

Supplement to “Instrumental Variables Estimation with Some Invalid Instruments and its Application to Mendelian Randomization”

Abstract

In this supplement we provide additional discussions, extended simulations, numerical results, and present all the technical details, including the proofs of Theorems 1, 2, and 3.

1 Additional Discussion About Theorem 1

1.1 Numerical Example

In Section 3.1 of the main manuscript, we discussed the identification result and illustrated it with a numerical example where $L = 4$, $\boldsymbol{\gamma}^* = (1, 2, 3, 4)$, $\boldsymbol{\Gamma}^* = (1, 2, 6, 8)$, and $s < U$ where $U = 3$. We showed that there are two sets $C_1 = \{1, 2\}$ and $C_2 = \{3, 4\}$ with $q_1 = 1$ and $q_2 = 2$. Since $q_1 \neq q_2$, by Theorem 1, identification is not possible with this numerical example.

One of the reviewers, however, mentioned an interesting numerical example where the setup is identical to our numerical example above, except $\boldsymbol{\Gamma}^*$ is perturbed by $\epsilon > 0$ such that $\tilde{\boldsymbol{\Gamma}}^* = (1, 2, 6, 8 + \epsilon)$. With $\tilde{\boldsymbol{\Gamma}}^*$, there is only one set $C_1 = \{1, 2\}$ where $q_1 = 1$ and we have

identification for any ϵ . However, we can shrink ϵ to be arbitrary small such that $\mathbf{\Gamma}^*$ and $\tilde{\mathbf{\Gamma}}^* = (1, 2, 6, 8 + \epsilon)$, are arbitrarily close to each other. As the reviewer stated “As a result, in any finite sample, it will be impossible to distinguish between the two cases, and hence no estimation or inference results that rely on Theorem 1 can be uniformly valid.”

However, consider the identical setup as before, except $\mathbf{\Gamma}^* = (1, 2, 7, 9)$. Then, there is only one subset $C_1 = \{1, 2\}$ where $q_1 = 1$ and identification is achieved. Furthermore, any small perturbation of $\mathbf{\Gamma}^*$ by $\delta > 0$ and $\epsilon > 0$, i.e. $\tilde{\mathbf{\Gamma}}^* = (1, 2, 7 + \delta, 9 + \epsilon)$, will still produce only subset $C_1 = \{1, 2\}$ and identification is maintained.

The two numerical examples with $\mathbf{\Gamma}^* = (1, 2, 6, 8)$ and $\mathbf{\Gamma}^* = (1, 2, 7, 9)$ illustrate what we call the *identification boundary*. The vector $\mathbf{\Gamma}^* = (1, 2, 6, 8)$ lies just at the identification boundary where any small perturbation can render the model unidentified or identified. In contrast, for $\mathbf{\Gamma}^* = (1, 2, 7, 9)$, the vector $\mathbf{\Gamma}^*$ lies far from the identification boundary and any small perturbation can still make the model identifiable. Exploration of the identification boundary for different values of $\mathbf{\Gamma}^*$ and $\boldsymbol{\gamma}^*$ is a topic for future research.

1.2 Normality Assumption and Identification

We consider two additional modeling assumptions which are not needed for identification, but are part of the classical linear simultaneous/structural equations model (Koopmans et al. 1950) and discuss the identification result in Section 3.1 of the main manuscript. First, we assume that the relationship between D_i and $\mathbf{Z}_{i.}$ is assumed to be linear

$$D_i = \mathbf{Z}_{i.}^T \boldsymbol{\gamma}^* + \xi_i, \quad E(\xi_i | \mathbf{Z}_{i.}) = 0 \quad (16)$$

where $\boldsymbol{\gamma}^*$ relates the instruments to the exposure and the error terms are bivariate Normal

$$(\epsilon_i, \xi_i) \stackrel{\text{iid}}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma}) \quad (17)$$

Under these assumptions in (16) and (17), the distributions of Y_i and D_i conditional on $\mathbf{Z}_{i\cdot}$ are fully characterized by finite-dimensional parameters $\boldsymbol{\alpha}^*, \beta^*, \boldsymbol{\gamma}^*$, and $\boldsymbol{\Sigma}$ known as “structural” parameters in econometrics (Wooldridge 2010). Let $\epsilon'_i = \beta^* \xi_i + \epsilon_i$. Then, we have the “reduced forms” (Wooldridge 2010)

$$Y_i = \mathbf{Z}_{i\cdot}^T \boldsymbol{\Gamma}^* + \epsilon'_i$$

$$D_i = \mathbf{Z}_{i\cdot}^T \boldsymbol{\gamma}^* + \xi_i$$

where $\boldsymbol{\Gamma}^* = \boldsymbol{\alpha}^* + \beta^* \boldsymbol{\gamma}^*$ and the covariance matrix of (ϵ'_i, ξ_i) is $\boldsymbol{\Sigma}' = \mathbf{M} \boldsymbol{\Sigma} \mathbf{M}^T$ with

$$\mathbf{M} = \begin{pmatrix} 1 & \beta^* \\ 0 & 1 \end{pmatrix}$$

We see that the distribution of Y_i and D_i are also fully characterized by the reduced form parameters $\boldsymbol{\Gamma}^*, \boldsymbol{\gamma}^*$ and $\boldsymbol{\Sigma}'$. By Rothenberg (1971), the reduced form parameters, $\boldsymbol{\Gamma}^*, \boldsymbol{\gamma}^*$, and $\boldsymbol{\Sigma}'$, are globally identified. Also, by Rothenberg (1971), the structural parameters, $\boldsymbol{\alpha}^*, \beta^*, \boldsymbol{\gamma}^*$, and $\boldsymbol{\Sigma}$, are identified if and only if the mapping between the reduced form parameters, $\boldsymbol{\Gamma}^*, \boldsymbol{\gamma}^*, \boldsymbol{\Sigma}'$, and the structural parameters, $\boldsymbol{\alpha}^*, \beta^*, \boldsymbol{\gamma}^*, \boldsymbol{\Sigma}$, represented by equations $\boldsymbol{\Sigma}' = \mathbf{M} \boldsymbol{\Sigma} \mathbf{M}^T$, $\boldsymbol{\gamma}^* = \boldsymbol{\gamma}^*$, and $\boldsymbol{\Gamma}^* = \boldsymbol{\alpha}^* + \beta^* \boldsymbol{\gamma}^*$, is bijective. We see that \mathbf{M} is an invertible matrix for any β^* and hence there is a bijective map between $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}'$. For $\boldsymbol{\gamma}^*$, it maps onto itself between the structural and reduced form parameters. Consequently, whether there is a bijection between the structural parameters and reduced form parameters is determined only by whether there is a unique solution $\boldsymbol{\alpha}^*$ and β^* to the equation

$$\boldsymbol{\Gamma}^* = \boldsymbol{\alpha}^* + \boldsymbol{\gamma}^* \beta^* \tag{18}$$

given $\boldsymbol{\gamma}^*$ and $\boldsymbol{\Gamma}^*$. Theorem 1 in the main manuscript states that a unique solution $\boldsymbol{\alpha}^*$ and β^* of (18) exists if and only if the consistency criterion holds, that $q_m = q_{m'}$ for all $m, m' \in$

$\{1, \dots, M\}$. Hence, with the modeling assumptions (16) and (17), we have identification of the structural parameters if and only if the consistency criterion holds.

