

# Testing Overidentifying Restrictions with High-Dimensional Data and Heteroskedasticity

Qingliang Fan\*

Zijian Guo<sup>†</sup>

Ziwei Mei<sup>‡</sup>

May 3, 2022

## Abstract

This paper proposes a new test of overidentifying restrictions (called the Q test) with high-dimensional data. This test is based on estimation and inference for a quadratic form of high-dimensional parameters. It is shown to have the desired asymptotic size and power properties under heteroskedasticity, even if the number of instruments and covariates is larger than the sample size. Simulation results show that the new test performs favorably compared to existing alternative tests (Chao et al., 2014; Kolesár, 2018; Carrasco and Doukali, 2021) under the scenarios when those tests are feasible or not. An empirical example of the trade and economic growth nexus manifests the usefulness of the proposed test.

**JEL classification:** C12, C21, C26, C55

**Keywords:** many instruments, overidentification test, quadratic form, heteroskedasticity, data-rich environment.

## 1 Introduction

A valid instrumental variable (IV) satisfies the exclusion restriction condition. Conditioning on the covariates, valid IVs do not have direct effects on the outcome and are independent of the unobserved confounders. Motivated by the increasing availability of large datasets, such as large survey data, clinical data, and administrative data, this paper proposes a new test of

---

\*Department of Economics, The Chinese University of Hong Kong. E-mail: michaelqfan@gmail.com.

<sup>†</sup>Department of Statistics, Rutgers University. Email: zijguo@stat.rutgers.edu.

<sup>‡</sup>Department of Economics, the Chinese University of Hong Kong. Email: zwmei@link.cuhk.edu.hk.

overidentifying restrictions that is robust to the presence of high-dimensional instruments and covariates with heteroskedastic errors.

The Sargan test (Sargan, 1958) is a special case of the J test by Hansen (1982), where the latter is robust to heteroskedastic errors in a GMM framework. These classical overidentification tests can be sensitive to the number of IVs. Under homoskedasticity, Lee and Okui (2012) propose a modified Sargan test compatible with an increasing number of instruments and show its equivalence to the Hahn-Hausman test (Hahn and Hausman, 2002). Chao et al. (2014) propose an overidentification test (called jackknife test hereafter) robust to many instruments and heteroskedasticity following the idea of jackknife instrumental variable estimation (JIVE, Angrist et al., 1999). However, these tests can only be applied if the number of instruments is smaller than the sample size. Carrasco and Doukali (2021) extend this method based on regularized JIVE (Hansen and Kozbur, 2014; Carrasco and Doukali, 2017) to allow the number of instruments to be larger than the sample size (called regularized jackknife test hereafter).

The abovementioned tests cannot allow a growing number of exogenous covariates. In applications, the empirical model often includes many potential covariates or the nonlinear terms of original variables, which are very likely to be high-dimensional (Belloni et al., 2014). Based on Cragg and Donald (1993), Kolesár (2018) extends Anatolyev and Gospodinov (2011) and proposes a modified Cragg-Donald test (MCD hereafter) test that allows many covariates and instruments. However, it requires homoscedasticity and the total number of exogenous variables  $p$  to be less than the sample size  $n$ , and the test has poor power when  $p \approx n$ .

Our paper extends the scope of Chao et al. (2014) and Kolesár (2018) by allowing the number of IVs and covariates to be both larger than the sample size. Unlike these  $\chi^2$  type of tests, the new test (called the  $Q$  test) is based on estimation and inference for a *Quadratic* form of high-dimensional parameters. This method has advantages over other existing tests in that it accommodates unknown sparsity patterns of the high-dimensional IV model, hence improving the power of the test while keeping the type I error under control. We apply Lasso estimation in the high-dimensional reduced form models and correct for its bias using the projection direction procedure similar to Guo et al. (2019). The instrument validity test we propose is flexible to the dimensionality of covariates  $p_x$  and instruments  $p_z$ . Specifically, the test allows different combinations of  $p_x$  and  $p_z$  to approach infinity with the sample size, as  $p = p_x + p_z \rightarrow \infty$ . Notably, we allow  $p_z$  to be *fixed*, and the covariates are of high dimension so that  $p > n$ . The proposed test can also check the validity of any overidentified subset of instruments, which makes it handy for empirical researchers. It is possible that when potential instruments are high-dimensional, the researcher might just want to test the validity of a small subset of IVs first.

**Main Results.** We show that under sparse settings, the  $Q$  test has the correct asymptotic size. Furthermore, the test has asymptotic power equal to one when the violation of the null hypothesis is strong enough. In contrast to other recently proposed tests (Chao et al., 2014;

(Kolesár, 2018; Carrasco and Doukali, 2021), our test is valid and powerful even when the numbers of both instruments and covariates are larger than the sample size. For sparse models, the test is shown to achieve higher asymptotic power than the jackknife-type overidentification tests by Chao et al. (2014) and Carrasco and Doukali (2021); see the discussions after Theorem 2 in Section 3 for more details. Our test can be easily implemented by practitioners with a high-dimensional dataset<sup>1</sup>.

In the simulation studies, the Q test shows apparent power improvement over other existing tests in strictly and approximately sparse settings and for both  $p > n$  and  $p \leq n$ . It has the highest empirical power even in the settings with no covariates and many instruments when all other tests are also feasible. In the settings with covariates, it is found that while other methods are sensitive to the dimension of covariates, the Q test has robust and satisfactory power performance under different combinations of  $n$  and  $p_x$ . Besides, even with relatively low individual IV strength, the test is still considerably more powerful than other tests though all tests suffer from power loss with relatively weak instruments. In the empirical example about the effect of trade on economic growth, we perform overidentifying restrictions tests on a large set of instruments, including several possibly invalid ones such as energy usage. The Q test strongly rejects the null hypothesis of correct specification, while other tests (Chao et al., 2014; Kolesár, 2018; Carrasco and Doukali, 2021) could not reject it, indicating the superiority of the new test under high-dimensional IV models.

**Our Contributions.** The main contributions of this paper are summarized as follows. First, the proposed Q test extends the current literature to the high-dimensional setting, which allows both the covariates and instruments to be high-dimensional with  $p > n$ . The test fits popular high-dimensional IV models with sparse settings, which has been commonly discussed in the literature (Belloni et al., 2012; Caner and Fan, 2015; Gold et al., 2020). The Q test is constructed based on a quadratic form of high-dimensional parameters and accommodates the heteroskedastic errors, which is common in IV models but has not been studied in earlier literature on quadratic form estimation or inference (Guo et al., 2019; Cai and Guo, 2020; Guo et al., 2021). Second, we establish the theoretical power of our test and compare it with the powers of other overidentification tests. For sparse models, we show that our test has higher asymptotic power than other recently proposed methods (Chao et al., 2014; Kolesár, 2018; Carrasco and Doukali, 2021) when the number of instruments grows proportionally to the sample size (i.e.,  $p_z/n \rightarrow \alpha_z > 0$ ). See the detailed comparison in the discussions after Theorem 2 in Section 3 and formal theoretical analysis in Appendix A. Third, we improve the bias correction of the Lasso-type estimator when estimating the quadratic functionals. Our debiasing method does not require sample splitting, therefore it has the advantage of finite sample efficiency. We propose a novel solution to deal with the degenerate variance issue, which also appears in

---

<sup>1</sup>The computer codes for the implementation of the above method are available at <https://github.com/ZiweiMEI/QTest>.

existing literature (Shi, 2015; Hsu and Shi, 2017; Liu and Lee, 2019; Cai and Guo, 2020; Guo et al., 2021), by further calibrating our debiased estimator of quadratic form and construct a test with uniformly correct asymptotic size without introducing extra randomness.

**Other Related Literature.** A strand of literature studies the estimation problem via valid IV selection (Kang et al., 2016; Guo et al., 2018; Windmeijer et al., 2019, 2021; Fan and Wu, 2020). These methods are based on demanding restrictions on either the decision matrix or the number of valid instruments (e.g., irrepresentable condition, majority rule, plurality rule). Among these works, Guo et al. (2018) discussed inference for the treatment effect in the presence of possibly invalid IVs and high-dimensional covariates. However, as pointed out in Guo (2021), the selection of valid IVs might suffer from the selection error in finite samples. The aforementioned selection methods require the majority or plurality rule, that is, a large proportion of IVs are valid. On the other hand, our current proposal works even if the majority or plurality rule does not hold. Some other literature (Liao, 2013; Cheng and Liao, 2015; Caner et al., 2018; Chang et al., 2021) studies moment condition selection under the GMM framework, requiring prior information about some moment conditions known to be valid. This paper also relates to a growing literature on studies of the quadratic form inference, such as Fan et al. (2021) with applications on the Markowitz mean-variance portfolio.

In addition, the Q test can be done without the oracle property of selecting relevant instruments from the reduced form equation and hence does not need the individual strength assumption on the instruments. This relates to the study of the high-dimensional IV models from the perspective of estimation (Donald and Newey, 2001; Okui, 2011; Bai and Ng, 2010; Carrasco, 2012; Bekker and Cruje, 2015; Fan and Zhong, 2018; Zhu, 2018). The sparsity on the reduced form equation relates to the literature on weak instruments. We refer to Andrews et al. (2019) for a recent and comprehensive review of that literature.

**Notations.** We clarify some notations before moving on. For any positive integer  $p$ , we define  $[p] = \{1, 2, \dots, p\}$ ,  $\lfloor p \rfloor$  to be the largest integer no greater than  $p$ , and  $\lceil p \rceil$  to be the smallest integer no less than  $p$ . For any  $G \subset [p]$ , let  $x_G$  denote the subvector of  $x$  given by  $(x_j)_{j \in G}$  where  $x$  is a  $p \times 1$  vector. Moreover, for any matrix  $X$  with  $p$  columns, let  $X_G$  denote the submatrix of  $X$  composed of the columns indexed by the subset  $G$ . For two positive sequences  $\{a_n\}$  and  $\{b_n\}$ ,  $a_n \lesssim b_n$  means  $a_n \leq Cb_n$  for some universal constant  $C > 0$  and all sufficiently large  $n$ , and  $b_n \gtrsim a_n$  means  $b_n \geq Ca_n$ . We use  $a_n \ll b_n$  or  $b_n \gg a_n$  to denote  $a_n = o(b_n)$ . The statement “ $\mathcal{A}$  w.p.a.1” for some event  $\mathcal{A}$  means that  $\lim_{n,p \rightarrow \infty} \Pr(\mathcal{A}) = 1$ , where “w.p.a.1” is an abbreviation of “with probability approaching one”. We use  $\mathbf{1}\{\cdot\}$  to denote the indicator function such that  $\mathbf{1}\{\mathcal{A}\} = 1$  if event  $\mathcal{A}$  is true and zero otherwise. For a vector  $x = (x_1, x_2, \dots, x_p)^\top \in \mathbb{R}^p$ , we define  $\|x\|_1 = \sum_{j=1}^p |x_j|$ ,  $\|x\|_2 = \sqrt{\sum_{j=1}^p |x_j|^2}$ ,  $\|x\|_\infty = \max_{1 \leq j \leq p} |x_j|$  and  $\|x\|_0 = |\{1 \leq j \leq p : x_j \neq 0\}|$ . For any  $p \times p$  square matrix  $A = (A_{ij})_{i,j=1}^p$ ,  $\|A\|_2 = \sup_{\|x\|_2 \leq 1} \|Ax\|_2$ ,  $\|A\|_1 = \max_{1 \leq i \leq p} \sum_{j=1}^p |A_{ij}|$ ,  $\|A\|_\infty = \max_{i,j} |A_{ij}|$ . For any matrix  $A$ ,  $A \succ 0$  means  $A$  is positive definite. We use  $\langle x_1, x_2 \rangle$  to denote the inner product of any two vectors  $x_1, x_2$  given by  $x_1^\top x_2$ .

The remainder of the paper is organized as follows. In Section 2, we present the Q test. Section 3 provides the main theoretical results. Section 4 demonstrates the finite sample performance of the proposed Q test, including an empirical example of trade on growth. Section 5 concludes the paper. Technical proofs and robustness simulation results are given in the Appendix.

## 2 Methodology

### 2.1 The Model and Problem Formulation

We consider the following classic linear IV model: for  $i = 1, 2, \dots, n$ ,

$$\begin{aligned} Y_i &= D_i\beta + X_{i\cdot}^\top \varphi + Z_{i\cdot}^\top \pi + e_i, & \mathbb{E}(e_i|Z_{i\cdot}, X_{i\cdot}) &= 0, \\ D_i &= X_{i\cdot}^\top \psi + Z_{i\cdot}^\top \gamma + \varepsilon_{i2}, & \mathbb{E}(\varepsilon_{i2}|Z_{i\cdot}, X_{i\cdot}) &= 0, \end{aligned} \quad (1)$$

where  $Y_i \in \mathbb{R}$  denotes the outcome variable,  $D_i \in \mathbb{R}$  denotes the endogenous variable,  $X_{i\cdot} \in \mathbb{R}^{p_x}$  and  $Z_{i\cdot} \in \mathbb{R}^{p_z}$  denote the covariates and the instrumental variables, respectively. We focus on the setting of a scalar endogenous variable  $D_i$ , which is common in empirical studies (Kolesár, 2018; Windmeijer et al., 2019; Mikusheva and Sun, 2020). The baseline model (1) can be extended to the scenarios with multiple and even high-dimensional endogenous variables with a satisfactory estimator for  $\beta$  (Breunig et al., 2020; Gold et al., 2020). The parameter of interest  $\beta$  has an interpretation as the treatment effect.  $\varphi$ ,  $\pi$ ,  $\psi$ , and  $\gamma$  are coefficients of conformable dimensions in the model. Since  $e_i$  and  $\varepsilon_{i2}$  are correlated,  $D_i$  is endogenous even conditional on all instruments and covariates. The test operates via a random sample  $\{Y_i, D_i, X_{i\cdot}, Z_{i\cdot}\}_{1 \leq i \leq n}$ . Denote  $Y = (Y_1, Y_2, \dots, Y_n)^\top$ ,  $D = (D_1, D_2, \dots, D_n)^\top$ ,  $X = [X_{1\cdot}, X_{2\cdot}, \dots, X_{n\cdot}]^\top$  and  $Z = [Z_{1\cdot}, Z_{2\cdot}, \dots, Z_{n\cdot}]^\top$ .

We allow the heteroskedastic errors in model (1) so that  $\text{Var}(e_i|Z_{i\cdot}, X_{i\cdot})$  and  $\text{Var}(\varepsilon_{i2}|Z_{i\cdot}, X_{i\cdot})$  may vary with  $i$ . As discussed in Section 1, the test allows  $p = p_x + p_z \rightarrow \infty$  in any combination of  $p_x$  and  $p_z$ . We also impose sparsity assumption on the model (1) following the literature on similar models with high-dimensional IVs and baseline covariates (e.g., Belloni et al., 2012, 2014; Kolesár et al., 2015; Guo et al., 2018; Kolesár, 2018).

Under model (1), if  $\pi = 0$ , all the instruments in  $Z$  are valid, that is, satisfying the exclusion restriction condition. If some elements of  $\pi$  are nonzero, then the instrumental variables associated with those nonzero  $\pi$ 's are invalid. We present the methodology by focusing on the correct specification of all instruments, which is of interest in many popular overidentification tests. Consequently, the test is reduced to testing the null hypothesis  $H_0 : \pi = 0$ , or equivalently,

$$H_0 : \|\pi\|_2^2 = 0. \quad (2)$$

Before moving on, we want to clarify that model (1) is more general than it appears in the following perspectives: a)  $X$  and  $Z$  are allowed to include the nonlinear terms of the original economic variables, such as B-splines, dummies, polynomials, and various interactions; b) We can consider the overidentifying restrictions of any subset  $G \subset [p_z]$  of instruments with  $|G| > 1$  by simply regarding all other instruments as covariates. In this case, the new instruments become  $\tilde{Z} = Z_{\cdot G}$ , and the covariates are given by  $\tilde{X} = [X, Z_{\cdot G^c}]$  where  $G^c = [p_z] \setminus G$  is the complement set. Without loss of generality,  $\pi_G = 0$  becomes the new condition for the instrument validity.

Our proposed Q test (detailed in Section 2.3) relates to Sargan's overidentification test, which regresses the residual  $Y - D\hat{\beta}_{2SLS}$  on  $(X, Z)$  and tests the joint significance of the coefficients of  $Z$ , where  $\hat{\beta}_{2SLS}$  is the two-stage least square estimator of  $\beta$ . Under linear model (1), it is equivalent to testing the null hypothesis (2). Following the same spirit, our proposed Q test is constructed by regressing the residual  $Y - D\hat{\beta}$  on all exogenous variables once we obtain a satisfactory estimator  $\hat{\beta}$ . For this purpose, we first look at the reduced form of model (1), from which we seek an estimator of  $\beta$ ,

$$\begin{aligned} Y &= X\Psi + Z\Gamma + \varepsilon_1, \\ D &= X\psi + Z\gamma + \varepsilon_2, \end{aligned} \tag{3}$$

where  $\Psi = \psi\beta + \varphi$ ,  $\Gamma = \gamma\beta + \pi$  and  $\varepsilon_1 = \varepsilon_2\beta + e$  with  $e = (e_1, e_2, \dots, e_n)^\top$  and  $\varepsilon_2 = (\varepsilon_{12}, \varepsilon_{22}, \dots, \varepsilon_{n2})^\top$ . Let  $s = \max\{\|\varphi\|_0, \|\psi\|_0, \|\pi\|_0, \|\gamma\|_0, \|\Gamma\|_0, \|\Psi\|_0\}$ . For simplicity, we also specify the sparsity level of  $\Gamma$  and  $\Psi$ , which can actually be induced by sparsity of other vectors. We specify the rate of sparsity level  $s$  in Assumption 4 of Section 3.

We define  $W = [X, Z]$  as the  $n \times p$  design matrix with  $p = p_z + p_x$ . Define  $\Sigma = \mathbb{E}(W_i W_i^\top)$ ,  $\Omega = \Sigma^{-1}$ ,  $\sigma_{i1}^2 = \text{Var}(\varepsilon_{i1}|W_i)$ ,  $\sigma_{i2}^2 = \text{Var}(\varepsilon_{i2}|W_i)$  and  $\sigma_{i12} = \text{cov}(\varepsilon_{i1}, \varepsilon_{i2}|W_i)$ , for  $1 \leq i \leq n$ . Let  $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n W_i W_i^\top = \frac{1}{n} W^\top W$ .

In the following, we derive how to identify a scaled version of  $\|\pi\|_2^2$  by the reduced form (3). If  $\|\gamma\|_2^2 > 0$ , we apply the expression  $\Gamma = \gamma\beta + \pi$ , and identify  $\beta$  as

$$\beta = \frac{\langle \Gamma, \gamma \rangle}{\|\gamma\|_2^2} - \frac{\langle \pi, \gamma \rangle}{\|\gamma\|_2^2}.$$

Define  $\beta_R \equiv \langle \Gamma, \gamma \rangle / \|\gamma\|_2^2$  where the subscript R means that  $\beta$  is identified by the *Reduced form* parameters. Note that  $\beta_R - \beta = \frac{\langle \pi, \gamma \rangle}{\|\gamma\|_2^2}$  for a general  $\pi$ , and under the null with  $\pi = 0$ , we have  $\beta = \beta_R$ . Subtracting  $D\beta_R$  from the structural equation, we have

$$Y - D\beta_R = X\tilde{\varphi} + Z\tilde{\pi} + \tilde{e}, \tag{4}$$

where  $\tilde{\varphi} = \varphi - \psi(\beta_R - \beta)$ ,  $\tilde{\pi} = \pi - \gamma(\beta_R - \beta)$  and  $\tilde{e} = \varepsilon_1 - \varepsilon_2\beta_R$ . By the definition of  $\beta_R$ , we

derive the following expression of the quadratic form of  $\tilde{\pi}$ ,

$$Q \equiv \|\tilde{\pi}\|_2^2 = \|\pi - \gamma(\beta_R - \beta)\|_2^2 = \|\pi\|_2^2 - \frac{\langle \pi, \gamma \rangle^2}{\|\gamma\|_2^2} = \|\pi\|_2^2 [1 - R^2(\pi, \gamma)], \quad (5)$$

where  $R(\pi, \gamma) = \frac{\langle \pi, \gamma \rangle}{\|\pi\|_2 \|\gamma\|_2} \mathbf{1}\{\|\pi\|_2 > 0, \|\gamma\|_2 > 0\}$  is the relatedness between  $\pi$  and  $\gamma$ . Note that  $|R(\pi, \gamma)| \neq 1$  means that  $\pi$  is not perfectly linearly dependent on  $\gamma$ , which is true when the overidentification condition is satisfied.

We shall point out that  $\|\tilde{\pi}\|_2^2 \neq 0$  if and only if  $\|\pi\|_2^2 \neq 0$  and  $|R(\pi, \gamma)| \neq 1$ . Hence, if  $|R(\pi, \gamma)| \neq 1$ , the null hypothesis (2) is equivalent to

$$H_0 : Q = 0. \quad (6)$$

So far, we have transformed our problem into an equivalent hypothesis testing problem in (6). We lay out testing procedures in the next subsection.

## 2.2 Estimation of $\beta_R$

We now discuss the estimation of  $\beta_R$  using its definition  $\beta_R \equiv \langle \Gamma, \gamma \rangle / \|\gamma\|_2^2$ . We apply the Lasso estimator (Tibshirani, 1996) to estimate  $\Gamma$  and  $\gamma$ :

$$\{\hat{\Psi}, \hat{\Gamma}\} = \arg \min_{\Psi, \Gamma} \frac{1}{n} \|Y - X\Psi - Z\Gamma\|_2^2 + \lambda_{1n}(\|\Psi\|_1 + \|\Gamma\|_1), \quad (7)$$

$$\{\hat{\psi}, \hat{\gamma}\} = \arg \min_{\psi, \gamma} \frac{1}{n} \|D - X\psi - Z\gamma\|_2^2 + \lambda_{2n}(\|\psi\|_1 + \|\gamma\|_1), \quad (8)$$

where  $\lambda_{1n}, \lambda_{2n}$  are positive tuning parameters selected by cross-validation in practice. We discuss the details of tuning parameter selection in Section 4. It is well known that estimation of  $\beta_R$  by plug-in Lasso estimators in (7) and (8), say  $\hat{\Gamma}^\top \hat{\gamma} / \|\hat{\gamma}\|_2^2$ , causes bias (Guo et al., 2019). Here we generalize the debiasing method for the quadratic form of high-dimensional parameters in the recent literature (Guo et al., 2019; Cai and Guo, 2020; Guo et al., 2021). Since  $\beta_R$  depends on quadratic transformations of the high-dimensional parameters, our debiasing procedure is different from the debiased Lasso methods such as Belloni et al. (2014), Zhang and Zhang (2014), and van de Geer et al. (2014). We specify our bias correction procedure in the following. First, for the numerator of  $\beta_R$ ,  $\langle \Gamma, \gamma \rangle = \Gamma^\top \gamma$ , the estimation error of the plug-in estimator  $\hat{\Gamma}^\top \hat{\gamma}$  is

$$\hat{\Gamma}^\top \hat{\gamma} - \Gamma^\top \gamma = \hat{\Gamma}^\top (\hat{\gamma} - \gamma) + \hat{\gamma}^\top (\hat{\Gamma} - \Gamma) - (\hat{\Gamma} - \Gamma)^\top (\hat{\gamma} - \gamma). \quad (9)$$

We mainly need to estimate the first two terms on the RHS of (9), which are the bias



components induced by plugging in the Lasso estimators. The last term in the RHS of (9) is negligible. The main idea is to estimate the bias term  $\widehat{\Gamma}^\top(\gamma - \widehat{\gamma})$  by  $u^\top \frac{1}{n} W^\top (D - X\widehat{\psi} - Z\widehat{\gamma})$ , where  $u \in \mathbb{R}^p$  is a projection direction vector to be constructed. For this purpose, we decompose the estimation error as follows,

$$u^\top \frac{1}{n} W^\top (D - X\widehat{\psi} - Z\widehat{\gamma}) - \widehat{\Gamma}^\top(\gamma - \widehat{\gamma}) = \left( u^\top \widehat{\Sigma} - (0_{p_x}^\top, \widehat{\Gamma}^\top) \right) \begin{pmatrix} \psi - \widehat{\psi} \\ \gamma - \widehat{\gamma} \end{pmatrix} + u^\top \frac{1}{n} W^\top \varepsilon_2, \quad (10)$$

where  $0_{p_x}$  is a  $p_x \times 1$  zero vector. To ensure that  $u^\top \frac{1}{n} W^\top (D - X\widehat{\psi} - Z\widehat{\gamma})$  accurately approximates  $\widehat{\Gamma}^\top(\widehat{\gamma} - \gamma)$ , we construct the following projection direction

$$\widehat{u}_1 = \arg \min_{u \in \mathbb{R}^p} \|u\|_1, \text{ s.t. } \|\widehat{\Sigma}u - (0_{p_x}^\top, \widehat{\Gamma}^\top)\|_\infty \leq \|\widehat{\Gamma}\|_2 \mu_{1n} \quad (11)$$

where  $\mu_{1n}$  is a positive tuning parameter. The  $\ell_1$  minimization ensures that the high-dimensional projection direction is stably constructed, while the constraint in (11) ensures that the first term on the right hand side of (9) is small. In contrast to minimizing the quadratic objective function in Guo et al. (2019, 2021), we use the  $\ell_1$  objective function to make our model compatible with heteroskedastic errors and also avoid sample splitting in estimating the quadratic form of high-dimensional parameters.

Similarly, we estimate  $\widehat{\gamma}^\top(\Gamma - \widehat{\Gamma})$  by  $\widehat{u}_2^\top \frac{1}{n} W^\top (Y - X\widehat{\Psi} - Z\widehat{\Gamma})$  where

$$\widehat{u}_2 = \arg \min_{u \in \mathbb{R}^p} \|u\|_1, \text{ s.t. } \|\widehat{\Sigma}u - (0_{p_x}^\top, \widehat{\gamma}^\top)\|_\infty \leq \|\widehat{\gamma}\|_2 \mu_{2n} \quad (12)$$

with  $\mu_{2n}$  being a positive tuning parameter. We then estimate  $\langle \Gamma, \gamma \rangle$  by

$$\widehat{\langle \Gamma, \gamma \rangle} = \widehat{\Gamma}^\top \widehat{\gamma} + \widehat{u}_1^\top \frac{1}{n} W^\top (D - X\widehat{\psi} - Z\widehat{\gamma}) + \widehat{u}_2^\top \frac{1}{n} W^\top (Y - X\widehat{\Psi} - Z\widehat{\Gamma}).$$

By applying a similar bias-correction idea, we estimate  $\|\gamma\|_2^2$  by

$$\widehat{\|\gamma\|_2^2} = \widehat{\gamma}^\top \widehat{\gamma} + \frac{2}{n} \widehat{u}_2^\top (D - X\widehat{\psi} - Z\widehat{\gamma}). \quad (13)$$

Finally, we propose the bias-corrected estimator of  $\beta_R \equiv \langle \Gamma, \gamma \rangle / \|\gamma\|_2^2$  as

$$\widehat{\beta}_R = \frac{\widehat{\langle \Gamma, \gamma \rangle}}{\widehat{\|\gamma\|_2^2}} \mathbf{1}_{\{\widehat{\|\gamma\|_2^2} > 0\}}. \quad (14)$$



### 2.3 The Q test

In the following, we construct a point estimator of  $Q$  defined in (5) and then propose a test of  $Q = 0$ . We substitute  $\beta_R$  by  $\hat{\beta}_R$  in equation (4), and get

$$Y - D\hat{\beta}_R = X\check{\varphi} + Z\check{\pi} + \check{\epsilon}, \quad (15)$$

where  $\check{\varphi} = \varphi - \psi(\hat{\beta}_R - \beta)$ ,  $\check{\pi} = \pi - \gamma(\hat{\beta}_R - \beta) = \tilde{\pi} - \gamma(\hat{\beta}_R - \beta_R)$  and  $\check{\epsilon} = \varepsilon_1 - \varepsilon_2\hat{\beta}_R$ . By the definition of  $\tilde{\pi} = \pi - \gamma(\beta_R - \beta)$  in (4), we obtain

$$\langle \tilde{\pi}, \gamma \rangle = \langle \pi, \gamma \rangle - \|\gamma\|_2^2(\beta_R - \beta) = \langle \pi, \gamma \rangle - \|\gamma\|_2^2 \langle \pi, \gamma \rangle / \|\gamma\|_2^2 = 0.$$

We apply the above equation and approximate  $Q = \|\tilde{\pi}\|_2^2$  by

$$\begin{aligned} \check{Q} &\equiv \|\check{\pi}\|_2^2 = \|\tilde{\pi}\|_2^2 + \|\gamma\|_2^2(\hat{\beta}_R - \beta_R)^2 - 2\langle \tilde{\pi}, \gamma \rangle(\hat{\beta}_R - \beta_R) \\ &= Q + \|\gamma\|_2^2(\hat{\beta}_R - \beta_R)^2. \end{aligned} \quad (16)$$

We show in (B3) in the Appendix that  $\hat{\beta}_R$  is a good estimator of  $\beta_R$  and hence  $\|\gamma\|_2^2(\hat{\beta}_R - \beta_R)^2$  is small, which ensures that  $\check{Q}$  approximates  $Q$  well.

In the following, we propose an estimator of the proxy  $\check{Q}$ , which is also an estimator of  $Q$ . We apply Lasso to estimate the  $\check{\pi}$  from (15),

$$\{\hat{\varphi}, \hat{\pi}\} = \arg \min_{\check{\varphi}, \check{\pi}} \frac{1}{n} \|Y - D\hat{\beta}_R - X\check{\varphi} - Z\check{\pi}\|_2^2 + \lambda_{3n}(\|\check{\varphi}\|_1 + \|\check{\pi}\|_1) \quad (17)$$

where  $\lambda_{3n}$  is a positive tuning parameter selected by cross-validation. Following the same idea about debiased estimators of  $\|\gamma\|_2^2$  and  $\langle \Gamma, \gamma \rangle$ , we construct a projection direction as

$$\hat{u}_3 = \arg \min_{u \in \mathbb{R}^p} \|u\|_1, \text{ s.t. } \|\hat{\Sigma}u - (0_{p_x}^\top, \hat{\pi}^\top)^\top\|_\infty \leq \|\hat{\pi}\|_2 \mu_{3n}, \quad (18)$$

where  $\mu_{3n}$  is a positive tuning parameter. We propose the following bias-corrected estimator of  $Q$  as

$$\hat{Q}^0 = \hat{\pi}^\top \hat{\pi} + \frac{2}{n} \hat{u}_3^\top W^\top (Y - D\hat{\beta}_R - X\hat{\varphi} - Z\hat{\pi}). \quad (19)$$

The conditional asymptotic variance of  $\hat{Q}^0$  can be estimated by

$$\hat{V}^0 = \frac{4}{n} \hat{u}_3^\top \sum_{i=1}^n W_i W_i^\top \hat{e}_i^2 \hat{u}_3, \quad (20)$$

where  $\hat{e}_i = Y_i - D_i\hat{\beta}_R - X_i^\top \hat{\varphi} - Z_i^\top \hat{\pi}$ .

We highlight that  $\widehat{Q}^0$  is not our proposed test statistic since  $\widehat{Q}^0$  suffers from the super-efficiency issue as other estimators of high-dimensional quadratic functions (Guo et al., 2021). The super-efficiency means that the asymptotic variance of  $\widehat{Q}^0$  might converge to zero faster than a parametric rate near the null hypothesis. In such a scenario, even if  $\widehat{Q}^0$  is still an accurate estimator,  $\widehat{V}^0$  might not accurately estimate the uncertainty of  $\widehat{Q}^0$ . To address this, we modify the estimator in (19) and estimate  $Q$  by

$$\widehat{Q} = \widehat{\pi}^\top \widehat{\pi} + \frac{2}{n} (W\widehat{u}_3 + \sqrt{\tau_n} \cdot \eta)^\top (Y - D\widehat{\beta}_R - X\widehat{\varphi} - Z\widehat{\pi}) \quad (21)$$

where  $\eta \in \mathbb{R}^n$  and  $\tau_n > 0$  will be specified in the following. Compared to  $\widehat{Q}^0$ , the additional term in  $\widehat{Q}$ ,  $n^{-1}2\sqrt{\tau_n}\eta^\top (Y - D\widehat{\beta}_R - X\widehat{\varphi} - Z\widehat{\pi})$ , inflates the variance of the estimator with a non-degenerate  $\tau_n$ . As a result, the super-efficiency issue no longer exists, and a correct asymptotic size is guaranteed. Accordingly, we estimate the asymptotic variance of  $\widehat{Q}$  by

$$\widehat{V} = \frac{4}{n} \sum_{i=1}^n (W_i^\top \widehat{u}_3 + \sqrt{\tau_n} \cdot \eta_i)^2 \widehat{e}_i^2. \quad (22)$$

Finally, the proposed  $Q$  test is given by

$$\mathcal{Q}_n(\alpha) = \mathbf{1} \left\{ \sqrt{n}\widehat{Q} > z_\alpha \sqrt{\widehat{V}} \right\} \quad (23)$$

where  $z_\alpha$  is the  $100(1-\alpha)$ -th percentile of the standard normal distribution. We use the one-sided test since  $Q$  is non-negative theoretically.

We now discuss the selection of  $\eta$  and  $\tau_n$  in the construction of  $\widehat{Q}$  in (21). We shall specify  $\eta \in \mathbb{R}^n$  as

$$\eta_i = \begin{cases} 1, & \text{if } i \in \mathcal{N}_0, \\ -1, & \text{otherwise} \end{cases} \quad (24)$$

for some subset  $\mathcal{N}_0 \subset [n]$  with cardinality  $\lceil n/2 \rceil$ . In simulation studies, we explore the robustness of our proposed method by different ways of specifying this set  $\mathcal{N}_0$ . See subsection 4.1 for details.

The determination of  $\tau_n$  plays a key role in the size-power balance of the  $Q$  test. Any fixed  $\tau_n > 0$  will lead to a testing procedure achieving the correct type I error asymptotically. For higher power, we define a data-dependent  $\tau_n$  as

$$\tau_n = \frac{\tau_0}{1 + \sqrt{n}\widehat{Q}_+^0 \cdot \log(\log(np))} \quad (25)$$

where  $\tau_0$  is a user-defined parameter set as 1 by default and  $\widehat{Q}_+^0 = \max\{Q^0, 0\}$  is the positive part of  $\widehat{Q}^0$ . We will discuss more about the choice of  $\tau_0$  in subsection 4.1.

We now explain the intuition of choosing  $\tau_n$  as in (25). First, under the null hypothesis that  $Q = 0$ ,  $\widehat{Q}^0$  converges to zero so that  $\tau_n$  is close to a constant  $\tau_0$  and  $\widehat{V}$  defined in (22) is non-degenerate. Second, under the alternative, together with some mild restrictions in Assumption 5 and B1, we can show that  $\tau_n$  converges to zero. In this way, the test has the asymptotic power one when  $Q = \Delta_n/\sqrt{n}$  for any fixed  $\Delta_n > 0$ . In other words, with a data-adaptive  $\tau_n$  chosen by (25), we can have a test with the correct size and also high power compared to a fixed  $\tau_n$ .

**Remark 1.** *We recommend (24) and (25) as our default choice of  $\eta$  and  $\tau_n$  and establish the desirable size and power properties for our recommended choices. However, our proposed method is compatible with many other choices of  $\eta$  and  $\tau_n$ . We will show that the desired size and power of our  $Q$  test can be achieved with a broader class of  $\eta$  and  $\tau_n$ ; see Assumption B1 and Proposition B7 in Appendix B for details.*

**Remark 2.** *The existing literature mainly provides two solutions to the super-efficiency problem. The first solution is to enlarge the estimated asymptotic variance (Cai and Guo, 2020; Guo et al., 2021; Shi, 2015) such that the estimated variance is an upper bound for the true variance when super-efficiency occurs. With this method, the test will become conservative with a downward size distortion. The second solution is to add an additional random normal variable to correct the asymptotic size (Cai and Guo, 2020; Hsu and Shi, 2017; Liu and Lee, 2019). However, under the null hypothesis, the size of the test is entirely determined by the extra randomized component, which is independent of the observed data. (24) and (25) allow us to construct an inference procedure that avoids the downward size distortion and inclusion of extra randomness independent of the observed data.*

### 3 Theoretical Justifications of the Q Test

Throughout the paper, we consider the asymptotic regime where  $p = p(n)$  goes to infinity as  $n$  diverges to infinity. All limits are achieved as  $n \rightarrow \infty$ . Before constructing the theory for our  $Q$  test, we first introduce the following regularity conditions for our theory.

**Assumption 1.** *Suppose that  $W_{i\cdot}$  is independent and identically distributed and has a sub-Gaussian norm  $\|W_{i\cdot}\|_{\psi_2}$  bounded by some constant  $K$ . The matrix  $\Sigma$  satisfies  $c_0 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C_0$  for some positive constant  $C_0 \geq c_0 > 0$ .*

**Assumption 2.** *Suppose that  $e_i$  and  $\varepsilon_{i2}$  are centered sub-Gaussian variables with sub-Gaussian norms bounded by some constant  $K$ . Suppose that  $\mathbb{E}(e_i|W_{i\cdot}) = 0$ ,  $\mathbb{E}(\varepsilon_{i2}|W_{i\cdot}) = 0$  and  $\sigma_{\min}^2 \leq \sigma_{i1}^2, \sigma_{i2}^2 \leq \sigma_{\max}^2$  for some positive constants  $\sigma_{\max}^2 \geq \sigma_{\min}^2 > 0$ . In addition, there exist some positive constants  $c^*$  and  $C^*$  such that  $\mathbb{E}(|\varepsilon_{im}|^{2+c^*}|W_{i\cdot}) \leq C^*$  for  $m = 1, 2$ . Further assume that  $|\sigma_{i12}|/(\sigma_{i1}\sigma_{i2}) \leq \rho_0 < 1$ .*

**Assumption 3.** Define the class of population precision matrices

$$\mathcal{U}(m_\omega, q, s_\omega) := \left\{ \Omega = (\omega_{jk})_{j,k=1}^p \succ 0 : \|\Omega\|_1 \leq m_\omega, \max_{1 \leq j \leq p} \sum_{k=1}^p |\omega_{jk}|^q \leq s_\omega \right\}. \quad (26)$$

where  $0 \leq q < 1$ . Suppose that  $\Omega \in \mathcal{U}(m_\omega, q, s_\omega)$  with  $m_\omega \geq 1$  and  $s_\omega \geq 1$ .

Assumptions 1 and 2 follow the common practice in the literature to regulate the population covariance matrix and the moment conditions of the errors. The sub-Gaussianity assumptions can be relaxed by the moment conditions as in Belloni et al. (2012) and Breunig et al. (2020) at the cost of more complications and a slower convergence rate of reduced form estimators. Assumption 3 considers a widely used class of precision matrices (Cai et al., 2011). Assumption 4 shows the required asymptotic regime.

**Assumption 4** (Asymptotic Regime). Suppose the following conditions hold

- (i)  $\frac{m_\omega^2 s \log(p)}{\sqrt{n}} + \frac{s_\omega m_\omega^{2-q} s^{3/2} \log p}{\sqrt{n}} + \frac{s_\omega m_\omega^{1-q} s^{(3-q)/2} (\log p)^{1-q/2}}{n^{(1-q)/2}} = o\left(\frac{1}{\log[\log(np)]}\right).$
- (ii)  $\frac{m_\omega s^2 \log p (\log(np))^2}{n} + \frac{(\log(np))^{3/2}}{\sqrt{n}} = o(1).$
- (iii) (Global strength of instruments)  $\|\gamma\|_2^2 \gtrsim n^{-1/2}.$

Assumption 4 (i) and (ii) show the restrictions on the sparsity levels and variable dimensions. (i) also guarantees that the bias term can be dominated by  $\sqrt{\tau_n}$  and hence by  $\sqrt{\widehat{V}}$  under the null hypothesis, which implies correct asymptotic size of the Q test. In the well-known literature on debiased Lasso (Javanmard and Montanari, 2014; Zhang and Zhang, 2014; van de Geer et al., 2014) a weaker condition  $s \log p / \sqrt{n} = o(1)$  is sufficient. Here we need slightly stronger restrictions on convergence rates due to the existence of heteroskedasticity.

Assumption 4 (iii) imposes restrictions on the global strength of all instruments. Note that it does not impose restrictions on the individual IV strength since we do not discuss variable selection here (nor does the theory rely on it) but treat all instruments as a whole instead. Other methodologies require stronger assumptions on the global strength. Kolesár (2018) assumes that  $\|\gamma\|_2^2$  converges to some constant. Assumption 2 in Chao et al. (2014) implies that  $\sqrt{n} \|\gamma\|_2^2 \rightarrow \infty$  for the linear setting when  $p_z$  is proportional to  $n$ .

**Assumption 5** (Tuning Parameters). Suppose the following conditions hold

- (i) The Lasso tuning parameters satisfy  $\lambda_{\ell n} = C_{\ell n} \sqrt{\log p / n}$ , for  $\ell = 1, 2, 3$ , where  $\min\{C_{1n}, C_{2n}\} \geq C$  and  $C_{3n} \geq (3 + |\beta| + 2\|\pi\|_2 / \|\gamma\|_2) C$  with a sufficiently large constant  $C > 0$ .

(ii) The tuning parameters for projection directions satisfy  $\mu_{jn} = C_j \sqrt{\log p/n}$  with sufficiently large constant  $C_j > 0$  for  $j = 1, 2, 3$ .

Assumption 5 demonstrates the rates of tuning parameters. (i) and (ii) follow the common practice to impose a rate of  $\sqrt{\log p/n}$  to tuning parameters to Lasso estimate and bias correction (Bickel et al., 2009; Bühlmann and van de Geer, 2011; Javanmard and Montanari, 2014).

