\{i\in\mathcal{S}^{*}:\left|\frac{\boldsymbol{\pi}_{j}}{\boldsymbol{\gamma}_{j}}\right|\leq 2C_{*}\frac{1}{\delta_{\min}}\sqrt{\frac{\log p_{z}}{n}}\right\}\quad\text{and}\quad\left|\widetilde{\mathcal{V}}\right|>\frac{1}{2}\left|\mathcal{S}^{*}\right|. \tag{60}$$

Under the extra assumption (IN3) *in the main paper, with probability larger than* $1-c\left(p^{-c}+\exp\left(-cn\right)\right)$

$$\widetilde{\mathcal{V}}=\mathcal{V}^{*}. \tag{61}$$

We have the following decomposition,

$$\begin{aligned}\widehat{\beta}-\beta&=\frac{\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_{j}\widetilde{\boldsymbol{\Gamma}}_{j}}{\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_{j}^{2}}-\frac{\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_{j}\boldsymbol{\Gamma}_{j}}{\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_{j}^{2}}+\frac{\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_{j}\boldsymbol{\Gamma}_{j}}{\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_{j}^{2}}-\beta\\&=\left(\frac{\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_{j}\widetilde{\boldsymbol{\Gamma}}_{j}}{\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_{j}^{2}}-\frac{\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_{j}\boldsymbol{\Gamma}_{j}}{\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_{j}^{2}}\right)+\frac{\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_{j}\boldsymbol{\pi}_{j}}{\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_{j}^{2}},\end{aligned} \tag{62}$$

where the first term is taken as the variance term and the second term is taken as the bias term. In the following, we are going to analyze the bias and the variance term separately. For the bias term, we have

$$\left|\frac{\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_{j}\boldsymbol{\pi}_{j}}{\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_{j}^{2}}\right|=\left|\frac{\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_{j}^{2}\frac{\boldsymbol{\pi}_{j}}{\boldsymbol{\gamma}_{j}}}{\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_{j}^{2}}\right|\leq\max_{j\in\widetilde{\mathcal{V}}}\left|\frac{\boldsymbol{\pi}_{j}}{\boldsymbol{\gamma}_{j}}\right|\leq 2C_{*}\frac{1}{\delta_{\min}}\sqrt{\frac{\log p_{z}}{n}}, \tag{63}$$

where the last inequality follows from (60). The following lemma controls the variance term.

LEMMA 8. *Suppose that* $\sqrt{s_{z1}}s\log p/\sqrt{n}\to 0$ *and the assumptions* (R1) − (R3) *and* (IN1) − (IN3) *are satisfied. On the event* $A\cap S\cap G$, *we have*

$$\left|\frac{\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_{j}\widetilde{\boldsymbol{\Gamma}}_{j}}{\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_{j}^{2}}-\frac{\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_{j}\boldsymbol{\Gamma}_{j}}{\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_{j}^{2}}\right|\leq C\frac{1}{\delta_{\min}}\sqrt{\frac{\log p_{z}}{n}}. \tag{64}$$

Combining (63) and (64), we can establish (23) in the main paper. Under the stronger assumption (22) in the main paper, we can establish (61) and the decomposition (62) holds as

$$\widehat{\beta}-\beta=\left(\frac{\sum_{j\in\mathcal{V}^{*}}\widetilde{\boldsymbol{\gamma}}_{j}\widetilde{\boldsymbol{\Gamma}}_{j}}{\sum_{j\in\mathcal{V}^{*}}\widetilde{\boldsymbol{\gamma}}_{j}^{2}}-\frac{\sum_{j\in\mathcal{V}^{*}}\boldsymbol{\gamma}_{j}\boldsymbol{\Gamma}_{j}}{\sum_{j\in\mathcal{V}^{*}}\boldsymbol{\gamma}_{j}^{2}}\right).$$

Based on the this expression, (24) in the main paper will follow the same argument with (27), which is presented in Section D.4. We introduce the following lemma to establish the coverage property.

LEMMA 9. *Suppose the assumptions* (R1) – (R5) *and* (IN1) – (IN3) *are satisfied. As* $\sqrt{s_{z1}} s \log p/\sqrt{n} \to 0$, *we have*

$$\frac{\widehat{\mathrm{V}}}{\mathrm{V}} \xrightarrow{p} 1. \tag{65}$$

Similarly to the proof of Theorem 4 in Section D.5, we establish

$$\sqrt{n}\frac{\widehat{\beta}-\beta}{\sqrt{\widehat{\mathrm{V}}}} = \frac{T^{\beta}+\Delta^{\beta}}{\sqrt{\mathrm{V}}} \times \frac{\sqrt{\mathrm{V}}}{\sqrt{\widehat{\mathrm{V}}}} \xrightarrow{d} N(0,1), \tag{66}$$

and hence establish the coverage property (25) in the main paper.

E. Supplementary Materials: Proof of Extra Lemmas

In this section, we prove extra lemmas used in the proof of main theorems.

E.1. Proof of Lemma 1

Define $\mathcal{I} = \{1, 2, \cdots, p_{\mathrm{z}}\}$. We first note the following expression for $\widetilde{\boldsymbol{\gamma}}_j$ and $\widetilde{\boldsymbol{\Gamma}}_j$ for $i \in \mathcal{I}$,

$$\sqrt{n}\left(\widetilde{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j\right) = \left(\widehat{\boldsymbol{\Sigma}}^{-1}\right)_{j,\cdot} \frac{1}{\sqrt{n}} \mathbf{W}^{\intercal}\boldsymbol{\Pi}_{\cdot 2} \quad \text{and} \quad \sqrt{n}\left(\widetilde{\boldsymbol{\Gamma}}_j - \boldsymbol{\Gamma}_j\right) = \left(\widehat{\boldsymbol{\Sigma}}^{-1}\right)_{j,\cdot} \frac{1}{\sqrt{n}} \mathbf{W}^{\intercal}\boldsymbol{\Pi}_{\cdot 1} \tag{67}$$

and the following limiting theorem (Theorem 3.1 in Wooldridge [2010]),

$$\widetilde{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma} \quad \text{and} \quad \widetilde{\boldsymbol{\Gamma}} \xrightarrow{p} \boldsymbol{\Gamma}, \tag{68}$$

$$\sqrt{n}\left(\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\right) \xrightarrow{d} N\left(0, \Theta_{22}\left(\boldsymbol{\Sigma}^{-1}\right)_{\mathcal{I},\mathcal{I}}\right) \quad \text{and} \quad \sqrt{n}\left(\widetilde{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}\right) \xrightarrow{d} N\left(0, \Theta_{11}\left(\boldsymbol{\Sigma}^{-1}\right)_{\mathcal{I},\mathcal{I}}\right). \tag{69}$$

Note that

$$\frac{\sqrt{\widehat{\boldsymbol{\Theta}}_{22}}\|\mathbf{W}(\widehat{\boldsymbol{\Sigma}}^{-1})_{\cdot j}\|_2}{\sqrt{n}} \xrightarrow{p} \sqrt{\Theta_{22}\left(\boldsymbol{\Sigma}^{-1}\right)_{jj}}. \tag{70}$$

We define the following events

$$\begin{aligned}
\mathcal{B}_1 &= \left\{\widetilde{\mathcal{S}} = \mathcal{S}^*\right\} \\
\mathcal{B}_2 &= \left\{\max_{j\in\mathcal{V}^*} \|\widetilde{\boldsymbol{\pi}}^{[j]}\|_0 < \frac{|\mathcal{S}^*|}{2} < \min_{j\in\mathcal{S}^*\setminus\mathcal{V}^*} \|\widetilde{\boldsymbol{\pi}}^{[j]}\|_0\right\} \\
\mathcal{B}_3 &= \left\{\mathrm{supp}\left(\widetilde{\boldsymbol{\pi}}^{[j]}_{\mathcal{S}^*}\right) = \mathrm{supp}\left(\boldsymbol{\pi}_{\mathcal{S}^*}\right) \quad \text{for } j \in \mathcal{V}^*\right\}
\end{aligned} \tag{71}$$

On the event $\mathcal{B} = \mathcal{B}_1 \cap \mathcal{B}_2 \cap \mathcal{B}_3$, we have $\widetilde{\mathcal{V}} = \mathcal{V}^*$ and it is sufficient to show that

$$\lim_{n\to\infty} \mathbf{P}(\mathcal{B}) = 0 \tag{72}$$

For $j \in \mathcal{S}$, we have

$$|\widetilde{\gamma}_j| - \frac{\sqrt{\widehat{\mathbf{\Theta}}_{22}}\|\mathbf{W}(\widehat{\mathbf{\Sigma}}^{-1})_{\cdot j}\|_2}{\sqrt{n}}\sqrt{\frac{a_0 \log n}{n}} \xrightarrow{p} |\boldsymbol{\gamma}_j| > 0, \tag{73}$$

where the convergence follows from (68) and (70). For $j \in \mathcal{S}^c$, we have

$$\sqrt{\frac{n}{a_0 \log n}}|\widetilde{\gamma}_j| - \frac{\sqrt{\widehat{\mathbf{\Theta}}_{22}}\|\mathbf{W}(\widehat{\mathbf{\Sigma}}^{-1})_{\cdot j}\|_2}{\sqrt{n}} \xrightarrow{p} -\sqrt{\Theta_{22}\left(\mathbf{\Sigma}^{-1}\right)_{jj}} < 0, \tag{74}$$

where the convergence follows from (69) and (70). Combining (73) and (74), we establish that

$$\lim_{n\to\infty} \mathbf{P}\left(\mathcal{B}_1\right) = 1. \tag{75}$$

In the following, we control $\lim_{n\to\infty} \mathbf{P}(\mathcal{B})$. Without loss of generality, we assume $1 \in \widetilde{\mathcal{S}}$ and focus on the case $i = 1$. In the following, we are going to analyze the performance of $\widehat{\beta}^{[1]}$ and $\widetilde{\boldsymbol{\pi}}_j^{[1]}$. Note that

$$\widehat{\beta}^{[1]} - \frac{\mathbf{\Gamma}_1}{\boldsymbol{\gamma}_1} \xrightarrow{p} 0. \tag{76}$$

and hence we have

$$\begin{aligned}
&\sqrt{\widehat{\mathbf{\Theta}}_{11} + (\widehat{\beta}^{[1]})^2\widehat{\mathbf{\Theta}}_{22} - 2\widehat{\beta}^{[1]}\widehat{\mathbf{\Theta}}_{12}}\frac{\|\mathbf{W}((\widehat{\mathbf{\Sigma}}^{-1})_{k\cdot} - \frac{\widetilde{\gamma}_k}{\widetilde{\gamma}_1}(\widehat{\mathbf{\Sigma}}^{-1})_{1\cdot})\|_2}{\sqrt{n}} \\
&\xrightarrow{p} \sqrt{\mathbf{\Theta}_{11} + \left(\frac{\mathbf{\Gamma}_1}{\boldsymbol{\gamma}_1}\right)^2\mathbf{\Theta}_{22} - 2\frac{\mathbf{\Gamma}_1}{\boldsymbol{\gamma}_1}\mathbf{\Theta}_{12}}\sqrt{\left(\mathbf{\Sigma}^{-1}\right)_{kk} + \left(\frac{\boldsymbol{\gamma}_k}{\boldsymbol{\gamma}_1}\right)^2\left(\mathbf{\Sigma}^{-1}\right)_{11} - 2\frac{\boldsymbol{\gamma}_k}{\boldsymbol{\gamma}_1}\left(\mathbf{\Sigma}^{-1}\right)_{k1}}.
\end{aligned} \tag{77}$$

We also have the following expression

$$\begin{aligned}
\widehat{\boldsymbol{\pi}}_k^{[1]} - \left(\mathbf{\Gamma}_k - \frac{\mathbf{\Gamma}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\gamma}_k\right) &= \left(\widetilde{\mathbf{\Gamma}}_k - \frac{\widetilde{\mathbf{\Gamma}}_1}{\widetilde{\boldsymbol{\gamma}}_1}\widetilde{\boldsymbol{\gamma}}_k\right) - \left(\mathbf{\Gamma}_k - \frac{\mathbf{\Gamma}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\gamma}_k\right) \\
&= \left(\widetilde{\mathbf{\Gamma}}_k - \mathbf{\Gamma}_k\right) - \frac{\mathbf{\Gamma}_1}{\boldsymbol{\gamma}_1}\left(\widetilde{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_k\right) - \frac{\boldsymbol{\gamma}_k}{\boldsymbol{\gamma}_1^2}\left(\boldsymbol{\gamma}_1\left(\widetilde{\mathbf{\Gamma}}_1 - \mathbf{\Gamma}_1\right) - \mathbf{\Gamma}_1\left(\widetilde{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_1\right)\right) \\
&+ \left(\frac{\widetilde{\mathbf{\Gamma}}_1}{\widetilde{\boldsymbol{\gamma}}_1} - \frac{\mathbf{\Gamma}_1}{\boldsymbol{\gamma}_1}\right)\left(\frac{\boldsymbol{\gamma}_k}{\boldsymbol{\gamma}_1}\left(\widetilde{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_1\right) - \left(\widetilde{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_k\right)\right)
\end{aligned} \tag{78}$$

Note that

$$\begin{aligned}
&\sqrt{n}\left(\left(\widetilde{\mathbf{\Gamma}}_k - \mathbf{\Gamma}_k\right) - \frac{\mathbf{\Gamma}_1}{\boldsymbol{\gamma}_1}\left(\widetilde{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_k\right) - \frac{\boldsymbol{\gamma}_k}{\boldsymbol{\gamma}_1^2}\left(\boldsymbol{\gamma}_1\left(\widetilde{\mathbf{\Gamma}}_1 - \mathbf{\Gamma}_1\right) - \mathbf{\Gamma}_1\left(\widetilde{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_1\right)\right)\right) \\
&= \left(\left(\widehat{\mathbf{\Sigma}}^{-1}\right)_{\cdot k} - \frac{\boldsymbol{\gamma}_j}{\boldsymbol{\gamma}_1}\left(\widehat{\mathbf{\Sigma}}^{-1}\right)_{\cdot 1}\right)\frac{1}{\sqrt{n}}\mathbf{W}^{\intercal}\left(\mathbf{\Pi}_{\cdot 2} - \frac{\mathbf{\Gamma}_1}{\boldsymbol{\gamma}_1}\mathbf{\Pi}_{\cdot 1}\right) \\
&\xrightarrow{d} N\left(0, \mathbf{\Theta}_{11} + \left(\frac{\mathbf{\Gamma}_1}{\boldsymbol{\gamma}_1}\right)^2\mathbf{\Theta}_{22} - 2\frac{\mathbf{\Gamma}_1}{\boldsymbol{\gamma}_1}\mathbf{\Theta}_{12}\left(\mathbf{\Sigma}^{-1}\right)_{kk} + \left(\frac{\boldsymbol{\gamma}_k}{\boldsymbol{\gamma}_1}\right)^2\left(\mathbf{\Sigma}^{-1}\right)_{11} - 2\frac{\boldsymbol{\gamma}_k}{\boldsymbol{\gamma}_1}\left(\mathbf{\Sigma}^{-1}\right)_{k1}\right),
\end{aligned} \tag{79}$$

where the convergence follows from Theorem 3.1 in Wooldridge [2010]. By (68) and (69), we have

$$\left(\frac{\widetilde{\mathbf{\Gamma}}_1}{\widetilde{\gamma}_1} - \frac{\mathbf{\Gamma}_1}{\gamma_1}\right)\left(\frac{\gamma_k}{\gamma_1}(\widetilde{\gamma}_1 - \gamma_1) - (\widetilde{\gamma}_k - \gamma_k)\right) \xrightarrow{p} 0.$$

Combined with (76) and (79), we have

$$\frac{\sqrt{n}}{\sqrt{\widehat{\mathbf{\Theta}}_{11} + (\widehat{\beta}^{[1]})^2\widehat{\mathbf{\Theta}}_{22} - 2\widehat{\beta}^{[1]}\widehat{\mathbf{\Theta}}_{12}}\frac{\|\mathbf{W}((\widehat{\mathbf{\Sigma}}^{-1})_{k\cdot} - \frac{\widetilde{\gamma}_k}{\widetilde{\gamma}_1}(\widehat{\mathbf{\Sigma}}^{-1})_{1\cdot})\|_2}{\sqrt{n}}}\left(\widehat{\boldsymbol{\pi}}_k^{[1]} - \left(\mathbf{\Gamma}_k - \frac{\mathbf{\Gamma}_1}{\gamma_1}\gamma_k\right)\right) \xrightarrow{p} N(0,1) \tag{80}$$

and hence

$$\widehat{\boldsymbol{\pi}}_k^{[1]} \xrightarrow{p} \mathbf{\Gamma}_k - \frac{\mathbf{\Gamma}_1}{\gamma_1}\gamma_k. \tag{81}$$

We divide the discussion into the following three cases,

- $1 \in \mathcal{S}^* \backslash \mathcal{V}^*$ and $k \in \mathcal{V}^*$;
- $1 \in \mathcal{V}^*$ and $k \in \mathcal{V}^*$;
- $1 \in \mathcal{V}^*$ and $k \in \mathcal{S}^* \backslash \mathcal{V}^*$.

$\underline{1 \in \mathcal{S}^* \backslash \mathcal{V}^* \text{ and } k \in \mathcal{V}^*}$
In this case, $\mathbf{\Gamma}_k - \frac{\mathbf{\Gamma}_1}{\gamma_1}\gamma_k = \frac{\boldsymbol{\pi}_1}{\gamma_1}\gamma_k \neq 0$. Hence, we have

$$\begin{aligned}&\left|\widehat{\boldsymbol{\pi}}_k^{[1]}\right| - 2.05\sqrt{\widehat{\mathbf{\Theta}}_{11} + (\widehat{\beta}^{[1]})^2\widehat{\mathbf{\Theta}}_{22} - 2\widehat{\beta}^{[1]}\widehat{\mathbf{\Theta}}_{12}}\frac{\|\mathbf{W}((\widehat{\mathbf{\Sigma}}^{-1})_{k\cdot} - \frac{\widetilde{\gamma}_k}{\widetilde{\gamma}_1}(\widehat{\mathbf{\Sigma}}^{-1})_{1\cdot})\|_2}{\sqrt{n}}\sqrt{\frac{\log n}{n}} \\ &\xrightarrow{p} \left|\frac{\boldsymbol{\pi}_1}{\gamma_1}\gamma_k\right| > 0\end{aligned} \tag{82}$$

and

$$\lim_{n\to\infty} \mathbf{P}\left(k \in \operatorname{supp}\left(\widetilde{\boldsymbol{\pi}}_{\mathcal{S}^*}^{[1]}\right)\right) = 1. \tag{83}$$

Hence, by the assumption (IN1),

$$\lim_{n\to\infty} \mathbf{P}\left(\|\widetilde{\boldsymbol{\pi}}_{\mathcal{S}^*}^{[1]}\|_0 > \frac{|\mathcal{V}^*|}{2}\right) = 1. \tag{84}$$

$\underline{1 \in \mathcal{V}^* \text{ and } k \in \mathcal{V}^*}$
In this case, $\mathbf{\Gamma}_k - \frac{\mathbf{\Gamma}_1}{\gamma_1}\gamma_k = 0$. By (80), we have

$$\frac{\left|\widehat{\boldsymbol{\pi}}_k^{[1]}\right|}{2.05\sqrt{\widehat{\mathbf{\Theta}}_{11} + (\widehat{\beta}^{[1]})^2\widehat{\mathbf{\Theta}}_{22} - 2\widehat{\beta}^{[1]}\widehat{\mathbf{\Theta}}_{12}}\frac{\|\mathbf{W}((\widehat{\mathbf{\Sigma}}^{-1})_{k\cdot} - \frac{\widetilde{\gamma}_k}{\widetilde{\gamma}_1}(\widehat{\mathbf{\Sigma}}^{-1})_{1\cdot})\|_2}{\sqrt{n}}\sqrt{\frac{\log n}{n}}} \xrightarrow{p} 0 \tag{85}$$

and

$$\lim_{n\to\infty} \mathbf{P}\left(k \notin \operatorname{supp}\left(\widetilde{\boldsymbol{\pi}}_{\mathcal{S}^*}^{[1]}\right)\right) = 1. \tag{86}$$

Hence, by the assumption (IN1),

$$\lim_{n\to\infty} \mathbf{P}\left(\|\widetilde{\boldsymbol{\pi}}^{[1]}_{\mathcal{S}^*}\|_0 < \frac{|\mathcal{V}^*|}{2}\right) = 1. \tag{87}$$

$\underline{1 \in \mathcal{V}^* \text{ and } k \in \mathcal{S}^*\backslash\mathcal{V}^*}$
In this case, $\boldsymbol{\Gamma}_k - \frac{\boldsymbol{\Gamma}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\gamma}_k = \boldsymbol{\pi}_k \neq 0$. Hence, we have

$$\left|\widehat{\boldsymbol{\pi}}^{[1]}_k\right| - 2.05\sqrt{\widehat{\boldsymbol{\Theta}}_{11} + (\widehat{\beta}^{[1]})^2\widehat{\boldsymbol{\Theta}}_{22} - 2\widehat{\beta}^{[1]}\widehat{\boldsymbol{\Theta}}_{12}}\frac{\|\mathbf{W}((\widehat{\boldsymbol{\Sigma}}^{-1})_{k\cdot} - \frac{\widetilde{\gamma}_k}{\widetilde{\gamma}_1}(\widehat{\boldsymbol{\Sigma}}^{-1})_{1\cdot})\|_2}{\sqrt{n}}\sqrt{\frac{\log n}{n}} \xrightarrow{p} |\boldsymbol{\pi}_k| > 0 \tag{88}$$

and hence

$$\lim_{n\to\infty} \mathbf{P}\left(k \in \operatorname{supp}\left(\widetilde{\boldsymbol{\pi}}^{[1]}_{\mathcal{S}^*}\right)\right) = 1. \tag{89}$$

Since we can replace the index 1 with any index $j \in \widetilde{\mathcal{V}}$, then (84) and (87) can be correspondingly replaced by

$$\lim_{n\to\infty} \mathbf{P}\left(\|\widetilde{\boldsymbol{\pi}}^{[j]}_{\mathcal{S}^*}\|_0 > \frac{|\mathcal{V}^*|}{2}\right) = 1 \quad \text{for} \quad j \in \mathcal{S}^*\backslash\mathcal{V}^*; \tag{90}$$

$$\lim_{n\to\infty} \mathbf{P}\left(\|\widetilde{\boldsymbol{\pi}}^{[j]}_{\mathcal{S}^*}\|_0 < \frac{|\mathcal{V}^*|}{2}\right) = 1 \quad \text{for} \quad j \in \mathcal{V}^*. \tag{91}$$

By (90) and (91), we can establish $\lim_{n\to\infty} \mathbf{P}(\mathcal{B}_1 \cap \mathcal{B}_2) = 1$. By (86) and (89), we have

$$\lim_{n\to\infty} \mathbf{P}\left(\operatorname{supp}\left(\widetilde{\boldsymbol{\pi}}^{[1]}_{\mathcal{S}^*}\right) = \operatorname{supp}(\boldsymbol{\pi}_{\mathcal{S}^*})\right) = 1. \tag{92}$$

Similarly, we can obtain that

$$\lim_{n\to\infty} \mathbf{P}\left(\operatorname{supp}\left(\widetilde{\boldsymbol{\pi}}^{[j]}_{\mathcal{S}^*}\right) = \operatorname{supp}(\boldsymbol{\pi}_{\mathcal{S}^*}) \quad \text{for} \quad j \in \mathcal{V}^*\right) = 1, \tag{93}$$

and hence $\lim_{n\to\infty} \mathbf{P}(\mathcal{B}_1 \cap \mathcal{B}_2 \cap \mathcal{B}_3) = 1$.