2 Simulation

2.1 Values of ρ and μ

In Section 4 of the main manuscript, we conduct a simulation study to study the performance of sisVIVE compared to other competitors such as two stage least squares. In addition, in Section 3.4 of the main manuscript, Corollary 2 characterizes the performance of sisVIVE theoretically if certain conditions based on constants ρ and μ are satisfied. In this section, we check whether these theoretical conditions are met for the simulation setup we considered in the main manuscript.

We first computed ρ from each simulated data set and take the median value of it after 1000 simulations. To compute μ , we use the true values of the correlation of $\mathbf{Z}_{i.}$, specifically $\mu = 0, 0.25, 0.5$, and 0.75 . Table 1 shows the value for μ and ρ for the simulation setup in the main manuscript. Second, based on the values of μ and ρ in Table 1, we check the

Table 1. Values of ρ defined in Corollary 2 for the Simulation Study

Instrument Corr. (μ)	Strong Instrument, Equal Strength	Strong Instrument, Variable Strength	Weak Instrument, Equal Strength	Weak Instrument, Variable Strength
0	0.31	0.39	0.20	0.22
0.25	0.54	0.58	0.36	0.37
0.5	0.72	0.73	0.53	0.53
0.75	0.87	0.87	0.73	0.73

condition required in Corollary 2, specifically the upper bound on s , $\min(1/(12\mu), 1/(10\rho^2))$, in equation (14) of the main manuscript. These upper bounds are evaluated in Table 2. Table 2 shows that in most settings, the condition for Corollary 2 is only satisfied when $s = 0$, i.e. when there are no invalid instruments. For example, when instrument are correlated and $\mu > 0$, Corollary 2 cannot be used to characterize the performance of sisVIVE

Table 2. Condition on s in Corollary 2 for the Simulation Study

Instrument Corr. (μ)	Strong Instrument, Equal Strength	Strong Instrument, Variable Strength	Weak Instrument, Equal Strength	Weak Instrument, Variable Strength
0	1.04	0.66	2.50	2.07
0.25	0.33	0.33	0.33	0.33
0.5	0.17	0.17	0.17	0.17
0.75	0.11	0.11	0.11	0.11

if invalid instruments are present. Table 2 also illustrates the point we illustrated in the main manuscript, that the condition for Corollary 2, even though it's interpretable, are strict. In the main manuscript, we provide a generalization of Corollary 2 in Theorem 2 at the expense of interpretability.

2.2 Varying Correlation Structure

In this section, we extend the simulation study in Section 4 of the main manuscript by considering other correlation structures between the instruments beyond those considered in the main manuscript. First, Figures 1 and 2 of the Supplementary Materials represent the setting where the pairwise correlation between valid instruments is set to μ and the pairwise correlation between invalid instruments is also set to μ . However, there is no correlation between any pair consisting of one valid and one invalid instrument. The new setup differs from the main manuscript where all the pairwise correlation between any two instruments is set to μ . Second, Figures 3 and 4 represent the setting where the pairwise correlation between a valid instrument and an invalid instrument is set to μ . However, there is no pairwise correlation between any pair of valid instruments or any pair of invalid instruments. Under the two new correlation structures, we rerun the simulation study in the main manuscript except we reduce the simulation number from 1000 to 500 and we only vary s with values $s = 1, 3, 4, 5, 7$, and 9 for computational reasons. Also, note that as a result of repeating the same simulation, the conditions for Corollary 2 in the main manuscript are similar to those in Tables 1 and 2 of Section 2.1 in the Supplementary Materials.

In both Figures 1 and 3 of the Supplementary Materials where we vary endogeneity, but the number of invalid instruments is fixed at $s = 3$, the behavior of all the estimators are similar to each other and to those in the main manuscript. OLS dominates naive TSLS, oracle TSLS, and sisVIVE when the endogeneity is small and close to zero, with the dominance being greater for weaker instruments. Once there is a sufficient amount of endogeneity, oracle TSLS, which knows exactly which instruments are valid and invalid, does best. sisVIVE also resembles the oracle in terms of performance. Naive TSLS, which assumes all the L instruments are valid, does worst since it assumes that all the L instruments are valid.

Similarly, in Figures 2 and 4 of the Supplementary Materials where we vary the number of invalid instruments, s , but fix the endogeneity to 0.8, the estimators behave similarly across the two Figures and to those in the main manuscript. We first see that at $s = 0$, i.e. when there are no invalid instruments, sisVIVE's performance is nearly identical to naive and oracle TSLS, although it degrades slightly for instruments with weak absolute strength. Also, when $s < L/2 = 5$, sisVIVE's performance is comparable to oracle TSLS and better than naive TSLS. Once we reach the identification boundary, $s < L/2 = 5$, sisVIVE's performance becomes similar to naive TSLS. This is the case regardless of the instruments' absolute and relative strength.

2.3 Performance of Estimate of $\hat{\alpha}_\lambda$

In this section, we extend the simulation study in Section 4 of the main manuscript by examining the estimation performance of α^* for sisVIVE. As we noted in the main manuscript, in Mendelian randomization, the target of estimation is β^* , the causal effect of the exposure on the outcome, and our procedure, sisVIVE, was designed to estimate β^* . However, in the process of estimating β^* , sisVIVE does produce an estimate for α^* . This section explores the relationship between this intermediate estimate for α^* , $\hat{\alpha}_\lambda$, and our desired estimate for β^* , $\hat{\beta}_\lambda$.