Before showing the main theoretical results, we first define the conditional variance

$$V = n^{-1} \sum_{i=1}^n (W_i^\top \Omega(0_{p_x}, \tilde{\pi}^\top)^\top + \sqrt{\tau_n^*} \cdot \eta_i)^2 \sigma_i^2 \quad (27)$$

with  $\sigma_i^2 = \sigma_{i1}^2 + \beta_R^2 \sigma_{i2}^2 - 2\beta_R \sigma_{i12}$  and  $\tau_n^* = \frac{\tau_0}{1 + \sqrt{n}Q \cdot \log[\log(np)]}$ . The following theorem shows the asymptotic normality of the Q test statistic.

**Theorem 1.** Suppose that Assumptions 1-5 hold,  $|R(\pi, \gamma)| \leq r_0 < 1$  for some constant  $r_0$ , and there exists some nonrandom  $V^*$  such that  $V/V^* \xrightarrow{p} 1$ . Let  $\Phi(\cdot)$  be the cumulative distribution function of standard normal distribution. If  $Q = \Delta_n/\sqrt{n}$  for some  $\Delta_n \rightarrow \Delta \in [0, \infty]$ , then for any  $z \in \mathbb{R}$ , the estimator  $\hat{Q}$  defined in (21) satisfies

$$\lim_{n \rightarrow \infty} \left| \Pr \left\{ \frac{\sqrt{n}\hat{Q}}{\sqrt{\hat{V}}} \leq z \right\} - \Phi \left( z - \frac{\Delta_n}{\sqrt{V^*}} \right) \right| = 0, \quad (28)$$

where  $\hat{V}$  is defined in (22). Thus, under the null hypothesis  $H_0 : \|\pi\|_2^2 = 0$ , we have for any  $\alpha \in (0, 1)$ ,

$$\lim_{n \rightarrow \infty} \Pr\{\mathcal{Q}_n(\alpha) = 1\} = \alpha. \quad (29)$$

As mentioned above, condition  $|R(\pi, \gamma)| \leq r_0 < 1$  is necessary to satisfy the overidentification condition. Our Q test is shown to have correct asymptotic size and the desired local power. We formalize the asymptotic power of our Q test in the next theorem.

**Theorem 2.** Suppose that Assumptions 1-5 hold,  $|R(\pi, \gamma)| \leq r_0 < 1$  for some constant  $r_0$ , and  $Q = \Delta_n/\sqrt{n}$  where  $\Delta_n \rightarrow \Delta \in (0, \infty]$ . Then the test  $\mathcal{Q}_n(\alpha)$  defined in (6) satisfies

$$\lim_{n \rightarrow \infty} \Pr\{\mathcal{Q}_n(\alpha) = 1\} = 1. \quad (30)$$

Note that with a fixed  $\tau_n$ , the size of the Q test can still be controlled, but the asymptotic power one is achieved only if  $\Delta_n \rightarrow \infty$ . Similar results (for a different problem of explained variance) have been shown in Cai and Guo (2020). In this paper, we compress the additional variance with a data-dependent choice of  $\tau_n$ , so that asymptotic power one of the Q test can be achieved when  $\Delta_n$  converges to any  $\Delta > 0$ .

**Comparison to Existing Methods.** We compare our methods to the (regularized) jackknife tests (Chao et al., 2014; Carrasco and Doukali, 2021) and the MCD test (Kolesár, 2018). The former is typically based on jackknife instrumental variable estimators, while the latter is a representative version of specification tests based on  $k$ -class estimators allowing high-dimensional covariates.

We show the advantages of the Q test over other methods from two perspectives. First, Theorem 2 shows that the Q test is not sensitive to variable dimensions, allowing  $p_z$  to be of either high or low dimension. In comparison, existing methods depend on stronger restrictions on the variable dimension. The MCD test is compatible with high-dimensional covariates, but requires  $p$  to be strictly less than  $n$ . The jackknife tests cannot handle high-dimensional covariates. In their framework, exogenous covariates are treated as both endogenous variables and included instruments (so they serve as their own instruments). Meanwhile, they assume a fixed number of endogenous variables, which fails to hold as  $p_x$  grows. In the simulation studies, we show that the jackknife test and regularized jackknife test break down with a large  $p_x$ . See Tables 1 and 2 for details.

Second, we show that other methods have lower theoretical power compared to ours under a special setting of  $p_x = 0$  and  $p_z/n \rightarrow \alpha_z \in (0, 1)$ . Both MCD and jackknife tests have asymptotic power one only if  $\sqrt{n}Q \rightarrow \infty$ . This result is consistent with the power analysis in Theorem 4 of Lee and Okui (2012). In comparison, our Q test only requires  $\sqrt{n}Q$  to converge to any fixed positive constant. See Appendix A for more detailed discussions and Appendix B.5 for proofs. In the simulation studies, we will show that the Q test is the most powerful even in the cases where all other tests can be applied theoretically, such as the setting with no baseline covariates and many IVs. Our test improves the power performance because we leverage the sparse model structure.

## 4 Numerical Experiments

### 4.1 Tuning Parameter Selection

We first clarify the tuning parameter selection before moving on. Applying the R package `glmnet`, we select the Lasso tuning parameters  $\lambda_{\ell n}$  for  $\ell = 1, 2, 3$  based on cross validation with the “one-standard-error rule”<sup>2</sup> (Hastie et al., 2009). We generalize the tuning parameter selection method in Gold et al. (2020) to select  $\mu_{kn}$  for  $k = 1, 2, 3$  used in the bias correction steps. When  $p < 0.5n$ , we follow the tuning parameter selection in Gold et al. (2020). Particularly, we set  $\mu_{1n} = 0$  when  $\|\hat{\Gamma}\|_2 = 0$ ; otherwise,  $\mu_{1n} = 1.2\|\hat{\Gamma}\|_2 m$  with  $m = \inf_{v \in \mathbb{R}^p} \|\hat{\Sigma}v - (0, \hat{\Gamma}^\top / \|\hat{\Gamma}\|_2)^\top\|_\infty$ . We select  $\mu_{2n}$  and  $\mu_{3n}$  in a similar way. Note that when  $p \approx n$ , the aforementioned selection

---

<sup>2</sup>This means it selects the largest value of  $\lambda$  such that the error is within one standard error of the minimum cross validation error.

criterion would lead to very small tuning parameters  $\mu_{kn}$ , since the sample covariance matrix  $\hat{\Sigma}$  is (nearly) invertible. When  $0.5n < p < 1.5n$  (which holds in all simulation settings in this paper), we first pick a random sub-sample of size  $n/2$  and follow the selection procedure for the setting with  $p < 0.5n$  to construct an initial  $\mu_{kn}^*$ . Then we scale it by a factor  $1/\sqrt{2}$ , i.e., use  $\mu_{kn} = \mu_{kn}^*/\sqrt{2}$  as the final selected tuning parameter.

The user-defined constant  $\tau_0$  in  $\tau_n$  specified as (25) plays a role in the size-power balance. We recommend  $\tau_0 = 1$  based on the simulation results. We will come back to the choice of  $\tau_0$  in the end of subsection 4.3 after we show the simulation results. For the choice of  $\eta$ , we suggest a simple and robust procedure: first, randomly generate  $K$  sets of  $\eta$ 's, e.g., let  $K = 5000$ . Each  $\eta$  is constructed by sampling subsets  $\mathcal{N}_0 \subset [n]$  defined in (24) where  $|\mathcal{N}_0| = \lceil n/2 \rceil$ . Then we pick the vector with the minimum  $\|W^\top \eta\|_\infty$  among the  $K$  generated  $\eta$ 's<sup>3</sup>, since the supremum norm is a factor bounding extra bias caused by the calibration term as specified by Assumption 5, and hence it is expected to be further controlled. Compared to a deterministic  $\eta$ , this method can also mitigate the effect of the specific ordering of the observations. For comparison, we also try other two deterministic  $\eta$ 's. Precisely, based on (24) we define  $\eta^{(1)}$  with  $\mathcal{N}_0^{(1)} = \{1, 2, 3, \dots, \lceil n/2 \rceil\}$  and  $\eta^{(2)}$  with  $\mathcal{N}_0^{(2)}$  being the collection of all positive odd numbers no greater than  $n$ . The tests based on these two  $\eta$  vectors are denoted as  $\mathcal{Q}_n^{(1)}(\alpha)$  and  $\mathcal{Q}_n^{(2)}(\alpha)$ , respectively. The simulation results show the random  $\eta$  has slightly better performance. We would also like to emphasize that some other choices of  $\eta$  and  $\tau_n$  satisfying Assumption B1 still lead to reasonably good results about the size and power of the Q test under our settings, albeit often not as good as the recommended choices in finite samples.

## 4.2 Simulation Setup

In the simulation study, we consider the data generating process specified by model (1). We investigate the finite sample performance using the following combinations of sample size and variable dimensions: (1) Large  $p_z$  and small  $p_x$  with  $(n, p_x, p_z) \in \{(200, 0, 150), (200, 0, 250), (300, 0, 250)\}$ ; (2) Small  $p_z$  and large  $p_x$  with  $(n, p_x, p_z) \in \{(200, 150, 10), (200, 250, 10), (300, 250, 10)\}$ ; and (3) Large  $p_z$  and large  $p_x$  with  $(n, p_x, p_z) \in \{(200, 150, 100), (300, 150, 100)\}$ . Throughout all settings, we generate  $W_i$  by a multivariate normal distribution with mean zero and covariances  $\Sigma_{jk} = 0.5^{|j-k|}$ , for  $j, k = 1, 2, \dots, p$ .

In terms of the coefficients, we set  $\beta = 1$  for all cases and  $\varphi^\top = (0.1, 0.2, \dots, 0.5, 0_{p_x-5}^\top)$  and  $\psi^\top = (0.3, 0.4, \dots, 0.7, 0_{p_x-5}^\top)$  for those settings with nonzero  $p_x$ . For IV relevancy, we consider the following three settings for  $\gamma$ : (1) Strictly sparse with relatively stronger instruments (labeled as ‘‘Strong’’ in the following tables). Specifically,  $\gamma^\top = (0.5\iota_{10}^\top, 0, 0, \dots, 0)$  where  $\iota_s$  is an  $s \times 1$

---

<sup>3</sup>In theory we can solve the problem  $\eta = \arg \min_{\eta} \|W^\top \eta\|_\infty$  s.t.  $|\eta_i| = 1$  for all  $i \in [n]$ . However, it is an integer programming problem with  $n$  binary variables with complexity of  $2^n$  in the worse case, which is extremely inefficient in practice.



all-one vector; (2) Strictly sparse with relatively weaker instruments (labeled as “Weak” in the following tables). Specifically,  $\gamma^\top = (0.2\iota_{10}^\top, 0, 0, \dots, 0)$ ; and (3) Approximately sparse (AS) with  $\gamma^\top = 0.5 \cdot (1, 0.8, 0.8^2, \dots, 0.8^{p_z-1})$ . Setting (1) above is a standard sparse setting. In comparison, setting (2) has the same sparsity level but with much weaker instruments, through which we can observe the robustness of different tests to relatively weak instruments. Setting (3) is to show that the Q test also works for approximately sparse coefficients, even though the theory is constructed based on exact sparsity. For scenarios with homoskedastic errors, we set  $(e_i, \varepsilon_{i2})^\top \sim N \left[ 0, \begin{pmatrix} 1.5 & 0.75 \\ 0.75 & 1.5 \end{pmatrix} \right]$ . For the heteroskedastic errors case, we set

$$e_i = a_0 U_i + \sqrt{1 - a_0^2} V_{i1}, \quad \varepsilon_{i2} = 0.5 e_i + \sqrt{1 - 0.5^2} V_{i2}$$

with  $U_i \sim N(0, z_{i1}^2)$  and  $V_{i1}, V_{i2}$  are independent standard normal random variables. We set  $a_0 = 2^{-1/4}$  so that the R-square<sup>4</sup> between  $e_i^2$  and the instruments equals 0.2.

### 4.3 Simulation Results

**Empirical Size.** We set  $\pi = 0$  in all settings to check the empirical type I error of different tests. Tables 1 and 2 show the type I error of different methods with homoskedastic and heteroskedastic errors under 5% nominal level, respectively. When  $p_x = 0$  and  $p < n$ , other existing methods show correct size. Particularly, the MCD test can control the type I error under heteroskedasticity, though it is only intended for homoskedastic errors theoretically. When  $p_x = 0$  and  $p > n$ , Carrasco and Doukali (2021) can control the size, albeit it is much more conservative than the nominal size. When  $p_x$  is large, the jackknife and regularized jackknife tests show severely distorted size under finite samples. Note that the asymptotic size of the jackknife tests is based on a fixed number of endogenous variables and covariates (included instruments). The simulation results show that their tests are not compatible with high-dimensional covariates.

In comparison, the Q test is robust to high-dimensional covariates and (or) instruments. The recommended choice  $\tau_0 = 1$  specified in (25) results in type I errors near the nominal size in all settings we consider. The Q test has a slightly upward bias in type I error in some settings (such as the approximate sparse case with  $n = 200$  and  $p_z = 250$ ), but the overall size is well controlled around the nominal size even if both covariates and instruments are of high dimensions. The results are robust with different choices of  $\eta$ .

**Power Comparison.** We set  $\pi^\top = (0_{s_1-s_\pi}^\top, \rho_\pi \iota_{s_\pi}^\top, 0, 0, \dots, 0)$  where  $0_s$  is an  $s \times 1$  zero vector. We consider two types of signals in  $\pi$ : (i) sparse signals with  $s_1 = 10$ ,  $s_\pi = 2$  (so that the ninth and tenth IVs are invalid) and  $\rho_\pi$  varying from -1 to 1 with step size 0.2; (ii) dense signals with many nearly invalid instruments for cases with large  $p_z$  and small  $p_x$ , specified by

---

<sup>4</sup>The R-square is  $R_{e^2|Z_1}^2 = \text{Var}[\mathbb{E}(e^2|Z_1)] / \{\text{Var}[\mathbb{E}(e^2|Z_1)] + \mathbb{E}[\text{Var}(e^2|Z_1)]\}$  (Bekker and Crudu, 2015).

Table 1: Type I error with homoskedastic errors under 5% level over 1000 simulations.

$(n, p_x, p_z)$	$\gamma$	Jackknife	R-Jackknife	MCD	$\mathcal{Q}_n$	$\mathcal{Q}_n^{(1)}$	$\mathcal{Q}_n^{(2)}$
(200,0,150)	Strong	0.043	0.037	0.038	0.063	0.063	0.064
	Weak	0.034	0.032	0.026	0.071	0.062	0.065
	AS	0.038	0.037	0.032	0.061	0.06	0.049
(200,0,250)	Strong	NA	0.024	NA	0.054	0.054	0.051
	Weak	NA	0.029	NA	0.058	0.052	0.066
	AS	NA	0.02	NA	0.061	0.056	0.04
(300,0,250)	Strong	0.035	0.038	0.047	0.054	0.054	0.048
	Weak	0.038	0.038	0.034	0.06	0.043	0.065
	AS	0.044	0.046	0.049	0.054	0.044	0.055
(200,150,10)	Strong	0.001	0.877	0.022	0.043	0.049	0.047
	Weak	0	0.367	0.028	0.05	0.054	0.049
	AS	0	0.559	0.017	0.044	0.05	0.048
(200,250,10)	Strong	NA	NA	NA	0.05	0.048	0.053
	Weak	NA	NA	NA	0.041	0.063	0.045
	AS	NA	NA	NA	0.04	0.05	0.038
(300,250,10)	Strong	0	0.962	0.03	0.053	0.045	0.049
	Weak	0	0.098	0.029	0.042	0.056	0.05
	AS	0	0.242	0.027	0.065	0.049	0.057
(200,150,100)	Strong	NA	1	NA	0.07	0.059	0.061
	Weak	NA	1	NA	0.064	0.058	0.056
	AS	NA	1	NA	0.046	0.054	0.057
(300,150,100)	Strong	0.51	0.799	0.04	0.057	0.057	0.074
	Weak	0.475	0.98	0.025	0.051	0.052	0.054
	AS	0.491	0.992	0.05	0.069	0.076	0.059

Note: “Jackknife”, “R-Jackknife” and “MCD” represent the jackknife test by [Chao et al. \(2014\)](#), the regularized jackknife test by [Carrasco and Doukali \(2021\)](#) and the MCD test by [Kolesár \(2018\)](#), respectively.  $\mathcal{Q}_n$ ,  $\mathcal{Q}_n^{(1)}$  and  $\mathcal{Q}_n^{(2)}$  denote the Q test with different  $\eta$ ’s. “NA” means the result is not available since the test is infeasible in relevant settings.

$s_1 = s_\pi = \lfloor 3\sqrt{p_z} \rfloor$  (so that the first  $\lfloor 3\sqrt{p_z} \rfloor$  instruments are invalid) and  $\rho_\pi$  varying from -0.1 to 0.1 with step size 0.02. In the latter case, though the individual signals are weak, the global signal for a violation of the null is strong enough to affect the estimation and inference on  $\beta$ . Hence, the desirable test should be able to detect invalidity in such a scenario.

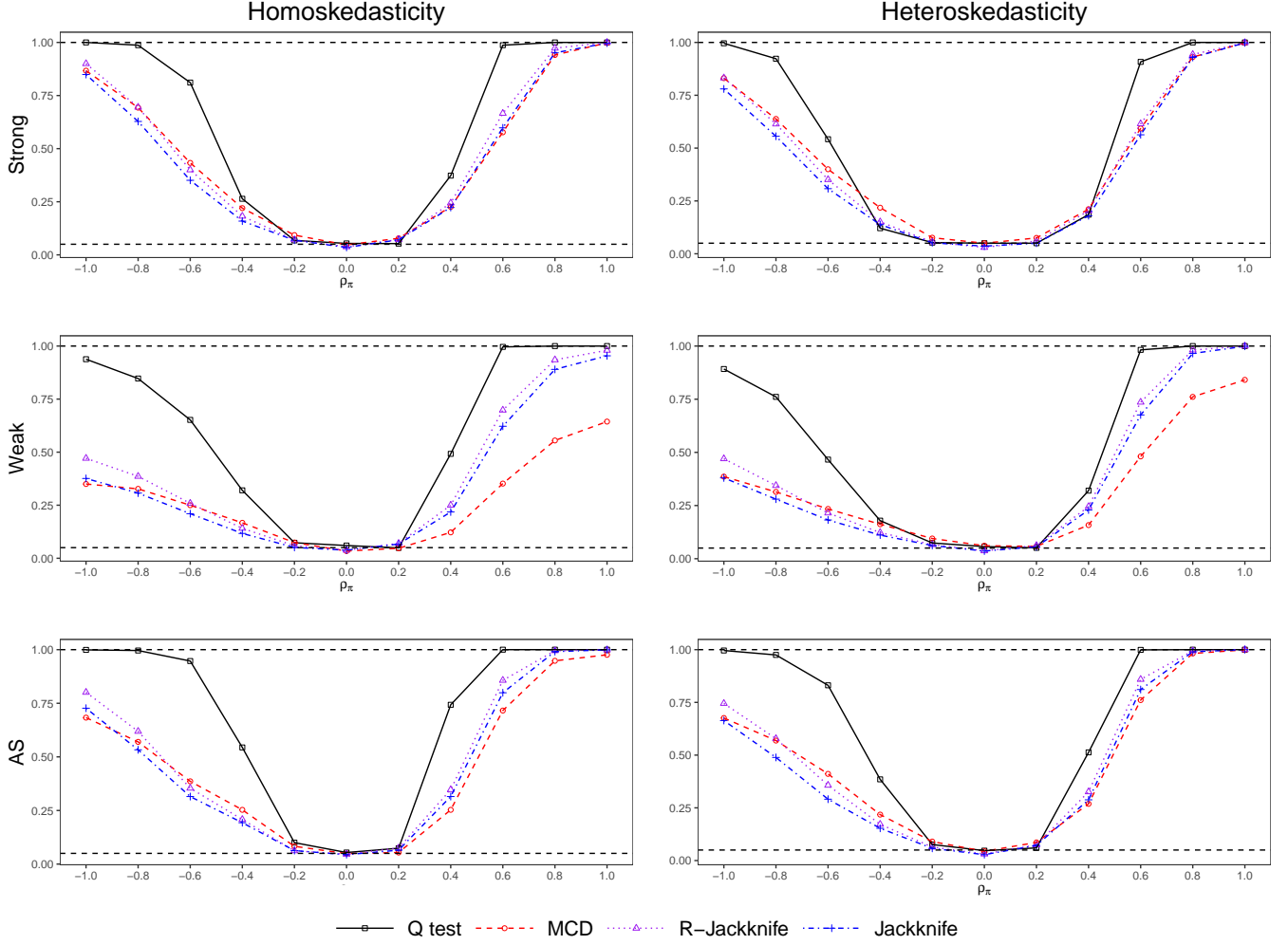
We report the power curves with  $n = 300$ . Figures 1 to 3 show the empirical power curves with a sparse  $\pi$ , while Figure 4 shows the empirical power curves with a dense  $\pi$  where a relatively large set of instruments are nearly valid. Simulation results with  $n = 200$  are similar and reported in Appendix C. Notice here we show the power curves of the Q test with the randomly generated  $\eta$ , and the power curves using the deterministic  $\eta$ ’s ( $\eta^{(1)}$  and  $\eta^{(2)}$  defined in 4.1) are quite similar.

Table 2: Type I error with heteroskedastic errors under 5% level over 1000 simulations.

$(n, p_x, p_z)$	$\gamma$	Jackknife	R-Jackknife	MCD	$\mathcal{Q}_n$	$\mathcal{Q}_n^{(1)}$	$\mathcal{Q}_n^{(2)}$
(200,0,150)	Strong	0.043	0.035	0.044	0.049	0.048	0.054
	Weak	0.053	0.054	0.055	0.055	0.05	0.06
	AS	0.023	0.025	0.027	0.041	0.062	0.049
(200,0,250)	Strong	NA	0.024	NA	0.041	0.048	0.059
	Weak	NA	0.022	NA	0.05	0.048	0.052
	AS	NA	0.019	NA	0.063	0.049	0.039
(300,0,250)	Strong	0.036	0.03	0.05	0.05	0.055	0.048
	Weak	0.036	0.037	0.061	0.057	0.059	0.051
	AS	0.028	0.031	0.044	0.047	0.054	0.052
(200,150,10)	Strong	0.001	0.878	0.042	0.053	0.052	0.037
	Weak	0	0.732	0.031	0.056	0.044	0.048
	AS	0	0.878	0.023	0.042	0.045	0.061
(200,250,10)	Strong	NA	NA	NA	0.04	0.044	0.036
	Weak	NA	NA	NA	0.038	0.04	0.041
	AS	NA	NA	NA	0.048	0.048	0.033
(300,250,10)	Strong	0.001	0.987	0.04	0.046	0.046	0.043
	Weak	0	0.365	0.031	0.047	0.053	0.052
	AS	0	0.644	0.034	0.033	0.053	0.049
(200,150,100)	Strong	NA	1	NA	0.043	0.057	0.046
	Weak	NA	1	NA	0.034	0.053	0.041
	AS	NA	1	NA	0.046	0.067	0.038
(300,150,100)	Strong	0.512	0.81	0.06	0.048	0.06	0.055
	Weak	0.507	0.999	0.049	0.05	0.05	0.045
	AS	0.505	0.996	0.04	0.055	0.049	0.053

Note: “Jackknife”, “R-Jackknife” and “MCD” represent the jackknife test by [Chao et al. \(2014\)](#), the regularized jackknife test by [Carrasco and Doukali \(2021\)](#) and MCD test by [Kolesár \(2018\)](#), respectively.  $\mathcal{Q}_n$ ,  $\mathcal{Q}_n^{(1)}$  and  $\mathcal{Q}_n^{(2)}$  denote the Q test with different  $\eta$ 's. “NA” means the result is not available since the test is infeasible in relevant settings.

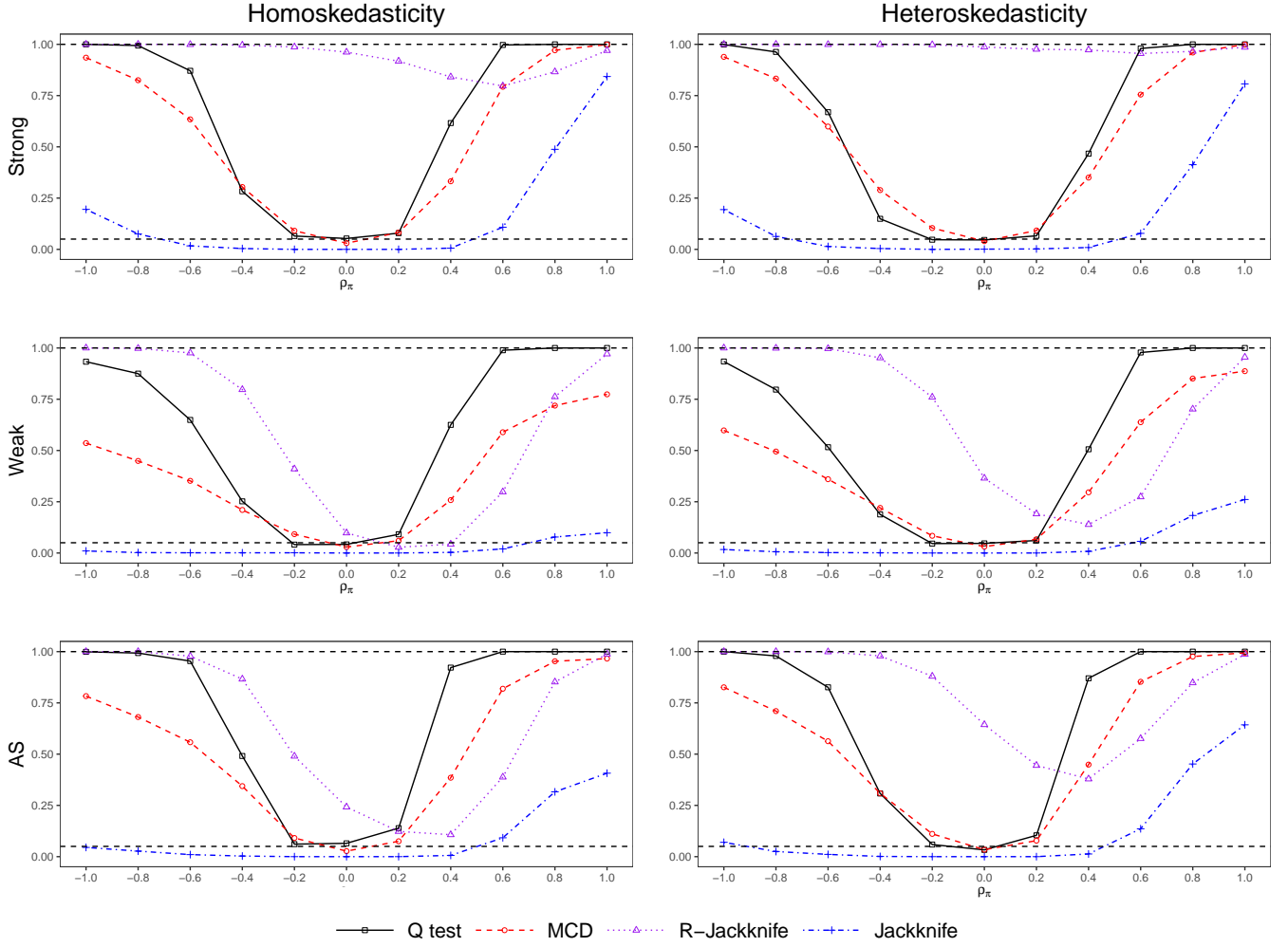
First of all, among all methods with a correct size, the Q test has the highest empirical power in all settings when  $|\rho_\pi|$  is away from zero (note that the jackknife tests have distorted size with a large number of covariates). As  $|\rho_\pi|$  grows, the Q test approaches power one much faster than other tests in all settings. Second, our test can handle different types of coefficients under (approximate) sparsity, and the power performance is more robust than other tests. The power Figures show that other tests are more sensitive to the strength of instruments than our test. Albeit all tests are negatively affected, the Q test still has the best size-power balanced performance when instruments are relatively weak (e.g., the middle subfigures of Figure 1). Besides, it shows robust performance for a dense  $\pi$  where a larger set of instruments are invalid with a much smaller individual magnitude. As shown in Figure 4, the power improvement of the Q test is more evident, especially when  $\rho_\pi < 0$ . Last but not least, in some settings where



**Figure 1.** Power of tests with  $(n, p_x, p_z) = (300, 0, 250)$  and sparse  $\pi$  under 5% level over 1000 simulations. “Jackknife”, “R-Jackknife” and “MCD” represent the jackknife test by [Chao et al. \(2014\)](#), the regularized jackknife test by [Carrasco and Doukali \(2021\)](#) and MCD test by [Kolesár \(2018\)](#), respectively. The nominal size 0.05 and power 1 are shown by the horizontal dashed lines. “Strong”, “Weak” and “AS” represent different settings of  $\gamma$ .

some or all of the other methods are infeasible (such as when  $p > n$ ), as shown in Figures C2, C4, C5 and C7 in the Appendix, the empirical power of the Q test is still comparable to that in  $p < n$  settings, indicating that the Q test can handle a wide range of variable dimensions.

In summary, the simulation results show strong evidence supporting the significant power improvement of the Q test with well controlled type I error under different empirically relevant settings. The Q test appears to be useful in applications with high-dimensional data with probably some relatively strong instruments and (approximate) sparsity. Section C of the Appendix presents the power curves of the Q test under 5% nominal level for  $\tau_0$  varying from 0 to 2 with step size 0.1 and for different settings of  $\rho_\pi$ 's. It shows that a smaller  $\tau_0$  results in higher empirical power but distorts the size. E.g., from the center subfigures ( $\rho_\pi = 0$ ) in Figures C8 to C18,

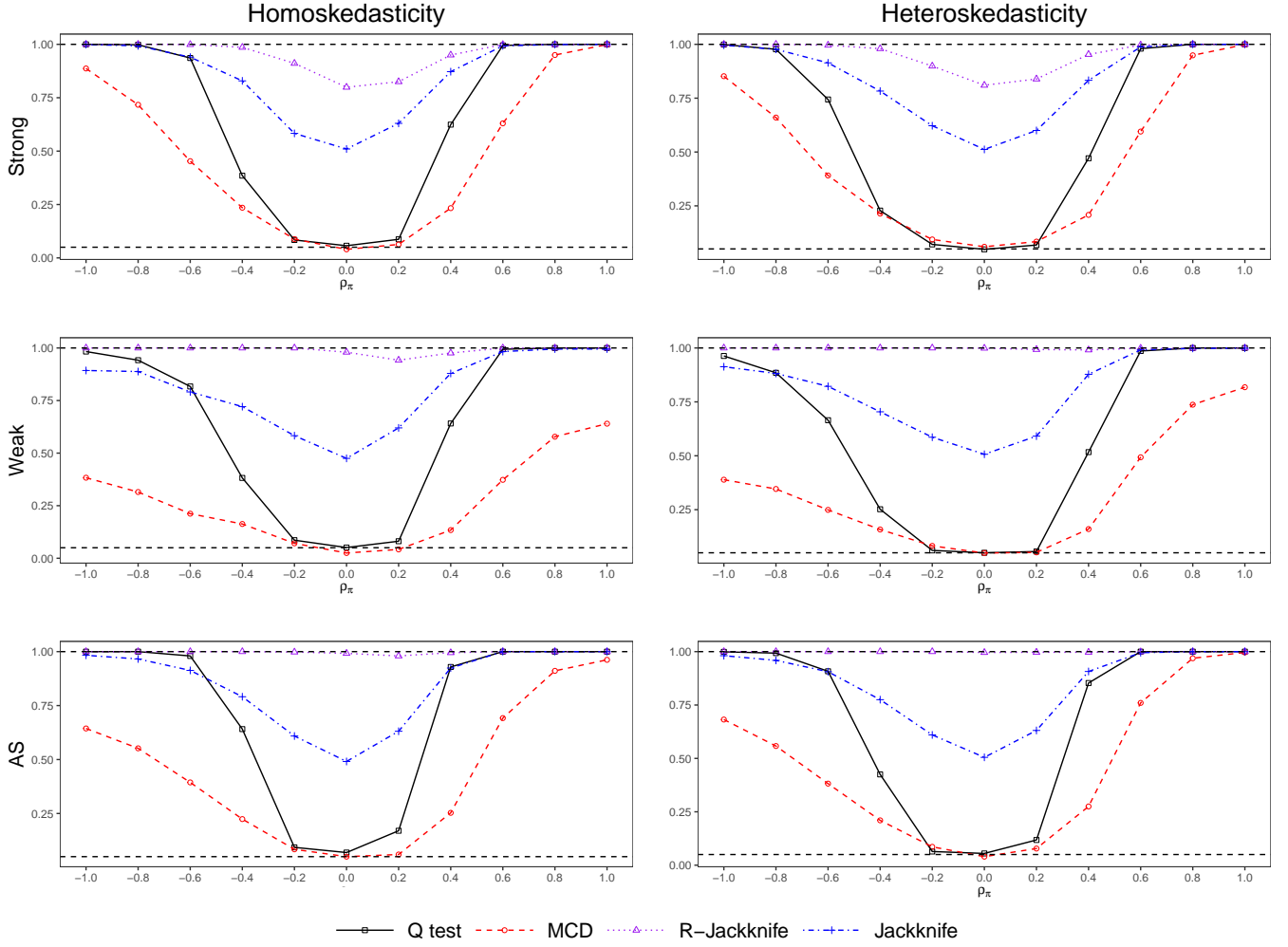


**Figure 2.** Power of tests with  $(n, p_x, p_z) = (300, 250, 10)$  and sparse  $\pi$  under 5% level over 1000 simulations. “Jackknife”, “R-Jackknife” and “MCD” represent the jackknife test by [Chao et al. \(2014\)](#), the regularized jackknife test by [Carrasco and Doukali \(2021\)](#) and MCD test by [Kolesár \(2018\)](#), respectively. The nominal size 0.05 and power 1 are shown by the horizontal dashed lines. “Strong”, “Weak” and “AS” represent different settings of  $\gamma$ .

we can see that empirical type I errors of the Q test can be controlled around the nominal size with any value of  $\tau_0$  above 0.5; and the local power can be further improved at the cost of a bit larger bias in the type I error by choosing an even smaller value for  $\tau_0$ .

#### 4.4 Empirical Study

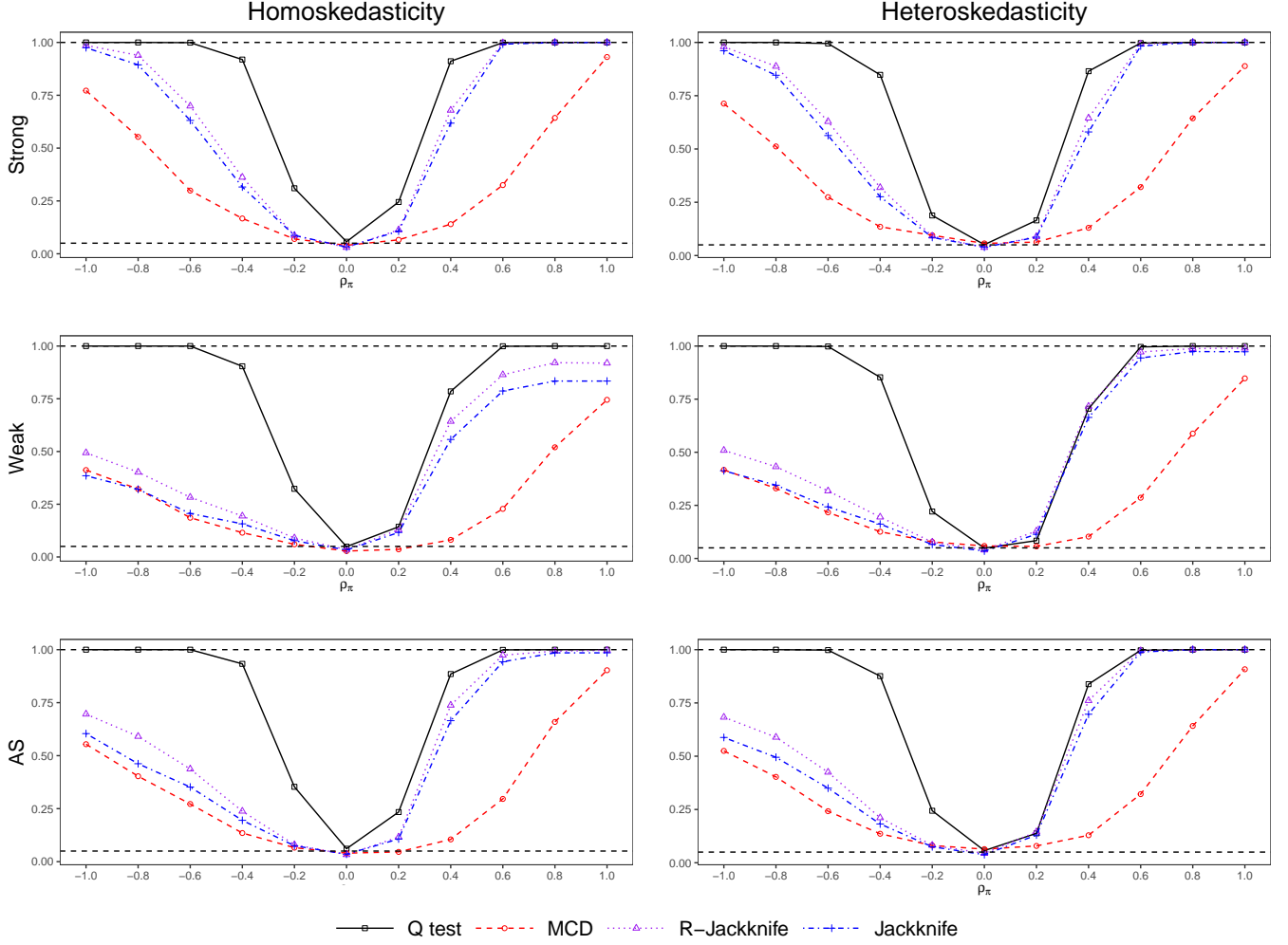
To illustrate the usefulness of the proposed test with high-dimensional data, we revisit the empirical analysis of the effect of trade on economic growth ([Frankel and Romer, 1999](#), FR99 hereafter). [Fan and Zhong \(2018\)](#) searched for more valid instruments (all geographical variables) following the celebrated gravity theory of trade. We update all variables to 2018 and expand



**Figure 3.** Power of tests with  $(n, p_x, p_z) = (300, 150, 100)$  and sparse  $\pi$  under 5% level over 1000 simulations. “Jackknife”, “R-Jackknife” and “MCD” represent the jackknife test by [Chao et al. \(2014\)](#), the regularized jackknife test by [Carrasco and Doukali \(2021\)](#) and MCD test by [Kolesár \(2018\)](#), respectively. The nominal size 0.05 and power 1 are shown by the horizontal dashed lines. “Strong”, “Weak” and “AS” represent different settings of  $\gamma$ .

the set of IVs from [Fan and Zhong \(2018\)](#) to include potentially invalid IVs from World Bank economic data. Following the literature, the outcome  $Y$  is the logarithm of GDP, covariates  $X$  include the logarithms of population and land area describing the size of the countries and hence  $p_x = 2$ . There are  $n = 159$  countries and  $p_z = 56$  instruments, including the constructed trade  $\hat{T}$  proposed by FR99 and other candidate IVs about some geographical characteristics, energy, environment and natural resources, and business activity variables. The outcome, endogenous variable, original FR99 covariates, and a subset of baseline instruments in [Fan and Zhong \(2018\)](#) are summarized in Table 3. This subset of instruments is used in the second part of the empirical study.

We standardize the data and compare the results of different testing methods. Table 4 shows



**Figure 4.** Power of tests with  $(n, p_x, p_z) = (300, 0, 250)$  and dense  $\pi$  under 5% level over 1000 simulations. “Jackknife”, “R-Jackknife” and “MCD” represent the jackknife test by [Chao et al. \(2014\)](#), the regularized jackknife test by [Carrasco and Doukali \(2021\)](#) and MCD test by [Kolesár \(2018\)](#), respectively. The nominal size 0.05 and power 1 are shown by the horizontal dashed lines. “Strong”, “Weak” and “AS” represent different settings of  $\gamma$ .

the p-values of different tests performed on the real data. We first test the correct specification of all instruments and expect the null to be rejected, since at least some instruments are likely to have a direct effect on economic growth. Those instruments include: (i) resources, such as freshwater; (ii) energy, such as electric power consumption and energy usage; (iii) environment, such as  $\text{CO}_2$ ,  $\text{SO}_2$ ,  $\text{NO}_2$ , and  $\text{PM}_{2.5}$  emissions; and (iv) business activities, such as ease of doing business index. We can see that only the Q test can strongly reject the null hypothesis and detect invalid instruments while other tests cannot, which is consistent with the simulation result that the Q test has the highest empirical power in relatively high dimensions.

Empirical researchers can also use our method to test whether a subset  $G$  of IVs is valid. The choice of  $G$  is often guided by earlier literature, or a subset of relatively strong IVs that



Table 3: Descriptive statistics.

Notation	Variable Name	Min	Median	Max	Mean	Std Dev.
$Y$	Log GDP	7.463	12.03	10.44	10.18	1.102
$D$	Trade	0.1981	4.129	0.7583	0.8694	0.5203
$X_1$	Ln Population	-3.037	6.674	1.472	1.355	1.83
$X_2$	Ln Area	5.193	16.61	11.96	11.68	2.312
$Z_1$	$\hat{T}$	0.01525	0.2968	0.07883	0.0922	0.05203
$Z_2$	Languages	1	16	1	1.887	2.129
$Z_3$	Area (Water)	0.0168	1.48	0.1126	0.1697	0.1988
$Z_4$	Land Boundaries	0	891200	2340	25220	100500
$Z_5$	% Forest	0	87560	201.3	1873	8160
$Z_6$	$Z_1 \cdot Z_2$	0	22150	1881	2820	3404
$Z_7$	$Z_1 \cdot Z_3$	0	2232	184.9	242.2	287.3
$Z_8$	$Z_1 \cdot Z_4$	0	98.26	30.32	29.71	22.42
$Z_9$	$Z_1 \cdot Z_5$	0	20.57	1.946	2.686	3.025

Note: Water area, and land boundaries are measured in square kilometers and kilometers, respectively. Data source: the World Bank, CIA World Factbook, R package `naiverreg`.