E.2. Proof of Lemma 2

Note that

$$\sqrt{n}\left(\frac{\widetilde{\boldsymbol{\gamma}}^{\intercal}_{\mathcal{V}^*}A(\mathcal{V}^*)\widetilde{\boldsymbol{\Gamma}}_{\mathcal{V}^*}}{\widetilde{\boldsymbol{\gamma}}^{\intercal}_{\mathcal{V}^*}A(\mathcal{V}^*)\widetilde{\boldsymbol{\gamma}}_{\mathcal{V}^*}} - \beta\right) = \frac{\widetilde{\boldsymbol{\gamma}}^{\intercal}_{\mathcal{V}^*}A(\mathcal{V}^*)\left(\widehat{\boldsymbol{\Sigma}}^{-1}\right)_{\mathcal{V}^*\cdot}}{\widetilde{\boldsymbol{\gamma}}^{\intercal}_{\mathcal{V}^*}A(\mathcal{V}^*)\widetilde{\boldsymbol{\gamma}}_{\mathcal{V}^*}}\frac{1}{\sqrt{n}}\mathbf{W}^{\intercal}\left(\boldsymbol{\Pi}_{\cdot 2} - \beta\boldsymbol{\Pi}_{\cdot 1}\right). \tag{94}$$

Since

$$\frac{\widetilde{\boldsymbol{\gamma}}^{\intercal}_{\mathcal{V}^*}A(\mathcal{V}^*)\left(\widehat{\boldsymbol{\Sigma}}^{-1}\right)_{\mathcal{V}^*\cdot}}{\widetilde{\boldsymbol{\gamma}}^{\intercal}_{\mathcal{V}^*}A(\mathcal{V}^*)\widetilde{\boldsymbol{\gamma}}_{\mathcal{V}^*}} \xrightarrow{p} \frac{\boldsymbol{\gamma}^{\intercal}_{\mathcal{V}^*}A^*(\mathcal{V}^*)\left(\boldsymbol{\Sigma}^{-1}\right)_{\mathcal{V}^*\cdot}}{\boldsymbol{\gamma}^{\intercal}_{\mathcal{V}^*}A^*(\mathcal{V}^*)\boldsymbol{\gamma}_{\mathcal{V}^*}} \tag{95}$$

and

$$\frac{1}{\sqrt{n}}\mathbf{W}^{\intercal}\left(\boldsymbol{\Pi}_{\cdot 2} - \beta\boldsymbol{\Pi}_{\cdot 1}\right) \xrightarrow{d} N\left(0, \left(\boldsymbol{\Theta}_{11} + \beta^2\boldsymbol{\Theta}_{22} - 2\beta\boldsymbol{\Theta}_{12}\right)\boldsymbol{\Sigma}\right), \tag{96}$$

we have

$$\begin{aligned}&\frac{\widetilde{\boldsymbol{\gamma}}_{\mathcal{V}^*}^{\intercal} A(\mathcal{V}^*)\left(\widehat{\boldsymbol{\Sigma}}^{-1}\right)_{\mathcal{V}^*\cdot}}{\widetilde{\boldsymbol{\gamma}}_{\mathcal{V}^*}^{\intercal} A(\mathcal{V}^*)\widetilde{\boldsymbol{\gamma}}_{\mathcal{V}^*}}\frac{1}{\sqrt{n}}\mathbf{W}^{\intercal}\left(\boldsymbol{\Pi}_{\cdot 2}-\beta\boldsymbol{\Pi}_{\cdot 1}\right)\\ &\xrightarrow{d} N\left(0,\frac{\boldsymbol{\gamma}_{\mathcal{V}^*}^{\intercal} A^*(\mathcal{V}^*)\left(\boldsymbol{\Sigma}^{-1}\right)_{\mathcal{V}^*\mathcal{V}^*} A^*(\mathcal{V}^*)\boldsymbol{\gamma}_{\mathcal{V}^*}}{\left(\boldsymbol{\gamma}_{\mathcal{V}^*}^{\intercal} A^*(\mathcal{V}^*)\boldsymbol{\gamma}_{\mathcal{V}^*}\right)^2}\left(\boldsymbol{\Theta}_{11}+\beta^2\boldsymbol{\Theta}_{22}-2\beta\boldsymbol{\Theta}_{12}\right)\right)\end{aligned} \tag{97}$$

Since $A^*(\mathcal{V}^*)\left(\boldsymbol{\Sigma}^{-1}\right)_{\mathcal{V}^*\mathcal{V}^*} A^*(\mathcal{V}^*) = A^*(\mathcal{V}^*)$, we establish (32).

E.3. Lemmas for scaled Lasso and de-biasing Lasso

We introduce the following lemmas for scaled Lasso and de-biasing Lasso used in the later proofs. Lemma 10 establishes the convergence rate of the scaled Lasso method, which is based on the analysis in Sun and Zhang [2012].

LEMMA 10. *On the event $G\cap S$, if $s\le cn/\log p$, then*

$$\|\widehat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma}\|_1+\|\widehat{\boldsymbol{\Psi}}-\boldsymbol{\Psi}\|_1\le Cs\sqrt{\frac{\log p}{n}}\sigma_1,\quad \|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\|_1+\|\widehat{\boldsymbol{\psi}}-\boldsymbol{\psi}\|_1\le Cs\sqrt{\frac{\log p}{n}}\sigma_2, \tag{98}$$

$$\frac{1}{\sqrt{n}}\|\mathbf{Z}(\widehat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma})+\mathbf{X}(\widehat{\boldsymbol{\Psi}}-\boldsymbol{\Psi})\|_2\le C\sqrt{\frac{s\log p}{n}}\sigma_1, \tag{99}$$

and

$$\frac{1}{\sqrt{n}}\|\mathbf{Z}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})+\mathbf{X}(\widehat{\boldsymbol{\psi}}-\boldsymbol{\psi})\|_2\le C\sqrt{\frac{s\log p}{n}}\sigma_2. \tag{100}$$

The following lemma is the key result for the de-biasing Lasso estimator, established in Zhang and Zhang [2014], Javanmard and Montanari [2014], van de Geer et al. [2014].

LEMMA 11. *We have the following expressions for the proposed de-biased estimator,*

$$\widetilde{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma}=D^{\boldsymbol{\Gamma}}+\Delta^{\boldsymbol{\Gamma}}, \tag{101}$$

where

$$D_j^{\boldsymbol{\Gamma}}=\frac{1}{n}(\widehat{\mathbf{v}}^{[j]})^{\intercal}\boldsymbol{\Pi}_{\cdot 1}\quad \textit{and}\quad \Delta_j^{\boldsymbol{\Gamma}}=\left(\frac{1}{n}(\widehat{\mathbf{u}}^{[j]})^{\intercal}\widehat{\boldsymbol{\Sigma}}-e_j^{\intercal}\right)\begin{pmatrix}\widehat{\boldsymbol{\Gamma}}-\boldsymbol{\Gamma}\\ \widehat{\boldsymbol{\Psi}}-\boldsymbol{\Psi}\end{pmatrix},\ i=1,\cdots,p_{\mathrm{z}}. \tag{102}$$

We also have

$$\widetilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}=D^{\boldsymbol{\gamma}}+\Delta^{\boldsymbol{\gamma}}, \tag{103}$$

where

$$D_j^{\boldsymbol{\gamma}}=\frac{1}{n}(\widehat{\mathbf{v}}^{[j]})^{\intercal}\boldsymbol{\Pi}_{\cdot 2}\quad \textit{and}\quad \Delta_j^{\boldsymbol{\gamma}}=\left(\frac{1}{n}(\widehat{\mathbf{u}}^{[j]})^{\intercal}\widehat{\boldsymbol{\Sigma}}-e_j^{\intercal}\right)\begin{pmatrix}\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\\ \widehat{\boldsymbol{\psi}}-\boldsymbol{\psi}\end{pmatrix},\ i=1,\cdots,p_{\mathrm{z}}. \tag{104}$$

On the event $S\cap G\cap A$, we have

$$\max\left\{\|\Delta^{\boldsymbol{\gamma}}\|_\infty,\|\Delta^{\boldsymbol{\Gamma}}\|_\infty\right\}\le Cs\frac{\log p}{n}\max\left\{\sigma_1,\sigma_2\right\}. \tag{105}$$

E.4. Proof of Lemma 3

The proof of Lemma 3 is a generalization of Lemma 4 in Cai and Guo [2016a]. In the following, we extend the Gaussian design in Cai and Guo [2016a] to sub-gaussian design considered in this paper. Since the error of the regression is still assumed to be Gaussian, it is sufficient to establish the probability bound of G_1, G_3, G_4 and A_1 for the sub-gaussian design matrix and control the events A_2 and A_3. The probability bound of the event A_1 for the sub-gaussian design is established in Lemma 4 of Cai and Guo [2016b]. By Corollary 5.17 in Vershynin [2012] and the union bound, we have

$$\mathbf{P}\left(\max_{1\le j\le p}\left|\frac{1}{n}\left(\|\mathbf{W}_{\cdot j}\|_2^2-\mathbf{E}\|\mathbf{W}_{\cdot j}\|_2^2\right)\right|\ge\epsilon\right)\le 2p\exp\left(-\frac{1}{6}\min\left\{\frac{\epsilon^2}{K^2},\frac{\epsilon}{K}\right\}n\right),$$

where $K=4M_1$. Taking $\epsilon=12M_1\sqrt{\log p/n}$, we have

$$\mathbf{P}\left(\max_{1\le j\le p}\left|\frac{1}{n}\left(\|\mathbf{W}_{\cdot j}\|_2^2-\mathbf{E}\|\mathbf{W}_{\cdot j}\|_2^2\right)\right|\ge 12M_1\sqrt{\frac{\log p}{n}}\right)\le 2p^{-\frac{1}{2}}\quad\text{and}\quad\mathbf{P}\left(G_1\right)\ge 1-2p^{-\frac{1}{2}}. \tag{106}$$

Similarly, we have $\mathbf{P}\left(\left|\frac{1}{n}\left(\|\mathbf{W}u\|_2^2-\mathbf{E}\|\mathbf{W}u\|_2^2\right)\right|\ge 12M_1\|u\|^2\sqrt{\log p/n}\right)\le 2p^{-\frac{3}{2}}$, and

$$\mathbf{P}\left(\left|\frac{1}{n}\left(\frac{\|\mathbf{W}u\|_2^2}{\mathbf{E}\|\mathbf{W}u\|_2^2}-1\right)\right|\ge 12M_1\frac{\|u\|^2}{\mathbf{E}\|\mathbf{W}u\|_2^2}\sqrt{\log p/n}\right)\le 2p^{-\frac{3}{2}},$$

and hence

$$\mathbf{P}\left(G_3\right)\ge 1-(p_{\mathrm{z}}+2)p^{-\frac{3}{2}}.$$

By Theorem 1.6 in Zhou [2009], if $n\ge 1/\theta^2\times c'M_1^3\max\left\{12(2+\gamma_0)^2M_1s\log\left(5ep/4s\right),9\log p\right\}$, then with probability at least $1-2\exp\left(-c\theta^2n/M_1^3\right)$, for all δ such that there exists $|J_0|\le 4s$ and $\|\delta_{J_0^c}\|_1\le\gamma_0\|\delta_{J_0}\|_1$, we have $\|Z\delta\|_2/(\sqrt{n}\|\mathbf{\Sigma}^{\frac{1}{2}}\delta\|_2)\ge 1-\theta$. By taking $\theta=\frac{1}{2}$, if $n\ge 4c'M_1^3\max\left\{12\left(2+\gamma_0\right)^2M_1s\log\left(5ep/4s\right),9\log p\right\}$, then $\mathbf{P}\left(G_4\right)\ge 1-2\exp\left(-cn/M_1^3\right)$. In the following, we control the events A_2 and A_3,

$$\begin{aligned}\mathbf{P}(A_2^c)\le&\mathbf{P}\left(\max_{1\le i\le q}\frac{\left|D_j^\gamma\right|}{\sqrt{\mathrm{Var}\left(D_j^\gamma\right)}}\ge\sqrt{2.02\log p_{\mathrm{z}}}\right)+\mathbf{P}\left(\max_{1\le j\le p_{\mathrm{z}}}\frac{\left|\Delta_j^\gamma\right|}{\sqrt{\mathrm{Var}\left(D_j^\gamma\right)}}\ge 0.01\sqrt{\log p_{\mathrm{z}}}\right)\\ \le&\frac{1}{2\sqrt{\pi\log p_{\mathrm{z}}}}p_{\mathrm{z}}^{-0.02}+\mathbf{P}\left((S\cap G\cap A_1)^c\right),\end{aligned}$$

where the first inequality follows from (103) and the second inequality follows from (104) and (105). The control of $\mathbf{P}(A_4\cap A_5)$ follows from (48) and (49). Note that

$$\begin{aligned}\mathbf{P}(A_3^c)\le&\mathbf{P}\left((A_1\cap G_1\cap G_3)^c\right)+\mathbf{P}\left(A_3^c\cap A_1\cap G_1\cap G_3\right)\\ \le&\mathbf{P}\left((A_1\cap G_1\cap G_3)^c\right)+2\mathbf{P}\left(\max_{1\le j\le p_{\mathrm{z}}}\frac{1}{\|\widehat{\mathbf{v}}^{[j]}\|_2\sigma_1}\left|(\widehat{\mathbf{v}}^{[j]})^{\intercal}\mathbf{\Pi}_{\cdot 1}\right|\ge\sqrt{2.05\log p_{\mathrm{z}}}\right)\\ \le&\mathbf{P}\left((A_1\cap G_1\cap G_3)^c\right)+\frac{1}{\sqrt{\pi\log p_{\mathrm{z}}}}p_{\mathrm{z}}^{-0.05},\end{aligned}$$

where the second inequality follows from (42) and the last inequality follows from the fact that $1/(\|\widehat{\mathbf{v}}^{[j]}\|_2\sigma_1) \times (\widehat{\mathbf{v}}^{[j]})^\intercal \mathbf{\Pi}_{\cdot 1}$ conditioning on $\mathbf{W}$ is normally distributed.

E.5. Proof of Lemma 4

In the following, we only establish the results for $\widehat{\mathbf{v}}^{[1]}$ and the same argument extends to $\widehat{\mathbf{v}}^{[j]}$ where $1 \le j \le p_z$. Since $\lambda_n = 2eC_0M_1^2\sqrt{\log p/n}$ is chosen such that $\mathbf{\Omega}_{1\cdot}$ belongs to the feasible set, we have

$$\frac{\|\widehat{\mathbf{v}}^{[1]}\|_2^2}{n} \le \frac{\|\mathbf{W}\mathbf{\Omega}_{1\cdot}\|_2^2}{n}. \tag{107}$$

By Lemma 12 in Javanmard and Montanari [2014], we have

$$\frac{\|\widehat{\mathbf{v}}^{[1]}\|_2^2}{n} \ge \frac{(1-\lambda_n)^2}{\widehat{\mathbf{\Sigma}}_{11}}. \tag{108}$$

By the definition of G_1 and G_3, we establish (42). Let $\mathcal{I} = \{1, 2, \cdots, p_z\}$ and assume that $M \in \mathbb{R}^{p\times p_z}$ belongs to the feasible set $\|\widehat{\mathbf{\Sigma}}\mathbf{\Omega} - \mathbf{I}_{\cdot\mathcal{I}}\|_\infty \le \lambda_n$, where $\mathbf{I}_{\cdot\mathcal{I}}$ denotes the sub-matrix of the identity matrix containing the column with index $i \in \mathcal{I}$, that is, $\|\widehat{\mathbf{\Sigma}}M - \mathbf{I}_{\cdot\mathcal{I}}\|_\infty \le \lambda_n$, and hence

$$\|\widehat{\mathbf{\Sigma}}M\boldsymbol{\gamma} - \boldsymbol{\gamma}\|_\infty = \|\left(\widehat{\mathbf{\Sigma}}M - \mathbf{I}_{\cdot\mathcal{I}}\right)\boldsymbol{\gamma}\|_\infty \le \|\widehat{\mathbf{\Sigma}}M - \mathbf{I}_{\cdot\mathcal{I}}\|_\infty\|\boldsymbol{\gamma}\|_1 \le \lambda_n\|\boldsymbol{\gamma}\|_1. \tag{109}$$

Note that

$$\left|\boldsymbol{\gamma}^\intercal\widehat{\mathbf{\Sigma}}M\boldsymbol{\gamma} - \|\boldsymbol{\gamma}\|_2^2\right| = \left|\boldsymbol{\gamma}^\intercal\left(\widehat{\mathbf{\Sigma}}M\boldsymbol{\gamma} - \boldsymbol{\gamma}\right)\right| \le \|\boldsymbol{\gamma}\|_1\|\widehat{\mathbf{\Sigma}}M\boldsymbol{\gamma} - \boldsymbol{\gamma}\|_\infty \le \lambda_n\|\boldsymbol{\gamma}\|_1^2, \tag{110}$$

where the last inequality follows from (109). The inequality (110) informs that $M\boldsymbol{\gamma}$ is in the feasible set

$$\left|\boldsymbol{\gamma}\widehat{\mathbf{\Sigma}}\left(M\boldsymbol{\gamma}\right) - \|\boldsymbol{\gamma}\|_2^2\right| \le \lambda_n\|\boldsymbol{\gamma}\|_1^2. \tag{111}$$

We define μ^* as

$$\begin{aligned} \mu^* &= \arg\min_\mu \mu^\intercal\widehat{\mathbf{\Sigma}}\mu \\ &\text{subject to} \quad \left|\boldsymbol{\gamma}^\intercal\widehat{\mathbf{\Sigma}}\mu - \|\boldsymbol{\gamma}\|_2^2\right| \le \lambda_n\|\boldsymbol{\gamma}\|_1^2 \end{aligned} \tag{112}$$

By (111), we have the following inequality,

$$\frac{1}{n}\left\|\sum_{j\in\mathcal{S}^*}\gamma_j\widehat{\mathbf{v}}^{[j]}\right\|_2^2 = \boldsymbol{\gamma}^\intercal M^\intercal\widehat{\mathbf{\Sigma}}M\boldsymbol{\gamma} \ge (\mu^*)^\intercal\,\widehat{\mathbf{\Sigma}}\mu^*. \tag{113}$$

In the following, we will show that $(\mu^*)^\intercal\,\widehat{\mathbf{\Sigma}}\mu^* = \langle\mu^*, \widehat{\mathbf{\Sigma}}\mu^*\rangle$ is further lower bounded. Since μ^* is feasible in the constrained set of (112), we have $\|\boldsymbol{\gamma}\|_2^2 - \boldsymbol{\gamma}^\intercal\widehat{\mathbf{\Sigma}}\mu^* - \lambda_n\|\boldsymbol{\gamma}\|_1^2 \le 0$, and hence for any positive constant $c > 0$, we have

$$\begin{aligned} &\langle\mu^*, \widehat{\mathbf{\Sigma}}\mu^*\rangle \ge \langle\mu^*, \widehat{\mathbf{\Sigma}}\mu^*\rangle + c\left(\|\boldsymbol{\gamma}\|_2^2 - \boldsymbol{\gamma}^\intercal\widehat{\mathbf{\Sigma}}\mu^* - \lambda_n\|\boldsymbol{\gamma}\|_1^2\right) \\ &\ge \min_\mu\left(\langle\mu, \widehat{\mathbf{\Sigma}}\mu\rangle + c\left(\|\boldsymbol{\gamma}\|_2^2 - \boldsymbol{\gamma}^\intercal\widehat{\mathbf{\Sigma}}\mu - \lambda_n\|\boldsymbol{\gamma}\|_1^2\right)\right) = -\frac{c^2}{4}\langle\boldsymbol{\gamma}, \widehat{\mathbf{\Sigma}}\boldsymbol{\gamma}\rangle + c\left(\|\boldsymbol{\gamma}\|_2^2 - \lambda_n\|\boldsymbol{\gamma}\|_1^2\right). \end{aligned} \tag{114}$$

Note that $\|\boldsymbol{\gamma}\|_1^2\lambda_n \leq s_{z1}\lambda_n\|\boldsymbol{\gamma}\|_2^2 = Cs_{z1}\sqrt{\log p/n}\|\boldsymbol{\gamma}\|_2^2 \ll \|\boldsymbol{\gamma}\|_2^2$, where the last inequality holds when $s_{z1}\sqrt{\log p/n} \to 0$. By (114), we have

$$\begin{aligned}\langle \mu^*, \widehat{\boldsymbol{\Sigma}}\mu^*\rangle &\geq \max_{c>0} -\frac{c^2}{4}\langle \boldsymbol{\gamma}, \widehat{\boldsymbol{\Sigma}}\boldsymbol{\gamma}\rangle + c\left(\|\boldsymbol{\gamma}\|_2^2 - \lambda_n\|\boldsymbol{\gamma}\|_1^2\right) \\ &= \frac{\left(\|\boldsymbol{\gamma}\|_2^2 - \lambda_n\|\boldsymbol{\gamma}\|_1^2\right)^2}{\langle \boldsymbol{\gamma}, \widehat{\boldsymbol{\Sigma}}\boldsymbol{\gamma}\rangle} \geq \frac{\|\boldsymbol{\gamma}\|_2^4\left(1 - s_{z1}\lambda_n\right)^2}{\langle \boldsymbol{\gamma}, \widehat{\boldsymbol{\Sigma}}\boldsymbol{\gamma}\rangle}.\end{aligned} \tag{115}$$

On the event G_3, we establish (43) for $\left\|\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j\widehat{\mathbf{v}}^{[j]}\right\|_2^2/n$. The same argument holds for $\left\|\sum_{j\in\mathcal{V}^*}\boldsymbol{\gamma}_j\widehat{\mathbf{v}}^{[j]}\right\|_2^2/n$. Note that

$$\frac{1}{M_2} \leq \boldsymbol{\Theta}_{11} + \beta^2\boldsymbol{\Theta}_{22} - 2\beta\boldsymbol{\Theta}_{12} = \begin{pmatrix}1 & -\beta\end{pmatrix}\boldsymbol{\Theta}\begin{pmatrix}1\\ -\beta\end{pmatrix} \leq M_2\left(1+\beta^2\right). \tag{116}$$

Combined with (43), we establish the first inequality of (44). Note that

$$\frac{1}{n}\left\|\sum_{j\in\mathcal{S}^*}\boldsymbol{\gamma}_j\widehat{\mathbf{v}}^{[j]}\right\|_2^2 \leq \left(2M_2\sum_{j\in\mathcal{V}^*}|\boldsymbol{\gamma}_j|\right)^2 \leq s_{z1}\|\boldsymbol{\gamma}\|_2^2. \tag{117}$$

Combined with (116), we establish the second inequality of (44). By the similar argument, we can establish (45).