To evaluate the estimate $\hat{\alpha}_\lambda$, we consider two metrics for error, the proportion of correctly selected valid instruments and the proportion of correctly selected invalid instruments. To illustrate these proportion-based error metrics, consider the following numerical example. Suppose there are $L = 10$ instruments of which the first three instruments are invalid, $\alpha_j^* \neq 0$ for $j = 1, 2, 3$ and the last seven instruments are valid, $\alpha_j^* = 0$ for $j = 4, 5, \dots, 10$. If sisVIVE estimates the first two instruments to be invalid, $\hat{\alpha}_j \neq 0$ for $j = 1, 2$ and the last eight to be valid, $\hat{\alpha}_j = 0$ for $j = 3, 4, \dots, 10$, the proportion of correctly selected valid instruments is $7/7 = 1$ and sisVIVE makes no error in estimating the valid instruments. However, the proportion of correctly selected invalid instruments is $2/3$ and sisVIVE makes an error in estimating the invalid instruments.

We rerun the simulation setup in Section 4 of the main manuscript and in Section 2.2 in the Supplementary Materials. However, instead of measuring the median absolute deviation, $|\hat{\beta}_\lambda - \beta^*|$, we instead measure the two proportion-based error metrics. Similar to Section 2.2 in the Supplementary Materials, we reduce the simulation from 1000 to 500 and only consider $s = 1, 3, 4, 5, 7$, and 9 for computational reasons. The results are in Figures 5 to 10.

When we vary endogeneity but fix the number of invalid instruments to be $s = 3$ (Figures 5, 7, and 9), the proportion of correctly selected invalid instruments is 1 and sisVIVE never makes a mistake in selecting the invalid instruments. However, sisVIVE does make mistakes in selecting the valid instruments as the proportion of correctly selected valid instruments is mostly below 1. Also, depending on the correlation structure between instruments, we get different behaviors for the proportion of correctly selected valid instruments. For example, when every pair of instruments has non-zero pairwise correlation (Figure 5), the proportion of correctly selected valid instruments remains roughly the same for different values of endogeneity. When there is only pairwise correlation within valid and invalid instruments (Figure 7), the proportion of correctly selected valid instruments decreases as endogeneity increases, most notably among weak instruments. Finally, when there is only pairwise corre-

lation between valid and invalid instruments (Figure 9), the proportion of correctly selected valid instruments increases as endogeneity increases. Despite these differences in the proportion of correctly selected valid instruments between different correlation structures, as the simulations in Section 4 of the main manuscript and Section 2.2 of the Supplementary Materials showed, sisVIVE’s median absolute deviation from the truth, $|\hat{\beta}_\lambda - \beta^*|$, remains relatively small and constant for all values of the endogeneity. This constant behavior is also present in the proportion of correctly selected invalid instruments, which remains at 1 for all correlation structures. This suggests that there is a strong relationship between correctly selecting the invalid instruments and sisVIVE’s median absolute deviation from β^* while there is at most a weak relationship between correctly selecting valid instruments and sisVIVE’s median absolute deviation from β^* . In fact, it appears that correctly selecting invalid instruments is more important than valid instruments if a small median absolute deviation is desired.

When we vary the number of invalid instruments s , but fix the endogeneity (Figures 6, 8, and 10), the proportion of correctly selected invalid instrument decreases significantly at the $s = 5$ boundary, regardless of the correlation structure between instruments. For example, for strong instruments in the three Figures, when $s < 5$, the proportion of correctly selected invalid instruments remain at 1. However, when $s \geq 5$, the proportion of correctly selected invalid instruments moves sharply away from 1. For weak instruments in the three Figures, when $s < 5$, the proportion of correctly selected invalid instruments remains close to 1, although there is a slightly decrease in the proportion when s moves from $s = 3$ to $s = 4$ and when μ is away from zero. However, similar to the strong instruments, when $s \geq 5$, the proportion of correctly selected invalid instruments moves away from 1. In contrast, the proportion of correctly selected valid instruments decreases steadily as s increases, regardless of the type of correlation structure between instruments. For strong instruments in the three Figures, the decrease in the proportion of correctly selected valid instruments begins

immediately after $s = 1$. For weak instruments in the three Figures, there is considerable fluctuation of the proportion of correctly selected valid instruments. For Figures 6 and Figures 8, the proportion of correctly selected valid instruments generally decreases as s increase, with the notable exception in the first row, third column of both Figures. For Figure 10, the proportion of correctly selected valid instruments decreases when $s < 5$, but increases again after $s \geq 5$.

The behaviors of the proportions of correctly selected invalid and valid instruments from Figures 6, 8, and 10 reaffirms our previous observation that there is a strong association between the proportion of correctly selected invalid instruments and the median absolute deviation of $\hat{\beta}_\lambda$, $|\hat{\beta}_\lambda - \beta^*|$. In particular, from Figure 3 of the main manuscript and Figures 2 and 4 of the Supplementary Materials, when $s < 5$, sisVIVE's median absolute deviation is just as small as the oracle two stage least squares. However, when $s \geq 5$, sisVIVE's median absolute deviation is just as large as the naive two stage least squares. The proportion of correctly selected invalid instruments in Figures 6, 8, and 10 closely corresponds to this sharp change in behavior between $s < 5$ and $s \geq 5$. In contrast, the proportion of correctly selected valid instruments does not have this sharp behavior at $s = 5$ across all the figures.

Overall, by measuring the estimation performance of $\hat{\alpha}_\lambda$ using the two proportion-based error metrics, we notice a strong relationship between the proportion of correctly selected invalid instruments and the median absolute deviation of $\hat{\beta}_\lambda$. For any type of correlation structure between instruments and different variations on endogeneity and s , sisVIVE deviates far from the truth if we incorrectly select the invalid instruments. Hence, it is much more important to correctly select invalid instruments at the expense of incorrectly selecting valid instruments for better estimation of β^* . This relationship makes sense since using invalid instruments creates bias whereas using at least one valid instrument and not using other valid instruments does not create bias, but just reduces efficiency. The relationship also suggests that when we choose the tuning parameter λ , which controls the number of

non-zero $\hat{\alpha}_\lambda$ and consequently, controls the proportion of correctly selected valid and invalid instruments, we should choose λ that correctly selects the invalid instruments, even if some valid instruments are selected as invalid. In particular, λ should generally be small so that there is less ℓ_1 penalty on $\|\alpha\|_1$, but not too small so that the penalty has no effect. As a result, few elements of $\hat{\alpha}_\lambda$ will be zero and more instruments will be selected as invalid. We discuss the choice of λ in more detail in Section 2.6.