Table 4: P-values of different tests.

Instrument Sets	Jackknife	R-Jackknife	MCD	$\mathcal{Q}_n$	$\mathcal{Q}_n^{(1)}$	$\mathcal{Q}_n^{(2)}$
All 56 Instruments	0.256	0.263	0.109	< 0.001	< 0.001	< 0.001
$\{Z_1, Z_2, \dots, Z_9\}$	1	0.255	0.276	0.372	0.115	0.376

Note: “Jackknife”, “R-Jackknife” and “MCD” represent the jackknife test by [Chao et al. \(2014\)](#), the regularized jackknife test by [Carrasco and Doukali \(2021\)](#) and MCD test by [Kolesár \(2018\)](#), respectively.

the researcher would want to use. Here we pick a subset of IVs used in [Fan and Zhong \(2018\)](#) as displayed in Table 3, and treat other instruments as covariates without loss of generality. Therefore the variable dimensions are now  $p_x = 49$  and  $p_z = 9$ . It shows that all tests do not reject the null, meaning there is no evidence for this subset of instruments being invalid.

The takeaway from this empirical exercise is that practitioners should be cautious in the interpretation of a failure in rejecting the null by existing overidentification tests when many covariates and (or) instruments are present. Using the tests with low power would result in further troubles in the estimation and inference of the endogenous treatment effect. Our proposed test improves the power in different model settings, and hence it is recommended in a data-rich environment to detect invalid instruments.

## 5 Conclusion

In this paper, we develop a new test of overidentifying restrictions that is robust to high-dimensional instruments and/or covariates and heteroskedasticity. It is based on the inference of a quadratic form of the high-dimensional parameters. We show that the test improves the power while having the size under control, even when  $p > n$ . The higher power stems from the utilization of the sparse model structure. The procedure can be used to test any overidentifying restrictions of a subset. Moreover, it is robust to relatively weak individual IV signal as long as the global signal is strong enough, which is common under high-dimensional settings. As high-dimensional data becomes more common in observational studies, the Q test should see many applications in detecting instrument misspecification.

From a technical perspective, this paper extends the inference of quadratic forms by disengaging the sample splitting procedure and allowing for heteroskedasticity. The test procedure overcomes the downward size distortion caused by the super-efficiency issue without introducing extra randomness.

## Bibliography

- Anatolyev, S. and N. Gospodinov (2011). Specification testing in models with many instruments. *Econometric Theory* 27(2), 427–441.
- Andrews, I., J. Stock, and L. Sun (2019). Weak instruments in IV regression: Theory and practice. *Annual Review of Economics* 11, 727–753.
- Angrist, J. D., G. W. Imbens, and A. B. Krueger (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics* 14(1), 57–67.
- Bai, J. and S. Ng (2010). Instrumental variable estimation in a data rich environment. *Econometric Theory* 26, 1577–1606.
- Bekker, P. A. and F. Cruadu (2015). Jackknife instrumental variable estimation with heteroskedasticity. *Journal of Econometrics* 185(2), 332–342.
- Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* 80, 2369–2429.
- Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies* 81(2), 608–650.
- Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics* 37(4), 1705–1732.

- Breunig, C., E. Mammen, and A. Simoni (2020). Ill-posed estimation in high-dimensional models with instrumental variables. *Journal of Econometrics* 219(1), 171–200.
- Bühlmann, P. and S. van de Geer (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.
- Cai, T. and Z. Guo (2020). Semisupervised inference for explained variance in high dimensional linear regression and its applications. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82(2), 391–419.
- Cai, T., W. Liu, and X. Luo (2011). A constrained  $\ell_1$  minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* 106(494), 594–607.
- Cai, T. T. and Z. Guo (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics* 45(2), 615–646.
- Caner, M. and Q. Fan (2015). Hybrid generalized empirical likelihood estimators: Instrument selection with adaptive lasso. *Journal of Econometrics* 187(1), 256–274.
- Caner, M., X. Han, and Y. Lee (2018). Adaptive elastic net GMM estimation with many invalid moment conditions: Simultaneous model and moment selection. *Journal of Business & Economic Statistics* 36(1), 24–46.
- Carrasco, M. (2012). A regularization approach to the many instruments problem. *Journal of Econometrics* 170, 383–398.
- Carrasco, M. and M. Doukali (2017). Efficient estimation using regularized jackknife IV estimator. *Annals of Economics and Statistics* (128), 109–149.
- Carrasco, M. and M. Doukali (2021). Testing overidentifying restrictions with many instruments and heteroskedasticity using regularized jackknife IV. *The Econometrics Journal*, 1–27.
- Chang, J., Z. Shi, and J. Zhang (2021). Culling the herd of moments with penalized empirical likelihood. *arXiv preprint arXiv:2108.03382*.
- Chao, J. C., J. A. Hausman, W. K. Newey, N. R. Swanson, and T. Woutersen (2014). Testing overidentifying restrictions with many instruments and heteroskedasticity. *Journal of Econometrics* 178, 15–21.
- Cheng, X. and Z. Liao (2015). Select the valid and relevant moments: An information-based lasso for GMM with many moments. *Journal of Econometrics* 186(2), 443–464.
- Cragg, J. G. and S. G. Donald (1993). Testing identifiability and specification in instrumental variable models. *Econometric Theory* 9(2), 222–240.

- Donald, S. and W. Newey (2001). Choosing the number of instruments. *Econometrica* 69, 1161–1191.
- Fan, J., H. Weng, and Y. Zhou (2021). Optimal estimation of functionals of high-dimensional mean and covariance matrix. *arXiv preprint arXiv:1908.07460*.
- Fan, Q. and Y. Wu (2020). Endogenous treatment effect estimation with some invalid and irrelevant instruments. *arXiv preprint arXiv:2006.14998*.
- Fan, Q. and W. Zhong (2018). Nonparametric additive instrumental variable estimator: A group shrinkage estimation perspective. *Journal of Business & Economic Statistics* 36(3), 388–399.
- Frankel, J. A. and D. H. Romer (1999). Does trade cause growth? *American Economic Review* 89(3), 379–399.
- Fuller, W. A. (1977). Some properties of a modification of the limited information estimator. *Econometrica*, 939–953.
- Gold, D., J. Lederer, and J. Tao (2020). Inference for high-dimensional instrumental variables regression. *Journal of Econometrics* 217(1), 79–111.
- Guo, Z. (2021). Post-selection problems for causal inference with invalid instruments: A solution using searching and sampling. *arXiv e-prints*, arXiv–2104.
- Guo, Z., H. Kang, T. T. Cai, and D. S. Small (2018). Testing endogeneity with high dimensional covariates. *Journal of Econometrics* 207(1), 175–187.
- Guo, Z., H. Kang, T. Tony Cai, and D. S. Small (2018). Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80(4), 793–815.
- Guo, Z., C. Renaux, P. Bühlmann, and T. Cai (2021). Group inference in high dimensions with applications to hierarchical testing. *Electronic Journal of Statistics* 15(2), 6633–6676.
- Guo, Z., W. Wang, T. T. Cai, and H. Li (2019). Optimal estimation of genetic relatedness in high-dimensional linear models. *Journal of the American Statistical Association* 114(525), 358–369.
- Hahn, J. and J. Hausman (2002). A new specification test for the validity of instrumental variables. *Econometrica* 70(1), 163–189.
- Hall, P. and C. C. Heyde (1980). *Martingale limit theory and its application*. Academic Press.
- Hansen, C. and D. Kozbur (2014). Instrumental variables estimation with many weak instruments using regularized jive. *Journal of Econometrics* 182(2), 290–308.

- Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 1029–1054.
- Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning* (2<sup>nd</sup> ed.). Springer.
- Hausman, J. A., W. K. Newey, T. Woutersen, J. C. Chao, and N. R. Swanson (2012). Instrumental variable estimation with heteroskedasticity and many instruments. *Quantitative Economics* 3(2), 211–255.
- Hsu, Y.-C. and X. Shi (2017). Model-selection tests for conditional moment restriction models. *The Econometrics Journal* 20(1), 52–85.
- Javanmard, A. and A. Montanari (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research* 15(1), 2869–2909.
- Kang, H., A. Zhang, T. T. Cai, and D. S. Small (2016). Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association* 111(513), 132–144.
- Kolesár, M. (2018). Minimum distance approach to inference with many instruments. *Journal of Econometrics* 204(1), 86–100.
- Kolesár, M., R. Chetty, J. Friedman, E. Glaeser, and G. W. Imbens (2015). Identification and inference with many invalid instruments. *Journal of Business & Economic Statistics* 33(4), 474–484.
- Lee, Y. and R. Okui (2012). Hahn–hausman test as a specification test. *Journal of Econometrics* 167(1), 133–139.
- Liao, Z. (2013). Adaptive GMM shrinkage estimation with consistent moment selection. *Econometric Theory* 29(5), 857–904.
- Liu, T. and L.-f. Lee (2019). A likelihood ratio test for spatial model selection. *Journal of Econometrics* 213(2), 434–458.
- Magnus, J. R. and H. Neudecker (1980). The elimination matrix: some lemmas and applications. *SIAM Journal on Algebraic Discrete Methods* 1(4), 422–449.
- Mikusheva, A. and L. Sun (2020). Inference with many weak instruments. *arXiv preprint arXiv:2004.12445*.
- Okui, R. (2011). Instrumental variable estimation in the presence of many moment conditions. *Journal of Econometrics* 165, 70–86.

- Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica*, 393–415.
- Shi, X. (2015). Model selection tests for moment inequality models. *Journal of Econometrics* 187(1), 1–17.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58(1), 267–288.
- van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics* 42(3), 1166–1202.
- Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.
- Windmeijer, F., H. Farbmacher, N. Davies, and G. Davey Smith (2019). On the use of the lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association* 114(527), 1339–1350.
- Windmeijer, F., X. Liang, F. P. Hartwig, and J. Bowden (2021). The confidence interval method for selecting valid instrumental variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 83(4), 752–776.
- Zhang, C.-H. and S. Zhang (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 76(1), 217–242.
- Zhu, Y. (2018). Sparse linear models and l1-regularized 2sls with high-dimensional endogenous regressors and instruments. *Journal of Econometrics* 202(2), 196–213.

# Appendices to “Testing Overidentifying Restrictions with High-Dimensional Data and Heteroskedasticity”

QINGLIANG FAN<sup>†</sup>, ZIJIAN GUO<sup>‡</sup>, ZIWEI MEI<sup>†</sup>

<sup>†</sup>Department of Economics, The Chinese University of Hong Kong

<sup>‡</sup>Department of Statistics, Rutgers University

The Appendices mainly include the following parts: Section A provides technical details of the power comparison discussions in Section 3 of the main paper. Section B contains all technical proofs. Section C collects the omitted robustness simulation results from the main text.

## A Power Analysis of Alternative Tests

For simplicity, we illustrate the power comparison in the case with no covariates  $X$  ( $p_x = 0$ ) and  $p_z < n$  unless there is further clarification. Under this setting, all other methods mentioned in Section 4 are feasible. Hence, the model we consider becomes

$$\begin{aligned} Y &= D\beta + Z\pi + e, \\ D &= Z\gamma + \varepsilon_2 \end{aligned} \tag{A1}$$

with a reduced from

$$\begin{aligned} Y &= Z\Gamma + \varepsilon_1. \\ D &= Z\gamma + \varepsilon_2 \end{aligned} \tag{A2}$$

where  $\Gamma = \gamma\beta + \pi$  and  $\varepsilon_1 = \varepsilon_2\beta + e$ .

### A.1 Modified Cragg-Donald Test

The modified Cragg-Donald (MCD) test in Kolesár (2018) is based on the limited information maximum likelihood (LIML) estimator. Let  $I_n$  denote the  $n \times n$  identity matrix. Define  $P = Z(Z^\top Z)^{-1}Z^\top$ ,  $M = I_n - P$  and  $\tilde{Y} = (Y, D)$ . Let  $m_{\min}$  be the minimum eigenvalue of  $S^{-1}T$ , where

$$S = \frac{1}{n - p_z} \tilde{Y}^\top M \tilde{Y}, \quad T = \frac{1}{n} \tilde{Y}^\top P \tilde{Y}. \tag{A3}$$

as defined by (5) and (6) in Kolesár (2018). From the discussions in the paragraphs right before and after Proposition 4 of Kolesár (2018), we know that as  $p_z \rightarrow \infty$  the MCD test is equivalent to a test rejecting whenever  $(n/\sqrt{p_z})(m_{\min} - p_z/n)/\sqrt{\frac{2}{1-\alpha_z} + \delta_0\kappa_0}$  is greater than the  $1 - \alpha$  quantile of  $N(0, 1)$  under significance level  $\alpha$ , where  $\delta_0$  is some constant and  $\kappa_0$  measures excess kurtosis



of  $e_i$ , as defined in Assumption RC and (22) of Kolesár (2018) respectively. For simplicity, we further assume Gaussianity of the error terms and hence  $\kappa_0 = 0$ . Therefore,  $\sqrt{\frac{2}{1-\alpha_z} + \delta_0 \kappa_0}$  is bounded away from zero.

Define  $\widehat{\Sigma} = Z^\top Z/n$ ,  $Q_{\widehat{\Sigma}} = \pi^\top \widehat{\Sigma} \pi (1 - R_{\widehat{\Sigma}}(\pi, \gamma))$  with  $R_{\widehat{\Sigma}}(\pi, \gamma) = \frac{\pi^\top \widehat{\Sigma} \gamma}{\sqrt{(\pi^\top \widehat{\Sigma} \pi)(\gamma^\top \widehat{\Sigma} \gamma)}}$ . Note

that when  $\widehat{\Sigma}$  has eigenvalues bounded away from zero and above, we have  $Q_{\widehat{\Sigma}} \asymp Q$  with  $Q$  defined as (5). We first specify some reasonable assumptions that coordinate the MCD test and simplify the analysis.

**Assumption A1.** *Suppose the following condition hold*

(i)  *$Z$  is nonrandom and  $\widehat{\Sigma} = Z^\top Z/n$  has eigenvalues bounded away from zero and above. Suppose that  $Q_\gamma = \gamma^\top \frac{Z^\top Z}{p_z} \gamma$  is a constant.*

(ii)  *$(\varepsilon_{i1}, \varepsilon_{i2})^\top \sim \text{i.i.d. } N(0, \Theta)$  with  $\Theta = \begin{pmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{12} & \Theta_{22} \end{pmatrix}$  being positive definite.*

(iii)  *$|R_{\widehat{\Sigma}}(\pi, \gamma)| \leq r_0 < 1$  for some positive  $r_0$ .*

The first part of Assumption A1 (i) is an analog of Assumption 1 with nonrandom  $Z$  and  $p < n$ . When  $p_z/n \rightarrow \alpha \in (0, 1)$ , the second part indicates the global strength of the instruments through  $\gamma^\top \widehat{\Sigma} \gamma \asymp \|\gamma\|_2^2$ . (ii) specifies the jointly normal distribution of the structural error terms. (iii) is the overidentification condition requiring no perfect correlation between  $\pi$  and  $\gamma$  as discussed for Theorem 1. We want to show that the MCD test has asymptotic power one only if  $(n/\sqrt{p_z})Q_{\widehat{\Sigma}} \rightarrow \infty$  and hence  $(n/\sqrt{p_z})Q \rightarrow \infty$ . Under Assumption A1, it suffices to show the following proposition.

**Proposition A1.** *If Assumptions A1 holds and  $Q_{\widehat{\Sigma}} = \sqrt{p_z} \Delta/n$ , then for any fixed  $\Delta > 0$  and any constant  $z > 0$ ,  $\limsup_{n \rightarrow \infty} \Pr\{(n/\sqrt{p_z})(m_{\min} - p_z/n) > z\} < 1$ .*

Therefore, under the simple case specified by Assumption A1 with  $p_z/n \rightarrow \alpha_z \in (0, 1)$ , MCD has asymptotic power one only if  $\sqrt{n}Q \rightarrow \infty$ . In comparison, our test has asymptotic power one when  $\sqrt{n}Q$  is any fixed positive constant away from zero under sparse model setting allowing  $p > n$ . Proposition A1 also shows that MCD can have higher asymptotic power if  $p_z \rightarrow \infty$  while  $p_z = o(n)$ . However, note that in the case with many covariates where  $\alpha_x = p_x/n + o(n^{-1/2})$ , the asymptotic variance of  $(n/\sqrt{p_z})(m_{\min} - p_z/n)$  becomes  $\frac{2\alpha_x}{1-\alpha_z-\alpha_x} + \delta_0 \kappa_0$  with  $\delta_0 \kappa_0 = 0$  under Gaussianity. When there are many covariates so that  $p_x + p_z \approx n$  or equivalently,  $\alpha_z + \alpha_x \approx 1$ , the finite sample power of MCD is further damaged with a large variance estimate even if  $p_z = o(n)$ . Proofs of the propositions in this section are available in B.5.

## A.2 Jackknife Specification Tests

We only discuss the jackknife test by [Chao et al. \(2014\)](#) for the case of  $p_z < n$ . The result for regularized jackknife test by [Carrasco and Doukali \(2021\)](#) is similar to  $P$  replaced by  $P_{\alpha_0} = Z(Z^\top Z + \alpha_0 I)^{-1}Z$  for some regularization parameter  $\alpha_0$ .

Define  $P^{\text{diag}} = \text{diag}(P_{11}, P_{22}, \dots, P_{nn})$ . For simplicity, the analysis in this part is based on the jackknife instrumental variable estimator (JIVE) given by

$$\hat{\beta}_{\text{JIVE}} = \frac{D(P - P^{\text{diag}})Y}{D(P - P^{\text{diag}})D} \quad (\text{A4})$$

proposed by [Angrist et al. \(1999\)](#). [Chao et al. \(2014\)](#) focus on HFUL<sup>1</sup> estimator by [Hausman et al. \(2012\)](#) for higher efficiency, which brings much complexity to power analysis. As [Chao et al. \(2014\)](#) point out, JIVE is also feasible for the jackknife specification test, and the asymptotic power is believed to be similar to relatively strong instruments. Define

$$\begin{aligned} \hat{T} &= \frac{\hat{v}^\top (P - P^{\text{diag}})\hat{v}}{\sqrt{\hat{S}}} + p_z \\ \hat{S} &= \frac{\sum_{i \neq j} \hat{v}_i^2 \hat{v}_j^2 P_{ij}^2}{p_z} \end{aligned} \quad (\text{A5})$$

with residual  $\hat{v}_i = Y_i - D\hat{\beta}_i$ . The jackknife IV test rejects the null whenever  $\hat{T}$  is greater than the  $1 - \alpha$  quantile of  $\chi_{p_z-1}^2$  under significance level  $\alpha$  (denoted by  $\chi_{p_z-1}^2(1 - \alpha)$ ). Since  $\frac{\chi_{p_z-1}^2(1 - \alpha) - p_z}{\sqrt{2p_z}} \rightarrow z_\alpha$  where  $z_\alpha$  is the  $1 - \alpha$  quantile of  $N(0, 1)$ , the test is asymptotically

equivalent to rejecting the null whenever  $(\hat{T} - p_z)/\sqrt{2p_z} = \hat{v}^\top (P - P^{\text{diag}})\hat{v}/\sqrt{2p_z \hat{S}}$  is greater than  $z_\alpha$ . Besides Assumption [A1](#), we need the following assumption for the jackknife test.

**Assumption A2.** *Suppose the following conditions hold*

1. *There is a positive constant  $C_P < 1$  such that  $P_{ii} < C_P$  for all  $i \in [n]$ .*
2.  *$\max\{\pi^\top Z_i, Z_i^\top \pi, \gamma^\top Z_i, Z_i^\top \gamma\} = O(1)$  uniformly for all  $i \in [n]$ .*

Assumption [A2](#) (i) is required in Assumption 1 of [Chao et al. \(2014\)](#). (ii) is an analog of a bounded  $\mathbb{E}[D_i^2]$  under the null, which is required in Lemma A3 of [Chao et al. \(2014\)](#) to bound the asymptotic variance estimate  $\hat{S}$  away from zero. Under random instruments with bounded eigenvalues of its population covariance matrix, (ii) automatically holds w.p.a.1. We follow [Chao et al. \(2014\)](#) to consider nonrandom instruments and impose an upper bound of the quantities in (ii). After their Theorem 2, [Chao et al. \(2014\)](#) discuss the power of the jackknife test increase at

---

<sup>1</sup>HFUL is an abbreviation of a heteroskedasticity robust version of [Fuller \(1977\)](#) estimator.

rate  $n/\sqrt{p}$  under Assumption A2. Here we formally show that under the simple case we consider, the jackknife test has asymptotic power one only if  $\sqrt{n}Q_{\hat{\Sigma}} \rightarrow \infty$  when  $p_z/n \rightarrow \alpha_z \in (0, 1)$ . It suffices to show the following proposition.

**Proposition A2.** *If Assumptions A1 and A2 hold and  $Q_{\hat{\Sigma}} = \sqrt{p_z}\Delta/n$ , then for any fixed  $\Delta > 0$  and any constant  $z > 0$ ,  $\limsup_{n \rightarrow \infty} \Pr\{\hat{v}^\top (P - P^{\text{diag}})\hat{v}/\sqrt{2p_z\hat{S}} > z\} < 1$ .*

Proposition A2 implies the same necessary condition as MCD for asymptotic power one and shows the superiority in asymptotic power of our test. Proposition A2 also implies that the jackknife test can have higher power if  $p_z \rightarrow \infty$  while  $p_z = o(n)$ . However, as discussed in Section 3, the jackknife test cannot handle high-dimensional covariates, since it treats all exogenous covariates as both endogenous variables and (included) instruments, and it assumes a fixed number of endogenous variables.

## B Proofs

We first show the preliminary propositions about some auxiliary and intermediate estimators in B.1. Then B.2 contains the proofs of theorems and propositions in Section 3. B.3 and B.4 respectively collect the proofs of preliminary propositions and technical lemmas applied in this section. B.5 provides the proofs of propositions in Section A about the power of other tests.

As discussed in Remark 1 of the main text, though we recommend some specific choices of tuning  $\eta$  and  $\tau_n$  by (24) and (25), respectively, we construct the theory with general restrictions on  $\eta$  and  $\tau_n$ , and show that our recommended choices satisfy these restrictions. The following assumption imposes such restrictions.

**Assumption B1.** *Suppose the following conditions hold.*

- (i)  $\eta = (\eta_i)_{i=1}^n$  contain  $n$  independent random variables with sub-Gaussian norms bounded by  $K$  satisfying  $\|\mathbb{E}[W^\top \eta]\|_\infty \lesssim \sqrt{n \log p}$ . Besides,  $\eta \perp (\varepsilon_1, \varepsilon_2)|W$ .
- (ii) As  $Q = \Delta_n/\sqrt{n}$ , there exists some uniformly bounded nonrandom sequence  $\tau_n^*$  such that  $\tau_n/\tau_n^* \xrightarrow{p} 1$  and

$$\frac{m_\omega s^{1/2}}{n^{1/4}} \vee \frac{m_\omega^2 s \log(np)}{\sqrt{n}} \vee \frac{s_\omega m_\omega^{2-q} s^{3/2} \log p}{\sqrt{n}} \vee \frac{s_\omega m_\omega^{1-q} s^{(3-q)/2} (\log p)^{1-q/2}}{n^{(1-q)/2}} = o(\sqrt{(1 \vee \Delta_n) \tau_n^*}).$$

Additionally, suppose that  $\tau_n^* = o(1)$  as  $\Delta_n \rightarrow \Delta > 0$ .

The first part of (i) controls the additional bias and variance caused by the calibration term, while the second part ensures the validity of the central limit theorem. B1 (ii) specifies the rate

of  $\tau_n$  to guarantee a correct size. The way we specify the condition in (ii) allows us to use a data-adaptive  $\tau_n$  for size-power balance. When  $\Delta_n$  is small, we need a larger  $\tau_n^*$  dominating the bias term to fix the “super-efficiency” issue. As  $\Delta_n$  gets larger, we allow smaller  $\tau_n^*$  to achieve higher power under the alternative. Similar ideas have appeared in the literature (Hsu and Shi, 2017; Liu and Lee, 2019). Proposition B7 shows that  $\eta$  and  $\tau_n$  specified by (24) and (25) satisfy Assumption B1.

For simplicity of notations, we use  $\xi_Z$  to denote  $(0_{p_x}^\top, \xi^\top)^\top$  for any  $p \times 1$  vector  $\xi$  throughout Section B.

## B.1 Preliminary Propositions

In this subsection, we provide several propositions about the properties of sub-Gaussian variables, initial Lasso estimators, the estimators of quadratic functionals of  $\Gamma$  and  $\gamma$ ,  $\hat{\beta}_R$  and the test statistic. These propositions are proved in subsection B.3.

**Proposition B1.** *Suppose that Assumptions 1, 2 and 5 (i) hold. If  $s \log p/n = o(1)$ , we have w.p.a.1*

$$\begin{aligned} \max\{\|\hat{\Gamma} - \Gamma\|_2, \|\hat{\gamma} - \gamma\|_2, \|\hat{\Psi} - \Psi\|_2, \|\hat{\psi} - \psi\|_2\} &\lesssim \sqrt{\frac{s \log p}{n}}, \\ \max\{\|\hat{\Gamma} - \Gamma\|_1, \|\hat{\gamma} - \gamma\|_1, \|\hat{\Psi} - \Psi\|_1, \|\hat{\psi} - \psi\|_1\} &\lesssim s \sqrt{\frac{\log p}{n}}. \end{aligned} \quad (\text{B1})$$

The proof is available in B.3.1.

**Proposition B2.** *Suppose that the conditions in Proposition B1 are satisfied, and Assumption 5 (ii) holds. If  $\|\Omega\|_1 \leq m_\omega$  for  $m_\omega \geq 1$ , then w.p.a.1,*

$$|\widehat{\langle \Gamma, \gamma \rangle} - \langle \Gamma, \gamma \rangle| \lesssim m_\omega \left[ (\|\Gamma\|_2 + \|\gamma\|_2) \sqrt{\frac{s \log p}{n}} + \frac{s \log p}{n} \right], \quad (\text{B2})$$

$$|\widehat{\|\gamma\|_2^2} - \|\gamma\|_2^2| \lesssim m_\omega \left[ \|\gamma\|_2 \sqrt{\frac{s \log p}{n}} + \frac{s \log p}{n} \right]. \quad (\text{B3})$$

Furthermore, when  $\|\gamma\|_2 \gg m_\omega \sqrt{\frac{s \log p}{n}}$ , we have  $\frac{\widehat{\|\gamma\|_2^2}}{\|\gamma\|_2^2} \xrightarrow{p} 1$ .

The proof is available in B.3.2.

**Proposition B3.** *Suppose that the conditions in Proposition B2 are satisfied. If  $\|\Omega\|_1 \leq m_\omega$  and  $\|\gamma\|_2 \gg m_\omega \sqrt{s \log p/n}$ , then w.p.a.1,*

$$|\hat{\beta}_R - \beta_R| \lesssim m_\omega \left( 1 + \frac{\|\pi\|_2}{\|\gamma\|_2} \right) \frac{1}{\|\gamma\|_2} \sqrt{\frac{s \log p}{n}}. \quad (\text{B4})$$

The proof is available in [B.3.3](#).

**Proposition B4.** *Under conditions in Proposition [B3](#), we have w.p.a.1*

$$\begin{aligned} \max\{\|\hat{\pi} - \tilde{\pi}\|_2, \|\hat{\varphi} - \tilde{\varphi}\|_2\} &\lesssim \left(1 + \frac{\|\pi\|_2}{\|\gamma\|_2}\right) \sqrt{\frac{s \log p}{n}}, \\ \max\{\|\hat{\pi} - \tilde{\pi}\|_1, \|\hat{\varphi} - \tilde{\varphi}\|_1\} &\lesssim \left(1 + \frac{\|\pi\|_2}{\|\gamma\|_2}\right) s \sqrt{\frac{\log p}{n}}, \end{aligned} \tag{B5}$$

The proof is available in [B.3.4](#).

**Proposition B5.** *Under Assumptions [1-5](#) and [B1](#), we have*

$$\hat{Q} - Q = M + B, \tag{B6}$$

where  $\sqrt{n}M/\sqrt{V} \xrightarrow{d} N(0, 1)$  with  $V$  defined in Theorem [1](#), and

$$|B| = o_p \left( \sqrt{\frac{Q}{n}} + \sqrt{\frac{((1 + \beta_R^2) \vee \Delta_n) \tau_n^*}{n}} + Q \right)$$

as  $Q = \Delta_n/\sqrt{n}$ . Specially,  $|B| = o_p(\sqrt{V/n} + Q)$  as  $\Delta_n$  is bounded.

The proof is available in [B.3.5](#).

**Proposition B6.** *Under Assumptions [1-5](#) and [B1](#), we have*

$$\frac{\hat{V}}{V} - 1 = o_p \left( 1 \vee \frac{\Delta_n}{1 + \beta_R^2} + \frac{\Delta_n^2}{V} \right), \tag{B7}$$

as  $Q = \Delta_n/\sqrt{n}$ , where  $V$  is specified in [\(27\)](#) and  $\hat{V}$  is given by [\(22\)](#).

The proof is available in [B.3.6](#).

**Proposition B7.** (a) *Under Assumption [1](#), the  $\eta$  defined by [\(24\)](#) satisfies Assumption [B1](#) (i).*

(b) *Suppose Assumptions [1-5](#) hold. Let  $Q = \Delta_n/\sqrt{n}$  where  $\Delta_n \rightarrow \Delta \in [0, \infty]$ . Then  $\tau_n$  defined by [\(25\)](#) satisfies Assumption [B1](#) (ii) with*

$$\tau_n^* = \frac{\tau_0}{1 + \Delta_n \cdot \log[\log(np)]}. \tag{B8}$$

The proof is available in [B.3.7](#).

## B.2 Proofs of Theorems in Section 3

### B.2.1 Proof of Theorem 1

By Proposition B7 it suffices to show that Theorem 1 holds with  $\eta$  and  $\tau_n^*$  satisfying Assumption B1. From Proposition B5 we know that  $\sqrt{n}|B|/\sqrt{V^*} = o_p\left(\sqrt{1 \vee \frac{\Delta_n}{1 + \beta_R^2}} + \frac{\Delta_n}{\sqrt{V^*}}\right)$ . Further applying Proposition B6 and Slutsky's theorem, we deduce that

$$\frac{\sqrt{n}B}{\sqrt{V^*}} - z \left( \sqrt{\frac{\widehat{V}}{V^*}} - 1 \right) = o_p \left( \sqrt{1 \vee \frac{\Delta_n}{1 + \beta_R^2}} + \frac{\Delta_n}{\sqrt{V^*}} \right).$$

Furthermore, we have  $\sqrt{n}M/\sqrt{V^*} \xrightarrow{d} N(0, 1)$  by Proposition B5 and Slutsky's theorem. Note that convergence to a continuous distribution implies uniform convergence of cumulative distribution functions. By Proposition B5, we have for any  $z \in \mathbb{R}$ ,

$$\Pr \left[ \frac{\sqrt{n}\widehat{Q}}{\sqrt{\widehat{V}}} \leq z \right] = \Pr \left[ \frac{\sqrt{n}M}{\sqrt{V^*}} + \frac{\sqrt{n}B}{\sqrt{V^*}} - z \left( \sqrt{\frac{\widehat{V}}{V^*}} - 1 \right) \leq z - \frac{\Delta_n}{\sqrt{V^*}} \right].$$

When  $\sqrt{1 \vee \frac{\Delta_n}{1 + \beta_R^2}} + \frac{\Delta_n}{\sqrt{V^*}}$  is bounded, we know that  $\frac{\sqrt{n}M}{\sqrt{V^*}} + \frac{\sqrt{n}B}{\sqrt{V^*}} - z \left( \sqrt{\frac{\widehat{V}}{V^*}} - 1 \right) \xrightarrow{d} N(0, 1)$  and hence

$$\Pr \left[ \frac{\sqrt{n}\widehat{Q}}{\sqrt{\widehat{V}}} \leq z \right] = \Phi \left( z - \frac{\Delta_n}{\sqrt{V^*}} \right) + o(1).$$

When  $\sqrt{1 \vee \frac{\Delta_n}{1 + \beta_R^2}} + \frac{\Delta_n}{\sqrt{V^*}} \rightarrow \infty$ , if  $\sqrt{1 \vee \frac{\Delta_n}{1 + \beta_R^2}} \lesssim \frac{\Delta_n}{\sqrt{V^*}}$ , we have  $\frac{\Delta_n}{\sqrt{V^*}} \rightarrow \infty$  and hence  $\Phi \left( z - \frac{\Delta_n}{\sqrt{V^*}} \right) = o(1)$  as  $\frac{\Delta_n}{\sqrt{V^*}} \rightarrow \infty$ . Then the conclusion follows if  $\Pr \left[ \frac{\sqrt{n}\widehat{Q}}{\sqrt{\widehat{V}}} \leq z \right] = o(1)$ .

Note that

$$\frac{\frac{\sqrt{n}M}{\sqrt{V^*}} + \frac{\sqrt{n}B}{\sqrt{V^*}} - z \left( \sqrt{\frac{\widehat{V}}{V^*}} - 1 \right)}{1 + \frac{\Delta_n}{\sqrt{V^*}}} \xrightarrow{p} 0, \quad \frac{z - \frac{\Delta_n}{\sqrt{V^*}}}{1 + \frac{\Delta_n}{\sqrt{V^*}}} \rightarrow -1$$

and hence

$$\Pr \left[ \frac{\sqrt{n}\hat{Q}}{\sqrt{\hat{V}}} \leq z \right] = \Pr \left[ \frac{\frac{\sqrt{n}M}{\sqrt{V^*}} + \frac{\sqrt{n}B}{\sqrt{V^*}} - z \left( \sqrt{\frac{\hat{V}}{V^*}} - 1 \right)}{1 + \frac{\Delta_n}{\sqrt{V^*}}} \leq \frac{z - \frac{\Delta_n}{\sqrt{V^*}}}{1 + \frac{\Delta_n}{\sqrt{V^*}}} \right] = o(1).$$

As a consequence, it suffices to show that  $\sqrt{1 \vee \frac{\Delta_n}{1 + \beta_R^2}} \lesssim \frac{\Delta_n}{\sqrt{V^*}}$ . It automatically holds when  $\Delta_n/\sqrt{V^*} \rightarrow \infty$  while  $\frac{\Delta_n}{1 + \beta_R^2}$  is bounded. Otherwise, when  $\frac{\Delta_n}{1 + \beta_R^2} \rightarrow \infty$ , by definition in (27) we know that w.p.a.1.

$$\begin{aligned} V &\lesssim (1 + \beta_R^2) \left[ \tilde{\pi}_Z^\top \Omega \hat{\Sigma} \Omega \tilde{\pi}_Z + \frac{\tau_n^*}{n} \|\eta\|_2^2 \right] \\ &\lesssim (1 + \beta_R^2) [Q + \tau_n^*] \lesssim (1 + \beta_R^2) \left[ \frac{\Delta_n}{\sqrt{n}} + 1 \right]. \end{aligned}$$

Then w.p.a.1

$$\frac{\Delta_n}{\sqrt{V^*}} \gtrsim \frac{\Delta_n}{\sqrt{(1 + \beta_R^2) \left[ \frac{\Delta_n}{\sqrt{n}} + 1 \right]}} \gtrsim \min \left( n^{1/4} \sqrt{\frac{\Delta_n}{1 + \beta_R^2}}, \frac{\Delta_n}{\sqrt{1 + \beta_R^2}} \right) \gg \sqrt{\frac{\Delta_n}{1 + \beta_R^2}}$$

where the last inequality applies the fact that  $\Delta_n \geq \frac{\Delta_n}{1 + \beta_R^2} \rightarrow \infty$ .

### B.2.2 Proof of Theorem 2

Similarly, it suffices to show that Theorem 2 holds with  $\eta$  and  $\tau_n^*$  satisfying Assumption B1. By Assumption 4 it suffices to show that  $\Delta_n/\sqrt{V} \xrightarrow{p} \infty$  when  $\Delta_n \rightarrow \Delta \in (0, \infty]$ . By definition in (27) we know that w.p.a.1.

$$\begin{aligned} V &\lesssim (1 + \beta_R^2) \left[ \tilde{\pi}_Z^\top \Omega \hat{\Sigma} \Omega \tilde{\pi}_Z + \frac{\tau_n^*}{n} \|\eta\|_2^2 \right] \\ &\lesssim \left( 1 + \frac{\|\pi\|_2^2}{\|\gamma\|_2^2} \right) [Q + \tau_n^*] \lesssim (1 + \Delta_n) \left( \frac{\Delta_n}{\sqrt{n}} + \tau_n^* \right) \end{aligned}$$



As  $\tau_n^* = o(1)$ , we have  $\frac{\Delta_n}{\sqrt{(1+\Delta_n)\tau_n^*}} \rightarrow \infty$  and  $\frac{\Delta_n}{\sqrt{(1+\Delta_n)\Delta_n/\sqrt{n}}} = \frac{n^{1/4}\Delta_n}{\sqrt{(1+\Delta_n)\Delta_n}} \rightarrow \infty$ .

Then w.p.a.1

$$\frac{\Delta_n}{\sqrt{V}} \gtrsim \frac{\Delta_n}{\sqrt{(1+\Delta_n)\tau_n^*}} \wedge \frac{\Delta_n}{\sqrt{(1+\Delta_n)\Delta_n/\sqrt{n}}} \rightarrow \infty$$

### B.3 Proofs of Preliminary Propositions

This subsection provides the proofs of Proposition B1-B7 in Subsection B.1.

#### B.3.1 Proof of Proposition B1

Based on Lemma 6.17 of [Bühlmann and van de Geer \(2011\)](#), Assumption 1 implies compatibility condition defined in (6.4) of the same reference with  $n$  large enough. Based on the Theorem 6.1 of [Bühlmann and van de Geer \(2011\)](#), it hence suffices to show that  $\mathcal{J}_m = \{4\|n^{-1}W^\top \varepsilon_m\|_\infty \leq \lambda_{mn}\}$  holds w.p.a.1 for  $m = 1, 2$ . By Assumptions 1 and 2 we know that  $W_{ij}\varepsilon_{im}$  for all  $j$  and  $m$  are centered sub-exponential random variables with sub-exponential norm bounded by  $2K^2$ . By Corollary 5.17 in [Vershynin \(2010\)](#) we know that there exists some  $C > 0$  such that for  $m = 1, 2$  and all  $t > 0$ ,

$$\Pr\left(4\left|n^{-1}\sum_{i=1}^n W_{ij}\varepsilon_{im}\right| > t\right) \leq 2\exp\left[-Cn\min\left(\frac{t^2}{64K^4}, \frac{t}{8K^2}\right)\right]. \quad (\text{B9})$$

By union bound, when  $\lambda_{mn} = \sqrt{\frac{128K^4 \log p}{Cn}}$  and  $\log(p) = o(n)$ ,

$$\Pr(4\|n^{-1}W^\top \varepsilon_m\|_\infty > \lambda_{mn}) \leq 2p\exp\left[-\min\left(2\log p, \sqrt{2Cn\log p}\right)\right] \leq 2p^{-1}. \quad (\text{B10})$$

It follows that  $4\|n^{-1}W^\top \varepsilon_m\|_\infty \leq \lambda_{mn}$  w.p.a.1.