E.6. Proof of Lemma 5

In the following proof, we will use the shorthand $\langle a, b\rangle_J = \sum_{j\in J} a_j b_j$. We have the following decompositions for $\widehat{\|\boldsymbol{\gamma}\|_2^2} - \|\boldsymbol{\gamma}\|_2^2$ and $\widehat{\boldsymbol{\gamma}^\intercal\boldsymbol{\Gamma}} - \boldsymbol{\gamma}^\intercal\boldsymbol{\Gamma}$,

$$\begin{aligned}\widehat{\|\boldsymbol{\gamma}\|_2^2} - \|\boldsymbol{\gamma}\|_2^2 =& 2\langle \boldsymbol{\gamma}, D^{\boldsymbol{\gamma}}\rangle_{\widetilde{\mathcal{S}}} + 2\langle \boldsymbol{\gamma}, \Delta^{\boldsymbol{\gamma}}\rangle_{\widetilde{\mathcal{S}}} + \langle D^{\boldsymbol{\gamma}}, D^{\boldsymbol{\gamma}}\rangle_{\widetilde{\mathcal{S}}} + \langle \Delta^{\boldsymbol{\gamma}}, \Delta^{\boldsymbol{\gamma}}\rangle_{\widetilde{\mathcal{S}}} + 2\langle D^{\boldsymbol{\gamma}}, \Delta^{\boldsymbol{\gamma}}\rangle_{\widetilde{\mathcal{S}}} \\ &- \left(\sum_{j\in\mathcal{S}^*\setminus\widetilde{\mathcal{S}}}\boldsymbol{\gamma}_j^2 - \sum_{j\in\widetilde{\mathcal{S}}\setminus\mathcal{S}^*}\boldsymbol{\gamma}_j^2\right),\end{aligned} \tag{118}$$

and

$$\begin{aligned}\widehat{\boldsymbol{\gamma}^\intercal\boldsymbol{\Gamma}} - \boldsymbol{\gamma}^\intercal\boldsymbol{\Gamma} =& \langle \boldsymbol{\gamma}, D^{\boldsymbol{\Gamma}}\rangle_{\widetilde{\mathcal{S}}} + \langle \boldsymbol{\Gamma}, D^{\boldsymbol{\gamma}}\rangle_{\widetilde{\mathcal{S}}} + \langle \boldsymbol{\gamma}, \Delta^{\boldsymbol{\Gamma}}\rangle_{\widetilde{\mathcal{S}}} + \langle \boldsymbol{\Gamma}, \Delta^{\boldsymbol{\gamma}}\rangle_{\widetilde{\mathcal{S}}} + \langle D^{\boldsymbol{\gamma}}, D^{\boldsymbol{\Gamma}}\rangle_{\widetilde{\mathcal{S}}} + \langle \Delta^{\boldsymbol{\gamma}}, \Delta^{\boldsymbol{\Gamma}}\rangle_{\widetilde{\mathcal{S}}} \\ &+ \langle D^{\boldsymbol{\gamma}}, \Delta^{\boldsymbol{\Gamma}}\rangle_{\widetilde{\mathcal{S}}} + \langle D^{\boldsymbol{\Gamma}}, \Delta^{\boldsymbol{\gamma}}\rangle_{\widetilde{\mathcal{S}}} - \left(\sum_{j\in\mathcal{S}^*\setminus\widetilde{\mathcal{S}}}\boldsymbol{\gamma}_j\boldsymbol{\Gamma}_j - \sum_{j\in\widetilde{\mathcal{S}}\setminus\mathcal{S}^*}\boldsymbol{\gamma}_j\boldsymbol{\Gamma}_j\right).\end{aligned} \tag{119}$$

Recall that $\widehat{\mathbf{v}}^{[j]} = \mathbf{W}^\intercal\widehat{\mathbf{u}}^{[j]}$, then we have the following expression

$$\langle \boldsymbol{\gamma}, D^{\boldsymbol{\gamma}}\rangle_{\widetilde{\mathcal{S}}} = \frac{1}{n}\sum_{j\in\widetilde{\mathcal{S}}}\boldsymbol{\gamma}_j(\widehat{\mathbf{v}}^{[j]})^\intercal\boldsymbol{\Pi}_{\cdot 2}, \tag{120}$$

and

$$\langle \boldsymbol{\gamma}, D^{\boldsymbol{\Gamma}} \rangle_{\widetilde{\mathcal{S}}} + \langle \boldsymbol{\Gamma}, D^{\boldsymbol{\gamma}} \rangle_{\widetilde{\mathcal{S}}} = \frac{1}{n} \sum_{j \in \widetilde{\mathcal{S}}} (\widehat{\mathbf{v}}^{[j]})^{\intercal} \left(\boldsymbol{\gamma}_j \boldsymbol{\Pi}_{\cdot 1} + \boldsymbol{\Gamma}_j \boldsymbol{\Pi}_{\cdot 2} \right). \tag{121}$$

Note that $\widetilde{\mathcal{S}}$ is correlated with the error $\boldsymbol{\Pi}_{\cdot 1}$ and $\boldsymbol{\Pi}_{\cdot 2}$. However, we can compare $\widetilde{\mathcal{S}}$ with the true support $\mathcal{S}^*$,

$$\begin{aligned} \langle \boldsymbol{\gamma}, D^{\boldsymbol{\gamma}} \rangle_{\widetilde{\mathcal{S}}} - \langle \boldsymbol{\gamma}, D^{\boldsymbol{\gamma}} \rangle_{\mathcal{S}^*} &= \frac{1}{n} \sum_{j \in \widetilde{\mathcal{S}}} \boldsymbol{\gamma}_j (\widehat{\mathbf{v}}^{[j]})^{\intercal} \boldsymbol{\Pi}_{\cdot 2} - \frac{1}{n} \sum_{j \in \mathcal{S}^*} \boldsymbol{\gamma}_j (\widehat{\mathbf{v}}^{[j]})^{\intercal} \boldsymbol{\Pi}_{\cdot 2} \\ &= \frac{1}{n} \sum_{j \in \widetilde{\mathcal{S}} \setminus \mathcal{S}^*} \boldsymbol{\gamma}_j (\widehat{\mathbf{v}}^{[j]})^{\intercal} \boldsymbol{\Pi}_{\cdot 2} - \frac{1}{n} \sum_{j \in \mathcal{S}^* \setminus \widetilde{\mathcal{S}}} \boldsymbol{\gamma}_j (\widehat{\mathbf{v}}^{[j]})^{\intercal} \boldsymbol{\Pi}_{\cdot 2}, \end{aligned} \tag{122}$$

and

$$\begin{aligned} & \left(\langle \boldsymbol{\gamma}, D^{\boldsymbol{\Gamma}} \rangle_{\widetilde{\mathcal{S}}} + \langle \boldsymbol{\Gamma}, D^{\boldsymbol{\gamma}} \rangle_{\widetilde{\mathcal{S}}} \right) - \left(\langle \boldsymbol{\gamma}, D^{\boldsymbol{\Gamma}} \rangle_{\mathcal{S}^*} + \langle \boldsymbol{\Gamma}, D^{\boldsymbol{\gamma}} \rangle_{\mathcal{S}^*} \right) \\ =& \frac{1}{n} \sum_{j \in \widetilde{\mathcal{S}}} (\widehat{\mathbf{v}}^{[j]})^{\intercal} \left(\boldsymbol{\gamma}_j \boldsymbol{\Pi}_{\cdot 1} + \boldsymbol{\Gamma}_j \boldsymbol{\Pi}_{\cdot 2} \right) - \frac{1}{n} \sum_{j \in \mathcal{S}^*} (\widehat{\mathbf{v}}^{[j]})^{\intercal} \left(\boldsymbol{\gamma}_j \boldsymbol{\Pi}_{\cdot 1} + \boldsymbol{\Gamma}_j \boldsymbol{\Pi}_{\cdot 2} \right) \\ =& \frac{1}{n} \sum_{j \in \widetilde{\mathcal{S}} \setminus \mathcal{S}^*} (\widehat{\mathbf{v}}^{[j]})^{\intercal} \left(\boldsymbol{\gamma}_j \boldsymbol{\Pi}_{\cdot 1} + \boldsymbol{\Gamma}_j \boldsymbol{\Pi}_{\cdot 2} \right) - \frac{1}{n} \sum_{j \in \mathcal{S}^* \setminus \widetilde{\mathcal{S}}} (\widehat{\mathbf{v}}^{[j]})^{\intercal} \left(\boldsymbol{\gamma}_j \boldsymbol{\Pi}_{\cdot 1} + \boldsymbol{\Gamma}_j \boldsymbol{\Pi}_{\cdot 2} \right). \end{aligned} \tag{123}$$

Hence, the residual terms are

$$\begin{aligned} R^{\boldsymbol{\gamma}} =& \sqrt{n} \left(2 \langle \boldsymbol{\gamma}, \Delta^{\boldsymbol{\gamma}} \rangle_{\widetilde{\mathcal{S}}} + \langle D^{\boldsymbol{\gamma}}, D^{\boldsymbol{\gamma}} \rangle_{\widetilde{\mathcal{S}}} + \langle \Delta^{\boldsymbol{\gamma}}, \Delta^{\boldsymbol{\gamma}} \rangle_{\widetilde{\mathcal{S}}} + 2 \langle D^{\boldsymbol{\gamma}}, \Delta^{\boldsymbol{\gamma}} \rangle_{\widetilde{\mathcal{S}}} \right) \\ &+ \sqrt{n} \left(\frac{1}{n} \sum_{j \in \widetilde{\mathcal{S}} \setminus \mathcal{S}^*} \boldsymbol{\gamma}_j (\widehat{\mathbf{v}}^{[j]})^{\intercal} \boldsymbol{\Pi}_{\cdot 2} - \frac{1}{n} \sum_{j \in \mathcal{S}^* \setminus \widetilde{\mathcal{S}}} \boldsymbol{\gamma}_j (\widehat{\mathbf{v}}^{[j]})^{\intercal} \boldsymbol{\Pi}_{\cdot 2} \right) - \sqrt{n} \left(\sum_{j \in \mathcal{S}^* \setminus \widetilde{\mathcal{S}}} \boldsymbol{\gamma}_j^2 - \sum_{j \in \widetilde{\mathcal{S}} \setminus \mathcal{S}^*} \boldsymbol{\gamma}_j^2 \right), \end{aligned} \tag{124}$$

and

$$\begin{aligned} R^{\text{inter}} =& \sqrt{n} \left(\langle \boldsymbol{\gamma}, \Delta^{\boldsymbol{\Gamma}} \rangle_{\widetilde{\mathcal{S}}} + \langle \boldsymbol{\Gamma}, \Delta^{\boldsymbol{\gamma}} \rangle_{\widetilde{\mathcal{S}}} + \langle D^{\boldsymbol{\gamma}}, D^{\boldsymbol{\Gamma}} \rangle_{\widetilde{\mathcal{S}}} + \langle \Delta^{\boldsymbol{\gamma}}, \Delta^{\boldsymbol{\Gamma}} \rangle_{\widetilde{\mathcal{S}}} + \langle D^{\boldsymbol{\gamma}}, \Delta^{\boldsymbol{\Gamma}} \rangle_{\widetilde{\mathcal{S}}} + \langle D^{\boldsymbol{\Gamma}}, \Delta^{\boldsymbol{\gamma}} \rangle_{\widetilde{\mathcal{S}}} \right) \\ &+ \sqrt{n} \left(\frac{1}{n} \sum_{j \in \widetilde{\mathcal{S}} \setminus \mathcal{S}^*} (\widehat{\mathbf{v}}^{[j]})^{\intercal} \left(\boldsymbol{\gamma}_j \boldsymbol{\Pi}_{\cdot 1} + \boldsymbol{\Gamma}_j \boldsymbol{\Pi}_{\cdot 2} \right) - \frac{1}{n} \sum_{j \in \mathcal{S}^* \setminus \widetilde{\mathcal{S}}} (\widehat{\mathbf{v}}^{[j]})^{\intercal} \left(\boldsymbol{\gamma}_j \boldsymbol{\Pi}_{\cdot 1} + \boldsymbol{\Gamma}_j \boldsymbol{\Pi}_{\cdot 2} \right) \right) \\ &- \sqrt{n} \left(\sum_{j \in \mathcal{S}^* \setminus \widetilde{\mathcal{S}}} \boldsymbol{\gamma}_j \boldsymbol{\Gamma}_j - \sum_{j \in \widetilde{\mathcal{S}} \setminus \mathcal{S}^*} \boldsymbol{\gamma}_j \boldsymbol{\Gamma}_j \right). \end{aligned} \tag{125}$$

Define $\mathcal{S}_0^* = \left\{ j : |\boldsymbol{\gamma}_j| > \sqrt{2.05 \log p_{\text{z}}} \sqrt{\text{Var}\left(D_j^{\boldsymbol{\gamma}} \right)} \right\}$ to be the set of strong signals, on the event A_2, we have

$$\mathcal{S}_0^* \subset \widetilde{\mathcal{S}} \subset \mathcal{S}^*, \quad \text{and} \quad \left| \widetilde{\mathcal{S}} \right| \leq s_{\text{z}1}. \tag{126}$$

On the event A_3, we have

$$\max\left\{\left\|D^{\mathbf{\Gamma}}\right\|_\infty, \|D^{\boldsymbol{\gamma}}\|_\infty\right\} \leq \left(1+12\sqrt{\frac{\log p}{n}}\right) M_1 \sqrt{\frac{2.05\log p_{\mathrm{z}}}{n}} \max\{\sigma_1,\sigma_2\}. \quad (127)$$

On the event $S\cap G\cap A$,

$$\max\left\{\|\Delta^{\boldsymbol{\gamma}}\|_\infty, \|\Delta^{\mathbf{\Gamma}}\|_\infty\right\} \leq Cs\frac{\log p}{n}\max\{\sigma_1,\sigma_2\}. \quad (128)$$

Combing (126), (127) and (128), we have on the event $S\cap G\cap A$,

$$\max\left\{\langle D^{\boldsymbol{\gamma}}, D^{\boldsymbol{\gamma}}\rangle_{\widetilde{\mathcal{S}}}, \langle D^{\mathbf{\Gamma}}, D^{\mathbf{\Gamma}}\rangle_{\widetilde{\mathcal{S}}}\right\} \leq Cs_{\mathrm{z}1}\frac{\log p_{\mathrm{z}}}{n}, \quad (129)$$

$$\max\left\{\langle \Delta^{\boldsymbol{\gamma}}, \Delta^{\boldsymbol{\gamma}}\rangle_{\widetilde{\mathcal{S}}}, \langle \Delta^{\mathbf{\Gamma}}, \Delta^{\mathbf{\Gamma}}\rangle_{\widetilde{\mathcal{S}}}\right\} \leq Cs_{\mathrm{z}1}\left(s\frac{\log p}{n}\right)^2. \quad (130)$$

Note that

$$\left|\langle D^{\boldsymbol{\gamma}}, \Delta^{\boldsymbol{\gamma}}\rangle_{\widetilde{\mathcal{S}}}\right| \leq \sqrt{\langle D^{\boldsymbol{\gamma}}, D^{\boldsymbol{\gamma}}\rangle_{\widetilde{\mathcal{S}}}\langle \Delta^{\boldsymbol{\gamma}}, \Delta^{\boldsymbol{\gamma}}\rangle_{\widetilde{\mathcal{S}}}} \leq \frac{1}{2}\left(\langle D^{\boldsymbol{\gamma}}, D^{\boldsymbol{\gamma}}\rangle_{\widetilde{\mathcal{S}}} + \langle \Delta^{\boldsymbol{\gamma}}, \Delta^{\boldsymbol{\gamma}}\rangle_{\widetilde{\mathcal{S}}}\right).$$

Hence, we have

$$\left|\langle D^{\boldsymbol{\gamma}}, D^{\boldsymbol{\gamma}}\rangle_{\widetilde{\mathcal{S}}} + \langle \Delta^{\boldsymbol{\gamma}}, \Delta^{\boldsymbol{\gamma}}\rangle_{\widetilde{\mathcal{S}}} + 2\langle D^{\boldsymbol{\gamma}}, \Delta^{\boldsymbol{\gamma}}\rangle_{\widetilde{\mathcal{S}}}\right| \leq Cs_{\mathrm{z}1}\frac{\log p_{\mathrm{z}}}{n} + Cs_{\mathrm{z}1}\left(s\frac{\log p}{n}\right)^2, \quad (131)$$

and

$$\left|\langle D^{\boldsymbol{\gamma}}, D^{\mathbf{\Gamma}}\rangle_{\widetilde{\mathcal{S}}} + \langle \Delta^{\boldsymbol{\gamma}}, \Delta^{\mathbf{\Gamma}}\rangle_{\widetilde{\mathcal{S}}} + \langle D^{\boldsymbol{\gamma}}, \Delta^{\mathbf{\Gamma}}\rangle_{\widetilde{\mathcal{S}}} + \langle D^{\mathbf{\Gamma}}, \Delta^{\boldsymbol{\gamma}}\rangle_{\widetilde{\mathcal{S}}}\right| \leq Cs_{\mathrm{z}1}\frac{\log p_{\mathrm{z}}}{n} + Cs_{\mathrm{z}1}\left(s\frac{\log p}{n}\right)^2. \quad (132)$$

We also have the following control

$$2\left|\langle \boldsymbol{\gamma}, \Delta^{\boldsymbol{\gamma}}\rangle_{\widetilde{\mathcal{S}}}\right| \leq \|\boldsymbol{\gamma}\|_2\sqrt{\langle \Delta^{\boldsymbol{\gamma}}, \Delta^{\boldsymbol{\gamma}}\rangle_{\widetilde{\mathcal{S}}}} \leq C\|\boldsymbol{\gamma}\|_2\sqrt{s_{\mathrm{z}1}}s\frac{\log p}{n}, \quad (133)$$

and

$$\left|\langle \boldsymbol{\gamma}, \Delta^{\mathbf{\Gamma}}\rangle_{\widetilde{\mathcal{S}}} + \langle \mathbf{\Gamma}, \Delta^{\boldsymbol{\gamma}}\rangle_{\widetilde{\mathcal{S}}}\right| \leq C\left(\|\boldsymbol{\gamma}\|_2 + \|\mathbf{\Gamma}\|_2\right)\sqrt{s_{\mathrm{z}1}}s\frac{\log p}{n}. \quad (134)$$

On the event $S\cap G\cap A$, we have $\widetilde{\mathcal{S}}\backslash\mathcal{S}^* = \varnothing$ and hence $\frac{1}{n}\sum_{j\in\widetilde{\mathcal{S}}\backslash\mathcal{S}^*}\boldsymbol{\gamma}_j(\widehat{\mathbf{v}}^{[j]})^\intercal\mathbf{\Pi}_{\cdot 2} = 0$, $\sum_{j\in\widetilde{\mathcal{S}}\backslash\mathcal{S}^*}\boldsymbol{\gamma}_j\mathbf{\Gamma}_j = 0$ and $\sum_{j\in\widetilde{\mathcal{S}}\backslash\mathcal{S}^*}\boldsymbol{\gamma}_j^2 = 0$; On the event $S\cap G\cap A$, we also have

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{j\in\mathcal{S}^*\backslash\widetilde{\mathcal{S}}}\boldsymbol{\gamma}_j(\widehat{\mathbf{v}}^{[j]})^\intercal\mathbf{\Pi}_{\cdot 2}\right| \leq \frac{1}{n}s_{\mathrm{z}1}\max_{j\in\mathcal{S}^*\backslash\widetilde{\mathcal{S}}}|\boldsymbol{\gamma}_j|\left|(\widehat{\mathbf{v}}^{[j]})^\intercal\mathbf{\Pi}_{\cdot 2}\right| \\
\leq& s_{\mathrm{z}1}\sqrt{2.05\log p_{\mathrm{z}}}\sqrt{\mathrm{Var}\left(D_j^{\boldsymbol{\gamma}}\right)}\left(1+12\sqrt{\frac{\log p}{n}}\right)M_1\sqrt{\frac{2.05\log p_{\mathrm{z}}}{n}}\sigma_2 \leq \frac{s_{\mathrm{z}1}\log p_{\mathrm{z}}}{n}.
\end{aligned} \quad (135)$$

On the event $S \cap G \cap A$, we get

$$\left| \sum_{j \in \mathcal{S}^* \setminus \widetilde{\mathcal{S}}} \boldsymbol{\gamma}_j^2 \right| \leq s_{\mathrm{z}1} \frac{\log p_{\mathrm{z}}}{n}. \tag{136}$$

By (124), (131), (133), (135) and (136), we establish that on the event $S \cap G \cap A$,

$$|R^{\boldsymbol{\gamma}}| \leq C s_{\mathrm{z}1} \frac{\log p_{\mathrm{z}}}{\sqrt{n}} + C \|\boldsymbol{\gamma}\|_2 \sqrt{s_{\mathrm{z}1}} s \frac{\log p}{\sqrt{n}}. \tag{137}$$

Similarly, we can establish that and

$$\left|R^{\mathrm{inter}}\right| \leq C s_{\mathrm{z}1} \frac{\log p_{\mathrm{z}}}{\sqrt{n}} + C \left(\|\boldsymbol{\gamma}\|_2 + \|\boldsymbol{\Gamma}\|_2\right) \sqrt{s_{\mathrm{z}1}} s \frac{\log p}{\sqrt{n}}. \tag{138}$$

We can establish (47) and (50) by taking $\boldsymbol{\Gamma}_j = \beta \boldsymbol{\gamma}_j$. Note that

$$\begin{aligned} &\frac{2\sqrt{\log p}}{\sqrt{n}} \left\| \sum_{j \in \mathcal{S}^*} \boldsymbol{\gamma}_j \widehat{\mathbf{v}}^{[j]} \right\|_2 \sqrt{\boldsymbol{\Theta}_{22}} = 2 \sqrt{\frac{\log p}{n}} \sqrt{\frac{\boldsymbol{\Theta}_{22}}{\boldsymbol{\Theta}_{11} + \beta^2 \boldsymbol{\Theta}_{22} - 2\beta \boldsymbol{\Theta}_{12}}} \sqrt{\mathrm{V}_H} \|\boldsymbol{\gamma}\|_2^2, \\ &\frac{\sqrt{\log p}}{\sqrt{n}} \left\| \sum_{j \in \mathcal{S}^*} \boldsymbol{\gamma}_j \widehat{\mathbf{v}}^{[j]} \right\|_2 \sqrt{\boldsymbol{\Theta}_{11} + \beta^2 \boldsymbol{\Theta}_{22} + 2\beta \boldsymbol{\Theta}_{12}} = \sqrt{\frac{\log p}{n}} \sqrt{\frac{\boldsymbol{\Theta}_{11} + \beta^2 \boldsymbol{\Theta}_{22} + 2\beta \boldsymbol{\Theta}_{12}}{\boldsymbol{\Theta}_{11} + \beta^2 \boldsymbol{\Theta}_{22} - 2\beta \boldsymbol{\Theta}_{12}}} \sqrt{\mathrm{V}_H} \|\boldsymbol{\gamma}\|_2^2. \end{aligned} \tag{139}$$

By the definition of A_4 and A_5 in (37) and Lemma 4, we establish

$$\max \left\{ \left| \frac{2}{n} \sum_{j \in \mathcal{S}^*} \boldsymbol{\gamma}_j (\widehat{\mathbf{v}}^{[j]})^\intercal \boldsymbol{\Pi}_{\cdot 2} \right|, \left| \frac{1}{n} \sum_{j \in \mathcal{S}^*} \boldsymbol{\gamma}_j (\widehat{\mathbf{v}}^{[j]})^\intercal \left(\boldsymbol{\Pi}_{\cdot 1} + \beta \boldsymbol{\Pi}_{\cdot 2} \right) \right| \right\} \leq C s_{\mathrm{z}1} \sqrt{\frac{\log p}{n}} \|\boldsymbol{\gamma}\|_2 \tag{140}$$

Combined with (50), we establish (51).

E.7. Proof of Lemma 8

The proof of Lemma 8 is similar to that of Lemma 5 and we will present it here. Similar to (118) and (119), we can obtain the following expressions,

$$\sum_{j \in \widetilde{\mathcal{V}}} \widetilde{\boldsymbol{\gamma}}_j^2 - \sum_{j \in \widetilde{\mathcal{V}}} \boldsymbol{\gamma}_j^2 = \frac{1}{n} \sum_{j \in \widetilde{\mathcal{V}}} \boldsymbol{\gamma}_j (\widehat{\mathbf{v}}^{[j]})^\intercal \boldsymbol{\Pi}_{\cdot 2} + \langle D^{\boldsymbol{\gamma}}, D^{\boldsymbol{\gamma}} \rangle_{\widetilde{\mathcal{V}}} + \langle \Delta^{\boldsymbol{\gamma}}, \Delta^{\boldsymbol{\gamma}} \rangle_{\widetilde{\mathcal{V}}} + 2 \langle D^{\boldsymbol{\gamma}}, \Delta^{\boldsymbol{\gamma}} \rangle_{\widetilde{\mathcal{V}}}, \tag{141}$$

and

$$\begin{aligned} \sum_{j \in \widetilde{\mathcal{V}}} \widetilde{\boldsymbol{\gamma}}_j \widetilde{\boldsymbol{\Gamma}}_j - \sum_{j \in \widetilde{\mathcal{V}}} \boldsymbol{\gamma}_j \boldsymbol{\Gamma}_j = &\frac{1}{n} \sum_{j \in \widetilde{\mathcal{V}}} (\widehat{\mathbf{v}}^{[j]})^\intercal \left(\boldsymbol{\gamma}_j \boldsymbol{\Pi}_{\cdot 1} + \boldsymbol{\Gamma}_j \boldsymbol{\Pi}_{\cdot 2} \right) + \langle \boldsymbol{\gamma}, \Delta^{\boldsymbol{\Gamma}} \rangle_{\widetilde{\mathcal{V}}} + \langle \boldsymbol{\Gamma}, \Delta^{\boldsymbol{\gamma}} \rangle_{\widetilde{\mathcal{V}}}, \\ &+ \langle D^{\boldsymbol{\gamma}}, D^{\boldsymbol{\Gamma}} \rangle_{\widetilde{\mathcal{V}}} + \langle \Delta^{\boldsymbol{\gamma}}, \Delta^{\boldsymbol{\Gamma}} \rangle_{\widetilde{\mathcal{V}}} + \langle D^{\boldsymbol{\gamma}}, \Delta^{\boldsymbol{\Gamma}} \rangle_{\widetilde{\mathcal{V}}} + \langle D^{\boldsymbol{\Gamma}}, \Delta^{\boldsymbol{\gamma}} \rangle_{\widetilde{\mathcal{V}}}, \end{aligned} \tag{142}$$

On the event $A \cap S \cap G$, we have

$$\frac{1}{n}\sum_{j\in\widetilde{\mathcal{V}}} \boldsymbol{\gamma}_j (\widehat{\mathbf{v}}^{[j]})^\intercal \boldsymbol{\Pi}_{\cdot 2} \leq \sqrt{|\widetilde{\mathcal{V}}|}\|\boldsymbol{\gamma}\|_2 \max_{j\in\widetilde{\mathcal{V}}} \frac{\|\widehat{\mathbf{v}}^{[j]}\|_2}{n}\sqrt{2.05\log p_{\mathrm{z}}} \leq C\|\boldsymbol{\gamma}\|_2\sqrt{\frac{|\widetilde{\mathcal{V}}|\log p_{\mathrm{z}}}{n}}, \tag{143}$$

where the last inequality follows from (42). Combined with the fact $\boldsymbol{\Gamma}_j = \boldsymbol{\gamma}_j(\beta + \boldsymbol{\pi}_j/\boldsymbol{\gamma}_j)$ and $|\boldsymbol{\pi}_j/\boldsymbol{\gamma}_j| \leq C\sqrt{\log p_{\mathrm{z}}/n}$, we also establish that

$$\frac{1}{n}\sum_{j\in\widetilde{\mathcal{V}}} (\widehat{\mathbf{v}}^{[j]})^\intercal \left(\boldsymbol{\gamma}_j \boldsymbol{\Pi}_{\cdot 1} + \boldsymbol{\Gamma}_j \boldsymbol{\Pi}_{\cdot 2}\right) \leq \sqrt{|\widetilde{\mathcal{V}}|}\|\boldsymbol{\gamma}\|_2 \max_{j\in\widetilde{\mathcal{V}}} \frac{\|\widehat{\mathbf{v}}^{[j]}\|_2}{n}\sqrt{2.05\log p_{\mathrm{z}}} \leq C\|\boldsymbol{\gamma}\|_2\sqrt{\frac{|\widetilde{\mathcal{V}}|\log p_{\mathrm{z}}}{n}}. \tag{144}$$

By (143), (144), (131) and (132), we obtain the following inequalities

$$\begin{aligned} &\max\left\{\left|\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_j\widetilde{\boldsymbol{\Gamma}}_j - \sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j\boldsymbol{\Gamma}_j\right|, \left|\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_j^2 - \sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j^2\right|\right\} \\ &\leq Cs_{\mathrm{z}1}\frac{\log p_{\mathrm{z}}}{n} + Cs_{\mathrm{z}1}\left(s\frac{\log p}{n}\right)^2 + C\|\boldsymbol{\gamma}_{\widetilde{\mathcal{V}}}\|_2\sqrt{\frac{2|\widetilde{\mathcal{V}}|\log p_{\mathrm{z}}}{n}}. \end{aligned} \tag{145}$$

Since $|\widetilde{\mathcal{V}}| > |\mathcal{V}^*|/2$ and $\min_{j\in\mathcal{S}^*}|\boldsymbol{\gamma}_j| \geq \delta_{\min} \gg \sqrt{\log p_{\mathrm{z}}/n}$, we have $\left|\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_j^2 - \sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j^2\right| \ll \sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j^2$. We have the following decomposition,

$$\begin{aligned} &\left|\frac{\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_j\widetilde{\boldsymbol{\Gamma}}_j}{\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_j^2} - \frac{\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j\boldsymbol{\Gamma}_j}{\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j^2}\right| \\ \leq &\frac{\left(\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j^2\right)\left|\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_j\widetilde{\boldsymbol{\Gamma}}_j - \sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j\boldsymbol{\Gamma}_j\right| + \left|\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j\boldsymbol{\Gamma}_j\right|\left|\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_j^2 - \sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j^2\right|}{\left(\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_j^2\right)\left(\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j^2\right)} \\ \leq &C\frac{\max\left\{\left|\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_j\widetilde{\boldsymbol{\Gamma}}_j - \sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j\boldsymbol{\Gamma}_j\right|, \left|\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_j^2 - \sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j^2\right|\right\}}{\sum_{j\in\widetilde{\mathcal{V}}}\boldsymbol{\gamma}_j^2} \leq C\frac{1}{\delta_{\min}}\sqrt{\frac{\log p_{\mathrm{z}}}{n}}, \end{aligned} \tag{146}$$

where the last inequality follows from (145) and the facts that $|\widetilde{\mathcal{V}}| > |\mathcal{V}^*|/2$, $\min_{j\in\mathcal{S}^*}|\boldsymbol{\gamma}_j| \geq \delta_{\min} \gg \sqrt{\log p_{\mathrm{z}}/n}$ and $s\log p/\sqrt{n} \to 0$.