2.4 Varying Instrument Strength

In this section, we extend the simulation study in Section 4 of the main manuscript by considering other types of instrument strength beyond those considered in the main manuscript. Specifically, we look at two cases where the invalid instruments are “stronger” than the valid instruments and the valid instruments are “stronger” than the invalid instruments. To simulate these two new cases, we first fix the concentration parameter, a global/overall measure of instrument strength, similar to the simulation setup in the main manuscript. Second, given a concentration parameter, for the case when the invalid instruments are stronger than the valid instruments, we find γ^* where $\gamma_j^* = 2 * \gamma_k^*$ for $j \in \text{supp}(\alpha^*)$ (i.e. set of invalid instruments) and $k \in \text{supp}(\alpha^*)^C$ (i.e. set of valid instruments). In other words, the γ_j^* s associated with invalid instruments have twice the magnitude of the γ_k^* s associated with the valid instruments. For the case when the valid instruments are stronger than the invalid instruments, we flip the roles of j and k where j now belongs to $\text{supp}(\alpha^*)^C$ and k belongs to $\text{supp}(\alpha^*)$. Finally, we rerun the simulation setup in Section 4 of the main manuscript and Sections 2.2 and 2.3 of the Supplementary Materials, except we replace the “Equal” and “Variable” strengths with the two new types of instrument strength introduced in this Section, denoted as “Stronger Invalid” (i.e. the case when the invalid instruments are stronger than the valid instruments) and “Stronger Valid” (i.e. the case when the valid instruments are stronger than the invalid instruments). We also reduce the number of simulations 1000

to 500 for computational reasons.

In addition, for each of the simulation setups, we repeat the exercise we did in Section 2.1 of the Supplementary Materials where we compute ρ and μ that appear in Corollary 2 of the main manuscript. Table 3 and 4 show the results when the instruments have the identical pairwise correlation; for other correlation structures, the condition on s is similar and hence, they are not presented (see Section 2.2 of the Supplementary Materials for discussion on this). The column and row labels in the two tables are identical as those found in Section 2.2 of the Supplementary Materials, except the new headings “Stronger Invalid” and “Stronger Valid.”

Table 3. Values of ρ defined in Corollary 2 for the Simulation Study

Instrument Corr. (μ)	Strong Instrument, Stronger Invalid	Strong Instrument, Stronger Valid	Weak Instrument, Stronger Invalid	Weak Instrument, Stronger Valid
0	0.41	0.33	0.28	0.18
0.25	0.60	0.54	0.47	0.33
0.5	0.75	0.71	0.64	0.49
0.75	0.88	0.86	0.81	0.70

Table 4. Condition on s in Corollary 2 for the Simulation Study

Instrument Corr. (μ)	Strong Instrument, Stronger Invalid	Strong Instrument, Stronger Valid	Weak Instrument, Stronger Invalid	Weak Instrument, Stronger Valid
0	0.60	0.90	1.27	3.02
0.25	0.28	0.33	0.33	0.33
0.5	0.17	0.17	0.17	0.17
0.75	0.11	0.11	0.11	0.11

Figures 11 to 14 represent the cases where the instruments have identical pairwise correlation μ . When we vary endogeneity, but fix $s = 3$ (Figure 11), sisVIVE performs as well as the oracle for strong instruments. For weak instruments, sisVIVE does better when the valid instruments are stronger than the invalid instruments (i.e. “Stronger Valid”) than when the invalid instruments are stronger than the valid instruments (i.e. “Stronger Invalid”). In both the strong and weak cases, sisVIVE does much better than the next best alternative, naive two stage least squares.

When we vary s , but fix endogeneity to 0.8 (Figure 12), sisVIVE deviates from the oracle at $s = 4$ for the case when the invalid instruments are stronger than the valid instruments (i.e. “Stronger Invalid”) and at $s = 7$ for the case when the valid instruments are stronger than the invalid instruments (i.e. “Stronger Valid”). When sisVIVE deviates from oracle TSLS, sisVIVE’s performance is no worse than naive two stage least squares.

When we look at the proportion-based error metrics for estimating α_λ^* (Figures 13 and 14), the behavior of the two curves are similar to what we observed in Section 2.3. That is, whenever sisVIVE performs badly, there is a large decrease in the proportion of correctly selected invalid instruments. Also, there is no relationship between sisVIVE’s median absolute bias of $\hat{\beta}_\lambda$ and the proportion of correctly selected valid instruments. When we vary endogeneity (Figure 13), the proportion of correctly selected invalid instruments remain at 1 except when the overall strength of the instruments is weak and the invalid instruments are stronger than the valid instruments (i.e. “Stronger Invalid”). However, in all cases, a smaller median absolute deviation in Figure 11 corresponds with having a high proportion of correctly selected invalid instruments in Figure 13. In contrast, the proportion of correctly selected valid instruments remains below 1 if the invalid instruments are stronger than the valid instruments (i.e. “Stronger Invalid”) and close to 1 if the valid instruments are stronger than the invalid instruments (i.e. “Stronger Valid”).

Similarly, when we vary s (Figure 14) and are under the case where the invalid instruments are stronger than the valid instruments (i.e. “Stronger Invalid”), the proportion of correctly selected invalid instruments move away from 1 at $s = 4$ when the overall strength of the instruments is strong and at $s = 3$ when the overall strength of the instruments is weak. When the valid instruments are stronger than the invalid instruments (i.e. “Stronger Valid”), the proportion of correctly selected invalid instruments move away from 1 at $s = 7$ for strong instruments and $s = 6$ for weak instruments. Again, similar to what we observed in Section 2.3 of the Supplementary Materials, these points of s correspond to sisVIVE’s deviation from

the oracle in Figure 12. In contrast, the proportion of correctly selected valid instruments vary widely in Figure 14 and there does not seem to be any relationship between it and sisVIVE’s deviation from the oracle.

For other correlation structures, specifically when (i) there is only correlation within valid and invalid instruments, and (ii) there is only correlation between valid and invalid instruments, we observe the same phenomena as the case where all the instruments are correlated. This is in alignment with Sections 2.2 and 2.3. The result from the two correlation structures under the different types of instrument strengths considered in this Section are in Figures 15 to 22.