#### B.3.2 Proof of Proposition B2

Recall that

$$\begin{aligned} \widehat{\langle \Gamma, \gamma \rangle} - \langle \Gamma, \gamma \rangle &= \widehat{u}_1^\top \frac{1}{n} W^\top \varepsilon_2 + \widehat{u}_2^\top \frac{1}{n} W^\top \varepsilon_1 - (\widehat{u}_2^\top \widehat{\Sigma} - (\mathbf{0}, \widehat{\gamma}^\top)) \begin{pmatrix} \widehat{\Psi} - \Psi \\ \widehat{\Gamma} - \Gamma \end{pmatrix} \\ &\quad - (\widehat{u}_2^\top \widehat{\Sigma} - (\mathbf{0}, \widehat{\Gamma}^\top)) \begin{pmatrix} \widehat{\psi} - \psi \\ \widehat{\gamma} - \gamma \end{pmatrix} - (\widehat{\Gamma} - \Gamma)^\top (\widehat{\gamma} - \gamma). \end{aligned} \quad (\text{B11})$$

By (7.33) in [Cai and Guo \(2017\)](#) we know that  $\Omega(\mathbf{0}, \widehat{\Gamma}^\top)^\top$  is a feasible solution to the minimization problem, and hence  $\|\widehat{u}_1\|_1 \leq \|\Omega(\mathbf{0}, \widehat{\Gamma}^\top)^\top\|_1 \leq m_\omega s \sqrt{\log p/n} + m_\omega \sqrt{s} \|\Gamma\|_2$ . Similarly,  $\|\widehat{u}_2\|_1 \leq m_\omega s \sqrt{\log p/n} + m_\omega \sqrt{s} \|\gamma\|_2$ . Again by (B9) and union bound, we have

$$\Pr(\|n^{-1}W^\top \varepsilon_2\|_\infty > t) \leq 2p \exp \left[ -Cn \min \left( \frac{t^2}{4K^4}, \frac{t}{2K^2} \right) \right].$$

Taking  $t = \sqrt{\frac{8K^4 \log p}{Cn}}$ , we conclude that w.p.a.1,  $\|n^{-1}W^\top \varepsilon_2\|_\infty \lesssim \sqrt{\log p/n}$ . Then

$$|\widehat{u}_1^\top \frac{1}{n} W^\top \varepsilon_2| \leq \|\widehat{u}_1\|_1 \|n^{-1}W^\top \varepsilon_2\|_\infty \lesssim m_\omega \left( \frac{s \log p}{n} + \sqrt{\frac{s \log p}{n}} \|\Gamma\|_2 \right).$$

Similarly,

$$|\widehat{u}_2^\top \frac{1}{n} W^\top \varepsilon_1| \leq \|\widehat{u}_2\|_1 \|n^{-1}W^\top \varepsilon_1\|_\infty \lesssim m_\omega \left( \frac{s \log p}{n} + \sqrt{\frac{s \log p}{n}} \|\gamma\|_2 \right).$$

For the remaining terms,

$$\begin{aligned} \left| (\widehat{u}_1^\top \widehat{\Sigma} - (\mathbf{0}, \widehat{\Gamma}^\top)) \begin{pmatrix} \widehat{\psi} - \psi \\ \widehat{\gamma} - \gamma \end{pmatrix} \right| &\leq \|\widehat{u}_1^\top \widehat{\Sigma} - (\mathbf{0}, \widehat{\Gamma}^\top)\|_\infty (\|\widehat{\psi} - \psi\|_1 + \|\widehat{\gamma} - \gamma\|_1) \\ &\lesssim \|\widehat{\Gamma}\|_2 \frac{s \log p}{n} \leq \|\widehat{\Gamma} - \Gamma\|_2 \frac{s \log p}{n} + \|\Gamma\|_2 \frac{s \log p}{n} \\ &\leq \left( \frac{s \log p}{n} \right)^{3/2} + \|\Gamma\|_2 \frac{s \log p}{n}, \end{aligned}$$

$$\begin{aligned} \left| (\widehat{u}_2^\top \widehat{\Sigma} - (\mathbf{0}, \widehat{\gamma}^\top)) \begin{pmatrix} \widehat{\Psi} - \Psi \\ \widehat{\Gamma} - \Gamma \end{pmatrix} \right| &\leq \|\widehat{u}_2^\top \widehat{\Sigma} - (\mathbf{0}, \widehat{\gamma}^\top)\|_\infty (\|\widehat{\Psi} - \Psi\|_1 + \|\widehat{\Gamma} - \Gamma\|_1) \\ &\lesssim \|\widehat{\gamma}\|_2 \frac{s \log p}{n} \leq \|\widehat{\gamma} - \gamma\|_2 \frac{s \log p}{n} + \|\gamma\|_2 \frac{s \log p}{n} \\ &\leq \left( \frac{s \log p}{n} \right)^{3/2} + \|\gamma\|_2 \frac{s \log p}{n}, \end{aligned}$$

and  $(\widehat{\Gamma} - \Gamma)^\top (\widehat{\gamma} - \gamma) \lesssim s \log p/n$ . When  $s \log p = o(n)$ ,  $\left( \frac{s \log p}{n} \right)^{3/2} \ll \frac{s \log p}{n}$  with  $n$  large enough, and then (B2) follows. By the following decomposition

$$\widehat{\|\gamma\|_2^2} - \|\gamma\|_2^2 = 2\widehat{u}_2^\top \frac{1}{n} W^\top \varepsilon_2 - 2(\widehat{u}_2^\top \widehat{\Sigma} - (\mathbf{0}, \widehat{\gamma}^\top)) \begin{pmatrix} \widehat{\Psi} - \Psi \\ \widehat{\Gamma} - \Gamma \end{pmatrix} - \|\widehat{\gamma} - \gamma\|_2^2 \quad (\text{B12})$$

and similar inequalities

$$|\widehat{u}_2^\top \frac{1}{n} W^\top \varepsilon_2| \lesssim m_\omega \left( \frac{s \log p}{n} + \sqrt{\frac{s \log p}{n}} \|\gamma\|_2 \right),$$

$$\left| (\widehat{u}_2^\top \widehat{\Sigma} - (\mathbf{0}, \widehat{\gamma}^\top)) \begin{pmatrix} \widehat{\psi} - \psi \\ \widehat{\gamma} - \gamma \end{pmatrix} \right| \leq \|(\widehat{\Sigma} \widehat{u}_2 - (\mathbf{0}, \widehat{\gamma}^\top)^\top)\|_\infty (\|\widehat{\psi} - \psi\|_1 + \|\widehat{\gamma} - \gamma\|_1) \lesssim \|\widehat{\gamma}\|_2 s \log p / n$$

and  $\|\widehat{\gamma} - \gamma\|_2^2 \lesssim s \log p / n$ , we can derive (B3) following similar arguments. By (B3) we know that

$$\left| \frac{\widehat{\|\gamma\|_2^2}}{\|\gamma\|_2^2} - 1 \right| = \frac{m_\omega \sqrt{s \log p / n}}{\|\gamma\|_2} + \frac{m_\omega s \log p / n}{\|\gamma\|_2^2}.$$

When  $\|\gamma\|_2 \gg m_\omega \sqrt{s \log p / n}$  and hence  $\|\gamma\|_2^2 \gg m_\omega^2 s \log p / n \geq m_\omega s \log p / n$ , we have  $\frac{\widehat{\|\gamma\|_2^2}}{\|\gamma\|_2^2} - 1 \xrightarrow{p} 0$ . This completes the proof.

### B.3.3 Proof of Proposition B3

We have the following decomposition for the estimation error of  $\beta_R$

$$\widehat{\beta}_R - \beta_R = \frac{\widehat{\langle \Gamma, \gamma \rangle} - \langle \Gamma, \gamma \rangle - \beta_R (\widehat{\|\gamma\|_2^2} - \|\gamma\|_2^2)}{\widehat{\|\gamma\|_2^2}}. \quad (\text{B13})$$

We first need the following lemmas about the estimation errors of the quadratic functionals.

By Proposition B2 we deduce that

$$\begin{aligned}
|\widehat{\beta}_R - \beta_R| &\lesssim m_\omega \frac{\left| (\|\Gamma\|_2 + (1 + |\beta_R|)\|\gamma\|_2) \sqrt{\frac{s \log p}{n}} + \frac{s \log p}{n} \right|}{\widehat{\|\gamma\|_2^2}} \\
&\lesssim \frac{m_\omega \|\gamma\|_2^2}{\widehat{\|\gamma\|_2^2}} \frac{\left| (1 + |\beta_R|)\|\gamma\|_2 \sqrt{\frac{s \log p}{n}} + \|\pi\|_2 \sqrt{\frac{s \log p}{n}} + \frac{s \log p}{n} \right|}{\|\gamma\|_2^2} \\
&\lesssim \frac{m_\omega \|\gamma\|_2^2}{\widehat{\|\gamma\|_2^2}} \left[ \left( 1 + \frac{\|\pi\|_2}{\|\gamma\|_2} \right) \frac{1}{\|\gamma\|_2} \sqrt{\frac{s \log p}{n}} + \frac{1}{\|\gamma\|_2^2} \frac{s \log p}{n} \right] \\
&\leq \frac{m_\omega \|\gamma\|_2^2}{\widehat{\|\gamma\|_2^2}} \left[ \left( 1 + \frac{\|\pi\|_2}{\|\gamma\|_2} \right) \frac{1}{\|\gamma\|_2} \sqrt{\frac{s \log p}{n}} + \left( 1 + \frac{\|\pi\|_2}{\|\gamma\|_2} \right) \frac{1}{\|\gamma\|_2^2} \frac{s \log p}{n} \right] \\
&\lesssim \frac{m_\omega \|\gamma\|_2^2}{\widehat{\|\gamma\|_2^2}} \left( 1 + \frac{\|\pi\|_2}{\|\gamma\|_2} \right) \frac{1}{\|\gamma\|_2} \sqrt{\frac{s \log p}{n}}
\end{aligned}$$

where the third inequality applies the fact that  $\|\Gamma\|_2 \leq \|\pi\|_2 + |\beta| \|\gamma\|_2 \lesssim \|\pi\|_2 + \|\gamma\|_2$ , while the last inequality follows the global strength assumption  $\|\gamma\|_2 \gg m_\omega \sqrt{s \log p / n}$  so that  $s \log p / (n \|\gamma\|_2^2) \leq \sqrt{s \log p / (n \|\gamma\|_2^2)}$  with large enough  $n$ . Then (B4) follows that  $\|\gamma\|_2^2 / \widehat{\|\gamma\|_2^2} \xrightarrow{p} 1$  as  $s \log p = o(n)$  and  $\|\gamma\|_2 \gg m_\omega \sqrt{s \log p / n}$ .

### B.3.4 Proof of Proposition B4

Similarly, it suffices to show that  $4\|n^{-1}W^\top \check{e}\|_\infty \leq \lambda_{3n}$ . Based on the result of Proposition B3, we can derive that  $|\widehat{\beta}_R| \leq |\beta| + |\beta_R - \beta| + |\widehat{\beta}_R - \beta_R| \leq |\beta| + 2(1 + \|\pi\|_2 / \|\gamma\|_2)$  when  $\|\gamma\|_2 \gg m_\omega \sqrt{s \log p / n}$  and hence

$$\begin{aligned}
\|n^{-1}W^\top \check{e}\|_\infty &\leq \|n^{-1}W^\top \varepsilon_1\|_\infty + |\widehat{\beta}_R| \|n^{-1}W^\top \varepsilon_2\|_\infty \\
&\leq \|n^{-1}W^\top \varepsilon_1\|_\infty + [|\beta| + 2(1 + \|\pi\|_2 / \|\gamma\|_2)] \|n^{-1}W^\top \varepsilon_2\|_\infty.
\end{aligned}$$

Then the conclusion follows the result that

$$\begin{aligned}
\Pr(4\|n^{-1}W^\top \check{e}\|_\infty > \lambda_{3n}) &\leq \Pr[\|n^{-1}W^\top \varepsilon_1\|_\infty + [|\beta| + 2(1 + \|\pi\|_2 / \|\gamma\|_2)] \|n^{-1}W^\top \varepsilon_2\|_\infty > \lambda_{3n}/4] \\
&\leq \Pr\left(\|n^{-1}W^\top \varepsilon_1\|_\infty > \sqrt{\frac{8K^4 \log p}{Cn}}\right) + \Pr\left(\|n^{-1}W^\top \varepsilon_2\|_\infty > \sqrt{\frac{8K^4 \log p}{Cn}}\right) \\
&\leq 4p^{-1}.
\end{aligned}$$

This completes the proof of Proposition B4.

### B.3.5 Proof of Proposition B5

We have the following decomposition

$$\begin{aligned}
\widehat{Q} - Q &= \widehat{Q} - \check{Q} + \check{Q} - Q \\
&= \frac{2}{n}(W\widehat{u}_3 + \sqrt{\tau_n}\eta)^\top \check{e} + 2(\widehat{\Sigma}\widehat{u}_3 - \widehat{\pi}_Z)^\top \begin{pmatrix} \widehat{\varphi} - \check{\varphi} \\ \widehat{\pi} - \check{\pi} \end{pmatrix} - \|\widehat{\pi} - \check{\pi}\|_2^2 + \|\gamma\|_2^2(\widehat{\beta}_R - \beta_R)^2 \quad (\text{B14}) \\
&= M + B_1 + B_2 + B_3,
\end{aligned}$$

$$M = \frac{2}{n}(W\Omega\check{\pi}_Z + \sqrt{\tau_n^*}\eta)^\top \check{e}, \quad (\text{B15})$$

$$B_1 = 2(\widehat{\Sigma}\widehat{u}_3 - \widehat{\pi}_Z)^\top \begin{pmatrix} \widehat{\varphi} - \check{\varphi} \\ \widehat{\pi} - \check{\pi} \end{pmatrix} - \|\widehat{\pi} - \check{\pi}\|_2^2 + \frac{2}{n}(0_{p_x}, \check{\pi} - \widehat{\pi})^\top \Omega W^\top \check{e}, \quad (\text{B16})$$

$$B_2 = \frac{2}{n}(\widehat{u}_3 - \Omega\check{\pi}_Z)^\top W^\top \check{e} + \frac{2}{n}(W\widehat{u}_3)^\top (\check{e} - \widetilde{e}) + \|\gamma\|_2^2(\widehat{\beta}_R - \beta_R)^2. \quad (\text{B17})$$

and

$$B_3 = \frac{2}{n}\sqrt{\tau_n^*} \cdot \eta^\top (\check{e} - \widetilde{e}) + \frac{2}{n}(\sqrt{\tau_n} - \sqrt{\tau_n^*})\eta^\top \check{e} + \frac{2}{n}\sqrt{\tau_n}\eta^\top (\widehat{e} - \check{e}). \quad (\text{B18})$$

We first point out that from (B10) we know  $\|n^{-1}W^\top \varepsilon_m\|_\infty \lesssim \sqrt{\log p/n}$  w.p.a.1. In the following proofs, we first prove the asymptotic normality of  $M$ . Then, we show the upper bounds of bias terms  $B_1$ ,  $B_2$ , and  $B_3$ .

Step I. Show asymptotic normality of  $M$ . Write

$$\frac{\sqrt{n}M}{\sqrt{V}} = \frac{2}{\sqrt{V}}\check{\pi}_Z^\top \Omega W^\top \check{e} = \frac{2}{n\sqrt{V}} \sum_{i=1}^n \check{\pi}_Z^\top \Omega W_i \check{e}_i = \frac{1}{\sqrt{n}} \sum_{i=1}^n \zeta_i \quad (\text{B19})$$

where  $\zeta_i = \frac{2}{\sqrt{V}}(\check{\pi}_Z^\top \Omega W_i + \sqrt{\tau_n^*} \cdot \eta_i)\check{e}_i$ . We know that  $\mathbb{E}[\zeta_i|W, \eta] = 0$ ,  $\sum_{i=1}^n \mathbb{E}[(\zeta_i/\sqrt{n})^2|W, \eta] = 1$ . By Corollary 3.1 of Hall and Heyde (1980), it sufficed to show the following conditional Lindeberg condition

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[ \zeta_i^2 \mathbf{1}(|\zeta_i| > \delta\sqrt{n}) \middle| W, \eta \right] \xrightarrow{p} 0.$$

For any  $b \geq 0$ , define

$$\mathcal{W}_{1n}(b) \equiv \left\{ \max_{i \in [n], j \in [p]} |W_{i,j}| \leq \sqrt{\frac{(b+2)\log(np)}{v}} \right\}, \quad \mathcal{W}_{2n}(b) \equiv \left\{ \max_{i \in [n]} |\eta_i| \leq \sqrt{\frac{(b+2)\log n}{v}} \right\},$$

$$\mathcal{W}_{3n}(b) \equiv \left\{ \|\widehat{\Sigma} - \Sigma\|_\infty \leq \sqrt{\frac{4(b+1)\log p}{Cn}} \right\}, \quad \mathcal{W}_{4n}(b) \equiv \left\{ \|W^\top \eta\|_\infty \leq \sqrt{\frac{4(b+1)\log p}{Cn}} \right\}$$

where  $C$  and  $K$  are specified in (B9) and  $v$  is the constant satisfying

$$\max \{ \Pr(|W_{ij}| > t), \Pr(|\eta_i| > t) \} \leq e^{1-vt^2} \text{ for any } t > 0$$

based on sub-Gaussianity of  $W$  and  $\eta$ . We claim the following lemma where the proof is available in B.4.1.

**Lemma B1.** *Under Assumptions 1 and B1 (i), we have for some constant  $c_j$  where  $j = 1, 2, 3, 4$ ,*

$$\Pr(\mathcal{W}_{1n}(b)) \geq 1 - e(np)^{-(b+1)}, \quad (\text{B20})$$

$$\Pr(\mathcal{W}_{2n}(b)) \geq 1 - en^{-(b+1)}, \quad (\text{B21})$$

$$\Pr(\mathcal{W}_{3n}(b)) \geq 1 - 2p^{-b}, \quad (\text{B22})$$

$$\Pr(\mathcal{W}_{4n}(b)) \geq 1 - 2p^{-b}. \quad (\text{B23})$$

Define  $\mathcal{W}_n(b) = \bigcap_{j=1}^4 \mathcal{W}_{jn}$ . We show that  $V \gtrsim (1 + \beta_R^2)(Q + \tau_n^*)$  under  $\mathcal{W}_n(b)$ . By Assumption 2 we know that

$$\begin{aligned} \sigma_i^2 &> \sigma_{i1}^2 + \beta_R^2 \sigma_{i2}^2 - 2\rho_0 \sigma_{i1} \sigma_{i2} \\ &\geq (1 - \rho_0)(\sigma_{i1}^2 + \beta_R^2 \sigma_{i2}^2) \geq (1 - \rho_0)(1 + \beta_R^2) \sigma_{\min}^2 \end{aligned} \quad (\text{B24})$$

where  $\sigma_i^2$  is defined right after (27). Hence, under  $\mathcal{W}_n(b)$  we have

$$\begin{aligned} V &\geq (1 - \rho_0)(1 + \beta_R^2) \sigma_{\min}^2 \frac{1}{n} \sum_{i=1}^n (W_{i\cdot}^\top \Omega \tilde{\pi}_Z + \sqrt{\tau_n^*} \eta_i)^2 \\ &\gtrsim (1 + \beta_R^2) \left[ \tilde{\pi}_Z^\top \Omega \tilde{\pi}_Z + \tilde{\pi}_Z^\top \Omega (\widehat{\Sigma} - \Sigma) \Omega \tilde{\pi}_Z + \frac{\sqrt{\tau_n^*}}{n} \tilde{\pi}_Z^\top W^\top \eta + \frac{\tau_n^*}{n} \|\eta\|_2^2 \right] \\ &\gtrsim (1 + \beta_R^2)(Q + \tau_n^*) \end{aligned} \quad (\text{B25})$$

with  $n$  large enough when  $m_\omega s \sqrt{\log(np)/n} = o(1)$ , where the last inequality applies the fact that  $|\tilde{\pi}_Z^\top \Omega (\widehat{\Sigma} - \Sigma) \Omega \tilde{\pi}_Z| \leq m_\omega^2 \|\tilde{\pi}_Z\|_1^2 \|\widehat{\Sigma} - \Sigma\|_\infty \lesssim Q \cdot m_\omega^2 s \sqrt{\log(np)/n} \ll Q$  and that  $|n^{-1} \sqrt{\tau_n^*} \tilde{\pi}_Z^\top \Omega W^\top \eta| \lesssim \sqrt{\tau_n^*} \|\tilde{\pi}_Z\|_2 m_\omega \sqrt{s \log p/n} \ll \tau_n^* + Q$ . By  $|\tilde{e}_i| \lesssim |\varepsilon_{i1}| + |\beta_R| |\varepsilon_{i2}| \lesssim (1 + |\beta_R|^2)^{1/2} (|\varepsilon_{i1}| + |\varepsilon_{i2}|)$ , we have w.p.a.1

$$\begin{aligned} |\zeta_i| &\lesssim V^{-1/2} \left( \|\Omega \tilde{\pi}_Z\|_1 \max_{i,j} |W_{i,j}| + \sqrt{\tau_n^*} \|\eta\|_\infty \right) |\tilde{e}_i| \\ &\lesssim V^{-1/2} \left( m_\omega \sqrt{sQ} \max_{i,j} |W_{i,j}| + \sqrt{\tau_n^*} \|\eta\|_\infty \right) \sqrt{1 + \beta_R^2 (|\varepsilon_{i1}| + |\varepsilon_{i2}|)} \\ &\equiv g_n(|\varepsilon_{i1}| + |\varepsilon_{i2}|) \end{aligned}$$

where  $g_n = V^{-1/2} (m_\omega \sqrt{sQ} \max_{i,j} |W_{i,j}| + \sqrt{\tau_n^*} \|\eta\|_\infty) \sqrt{1 + \beta_R^2}$ . Under  $\mathcal{W}_n(b)$  by (B25) we have  $g_n \lesssim m_\omega \sqrt{s \log(np)}$ . Therefore, given any  $b > 0$  under  $\mathcal{W}_n(b)$  we have for any  $\delta > 0$

$$\begin{aligned}
& \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[ \zeta_i^2 \mathbf{1}(|\zeta_i| > \delta \sqrt{n}) \middle| W, \eta \right] \\
& \leq \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left[ \zeta_i^2 \mathbf{1} \left( |\varepsilon_{i1}| + |\varepsilon_{i2}| \gtrsim \delta \frac{\sqrt{n}}{g_n} \right) \middle| W, \eta \right] \\
& \leq \frac{1}{(1 - \rho_0)(1 + \beta_R^2) \sigma_{\min}^2} \sum_{i=1}^n \frac{(\tilde{\pi}_Z^\top \Omega W_i + \sqrt{\tau_n^*} \eta_i)^2}{\sum_{i=1}^n (\tilde{\pi}_Z^\top \Omega W_i + \sqrt{\tau_n^*} \eta_i)^2} \mathbb{E} \left[ \tilde{e}_i^2 \mathbf{1} \left( |\varepsilon_{i1}| + |\varepsilon_{i2}| \gtrsim \delta \frac{\sqrt{n}}{g_n} \right) \middle| W, \eta \right] \\
& \lesssim \sum_{i=1}^n \frac{(\tilde{\pi}_Z^\top \Omega W_i + \sqrt{\tau_n^*} \eta_i)^2}{\sum_{i=1}^n (\tilde{\pi}_Z^\top \Omega W_i + \sqrt{\tau_n^*} \eta_i)^2} \mathbb{E} \left[ (|\varepsilon_{i1}| + |\varepsilon_{i2}|)^2 \mathbf{1} \left( |\varepsilon_{i1}| + |\varepsilon_{i2}| \gtrsim \delta \frac{\sqrt{n}}{g_n} \right) \middle| W \right] \\
& \leq \left( \frac{m_\omega}{\delta} \sqrt{\frac{s \log(np)}{n}} \right)^{c^*} \sum_{i=1}^n \frac{(\tilde{\pi}_Z^\top \Omega W_i + \sqrt{\tau_n^*} \eta_i)^2}{\sum_{i=1}^n (\tilde{\pi}_Z^\top \Omega W_i + \sqrt{\tau_n^*} \eta_i)^2} \mathbb{E} \left[ (|\varepsilon_{i1}| + |\varepsilon_{i2}|)^{2+c^*} \middle| W, \eta \right] \\
& \lesssim \left( \frac{m_\omega}{\delta} \sqrt{\frac{s \log(np)}{n}} \right)^{c^*}
\end{aligned} \tag{B26}$$

where the third inequality applies (B24), and the last two inequalities apply Assumption 2. The RHS approaches zero as  $n \rightarrow \infty$ . As  $\Pr[\mathcal{W}_n(b)] \rightarrow 0$ , the conditional Lindeberg condition is verified, which implies  $\sqrt{n}M/\sqrt{V} \xrightarrow{d} N(0, 1)$ .

Step II. Show the upper bound of  $B_1$ . We deduce by the construction of  $\hat{u}_3$  and Proposition B4 that w.p.a.1,

$$\begin{aligned}
\left| (\hat{\Sigma} \hat{u} - \hat{\pi}_Z)^\top \begin{pmatrix} \hat{\varphi} - \check{\varphi} \\ \hat{\pi} - \check{\pi} \end{pmatrix} \right| & \leq \|\hat{\Sigma} \hat{u} - \hat{\pi}_Z\|_\infty (\|\hat{\varphi} - \check{\varphi}\|_1 + \|\hat{\pi} - \check{\pi}\|_1) \leq \|\hat{\pi}\|_2 \left( 1 + \frac{\|\pi\|_2}{\|\gamma\|_2} \right) \frac{s \log p}{n} \\
& \lesssim (1 + \|\pi\|_2^2 / \|\gamma\|_2^2) \frac{s \log p}{n} + \left( 1 + \frac{\|\pi\|_2}{\|\gamma\|_2} \right) \|\pi\|_2 \frac{s \log p}{n}
\end{aligned}$$

with  $n$  large enough as  $s \log p / n = o(1)$ ,

$$\|\hat{\pi} - \check{\pi}\|_2^2 \lesssim (1 + \|\pi\|_2^2 / \|\gamma\|_2^2) \frac{s \log p}{n},$$

and by  $\|n^{-1}W\tilde{e}\|_\infty \lesssim (1 + \|\pi\|_2/\|\gamma\|_2)\sqrt{\log p/n}$  we deduce that

$$\begin{aligned} |(\hat{\pi}_Z - \tilde{\pi}_Z)^\top \Omega n^{-1}W\tilde{e}| &\leq m_\omega \|\hat{\pi} - \tilde{\pi}\|_1 \cdot \|n^{-1}W(\varepsilon_1 - \beta_R \varepsilon_2)\|_\infty \\ &\lesssim m_\omega \left(1 + \frac{\|\pi\|_2}{\|\gamma\|_2}\right) \sqrt{\frac{s^2 \log p}{n}} \cdot \left(1 + \frac{\|\pi\|_2}{\|\gamma\|_2}\right) \sqrt{\frac{\log p}{n}} \\ &\lesssim \left(1 + \frac{\|\pi\|_2^2}{\|\gamma\|_2^2}\right) \frac{m_\omega s \log p}{n}. \end{aligned}$$

Hence,

$$\begin{aligned} |B_1| &\lesssim s \log p/n + \|\pi\|_2 s \log p/n + \|\pi\|_2^2 [s \log p/(n\|\gamma\|_2) + m_\omega s \log p/(n\|\gamma\|_2^2)] \\ &\lesssim s \log p/n + (1 + \|\pi\|_2^2) s \log p/n + \|\pi\|_2^2 \cdot s \log p/\sqrt{n} \\ &\lesssim s \log p/n + \|\pi\|_2^2 \cdot s \log p/\sqrt{n} = o_p(\sqrt{\tau_n^*/n} + Q). \end{aligned}$$

Step III. Show the upper bound of  $B_2$ . We first need the following lemma about the approximation error of  $\hat{u}_3$  compared to  $\Omega\tilde{\pi}_Z$ , where the proof is available in [B.4.2](#).

**Lemma B2.** *Let  $\xi$  be a arbitrary  $p \times 1$  and  $\|\xi\|_0 \leq s$ . Suppose  $\hat{\xi}$  is an estimator of  $\xi$  satisfying  $\|\hat{\xi} - \xi\|_2 \lesssim \sqrt{s}\lambda$  and  $\|\hat{\xi} - \xi\|_1 \lesssim s\lambda$  for some  $\lambda > 0$ . If  $\mu_n \asymp \sqrt{\log p/n}$ , then w.p.a.1, the projection direction constructed by*

$$\hat{u} = \arg \min_u \|u\|_1 \text{ s.t. } \|\hat{\Sigma}u - \hat{\xi}\|_\infty \leq \|\hat{\xi}\|_2 \mu_n \quad (\text{B27})$$

satisfies

$$\|\hat{u} - \Omega\xi\|_\infty \lesssim m_\omega \left( \|\xi\|_2 \sqrt{\frac{s \log p}{n}} + s\lambda \sqrt{\frac{\log p}{n}} + \sqrt{s}\lambda \right). \quad (\text{B28})$$

Additionally under Assumption [3](#), we have w.p.a.1

$$\|\hat{u} - \Omega\xi\|_1 \lesssim m_\omega s\lambda + s_\omega s \|\xi\|_2^q \|\hat{u} - \Omega\xi\|_\infty^{1-q}. \quad (\text{B29})$$

By Lemma [B2](#) we know that with  $\xi = \tilde{\pi}_Z$  and  $\lambda = \lambda_{3n} \asymp (1 + \|\pi\|_2/\|\gamma\|_2)\sqrt{\log p/n}$ , when



$$s \log p = o(n),$$

$$\begin{aligned}
\|\widehat{u}_3 - \Omega \widetilde{\pi}_Z\|_1 &\lesssim m_\omega \left(1 + \frac{\|\pi\|_2}{\|\gamma\|_2}\right) \sqrt{\frac{s^2 \log p}{n}} \\
&\quad + s_\omega s \|\widetilde{\pi}\|_2^q m_\omega^{1-q} \left[ \|\widetilde{\pi}\|_2 \sqrt{\frac{s \log p}{n}} + \left(1 + \frac{\|\pi\|_2}{\|\gamma\|_2}\right) \sqrt{\frac{s \log p}{n}} \right]^{1-q} \\
&\leq m_\omega \left(1 + \frac{\|\pi\|_2}{\|\gamma\|_2}\right) \sqrt{\frac{s^2 \log p}{n}} + s_\omega s m_\omega^{1-q} \|\widetilde{\pi}\|_2 \left(\frac{s \log p}{n}\right)^{(1-q)/2} \\
&\quad + s_\omega s m_\omega^{1-q} \|\widetilde{\pi}\|_2^q \left(1 + \frac{\|\pi\|_2}{\|\gamma\|_2}\right)^{1-q} \left(\frac{s \log p}{n}\right)^{(1-q)/2}
\end{aligned}$$

where the second and the third inequality applies the fact that

$$(a + b)^x \leq a^x + b^x \text{ for any } a, b > 0 \text{ and } 0 < x < 1. \quad (\text{B30})$$

Since by Proposition B3

$$\begin{aligned}
\|\widetilde{\pi}\|_2 &\leq \|\widetilde{\pi} - \pi\|_2 + \|\pi\|_2 \\
&= |\widehat{\beta}_R - \beta| \|\gamma\|_2 + \|\pi\|_2 \\
&\leq |\widehat{\beta}_R - \beta_R| \|\gamma\|_2 + 2\|\pi\|_2 \\
&\leq m_\omega \left(1 + \frac{\|\pi\|_2}{\|\gamma\|_2}\right) \sqrt{\frac{s \log p}{n}} + 2\|\pi\|_2 \\
&\leq 2m_\omega \sqrt{\frac{s \log p}{n}} + 3\|\pi\|_2
\end{aligned}$$

with  $n$  large enough when  $m_\omega \sqrt{s \log p / n} = o(\|\gamma\|_2)$ . Hence, as  $s \log p = o(n)$

$$\begin{aligned}
\|\widehat{u}_3 - \Omega \widetilde{\pi}_Z\|_1 &\lesssim s_\omega s m_\omega^{1-q} \|\pi\|_2 \left(\frac{s \log p}{n}\right)^{(1-q)/2} + s_\omega s m_\omega^{1-q} \|\pi\|_2^q \left(1 + \frac{\|\pi\|_2}{\|\gamma\|_2}\right)^{1-q} \left(\frac{s \log p}{n}\right)^{(1-q)/2} \\
&\quad + s_\omega s m_\omega^{2-q} \left(1 + \frac{\|\pi\|_2}{\|\gamma\|_2}\right) \left(\frac{s \log p}{n}\right)^{1/2}.
\end{aligned} \quad (\text{B31})$$

From the derivations in Step II we have shown that  $\|n^{-1} W \widetilde{e}\|_\infty \lesssim (1 + \beta_R^2)^{1/2} \sqrt{\log p / n} \lesssim$

$(1 + \|\pi\|_2/\|\gamma\|_2)\sqrt{s \log p/n}$ . Then w.p.a.1,

$$\begin{aligned}
& \left| \frac{2}{n} (\hat{u}_3 - \Omega \tilde{\pi}_Z)^\top W^\top \tilde{e} \right| \\
& \leq \|\hat{u}_3 - \Omega \tilde{\pi}_Z\|_1 \|n^{-1} W^\top \tilde{e}\|_\infty \\
& \lesssim \|\pi\|_2 (1 + |\beta_R|^2)^{1/2} \frac{m_\omega^{1-q} s_\omega s^{(3-q)/2} (\log p)^{1-q/2}}{n^{1-q/2}} \\
& \quad + \left(1 + \frac{\|\pi\|_2^2}{\|\gamma\|_2^2}\right)^{1-q/2} \|\pi\|_2^q \frac{m_\omega^{1-q} s_\omega s^{(3-q)/2} (\log p)^{1-q/2}}{n^{1-q/2}} + \left(1 + \frac{\|\pi\|_2^2}{\|\gamma\|_2^2}\right) \frac{s_\omega m_\omega^{2-q} s^{3/2} \log p}{n} \\
& \lesssim \left(1 + \frac{\|\pi\|_2^2}{\|\gamma\|_2^2}\right) \frac{s_\omega m_\omega^{2-q} s^{3/2} \log p}{n} + (\sqrt{Q/n} + Q) \frac{m_\omega^{1-q} s_\omega s^{(3-q)/2} (\log p)^{1-q/2}}{n^{(1-q)/2}} \\
& \quad + \left(1 + \frac{\|\pi\|_2^2}{\|\gamma\|_2^2}\right)^{1-q/2} \|\pi\|_2^q \frac{m_\omega^{1-q} s_\omega s^{(3-q)/2} (\log p)^{1-q/2}}{n^{1-q/2}} \\
& \lesssim \left(1 + \frac{\|\pi\|_2^2}{\|\gamma\|_2^2}\right) \frac{s_\omega m_\omega^{2-q} s^{3/2} \log p}{n} + (\sqrt{Q/n} + Q) \frac{m_\omega^{1-q} s_\omega s^{(3-q)/2} (\log p)^{1-q/2}}{n^{(1-q)/2}} \\
& \quad + \max\{1, \|\pi\|_2\} \frac{m_\omega^{1-q} s_\omega s^{(3-q)/2} (\log p)^{1-q/2}}{n^{1-q/2}} + \frac{\|\pi\|_2^2}{\|\gamma\|_2^{2-q}} \frac{m_\omega^{1-q} s_\omega s^{(3-q)/2} (\log p)^{1-q/2}}{n^{1-q/2}} \\
& \lesssim \frac{s_\omega m_\omega^{2-q} s^{3/2} \log p}{n} + \frac{m_\omega^{1-q} s_\omega s^{(3-q)/2} (\log p)^{1-q/2}}{n^{1-q/2}} + (\sqrt{Q/n} + Q) \frac{m_\omega^{1-q} s_\omega s^{(3-q)/2} (\log p)^{1-q/2}}{n^{(1-q)/2}} \\
& \quad + \|\pi\|_2^2 \left[ \frac{1}{\|\gamma\|_2^2} \left( \frac{m_\omega s \log p}{n} + \frac{s_\omega m_\omega^{2-q} s^{3/2} \log p}{n} \right) + \frac{1}{\|\gamma\|_2^{2-q}} \frac{m_\omega^{1-q} s_\omega s^{(3-q)/2} (\log p)^{1-q/2}}{n^{1-q/2}} \right] \\
& \lesssim \frac{s_\omega m_\omega^{2-q} s^{3/2} \log p}{n} + \frac{m_\omega^{1-q} s_\omega s^{(3-q)/2} (\log p)^{1-q/2}}{n^{1-q/2}} + (\sqrt{Q/n} + Q) \frac{m_\omega^{1-q} s_\omega s^{(3-q)/2} (\log p)^{1-q/2}}{n^{(1-q)/2}} \\
& \quad + \|\pi\|_2^2 \left[ \frac{s_\omega m_\omega^{2-q} s^{3/2} \log p}{\sqrt{n}} + \frac{m_\omega^{1-q} s_\omega s^{(3-q)/2} (\log p)^{1-q/2}}{n^{(1-q)/2}} \right]
\end{aligned}$$

where the third inequality applies  $\|\pi\|_2 |\beta_R| \lesssim \|\pi\|_2^2 / \|\gamma\|_2 \lesssim n^{1/4} Q$ , the fourth inequality applies (B30) the fact that  $\|\pi\|_2^q \leq \max\{1, \|\pi\|_2\}$  and the fifth applies  $\max\{1, \|\pi\|_2\} \leq 1 + \|\pi\|_2 \lesssim 1 + \sqrt{Q}$  w.p.a.1. The last inequality applies the global strength of instruments. By Assumptions 4 (i) and B1 (ii) we know that

$$\frac{s_\omega m_\omega^{2-q} s^{3/2} \log p}{\sqrt{n}} + \frac{m_\omega^{1-q} s_\omega s^{(3-q)/2} (\log p)^{1-q/2}}{n^{(1-q)/2}} = o(\sqrt{(1 \vee \Delta_n) \tau_n^*} \wedge 1).$$

Then we have  $\left| \frac{2}{n} (\hat{u}_3 - \Omega \tilde{\pi}_Z)^\top W^\top \tilde{e} \right| = o_p(\sqrt{(1 \vee \Delta_n) \tau_n^* / n} + \sqrt{Q/n}) + o_p(Q)$ .

For the second term of  $B_2$  in (B17), we have the following inequality

$$\begin{aligned}
\left| \frac{2}{n} \widehat{u}_3 W^\top (\tilde{e} - \widetilde{e}) \right| &\leq 2 |\widehat{\beta}_R - \beta_R| \|\widehat{u}_3\|_1 \left\| \frac{1}{n} W^\top \varepsilon_2 \right\|_\infty \\
&\lesssim |\widehat{\beta}_R - \beta_R| \|\Omega \widehat{\pi}_Z\|_1 \sqrt{\frac{\log p}{n}} \\
&\leq |\widehat{\beta}_R - \beta_R| \sqrt{\frac{\log p}{n}} \cdot m_\omega (\|\tilde{\pi}\|_1 + \|\widehat{\pi} - \tilde{\pi}\|_1) \\
&\lesssim |\widehat{\beta}_R - \beta_R| m_\omega \|\tilde{\pi}\|_2 \sqrt{\frac{s \log p}{n}} + |\widehat{\beta}_R - \beta_R| m_\omega \left( 1 + \frac{\|\pi\|_2}{\|\gamma\|_2} \right) \frac{s \log p}{n} \\
&\lesssim |\widehat{\beta}_R - \beta_R| m_\omega \|\pi\|_2 \sqrt{\frac{s \log p}{n}} + |\widehat{\beta}_R - \beta_R| \frac{m_\omega s \log p}{n}
\end{aligned}$$

where the second inequality by the fact that  $\Omega \widehat{\pi}_Z$  is also in the feasible set of optimization problem (B27), and the last inequality applies global strength of instruments. By Proposition B3, the first term on the right hand side can be respectively bounded by

$$\begin{aligned}
&|\widehat{\beta}_R - \beta_R| m_\omega \|\pi\|_2 \sqrt{\frac{s \log p}{n}} \\
&\lesssim m_\omega^2 \left[ \left( 1 + \frac{\|\pi\|_2}{\|\gamma\|_2} \right) \frac{\|\pi\|_2 s \log p}{\|\gamma\|_2 n} + \frac{\|\pi\|_2}{\|\gamma\|_2^2} \left( \frac{s \log p}{n} \right)^{3/2} \right] \\
&\lesssim m_\omega^2 \left[ \left( 1 + \frac{\|\pi\|_2^2}{\|\gamma\|_2^2} \right) \frac{s \log p}{n} + \frac{1 + \|\pi\|_2^2}{\|\gamma\|_2^2} \left( \frac{s \log p}{n} \right)^{3/2} \right] \\
&\lesssim m_\omega^2 \left[ \left( 1 + \frac{\|\pi\|_2^2}{\|\gamma\|_2^2} \right) \frac{s \log p}{n} + \frac{1}{\|\gamma\|_2^2} \left( \frac{s \log p}{n} \right)^{3/2} \right] \\
&\lesssim m_\omega^2 \left[ \frac{(s \log p)^{3/2}}{n} + Q \frac{s \log p}{\sqrt{n}} \right]
\end{aligned}$$

by the global strength of instruments. Based on Assumptions 4 and B1 (ii) we have  $|\widehat{\beta}_R - \beta_R| m_\omega \|\pi\|_2 \sqrt{\frac{s \log p}{n}} = o_p(\sqrt{\tau_n^*/n}) + o_p(Q)$ . Besides,

$$\begin{aligned}
|\widehat{\beta}_R - \beta_R| m_\omega \frac{s \log p}{n} &\lesssim \left( 1 + \frac{\|\pi\|_2}{\|\gamma\|_2} \right) \frac{m_\omega^2}{\|\gamma\|_2} \left( \frac{s \log p}{n} \right)^{3/2} \\
&\lesssim \left( 1 + \frac{\|\pi\|_2^2}{\|\gamma\|_2^2} \right) \frac{m_\omega^2}{\|\gamma\|_2} \left( \frac{s \log p}{n} \right)^{3/2} \\
&\lesssim \frac{m_\omega^2 (s \log p)^{3/2}}{n^{5/4}} + Q \frac{m_\omega^2 (s \log p)^{3/2}}{n^{3/4}}.
\end{aligned}$$

By Assumptions 4 and B1 (ii)  $|\hat{\beta}_R - \beta_R| m_\omega \left(1 + \frac{\|\pi\|_2}{\|\gamma\|_2}\right) \frac{s \log p}{n} = o_p(\sqrt{(1 \vee \Delta_n) \tau_n^*/n}) + o_p(Q)$ .

For the remaining component we have

$$(\hat{\beta}_R - \beta_R)^2 \|\gamma\|_2 \sqrt{\frac{m_\omega^2 s \log p}{n}} = (\hat{\beta}_R - \beta_R)^2 \|\gamma\|_2^2 \sqrt{\frac{m_\omega^2 s \log p}{n \|\gamma\|_2^2}} \ll (\hat{\beta}_R - \beta_R)^2 \|\gamma\|_2^2$$

and hence it suffices to bound the  $\|\gamma\|_2^2 (\hat{\beta}_R - \beta_R)^2$ , which is exactly the third term of  $B_2$ .

In addition, by Proposition B3 under the global strength assumption, we have w.p.a.1

$$\|\gamma\|_2^2 (\hat{\beta}_R - \beta_R)^2 \lesssim (1 + \|\pi\|_2^2 / \|\gamma\|_2^2) m_\omega^2 s \log p / n \lesssim m_\omega^2 s \log p / n + Q \cdot m_\omega^2 s \log p / \sqrt{n}. \quad (\text{B32})$$

By Assumptions 4 (i) and B1 (ii) we have  $m_\omega^2 s \log p / n = o(\sqrt{(1 \vee \Delta_n) \tau_n^*})$  and  $Q \cdot m_\omega^2 s \log p / \sqrt{n} = o(Q)$ . Then  $\|\gamma\|_2^2 (\hat{\beta}_R - \beta_R)^2 = o_p(\sqrt{(1 \vee \Delta_n) \tau_n^* / n}) + o_p(Q)$ .