E.8. Proof of Lemma 7

Since $\min_{j\in\mathcal{S}^*}|\boldsymbol{\gamma}_j| \geq \delta_{\min} \gg \sqrt{\log p/n}$, on the event $A \cap S \cap G$, we have

$$\widetilde{\mathcal{S}} = \mathcal{S}^*.$$

Without loss of generality, we assume $1 \in \widetilde{\mathcal{S}}$ and focus on the case $i = 1$. In the following, we are going to analyze the performance of $\widehat{\beta}^{[1]}$ and $\widetilde{\boldsymbol{\pi}}_j^{[1]}$. In the following,

we first analyze $\widehat{\beta}^{[1]}$,

$$\sqrt{n}\left(\widehat{\beta}^{[1]}-\left(\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\right)=T^{\beta,1}+\Delta^{\beta,1}, \tag{147}$$

where

$$T^{\beta,1}=\frac{1}{\sqrt{n}\boldsymbol{\gamma}_1}(\widehat{\mathbf{v}}^{[1]})^{\intercal}\left(\boldsymbol{\Pi}_{\cdot 1}-\left(\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\boldsymbol{\Pi}_{\cdot 2}\right) \quad \text{and} \quad \Delta^{\beta,1}=R_1+R_2, \tag{148}$$

with

$$R_1=\frac{\sqrt{n}}{\boldsymbol{\gamma}_1}\left(\Delta_1^{\boldsymbol{\Gamma}}-\left(\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\Delta_1^{\gamma}\right), \text{ and } R_2=\frac{-\left(D_1^{\gamma}+\Delta_1^{\gamma}\right)}{\boldsymbol{\gamma}_1+\left(D_1^{\gamma}+\Delta_1^{\gamma}\right)}\left(T^{\beta,1}+R_1\right). \tag{149}$$

To analyze $\widetilde{\boldsymbol{\pi}}^{[1]}$, we first analyze the following estimator,

$$\widehat{\boldsymbol{\pi}}^{[1]}=\widetilde{\boldsymbol{\Gamma}}-\widehat{\beta}^{[1]}\widetilde{\boldsymbol{\gamma}}. \tag{150}$$

Note that

$$\begin{aligned}\widehat{\boldsymbol{\pi}}_j^{[1]}-\boldsymbol{\pi}_j=&-\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\gamma}_j+\left(\widetilde{\boldsymbol{\Gamma}}_j-\boldsymbol{\Gamma}_j\right)-\left(\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\left(\widetilde{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j\right)-\boldsymbol{\gamma}_j\left(\widehat{\beta}^{[1]}-\left(\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\right)\\&-\left(\widehat{\beta}^{[1]}-\left(\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\right)\left(\widetilde{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j\right).\end{aligned} \tag{151}$$

By (101) and (102), we have

$$\sqrt{n}\left(\widetilde{\boldsymbol{\Gamma}}_j-\boldsymbol{\Gamma}_j\right)=\frac{1}{\sqrt{n}}(\widehat{\mathbf{v}}^{[j]})^{\intercal}\boldsymbol{\Pi}_{\cdot 1}+\sqrt{n}\Delta_j^{\boldsymbol{\Gamma}}. \tag{152}$$

By (103) and (104), we have

$$\sqrt{n}\left(\widetilde{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j\right)=\frac{1}{\sqrt{n}}(\widehat{\mathbf{v}}^{[j]})^{\intercal}\boldsymbol{\Pi}_{\cdot 2}+\sqrt{n}\Delta_j^{\gamma}. \tag{153}$$

By plugging (148), (152) and (153) into (151), we have the following decomposition of $\widehat{\boldsymbol{\pi}}_j^{[1]}-\boldsymbol{\pi}_j$

$$\sqrt{n}\left(\widehat{\boldsymbol{\pi}}_j^{[1]}-\boldsymbol{\pi}_j\right)=-\sqrt{n}\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\gamma}_j+T^{\boldsymbol{\pi}_j}+\Delta^{\boldsymbol{\pi}_j}, \tag{154}$$

where

$$T^{\boldsymbol{\pi}_j}=\frac{1}{\sqrt{n}}\left((\widehat{\mathbf{v}}^{[j]})^{\intercal}-\frac{\boldsymbol{\gamma}_j}{\boldsymbol{\gamma}_1}(\widehat{\mathbf{v}}^{[1]})^{\intercal}\right)\left(\boldsymbol{\Pi}_{\cdot 1}-\left(\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\boldsymbol{\Pi}_{\cdot 2}\right),$$

and

$$\Delta^{\boldsymbol{\pi}_j}=\sqrt{n}\left(\Delta_j^{\boldsymbol{\Gamma}}-\left(\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\Delta_j^{\gamma}-\boldsymbol{\gamma}_j\Delta^{\beta,1}\right)-\left(T^{\beta,1}+\Delta^{\beta,1}\right)\left(\widetilde{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j\right). \tag{155}$$

Define the events for $i \in \mathcal{S}^*$,

$$F^i = \left\{ \max_{j \in \mathcal{S}^*, j \neq i} |T^{\boldsymbol{\pi}_j}| \leq 2.02\sqrt{\log p_{\mathrm{z}}} \sqrt{\boldsymbol{\Theta}_{11} + \left(\beta + \frac{\boldsymbol{\pi}_i}{\boldsymbol{\gamma}_i}\right)^2 \boldsymbol{\Theta}_{22} - 2\left(\beta + \frac{\boldsymbol{\pi}_i}{\boldsymbol{\gamma}_i}\right)\boldsymbol{\Theta}_{12}} \frac{\left\| \widehat{\mathbf{v}}^{[j]} - \frac{\boldsymbol{\gamma}_j}{\boldsymbol{\gamma}_i}\widehat{\mathbf{v}}^{[i]} \right\|_2}{\sqrt{n}}. \right\}$$

Then for $F = \cap_{i \in \mathcal{S}^*} F^i$, we have

$$\mathbf{P}(F) \geq 1 - C s_{\mathrm{z}1}^2 p_{\mathrm{z}}^{-2.04} \geq 1 - c p^{-c}. \tag{156}$$

The proof of Lemma 7 relies on the following lemmas. The following lemma provides upper bound and lower bound for the variance term and the proof of the following lemma can be found in Section E.11.

LEMMA 12. *On the event $A \cap S \cap G$, we have*

$$\begin{aligned} &\sqrt{\boldsymbol{\Theta}_{11} + \left(\beta + \left|\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right|\right)^2 \boldsymbol{\Theta}_{22} - 2\left(\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\boldsymbol{\Theta}_{12}} \frac{\left\| \widehat{\mathbf{v}}^{[j]} - \frac{\boldsymbol{\gamma}_j}{\boldsymbol{\gamma}_1}\widehat{\mathbf{v}}^{[1]} \right\|_2}{\sqrt{n}} \\ &\leq 1.1\sqrt{M_1 M_2}\left(1 + \left|\frac{\boldsymbol{\gamma}_j}{\boldsymbol{\gamma}_1}\right|\right)\sqrt{1 + \left(\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)^2}, \end{aligned} \tag{157}$$

and

$$\begin{aligned} &\sqrt{\boldsymbol{\Theta}_{11} + \left(\beta + \left|\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right|\right)^2 \boldsymbol{\Theta}_{22} - 2\left(\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\boldsymbol{\Theta}_{12}} \frac{\left\| \widehat{\mathbf{v}}^{[j]} - \frac{\boldsymbol{\gamma}_j}{\boldsymbol{\gamma}_1}\widehat{\mathbf{v}}^{[1]} \right\|_2}{\sqrt{n}} \\ &\geq 0.45\sqrt{\frac{M_1}{M_2}}\left(1 + \left|\frac{\boldsymbol{\gamma}_j}{\boldsymbol{\gamma}_1}\right|\right)\sqrt{1 + \left(\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)^2}. \end{aligned} \tag{158}$$

LEMMA 13. *On the event $A \cap S \cap G \cap F^1$, for large n, we have*

$$0.995 \leq \frac{\sqrt{\widehat{\boldsymbol{\Theta}}_{11} + \left(\widehat{\beta}^{[1]}\right)^2 \widehat{\boldsymbol{\Theta}}_{22} - 2\widehat{\beta}^{[1]}\widehat{\boldsymbol{\Theta}}_{12}} \left\| \widehat{\mathbf{v}}^{[j]} - \frac{\widetilde{\boldsymbol{\gamma}}_j}{\widetilde{\boldsymbol{\gamma}}_1}\widehat{\mathbf{v}}^{[1]} \right\|_2}{\sqrt{\boldsymbol{\Theta}_{11} + \left(\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)^2 \boldsymbol{\Theta}_{22} - 2\left(\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\boldsymbol{\Theta}_{12}} \left\| \widehat{\mathbf{v}}^{[j]} - \frac{\boldsymbol{\gamma}_j}{\boldsymbol{\gamma}_1}\widehat{\mathbf{v}}^{[1]} \right\|_2} \leq 1.005. \tag{159}$$

On the event $A \cap S \cap G \cap F^1$, we have

$$\max_{j \in \mathcal{S}^*} \frac{1}{\sqrt{n}} |T^{\boldsymbol{\pi}_j}| \leq 2.02\sqrt{\frac{\log p_{\mathrm{z}}}{n}} \sqrt{\boldsymbol{\Theta}_{11} + \left(\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)^2 \boldsymbol{\Theta}_{22} - 2\left(\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\boldsymbol{\Theta}_{12}} \frac{\left\| \widehat{\mathbf{v}}^{[j]} - \frac{\boldsymbol{\gamma}_j}{\boldsymbol{\gamma}_1}\widehat{\mathbf{v}}^{[1]} \right\|_2}{\sqrt{n}}; \tag{160}$$

and

$$\max_{j \in \mathcal{S}^*} \frac{1}{\sqrt{n}} |\Delta^{\boldsymbol{\pi}_j}| \leq \frac{1}{300}\sqrt{\frac{\log p_{\mathrm{z}}}{n}} \sqrt{\boldsymbol{\Theta}_{11} + \left(\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)^2 \boldsymbol{\Theta}_{22} - 2\left(\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\boldsymbol{\Theta}_{12}} \frac{\left\| \widehat{\mathbf{v}}^{[j]} - \frac{\boldsymbol{\gamma}_j}{\boldsymbol{\gamma}_1}\widehat{\mathbf{v}}^{[1]} \right\|_2}{\sqrt{n}}. \tag{161}$$

The ratio $\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}$ can be divided into the following three cases,

(a) strongly invalid instrument case, $|\boldsymbol{\pi}_1/\boldsymbol{\gamma}_1| \geq C_*(1/\delta_{\min})\sqrt{\log p_z/n}$, where $C_* = 12(1+|\beta|)\sqrt{M_1/M_2}$;

(b) weakly invalid instrument case, $|\boldsymbol{\pi}_1/\boldsymbol{\gamma}_1| < C_*(1/\delta_{\min})\sqrt{\log p_z/n}$;

(c) valid instrument case, $\boldsymbol{\pi}_1/\boldsymbol{\gamma}_1 = 0$.

We are going to show that our procedure (12) in the main paper will rule out the strong invalid instrument case and a stronger assumption (22) in the main paper will help use rule out the weakly invalid instrument case. In the following, we will analyze the three cases separately.

strongly invalid instrument case

In this case, we assume that $|\boldsymbol{\pi}_1/\boldsymbol{\gamma}_1| \geq C_*(1/\delta_{\min})\sqrt{\log p_z/n}$. For $j \in \mathcal{V}^*$, (154) can be re-expressed as

$$\sqrt{n}\left(\widehat{\boldsymbol{\pi}}_j^{[1]} - 0\right) = -\sqrt{n}\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\gamma}_j + T^{\boldsymbol{\pi}_j} + \Delta^{\boldsymbol{\pi}_j}. \tag{162}$$

We are going to show that on the event $A \cap S \cap G \cap F^1$,

$$\|\widetilde{\boldsymbol{\pi}}^{[1]}\|_0 > \frac{|\mathcal{S}^*|}{2}. \tag{163}$$

It is sufficient to show for $j \in \mathcal{V}^*$

$$\left|-\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\gamma}_j + \frac{1}{\sqrt{n}}\left(T^{\boldsymbol{\pi}_j} + \Delta^{\boldsymbol{\pi}_j}\right)\right| \geq 2.05\sqrt{\widehat{\boldsymbol{\Theta}}_{11} + \left(\widehat{\beta}^{[1]}\right)^2\widehat{\boldsymbol{\Theta}}_{22} - 2\widehat{\beta}^{[1]}\widehat{\boldsymbol{\Theta}}_{12}}\sqrt{\frac{\log p_z}{n}}\frac{\left\|\widehat{\mathbf{v}}^{[j]} - \frac{\widetilde{\boldsymbol{\gamma}}_j}{\widetilde{\boldsymbol{\gamma}}_1}\widehat{\mathbf{v}}^{[1]}\right\|_2}{\sqrt{n}}, \tag{164}$$

which can be reduced to

$$\max_{j\in\mathcal{V}^*}\frac{1}{\sqrt{n}}\left|T^{\boldsymbol{\pi}_j} + \Delta^{\boldsymbol{\pi}_j}\right| \leq 2.05\sqrt{\widehat{\boldsymbol{\Theta}}_{11} + \left(\widehat{\beta}^{[1]}\right)^2\widehat{\boldsymbol{\Theta}}_{22} - 2\widehat{\beta}^{[1]}\widehat{\boldsymbol{\Theta}}_{12}}\sqrt{\frac{\log p_z}{n}}\frac{\left\|\widehat{\mathbf{v}}^{[j]} - \frac{\widetilde{\boldsymbol{\gamma}}_j}{\widetilde{\boldsymbol{\gamma}}_1}\widehat{\mathbf{v}}^{[1]}\right\|_2}{\sqrt{n}}; \tag{165}$$

and

$$\left|\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\gamma}_j\right| \geq 4.1\sqrt{\widehat{\boldsymbol{\Theta}}_{11} + \left(\widehat{\beta}^{[1]}\right)^2\widehat{\boldsymbol{\Theta}}_{22} - 2\widehat{\beta}^{[1]}\widehat{\boldsymbol{\Theta}}_{12}}\sqrt{\frac{\log p_z}{n}}\frac{\left\|\widehat{\mathbf{v}}^{[j]} - \frac{\widetilde{\boldsymbol{\gamma}}_j}{\widetilde{\boldsymbol{\gamma}}_1}\widehat{\mathbf{v}}^{[1]}\right\|_2}{\sqrt{n}}. \tag{166}$$

By (159), (160) and (161), we establish (165). By (157) and (159), we have

$$\begin{aligned}
&4.1\sqrt{\widehat{\boldsymbol{\Theta}}_{11} + \left(\widehat{\beta}^{[1]}\right)^2\widehat{\boldsymbol{\Theta}}_{22} - 2\widehat{\beta}^{[1]}\widehat{\boldsymbol{\Theta}}_{12}}\sqrt{\frac{\log p_z}{n}}\frac{\left\|\widehat{\mathbf{v}}^{[j]} - \frac{\widetilde{\boldsymbol{\gamma}}_j}{\widetilde{\boldsymbol{\gamma}}_1}\widehat{\mathbf{v}}^{[1]}\right\|_2}{\sqrt{n}} \\
\leq& 4.1 \times 1.005 \times 1.1\sqrt{M_1M_2}\left(1 + \left|\frac{\boldsymbol{\gamma}_j}{\boldsymbol{\gamma}_1}\right|\right)\sqrt{1 + \left(\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)^2}\sqrt{\frac{\log p_z}{n}} \\
\leq& |\boldsymbol{\gamma}_j|\, 4.1 \times 1.005 \times 1.1\sqrt{M_1M_2}\left(\left|\frac{1}{\boldsymbol{\gamma}_j}\right| + \left|\frac{1}{\boldsymbol{\gamma}_1}\right|\right)\left(1 + \left|\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right|\right)\sqrt{\frac{\log p_z}{n}}.
\end{aligned} \tag{167}$$

The last term can be further upper bounded by

$$\begin{aligned}
&|\boldsymbol{\gamma}_j| \frac{1}{\delta_{\min}} 8.2 \times 1.005 \times 1.1 \sqrt{M_1 M_2} \left(1 + \left|\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right|\right) \sqrt{\frac{\log p_{\mathrm{z}}}{n}} \\
\leq & |\boldsymbol{\gamma}_j| \frac{1}{\delta_{\min}} 8.2 \times 1.005 \times 1.1 \sqrt{M_1 M_2} \left(1 + |\beta|\right) \sqrt{\frac{\log p_{\mathrm{z}}}{n}} \\
&+ |\boldsymbol{\gamma}_j| \frac{1}{\delta_{\min}} 8.2 \times 1.005 \times 1.1 \sqrt{M_1 M_2} |\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}| \sqrt{\frac{\log p_{\mathrm{z}}}{n}} \\
\leq & 0.99 \frac{C_*}{\delta_{\min}} |\boldsymbol{\gamma}_j| \sqrt{\frac{\log p_{\mathrm{z}}}{n}} + C \frac{\sqrt{\frac{\log p_{\mathrm{z}}}{n}}}{\delta_{\min}} \left|\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1} \boldsymbol{\gamma}_j\right|,
\end{aligned} \tag{168}$$

where the first inequality follows from triangle inequality and the second inequality follows from the definition of C_*. Since $\delta_{\min} \gg \sqrt{\log p/n}$ and $|\boldsymbol{\pi}_1/\boldsymbol{\gamma}_1| \geq C_*(1/\delta_{\min})\sqrt{\log p_{\mathrm{z}}/n}$, by (167) and (168), we conclude (166).

weakly invalid instrument case

In this case, we assume $0 < |\boldsymbol{\pi}_1/\boldsymbol{\gamma}_1| < C_*(1/\delta_{\min})\sqrt{\log p_{\mathrm{z}}/n}$. We have the following expression of (154),

$$\widehat{\boldsymbol{\pi}}_j^{[1]} = \boldsymbol{\pi}_j - \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\gamma}_j + \frac{1}{\sqrt{n}}\left(T^{\boldsymbol{\pi}_j} + \Delta^{\boldsymbol{\pi}_j}\right). \tag{169}$$

We are going to show that on the event $A \cap S \cap G \cap F^1$,

$$\left\{ j \in \mathcal{S}^* : \left|\frac{\boldsymbol{\pi}_j}{\boldsymbol{\gamma}_j}\right| \geq 2C_* \frac{1}{\delta_{\min}} \sqrt{\frac{\log p_{\mathrm{z}}}{n}} \right\} \subset \operatorname{supp}\left(\widetilde{\boldsymbol{\pi}}^{[1]}\right). \tag{170}$$

It is sufficient to show the following inequality if $|\boldsymbol{\pi}_j/\boldsymbol{\gamma}_j| > 2C_*(1/\delta_{\min})\sqrt{\log p_{\mathrm{z}}/n}$,

$$\left|\boldsymbol{\pi}_j - \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\gamma}_j + \frac{1}{\sqrt{n}}\left(T^{\boldsymbol{\pi}_j} + \Delta^{\boldsymbol{\pi}_j}\right)\right| \geq 2.05 \sqrt{\widehat{\boldsymbol{\Theta}}_{11} + \left(\widehat{\beta}^{[1]}\right)^2 \widehat{\boldsymbol{\Theta}}_{22} - 2\widehat{\beta}^{[1]}\widehat{\boldsymbol{\Theta}}_{12}} \sqrt{\frac{\log p_{\mathrm{z}}}{n}} \frac{\left\|\widehat{\mathbf{v}}^{[j]} - \frac{\widetilde{\boldsymbol{\gamma}}_j}{\widetilde{\boldsymbol{\gamma}}_1}\widehat{\mathbf{v}}^{[1]}\right\|_2}{\sqrt{n}}, \tag{171}$$

which can be reduced to

$$\max_{j \in \operatorname{supp}(\boldsymbol{\pi}) \cap \mathcal{S}^*} \frac{1}{\sqrt{n}} |T^{\boldsymbol{\pi}_j} + \Delta^{\boldsymbol{\pi}_j}| \leq 2.05 \sqrt{\widehat{\boldsymbol{\Theta}}_{11} + \left(\widehat{\beta}^{[1]}\right)^2 \widehat{\boldsymbol{\Theta}}_{22} - 2\widehat{\beta}^{[1]}\widehat{\boldsymbol{\Theta}}_{12}} \sqrt{\frac{\log p_{\mathrm{z}}}{n}} \frac{\left\|\widehat{\mathbf{v}}^{[j]} - \frac{\widetilde{\boldsymbol{\gamma}}_j}{\widetilde{\boldsymbol{\gamma}}_1}\widehat{\mathbf{v}}^{[1]}\right\|_2}{\sqrt{n}}, \tag{172}$$

$$\left|\boldsymbol{\pi}_j - \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\gamma}_j\right| \geq 4.1 \sqrt{\widehat{\boldsymbol{\Theta}}_{11} + \left(\widehat{\beta}^{[1]}\right)^2 \widehat{\boldsymbol{\Theta}}_{22} - 2\widehat{\beta}^{[1]}\widehat{\boldsymbol{\Theta}}_{12}} \sqrt{\frac{\log p_{\mathrm{z}}}{n}} \frac{\left\|\widehat{\mathbf{v}}^{[j]} - \frac{\widetilde{\boldsymbol{\gamma}}_j}{\widetilde{\boldsymbol{\gamma}}_1}\widehat{\mathbf{v}}^{[1]}\right\|_2}{\sqrt{n}}. \tag{173}$$

By (159), (160) and (161), we establish(172). Since

$$\frac{1}{|\boldsymbol{\gamma}_j|}\left|\boldsymbol{\pi}_j - \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\boldsymbol{\gamma}_j\right| \geq \left|\frac{\boldsymbol{\pi}_j}{\boldsymbol{\gamma}_j}\right| - \left|\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right| \geq C_* \frac{1}{\delta_{\min}} \sqrt{\frac{\log p_{\mathrm{z}}}{n}},$$

the assumption $0 < |\boldsymbol{\pi}_1/\boldsymbol{\gamma}_1| < C_*(1/\delta_{\min})\sqrt{\log p_z/n}$ and (168) lead to (173).

valid instrument case

In this case, the instrumental variable is valid with $\boldsymbol{\pi}_1/\boldsymbol{\gamma}_1 = 0$. For $j \in \mathcal{V}^*$, (154) can be re-expressed as

$$\sqrt{n}\left(\widehat{\boldsymbol{\pi}}_j^{[1]} - 0\right) = T^{\boldsymbol{\pi}_j} + \Delta^{\boldsymbol{\pi}_j}. \tag{174}$$

By (159), (160) and (161), on the event $A \cap S \cap G \cap F^1$,

$$\max_{j\in\mathcal{V}^*} \frac{1}{\sqrt{n}} |T^{\boldsymbol{\pi}_j} + \Delta^{\boldsymbol{\pi}_j}| \le 2.05\sqrt{\widehat{\boldsymbol{\Theta}}_{11} + \left(\widehat{\beta}^{[1]}\right)^2 \widehat{\boldsymbol{\Theta}}_{22} - 2\widehat{\beta}^{[1]}\widehat{\boldsymbol{\Theta}}_{12}}\sqrt{\frac{\log p_z}{n}} \frac{\left\|\widehat{\mathbf{v}}^{[j]} - \frac{\widetilde{\boldsymbol{\gamma}}_j}{\widetilde{\boldsymbol{\gamma}}_1}\widehat{\mathbf{v}}^{[1]}\right\|_2}{\sqrt{n}}. \tag{175}$$

and hence

$$\operatorname{supp}\left(\widetilde{\boldsymbol{\pi}}^{[1]}\right) \subset \operatorname{supp}(\boldsymbol{\pi}) \quad \text{and} \quad \|\widetilde{\boldsymbol{\pi}}^{[1]}\|_0 < \frac{|\mathcal{S}^*|}{2}. \tag{176}$$

For $|\boldsymbol{\pi}_j/\boldsymbol{\gamma}_j| \ge C_*(1/\delta_{\min})\sqrt{\log p_z/n}$, by (168), we obtain

$$|\boldsymbol{\pi}_j| \ge 4.1\sqrt{\widehat{\boldsymbol{\Theta}}_{11} + \left(\widehat{\beta}^{[1]}\right)^2 \widehat{\boldsymbol{\Theta}}_{22} - 2\widehat{\beta}^{[1]}\widehat{\boldsymbol{\Theta}}_{12}}\sqrt{\frac{\log p_z}{n}} \frac{\left\|\widehat{\mathbf{v}}^{[j]} - \frac{\widetilde{\boldsymbol{\gamma}}_j}{\widetilde{\boldsymbol{\gamma}}_1}\widehat{\mathbf{v}}^{[1]}\right\|_2}{\sqrt{n}}. \tag{177}$$

Combined with (175), we have

$$\left|\widehat{\boldsymbol{\pi}}_j^{[1]}\right| \ge 2.05\sqrt{\widehat{\boldsymbol{\Theta}}_{11} + \left(\widehat{\beta}^{[1]}\right)^2 \widehat{\boldsymbol{\Theta}}_{22} - 2\widehat{\beta}^{[1]}\widehat{\boldsymbol{\Theta}}_{12}}\sqrt{\frac{\log p_z}{n}} \frac{\left\|\widehat{\mathbf{v}}^{[j]} - \frac{\widetilde{\boldsymbol{\gamma}}_j}{\widetilde{\boldsymbol{\gamma}}_1}\widehat{\mathbf{v}}^{[1]}\right\|_2}{\sqrt{n}}. \tag{178}$$

Hence

$$\left\{j \in \mathcal{S}^* : \left|\frac{\boldsymbol{\pi}_j}{\boldsymbol{\gamma}_j}\right| \ge C_* \frac{1}{\delta_{\min}}\sqrt{\frac{\log p_z}{n}}\right\} \subset \operatorname{supp}\left(\widetilde{\boldsymbol{\pi}}^{[1]}\right). \tag{179}$$

By comparing (163) and (176), we rule out the strong invalid instrumental variable case and obtain $\left|\widetilde{\mathcal{V}}\right| > |\mathcal{S}^*|/2$ in (60). Further by (170) and (179), we establish (60). With a stronger assumption (22) in the main paper, the weak invalid instrument case is also ruled out and (179) leads to (61).