The simulation study in this Section showed that in vast majority of cases, sisVIVE estimates the causal effect of interest better than the next best alternative, naive two stage least squares and in many cases, sisSIVE’s performance is similar to the oracle. However, when the invalid instruments are stronger than the valid instruments (i.e. “Stronger Invalid”), sisVIVE’s performance does not do as well relative to the oracle, even though by the identification result in Corollary 1 of the main manuscript, at $s = 4$, identification is guaranteed. The degradation in performance of sisVIVE may be due to a number of reasons. It may follow from the fact that the condition in Corollary 2 are not met since Table 4 shows that in the “Stronger Invalid” case, s has to be less than 1 or 2. It may be that we chose a bad tuning parameter λ ; based on the results on the proportion of correctly selected invalid instruments, we may need a smaller λ than what we used was chosen by cross validation. A closer analysis of this particular case more closely is a topic for future research. Regardless, even when sisVIVE’s performance degrades, it does no worse than the next best alternative, naive two stage least squares.

In addition, the simulation study reaffirmed the points mentioned in Sections 2.2 and 2.3 of the Supplementary Materials that (i) sisVIVE seems to do well under different correlation structures, and (ii) $\hat{\beta}_\lambda$ ’s deviation from β^* depends heavily on the proportion of

correctly selected invalid instruments more so than the proportion of correctly selected valid instruments.

2.5 Number of potential instruments

In this section, we extend the simulation study in Section 4 of the main manuscript by increasing the potential number of instruments from $L = 10$ to $L = 100$. We note that in Mendelian randomization settings, it is rare to have 100 potential genetic instruments since all 100 of the genetic instruments must affect the exposure (see the Introduction and Section 3.1 of the main manuscript for details). Usually, the number of potential instruments is far less than 100 (see citations in the Introduction of our main manuscript for examples). However, for completeness, we demonstrate sisVIVE’s performance when $L = 100$ potential instruments are present.

We rerun the simulation setup in Section 4 of the main manuscript and Section 2.3 in the Supplementary Materials except $L = 100$ and when we vary endogeneity, we fix the number of invalid instruments to be 30 (instead of 3); note that based on the simulation results in Section 2.2 where other correlation structures did not impact the performance of sisVIVE, we only consider the correlation structure in the main manuscript, specifically where all the instruments are correlated to each other with pairwise correlation μ . Also, for computational reasons, we reduce the simulation number from 1000 to 500. Finally, we repeat the exercise in Section 2.1 by computing ρ and μ defined in Corollary 2. Table 5 and 6 show the results.

Table 5. Values of ρ defined in Corollary 2 for the Simulation Study

Instrument Corr. (μ)	Strong Instrument, Equal Strength	Strong Instrument, Variable Strength	Weak Instrument, Equal Strength	Weak Instrument, Variable Strength
0	0.15	0.17	0.16	0.17
0.25	0.54	0.54	0.53	0.53
0.5	0.73	0.73	0.53	0.73
0.75	0.87	0.87	0.88	0.87

Table 6. Condition on s in Corollary 2 for the Simulation Study

Instrument Corr. (μ)	Strong Instrument, Equal Strength	Strong Instrument, Variable Strength	Weak Instrument, Equal Strength	Weak Instrument, Variable Strength
0	4.2	3.3	4.0	3.4
0.25	0.33	0.33	0.33	0.33
0.5	0.17	0.17	0.17	0.17
0.75	0.11	0.11	0.11	0.11

Figures 23 and 24 represent the results from the simulation setup when we measure the median of $|\hat{\beta}^* - \beta^*|$ over 500 simulations; this setup is identical to Section 4 in the main manuscript except for the exceptions mentioned in the previous paragraph. The behavior of all four estimators are similar to Figures 2 and 3 in the main manuscript. For example, when we vary endogeneity (Figure 23), sisVIVE tends to perform slightly worse when the overall strength of the instruments is weak. Also, when the number of invalid instruments, s , is varied (Figure 24), sisVIVE has a sharp peak at $s = 50$, similar to the sharp peak at $s = 5$ in Figures 3 of the main manuscript.

Figures 25 and 26 represent the simulation setups in Section 2.3 of the Supplementary Materials. Similar to what we observed in Section 2.3 when $L = 10$, when we vary endogeneity (Figure 25), but fix the number of invalid instruments to 30, we see that the proportion of correctly selected invalid instruments are 1. When we vary s (Figure 26), we again notice a sharp decrease in the proportion of correctly selected valid invalid instruments around $s = 50$ for all instrument strength and magnitude of the correlation.

Overall, the simulation study suggests that sisVIVE does scale as L increases and that its performance at large values of L is similar to its performance at smaller values of L , such as $L = 10$.

2.6 Choice of λ

In this section, we look at different ways to select λ . As discussed in the main manuscript, the choice of λ impacts the performance of sisVIVE where a high value of λ will push most

elements of $\hat{\alpha}_\lambda$ to zero while a low value of λ will do the opposite. In Section 3.3 of the main manuscript, we suggested cross-validation with the “one standard error” rule as a data-driven method to choosing the tuning parameter. In addition, in Section 3.4, we provided theoretical results which suggested choosing a λ that is greater than $3\|\mathbf{Z}^T\mathbf{P}_{\hat{\mathbf{D}}^\perp}\boldsymbol{\epsilon}\|_\infty$. We explore these two possible choices of λ and their impact on estimation.

We begin with a simulation study similar to the one in the main manuscript. In particular, we have $L = 10$ instruments of which the pairwise correlation between all instruments is 0.75 and the endogeneity is fixed at 0.8. We vary s , the number of invalid instruments and vary instruments’ absolute strength, relative strength, and other strengths considered in Section 2.4 of the Supplementary Materials. In short, the simulation setups we consider correspond to the last row of Figure 3 in the main manuscript and the last row of Figure 12 in the Supplementary Materials. We do not simulate other correlation structures or different L because the simulation results in Sections 2.2 and 2.5 of the Supplementary Materials showed sisVIVE behaves similarly as the cases we consider in this Section.

Table 7 shows the different values of λ averaged across 500 simulations where the overall, absolute instrument strength is strong (see Section 4 of the main manuscript for details on the definition of an absolute instrument strength). We use the same column heading labels in Figure 3 of the main manuscript and Figure 12 in the Supplementary Materials. We also use the column labeled “CV” to denote the average λ s based on cross validation laid out in Section 3.3 of the main manuscript. Also, the column labeled “Theory” denotes the average λ s based on Theorem 2, specifically the average of $3\|\mathbf{Z}^T\mathbf{P}_{\hat{\mathbf{D}}^\perp}\boldsymbol{\epsilon}\|_\infty$ over 500 simulations. In almost all cases, cross validation tends to choose a smaller λ than one prescribed by Theorem 2, with the exception of $s = 9$ in the “Equal” column and $s = 7, 8$, and 9 in the “Stronger Valid” column. Except for these cases, cross validation tends to prefer a small λ , thereby preferring $\hat{\alpha}_\lambda$ to have more non-zero entries than zero entries and more instruments selected as invalid instruments than valid instruments.