Step IV. Show the upper bound of  $B_3$ . We first have  $n^{-1} \sum_{i=1}^n \eta_i \varepsilon_{im} \lesssim n^{-1/2}$  for  $m = 1, 2$  w.p.a.1 applying the fact that  $\eta_i \varepsilon_{im}$  is sub-exponential with  $\mathbb{E}(\eta_i \varepsilon_{im}) = \mathbb{E}[\eta_i \mathbb{E}(\varepsilon_{im} | \eta_i)] = 0$  by Assumptions 2 and B1 (i) and Bernstein-type inequality in Proposition 5.16 of Vershynin (2010) by taking  $a_i = 1$  and  $t = \sqrt{n}$ . Then

$$\begin{aligned} \left| \frac{2\sqrt{\tau_n^*}}{n} \eta^\top (\tilde{e} - \tilde{e}) \right| &\lesssim \sqrt{\tau_n^*} |n^{-1} \eta^\top \varepsilon_2| \cdot |\hat{\beta}_R - \beta_R| \\ &\lesssim \frac{m_\omega \sqrt{\tau_n^*}}{\sqrt{n}} \left(1 + \frac{\|\pi\|_2}{\|\gamma\|_2}\right) \frac{1}{\|\gamma\|_2} \sqrt{\frac{s \log p}{n}} \\ &\lesssim \sqrt{\frac{\tau_n^*}{n}} \cdot \frac{1}{\|\gamma\|_2} \sqrt{\frac{m_\omega^2 s \log p}{n}} + \sqrt{\frac{\tau_n^*}{n}} \frac{\|\pi\|_2}{\sqrt{n} \|\gamma\|_2} \frac{\sqrt{m_\omega^2 s \log p}}{\|\gamma\|_2} \\ &\lesssim o_p \left( \sqrt{\frac{\tau_n^*}{n}} \right) + \sqrt{\frac{\tau_n^*}{n}} \left( \|\pi\|_2^2 + \frac{1}{n \|\gamma\|_2^2} \right) \frac{\sqrt{m_\omega^2 s \log p}}{\|\gamma\|_2} \\ &= o_p \left( \sqrt{\frac{\tau_n^*}{n}} \right) + o_p(Q) + \sqrt{\frac{\tau_n^*}{n}} \cdot \frac{\sqrt{m_\omega^2 s \log p}}{n \|\gamma\|_2^3} \\ &= o_p \left( \sqrt{\frac{\tau_n^*}{n}} \right) + o_p(Q) \end{aligned}$$

w.p.a.1 by Assumptions 4 and B1 (ii). Therefore, we show that the first term of  $B_3 = o_p(\sqrt{V/n} + Q)$ . For the second term, noting that with  $\tau_n^*$  uniformly bounded and  $\tau_n / \tau_n^* \xrightarrow{p} 1$ , we follow

similar arguments to deduce that

$$\begin{aligned} \left| \frac{2}{n} (\sqrt{\tau_n} - \sqrt{\tau_n^*}) \eta^\top (\varepsilon_1 - \hat{\beta}_R \varepsilon_2) \right| &\lesssim \left| \sqrt{\frac{\tau_n}{\tau_n^*}} - 1 \right| \frac{(1 + |\beta_R|) \sqrt{\tau_n^*} + |\hat{\beta}_R - \beta_R| \sqrt{\tau_n^*}}{\sqrt{n}} \\ &= o_p \left( \sqrt{\frac{(1 + \beta_R^2) \tau_n^*}{n}} \right) + o_p(Q). \end{aligned}$$

w.p.a.1 by Proposition B3 and global strength of instruments. Finally, by the fact that  $\tau_n = O_p(\tau_n^*) = O_p(1)$  and Assumption B1 (i), Proposition B4 and (B23), we have w.p.a.1

$$\begin{aligned} \left| \frac{2}{n} \sqrt{\tau_n} \eta^\top (\hat{e} - \check{e}) \right| &= \frac{1}{n} \|W^\top \eta\|_\infty (\|\hat{\varphi} - \check{\varphi}\|_1 + \|\hat{\pi} - \check{\pi}\|_1) \\ &\lesssim \left( 1 + \frac{\|\pi\|_2}{\|\gamma\|_2} \right) \frac{s \log p}{n} = o_p[\sqrt{(1 \vee \Delta_n) \tau_n^*/n} + \sqrt{Q/n}]. \end{aligned}$$

Combining all results in Steps 2 to 4, we have

$$|B| \leq |B_1| + |B_2| + |B_3| = o_p \left[ \sqrt{\frac{Q}{n}} + \sqrt{\frac{((1 + \beta_R^2) \vee \tau_n^*)}{n}} + Q \right].$$

### B.3.6 Proof of Proposition B6

It suffices to show that

$$\hat{V} - V = o_p \left[ \left( 1 \vee \frac{\Delta_n}{1 + \beta_R^2} \right) V + \Delta_n^2 \right].$$

Note that

$$\begin{aligned} \hat{V} - V &= \frac{1}{n} \sum_{i=1}^n (\hat{u}_3^\top W_i + \sqrt{\tau_n} \eta_i)^2 \hat{e}_i^2 - \frac{1}{n} \sum_{i=1}^n (\tilde{\pi}_Z^\top \Omega W_i + \sqrt{\tau_n^*} \eta_i)^2 \sigma_i^2 \\ &= I_1^V + I_2^V + I_3^V, \end{aligned}$$

where

$$I_1^V = \frac{1}{n} \sum_{i=1}^n [(\hat{u}_3 - \Omega \tilde{\pi}_Z)^\top W_i + (\sqrt{\tau_n} - \sqrt{\tau_n^*}) \eta_i]^2 \sigma_i^2, \quad (\text{B33})$$

$$I_2^V = \frac{1}{n} \sum_{i=1}^n (\tilde{\pi}_Z^\top \Omega W_i + \sqrt{\tau_n^*} \eta_i) [(\hat{u}_3 - \Omega \tilde{\pi}_Z)^\top W_i + (\sqrt{\tau_n} - \sqrt{\tau_n^*}) \eta_i] \sigma_i^2, \quad (\text{B34})$$

and

$$I_3^V = \frac{1}{n} \sum_{i=1}^n [\hat{u}_3^\top W_i + \sqrt{\tau_n} \eta_i]^2 (\hat{e}_i^2 - \sigma_i^2). \quad (\text{B35})$$

We first show a lemma useful for finding the upper bounds of some quantities, and the proof is available in [B.4.3](#).

**Lemma B3.** *Under Assumptions [1](#), [2](#), [4](#) and [B1](#) (i), we have w.p.a.1*

$$\max \left( \left\| \frac{1}{n} \sum_{i=1}^n W_i W_i^\top \sigma_i^2 \right\|_\infty, \left\| \frac{1}{n} \sum_{i=1}^n W_i \eta_i \sigma_i^2 \right\|_\infty, \left| \frac{1}{n} \sum_{i=1}^n \eta_i^2 \sigma_i^2 \right| \right) \lesssim 1 + \beta_R^2. \quad (\text{B36})$$

Thus, for any  $\xi_1, \xi_2 \in \mathbb{R}^p$ , w.p.a.1

$$\left| \frac{1}{n} \sum_{i=1}^n \xi_1^\top W_i W_i^\top \xi_2 \sigma_i^2 \right| \lesssim \|\xi_1\|_1 \|\xi_2\|_1 (1 + \beta_R^2), \quad \left| \frac{1}{n} \sum_{i=1}^n \xi_2^\top W_i \eta_i \sigma_i^2 \right| \lesssim \|\xi_2\|_1 (1 + \beta_R^2). \quad (\text{B37})$$

Furthermore, if  $\|\xi_2\|_0 \leq s$ , we have

$$\frac{1}{n} \sum_{i=1}^n (\xi_2^\top \Omega W_i)^2 \sigma_i^2 \lesssim \|\xi_2\|_2^2 (1 + \beta_R^2), \quad \left| \frac{1}{n} \sum_{i=1}^n \xi_1^\top W_i W_i^\top \Omega \xi_2 \sigma_i^2 \right| \lesssim \|\xi_1\|_1 \|\xi_2\|_2 (1 + \beta_R^2). \quad (\text{B38})$$

Step I. Bound  $I_1^V$ . By [\(B31\)](#), [\(B32\)](#), Assumptions [4](#) and [B1](#) and the fact that  $\|\pi\|_2^q \leq$

$\max\{1, \|\pi\|_2\} \leq 1 + \|\pi\|_2$ , we have w.p.a.1

$$\begin{aligned}
& \|\widehat{u}_3 - \Omega\check{\pi}_Z\|_1^2 + \|\check{\pi}_Z - \widetilde{\pi}_Z\|_2^2 \\
& \leq Q \left[ s_\omega s m_\omega^{1-q} \left( \frac{s \log p}{n} \right)^{(1-q)/2} \right]^2 + \left[ s_\omega s m_\omega^{1-q} \|\pi\|_2^q \left( 1 + \frac{\|\pi\|_2}{\|\gamma\|_2} \right)^{1-q} \left( \frac{s \log p}{n} \right)^{(1-q)/2} \right]^2 \\
& \quad + \left[ s_\omega s m_\omega^{2-q} \left( 1 + \frac{\|\pi\|_2}{\|\gamma\|_2} \right) \left( \frac{s \log p}{n} \right)^{1/2} \right]^2 + \|\gamma\|_2^2 (\widehat{\beta}_R - \beta_R)^2 \\
& \lesssim Q \left\{ \frac{m_\omega^2 s \log p}{\sqrt{n}} + \left( \frac{s_\omega s^{(3-q)/2} m_\omega^{1-q} (\log p)^{(1-q)/2}}{n^{(1-q)/2}} \right)^2 \right\} \\
& \quad + \frac{Q}{\|\gamma\|_2^{2(1-q)}} \left( \frac{s_\omega s^{(3-q)/2} m_\omega^{1-q} (\log p)^{(1-q)/2}}{n^{(1-q)/2}} \right)^2 + \frac{Q}{\|\gamma\|_2^2} \left( \frac{s_\omega s^{3/2} m_\omega^{2-q} (\log p)^{1/2}}{n^{1/2}} \right)^2 \\
& \quad + \left( \frac{s_\omega m_\omega^{2-q} s^{3/2} (\log p)^{1/2}}{n^{1/2}} \right)^2 + \left( \frac{s_\omega s^{(3-q)/2} m_\omega^{1-q} (\log p)^{(1-q)/2}}{n^{(1-q)/2}} \right)^2 + \frac{m_\omega^2 s \log p}{n} \\
& \lesssim Q \left\{ \frac{m_\omega^2 s \log p}{\sqrt{n}} + \left( \frac{s_\omega s^{(3-q)/2} m_\omega^{1-q} (\log p)^{(1-q)/2}}{n^{(1-q)/2}} \right)^2 \right\} \\
& \quad + \Delta_n \left( \frac{s_\omega s^{(3-q)/2} m_\omega^{1-q} (\log p)^{(1-q)/2}}{n^{(1-q)/2}} \right)^2 + \Delta_n \left( \frac{s_\omega s^{3/2} m_\omega^{2-q} (\log p)^{1/2}}{n^{1/2}} \right)^2 \\
& \quad + 2 \left( \frac{s_\omega m_\omega^{2-q} s^{3/2} (\log p)^{1/2}}{n^{1/2}} \right)^2 + \left( \frac{s_\omega s^{(3-q)/2} m_\omega^{1-q} (\log p)^{(1-q)/2}}{n^{(1-q)/2}} \right)^2 \\
& = o_p(Q) + \frac{\Delta_n + 1}{1 + \beta_R^2} \left( 1 + \frac{\|\pi\|_2^2}{\|\gamma\|_2^2} \right) \left[ \left( \frac{s_\omega s^{(3-q)/2} m_\omega^{1-q} (\log p)^{(1-q)/2}}{n^{(1-q)/2}} \right)^2 + \left( \frac{s_\omega s^{3/2} m_\omega^{2-q} (\log p)^{1/2}}{n^{1/2}} \right)^2 \right] \\
& \lesssim o_p(Q) + \frac{\Delta_n^2 + 1}{1 + \beta_R^2} \left[ \left( \frac{s_\omega s^{(3-q)/2} m_\omega^{1-q} (\log p)^{(1-q)/2}}{n^{(1-q)/2}} \right)^2 + \left( \frac{s_\omega s^{3/2} m_\omega^{2-q} (\log p)^{1/2}}{n^{1/2}} \right)^2 \right] \\
& = o_p \left( Q + \frac{(1 \vee \Delta_n) \tau_n^*}{1 + \beta_R^2} + \frac{\Delta_n^2}{1 + \beta_R^2} \right)
\end{aligned} \tag{B39}$$

where the last inequality applies global strength of instruments so that  $\|\pi\|_2^2/\|\gamma\|_2^2 \lesssim \Delta_n$  and  $(1 + \Delta_n)^2 \lesssim 1 + \Delta_n^2$ . Besides, w.p.a.1.

$$(\sqrt{\tau_n} - \sqrt{\tau_n^*})^2 \frac{1}{n} \sum_{i=1}^n \eta_i^2 \sigma_i^2 \lesssim (\sqrt{\tau_n} - \sqrt{\tau_n^*})^2 (1 + \beta_R^2) \|n^{-1/2} \eta\|_2^2 \lesssim (\sqrt{\tau_n} - \sqrt{\tau_n^*})^2 (1 + \beta_R^2).$$

By Lemma B3 we deduce that w.p.a.1

$$\begin{aligned}
|I_1^V| &\lesssim \frac{1}{n} \sum_{i=1}^n [(\hat{u}_3 - \Omega \tilde{\pi}_Z)^\top W_i]^2 \sigma_i^2 + \frac{1}{n} \sum_{i=1}^n [(\tilde{\pi}_Z - \tilde{\pi}_Z)^\top \Omega W_i]^2 \sigma_i^2 + \frac{1}{n} \sum_{i=1}^n (\sqrt{\tau_n} - \sqrt{\tau_n^*})^2 \eta_i^2 \sigma_i^2 \\
&\lesssim (1 + \beta_R^2) [\|\hat{u}_3 - \Omega \tilde{\pi}_Z\|_1^2 + \|\tilde{\pi}_Z - \tilde{\pi}_Z\|_1^2 + (\sqrt{\tau_n} - \sqrt{\tau_n^*})^2] \\
&= o_p[((1 + \beta_R^2) \vee \Delta_n)(Q + \tau_n^*) + \Delta_n^2] = o_p \left[ \left(1 \vee \frac{\Delta_n}{1 + \beta_R^2}\right) V + \Delta_n^2 \right].
\end{aligned}$$

Step II. Bound  $I_2^V$ . we deduce that w.p.a.1

$$\begin{aligned}
|I_2^V| &\lesssim \left| \tilde{\pi}_Z^\top \Omega \frac{1}{n} \sum_{i=1}^n W_i W_i^\top \sigma_i^2 (\hat{u}_3 - \Omega \tilde{\pi}_Z) \right| + \left| \sqrt{\tau_n^*} (\sqrt{\tau_n} - \sqrt{\tau_n^*}) \frac{1}{n} \sum_{i=1}^n \eta_i^2 \sigma_i^2 \right| \\
&\quad + \left| \left[ \sqrt{\tau_n^*} (\hat{u}_3 - \Omega \tilde{\pi}_Z) + (\sqrt{\tau_n} - \sqrt{\tau_n^*}) \Omega \tilde{\pi}_Z \right]^\top \frac{1}{n} \sum_{i=1}^n W_i \eta_i \sigma_i^2 \right| \\
&\lesssim \left| \tilde{\pi}_Z^\top \Omega \frac{1}{n} \sum_{i=1}^n W_i W_i^\top \sigma_i^2 (\hat{u}_3 - \Omega \tilde{\pi}_Z) \right| + \left| \tilde{\pi}_Z^\top \Omega \frac{1}{n} \sum_{i=1}^n W_i W_i^\top \sigma_i^2 \Omega (\tilde{\pi}_Z - \tilde{\pi}_Z) \right| + (1 + \beta_R^2) \tau_n^* \left| \sqrt{\frac{\tau_n}{\tau_n^*}} - 1 \right| \\
&\quad + \sqrt{\tau_n^*} \left| [(\hat{u}_3 - \Omega \tilde{\pi}_Z) + (\tilde{\pi}_Z - \tilde{\pi}_Z)]^\top n^{-1} \sum_{i=1}^n W_i \eta_i \sigma_i^2 \right| + \tau_n^* \cdot \left| \sqrt{\frac{\tau_n}{\tau_n^*}} - 1 \right| \cdot \left| \frac{1}{n} \sum_{i=1}^n \tilde{\pi}_Z^\top \Omega W_i \eta_i \sigma_i^2 \right| \\
&\lesssim (1 + \beta_R^2) \|\tilde{\pi}\|_2 (\|\hat{u}_3 - \Omega \tilde{\pi}_Z\|_1 + \|\tilde{\pi}_Z - \tilde{\pi}_Z\|_2) + (1 + \beta_R^2) \tau_n^* \left| \sqrt{\frac{\tau_n}{\tau_n^*}} - 1 \right| \\
&\quad + \sqrt{\tau_n^*} (\|\hat{u}_3 - \Omega \tilde{\pi}_Z\|_1 + \|\tilde{\pi}_Z - \tilde{\pi}_Z\|_2) + \tau_n^* \cdot \left| \sqrt{\frac{\tau_n}{\tau_n^*}} - 1 \right| \cdot \left( \frac{1}{n} \sum_{i=1}^n (\tilde{\pi}_Z^\top \Omega W_i)^2 \frac{\sigma_i^2}{\tau_n^*} + \frac{1}{n} \sum_{i=1}^n \eta_i^2 \sigma_i^2 \right) \\
&\lesssim (1 + \beta_R^2) (\|\pi\|_2 + \sqrt{\tau_n^*}) (\|\hat{u}_3 - \Omega \tilde{\pi}_Z\|_1 + \|\tilde{\pi}_Z - \tilde{\pi}_Z\|_2) + (1 + \beta_R^2) \tau_n^* \left| \sqrt{\frac{\tau_n}{\tau_n^*}} - 1 \right| \\
&\quad + (1 + \beta_R^2) \left| \sqrt{\frac{\tau_n}{\tau_n^*}} - 1 \right| \cdot (Q + \tau_n^*) \\
&= (1 + \beta_R^2) (\|\pi\|_2 + \sqrt{\tau_n^*}) (\|\hat{u}_3 - \Omega \tilde{\pi}_Z\|_1 + \|\tilde{\pi}_Z - \tilde{\pi}_Z\|_2) + o_p(V) \\
&= o_p \left[ (1 + \beta_R^2) \sqrt{Q + \tau_n^*} \sqrt{Q + \frac{(1 \vee \Delta_n) \tau_n^*}{1 + \beta_R^2} + \frac{\Delta_n^2}{1 + \beta_R^2}} \right] + o_p(V) \\
&= o_p \left[ \left(1 \vee \frac{\Delta_n}{1 + \beta_R^2}\right) (1 + \beta_R^2) (Q + \tau_n^*) + \Delta_n^2 \right] + o_p(V) = o_p \left[ \left(1 \vee \frac{\Delta_n}{1 + \beta_R^2}\right) V + \Delta_n^2 \right].
\end{aligned}$$

where the second equality applies (B39).



Step III. Bound  $I_3^V$ . Note that

$$\begin{aligned}
|I_3^V| &\leq \left| \frac{1}{n} \sum_{i=1}^n (\hat{u}_3^\top W_i) (\hat{e}_i^2 - \sigma_i^2) \right| + \left| \frac{\tau_n}{n} \sum_{i=1}^n \eta_i^2 (\hat{e}_i^2 - \sigma_i^2) \right| + \left| \frac{\sqrt{\tau_n}}{n} \sum_{i=1}^n \hat{u}_3^\top W_i \eta_i (\hat{e}_i^2 - \sigma_i^2) \right| \\
&\leq \|\hat{u}_3\|_1^2 \left\| \frac{1}{n} \sum_{i=1}^n W_i W_i^\top (\hat{e}_i^2 - \sigma_i^2) \right\|_\infty + \tau_n \left| \frac{1}{n} \sum_{i=1}^n \eta_i^2 (\hat{e}_i^2 - \sigma_i^2) \right| + \sqrt{\tau_n} \|\hat{u}_3\|_1 \left\| \frac{2}{n} \sum_{i=1}^n W_i \eta_i (\hat{e}_i^2 - \sigma_i^2) \right\|_\infty \\
&\lesssim (\|\hat{u}_3\|_1^2 + \tau_n^*) \left[ \left\| \frac{1}{n} \sum_{i=1}^n W_i W_i^\top (\hat{e}_i^2 - \sigma_i^2) \right\|_\infty + \left| \frac{1}{n} \sum_{i=1}^n \eta_i^2 (\hat{e}_i^2 - \sigma_i^2) \right| + \left\| \frac{2}{n} \sum_{i=1}^n W_i \eta_i (\hat{e}_i^2 - \sigma_i^2) \right\|_\infty \right].
\end{aligned}$$

We have the following Lemma and the proof is available in [B.4.4](#).

**Lemma B4.** *Under Assumptions 1 to 5, w.p.a.1*

$$\begin{aligned}
&\left\| \frac{1}{n} \sum_{i=1}^n W_i W_i^\top (\hat{e}_i^2 - \sigma_i^2) \right\|_\infty + \left| \frac{1}{n} \sum_{i=1}^n \eta_i^2 (\hat{e}_i^2 - \sigma_i^2) \right| + \left\| \frac{2}{n} \sum_{i=1}^n W_i \eta_i (\hat{e}_i^2 - \sigma_i^2) \right\|_\infty \\
&\lesssim \left( 1 + \frac{\|\pi\|_2^2}{\|\gamma\|_2^2} \right) \left[ \frac{m_\omega s^{1/2} (\log p)^{1/2}}{n^{1/4}} + \frac{s^2 \log p [\log(np)]^2}{n} + \sqrt{\frac{[\log(np)]^3}{n}} \right].
\end{aligned} \tag{B40}$$

Furthermore, w.p.a.1

$$\begin{aligned}
\|\hat{u}_3\|_1^2 &\leq \|\Omega \hat{\pi}_Z\|_1^2 \leq m_\omega^2 (\|\hat{\pi} - \tilde{\pi}\|_1^2 + s \|\tilde{\pi} - \tilde{\pi}\|_2^2 + sQ) \\
&\lesssim m_\omega^2 \left[ \left( 1 + \frac{\|\pi\|_2^2}{\|\gamma\|_2^2} \right) \frac{s^2 \log p}{n} + \|\gamma\|_2^2 (\hat{\beta}_R - \beta_R) + sQ \right] \\
&\lesssim \left( 1 + \frac{\|\pi\|_2^2}{\|\gamma\|_2^2} \right) \frac{m_\omega^2 s^2 \log p}{n} + m_\omega^2 sQ,
\end{aligned}$$

and hence

$$\begin{aligned}
& \|\widehat{u}_3\|_1^2 \left[ \left\| \frac{1}{n} \sum_{i=1}^n W_i W_i^\top (\widehat{e}_i^2 - \sigma_i^2) \right\|_\infty + \left| \frac{1}{n} \sum_{i=1}^n \eta_i^2 (\widehat{e}_i^2 - \sigma_i^2) \right| + \left\| \frac{2}{n} \sum_{i=1}^n W_i \eta_i (\widehat{e}_i^2 - \sigma_i^2) \right\|_\infty \right] \\
& \lesssim \left( 1 + \frac{\|\pi\|_2^4}{\|\gamma\|_2^4} \right) \frac{m_\omega^2 s^2 \log p}{n} \left[ \frac{m_\omega s^{1/2} (\log p)^{1/2}}{n^{1/4}} + \frac{s^2 \log p [\log(np)]^2}{n} + \sqrt{\frac{[\log(np)]^3}{n}} \right] \\
& \quad + m_\omega^2 s Q \left( 1 + \frac{\|\pi\|_2^2}{\|\gamma\|_2^2} \right) \left[ \frac{m_\omega s^{1/2} (\log p)^{1/2}}{n^{1/4}} + \frac{s^2 \log p [\log(np)]^2}{n} + \frac{s [\log(np)]^{3/2}}{\sqrt{n}} \right] \\
& \lesssim (1 + \Delta_n^2) \frac{m_\omega^2 s^2 \log p}{n} \left[ \frac{m_\omega s^{1/2} (\log p)^{1/2}}{n^{1/4}} + \frac{s^2 \log p [\log(np)]^2}{n} + \sqrt{\frac{[\log(np)]^3}{n}} \right] \\
& \quad + \frac{m_\omega^2 s \Delta_n}{\sqrt{n}} (1 + \Delta_n) \left[ \frac{m_\omega s^{1/2} (\log p)^{1/2}}{n^{1/4}} + \frac{s^2 \log p [\log(np)]^2}{n} + \frac{s [\log(np)]^{3/2}}{\sqrt{n}} \right] \\
& \lesssim (1 + \Delta_n^2) \left[ \frac{m_\omega^2 s \log p}{n} + \frac{m_\omega^2 s^2}{\sqrt{n}} \right] \cdot \left[ \frac{m_\omega s^{1/2} (\log p)^{1/2}}{n^{1/4}} + \frac{s^2 \log p [\log(np)]^2}{n} + \sqrt{\frac{[\log(np)]^3}{n}} \right] \\
& = o_p \left[ (1 \vee \Delta_n) \tau_n^* + \Delta_n^2 \right] = o_p \left[ \left( 1 \vee \frac{\Delta_n}{1 + \beta_R^2} \right) V + \Delta_n^2 \right]
\end{aligned}$$

where the first equality applies Assumption 4 (i), (ii) and the fact that  $\left[ \frac{m_\omega^2 s^2 \log p}{n} + \frac{m_\omega^2 s}{\sqrt{n}} \right] = o_p[1 \wedge ((1 \vee \Delta_n) \tau_n^*)]$ . Finally, we deduce that w.p.a.1

$$\begin{aligned}
& \tau_n^* \left[ \left\| \frac{1}{n} \sum_{i=1}^n W_i W_i^\top (\widehat{e}_i^2 - \sigma_i^2) \right\|_\infty + \left| \frac{1}{n} \sum_{i=1}^n \eta_i^2 (\widehat{e}_i^2 - \sigma_i^2) \right| + \left\| \frac{2}{n} \sum_{i=1}^n W_i \eta_i (\widehat{e}_i^2 - \sigma_i^2) \right\|_\infty \right] \\
& \lesssim (\tau_n^* + \Delta_n^2) \left[ \frac{m_\omega s^{1/2} (\log p)^{1/2}}{n^{1/4}} + \frac{s^2 \log p [\log(np)]^2}{n} + \sqrt{\frac{[\log(np)]^3}{n}} \right] \\
& = o_p(\tau_n^* + \Delta_n^2) = o_p(V + \Delta_n^2).
\end{aligned}$$

Combining all results in Step III, we conclude that  $|I_3^V| = o_p \left[ \left( 1 \vee \frac{\Delta_n}{1 + \beta_R^2} \right) V + \Delta_n^2 \right]$ , which completes the proof of Proposition B6.

### B.3.7 Proof of Proposition B7

We first prove (a). Note that any nonrandom variables can be regarded as independent random variables, and sub-Gaussianity is also satisfied since  $\eta_i$  is uniformly bounded. From (24) we directly obtain  $\|\eta\|_2 = \sqrt{n}$ . By Proposition 5.10 of Vershynin (2010) and union bound, we have

for some constant  $c > 0$  that

$$\Pr \left( \left\| \frac{1}{n} [W - \mathbb{E}(W)]^\top \eta \right\|_\infty > \sqrt{\frac{2K^2 \log p}{cn}} \right) \leq p^{-1}.$$

Since  $|n^{-1} \sum_{i=1}^n \eta_i \mathbb{E}(W_{ij})| = |n^{-1}(n \bmod 2) \mathbb{E}(W_{ij})| \lesssim n^{-1}$ , we conclude that  $\left\| \frac{1}{n} W^\top \eta \right\|_\infty \lesssim \sqrt{\log p/n}$  w.p.a.1.

For (b), under the conditions in Proposition B7 it is straightforward to show that Assumption B1 (ii) holds. Besides,  $\tau_n$  defined in (25) converges to zero when  $\Delta_n > 0$ . It hence suffices to show that  $\tau_n/\tau_n^* \xrightarrow{p} 1$ . We have the following decomposition

$$\widehat{Q}^0 - Q = M^0 + B_1 + B_2$$

where  $M^0 = n^{-1} \tilde{\pi}_Z^\top W^\top \tilde{e}$  and  $B_1, B_2$  are specified in (B16) and (B17). Following the same arguments in Steps II and III of the proof for Proposition B5, we can deduce that when Assumption 4 (i) holds,

$$\begin{aligned} \sqrt{n}(B_1 + B_2) &= o_p \left( \frac{1}{\log[\log(np)]} + \sqrt{Q} + \sqrt{n}Q \right) = o_p \left( \frac{1}{\log[\log(np)]} + \frac{\sqrt{\Delta_n}}{n^{1/4}} + \Delta_n \right) \\ &= o_p \left( \frac{1}{\log[\log(np)]} + \sqrt{\frac{\Delta_n}{\log[\log(np)]}} + \Delta_n \right). \end{aligned}$$

We also have

$$\begin{aligned} |\sqrt{n}M^0| &\lesssim \sqrt{n}(1 + |\beta_R|) \|\tilde{\pi}\|_1 [\|n^{-1}W^\top \varepsilon_1\|_\infty + \|n^{-1}W^\top \varepsilon_2\|_\infty] \\ &\lesssim \sqrt{n} \|\tilde{\pi}\|_2 \left( 1 + \frac{\|\pi\|_2}{\|\gamma\|_2} \right) \sqrt{\frac{s \log p}{n}} \\ &\lesssim (\sqrt{\Delta_n} + \Delta_n) \frac{(s \log p)^{1/2}}{n^{1/4}} = o_p \left( \sqrt{\frac{\Delta_n}{\log[\log(np)]}} + \Delta_n \right) \end{aligned}$$

where the last inequality applies global strength of instruments, and the equality applies Assumption 4 (ii).

Case I. When  $\Delta_n \log[\log(np)] \rightarrow 0$ , we have

$$|\sqrt{n}\widehat{Q}^0| \log[\log(np)] \leq \Delta_n \log[\log(np)] + |\sqrt{n}M^0| \log[\log(np)] + \sqrt{n}|B_1 + B_2| \log[\log(np)] = o_p(1).$$

Then  $\tau_n$  and  $\tau_n^*$  have the same probability limit  $\tau_0$  and the conclusion follows.

Case II. When  $\Delta_n \log[\log(np)]$  is bounded away from zero, it suffices to show that  $\sqrt{n}\widehat{Q}^0/\Delta_n \xrightarrow{p}$

1. Noting that

$$\frac{\sqrt{n}(B_1 + B_2)}{\Delta_n} = o_p \left( \frac{1}{\Delta_n \log[\log(np)]} + \sqrt{\frac{1}{\Delta_n \log[\log(np)]}} + 1 \right) = o_p(1)$$

and

$$\frac{|\sqrt{n}M^0|}{\Delta_n} = o_p \left( \sqrt{\frac{1}{\Delta_n \log[\log(np)]}} + 1 \right) = o_p(1).$$

Then the conclusion follows.

## B.4 Proofs of Technical Lemmas

### B.4.1 Proof of Lemma B1

Under Assumption 1, we know that there exists some constants  $C > 0$  and  $v > 0$  such that  $\Pr(|W_{ij}| > t) \leq e^{1-vt^2}$  for all  $t > 0$  and all  $i, j$ . By union bound, we can derive

$$\Pr \left( \max_{i \in [n], j \in [p]} |W_{ij}| > \sqrt{\frac{t + \log(np)}{v}} \right) \leq (np)e^{1-t-\log(np)} = e^{1-t}. \quad (\text{B41})$$

Then (B20) follows by taking  $t = (b+1)\log(np)$ . The proof of (B21) for  $\mathcal{W}_{2n}(b)$  follows entirely the same idea.

Following similar ideas about sub-exponential variables as (B9) by taking  $t = \sqrt{\frac{4bK^4 \log p}{Cn}}$  and union bound we have

$$\Pr \left( \|\hat{\Sigma} - \Sigma\|_\infty > \sqrt{\frac{4(b+1)K^4 \log p}{Cn}} \right) \leq 2p^{-b}$$

and

$$\Pr \left( \|n^{-1}W^\top \eta - n^{-1}\mathbb{E}[W^\top \eta]\|_\infty > \sqrt{\frac{4(b+1)K^4 \log p}{Cn}} \right) \leq 2p^{-b}.$$

Then (B22) and (B23) follow.

### B.4.2 Proof of Lemma B2

This proof follows similar spirits in Cai et al. (2011). By simple inequalities of sup norm, we have

$$\|\hat{u} - \Omega\xi\|_\infty \leq \|(I - \Omega\hat{\Sigma})\hat{u}\|_\infty + \|\Omega(\hat{\Sigma}\hat{u} - \hat{\xi})\|_\infty + \|\Omega(\hat{\xi} - \xi)\|_\infty. \quad (\text{B42})$$

By Assumption 1 and (7.33) in Cai and Guo (2017) we know that w.p.a.1,  $\Omega\hat{\xi}$  belongs to the feasible set of minimizing problem (B27) so that  $\|\hat{u}\|_1 \leq \|\Omega\hat{\xi}\|_1$ , and hence

$$\begin{aligned}
\|(I - \Omega\hat{\Sigma})\hat{u}\|_\infty &\leq \|I - \Omega\hat{\Sigma}\|_\infty \|\hat{u}\|_1 \lesssim \|\Omega\hat{\xi}\|_1 \sqrt{\frac{\log p}{n}} \\
&\leq m_\omega \left( \sqrt{\frac{\log p}{n}} \|\xi\|_1 + \sqrt{\frac{\log p}{n}} \|\hat{\xi} - \xi\|_1 \right) \\
&\lesssim m_\omega \left( \sqrt{\frac{s \log p}{n}} \|\xi\|_2 + s\lambda \sqrt{\frac{\log p}{n}} \right).
\end{aligned} \tag{B43}$$

By the definition of  $\hat{u}$  we have

$$\begin{aligned}
\|\Omega(\hat{\Sigma}\hat{u} - \hat{\xi})\|_\infty &\leq m_\omega \|\hat{\Sigma}\hat{u} - \hat{\xi}\|_\infty \lesssim m_\omega \|\hat{\xi}\|_2 \sqrt{\log p/n} \\
&\lesssim m_\omega \|\xi\|_2 \sqrt{\log p/n} + m_\omega \|\hat{\xi} - \xi\|_2 \sqrt{\log p/n} \lesssim m_\omega \|\xi\|_2 \sqrt{s \log p/n} + m_\omega s\lambda \sqrt{\log p/n}
\end{aligned} \tag{B44}$$

by the fact that  $\|\hat{\xi} - \xi\|_2 \lesssim \sqrt{s \log p/n}$  and  $s \geq \sqrt{s}$ . For the third term,

$$\|\Omega(\hat{\xi} - \xi)\|_\infty \leq m_\omega \|\hat{\xi} - \xi\|_\infty \lesssim m_\omega \sqrt{s\lambda}. \tag{B45}$$

Then (B28) follows (B42) to (B45). To construct (B29), let  $c_n = \|\hat{u} - \Omega\hat{\xi}\|_\infty$  and consider the following quantities

$$\begin{aligned}
u &= \Omega\hat{\xi}, \quad \hat{u}^1 = (\hat{u}_j \cdot \mathbf{1}\{|\hat{u}_j| \geq 2c_n\}, 1 \leq j \leq p)^\top, \\
h &= \hat{u} - u, \quad h^1 = \hat{u}^1 - u, \quad h^2 = h - h^1.
\end{aligned}$$

We first have  $\|\hat{u}\|_1 \leq \|u\|_1$  since  $u = \Omega\hat{\xi}$  belongs to the feasible set. Then

$$\|u\|_1 - \|h^1\|_1 + \|h^2\|_1 \leq \|u + h^1\|_1 + \|h^2\|_1 = \|\hat{u}\|_1 \leq \|u\|_1,$$

which indicates  $\|h^2\|_1 \leq \|h^1\|_1$ . It follows that  $\|h\|_1 \leq 2\|h^1\|_1$ . Therefore, it suffices to find an upper bound of  $\|h^1\|_1$ . Let  $u^0 = \Omega\xi$ . We have

$$\|h^1\|_1 \leq \|\hat{u}^1 - u^0\|_1 + m_\omega \|\hat{\xi} - \xi\|_1 \lesssim \|\hat{u}^1 - u^0\|_1 + m_\omega s\lambda,$$

$$\begin{aligned}
\|\widehat{u}^1 - u^0\|_1 &= \sum_{j=1}^p |\widehat{u}_j \mathbf{1}\{|\widehat{u}_j| \geq 2c_n\} - u_j^0| \\
&\leq \sum_{j=1}^p |u_j^0 \mathbf{1}\{|u_j^0| \leq 2c_n\}| + \sum_{j=1}^p |\widehat{u}_j \mathbf{1}\{|\widehat{u}_j| \geq 2c_n\} - u_j^0 \mathbf{1}\{|u_j^0| \geq 2c_n\}| \\
&\leq \sum_{j=1}^p |u_j^0 \mathbf{1}\{|u_j^0| \leq 2c_n\}| + \sum_{j=1}^p |\widehat{u}_j - u_j| \mathbf{1}\{|\widehat{u}_j| \geq 2c_n\} + \sum_{j=1}^p |u_j^0| \mathbf{1}\{|\widehat{u}_j| \geq 2c_n\} - \mathbf{1}\{|u_j^0| \geq 2c_n\}| \\
&\leq \sum_{j=1}^p |u_j^0 \mathbf{1}\{|u_j^0| \leq 2c_n\}| + c_n \sum_{j=1}^p \mathbf{1}\{|u_j^0| \geq c_n\} + \sum_{j=1}^p |u_j^0| \mathbf{1}\{||u_j^0| - 2c_n| \leq |\widehat{u}_j - u_j^0|\}|
\end{aligned}$$

where we use the following inequality: for any  $a, b, c \in \mathbb{R}$ , we have

$$|I\{a < c\} - I\{b < c\}| \leq I\{|b - c| < |a - b|\}.$$

By the fact that

$$\begin{aligned}
\sum_{j=1}^p |u_j^0|^q &\leq \sum_{j=1}^p \left( \sum_{k=1}^p |\xi_k| |\omega_{jk}| \right)^q \leq \sum_{j=1}^p \sum_{k=1}^p |\xi_k|^q |\omega_{jk}|^q \\
&= \sum_{k=1}^p |\xi_k|^q \cdot s_\omega \leq s_\omega \cdot s \|\xi\|_\infty^q \leq s_\omega \cdot s \|\xi\|_2^q,
\end{aligned}$$

we have for each term of the upper bound derived above,

$$\begin{aligned}
\sum_{j=1}^p |u_j^0 \mathbf{1}\{|u_j^0| \leq 2c_n\}| &\leq (2c_n)^{1-q} \sum_{j=1}^p |u_j^0|^q \leq (2c_n)^{1-q} s_\omega s \|\xi\|_2^q, \\
c_n \sum_{j=1}^p \mathbf{1}\{|u_j^0| \geq c_n\} &\leq (c_n)^{1-q} \sum_{j=1}^p |u_j^0|^q \leq (c_n)^{1-q} s_\omega s \|\xi\|_2^q, \\
\sum_{j=1}^p |u_j^0| \mathbf{1}\{||u_j^0| - 2c_n| \leq |\widehat{u}_j - u_j^0|\} &\leq \sum_{j=1}^p |u_j^0 \mathbf{1}\{|u_j^0| \leq 3c_n\}| \leq (3c_n)^{1-q} s_\omega s \|\xi\|_2^q.
\end{aligned}$$

Then we deduce that

$$\begin{aligned}
\|\widehat{u} - \Omega\xi\|_1 &\leq \|\Omega(\widehat{\xi} - \xi)\|_1 + 2\|h^1\|_1 \lesssim m_\omega s \lambda + s_\omega s \|\xi\|_2^q c_n^{1-q} \\
&\lesssim m_\omega s \lambda + s_\omega s \|\xi\|_2^q (\|\widehat{u} - \Omega\xi\|_\infty)^{1-q},
\end{aligned}$$

which establishes (B29).