E.9. Proof of Lemma 6

The proof of this lemma follows from the following results. Under the regularity assumptions (R1) – (R3), as $\sqrt{s_{z1}}s\log p/\sqrt{n} \to 0$, we have

$$\max_{1\le i,j\le 2} |\widehat{\boldsymbol{\Theta}}_{ij} - \boldsymbol{\Theta}_{ij}| \xrightarrow{p} 0; \tag{180}$$

and

$$\frac{\widehat{\|\boldsymbol{\gamma}\|_2^2}}{\|\boldsymbol{\gamma}\|_2^2} \xrightarrow{p} 1 \quad \text{and} \quad \frac{\left\|\sum_{j\in\widetilde{\mathcal{S}}} \widetilde{\boldsymbol{\gamma}}_j \widehat{\mathbf{v}}^{[j]}\right\|_2}{\left\|\sum_{j\in\mathcal{S}^*} \boldsymbol{\gamma}_j \widehat{\mathbf{v}}^{[j]}\right\|_2} \xrightarrow{p} 1. \tag{181}$$

By (27) and (180), we establish that

$$\frac{\sqrt{\widehat{\mathbf{\Theta}}_{11}+\widehat{\beta}^2\widehat{\mathbf{\Theta}}_{22}-2\widehat{\beta}\widehat{\mathbf{\Theta}}_{12}}}{\sqrt{\mathbf{\Theta}_{11}+(\beta)^2\mathbf{\Theta}_{22}-2\beta\mathbf{\Theta}_{12}}} \xrightarrow{p} 1.$$

Combined with (181), we establish (58).

Proof of (180) A stronger version of this proposition has already been proved in Ren et al. [2013], where part of it was already established in Sun and Zhang [2012]. To be self-contained, we will provide the sketch of the proof in the following.

The difference between $\widehat{\mathbf{\Theta}}-\mathbf{\Theta}$ can be decomposed as,

$$\widehat{\mathbf{\Theta}}-\mathbf{\Theta}=\mathbf{\Theta}^{\text{ora}}-\mathbf{\Theta}+\widehat{\mathbf{\Theta}}-\mathbf{\Theta}^{\text{ora}}, \tag{182}$$

where $\mathbf{\Theta}_{11}^{\text{ora}}=\frac{1}{n}\|Y-\mathbf{Z\Gamma}-\mathbf{X\Psi}\|_2^2$, $\mathbf{\Theta}_{22}^{\text{ora}}=\frac{1}{n}\|D-\mathbf{Z}\boldsymbol{\gamma}-\mathbf{X}\boldsymbol{\psi}\|_2^2$ and $\mathbf{\Theta}_{12}^{\text{ora}}=\frac{1}{n}(Y-\mathbf{Z\Gamma}-\mathbf{X\Psi})^\intercal(D-\mathbf{Z}\boldsymbol{\gamma}-\mathbf{X}\boldsymbol{\psi})$. In the following, we only provide the detailed analysis of $\widehat{\mathbf{\Theta}}_{12}-\mathbf{\Theta}_{12}^{\text{ora}}$. The other differences can be established in a similar way and the difference between $\mathbf{\Theta}^{\text{ora}}-\mathbf{\Theta}$ can be established by central limit theorem.

$$\widehat{\mathbf{\Theta}}_{12}-\mathbf{\Theta}_{12}^{\text{ora}}=\frac{1}{n}\begin{pmatrix}\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\\ \widehat{\boldsymbol{\psi}}-\boldsymbol{\psi}\end{pmatrix}^\intercal\mathbf{W}^\intercal\mathbf{W}\begin{pmatrix}\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}\\ \widehat{\mathbf{\Psi}}-\mathbf{\Psi}\end{pmatrix}+\frac{1}{n}\mathbf{\Pi}_{\cdot 2}^\intercal\mathbf{W}\begin{pmatrix}\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}\\ \widehat{\mathbf{\Psi}}-\mathbf{\Psi}\end{pmatrix}+\frac{1}{n}\mathbf{\Pi}_{\cdot 1}^\intercal\mathbf{W}\begin{pmatrix}\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\\ \widehat{\boldsymbol{\psi}}-\boldsymbol{\psi}\end{pmatrix}. \tag{183}$$

By (183), we have

$$\begin{aligned}\left|\widehat{\mathbf{\Theta}}_{12}-\mathbf{\Theta}_{12}^{\text{ora}}\right| \le & \frac{1}{\sqrt{n}}\|\mathbf{W}\begin{pmatrix}\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\\ \widehat{\boldsymbol{\psi}}-\boldsymbol{\psi}\end{pmatrix}\|_2\frac{1}{\sqrt{n}}\|\mathbf{W}\begin{pmatrix}\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}\\ \widehat{\mathbf{\Psi}}-\mathbf{\Psi}\end{pmatrix}\|_2+\frac{1}{n}\|\mathbf{\Pi}_{\cdot 2}^\intercal\mathbf{W}\|_\infty\|\begin{pmatrix}\widehat{\mathbf{\Gamma}}-\mathbf{\Gamma}\\ \widehat{\mathbf{\Psi}}-\mathbf{\Psi}\end{pmatrix}\|_1\\ & +\frac{1}{n}\|\mathbf{\Pi}_{\cdot 1}^\intercal\mathbf{W}\|_\infty\|\begin{pmatrix}\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\\ \widehat{\boldsymbol{\psi}}-\boldsymbol{\psi}\end{pmatrix}\|_1.\end{aligned} \tag{184}$$

The following of the proof follows from Lemma 10 and definition of event G.

Proof of (181) For a given $0<\epsilon_0<1$, we have

$$\mathbf{P}\left(\left|\frac{\|\boldsymbol{\gamma}\|_2^2}{\widehat{\|\boldsymbol{\gamma}\|_2^2}}-1\right|\ge\epsilon_0\right)\le\mathbf{P}\left(\left|\frac{\widehat{\|\boldsymbol{\gamma}\|_2^2}-\|\boldsymbol{\gamma}\|_2^2}{\|\boldsymbol{\gamma}\|_2^2}\right|\ge\frac{\epsilon_0}{1-\epsilon_0}\right).$$

By Lemma 5, on the event $A\cap S\cap G$, we have

$$\left|\frac{\widehat{\|\boldsymbol{\gamma}\|_2^2}-\|\boldsymbol{\gamma}\|_2^2}{\|\boldsymbol{\gamma}\|_2^2}\right|\le C\frac{1}{\|\boldsymbol{\gamma}\|_2^2}\left(s_{\text{z}1}\frac{\log p_{\text{z}}}{n}+C\|\boldsymbol{\gamma}\|_2\sqrt{\frac{2s_{\text{z}1}\log p_{\text{z}}}{n}}\right).$$

Since $\|\boldsymbol{\gamma}\|_2^2\gg(s\log p/\sqrt{n})^2$, we obtain that

$$\left|\frac{\widehat{\|\boldsymbol{\gamma}\|_2^2}-\|\boldsymbol{\gamma}\|_2^2}{\|\boldsymbol{\gamma}\|_2^2}\right|\le\frac{\epsilon_0}{1-\epsilon_0}\text{ and }\mathbf{P}\left(\left|\frac{\|\boldsymbol{\gamma}\|_2^2}{\widehat{\|\boldsymbol{\gamma}\|_2^2}}-1\right|\ge\epsilon_0\right)\le\mathbf{P}\left((A\cap S\cap G)^c\right).$$

Combined with Lemma 3, we establish the first convergence result of (181). On the event $S \cap G \cap A$, we have

$$\left| \frac{\left\| \sum_{j\in\widetilde{\mathcal{S}}} \widetilde{\gamma}_j \widehat{\mathbf{v}}^{[j]} \right\|_2}{\left\| \sum_{j\in\mathcal{S}^*} \gamma_j \widehat{\mathbf{v}}^{[j]} \right\|_2} - 1 \right| \leq \frac{\sum_{j\in\widetilde{\mathcal{S}}} |\widetilde{\gamma}_j - \gamma_j| \frac{\|\widehat{\mathbf{v}}^{[j]}\|}{\sqrt{n}} + \sum_{j\in\mathcal{S}^*\setminus\widetilde{\mathcal{S}}} |\gamma_j| \frac{\|\widehat{\mathbf{v}}^{[j]}\|}{\sqrt{n}}}{\frac{1}{\sqrt{n}} \left\| \sum_{j\in\mathcal{S}^*} \gamma_j \widehat{\mathbf{v}}^{[j]} \right\|_2}$$

By Lemma 4, we have

$$\begin{aligned}
&\frac{\sum_{j\in\widetilde{\mathcal{S}}} |\widetilde{\gamma}_j - \gamma_j| \frac{\|\widehat{\mathbf{v}}^{[j]}\|}{\sqrt{n}} + \sum_{j\in\mathcal{S}^*\setminus\widetilde{\mathcal{S}}} |\gamma_j| \frac{\|\widehat{\mathbf{v}}^{[j]}\|}{\sqrt{n}}}{\frac{1}{\sqrt{n}} \left\| \sum_{j\in\mathcal{S}^*} \gamma_j \widehat{\mathbf{v}}^{[j]} \right\|_2} \\
\leq &\frac{\left(\sum_{j\in\widetilde{\mathcal{S}}} |\widetilde{\gamma}_j - \gamma_j| + \sum_{i\in\mathcal{S}^*\setminus\widetilde{\mathcal{S}}} |\gamma_j| \right) \sqrt{\left(1 + 12\sqrt{\frac{\log p}{n}} \right) M_1}}{\sqrt{\frac{M_1 \|\gamma\|_2^2 (1 - s_{z1}\lambda_n)^2}{1 - 12\sqrt{\frac{\log p}{n}}}}} \leq C s_{z1} \sqrt{\frac{\log p}{n}} \leq \epsilon_0,
\end{aligned}$$

and hence the second convergence result of (181) follows from the following inequality.

E.10. Proof of Lemma 9

Define $\widetilde{\|\gamma\|_2^2} = \sum_{j\in\widetilde{\mathcal{V}}} \widetilde{\gamma}_j^2$ and $\|\gamma_{\mathcal{V}^*}\|_2^2 = \sum_{j\in\mathcal{V}^*} \gamma_j^2$. The proof of this lemma is further based on the following results. Under the assumptions (R1) – (R5) and (IN1)-(IN3). As $\sqrt{s_{z1}} s \log p / \sqrt{n} \to 0$, we have

$$\frac{\widetilde{\|\gamma\|_2^2}}{\|\gamma_{\mathcal{V}^*}\|_2^2} \xrightarrow{p} 1 \quad \text{and} \quad \frac{\left\| \sum_{j\in\widetilde{\mathcal{V}}} \widetilde{\gamma}_j \widehat{\mathbf{v}}^{[j]} \right\|_2}{\left\| \sum_{j\in\mathcal{V}^*} \gamma_j \widehat{\mathbf{v}}^{[j]} \right\|_2} \xrightarrow{p} 1. \tag{185}$$

By (180) and (23) in the main paper, we establish that

$$\frac{\sqrt{\widehat{\mathbf{\Theta}}_{11} + \widehat{\beta}^2 \widehat{\mathbf{\Theta}}_{22} - 2\widehat{\beta}\widehat{\mathbf{\Theta}}_{12}}}{\sqrt{\mathbf{\Theta}_{11} + (\beta)^2 \mathbf{\Theta}_{22} - 2\beta\mathbf{\Theta}_{12}}} \xrightarrow{p} 1.$$

Combined with (185), we establish (65).

Proof of (185) For a given $0 < \epsilon_0 < 1$, we have

$$\mathbf{P}\left(\left| \frac{\|\gamma_{\mathcal{V}^*}\|_2^2}{\widetilde{\|\gamma\|_2^2}} - 1 \right| \geq \epsilon_0 \right) \leq \mathbf{P}\left(\left| \frac{\widetilde{\|\gamma\|_2^2} - \|\gamma_{\mathcal{V}^*}\|_2^2}{\|\gamma_{\mathcal{V}^*}\|_2^2} \right| \geq \frac{\epsilon_0}{1 - \epsilon_0} \right).$$

By Lemma 7, on the event $A \cap S \cap G \cap F$, we have $\widetilde{\mathcal{V}} = \mathcal{V}^*$ and (145) leads to

$$\left| \frac{\widetilde{\|\gamma\|_2^2} - \|\gamma_{\mathcal{V}^*}\|_2^2}{\|\gamma_{\mathcal{V}^*}\|_2^2} \right| \leq C \frac{1}{\|\gamma_{\mathcal{V}^*}\|_2^2} \left(s_{z1} \frac{\log p_z}{n} + C s_{z1} \left(s \frac{\log p}{n} \right)^2 + C \|\gamma_{\widetilde{\mathcal{V}}}\|_2 \sqrt{\frac{2|\widetilde{\mathcal{V}}| \log p_z}{n}} \right). \tag{186}$$

Since $\|\boldsymbol{\gamma}_{\mathcal{V}^*}\|_2^2 \gg (s\log p/\sqrt{n})^2$, we obtain that

$$\left|\frac{\widetilde{\|\boldsymbol{\gamma}\|_2^2} - \|\boldsymbol{\gamma}_{\mathcal{V}^*}\|_2^2}{\|\boldsymbol{\gamma}_{\mathcal{V}^*}\|_2^2}\right| \leq \frac{\epsilon_0}{1-\epsilon_0} \text{ and } \mathbf{P}\left(\left|\frac{\|\boldsymbol{\gamma}_{\mathcal{V}^*}\|_2^2}{\widetilde{\|\boldsymbol{\gamma}\|_2^2}} - 1\right| \geq \epsilon_0\right) \leq \mathbf{P}\left((A\cap S\cap G\cap F)^c\right).$$

Combined with Lemma 3 and (156), we establish the first converge result of (185).

On the event $S\cap G\cap A\cap F$, we have

$$\left|\frac{\left\|\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_j\widehat{\mathbf{v}}^{[j]}\right\|_2}{\left\|\sum_{j\in\mathcal{V}^*}\boldsymbol{\gamma}_j\widehat{\mathbf{v}}^{[j]}\right\|_2} - 1\right| \leq \frac{\sum_{j\in\mathcal{V}^*}|\widetilde{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j|\frac{\|\widehat{\mathbf{v}}^{[j]}\|}{\sqrt{n}}}{\frac{1}{\sqrt{n}}\left\|\sum_{j\in\mathcal{V}^*}\boldsymbol{\gamma}_j\widehat{\mathbf{v}}^{[j]}\right\|_2}.$$

By Lemma 4, we have

$$\frac{\sum_{j\in\mathcal{V}^*}|\widetilde{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j|\frac{\|\widehat{\mathbf{v}}^{[j]}\|}{\sqrt{n}}}{\frac{1}{\sqrt{n}}\left\|\sum_{j\in\mathcal{V}^*}\boldsymbol{\gamma}_j\widehat{\mathbf{v}}^{[j]}\right\|_2} \leq \frac{\left(\sum_{j\in\mathcal{V}^*}|\widetilde{\boldsymbol{\gamma}}_j - \boldsymbol{\gamma}_j|\right)\sqrt{\left(1+12\sqrt{\frac{\log p}{n}}\right)M_1}}{\sqrt{\frac{M_1\|\boldsymbol{\gamma}\|_2^2(1-s_{z1}\lambda_n)^2}{1-12\sqrt{\frac{\log p}{n}}}}} \leq Cs_{z1}\sqrt{\frac{\log p}{n}}. \tag{187}$$

Hence the second converge result of (185) follows from

$$\mathbf{P}\left(\left|\frac{\left\|\sum_{j\in\widetilde{\mathcal{V}}}\widetilde{\boldsymbol{\gamma}}_j\widehat{\mathbf{v}}^{[j]}\right\|_2}{\left\|\sum_{j\in\mathcal{V}^*}\boldsymbol{\gamma}_j\widehat{\mathbf{v}}^{[j]}\right\|_2} - 1\right| \geq \epsilon_0\right) \leq \mathbf{P}\left((S\cap G\cap A\cap F)^c\right).$$

E.11. Proof of Lemmas 10, 11, 12 and 13

Proof of Lemma 10 We only estalbish the first half of (98) and (99). The proof of the second half of (98) and (100) will be similar. The proof has been established in Sun and Zhang [2012] for fixed designs under certain assumptions for the design. In the following, we will check that the assumptions in Corollary 1 in Sun and Zhang [2012] are satisfied with high probability for the subgaussian random designs considered in this paper and then apply equation (23) in Sun and Zhang [2012]. By the definition of τ^* in Sun and Zhang [2012], we have $\tau^* \leq \tau$ where τ is defined in (30). Hence, on the event S_1, equation (23) in Sun and Zhang [2012] holds. By the relationship between ℓ_1 cone invertibility factor and the restricted eigenvalue established in Lemma 13 of Cai and Guo [2016b], we obtain that on the event $S\cap G$,

$$\|\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}\|_1 + \|\widehat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}\|_1 \leq C\frac{s\lambda_0\sigma_1}{\kappa^2(\mathbf{H}, 4s, 1+2\epsilon_0)}. \tag{188}$$

Similar to the proof of Lemma 13 in Cai and Guo [2016b], we establish

$$\kappa^2\left(\mathbf{H}, 4s, 1+2\epsilon_0\right) \geq \frac{n}{\max\|\mathbf{W}_{\cdot j}\|_2^2}\kappa^2\left(\mathbf{W}, 4s, (1+2\epsilon_0)\left(\frac{\max\|\mathbf{W}_{\cdot j}\|_2}{\min\|\mathbf{W}_{\cdot j}\|_2}\right)\right). \tag{189}$$

Hence, on the event $G \cap S$, we establish the first half of (98). Since

$$\frac{1}{n}\|\mathbf{Z}(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}) + \mathbf{X}(\widehat{\mathbf{\Psi}} - \mathbf{\Psi})\|_2^2 \leq \|\frac{1}{n}\mathbf{W}^\intercal\mathbf{W}\begin{pmatrix}\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma} \\ \widehat{\mathbf{\Psi}} - \mathbf{\Psi}\end{pmatrix}\|_\infty \left(\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_1 + \|\widehat{\mathbf{\Psi}} - \mathbf{\Psi}\|_1\right),$$

we establish (99).

Proof of Lemma 11 The decompositions (101) and (103) are established by the definitions of $\widetilde{\mathbf{\Gamma}}$ and $\widetilde{\boldsymbol{\gamma}}$. The error bound (105) follows from the following inequality

$$|\Delta_j^{\boldsymbol{\gamma}}| \leq \|\left(\frac{1}{n}(\widehat{\mathbf{u}}^{[j]})^\intercal\widehat{\mathbf{\Sigma}} - e_j^\intercal\right)\|_\infty\|\begin{pmatrix}\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \\ \widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}\end{pmatrix}\|_1.$$

Proof of Lemma 12

This lemma can be established by a similar argument with Lemma 4. On the event $A \cap S \cap G$, we have

$$\frac{1}{\sqrt{M_2}}\sqrt{1 + \left(\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)^2} \leq \sqrt{\mathbf{\Theta}_{11} + \left(\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)^2\mathbf{\Theta}_{22} - 2\left(\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\mathbf{\Theta}_{12}} \leq \sqrt{M_2}\sqrt{1 + \left(\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)^2}, \tag{190}$$

and

$$\frac{\sqrt{M_1}\sqrt{1 + \left(\frac{\boldsymbol{\gamma}_j}{\boldsymbol{\gamma}_1}\right)^2}|1 - 2\lambda_n|}{\sqrt{1 - 12\sqrt{\frac{\log p}{n}}}} \leq \frac{\left\|\widehat{\mathbf{v}}^{[j]} - \frac{\boldsymbol{\gamma}_j}{\boldsymbol{\gamma}_1}\widehat{\mathbf{v}}^{[1]}\right\|_2}{\sqrt{n}} \leq \sqrt{M_1}\left(1 + \left|\frac{\boldsymbol{\gamma}_j}{\boldsymbol{\gamma}_1}\right|\right)\sqrt{1 + 12\sqrt{\frac{\log p}{n}}}. \tag{191}$$

Hence, (157) and (158) follow from the above inequalities (190) and (191).