Table 7. Average λ from cross validation and Theorem 2 after 500 simulations for instruments whose overall strength is strong.

s	Equal		Variable		Stronger Invalid		Stronger Valid	
	CV	Theory	CV	Theory	CV	Theory	CV	Theory
1	1.88	2.70	2.04	2.71	1.53	2.70	2.06	2.72
2	1.36	2.66	1.39	2.67	0.95	2.65	1.58	2.68
3	1.06	2.64	1.12	2.66	0.84	2.64	1.33	2.68
4	0.84	2.64	0.86	2.65	1.08	2.63	1.16	2.68
5	1.70	2.63	1.33	2.64	0.87	2.62	0.99	2.67
6	1.78	2.62	1.10	2.63	0.85	2.61	0.96	2.67
7	2.02	2.62	0.79	2.64	0.91	2.61	3.40	2.68
8	2.41	2.62	0.86	2.62	1.01	2.61	3.74	2.67
9	3.19	2.62	0.45	2.62	1.31	2.60	6.03	2.67

Table 8 shows the estimation performance of sisVIVE, the median of $|\beta^* - \hat{\beta}_\lambda|$ over 500 simulations, based on two different λ s, one based on cross validation and one based on Theorem 2. In most cases, sisVIVE with a cross validated λ performs just as well as sisVIVE with a theory-based λ . For the “Equal” and “Variable” case, when $s < 5$, sisVIVE with a cross-validated λ performs better than sisVIVE with a theory-based λ . For the “Stronger Invalid” case, when $s < 3$, sisVIVE with a cross validated λ performs better than sisVIVE with a theory-based λ . However, when $s \geq 3$, sisVIVE with a cross validated λ performs worse than sisVIVE with a theory-based λ , although the differences between the two decrease as s increases. For the “Stronger Valid” case, sisVIVE with a cross validated λ always dominates sisVIVE with a theory-based λ , although the differences between the two are slight when $s \geq 7$.

Table 9 considers the same setup as Table 7, except we now look at instruments where their overall, absolute strength is weak. Under this case, we see drastic differences between λ s chosen based on cross validation and Theorem 2. For example, for the “Equal” and “Variable” cases, when $s < 5$, λ chosen based on cross validation is, on average, smaller than λ chosen based on Theorem 2. When $s \geq 5$, λ chosen based on cross validation is, on average, bigger than λ chosen based on Theorem 2. For the “Stronger Invalid” case, when $s < 3$, λ based on cross validation is, on average, smaller than λ based on Theorem 2. But,

Table 8. Median absolute estimation error ($|\beta^* - \hat{\beta}_\lambda|$) after 500 simulations from λ chosen by cross-validation and Theorem 2. The table only considers instruments whose overall strength is strong.

s	Equal		Variable		Stronger Invalid		Stronger Valid	
	CV	Theory	CV	Theory	CV	Theory	CV	Theory
1	0.13	0.17	0.14	0.16	0.13	0.19	0.14	0.16
2	0.16	0.27	0.16	0.27	0.16	0.34	0.16	0.24
3	0.18	0.39	0.18	0.37	0.24	0.54	0.18	0.32
4	0.21	0.53	0.22	0.53	1.57	1.34	0.20	0.41
5	0.71	1.15	0.76	1.43	1.43	1.25	0.23	0.55
6	2.43	2.34	2.05	1.93	1.35	1.23	0.28	0.71
7	2.42	2.37	1.83	1.95	1.28	1.21	3.83	3.95
8	2.35	2.34	1.98	2.05	1.22	1.18	4.24	4.39
9	2.29	3.01	1.23	1.37	1.17	1.16	4.34	4.51

when $s \geq 3$, the opposite is the case. Finally, for the “Stronger Valid” case, this phenomena occurs at $s = 6$.

Table 9. Average λ from cross validation and Theorem 2 after 500 simulations for instruments whose overall strength is weak.

s	Equal		Variable		Stronger Invalid		Stronger Valid	
	CV	Theory	CV	Theory	CV	Theory	CV	Theory
1	1.36	3.20	1.56	3.23	1.05	3.13	1.52	3.24
2	1.25	3.00	1.22	3.01	0.93	2.92	1.47	3.07
3	1.12	2.91	1.11	2.94	3.67	2.81	1.26	3.00
4	2.06	2.86	1.83	2.89	9.47	2.75	1.13	2.97
5	6.30	2.80	4.34	2.84	10.52	2.71	1.20	2.92
6	11.99	2.78	7.48	2.80	10.74	2.69	3.36	2.93
7	14.14	2.76	5.92	2.77	10.58	2.67	7.79	2.93
8	14.04	2.75	5.94	2.75	9.92	2.66	9.70	2.93
9	13.16	2.74	2.02	2.68	9.47	2.64	7.09	2.96

Table 10 considers the same setup as Table 8, except we now look at instruments where their overall, absolute strength is weak. Similar to Table 8, sisVIVE with a cross validated λ performs better than sisVIVE with a theory-based λ , with the only exception at $s = 5$ under “Equal” column. In fact, sisVIVE with a cross validated λ performs drastically better than sisVIVE based on Theorem 2 in the following cases: $s < 5$ (for “Equal” and “Variable” cases), $s < 3$ (for “Stronger Invalid” case), and $s < 7$ (for “Stronger Valid” case).

Based on these simulations, sisVIVE based on cross-validation generally performs better

Table 10. Median absolute estimation error ($|\beta^* - \hat{\beta}_\lambda|$) after 500 simulations from λ chosen by cross-validation and Theorem 2. The table only considers instruments whose overall strength is weak.

s	Equal		Variable		Stronger Invalid		Stronger Valid	
	CV	Theory	CV	Theory	CV	Theory	CV	Theory
1	0.44	0.63	0.44	0.60	0.43	0.69	0.44	0.61
2	0.51	0.96	0.50	0.94	0.50	1.13	0.52	0.88
3	0.55	1.30	0.55	1.26	0.70	1.86	0.56	1.13
4	0.61	1.74	0.61	1.75	3.19	3.77	0.58	1.43
5	4.10	3.80	3.98	3.93	3.25	3.78	0.62	1.83
6	5.28	6.03	5.28	5.54	3.36	3.79	0.73	2.52
7	5.84	6.55	5.58	5.63	3.47	3.77	7.51	7.68
8	6.29	6.75	6.19	6.19	3.52	3.70	9.69	9.77
9	6.72	6.90	4.18	4.34	3.56	3.64	10.86	10.91

than sisVIVE based on Theorem 2, especially when the overall instrument strength is weak. We also note that cross validation tends to choose a smaller λ than the one based on Theorem 2, suggesting that for better estimation, it is preferable to set only a few elements of $\hat{\alpha}_\lambda$ to zero and declare more instruments to be invalid than valid. This observation was also seen in our simulation in Section 2.3 where low median absolute error, $|\beta^* - \hat{\beta}_\lambda|$, was tied to high proportion of correctly chosen invalid instruments. We note that this observation is in contrast with estimating sparse vectors in typical high dimensional regression settings where many zeroed elements are desirable in the estimated sparse vector.