### B.4.3 Proof of Lemma B3

By the fact that  $\|\widehat{\Sigma} - \Sigma\|_\infty = o_p(1)$  and  $\|\Sigma\|_\infty$  are bounded by the maximum eigenvalue, we have w.p.a.1

$$\begin{aligned} \|n^{-1} \sum_{i=1}^n W_{i\cdot} W_{i\cdot}^\top \sigma_i^2\|_\infty &= \max_{j,k \in [p]} |n^{-1} \sum_{i=1}^n W_{ij} W_{ik} \sigma_i^2| \lesssim \max_{j \in [p]} |n^{-1} \sum_{i=1}^n W_{ij}^2| (1 + \beta_R^2) \sigma_{\max}^2 \\ &\lesssim \|\widehat{\Sigma}\|_\infty (1 + \beta_R^2) \\ &\lesssim (\|\widehat{\Sigma} - \Sigma\|_\infty + \|\Sigma\|_\infty) (1 + \beta_R^2) \lesssim 1 + \beta_R^2. \end{aligned}$$

Similarly,

$$\begin{aligned} |n^{-1} \sum_{i=1}^n \eta_i^2 \sigma_i^2| &\lesssim n^{-1} \sum_{i=1}^n \eta_i^2 (1 + \beta_R^2) \sigma_{\max}^2 \lesssim 1 + \beta_R^2, \\ \|n^{-1} \sum_{i=1}^n W_{i\cdot} \eta_i \sigma_i^2\|_\infty &= \max_{j \in [p]} |n^{-1} \sum_{i=1}^n W_{ij} \eta_i \sigma_i^2| \lesssim \left[ \max_{j \in [p]} n^{-1} \sum_{i=1}^n W_{ij}^2 + n^{-1} \sum_{i=1}^n \eta_i^2 \right] (1 + \beta_R^2) \sigma_{\max}^2 \\ &\lesssim (\|\widehat{\Sigma}\|_\infty + 1) (1 + \beta_R^2) \lesssim 1 + \beta_R^2. \end{aligned}$$

where the last inequality applies the result that  $\|\widehat{\Sigma}\|_\infty \leq \|\widehat{\Sigma} - \Sigma\|_\infty + \|\Sigma\|_\infty \lesssim 1$ . Then (B37) is resulted from the fact that

$$\begin{aligned} \left| \frac{1}{n} \sum_{i=1}^n \xi_1^\top W_{i\cdot} W_{i\cdot}^\top \xi_2 \sigma_i^2 \right| &\leq \|\xi_1\|_1 \|\xi_2\|_1 \|n^{-1} \sum_{i=1}^n W_{i\cdot} W_{i\cdot}^\top \sigma_i^2\|_\infty, \\ \left| \frac{1}{n} \sum_{i=1}^n \xi_2^\top W_{i\cdot} \eta_i \sigma_i^2 \right| &\leq \|\xi_2\|_1 \|n^{-1} \sum_{i=1}^n W_{i\cdot}^\top \eta_i \sigma_i^2\|_\infty. \end{aligned}$$

As for (B38),

$$\begin{aligned} \left| \frac{1}{n} \sum_{i=1}^n \xi_2^\top \Omega W_{i\cdot} W_{i\cdot}^\top \Omega \xi_2 \sigma_i^2 \right| &\leq (1 + \beta_R^2) \sigma_{\max}^2 \left| \frac{1}{n} \sum_{i=1}^n \xi_2^\top \Omega W_{i\cdot} W_{i\cdot}^\top \Omega \xi_2 \right| \\ &\lesssim (1 + \beta_R^2) [\xi_2^\top \Omega (\widehat{\Sigma} - \Sigma) \Omega \xi_2 + \xi_2^\top \Omega \xi_2] \\ &\lesssim (1 + \beta_R^2) \|\xi_2\|_2^2 \cdot m_\omega^2 \sqrt{\frac{s^2 \log p}{n}} + (1 + \beta_R^2) \|\xi_2\|_2^2 \lesssim (1 + \beta_R^2) \|\xi_2\|_2^2 \end{aligned}$$

and

$$\left| \frac{1}{n} \sum_{i=1}^n \xi_1^\top W_{i\cdot} W_{i\cdot}^\top \Omega \xi_2 \sigma_i^2 \right| \leq \sqrt{\frac{1}{n} \sum_{i=1}^n \xi_1^\top W_{i\cdot} W_{i\cdot}^\top \xi_1 \sigma_i^2} \sqrt{\frac{1}{n} \sum_{i=1}^n \xi_2^\top \Omega W_{i\cdot} W_{i\cdot}^\top \Omega \xi_2 \sigma_i^2}$$

since  $\frac{1}{n} \sum W_i W_i^\top \sigma_i^2$  is positive semi-definite.

#### B.4.4 Proof of Lemma B4

We only show that upper bound of the first term in (B40) since the other two terms can be bounded in the same way by the sub-Gaussianity of  $\eta$  and the conditional independence clarified in Assumption B1 (i).

We first decompose and bound  $\widehat{e}_i^2 - \sigma_i^2$  as

$$\begin{aligned} \widehat{e}_i^2 - \sigma_i^2 &= \widehat{e}_i^2 - \check{e}_i^2 + \check{e}_i^2 - \widetilde{e}_i^2 + \widetilde{e}_i^2 - \sigma_i^2 \\ &= \left[ W_{i\cdot}^\top \begin{pmatrix} \widehat{\varphi} - \check{\varphi} \\ \widehat{\pi} - \check{\pi} \end{pmatrix} + \check{e}_i \right]^2 - \check{e}_i^2 + (\widehat{\beta}_R^2 - \beta_R^2) \varepsilon_{i2}^2 - 2(\widehat{\beta}_R - \beta_R) \varepsilon_{i1} \varepsilon_{i2} + \widetilde{e}_i^2 - \sigma_i^2 \\ &= \left[ W_{i\cdot}^\top \begin{pmatrix} \widehat{\varphi} - \check{\varphi} \\ \widehat{\pi} - \check{\pi} \end{pmatrix} \right]^2 - 2W_{i\cdot}^\top \begin{pmatrix} \widehat{\varphi} - \check{\varphi} \\ \widehat{\pi} - \check{\pi} \end{pmatrix} (\varepsilon_{i1} - \widehat{\beta}_R \varepsilon_{i2}) + (\widehat{\beta}_R^2 - \beta_R^2) (\varepsilon_{i2}^2 - \sigma_{i2}^2) \\ &\quad - 2(\widehat{\beta}_R - \beta_R) (\varepsilon_{i1} \varepsilon_{i2} - \sigma_{i12}) + \widetilde{e}_i^2 - \sigma_i^2 + (\widehat{\beta}_R^2 - \beta_R^2) \sigma_{i2}^2 - 2(\widehat{\beta}_R - \beta_R) \sigma_{i12}. \end{aligned}$$

Then

$$\frac{1}{n} \sum_{i=1}^n W_i W_i^\top (\widehat{e}_i^2 - \sigma_i^2) = \Delta_1^V + \Delta_2^V + \Delta_3^V,$$

where

$$\begin{aligned} \Delta_1^V &= \frac{1}{n} \sum_{i=1}^n W_{i\cdot} \left[ W_{i\cdot}^\top \begin{pmatrix} \widehat{\varphi} - \check{\varphi} \\ \widehat{\pi} - \check{\pi} \end{pmatrix} \right]^2 W_{i\cdot}^\top, \\ \Delta_2^V &= \frac{1}{n} \sum_{i=1}^n W_{i\cdot} W_{i\cdot}^\top \begin{pmatrix} \widehat{\varphi} - \check{\varphi} \\ \widehat{\pi} - \check{\pi} \end{pmatrix} (\varepsilon_{i1} - \widehat{\beta}_R \varepsilon_{i2}) W_{i\cdot}^\top, \\ \Delta_3^V &= \frac{1}{n} \sum_{i=1}^n W_{i\cdot} W_{i\cdot}^\top \left[ (\widehat{\beta}_R^2 - \beta_R^2) (\varepsilon_{i2}^2 - \sigma_{i2}^2) - 2(\widehat{\beta}_R - \beta_R) (\varepsilon_{i1} \varepsilon_{i2} - \sigma_{i12}) + \widetilde{e}_i^2 - \sigma_i^2 \right], \\ \Delta_4^V &= \frac{1}{n} \sum_{i=1}^n W_{i\cdot} W_{i\cdot}^\top \left[ (\widehat{\beta}_R^2 - \beta_R^2) \sigma_{i2}^2 - 2(\widehat{\beta}_R - \beta_R) \sigma_{i12} \right]. \end{aligned}$$



To bound  $\Delta_1^V$ , we apply Proposition B5 and (B20) to deduce that w.p.a.1

$$\begin{aligned}
\|\Delta_1^V\|_\infty &= \left\| \frac{1}{n} \sum_{i=1}^n W_{i\cdot} \left[ W_{i\cdot}^\top \begin{pmatrix} \hat{\varphi} - \check{\varphi} \\ \hat{\pi} - \check{\pi} \end{pmatrix} \right]^2 W_{i\cdot}^\top \right\|_\infty \\
&= \left\| \frac{1}{n} \sum_{i=1}^n \text{vec} \left( W_{i\cdot} W_{i\cdot}^\top \begin{pmatrix} \hat{\varphi} - \check{\varphi} \\ \hat{\pi} - \check{\pi} \end{pmatrix} \begin{pmatrix} \hat{\varphi} - \check{\varphi} \\ \hat{\pi} - \check{\pi} \end{pmatrix}^\top W_{i\cdot} W_{i\cdot}^\top \right) \right\|_\infty \\
&= \left\| \frac{1}{n} \sum_{i=1}^n W_{i\cdot} W_{i\cdot}^\top \otimes W_{i\cdot} W_{i\cdot}^\top \text{vec} \left[ \begin{pmatrix} \hat{\varphi} - \check{\varphi} \\ \hat{\pi} - \check{\pi} \end{pmatrix} \begin{pmatrix} \hat{\varphi} - \check{\varphi} \\ \hat{\pi} - \check{\pi} \end{pmatrix}^\top \right] \right\|_\infty \\
&\leq \left\| \frac{1}{n} \sum_{i=1}^n W_{i\cdot} W_{i\cdot}^\top \otimes W_{i\cdot} W_{i\cdot}^\top \right\|_\infty (\|\hat{\varphi} - \check{\varphi}\|_1 + \|\hat{\pi} - \check{\pi}\|_1)^2 \\
&\lesssim \max_{j,k,\ell,m \in [p]} |W_{ij} W_{ik} W_{i\ell} W_{im}| \cdot \left( 1 + \frac{\|\pi\|_2^2}{\|\gamma\|_2^2} \right) \frac{s^2 \log p}{n} = \left( 1 + \frac{\|\pi\|_2^2}{\|\gamma\|_2^2} \right) \frac{s^2 [\log(np)]^2 \log p}{n}.
\end{aligned}$$

To bound  $\Delta_2^V$  and  $\Delta_3^V$ , we need the following lemma where the proof is available in B.4.5.

**Lemma B5.** *Let  $(\nu_i)_{i=1}^n$  be a sequence of independent centered sub-exponential variables satisfying  $\mathbb{E}[\nu_i | W_i] = 0$  where  $W_i \in \mathbb{R}^p$ ,  $i = 1, 2, \dots, n$  are i.i.d. and sub-Gaussian. For any fixed  $b \geq 0$ , there exists some constant  $C(b)$  such that for any  $(k_1, k_2, \dots, k_b) \in \mathbb{N}_+^b$ ,*

$$\Pr \left( \left| \frac{1}{n} \sum_{i=1}^n W_{ik_1} W_{ik_2} \cdots W_{ik_b} \nu_i \right| > C(b) \sqrt{\frac{(\log(np))^{b+1}}{n}} \right) \leq 2(1+e)(np)^{-(b+1)}. \quad (\text{B46})$$

with  $n$  large enough for any constant  $c > 0$ . Hence, under Assumptions 1 and 2, w.p.a.1.

$$\max_{j,k,h \in [p]} \left| \frac{1}{n} \sum_{i=1}^n W_{ij} W_{ik} W_{ih} \varepsilon_{im} \right| \lesssim \sqrt{\frac{(\log np)^3}{n}} \quad (\text{B47})$$

and

$$\max_{j,k \in [p]} \left| \frac{1}{n} \sum_{i=1}^n W_{ij} W_{ik} \varepsilon_{i\ell} \varepsilon_{im} - \frac{1}{n} \mathbb{E} \left[ \sum_{i=1}^n W_{ij} W_{ik} \varepsilon_{i\ell} \varepsilon_{im} | W \right] \right| \lesssim \sqrt{\frac{(\log np)^3}{n}} \quad (\text{B48})$$

for  $\ell, m = 1, 2$  for  $n$  large enough under an additional assumption  $\log p = o(n)$ .

Then by (B47), Proposition B3 and Assumption 4, we deduce that w.p.a.1

$$\begin{aligned}
\|\Delta_2^V\|_\infty &\leq \left\| \frac{1}{n} \sum_{i=1}^n W_i W_i^\top \begin{pmatrix} \hat{\varphi} - \check{\varphi} \\ \hat{\pi} - \check{\pi} \end{pmatrix} (\varepsilon_{i1} - \beta_R \varepsilon_{i2}) W_i^\top \right\|_\infty + |\hat{\beta}_R - \tilde{\beta}_R| \cdot \left\| \frac{1}{n} \sum_{i=1}^n W_i W_i^\top \begin{pmatrix} \hat{\varphi} - \check{\varphi} \\ \hat{\pi} - \check{\pi} \end{pmatrix} \varepsilon_{i2} W_i^\top \right\|_\infty \\
&\leq \left[ \left\| \frac{1}{n} \sum_{i=1}^n W_i W_i^\top \otimes W_i (\varepsilon_{i1} - \beta_R \varepsilon_{i2}) \right\|_\infty + |\hat{\beta}_R - \tilde{\beta}_R| \cdot \left\| \frac{1}{n} \sum_{i=1}^n W_i W_i^\top \otimes W_i \varepsilon_{i2} \right\|_\infty \right] \cdot \left\| \begin{pmatrix} \hat{\varphi} - \check{\varphi} \\ \hat{\pi} - \check{\pi} \end{pmatrix} \right\|_1 \\
&\lesssim \left[ 1 + |\beta_R| + \left( 1 + \frac{\|\pi\|_2}{\|\gamma\|_2} \right) \frac{1}{\|\gamma\|_2} \sqrt{\frac{m_\omega^2 s \log p}{n}} \right] \cdot \sqrt{\frac{[\log(np)]^3}{n}} \cdot \left( 1 + \frac{\|\pi\|_2}{\|\gamma\|_2} \right) \sqrt{\frac{s^2 \log p}{n}} \\
&\lesssim \left[ 1 + \frac{\|\pi\|_2}{\|\gamma\|_2} + \left( 1 + \frac{\|\pi\|_2}{\|\gamma\|_2} \right) \sqrt{\frac{m_\omega^2 s \log p}{n^{1/2}}} \right] \cdot \sqrt{\frac{[\log(np)]^3}{n}} \cdot \left( 1 + \frac{\|\pi\|_2}{\|\gamma\|_2} \right) \\
&\lesssim \left( 1 + \frac{\|\pi\|_2^2}{\|\gamma\|_2^2} \right) \sqrt{\frac{[\log(np)]^3}{n}}
\end{aligned}$$

with  $n$  large enough.

Besides, from Proposition B3 we deduce that

$$\begin{aligned}
|\hat{\beta}_R^2 - \beta_R^2| &= (\hat{\beta}_R - \beta_R)^2 + 2\beta_R(\hat{\beta}_R - \beta_R) \\
&\lesssim \left[ 1 + \frac{\|\pi\|_2^2}{\|\gamma\|_2^2} \right] \frac{m_\omega^2 s \log p}{\|\gamma\|_2^2 n} + |\beta_R| \cdot \left( 1 + \frac{\|\pi\|_2}{\|\gamma\|_2} \right) \sqrt{\frac{m_\omega^2 s \log p}{\|\gamma\|_2^2 n}} \\
&\lesssim \left[ 1 + \frac{\|\pi\|_2^2}{\|\gamma\|_2^2} \right] \frac{m_\omega^2 s \log p}{n^{1/2}} + \frac{\|\pi\|_2}{\|\gamma\|_2} \cdot \left( 1 + \frac{\|\pi\|_2}{\|\gamma\|_2} \right) \sqrt{\frac{m_\omega^2 s \log p}{n^{1/2}}} \\
&\lesssim \left[ 1 + \frac{\|\pi\|_2^2}{\|\gamma\|_2^2} \right] \frac{m_\omega s^{1/2} (\log p)^{1/2}}{n^{1/4}}
\end{aligned}$$

and  $|\hat{\beta}_R - \beta_R| \leq \left( 1 + \frac{\|\pi\|_2}{\|\gamma\|_2} \right) \frac{1}{\|\gamma\|_2} \sqrt{\frac{m_\omega^2 s \log p}{n}} \lesssim \left( 1 + \frac{\|\pi\|_2^2}{\|\gamma\|_2^2} \right) \frac{m_\omega s^{1/2} (\log p)^{1/2}}{n^{1/4}}$ . Then by (B48)

we have

$$\begin{aligned}
\|\Delta_3^V\|_\infty &\lesssim \left[ |\hat{\beta}_R^2 - \beta_R^2| + |\hat{\beta}_R - \beta_R| + |\beta_R| + 1 \right] \cdot \sqrt{\frac{[\log(np)]^3}{n}} \\
&\lesssim \left[ \left( 1 + \frac{\|\pi\|_2^2}{\|\gamma\|_2^2} \right) \frac{m_\omega s^{1/2} (\log p)^{1/2}}{n^{1/4}} + \frac{\|\pi\|_2}{\|\gamma\|_2} + 1 \right] \sqrt{\frac{[\log(np)]^3}{n}} \\
&\lesssim \left( 1 + \frac{\|\pi\|_2^2}{\|\gamma\|_2^2} \right) \sqrt{\frac{[\log(np)]^3}{n}}.
\end{aligned}$$

Finally, by Assumptions 1 and 2 we deduce that

$$\begin{aligned}
\max \left[ \left\| n^{-1} \sum_{i=1}^n W_i \cdot W_i^\top \sigma_{i2}^2 \right\|_\infty, \left\| n^{-1} \sum_{i=1}^n W_i \cdot W_i^\top \sigma_{i12} \right\|_\infty \right] &\lesssim \max_{j \in [p]} \left[ n^{-1} \sum_{i=1}^n W_{ij}^2 \right] \\
&\leq \|\widehat{\Sigma}\|_\infty \\
&\leq \|\widehat{\Sigma} - \Sigma\|_\infty + \|\Sigma\|_\infty \lesssim 1
\end{aligned}$$

and hence,

$$\begin{aligned}
\|\Delta_3^V\|_\infty &\lesssim |\widehat{\beta}_R^2 - \beta_R^2| + |\widehat{\beta}_R - \beta_R| \\
&\lesssim \left( 1 + \frac{\|\pi\|_2^2}{\|\gamma\|_2^2} \right) \frac{m_\omega s^{1/2} (\log p)^{1/2}}{n^{1/4}}.
\end{aligned}$$

Then we complete the proof of Proposition B6 by summing up the upper bounds of  $\Delta_1^V$ ,  $\Delta_2^V$ ,  $\Delta_3^V$  and  $\Delta_4^V$ .

#### B.4.5 Proof of Lemma B5

Since  $\nu_i$  is sub-exponential and conditionally centered, we have for any fixed  $k = (k_1, k_2, \dots, k_b)$  with  $b \geq 1$

$$\mathbb{E} \left[ \sum_{i=1}^n W_{ik_1} W_{ik_2} \cdots W_{ik_b} \nu_i \mathbf{1}(\mathcal{W}_{1n}(b)) \right] = \mathbb{E} \left[ \sum_{i=1}^n W_{ik_1} W_{ik_2} \cdots W_{ik_b} \mathbf{1}(\mathcal{W}_{1n}(b)) \mathbb{E}[\nu_i | W] \right] = 0,$$

which implies that

$$\begin{aligned}
\mathbb{E} \left[ \exp \left( \lambda \cdot \sum_{i=1}^n W_{ik_1} W_{ik_2} \cdots W_{ik_b} \nu_i \right) \mathbf{1}(\mathcal{W}_{1n}(b)) \right] &\leq 1 + \sum_{d=2}^{\infty} \frac{1}{d!} \mathbb{E} \left[ |\lambda|^d \left| \sum_{i=1}^n W_{ik_1} W_{ik_2} \cdots W_{ik_b} \nu_i \right|^d \mathbf{1}(\mathcal{W}_{1n}(b)) \right] \\
&\leq 1 + \sum_{d=2}^{\infty} \frac{1}{d!} \mathbb{E} \left[ |\lambda|^d \sum_{i=1}^n |\nu_i|^d \left( \sqrt{\frac{(b+2) \log(np)}{v}} \right)^{bd} \right] \\
&\leq 1 + n \sum_{d=2}^{\infty} (2eK^2 \lambda)^d \left( \sqrt{\frac{(b+2) \log(np)}{v}} \right)^{bd} \\
&\leq 1 + 8e^2 n K^4 \lambda^2 \left( \frac{(b+2) \log(np)}{v} \right)^b \\
&\leq \exp \left[ 8e^2 n K^4 \lambda^2 \left( \frac{(b+2) \log(np)}{v} \right)^b \right]
\end{aligned}$$

when  $0 < \lambda \leq 1/\left[eK^2\left(\frac{(b+2)\log(np)}{v}\right)^{b/2}\right]$ , where the third inequality applies  $d! \geq (d/e)^d$  the fact that the sup-exponential norm of  $\nu_i$  is bounded by  $2K^2$ . As a consequence,

$$\begin{aligned}
& \Pr\left\{\sum_{i=1}^n W_{ik_1} W_{ik_2} \cdots W_{ik_b} \nu_i > t, \mathcal{W}_{1n}(b)\right\} \\
&= \Pr\left\{\exp\left(\lambda \cdot \sum_{i=1}^n W_{ik_1} W_{ik_2} \cdots W_{ik_b} \nu_i\right) > \exp(\lambda t) \mid \mathcal{W}_{1n}(b)\right\} \Pr(\mathcal{W}_{1n}(b)) \\
&\leq e^{-\lambda t} \mathbb{E}\left[\exp\left(\lambda \cdot \sum_{i=1}^n W_{ik_1} W_{ik_2} \cdots W_{ik_b} \nu_i\right) \mid \mathcal{W}_{1n}(b)\right] \Pr(\mathcal{W}_{1n}(b)) \\
&= e^{-\lambda t} \mathbb{E}\left[\exp\left(\lambda \cdot \sum_{i=1}^n W_{ik_1} W_{ik_2} \cdots W_{ik_b} \nu_i\right) \mathbf{1}(\mathcal{W}_{1n}(b))\right] \\
&\leq \exp\left[-\lambda t + 8e^2 n K^4 \lambda^2 \left(\frac{(b+2)\log(np)}{v}\right)^b\right]
\end{aligned}$$

where the first inequality applies Markov inequality. Choosing

$$\lambda = \min\left\{\frac{t}{16e^2 n K^4} \left(\frac{v}{(b+2)\log(np)}\right)^b, \frac{v^{b/2}}{eK^2((b+2)\log(np))^{b/2}}\right\},$$

we deduce that

$$\begin{aligned}
& \Pr\left\{\sum_{i=1}^n W_{ik_1} W_{ik_2} \cdots W_{ik_b} \nu_i > t\right\} \\
&\leq \Pr\left\{\sum_{i=1}^n W_{ik_1} W_{ik_2} \cdots W_{ik_b} \nu_i > t, \mathcal{W}_{1n}(b)\right\} + \Pr(\mathcal{W}_{1n}(b)^c) \\
&\leq \exp\left[-\min\left(\frac{t^2}{16e^2 n K^4} \left(\frac{v}{(b+2)\log(np)}\right)^b, \frac{v^{b/2} t}{eK^2((b+2)\log(np))^{b/2}}\right)\right] + b \cdot e(np)^{-(b+1)}.
\end{aligned}$$

Then (B46) is verified by taking  $t = \sqrt{\frac{16e^2 n K^4 (b+2)^b (b+1) [\log(np)]^{b+1}}{v^b}}$  and repeating the same arguments for  $-\nu_i$ . (B47) follows by taking  $b = 2$ ,  $\nu_i = W_{ih} \varepsilon_{im}$  and union bound of  $p^3$  events. (B48) follows by taking  $b = 2$ ,  $\nu_i = \varepsilon_{il} \varepsilon_{im} - \mathbb{E}[\varepsilon_{il} \varepsilon_{im} | W_i]$  and union bound of  $p^2$  events.

## B.5 Proofs of Propositions in Section A

### B.5.1 Proof of Proposition A1

First we show that  $S \xrightarrow{p} \Theta$  so that  $S^{-1} \xrightarrow{p} \Theta^{-1}$  by positive definiteness. Note that the  $(j, k)$ -th element of  $S$  is given by  $S_{jk} = (n - p_z)^{-1} \varepsilon_j^\top M \varepsilon_k$  for  $j, k = 1, 2$ . The result is implied by the fact that  $\mathbb{E}(S_{jk}) = (n - p_z)^{-1} \Theta_{jk} \text{tr}(M) = \Theta_{jk}$  and  $\text{Var}(S_{jk})$

Next, note that  $m_{\min} - p_z/n = \lambda_{\min}(S^{-1}T - (p_z/n)I_2) = \lambda_{\min}(S^{-1}(T - (p_z/n)S))$ . Matrix  $T - (p_z/n)S$  can be decomposed as

$$\begin{aligned} T - \frac{p_z}{n}S &= \begin{pmatrix} \Gamma^\top \widehat{\Sigma} \Gamma & \Gamma^\top \widehat{\Sigma} \gamma \\ \Gamma^\top \widehat{\Sigma} \gamma & \gamma^\top \widehat{\Sigma} \gamma \end{pmatrix} + \frac{1}{n} \begin{pmatrix} \varepsilon_1^\top H \varepsilon_1 & \varepsilon_1^\top H \varepsilon_2 \\ \varepsilon_1^\top H \varepsilon_2 & \varepsilon_2^\top H \varepsilon_2 \end{pmatrix} + \frac{1}{n} \begin{pmatrix} 2\Gamma^\top Z^\top \varepsilon_1 & \Gamma^\top Z^\top \varepsilon_2 + \gamma^\top Z^\top \varepsilon_1 \\ \gamma^\top Z^\top \varepsilon_1 + \Gamma^\top Z^\top \varepsilon_2 & 2\gamma^\top Z^\top \varepsilon_2 \end{pmatrix} \\ &\equiv O_1 + O_2 + O_3 \end{aligned}$$

with  $H = P - \frac{p_z}{n - p_z} M$ . Hence,

$$\begin{aligned} m_{\min} - \frac{p_z}{n} &= \lambda_{\min}(S^{-1/2}(O_1 + O_2 + O_3)S^{-1/2}) \\ &\leq \lambda_{\min}(S^{-1/2}O_1S^{-1/2}) + \lambda_{\max}(S^{-1/2}(O_2 + O_3)S^{-1/2}) \\ &= \lambda_{\min}(S^{-1}O_1) + \lambda_{\max}(S^{-1/2}(O_2 + O_3)S^{-1/2}) \end{aligned}$$

by the inequalities<sup>2</sup>  $\lambda_{\min}(A + B) \leq \lambda_{\min}(A) + \lambda_{\max}(B)$  and  $\lambda_{\max}(A + B) \leq \lambda_{\max}(A) + \lambda_{\max}(B)$  for any symmetric square matrices  $A$  and  $B$ . Before moving on, we first introduce an easily verified lemma.

**Lemma B6.** *The eigenvalues of a  $2 \times 2$  matrix  $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$  are given by*

$$\lambda_{1,2}(A) = \frac{A_{11} + A_{22} \pm \sqrt{(A_{11} - A_{22})^2 + 4A_{12}A_{21}}}{2}.$$

This lemma implies that

$$\frac{n}{\sqrt{p_z}} \lambda_{\min}(S^{-1}O_1) = \frac{n}{2\sqrt{p_z}\det(S)} \left[ S_{22}\Gamma^\top \widehat{\Sigma} \Gamma + S_{11}\gamma^\top \widehat{\Sigma} \gamma - 2S_{12}\gamma^\top \widehat{\Sigma} \Gamma - \sqrt{\mathcal{C}} \right]$$

where

$$\mathcal{C} = (S_{22}\Gamma^\top \widehat{\Sigma} \Gamma - S_{11}\gamma^\top \widehat{\Sigma} \gamma)^2 + 4(S_{22}\gamma^\top \widehat{\Sigma} \Gamma - S_{12}\gamma^\top \widehat{\Sigma} \gamma)(S_{11}\gamma^\top \widehat{\Sigma} \Gamma - S_{12}\Gamma^\top \widehat{\Sigma} \Gamma).$$

---

<sup>2</sup>These inequalities are implied by the convexity of function  $\lambda_{\max}(\cdot)$  and the concavity of  $\lambda_{\min}(\cdot)$ .

We have  $\mathcal{C} > 0$  since  $S$  is positive definite, and it is easy to verify that  $O_1$  is positive semi-definite. The positive definiteness also implies  $S_{22}\Gamma^\top \widehat{\Sigma}\Gamma + S_{11}\gamma^\top \widehat{\Sigma}\gamma - 2S_{12}\gamma^\top \widehat{\Sigma}\Gamma \geq 0$ . Then

$$\begin{aligned} \frac{n}{\sqrt{p_z}} \lambda_{\min}(S^{-1}O_1) &= \frac{4n}{2\sqrt{p_z}\det(S)} \left[ \frac{(S_{11}S_{22} - S_{12}^2)(\Gamma^\top \widehat{\Sigma}\Gamma \cdot \gamma^\top \widehat{\Sigma}\gamma - (\gamma^\top \widehat{\Sigma}\Gamma)^2)}{S_{22}\Gamma^\top \widehat{\Sigma}\Gamma + S_{11}\gamma^\top \widehat{\Sigma}\gamma - 2S_{12}\gamma^\top \widehat{\Sigma}\Gamma + \sqrt{\mathcal{C}}} \right] \\ &\leq \frac{2n}{\sqrt{p_z}} \frac{\gamma^\top \widehat{\Sigma}\gamma \cdot Q_{\widehat{\Sigma}}}{(S_{11} + \beta^2 S_{22} - 2\beta S_{12})\gamma^\top \widehat{\Sigma}\gamma + S_{22}\pi^\top \widehat{\Sigma}\pi + 2\beta(S_{22} - S_{12})\pi^\top \widehat{\Sigma}\gamma} \\ &\xrightarrow{p} \frac{2\Delta}{(\Theta_{11} + \beta^2 \Theta_{22} - 2\beta \Theta_{12})}, \end{aligned}$$

where the limit applies the fact that  $\pi^\top \widehat{\Sigma}\pi/\gamma^\top \widehat{\Sigma}\gamma \leq [(1-r_0^2)Q_\gamma]^{-1}\Delta/\sqrt{p_z} \rightarrow 0$  and  $|\pi^\top \widehat{\Sigma}\gamma|/\gamma^\top \widehat{\Sigma}\gamma \leq \sqrt{\pi^\top \widehat{\Sigma}\pi/\gamma^\top \widehat{\Sigma}\gamma} \rightarrow 0$ . Noting that by positive definiteness we have  $\Theta_{11} + \beta^2 \Theta_{22} - 2\beta \Theta_{12} > \Theta_{11} + \beta^2 \Theta_{22} - 2|\beta|\sqrt{\Theta_{11}\Theta_{22}} \geq 0$ , we know that there exists a positive constant  $C = 2/(\Theta_{11} + \beta^2 \Theta_{22} - 2\beta \Theta_{12})$  such that

$$\Pr \left( \frac{n}{\sqrt{p_z}} \lambda_{\min}(S^{-1}O_1) \leq 2C\Delta \right) \rightarrow 1.$$

Thus, it suffices to show that  $\liminf_{n \rightarrow \infty} \Pr \left( \frac{n}{\sqrt{p_z}} \lambda_{\min}[S^{-1/2}(O_2 + O_3)S^{-1/2}] \leq z - 2C\Delta \right) > 0$  for any  $z > 0$ .

Let  $U_n = (\varepsilon_1, \varepsilon_2)$ ,  $P_n = H/\sqrt{p_z}$  and  $M_n = (Z\Gamma, Z\gamma)$ . When  $Q_{\widehat{\Sigma}} = \sqrt{p_z}\Delta/n$  with any fixed  $\Delta > 0$ , we have  $M_n^\top P_n P_n M_n = (n/p_z)O_1 \rightarrow \Lambda = Q_\gamma \begin{pmatrix} \beta^2 & \beta \\ \beta & 1 \end{pmatrix}$ . By Lemma A.2 in [Kolesár \(2018\)](#), we know that

$$\frac{n}{\sqrt{p_z}} \text{vec}(O_2 + O_3) \xrightarrow{d} N \left( 0, 2N_2 \left( \frac{1}{1 - \alpha_z} \Theta \otimes \Theta + \Theta \otimes \Lambda + \Lambda \otimes \Theta \right) \right)$$

where  $N_2$  is a  $4 \times 4$  symmetrizer matrix<sup>3</sup> Then

$$\frac{n}{\sqrt{p_z}} \text{vec}(S^{-1/2}(O_2 + O_3)S^{-1/2}) = \frac{n}{\sqrt{p_z}} (S^{-1/2} \otimes S^{-1/2}) \text{vec}(O_2 + O_3) \xrightarrow{d} \text{vec}\mathcal{O} = \text{vec}(\mathcal{O}_{ij})_{i,j=1,2},$$

where  $\mathcal{O}$  is a  $2 \times 2$  random matrix satisfying

$$\text{vec}\mathcal{O} \sim N \left( 0, 2N_2 \left( \frac{1}{1 - \alpha_z} I_4 + I_2 \otimes (\Theta^{-1/2} \Lambda \Theta^{-1/2}) + (\Theta^{-1/2} \Lambda \Theta^{-1/2}) \otimes I_2 \right) \right) \quad (\text{B49})$$

---

<sup>3</sup>See Definition 2.2a, 2.2b and Lemma 2.1 about matrix  $N$  in [Magnus and Neudecker \(1980\)](#).

applying Lemma 2.1 (v) of Magnus and Neudecker (1980). Note that the  $4 \times 4$  asymptotic covariance matrix in (B49) has rank 3, and its second and third rows (and also columns) are identical. Therefore, we know that  $\mathcal{O}_{11}$ ,  $\mathcal{O}_{12}$  and  $\mathcal{O}_{22}$  are jointly normal with a positive definite covariance matrix. Then by Lemma B6 and the inequality  $\sqrt{a^2 + b^2} \leq |a| + |b|$ , we have for any  $z \in \mathbb{R}$ ,

$$\begin{aligned}
& \lim_{n \rightarrow \infty} \Pr \left( \frac{n}{\sqrt{p_z}} \lambda_{\max}(S^{-1/2}(O_2 + O_3)S^{-1/2}) \leq z - 2C\Delta \right) \\
&= \Pr \left( \mathcal{O}_{11} + \mathcal{O}_{22} + \sqrt{|\mathcal{O}_{11} - \mathcal{O}_{22}|^2 + 4\mathcal{O}_{12}^2} \leq 2z - 4C\Delta \right) \\
&\geq \Pr(\mathcal{O}_{11} + \mathcal{O}_{22} + |\mathcal{O}_{11} - \mathcal{O}_{22}| \leq 2z - 2 - 4C\Delta, |\mathcal{O}_{12}| < 1) \\
&= \Pr(\max\{\mathcal{O}_{11}, \mathcal{O}_{22}\} \leq 2z - 2 - 4C\Delta, |\mathcal{O}_{12}| < 1) > 0.
\end{aligned}$$

This completes the proof of Proposition A1.

### B.5.2 Proof of Proposition A2

By decomposition of the test statistic we have

$$\begin{aligned}
\widehat{v}^\top (P - P^{\text{diag}}) \widehat{v} &= (\pi + \gamma(\beta - \widehat{\beta}_{\text{JIVE}}))^\top Z^\top (I - P^{\text{diag}}) Z (\pi + \gamma(\beta - \widehat{\beta}_{\text{JIVE}})) \\
&\quad + (\varepsilon_1 - \widehat{\beta}_{\text{JIVE}} \varepsilon_2)^\top (P - P^{\text{diag}}) (\varepsilon_1 - \widehat{\beta}_{\text{JIVE}} \varepsilon_2) + \\
&\quad 2(\varepsilon_1 - \widehat{\beta}_{\text{JIVE}} \varepsilon_2)^\top (I - P^{\text{diag}}) Z (\pi + \gamma(\beta - \widehat{\beta}_{\text{JIVE}}))
\end{aligned}$$

where

$$\begin{aligned}
\widehat{\beta}_{\text{JIVE}} &= \frac{D^\top (P - P^{\text{diag}}) Y}{D^\top (P - P^{\text{diag}}) D} \\
&= \beta + \frac{D^\top (P - P^{\text{diag}}) (Z\pi + e)}{D^\top (P - P^{\text{diag}}) D} \\
&= \beta + \frac{\gamma^\top [Z^\top (I - P^{\text{diag}}) Z] \pi + \varepsilon_2^\top (P - P^{\text{diag}}) e + \varepsilon_2^\top (I - P^{\text{diag}}) Z \pi + e^\top (I - P^{\text{diag}}) Z \gamma}{\gamma^\top [Z^\top (I - P^{\text{diag}}) Z] \gamma + \varepsilon_2^\top (P - P^{\text{diag}}) \varepsilon_2 + 2\varepsilon_2^\top (I - P^{\text{diag}}) Z \gamma}.
\end{aligned}$$

Note that  $\gamma^\top Z^\top (I - P^{\text{diag}}) Z \gamma \geq (1 - C_P) p_z Q_\gamma$ . By Lemma A.1 of Chao et al. (2014) we have  $\varepsilon_2^\top (P - P^{\text{diag}}) e = O_p(\sqrt{p_z}) = o_p(\gamma^\top Z^\top (I - P^{\text{diag}}) Z \gamma)$  and similarly  $e^\top (P - P^{\text{diag}}) e, \varepsilon_2^\top (P - P^{\text{diag}}) \varepsilon_2 = O_p(\sqrt{p_z}) = o_p(\gamma^\top Z^\top (I - P^{\text{diag}}) Z \gamma)$ . Besides, by normality we know that  $p_z^{-1/2} \gamma^\top Z^\top (I - P^{\text{diag}}) \varepsilon_2 \sim N(0, p_z^{-1} \Theta_{22} \gamma^\top Z^\top (I - P^{\text{diag}}) Z \gamma) = O_p(p_z^{-1/2} \sqrt{\gamma^\top Z^\top (I - P^{\text{diag}}) Z \gamma}) = O_p(1)$  and hence  $\gamma^\top Z^\top (I - P^{\text{diag}}) \varepsilon_2 = o_p(\gamma^\top Z^\top (I - P^{\text{diag}}) Z \gamma)$ . Similarly we can show that  $\sqrt{p_z} e^\top (P - P^{\text{diag}}) Z \gamma = O_p(\gamma^\top Z^\top (I - P^{\text{diag}}) Z \gamma)$  and  $\sqrt{p_z} e^\top (P - P^{\text{diag}}) Z \pi = O_p(\gamma^\top Z^\top (I - P^{\text{diag}}) Z \gamma)$  as

$Q_{\hat{\Sigma}} = \sqrt{p_z}\Delta/n$ . Consequently,

$$\hat{\beta}_{\text{JIVE}} = \beta + \frac{\gamma^\top Z^\top (1 - P^{\text{diag}}) Z \pi}{\gamma^\top Z^\top (1 - P^{\text{diag}}) Z \gamma} + O_p(p_z^{-1/2}) = b_{\text{JIVE}} + O_p(p_z^{-1/2}).$$

where  $b_{\text{JIVE}} = \frac{\gamma^\top Z^\top (1 - P^{\text{diag}}) Z \Gamma}{\gamma^\top Z^\top (1 - P^{\text{diag}}) Z \gamma}$ . Then

$$\begin{aligned} \frac{\hat{v}^\top (P - P^{\text{diag}}) \hat{v}}{\sqrt{p_z}} &= \frac{\pi^\top Z^\top (I - P^{\text{diag}}) Z \pi}{\sqrt{p_z}} \left( 1 - \frac{[\pi^\top Z^\top (I - P^{\text{diag}}) Z \gamma]^2}{\pi^\top Z^\top (I - P^{\text{diag}}) Z \pi \cdot \gamma^\top Z^\top (I - P^{\text{diag}}) Z \gamma} \right) \\ &\quad + \left( \varepsilon_1 - \frac{\gamma^\top Z^\top (I - P^{\text{diag}}) Z \Gamma}{\gamma^\top Z^\top (I - P^{\text{diag}}) Z \gamma} \varepsilon_2 \right)^\top \frac{P - P^{\text{diag}}}{\sqrt{p_z}} \left( \varepsilon_1 - \frac{\gamma^\top Z^\top (I - P^{\text{diag}}) Z \Gamma}{\gamma^\top Z^\top (I - P^{\text{diag}}) Z \gamma} \varepsilon_2 \right) \\ &\quad + 2 \left( \pi - \frac{\gamma^\top Z^\top (I - P^{\text{diag}}) Z \pi}{\gamma^\top Z^\top (I - P^{\text{diag}}) Z \gamma} \gamma \right)^\top Z^\top \frac{I - P^{\text{diag}}}{\sqrt{p_z}} \left( \varepsilon_1 - \frac{\gamma^\top Z^\top (I - P^{\text{diag}}) Z \Gamma}{\gamma^\top Z^\top (I - P^{\text{diag}}) Z \gamma} \varepsilon_2 \right) \\ &\quad + O_p(p_z^{-1/2}) \\ &\equiv R_1 + R_2 + R_3 + O_p(p_z^{-1/2}) \end{aligned}$$

where  $R_1 \asymp \Delta$ . By Lemma A.1 of [Chao et al. \(2014\)](#) we know that  $(\Xi)^{-1} R_2 \xrightarrow{d} N(0, 1)$  where  $\Xi = p_z^{-1} \sum_{i \neq j} P_{ij}^2 (\Theta_{11} + b_{\text{JIVE}}^2 \Theta_{22} - 2\Theta_{12} b_{\text{JIVE}}) \geq (\Theta_{11} + b_{\text{JIVE}}^2 \Theta_{22} - 2\Theta_{12} b_{\text{JIVE}})(1 - C_P^2)$  by the fact that

$$\begin{aligned} \sum_{i \neq j} P_{ij}^2 &= \sum_{i,j=1}^n P_{ij}^2 - \sum_{i=1}^n P_{ii}^2 = \text{tr}(P^2) - \sum_{i=1}^n P_{ii}^2 \\ &= \text{tr}(P) - \sum_{i=1}^n P_{ii}^2 \geq p_z - p_z C_P^2. \end{aligned}$$

Then  $\limsup_{n \rightarrow \infty} \Pr(R_2 > z_0) < 1$  for any  $z_0 \in \mathbb{R}$ . In addition, given  $Q_{\hat{\Sigma}} = \sqrt{p_z}\Delta/n$  and normality of  $(\varepsilon_{i1}, \varepsilon_{i2})^\top$  it is easy to show that  $R_3 = o_p(1)$ . It hence suffices to show that  $\hat{S}$  is bounded away from zero. Define  $v_i = Y_i - D_i b_{\text{JIVE}} = D_i(\beta - b_{\text{JIVE}}) + Z_i^\top \pi + e_i$ . Noting that

$$\begin{aligned} |\hat{v}_i^2 - v_i^2| &= \left| \left( D_i(\beta - \hat{\beta}_{\text{JIVE}}) + Z_i^\top \pi + e_i \right)^2 - \left( D_i(\beta - b_{\text{JIVE}}) + Z_i^\top \pi + e_i \right)^2 \right| \\ &= D_i^2 \left| (\beta - \hat{\beta}_{\text{JIVE}})^2 - (\beta - b_{\text{JIVE}})^2 \right| + \left| D_i(Z_i^\top \pi + e_i)(b - \hat{\beta}_{\text{JIVE}}) \right| \\ &\leq C[D_i^2 + (Z_i^\top \pi + e_i)^2] \cdot |\hat{\beta}_{\text{JIVE}} - b_{\text{JIVE}}| \end{aligned}$$



for some constant  $C$  dependent on  $\beta$ ,  $Q_{\hat{\Sigma}}$  and  $Q_{\gamma}$ . By Assumption A2 (ii) we know that  $\mathbb{E}[D_i^2 + (Z_{i\cdot}^\top \pi + e_i)^2] = O(1)$ . Then by sampling independence,

$$\begin{aligned} p_z^{-1} \mathbb{E} \left[ \sum_{i \neq j} P_{ij}^2 (D_i^2 + (Z_{i\cdot}^\top \pi + e_i)^2) (D_j^2 + (Z_{j\cdot}^\top \pi + e_j)^2) \right] &= O(1) \cdot p_z^{-1} \sum_{i \neq j} P_{ij}^2 \\ &= O(1) \cdot p_z^{-1} \left[ \text{tr}(P^2) - \sum_{i=1}^n P_{ii}^2 \right] \\ &\leq O(1) \cdot p_z^{-1} \sum_{i=1}^n P_{ii} = O(1). \end{aligned}$$

and similarly  $p_z^{-1} \mathbb{E} \left[ \sum_{i \neq j} P_{ij}^2 (D_i \beta + Z_{i\cdot}^\top \pi + e_i)^2 v_j^2 \right] = O(1)$ , we know by Markov inequality that

$$p_z^{-1} \sum_{i \neq j} P_{ij}^2 (D_i^2 + (Z_{i\cdot}^\top \pi + e_i)^2) (D_j^2 + (Z_{j\cdot}^\top \pi + e_j)^2) = O_p(1)$$

and

$$p_z^{-1} \sum_{i \neq j} P_{ij}^2 (D_i \beta + Z_{i\cdot}^\top \pi + e_i)^2 v_j^2 = O_p(1).$$

Since  $|\hat{\beta}_{\text{JIVE}} - b_{\text{JIVE}}| = O_p(p_z^{-1/2})$ , we deduce that

$$\begin{aligned} \left| p_z^{-1} \sum_{i \neq j} P_{ij}^2 \hat{v}_i^2 \hat{v}_j^2 - p_z^{-1} \sum_{i \neq j} P_{ij}^2 v_i^2 v_j^2 \right| &\leq p_z^{-1} \sum_{i \neq j} P_{ij}^2 |\hat{v}_i^2 \hat{v}_j^2 - v_i^2 v_j^2| \\ &\leq p_z^{-1} \sum_{i \neq j} P_{ij}^2 (|\hat{v}_i^2 - v_i^2| \cdot |\hat{v}_j^2 - v_j^2| + v_i^2 \cdot |\hat{v}_j^2 - v_j^2| + v_j^2 |\hat{v}_i^2 - v_i^2|) \\ &\leq C^2 p_z^{-1} \sum_{i \neq j} P_{ij}^2 [(D_i^2 + Z_{i\cdot}^\top \pi + e_i)^2 (D_j^2 + Z_{j\cdot}^\top \pi + e_j)^2] \cdot (\hat{\beta}_{\text{JIVE}} - b_{\text{JIVE}})^2 \\ &\quad + 2C p_z^{-1} \sum_{i \neq j} P_{ij}^2 (D_i^2 + Z_{i\cdot}^\top \pi + e_i)^2 x_j^2 \cdot |\hat{\beta}_{\text{JIVE}} - b_{\text{JIVE}}| = o_p(1). \end{aligned}$$

We finally show that  $p_z^{-1} \sum_{i \neq j} P_{ij}^2 v_i^2 v_j^2$  is bounded away from zero. Let  $d_i = v_i^2 - \mathbb{E}[v_i^2]$ . Then

$$p_z^{-1} \sum_{i \neq j} P_{ij}^2 v_i^2 v_j^2 - p_z^{-1} \sum_{i \neq j} P_{ij}^2 \mathbb{E}[v_i^2] \mathbb{E}[v_j^2] = p_z^{-1} \sum_{i \neq j} P_{ij}^2 d_i d_j + 2p_z^{-1} \sum_{i \neq j} P_{ij}^2 d_i v_j^2.$$

By Assumption A2 and normality of the error terms we know that  $\mathbb{E}[d_i^2] \lesssim \mathbb{E}[v_i^4] = O(1)$  and  $\mathbb{E}[v_i^2] = O(1)$ . Then following the same arguments in the proof of Lemma A3 of Chao et al.