Proof of Lemma 13

(159) follows from the standard convergence analysis and (160) follows from high probability statement of Gaussian random variable. It remains to establish (161). We will analyze the expression (155) term by term. Note that on the event $A \cap S \cap G$,

$$\begin{aligned}
\left|T^{\beta,1}\right| &\leq \frac{\sqrt{\log p_{\mathrm{z}}}}{\sqrt{n}\,|\boldsymbol{\gamma}_1|}\|\widehat{\mathbf{v}}^{[1]}\|_2\sqrt{\mathbf{\Theta}_{11} + \left(\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)^2\mathbf{\Theta}_{22} - 2\left(\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\mathbf{\Theta}_{12}} \\
&\leq C\frac{1}{|\boldsymbol{\gamma}_1|}\frac{\|\widehat{\mathbf{v}}^{[1]}\|_2}{\sqrt{n}}\sqrt{\log p_{\mathrm{z}}}\left(1 + \left|\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right|\right); \\
|R_1| &\leq C\frac{1}{|\boldsymbol{\gamma}_1|}\left(\left|\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right| + 1\right)s\frac{\log p}{\sqrt{n}}; \\
|R_2| &\leq C\frac{1}{|\boldsymbol{\gamma}_1|}\left(\sqrt{\frac{\log p}{n}} + s\frac{\log p}{n}\right)\left(\left|T^{\beta,1}\right| + |R_1|\right) \\
&\leq C\frac{1}{|\boldsymbol{\gamma}_1|}\sqrt{\frac{\log p}{n}}\frac{1}{|\boldsymbol{\gamma}_1|}\frac{\|\widehat{\mathbf{v}}^{[1]}\|_2}{\sqrt{n}}\sqrt{\log p_{\mathrm{z}}}\left(1 + \left|\beta + \frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right|\right).
\end{aligned} \tag{192}$$

Hence

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\left|\boldsymbol{\gamma}_j\Delta^{\beta,1}+\left(T^{\beta,1}+\Delta^{\beta,1}\right)\left(\widetilde{\boldsymbol{\gamma}}_j-\boldsymbol{\gamma}_j\right)\right| \leq C\frac{|\boldsymbol{\gamma}_j|}{\sqrt{n}}\left(|R_1|+|R_2|\right)+\sqrt{\frac{\log p_{\mathrm{z}}}{n}}\left|T^{\beta,1}\right| \\
\leq & C\frac{|\boldsymbol{\gamma}_j|}{|\boldsymbol{\gamma}_1|}\left(\left|\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right|+1\right)s\frac{\log p}{n}+C\frac{|\boldsymbol{\gamma}_j|}{|\boldsymbol{\gamma}_1|}\frac{\sqrt{\log p\log p_{\mathrm{z}}}}{n}\frac{1}{|\boldsymbol{\gamma}_1|}\frac{\|\widehat{\mathbf{v}}^{[1]}\|_2}{\sqrt{n}}\left(1+\left|\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right|\right) \\
&+C\frac{1}{|\boldsymbol{\gamma}_1|}\frac{\|\widehat{\mathbf{v}}^{[1]}\|_2}{\sqrt{n}}\frac{\sqrt{\log p\log p_{\mathrm{z}}}}{n}\left(1+\left|\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right|\right) \\
\leq & C\left(1+\left|\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right|\right)\left(\frac{|\boldsymbol{\gamma}_j|}{|\boldsymbol{\gamma}_1|}s\frac{\log p}{n}+\left(1+\frac{|\boldsymbol{\gamma}_j|}{|\boldsymbol{\gamma}_1|}\right)\frac{1}{|\boldsymbol{\gamma}_1|}\frac{\sqrt{\log p\log p_{\mathrm{z}}}}{n}\right).
\end{aligned} \tag{193}$$

Since

$$\left|\Delta_j^{\boldsymbol{\Gamma}}-\left(\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right)\Delta_j^{\boldsymbol{\gamma}}\right| \leq C\left(\left|\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right|+1\right)s\frac{\log p}{n},$$

we have

$$\max_{j\in\mathcal{S}^*}\frac{1}{\sqrt{n}}\left|\Delta^{\boldsymbol{\pi}_j}\right| \leq \left(1+\left|\beta+\frac{\boldsymbol{\pi}_1}{\boldsymbol{\gamma}_1}\right|\right)\left(1+\frac{|\boldsymbol{\gamma}_j|}{|\boldsymbol{\gamma}_1|}\right)\left(s\frac{\log p}{n}+\frac{1}{|\boldsymbol{\gamma}_1|}\frac{\sqrt{\log p\log p_{\mathrm{z}}}}{n}\right) \tag{194}$$

By the assumption $\min_{j\in\mathcal{S}^*}|\boldsymbol{\gamma}_j| \gg \sqrt{\log p/n}$ and (158), we establish (161).

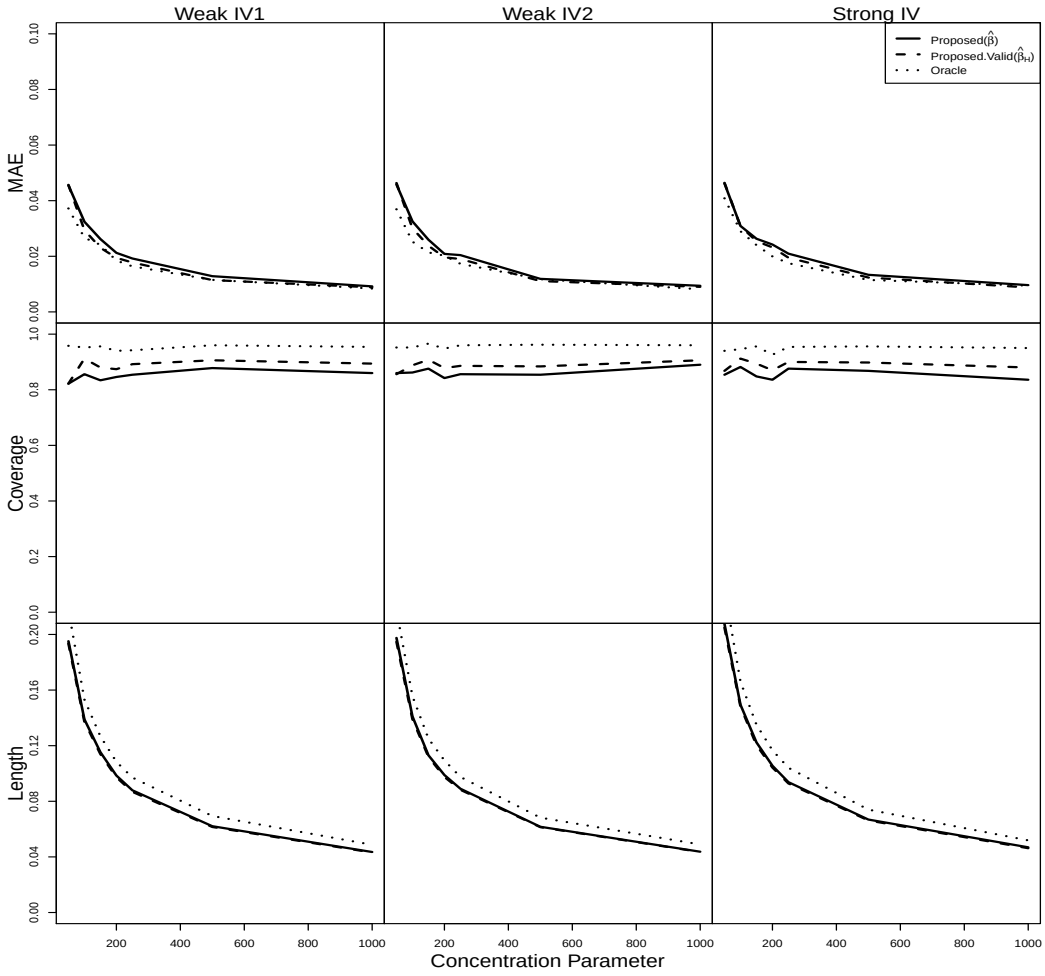

Fig. A1: Comparison of different methods in the case $\rho_2 = 0$, $p_z = 100$, $p_x = 150$ and $n = 100$. The x-axis represents the concentration parameter. On the y-axis, MAE represents Median Absolute Error of the estimators, Coverage represents coverage of confidence intervals and Length represents the average length of confidence intervals. Proposed is our method allowing for invalid IVs and is represented by the solid line. Proposed.valid is our method that assumes all the IVs are valid and is represented by the dashed line. Oracle is the method that knows exactly which instruments are valid and is represented by the dotted line. The column labeled with Weak IV1 represents the case $\rho_1 = 0.1$, the column labeled with Weak IV2 represents the case $\rho_1 = 0.2$ and the column labeled with Strong IV represents the case $\rho_1 = 0$.

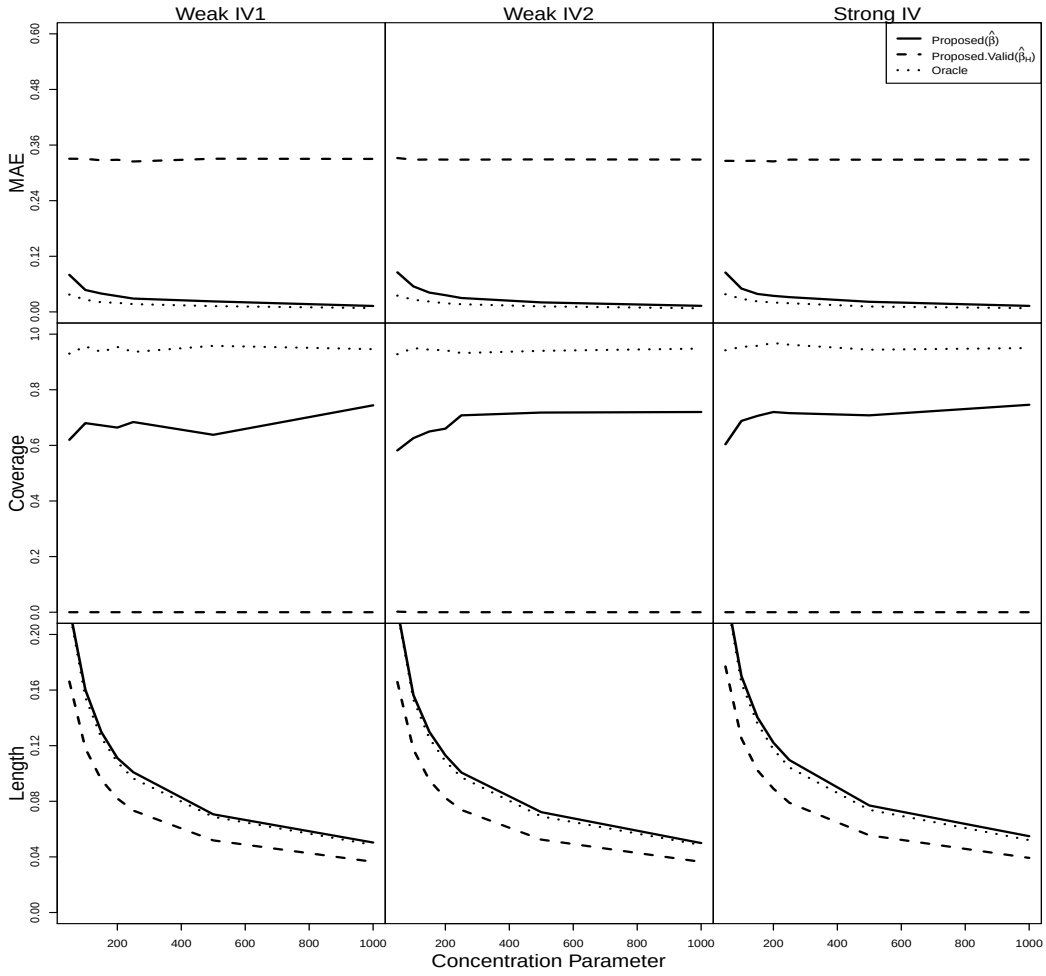

Fig. A2: Comparison of different methods in the case $\rho_2 = 1$, $p_z = 100$, $p_x = 150$ and $n = 100$. The x-axis represents the concentration parameter. On the y-axis, MAE represents Median Absolute Error of the estimators, Coverage represents coverage of confidence intervals and Length represents the average length of confidence intervals. Proposed is our method allowing for invalid IVs and is represented by the solid line. Proposed.valid is our method that assumes all the IVs are valid and is represented by the dashed line. Oracle is the method that knows exactly which instruments are valid and is represented by the dotted line. The column labeled with Weak IV1 represents the case $\rho_1 = 0.1$, the column labeled with Weak IV2 represents the case $\rho_1 = 0.2$ and the column labeled with Strong IV represents the case $\rho_1 = 0$.

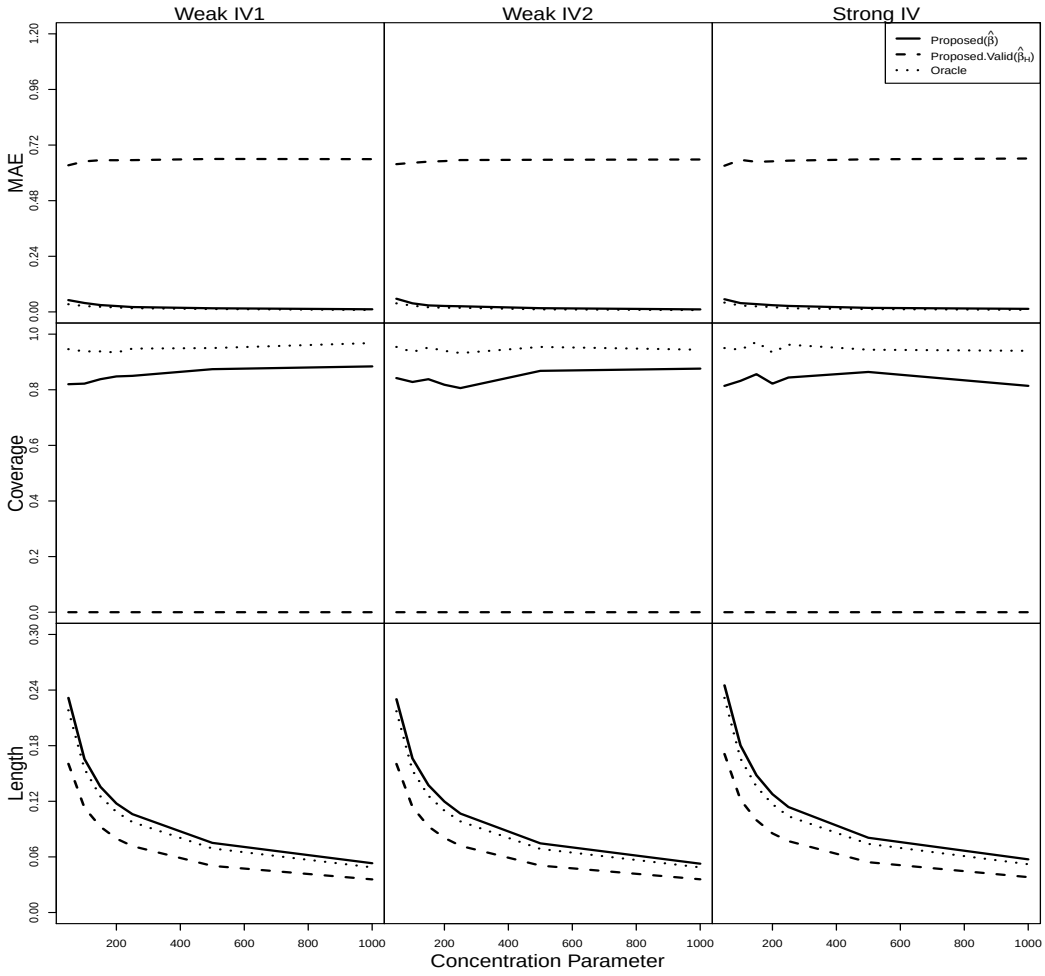

Fig. A3: Comparison of different methods in the case $\rho_2 = 2$, $p_z = 100$, $p_x = 150$ and $n = 100$. The x-axis represents the concentration parameter. On the y-axis, MAE represents Median Absolute Error of the estimators, Coverage represents coverage of confidence intervals and Length represents the average length of confidence intervals. Proposed is our method allowing for invalid IVs and is represented by the solid line. Proposed.valid is our method that assumes all the IVs are valid and is represented by the dashed line. Oracle is the method that knows exactly which instruments are valid and is represented by the dotted line. The column labeled with Weak IV1 represents the case $\rho_1 = 0.1$, the column labeled with Weak IV2 represents the case $\rho_1 = 0.2$ and the column labeled with Strong IV represents the case $\rho_1 = 0$.

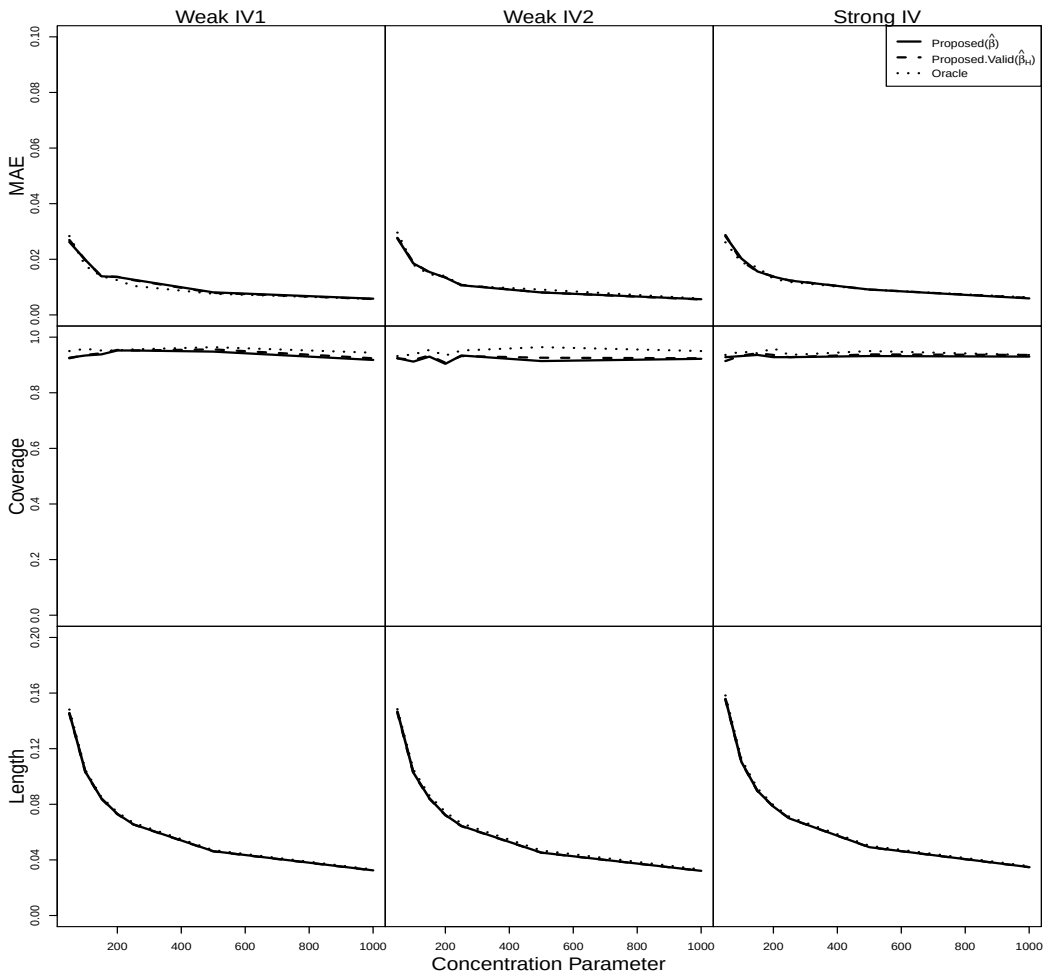

Fig. A4: Comparison of different methods in the case $\rho_2 = 0$, $p_z = 100$, $p_x = 150$ and $n = 200$. The x-axis represents the concentration parameter. On the y-axis, MAE represents Median Absolute Error of the estimators, Coverage represents coverage of confidence intervals and Length represents the average length of confidence intervals. Proposed is our method allowing for invalid IVs and is represented by the solid line. Proposed.valid is our method that assumes all the IVs are valid and is represented by the dashed line. Oracle is the method that knows exactly which instruments are valid and is represented by the dotted line. The column labeled with Weak IV1 represents the case $\rho_1 = 0.1$, the column labeled with Weak IV2 represents the case $\rho_1 = 0.2$ and the column labeled with Strong IV represents the case $\rho_1 = 0$.

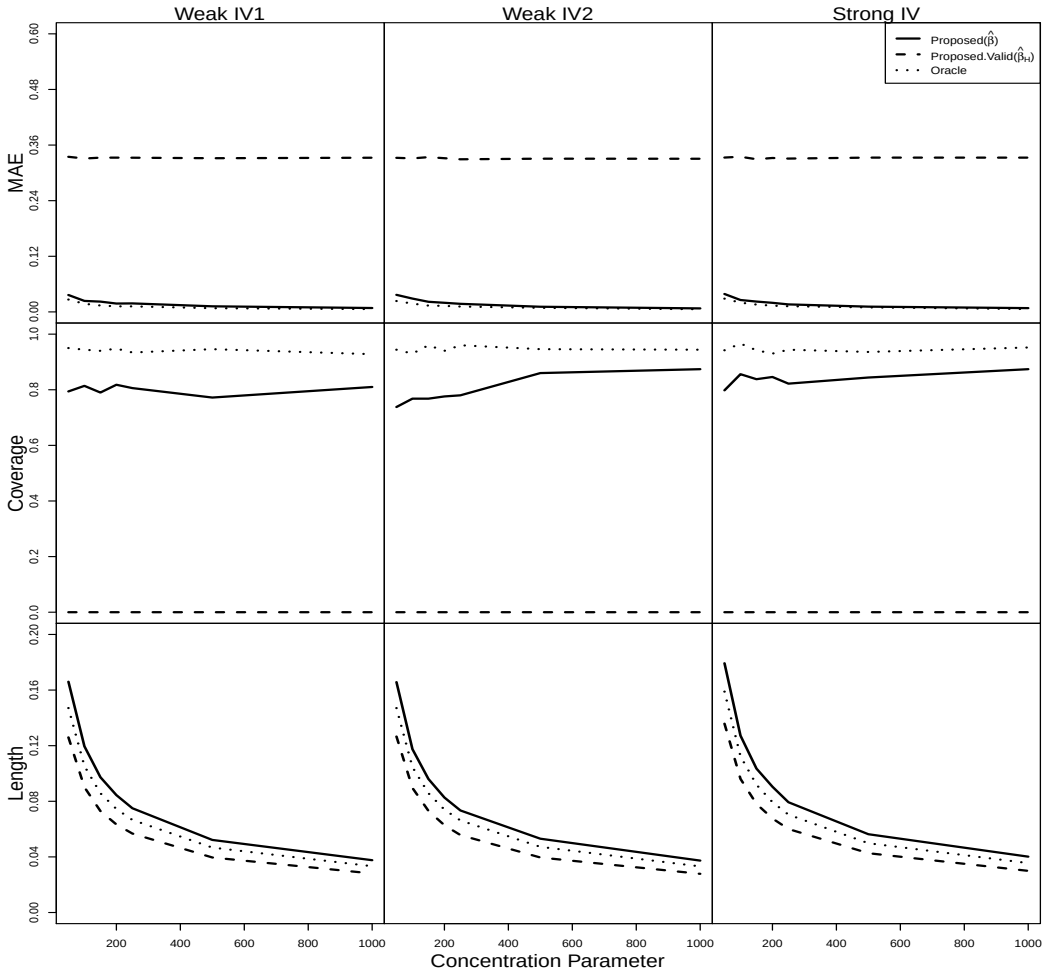

Fig. A5: Comparison of different methods in the case $\rho_2 = 1$, $p_z = 100$, $p_x = 150$ and $n = 200$. The x-axis represents the concentration parameter. On the y-axis, MAE represents Median Absolute Error of the estimators, Coverage represents coverage of confidence intervals and Length represents the average length of confidence intervals. Proposed is our method allowing for invalid IVs and is represented by the solid line. Proposed.valid is our method that assumes all the IVs are valid and is represented by the dashed line. Oracle is the method that knows exactly which instruments are valid and is represented by the dotted line. The column labeled with Weak IV1 represents the case $\rho_1 = 0.1$, the column labeled with Weak IV2 represents the case $\rho_1 = 0.2$ and the column labeled with Strong IV represents the case $\rho_1 = 0$.