Despite the simulation evidence suggesting the use of cross validation to choose λ over Theorem 2 to choose λ , unfortunately, there is little theory to justify the use of cross validation in ℓ_1 penalization settings (Hastie et al. 2009; Bühlmann and van de Geer 2011). However, Section 2.5.1 of Bühlmann and van de Geer (2011) does provide limited theoretical results suggesting that λ based on cross validation tends to set few elements of $\hat{\alpha}_\lambda$ to zero, a desirable property in our setting where we want to select more instruments to be invalid than valid for better estimation performance of $\hat{\beta}_\lambda$.

Besides cross validation and Theorem 2, there is another way to choose λ if we assume Corollary 1 holds for our data. That is, if we are in the always identified region where

$s < U \leq L/2$, one possible method of choosing λ would be to find the λ where exactly $U = L/2$, say $\lambda_{L/2}$. From there, we grid the values of potential λ s between 0 and $\lambda_{L/2}$ and choose the λ that minimizes the estimating equation $\|\mathbf{P}_Z(\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha} - \mathbf{D}\beta)\|_2$. It would be interesting to investigate this method in future research.

3 Additional Discussion about Theorem 2

Theorem 2 is written in terms of the restricted isometry type (RIP) condition while its corresponding Corollary 2 is written in terms of the mutual incoherence property (MIP) condition. As the main text states, the RIP condition implies the MIP condition, but not vice versa. We illustrate this relationship with the following simple example. Suppose the matrix of instruments \mathbf{Z} is an n by L matrix where each entry Z_{ij} are from i.i.d. standard Normal. Based on Theorem 5.2 in Baraniuk et al. (2008), when $n \geq Cs \log(L/s)$ for some C not dependent on L and s , we are able to ensure the RIP condition $2\delta_{2s}^-(\mathbf{Z}) > 3\delta_{3s}^+(\mathbf{Z})$ with high probability. Here, $2\delta_{2s}^-(\mathbf{Z}) > 3\delta_{3s}^+(\mathbf{Z})$ is a stronger condition than $2\delta_{2s}^-(\mathbf{Z}) > \delta_{3s}^+(\mathbf{Z}) + 2\delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z})$, the RIP condition we need for Theorem 2. However, based on Theorem 8 in Cai et al. (2013), to guarantee our MIP condition $\mu < \frac{1}{12s}$, we need $n \geq Cs^2 \log L$ for some C not dependent on L and s . In short, when the order of n is between $s \log(L/s)$ and $s^2 \log L$, \mathbf{Z} meet the RIP condition but not the MIP condition, with high probability.

4 Wisconsin Longitudinal Data

4.1 Background of Data

This research uses data from the Wisconsin Longitudinal Study (WLS) of the University of Wisconsin-Madison. Since 1991, the WLS has been supported principally by the National Institute on Aging (AG-9775, AG-21079, AG-033285, and AG-041868), with additional sup-

port from the Vilas Estate Trust, the National Science Foundation, the Spencer Foundation, and the Graduate School of the University of Wisconsin-Madison. Since 1992, data have been collected by the University of Wisconsin Survey Center. A public use file of data from the Wisconsin Longitudinal Study is available from the Wisconsin Longitudinal Study, University of Wisconsin-Madison, 1180 Observatory Drive, Madison, Wisconsin 53706 and at <http://www.ssc.wisc.edu/wlsresearch/data/>. The opinions expressed herein are those of the authors.

4.2 Reduced form estimates

Tables 11 and 12 summarize the reduced form estimates for the data analysis in the main manuscript. The reduced form estimates are computed by using ordinary least squares (OLS) where the genetic instruments are the explanatory variables and the dependent variables is the exposure variable used in Section 5 of the main manuscript and Health Utility Index Mark 3 (HUI-3).

Table 11. Reduced Form Estimates for HUI-3 and BMI for Three Instruments

Instruments	BMI (SE)	HUI-3 (SE)
rs1421085	-0.05 (0.02)	0.0003 (0.004)
rs1501299	0.01 (0.02)	0.002 (0.005)
rs2241766	-0.0007 (0.03)	-0.0001 (0.007)

Table 12. Reduced Form Estimates for HUI-3 and BMI for Four Instruments

Instruments	BMI (SE)	HUI-3 (SE)
rs1421085	-0.05 (0.02)	0.0004 (0.004)
rs1501299	0.01 (0.02)	0.002 (0.005)
rs2241766	-0.0006 (0.03)	-0.0004 (0.007)
rs6265	-0.004 (0.02)	-0.008 (0.005)

4.3 First stage F statistic and structural correlations

The first stage F statistic with three instruments and the binary exposure in Table 11 is 3.16. The first stage F statistic with four instruments and the exposure BMI in Table 12 is

2.38. Based on the two F statistics, the instruments are generally weak.

We also estimate the implied structural correlation from our model, specifically the correlation between D_i , the exposure, and ϵ_i . We estimate ϵ_i by taking the residual from the estimates of β^* and α^* , $\hat{\epsilon}_i = Y_i - D_i\hat{\beta}_\lambda - \mathbf{Z}_i^T\hat{\alpha}_\lambda$ where λ is chosen by cross-validation described in Section 3.3 of the main manuscript. We find that our estimate of this correlation is -0.2 , suggesting a mild form of endogeneity.

4.4 Sargan overidentification test

For the data analysis with three SNPs, the Sargan overidentification test Sargan (1958), which tests assumptions (A2) and (A3) in the presence of multiple instruments, gives a Chi-squared value of 0.12 (p-value: 0.94), retaining the null hypothesis that the instruments are all valid under the 0.05 significance level. For the data analysis with four SNPs, the Sargan overidentification test gives a Chi-squared value of 2.53 (p-value: 0.47).