(2014), we have

$$p_z^{-1} \sum_{i \neq j} P_{ij}^2 d_i d_j = o_p(1), \quad p_z^{-1} \sum_{i \neq j} P_{ij}^2 d_i v_j^2 = o_p(1).$$

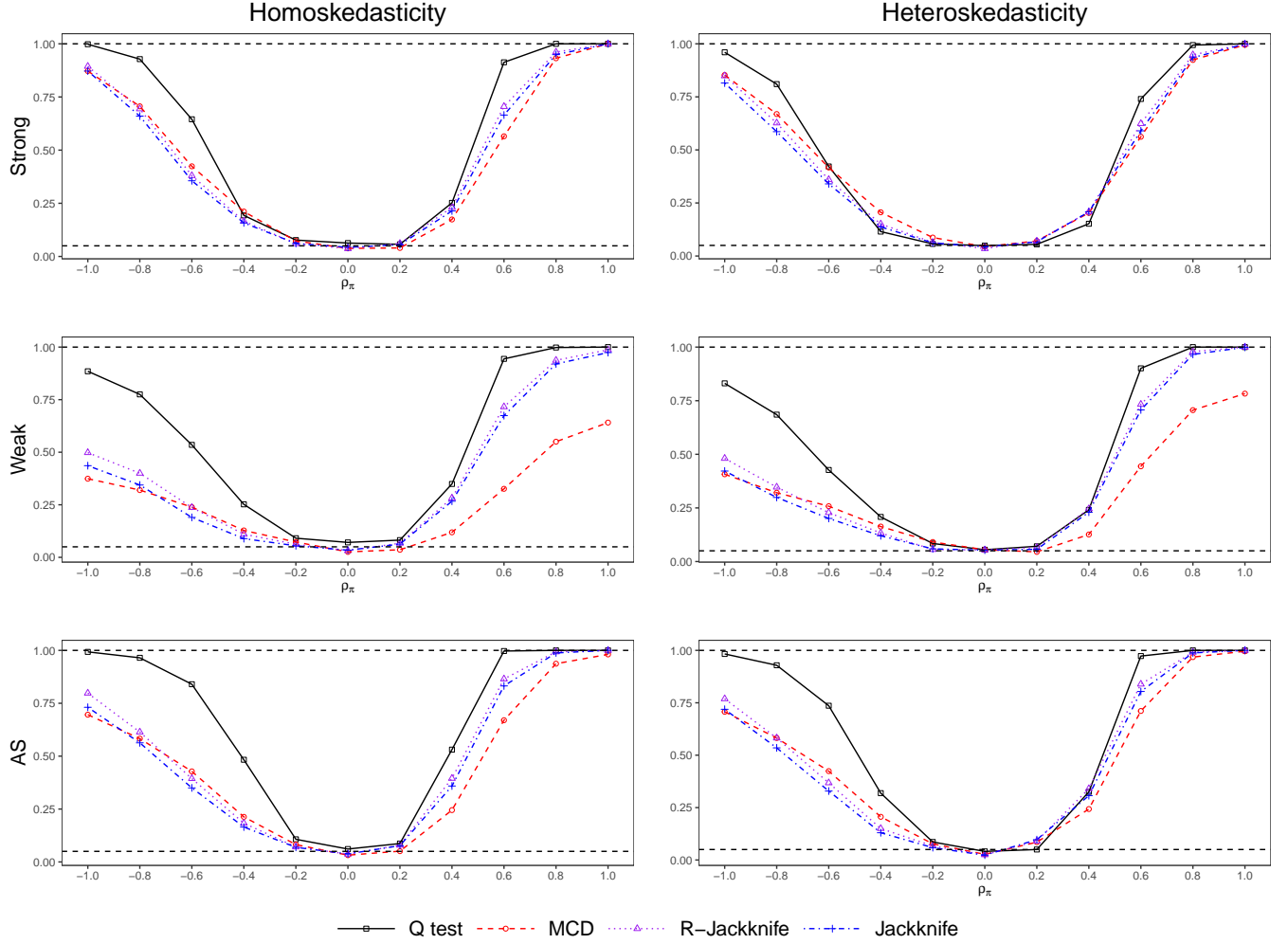
Since  $\mathbb{E}[v_i^2] = \pi + \gamma(\beta - b_{\text{JIVE}}))^\top Z_i Z_i^\top (\pi + \gamma(\beta - b_{\text{JIVE}})) + \mathbb{E}[(e_i + \varepsilon_{i2}(\beta - b_{\text{JIVE}}))^2] \geq \Theta_{11} + b_{\text{JIVE}}^2 \Theta_{22} - 2b_{\text{JIVE}} \Theta_{12} > 0$ , we know that

$$\begin{aligned} p_z^{-1} \sum_{i \neq j} P_{ij}^2 v_i^2 v_j^2 &= p_z^{-1} \sum_{i \neq j} P_{ij}^2 \mathbb{E}[v_i^2] \mathbb{E}[v_j^2] + o_p(1) \\ &\geq (2p_z)^{-1} \sum_{i \neq j} P_{ij}^2 (\Theta_{11} + b_{\text{JIVE}}^2 \Theta_{22} - 2b_{\text{JIVE}} \Theta_{12})^2 \\ &= \frac{1}{2} (\Theta_{11} + b_{\text{JIVE}}^2 \Theta_{22} - 2b_{\text{JIVE}} \Theta_{12})^2 \end{aligned}$$

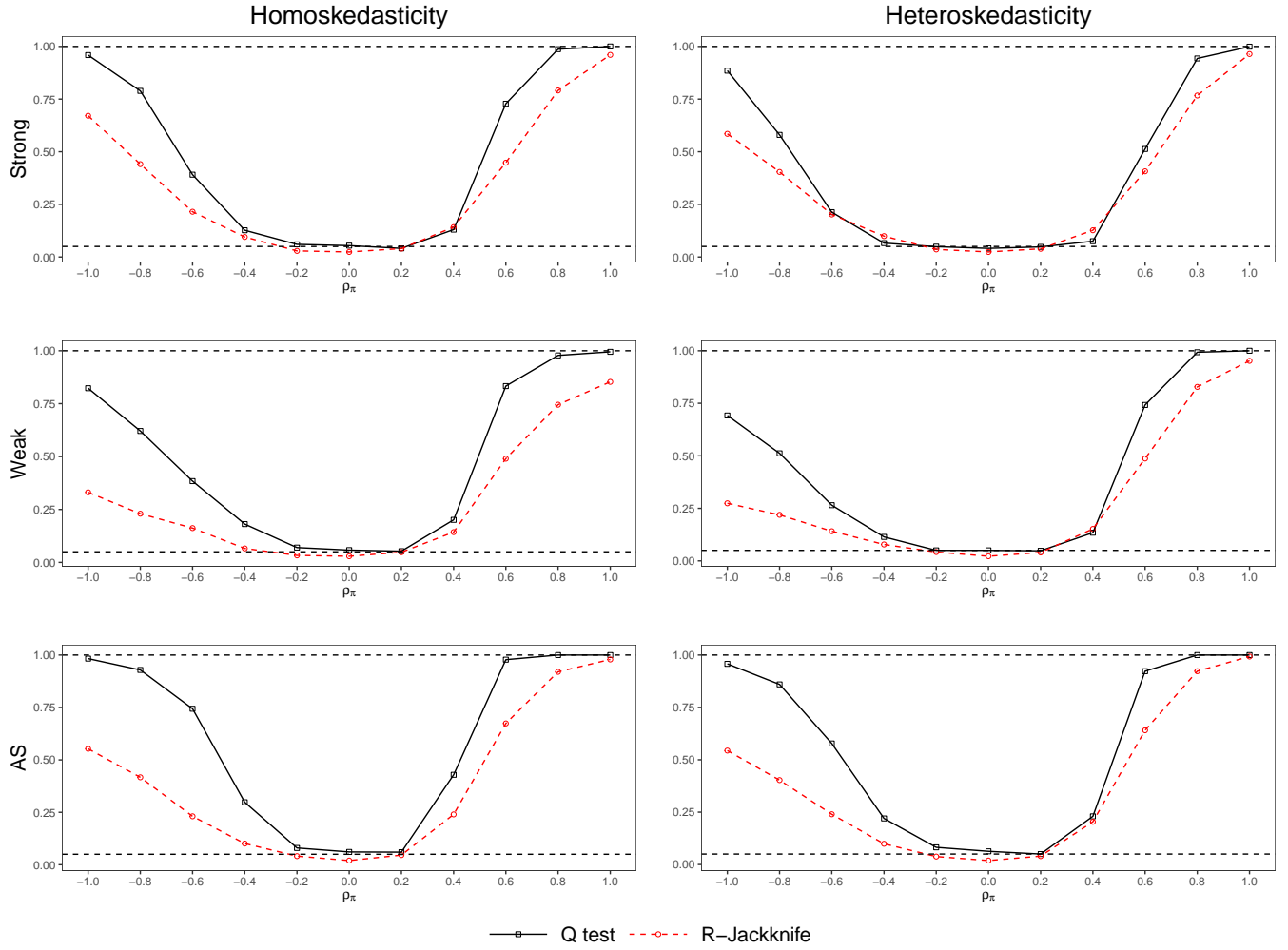
bounded away from zero by positive definiteness of  $\Theta$ . This shows that  $R_3$  is bounded away from zero and completes the proof of Proposition A2.

## C Omitted Simulation Results in Section 4

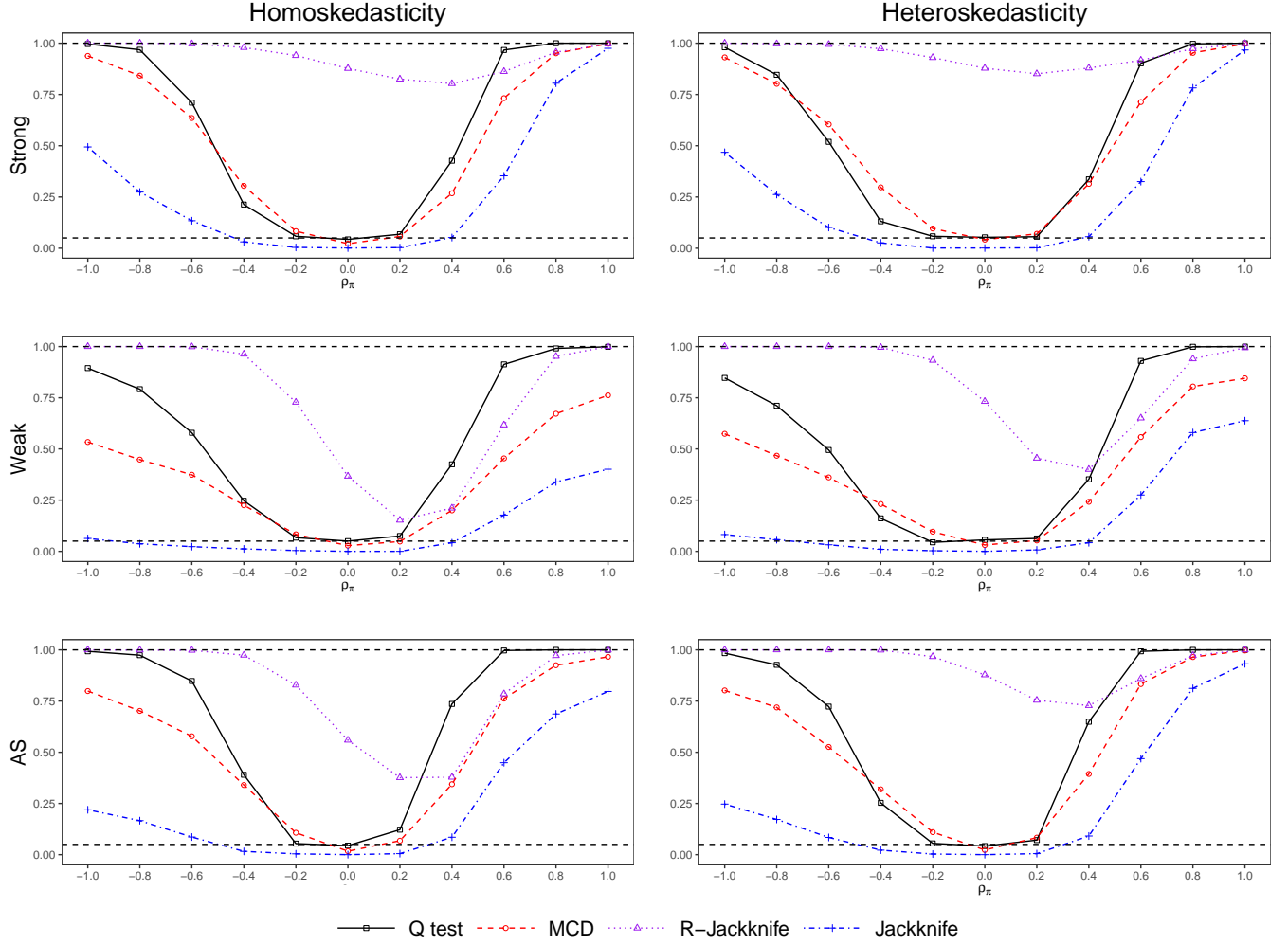
This section shows the simulation results omitted in the main paper. Figures C1 to C7 show the power curves of the cases with  $n = 200$ , and Figures C8 to C18 show the rejection rates of the Q test under different choices of  $\tau_0$  and different magnitudes of  $\rho_\pi$ .



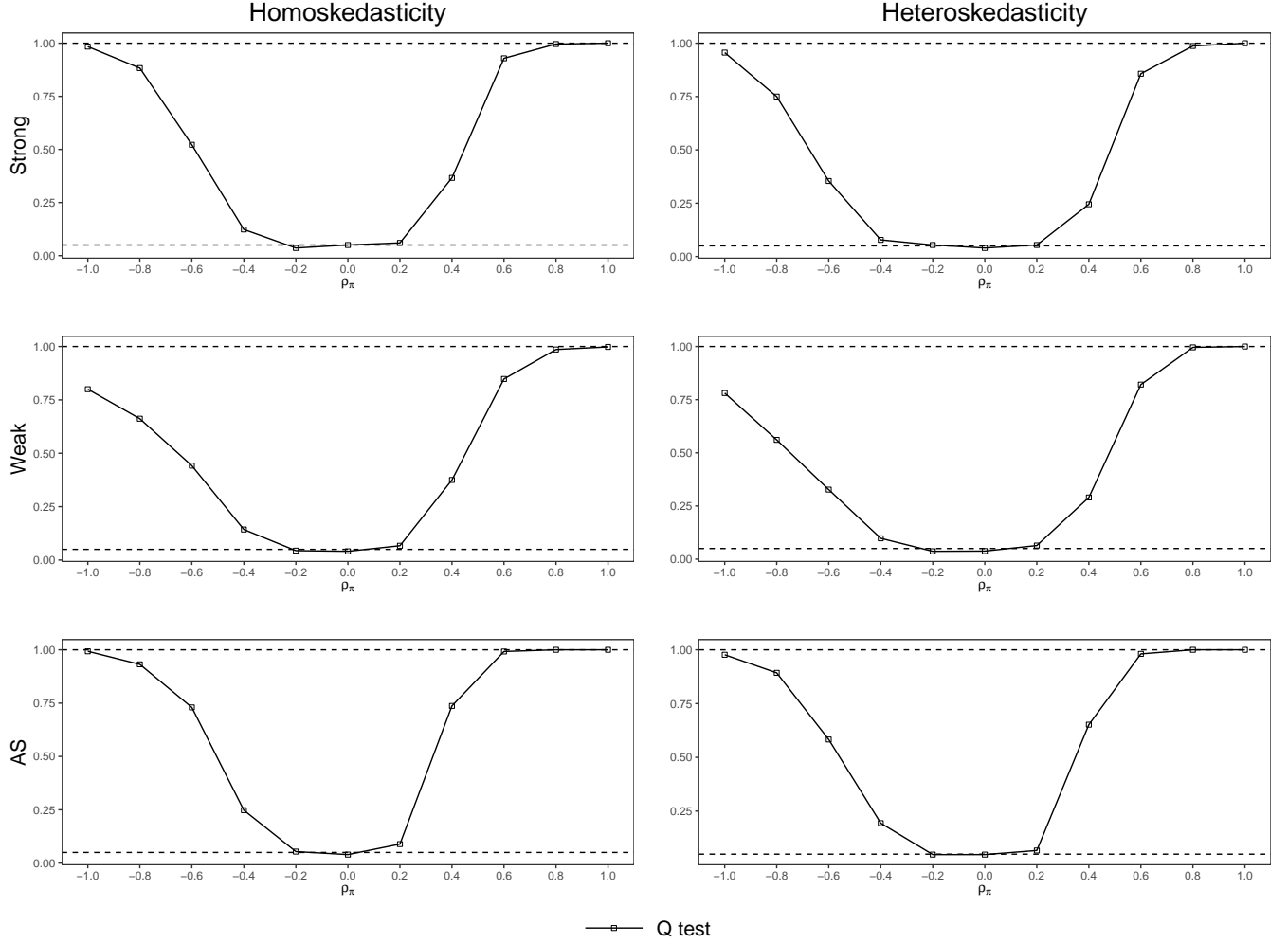
**Figure C1.** Power of instrument validity tests with  $(n, p_x, p_z) = (200, 0, 150)$  and sparse  $\pi$  under 5% level over 1000 simulations. “Jackknife”, “R-Jackknife” and “MCD” represent the jackknife test by [Chao et al. \(2014\)](#), the regularized jackknife test by [Carrasco and Doukali \(2021\)](#) and MCD test by [Kolesár \(2018\)](#), respectively. The nominal size 0.05 and power 1 are shown by the horizontal dashed lines. “Strong”, “Weak” and “AS” represent different settings of  $\gamma$ .



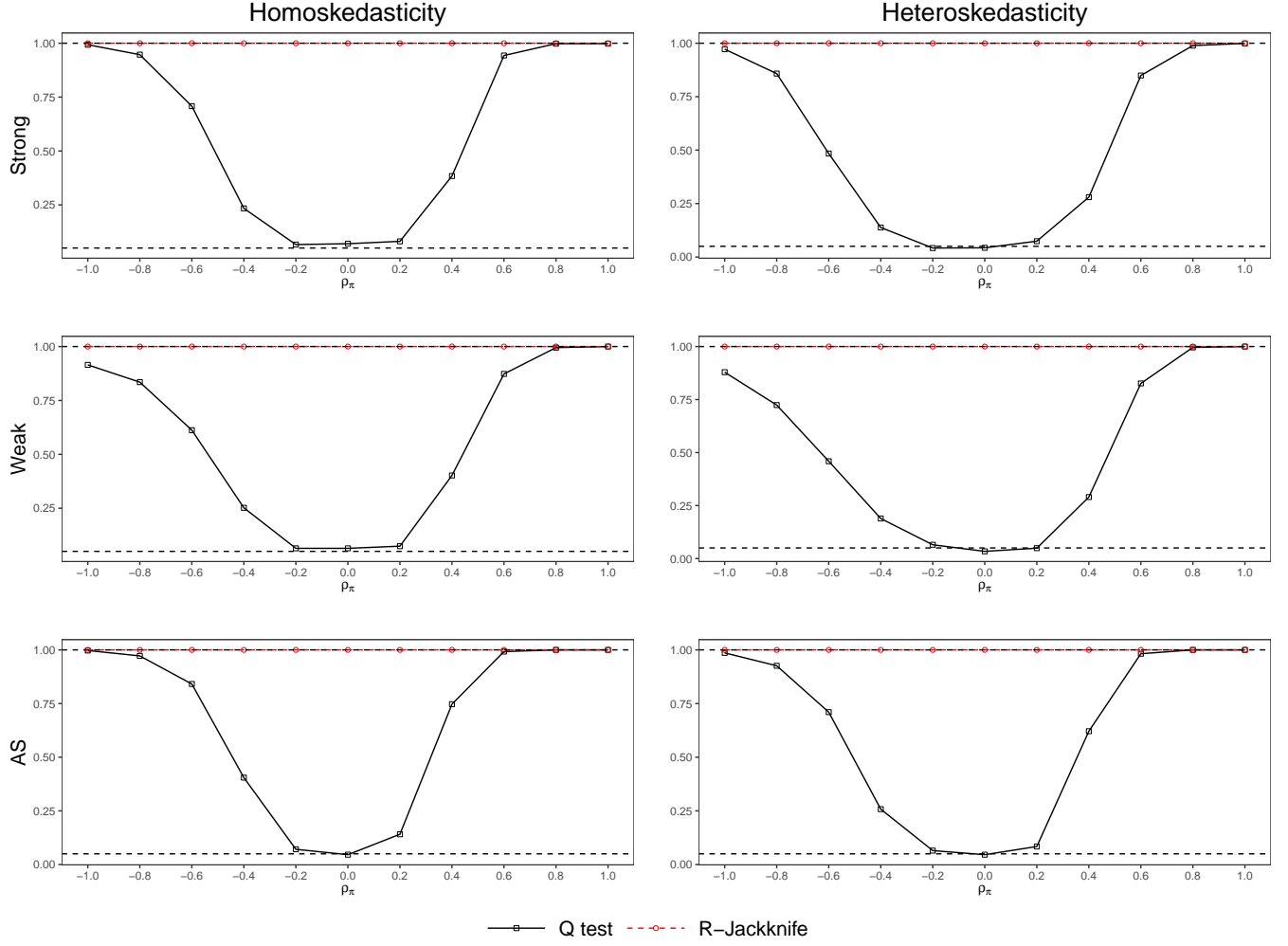
**Figure C2.** Power of instrument validity tests with  $(n, p_x, p_z) = (200, 0, 250)$  and sparse  $\pi$  under 5% level over 1000 simulations. “Jackknife”, “R-Jackknife” and “MCD” represent the jackknife test by [Chao et al. \(2014\)](#), the regularized jackknife test by [Carrasco and Doukali \(2021\)](#) and MCD test by [Kolesár \(2018\)](#), respectively. The nominal size 0.05 and power 1 are shown by the horizontal dashed lines. “Strong”, “Weak” and “AS” represent different settings of  $\gamma$ .



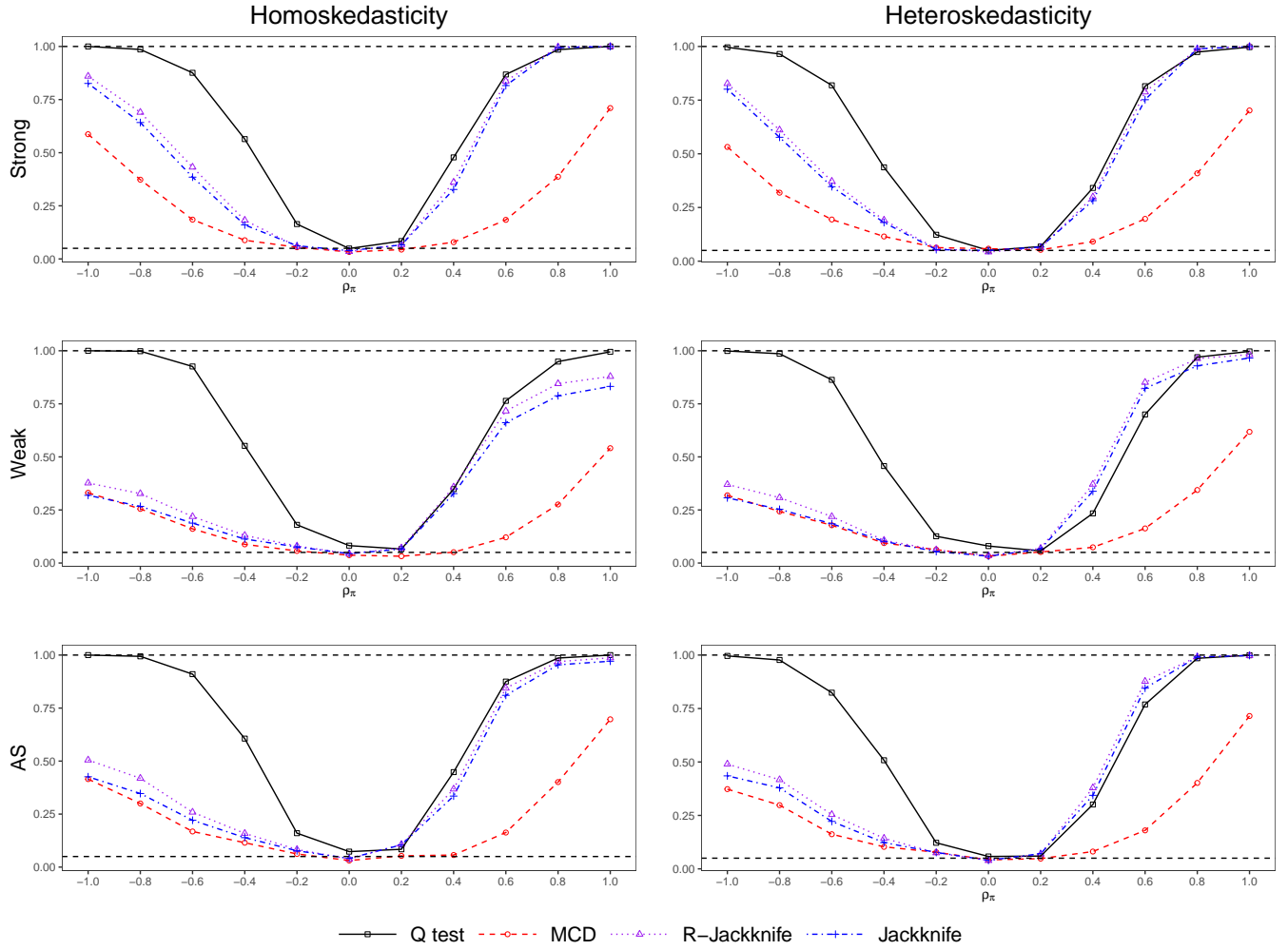
**Figure C3.** Power of instrument validity tests with  $(n, p_x, p_z) = (200, 150, 10)$  and sparse  $\pi$  under 5% level over 1000 simulations. “Jackknife”, “R-Jackknife” and “MCD” represent the jackknife test by [Chao et al. \(2014\)](#), the regularized jackknife test by [Carrasco and Doukali \(2021\)](#) and MCD test by [Kolesár \(2018\)](#), respectively. The nominal size 0.05 and power 1 are shown by the horizontal dashed lines. “Strong”, “Weak” and “AS” represent different settings of  $\gamma$ .



**Figure C4.** Power of instrument validity tests with  $(n, p_x, p_z) = (200, 250, 10)$  and sparse  $\pi$  under 5% level over 1000 simulations. “Jackknife”, “R-Jackknife” and “MCD” represent the jackknife test by [Chao et al. \(2014\)](#), the regularized jackknife test by [Carrasco and Doukali \(2021\)](#) and MCD test by [Kolesár \(2018\)](#), respectively. The nominal size 0.05 and power 1 are shown by the horizontal dashed lines. “Strong”, “Weak” and “AS” represent different settings of  $\gamma$ .

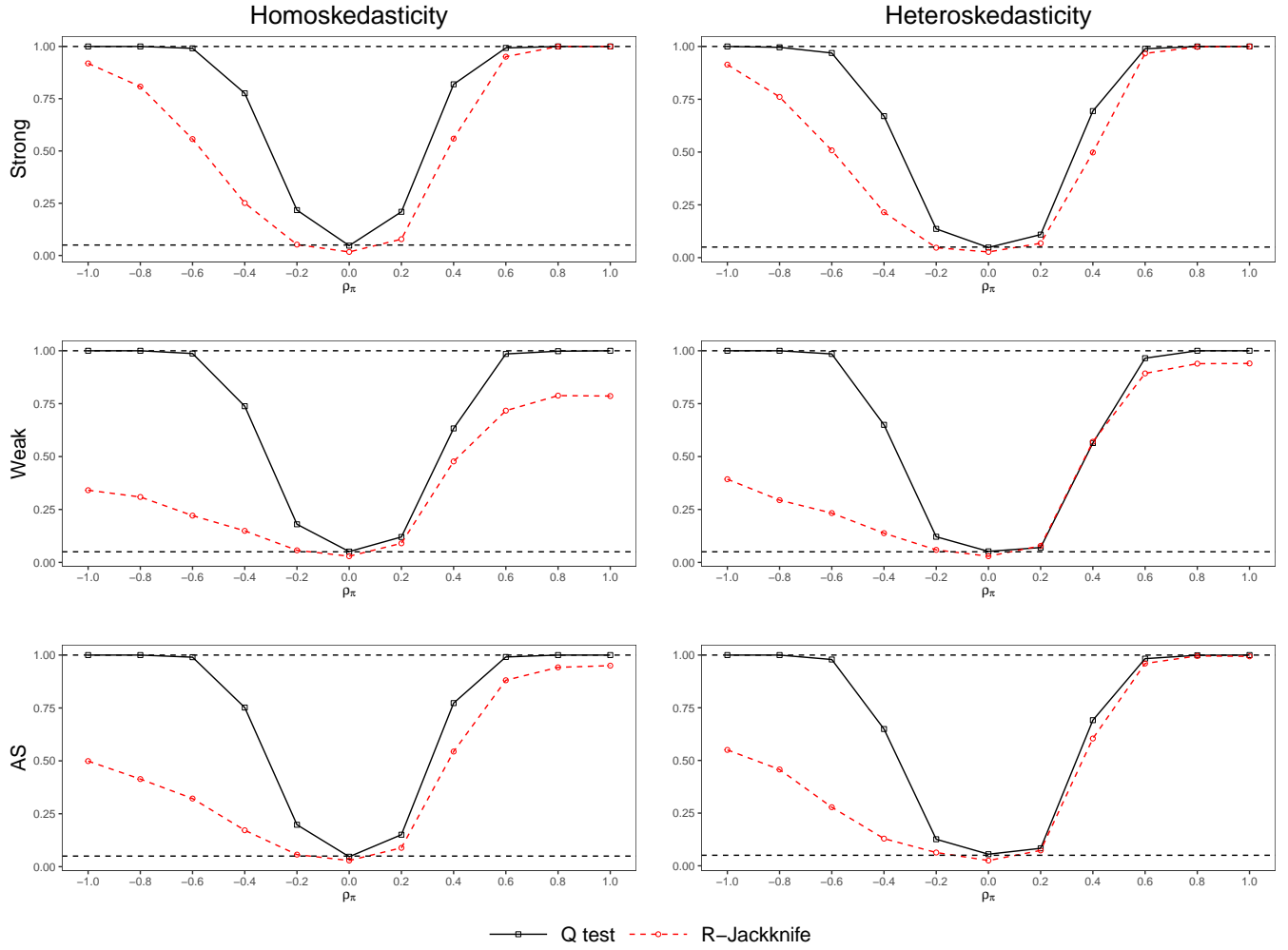


**Figure C5.** Power of instrument validity tests with  $(n, p_x, p_z) = (200, 150, 100)$  and sparse  $\pi$  under 5% level over 1000 simulations. “Jackknife”, “R-Jackknife” and “MCD” represent the jackknife test by [Chao et al. \(2014\)](#), the regularized jackknife test by [Carrasco and Doukali \(2021\)](#) and MCD test by [Kolesár \(2018\)](#), respectively. The nominal size 0.05 and power 1 are shown by the horizontal dashed lines. “Strong”, “Weak” and “AS” represent different settings of  $\gamma$ .

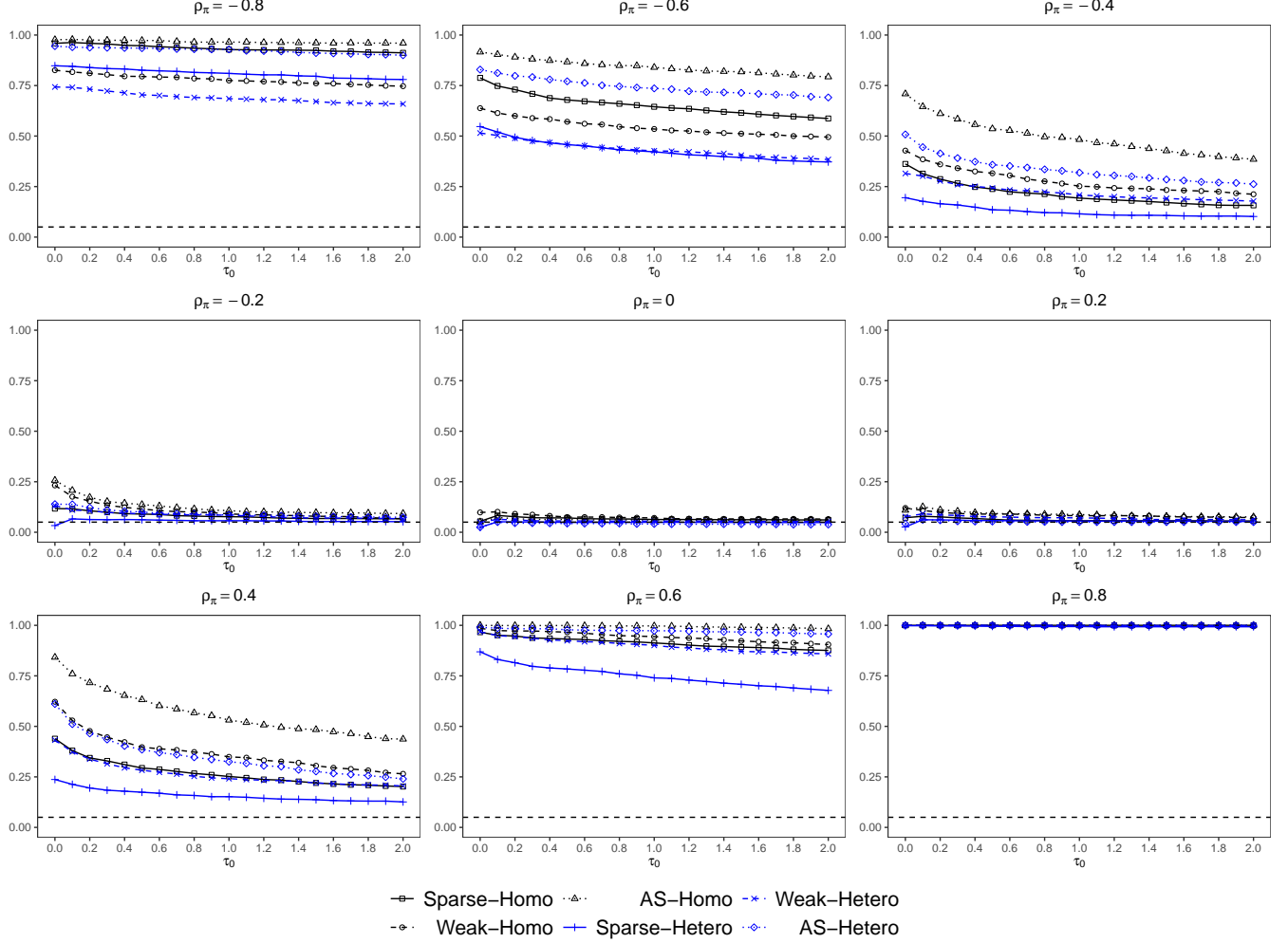


**Figure C6.** Power of instrument validity tests with  $(n, p_x, p_z) = (200, 0, 150)$  and dense  $\pi$  under 5% level over 1000 simulations. “Jackknife”, “R-Jackknife” and “MCD” represent the jackknife test by [Chao et al. \(2014\)](#), the regularized jackknife test by [Carrasco and Doukali \(2021\)](#) and MCD test by [Kolesár \(2018\)](#), respectively. The nominal size 0.05 and power 1 are shown by the horizontal dashed lines. “Strong”, “Weak” and “AS” represent different settings of  $\gamma$ .