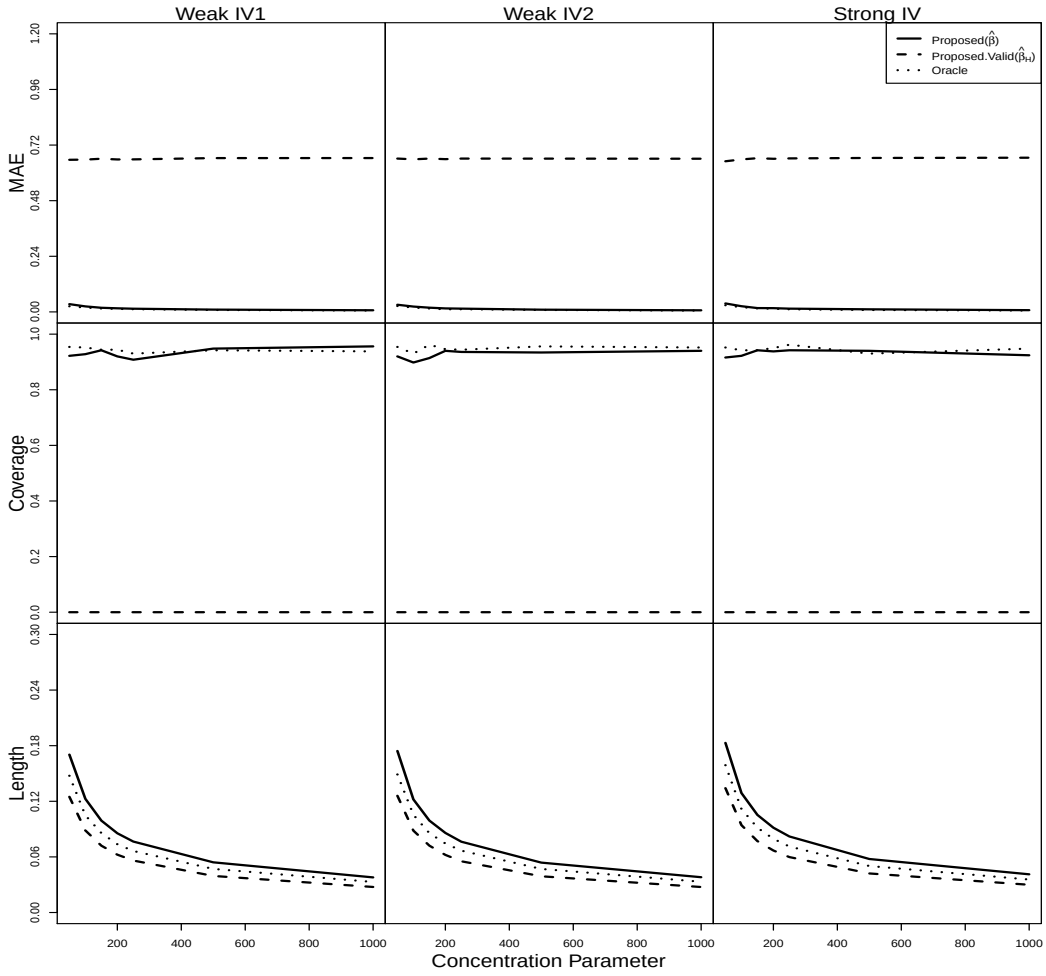

Fig. A6: Comparison of different methods in the case $\rho_2 = 2$, $p_z = 100$, $p_x = 150$ and $n = 200$. The x-axis represents the concentration parameter. On the y-axis, MAE represents Median Absolute Error of the estimators, Coverage represents coverage of confidence intervals and Length represents the average length of confidence intervals. Proposed is our method allowing for invalid IVs and is represented by the solid line. Proposed.valid is our method that assumes all the IVs are valid and is represented by the dashed line. Oracle is the method that knows exactly which instruments are valid and is represented by the dotted line. The column labeled with Weak IV1 represents the case $\rho_1 = 0.1$, the column labeled with Weak IV2 represents the case $\rho_1 = 0.2$ and the column labeled with Strong IV represents the case $\rho_1 = 0$.

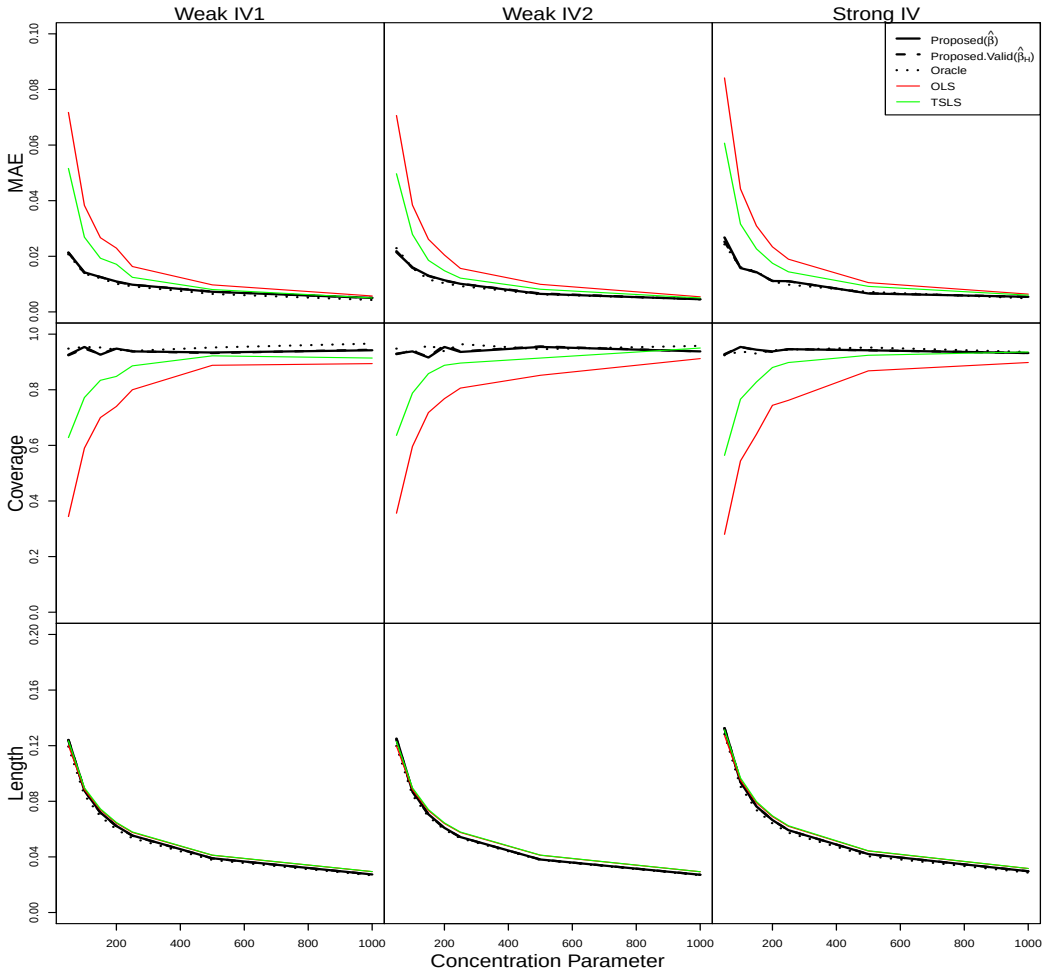

Fig. A7: Comparison of different methods in the case $\rho_2 = 0$, $p_{\mathrm{z}} = 100$, $p_{\mathrm{x}} = 150$ and $n = 300$. The x-axis represents the concentration parameter. On the y-axis, MAE represents Median Absolute Error of the estimators, Coverage represents coverage of confidence intervals and Length represents the average length of confidence intervals. Proposed is our method allowing for invalid IVs and is represented by the solid line. Proposed.valid is our method that assumes all the IVs are valid and is represented by the dashed line. Oracle is the method that knows exactly which instruments are valid and is represented by the dotted line. The column labeled with Weak IV1 represents the case $\rho_1 = 0.1$, the column labeled with Weak IV2 represents the case $\rho_1 = 0.2$ and the column labeled with Strong IV represents the case $\rho_1 = 0$.

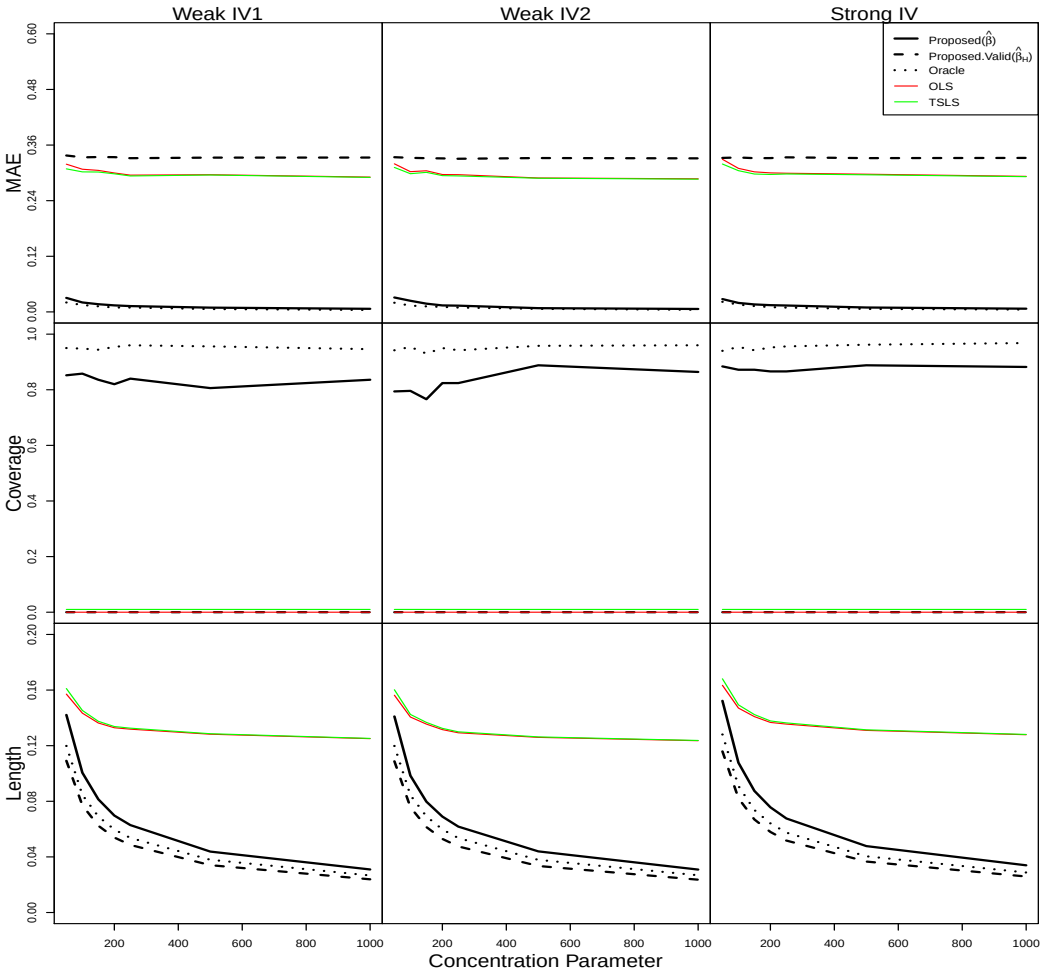

Fig. A8: Comparison of different methods in the case $\rho_2 = 1$, $p_z = 100$, $p_x = 150$ and $n = 300$. The x-axis represents the concentration parameter. On the y-axis, MAE represents Median Absolute Error of the estimators, Coverage represents coverage of confidence intervals and Length represents the average length of confidence intervals. Proposed is our method allowing for invalid IVs and is represented by the solid line. Proposed.valid is our method that assumes all the IVs are valid and is represented by the dashed line. Oracle is the method that knows exactly which instruments are valid and is represented by the dotted line. The column labeled with Weak IV1 represents the case $\rho_1 = 0.1$, the column labeled with Weak IV2 represents the case $\rho_1 = 0.2$ and the column labeled with Strong IV represents the case $\rho_1 = 0$.

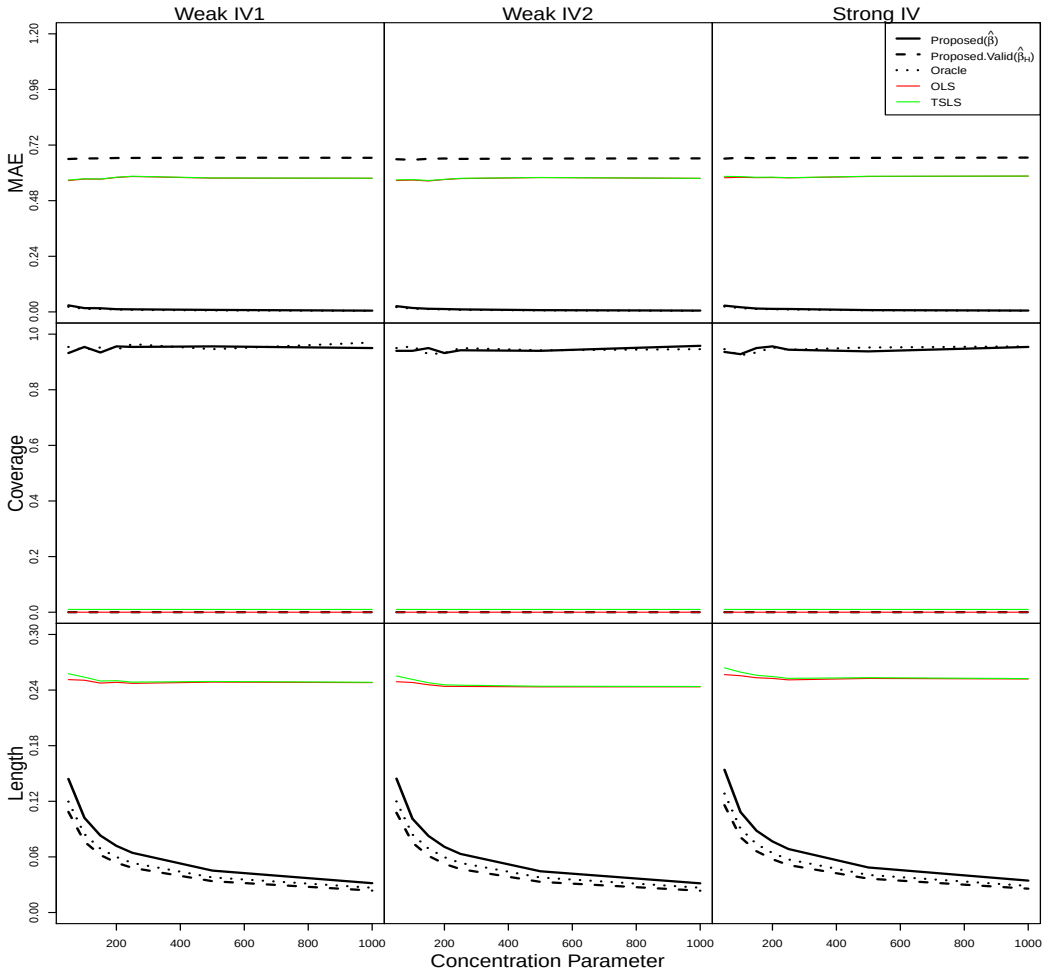

Fig. A9: Comparison of different methods in the case $\rho_2 = 1$, $p_{\mathrm{z}} = 100$, $p_{\mathrm{x}} = 150$ and $n = 300$. The x-axis represents the concentration parameter. On the y-axis, MAE represents Median Absolute Error of the estimators, Coverage represents coverage of confidence intervals and Length represents the average length of confidence intervals. Proposed is our method allowing for invalid IVs and is represented by the solid line. Proposed.valid is our method that assumes all the IVs are valid and is represented by the dashed line. Oracle is the method that knows exactly which instruments are valid and is represented by the dotted line. The column labeled with Weak IV1 represents the case $\rho_1 = 0.1$, the column labeled with Weak IV2 represents the case $\rho_1 = 0.2$ and the column labeled with Strong IV represents the case $\rho_1 = 0$.

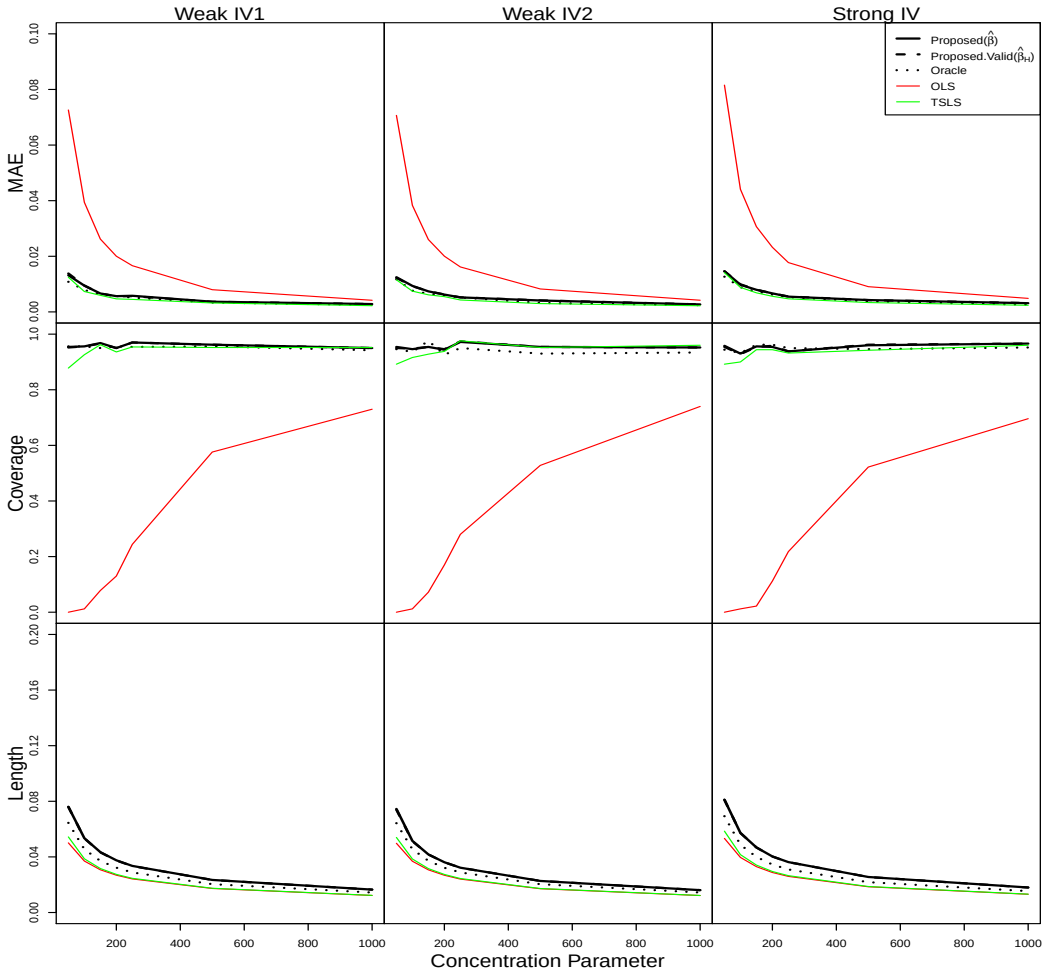

Fig. A10: Comparison of different methods in the case $\rho_2 = 0$, $p_z = 100$, $p_x = 150$ and $n = 1000$. The x-axis represents the concentration parameter. On the y-axis, MAE represents Median Absolute Error of the estimators, Coverage represents coverage of confidence intervals and Length represents the average length of confidence intervals. Proposed is our method allowing for invalid IVs and is represented by the solid line. Proposed.valid is our method that assumes all the IVs are valid and is represented by the dashed line. Oracle is the method that knows exactly which instruments are valid and is represented by the dotted line. The column labeled with Weak IV1 represents the case $\rho_1 = 0.1$, the column labeled with Weak IV2 represents the case $\rho_1 = 0.2$ and the column labeled with Strong IV represents the case $\rho_1 = 0$.

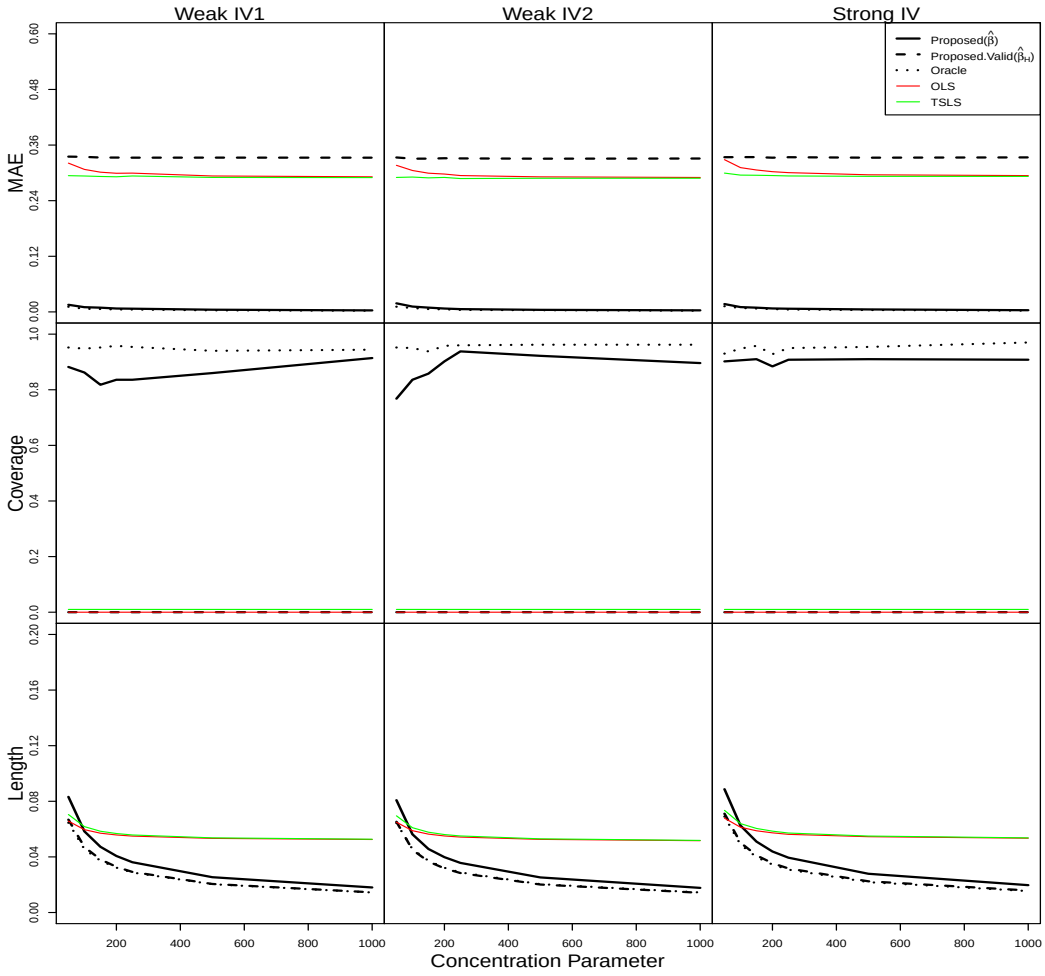

Fig. A11: Comparison of different methods in the case $\rho_2 = 1$, $p_z = 100$, $p_x = 150$ and $n = 1000$. The x-axis represents the concentration parameter. On the y-axis, MAE represents Median Absolute Error of the estimators, Coverage represents coverage of confidence intervals and Length represents the average length of confidence intervals. Proposed is our method allowing for invalid IVs and is represented by the solid line. Proposed.valid is our method that assumes all the IVs are valid and is represented by the dashed line. Oracle is the method that knows exactly which instruments are valid and is represented by the dotted line. The column labeled with Weak IV1 represents the case $\rho_1 = 0.1$, the column labeled with Weak IV2 represents the case $\rho_1 = 0.2$ and the column labeled with Strong IV represents the case $\rho_1 = 0$.

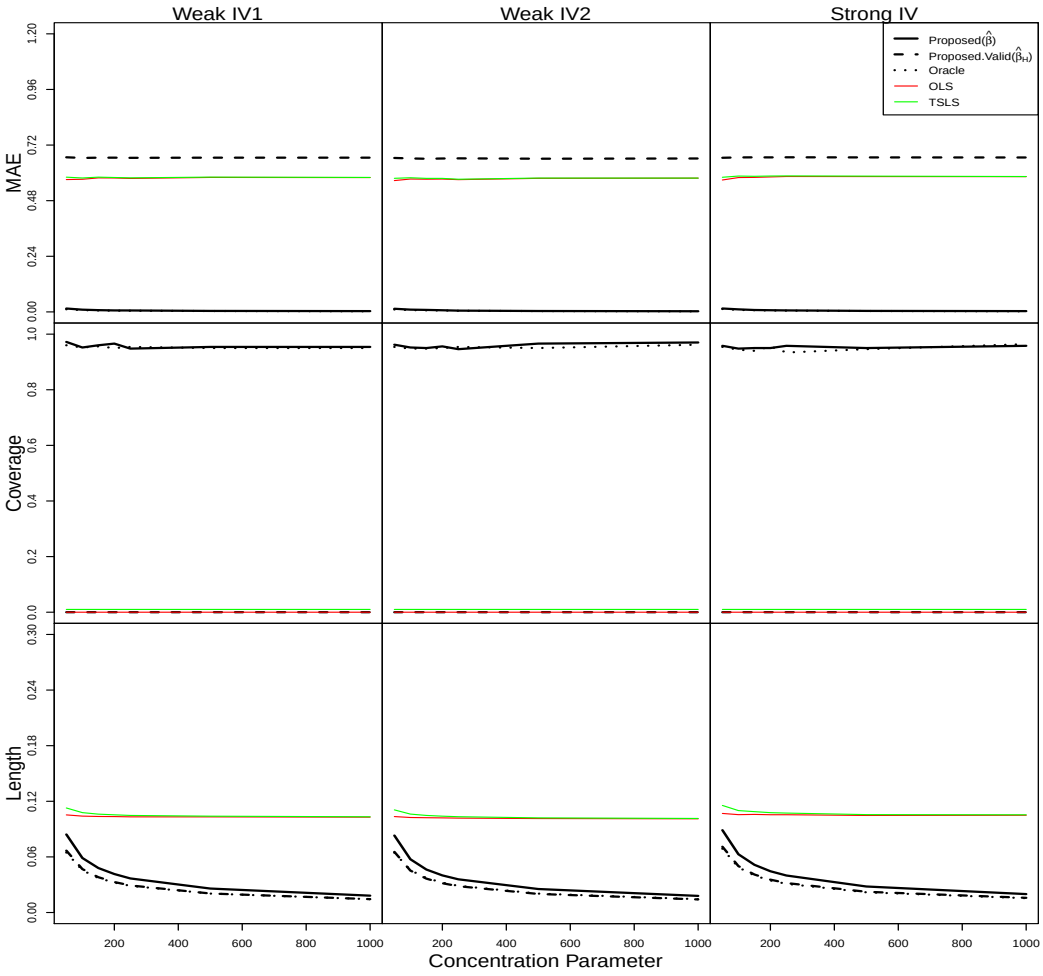

Fig. A12: Comparison of different methods in the case $\rho_2 = 2$, $p_z = 100$, $p_x = 150$ and $n = 1000$. The x-axis represents the concentration parameter. On the y-axis, MAE represents Median Absolute Error of the estimators, Coverage represents coverage of confidence intervals and Length represents the average length of confidence intervals. Proposed is our method allowing for invalid IVs and is represented by the solid line. Proposed.valid is our method that assumes all the IVs are valid and is represented by the dashed line. Oracle is the method that knows exactly which instruments are valid and is represented by the dotted line. The column labeled with Weak IV1 represents the case $\rho_1 = 0.1$, the column labeled with Weak IV2 represents the case $\rho_1 = 0.2$ and the column labeled with Strong IV represents the case $\rho_1 = 0$.