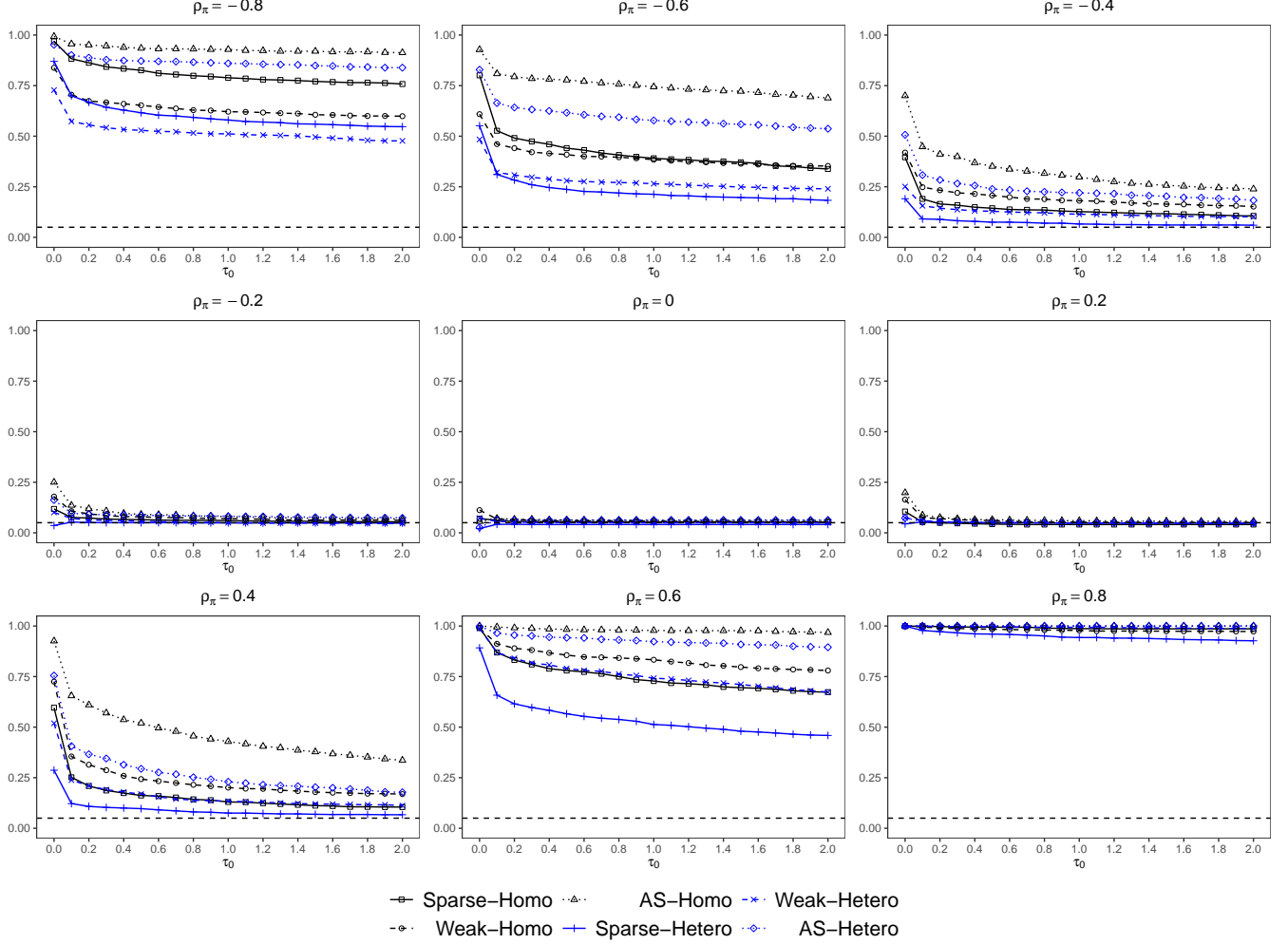




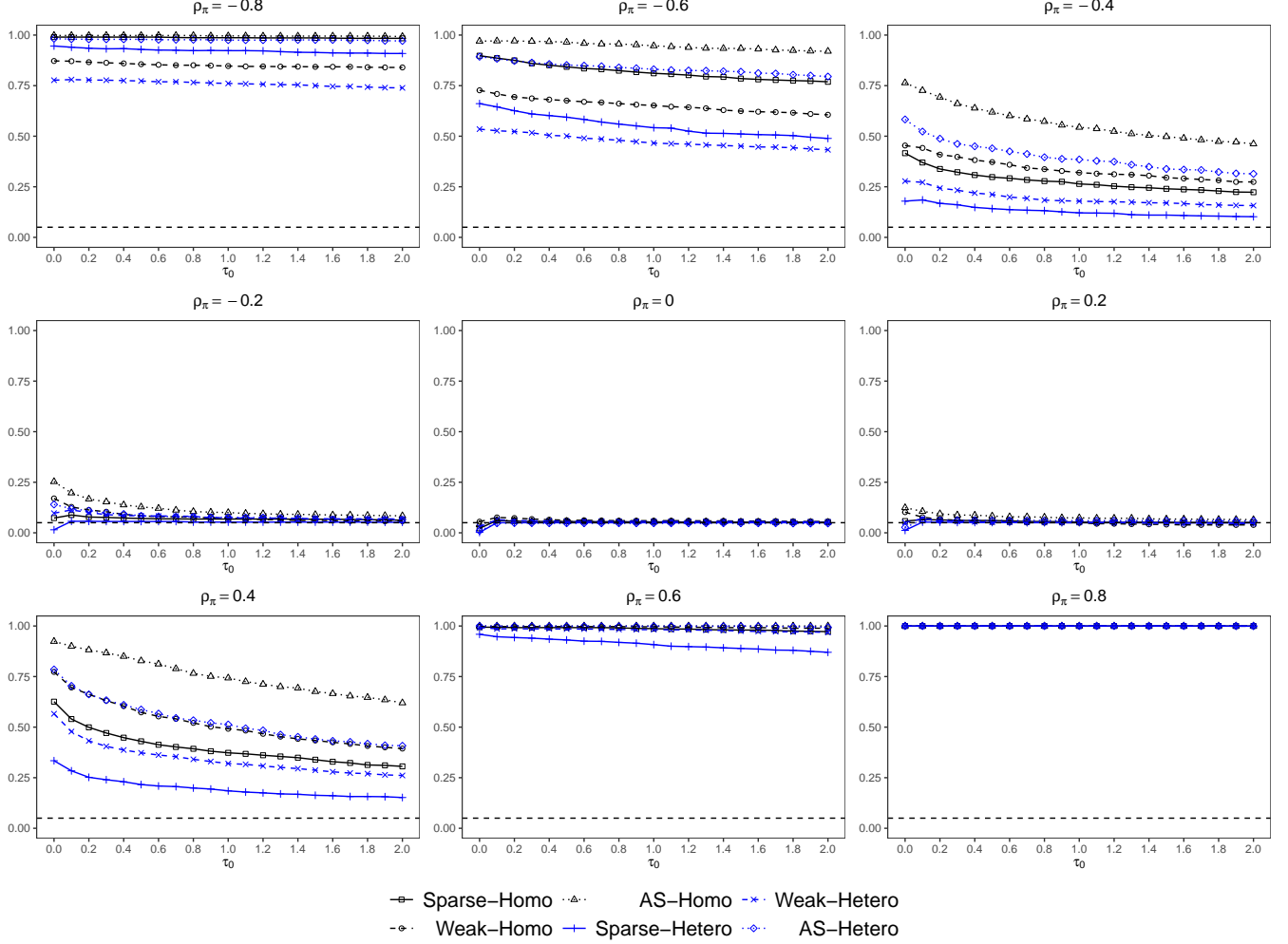
**Figure C7.** Power of instrument validity tests with  $(n, p_x, p_z) = (200, 0, 250)$  and dense  $\pi$  under 5% level over 1000 simulations. “Jackknife”, “R-Jackknife” and “MCD” represent the jackknife test by [Chao et al. \(2014\)](#), the regularized jackknife test by [Carrasco and Doukali \(2021\)](#) and MCD test by [Kolesár \(2018\)](#), respectively. The nominal size 0.05 and power 1 are shown by the horizontal dashed lines. “Strong”, “Weak” and “AS” represent different settings of  $\gamma$ .



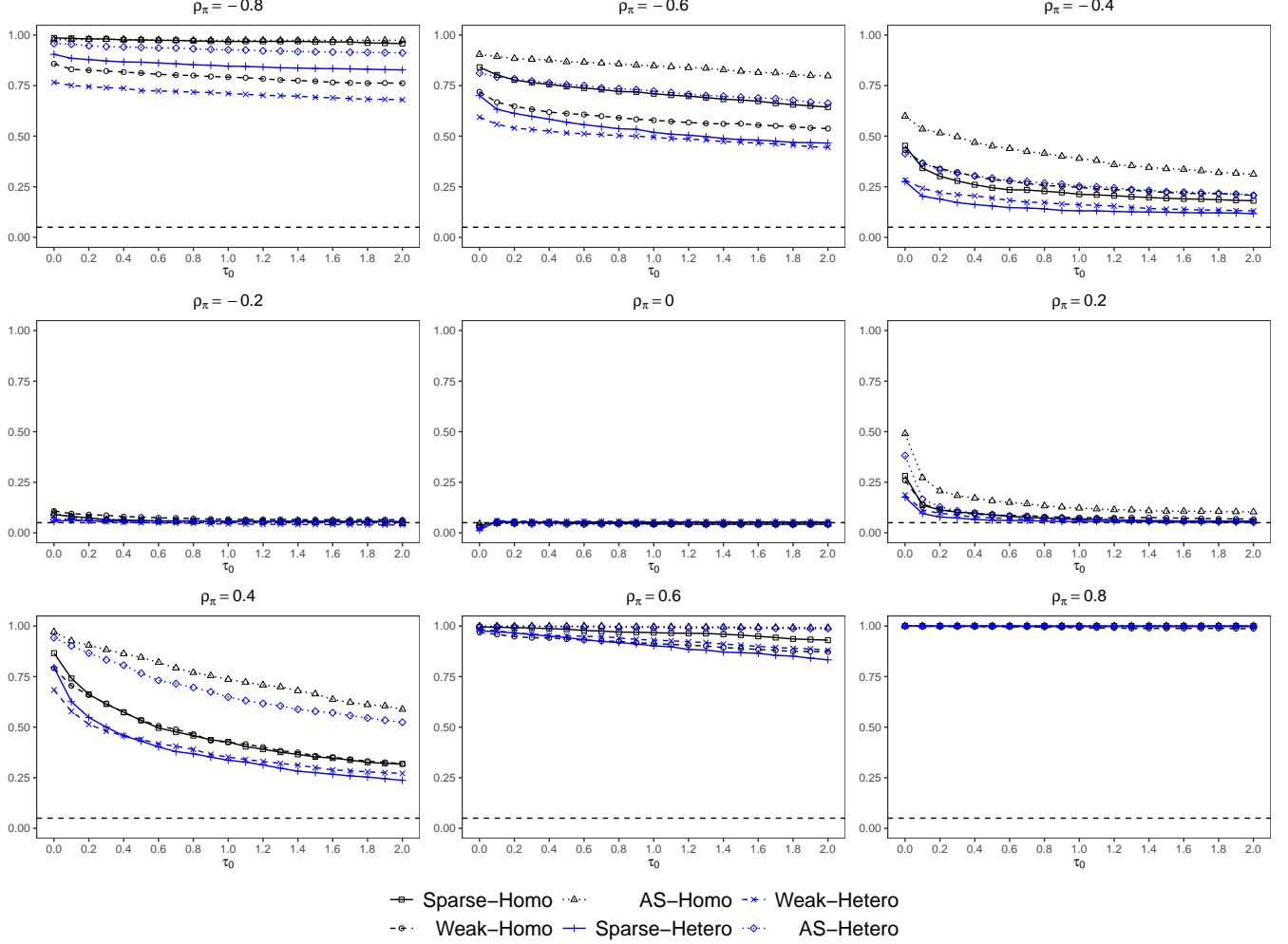
**Figure C8.** Empirical rejection rates of the Q test with  $(n, p_x, p_z) = (200, 0, 150)$  and sparse  $\pi$  under 5% level over 1000 simulations. The horizontal axis represents values of  $\tau_0$ . The nominal size 0.05 is shown by the horizontal dashed line. “Strong”, “Weak” and “AS” represent different settings of  $\gamma$ . “Homo” is short for homoskedasticity while “hetero” for heteroskedasticity.



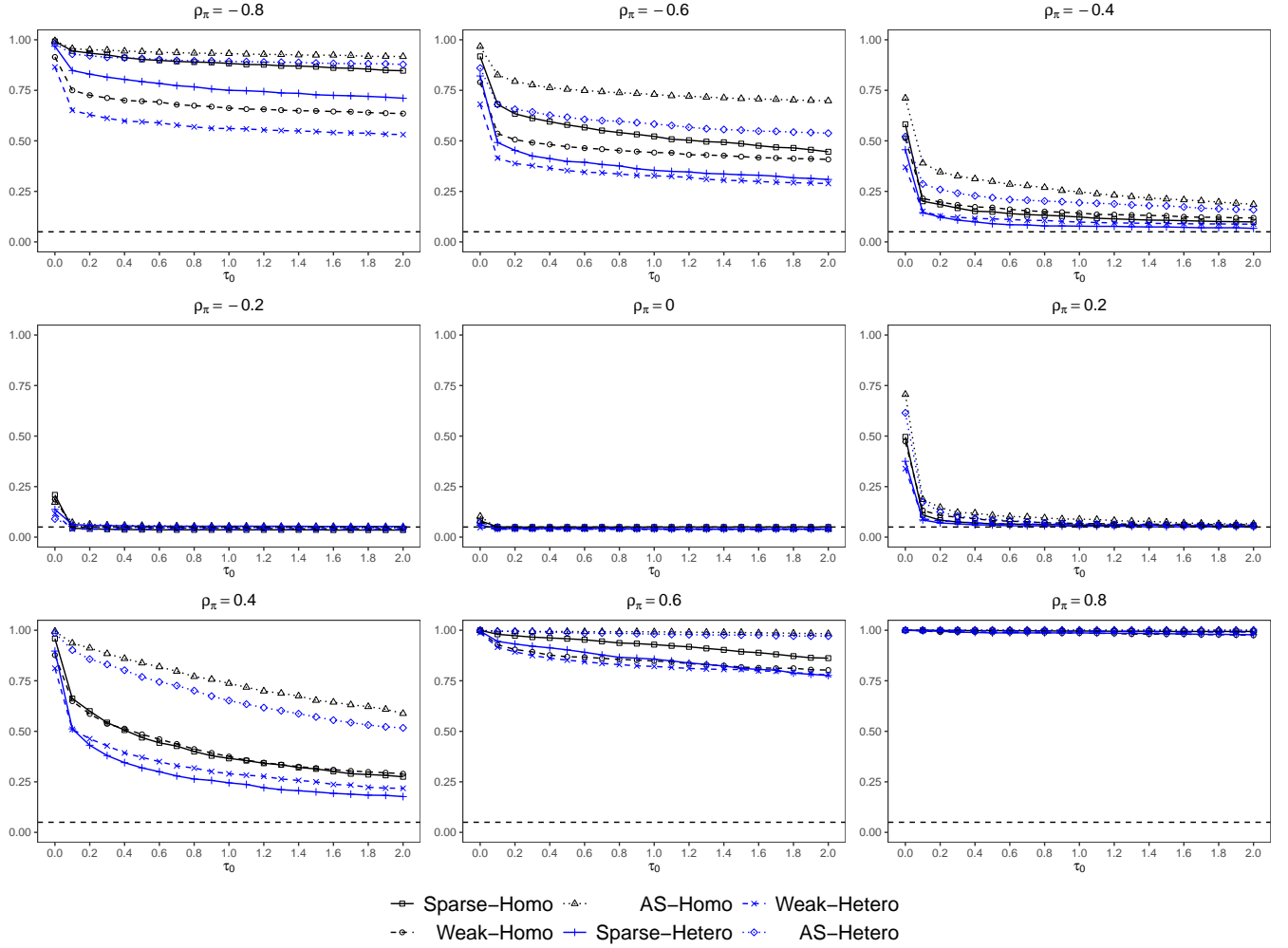
**Figure C9.** Empirical rejection rates of the Q test with  $(n, p_x, p_z) = (200, 0, 250)$  and sparse  $\pi$  under 5% level over 1000 simulations. The horizontal axis represents values of  $\tau_0$ . The nominal size 0.05 is shown by the horizontal dashed line. “Strong”, “Weak” and “AS” represent different settings of  $\gamma$ . “Homo” is short for homoskedasticity while “hetero” for heteroskedasticity.



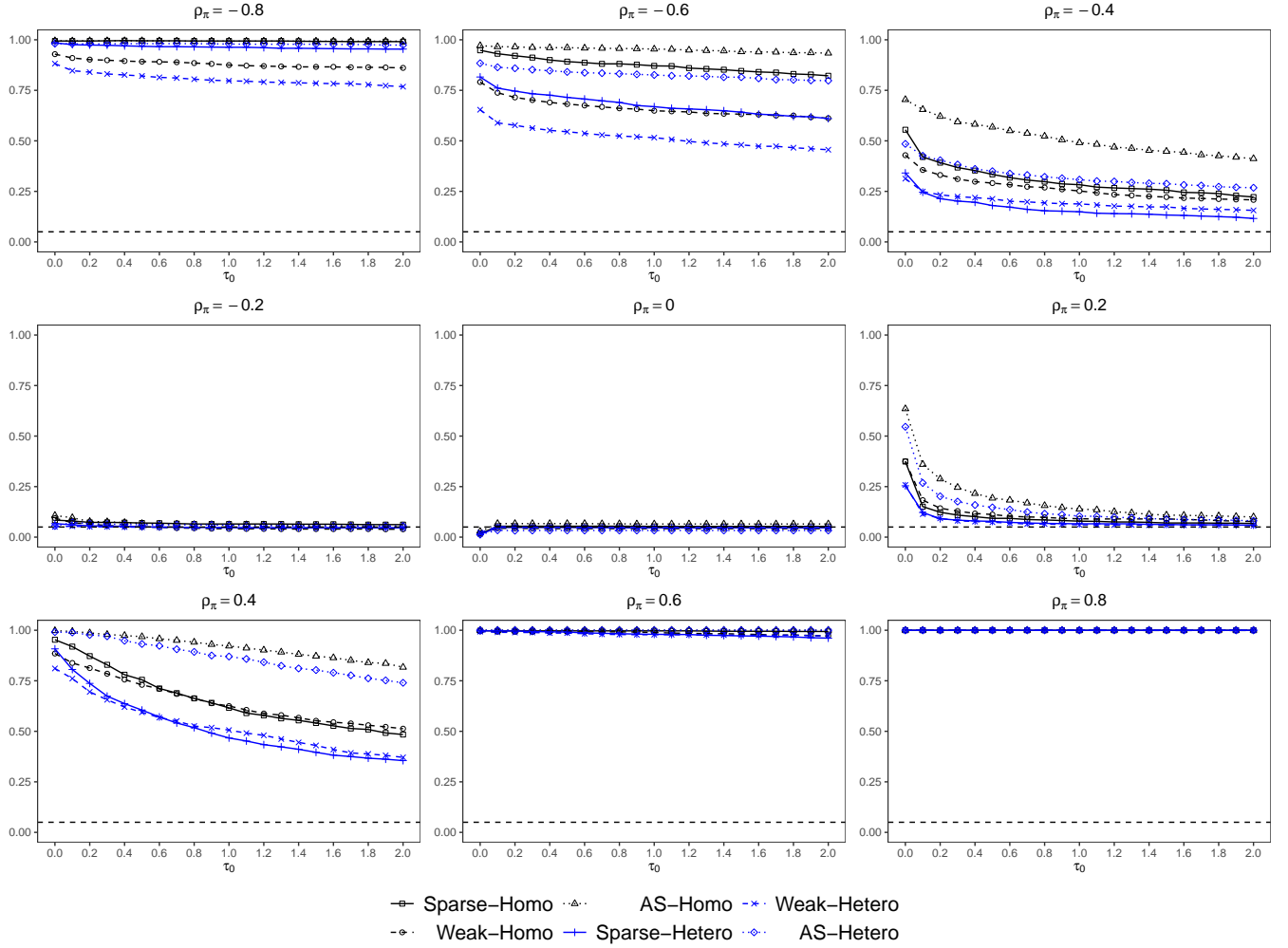
**Figure C10.** Empirical rejection rates of the Q test with  $(n, p_x, p_z) = (300, 0, 250)$  and sparse  $\pi$  under 5% level over 1000 simulations. The horizontal axis represents values of  $\tau_0$ . The nominal size 0.05 is shown by the horizontal dashed line. “Strong”, “Weak” and “AS” represent different settings of  $\gamma$ . “Homo” is short for homoskedasticity while “hetero” for heteroskedasticity.



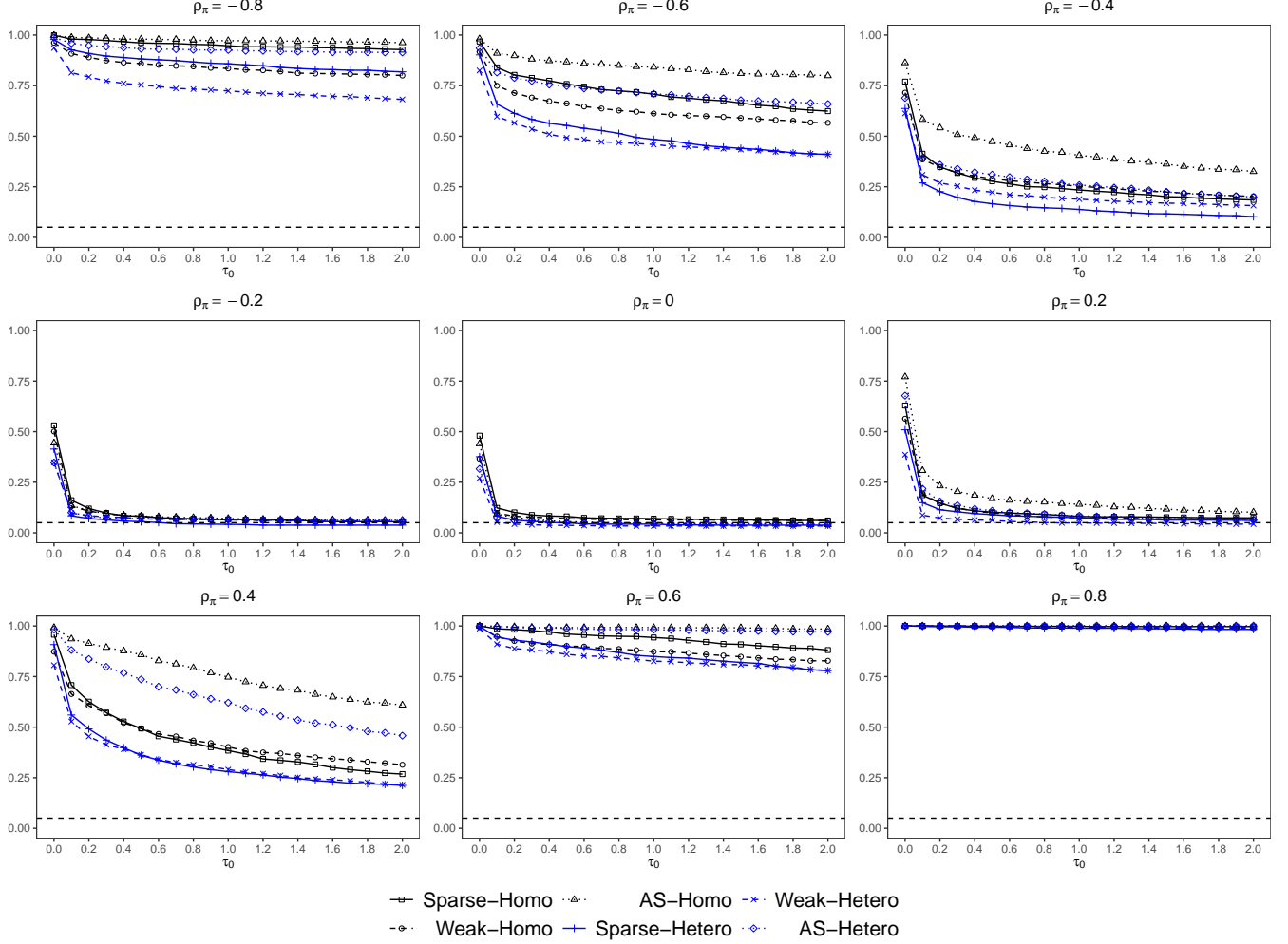
**Figure C11.** Empirical rejection rates of the Q test with  $(n, p_x, p_z) = (200, 150, 10)$  and sparse  $\pi$  under 5% level over 1000 simulations. The horizontal axis represents values of  $\tau_0$ . The nominal size 0.05 is shown by the horizontal dashed line. “Strong”, “Weak” and “AS” represent different settings of  $\gamma$ . “Homo” is short for homoskedasticity while “hetero” for heteroskedasticity.



**Figure C12.** Empirical rejection rates of the Q test with  $(n, p_x, p_z) = (200, 250, 10)$  and sparse  $\pi$  under 5% level over 1000 simulations. The horizontal axis represents values of  $\tau_0$ . The nominal size 0.05 is shown by the horizontal dashed line. “Strong”, “Weak” and “AS” represent different settings of  $\gamma$ . “Homo” is short for homoskedasticity while “hetero” for heteroskedasticity.

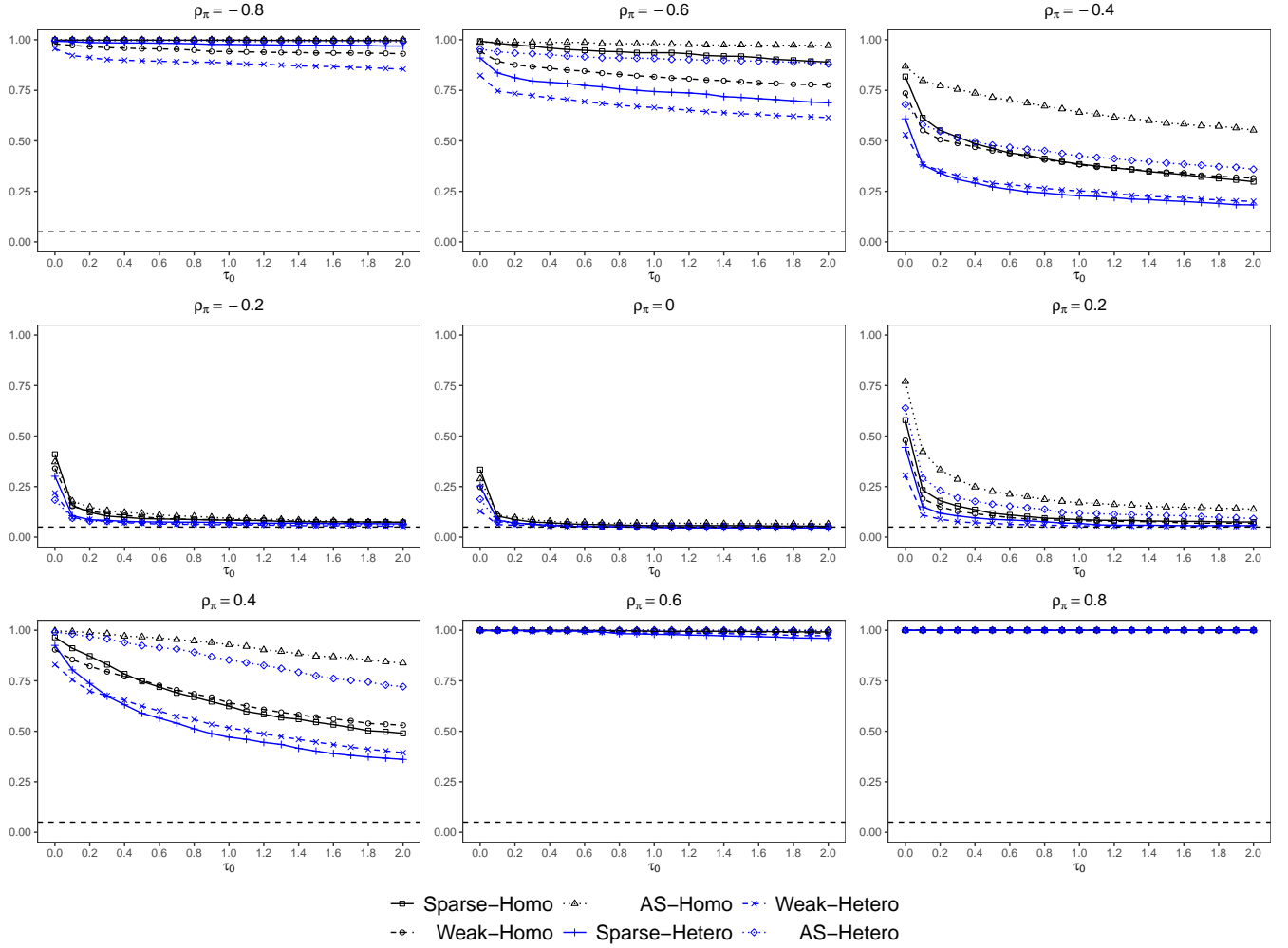


**Figure C13.** Empirical rejection rates of the Q test with  $(n, p_x, p_z) = (300, 250, 10)$  and sparse  $\pi$  under 5% level over 1000 simulations. The horizontal axis represents values of  $\tau_0$ . The nominal size 0.05 is shown by the horizontal dashed line. “Strong”, “Weak” and “AS” represent different settings of  $\gamma$ . “Homo” is short for homoskedasticity while “hetero” for heteroskedasticity.

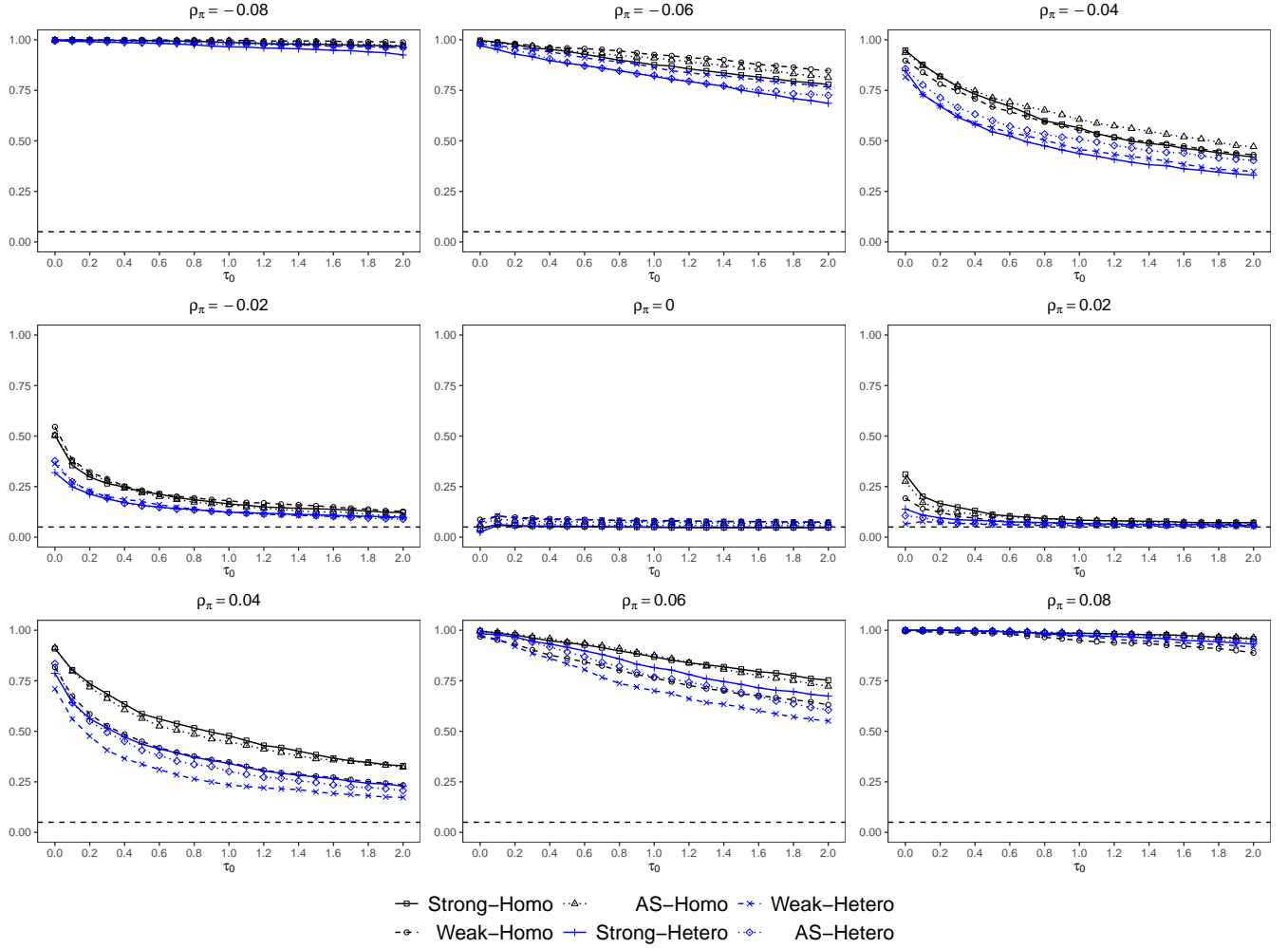


**Figure C14.** Empirical rejection rates of the Q test with  $(n, p_x, p_z) = (200, 150, 100)$  and sparse  $\pi$  under 5% level over 1000 simulations. The horizontal axis represents values of  $\tau_0$ . The nominal size 0.05 is shown by the horizontal dashed line. “Strong”, “Weak” and “AS” represent different settings of  $\gamma$ . “Homo” is short for homoskedasticity while “hetero” for heteroskedasticity.

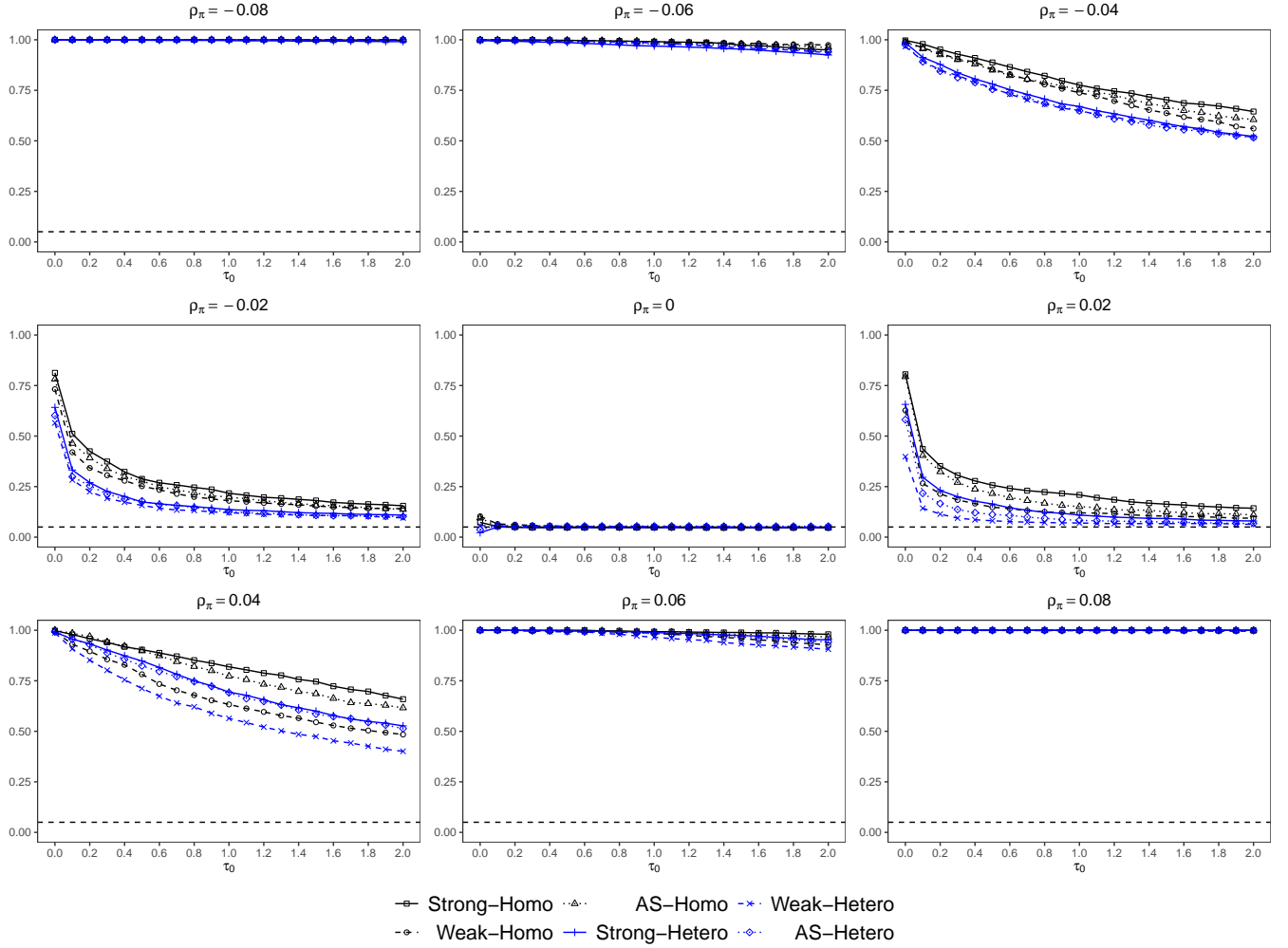




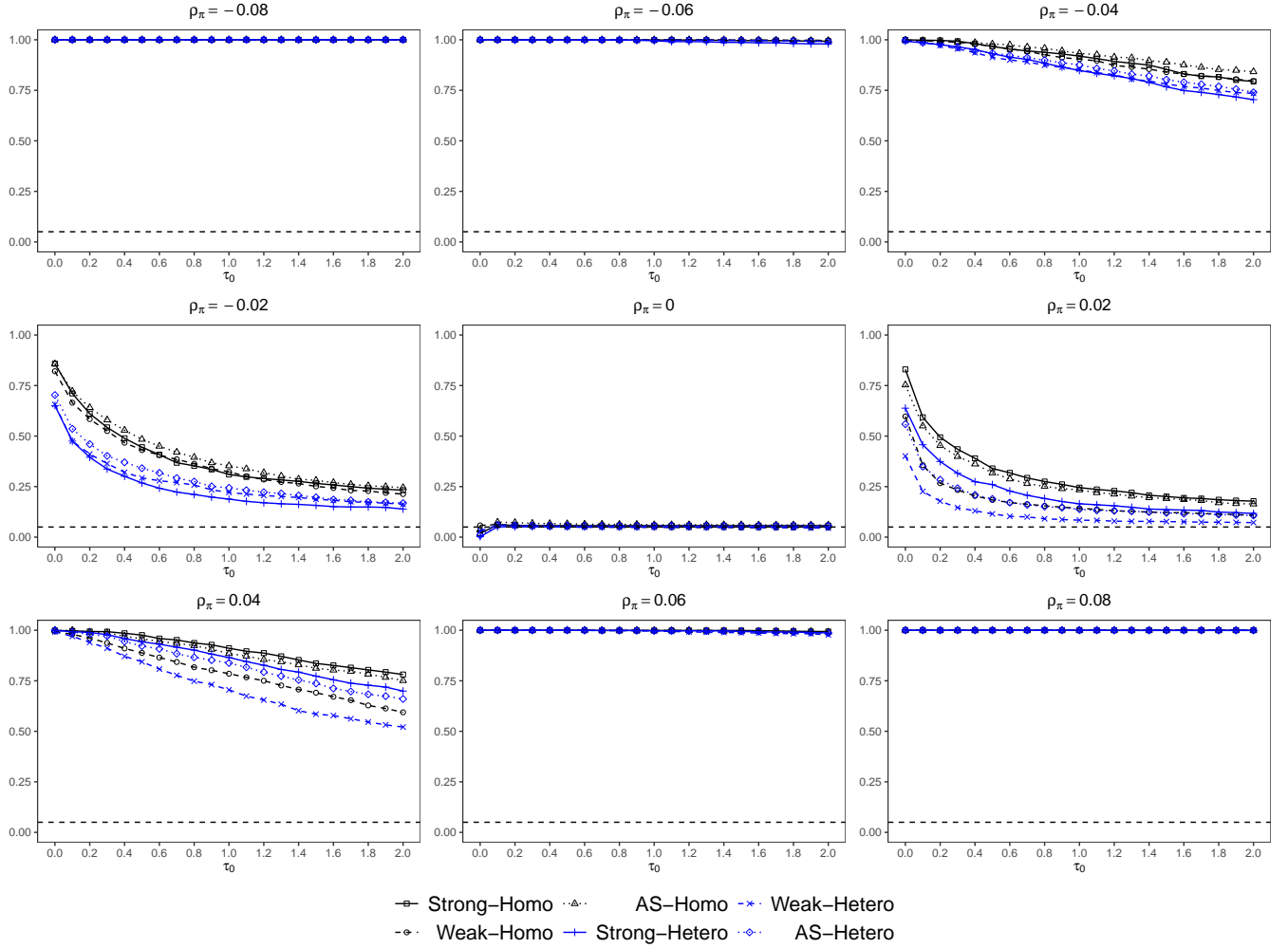
**Figure C15.** Empirical rejection rates of the Q test with  $(n, p_x, p_z) = (300, 150, 100)$  and sparse  $\pi$  under 5% level over 1000 simulations. The horizontal axis represents values of  $\tau_0$ . The nominal size 0.05 is shown by the horizontal dashed line. “Strong”, “Weak” and “AS” represent different settings of  $\gamma$ . “Homo” is short for homoskedasticity while “hetero” for heteroskedasticity.



**Figure C16.** Empirical rejection rates of the Q test with  $(n, p_x, p_z) = (200, 0, 150)$  and dense  $\pi$  under 5% level over 1000 simulations. The horizontal axis represents values of  $\tau_0$ . The nominal size 0.05 is shown by the horizontal dashed line. “Strong”, “Weak” and “AS” represent different settings of  $\gamma$ . “Homo” is short for homoskedasticity while “hetero” for heteroskedasticity.



**Figure C17.** Empirical rejection rates of the Q test with  $(n, p_x, p_z) = (200, 0, 250)$  and dense  $\pi$  under 5% level over 1000 simulations. The horizontal axis represents values of  $\tau_0$ . The nominal size 0.05 is shown by the horizontal dashed line. “Strong”, “Weak” and “AS” represent different settings of  $\gamma$ . “Homo” is short for homoskedasticity while “hetero” for heteroskedasticity.



**Figure C18.** Empirical rejection rates of the Q test with  $(n, p_x, p_z) = (300, 0, 250)$  and dense  $\pi$  under 5% level over 1000 simulations. The horizontal axis represents values of  $\tau_0$ . The nominal size 0.05 is shown by the horizontal dashed line. “Strong”, “Weak” and “AS” represent different settings of  $\gamma$ . “Homo” is short for homoskedasticity while “hetero” for heteroskedasticity.