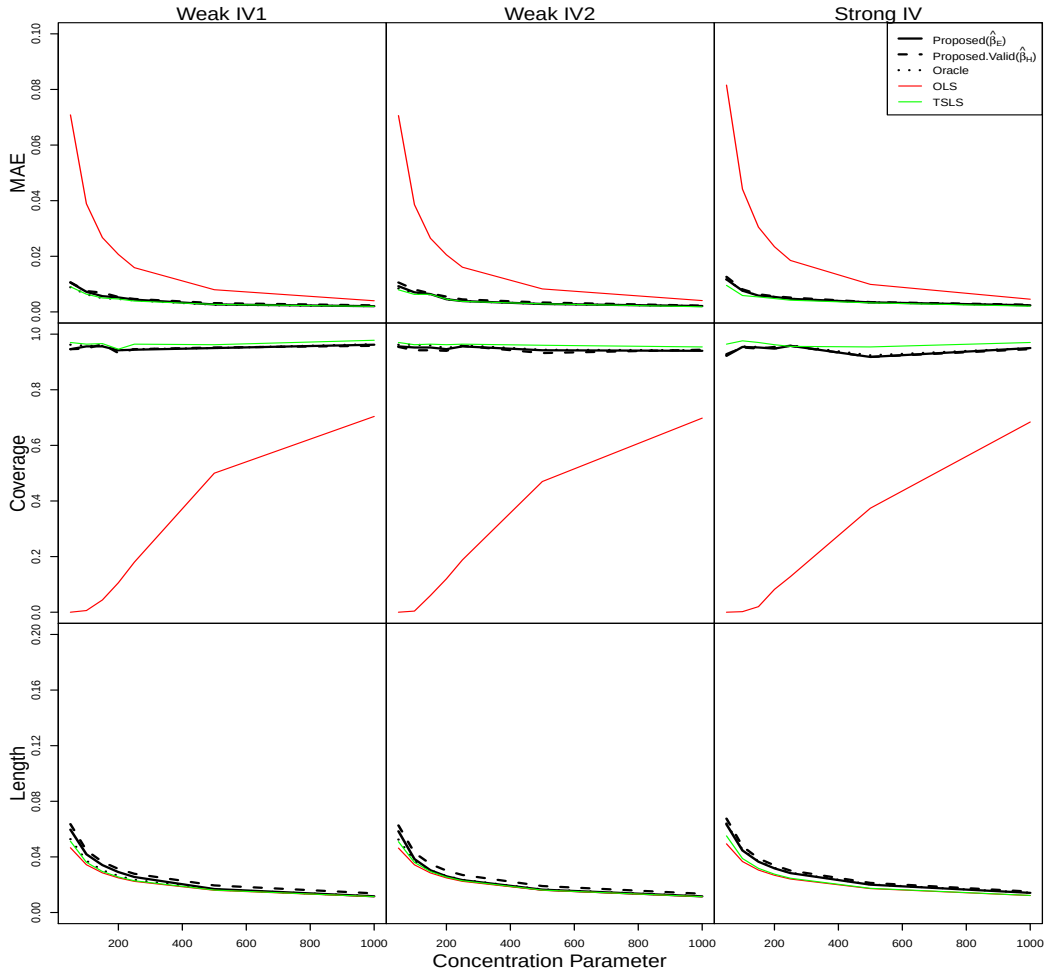

Fig. A13: Comparison of different methods in the case $\rho_2 = 0$, $p_z = 9$, $p_x = 10$ and $n = 1000$. The x-axis represents the concentration parameter. On the y-axis, MAE represents Median Absolute Error of the estimators, Coverage represents coverage of confidence intervals and Length represents the average length of confidence intervals. Proposed is our method allowing for invalid IVs and is represented by the solid line. Proposed.valid is our method that assumes all the IVs are valid and is represented by the dashed line. Oracle is the method that knows exactly which instruments are valid and is represented by the dotted line. The column labeled with Weak IV1 represents the case $\rho_1 = 0.1$, the column labeled with Weak IV2 represents the case $\rho_1 = 0.2$ and the column labeled with Strong IV represents the case $\rho_1 = 0$.

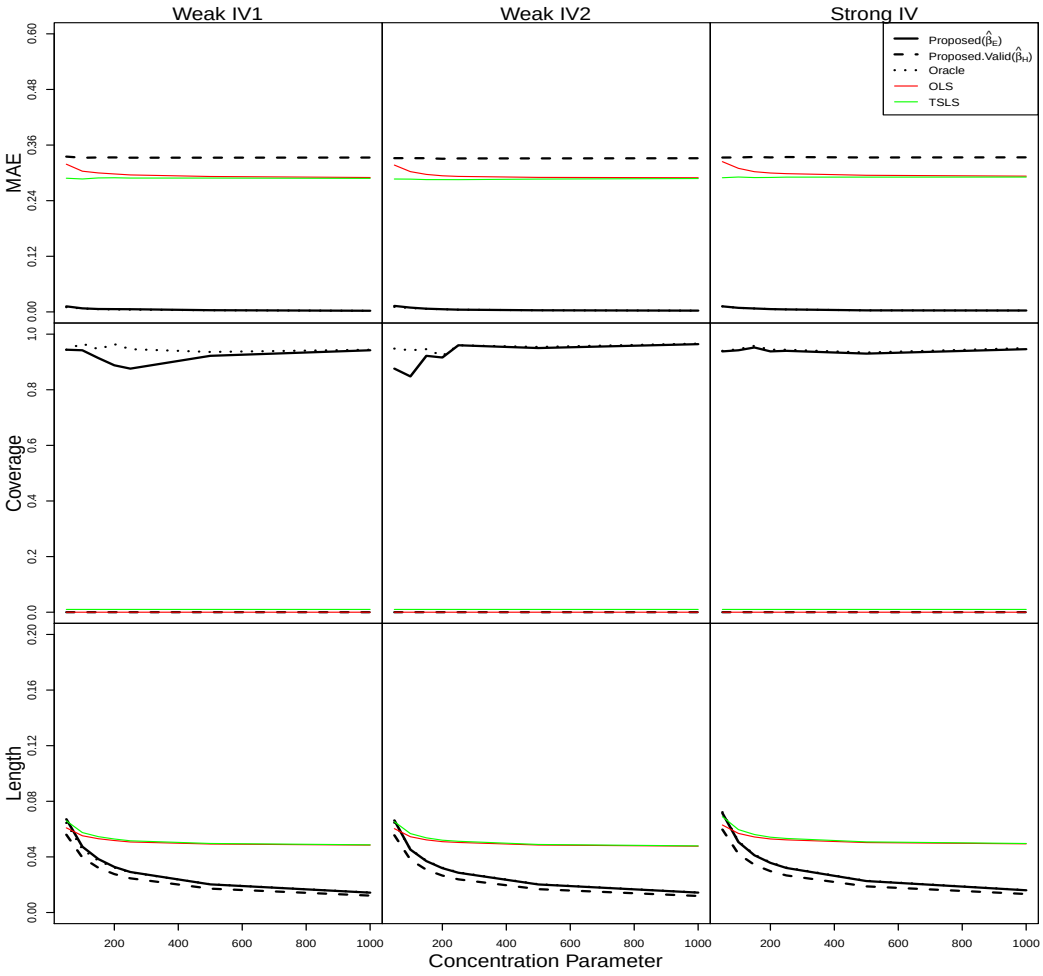

Fig. A14: Comparison of different methods in the case $\rho_2 = 1$, $p_z = 9$, $p_x = 10$ and $n = 1000$. The x-axis represents the concentration parameter. On the y-axis, MAE represents Median Absolute Error of the estimators, Coverage represents coverage of confidence intervals and Length represents the average length of confidence intervals. Proposed is our method allowing for invalid IVs and is represented by the solid line. Proposed.valid is our method that assumes all the IVs are valid and is represented by the dashed line. Oracle is the method that knows exactly which instruments are valid and is represented by the dotted line. The column labeled with Weak IV1 represents the case $\rho_1 = 0.1$, the column labeled with Weak IV2 represents the case $\rho_1 = 0.2$ and the column labeled with Strong IV represents the case $\rho_1 = 0$.

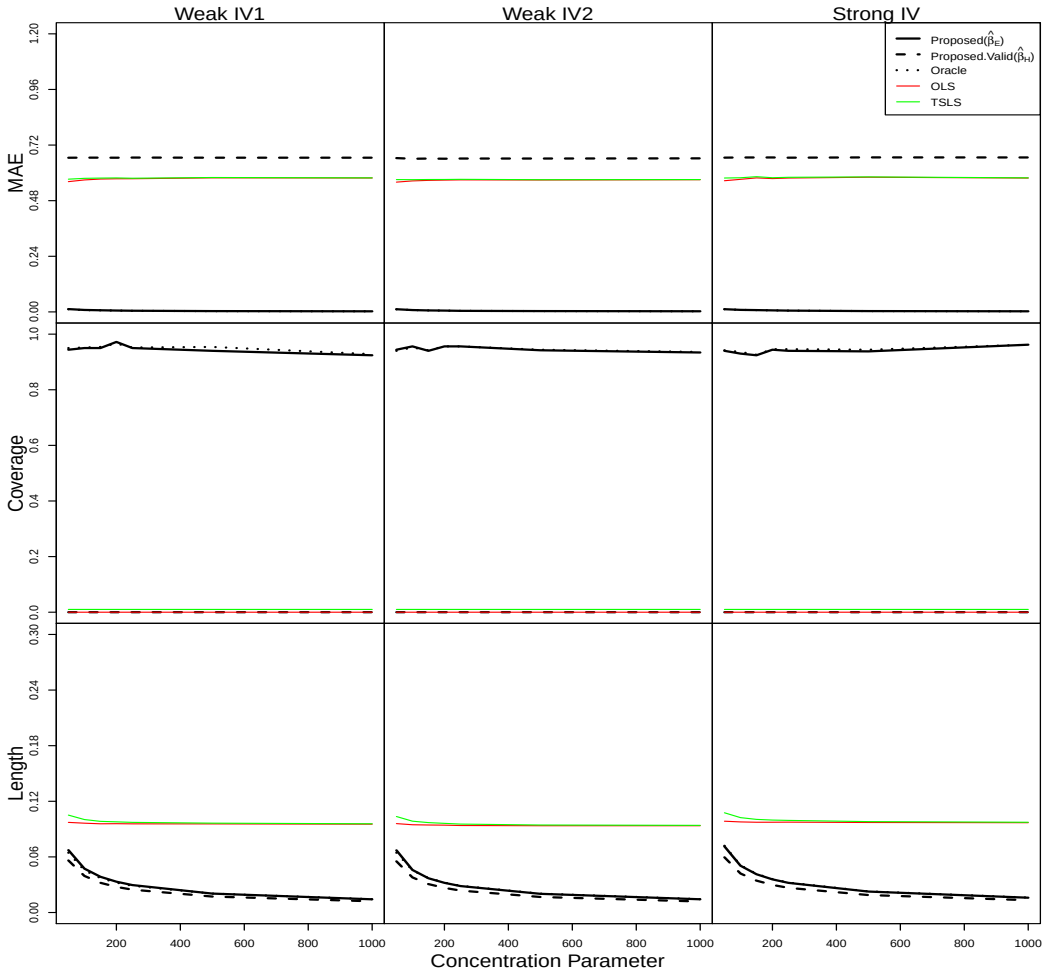

Fig. A15: Comparison of different methods in the case $\rho_2 = 2$, $p_z = 9$, $p_x = 10$ and $n = 1000$. The x-axis represents the concentration parameter. On the y-axis, MAE represents Median Absolute Error of the estimators, Coverage represents coverage of confidence intervals and Length represents the average length of confidence intervals. Proposed is our method allowing for invalid IVs and is represented by the solid line. Proposed.valid is our method that assumes all the IVs are valid and is represented by the dashed line. Oracle is the method that knows exactly which instruments are valid and is represented by the dotted line. The column labeled with Weak IV1 represents the case $\rho_1 = 0.1$, the column labeled with Weak IV2 represents the case $\rho_1 = 0.2$ and the column labeled with Strong IV represents the case $\rho_1 = 0$.

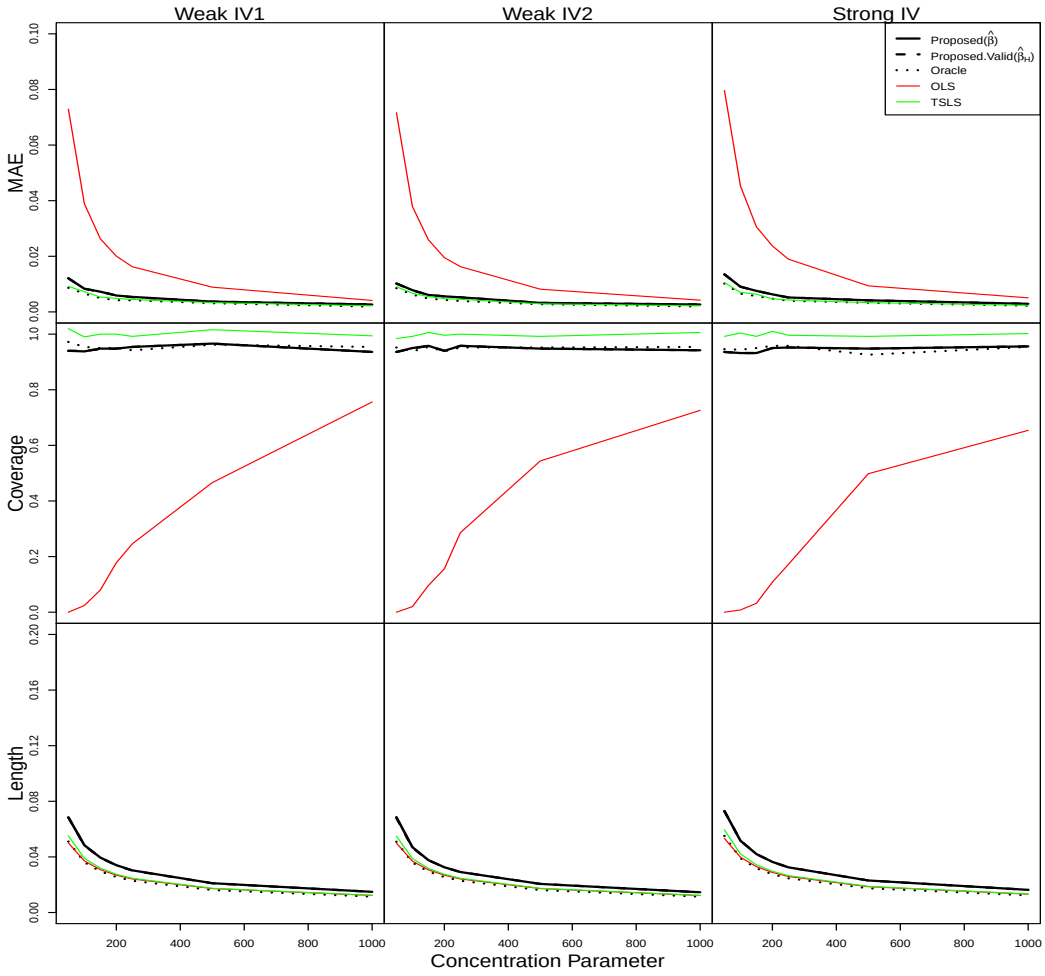

Fig. A16: Comparison of different methods in the case $\rho_2 = 0$, $p_z = 9$, $p_x = 150$ and $n = 1000$. The x-axis represents the concentration parameter. On the y-axis, MAE represents Median Absolute Error of the estimators, Coverage represents coverage of confidence intervals and Length represents the average length of confidence intervals. Proposed is our method allowing for invalid IVs and is represented by the solid line. Proposed.valid is our method that assumes all the IVs are valid and is represented by the dashed line. Oracle is the method that knows exactly which instruments are valid and is represented by the dotted line. The column labeled with Weak IV1 represents the case $\rho_1 = 0.1$, the column labeled with Weak IV2 represents the case $\rho_1 = 0.2$ and the column labeled with Strong IV represents the case $\rho_1 = 0$.

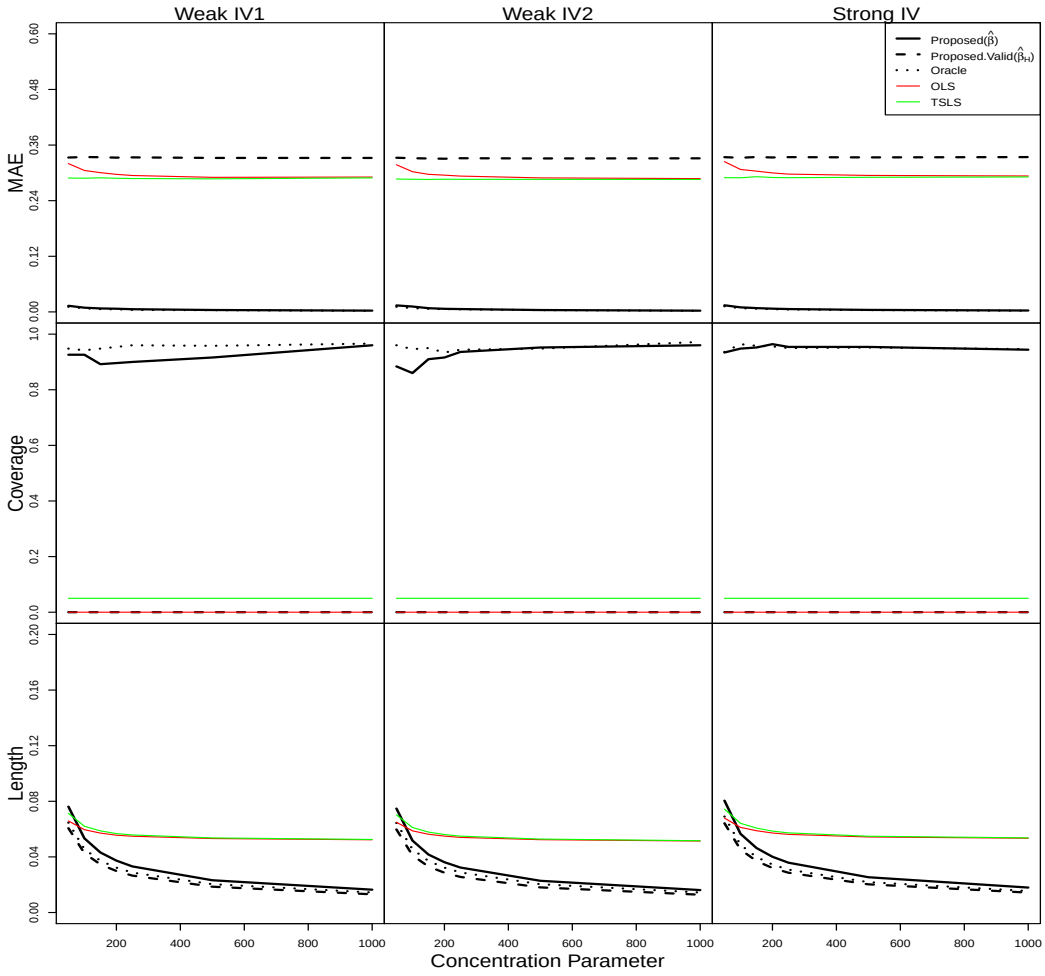

Fig. A17: Comparison of different methods in the case $\rho_2 = 1$, $p_z = 9$, $p_x = 150$ and $n = 1000$. The x-axis represents the concentration parameter. On the y-axis, MAE represents Median Absolute Error of the estimators, Coverage represents coverage of confidence intervals and Length represents the average length of confidence intervals. Proposed is our method allowing for invalid IVs and is represented by the solid line. Proposed.valid is our method that assumes all the IVs are valid and is represented by the dashed line. Oracle is the method that knows exactly which instruments are valid and is represented by the dotted line. The column labeled with Weak IV1 represents the case $\rho_1 = 0.1$, the column labeled with Weak IV2 represents the case $\rho_1 = 0.2$ and the column labeled with Strong IV represents the case $\rho_1 = 0$.

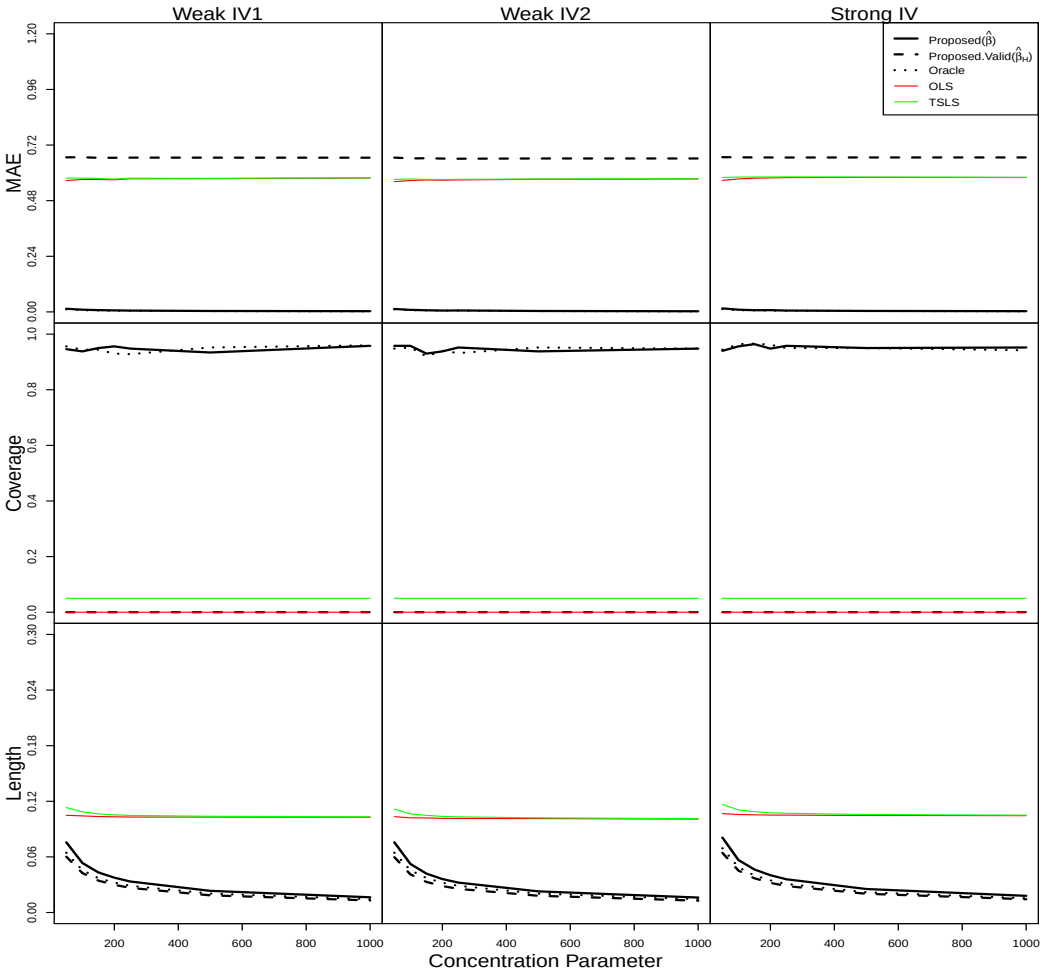

Fig. A18: Comparison of different methods in the case $\rho_2 = 2$, $p_z = 9$, $p_x = 150$ and $n = 1000$. The x-axis represents the concentration parameter. On the y-axis, MAE represents Median Absolute Error of the estimators, Coverage represents coverage of confidence intervals and Length represents the average length of confidence intervals. Proposed is our method allowing for invalid IVs and is represented by the solid line. Proposed.valid is our method that assumes all the IVs are valid and is represented by the dashed line. Oracle is the method that knows exactly which instruments are valid and is represented by the dotted line. The column labeled with Weak IV1 represents the case $\rho_1 = 0.1$, the column labeled with Weak IV2 represents the case $\rho_1 = 0.2$ and the column labeled with Strong IV represents the case $\rho_1 = 0$.