4.5 Quantifying obesity with BMI

To quantify obesity using BMI, we looked at BMI across several categories of obesity. The categories were based on US National Institute of Health clinical guidelines (National Institute of Health 1998) and were also used in Trakas et al. (2001) and Sach et al. (2007) in their analysis. Table 1 summarizes the different classes of obesity and their associations to HUI-3.

Table 13. Relationship Between Obesity and Health Utility Index Mark 3 (HUI-3)

Obesity Categories	N	Health Utility Index Mark 3		
		1st quartile	Median	3rd quartile
Not obese (BMI < 30)	2581	0.84	0.92	0.97
Obese class I ($30 \leq \text{BMI} < 35$)	777	0.73	0.91	0.97
Obese class II ($35 \leq \text{BMI} < 40$)	246	0.66	0.85	0.97
Obese class III ($40 \leq \text{BMI}$)	108	0.51	0.72	0.91
All categories	3712	0.78	0.92	0.97

We notice that among different obese classes, the median HUI-3 scores are different. Hence, simply classifying individuals as obese vs. not obese ignores the variation of HUI-3 scores among different obese classes. To be comparable to prior literature, the method we use in the main manuscript to quantify obesity is the closest equivalent to Trakas et al. (2001), which also used our utility score measure HUI-3. We also explore different methods of quantifying obesity through BMI. Specifically, we consider the following methods listed below.

1. The binary BMI is takes a value of one if BMI is greater than or equal to 30 (i.e. obese) and zero otherwise.
2. The exposure BMI A is what we use in the main manuscript.
3. The exposure BMI B is defined to be similar to Trakas et al. (2001), except the magnitude of the BMIs is taken into consideration. Specifically, if an individual's BMI is less than 30, the individual's exposure is assigned a value of zero. If an individual's BMI is between 30 and 35 (i.e. Obese Class I), the individual's exposure is assigned a value of one. If an individual's BMI is between 35 and 40 (i.e. Obese Class II), the individual's exposure is assigned a value of three. If an individual's BMI is above 40 (i.e. Obese Class III), the individual's exposure is assigned a value of six.
4. The censored BMI takes into account the actual value of BMI at the obese range so that it not only indicate obesity, but also to measure its severity. Specifically, the censored BMI is defined the maximum of $(\text{BMI} - 30)$ and 0 (i.e. $\max(\text{BMI} - 30, 0)$).

For each method of quantifying obesity, we estimate β^* by using ordinary least squares (OLS), two stage least squares (TSLS) under the assumption that all the instruments are valid, and sisVIVE. These results are reported in Tables 14 and 15 for the cases of three and four instruments studied in Section 5 of the main manuscript. Overall, we notice that the estimates of OLS, TSLS, and sisVIVE tend to be similar across different types of exposures.

Granted, it is difficult to compare the estimates since each exposure variable measures slightly different aspects about obesity and its impact on HUI-3. We also note that in the case of four instruments where one of the instrument, rs6265, was suspect, sisVIVE correctly picks rs6265 to be an invalid instrument in every method of quantifying obesity.

Table 14. Different Exposures and Their Estimates With Three Instruments

Exposure	OLS (SE)	TSLs (SE)	sisVIVE, Invalid Instrument
Binary BMI	-0.074 (SE: 0.0070)	-0.012 (SE: 0.18)	-0.012, None
BMI A	-0.052 (SE: 0.0040)	-0.00094 (SE: 0.081)	-0.00094, None
BMI B	-0.031 (SE: 0.0024)	-0.0011 (SE: 0.051)	-0.0011, None
Censored BMI	-0.013 (SE: 0.0010)	-0.00019 (SE: 0.022)	-0.00019, None

Table 15. Different Exposures and Their Estimates With Four Instruments

Exposure	OLS (SE)	TSLs (SE)	sisVIVE, Invalid Instrument
Binary BMI	-0.074 (SE: 0.0070)	-0.097 (SE: 0.17)	-0.039, rs6265
BMI A	-0.052 (SE: 0.0040)	-0.0086 (SE: 0.080)	-0.0037, rs6265
BMI B	-0.031 (SE: 0.0024)	-0.0012 (SE: 0.051)	-0.0017, rs6265
Censored BMI	-0.013 (SE: 0.0010)	0.00091 (SE: 0.022)	-0.00011, rs6265

5 Proofs

We adopt the following notations for the proofs. For any sets $A, B \subseteq \{1, \dots, L\}$, denote $A \cap B$ to be the intersection of sets A and B , $A \cup B$ to be the union of sets A and B , and A^C and B^C to be the complement of sets A and B , respectively. If $A \subseteq B$, denote $B \setminus A$ to be the set that comprises of all the elements of B except those that are in A . Let $|A|$ and $|B|$ denote the cardinality of the sets A and B , respectively.

For any vector $\alpha \in \mathbb{R}^L$ and set $A \subseteq \{1, \dots, L\}$, denote $\alpha_A \in \mathbb{R}^L$ to be the vector where all the elements except whose indices are in A are zero. Also, denote the j th element as α_j . Let $\text{supp}(\alpha) \subseteq \{1, \dots, L\}$ to be the support of the vector α and $\text{supp}(\alpha)^C$ be the complement set. For any matrix $\mathbf{M} \in \mathbb{R}^{n \times L}$ and set $A \subseteq \{1, \dots, p\}$, let $\mathbf{M}_A \in \mathbb{R}^{n \times L}$ be an n by $|A|$ matrix where the columns are specified by set A .

5.1 Proof of Theorem 1

First, we prove that, β^* is a unique solution if and only if α^* is a unique solution. Suppose β^* has a unique solution; that is, for any two solutions $\alpha^{(1)}, \beta^{(1)}$ and $\alpha^{(2)}, \beta^{(2)}$, in equation (7)

$$\alpha^{(1)} + \gamma^* \beta^{(1)} = \Gamma^* \quad (19a)$$

$$\alpha^{(2)} + \gamma^* \beta^{(2)} = \Gamma^* \quad (19b)$$

we have $\beta^{(1)} = \beta^{(2)}$. Subtracting $\gamma^* \beta^{(1)}$ from equations (19) gives $\alpha^{(1)} = \alpha^{(2)}$. Now, suppose α^* is unique, which implies $\alpha^{(1)} = \alpha^{(2)}$. Again, subtracting $\alpha^{(1)}$ from (19) reveals $\beta^{(1)} = \beta^{(2)}$.

Second, we prove the necessary and sufficient conditions for Theorem 1. Suppose the subspace conditions on γ^* and Γ^* hold, specifically $q_m = q_{m'}$ for any $m \neq m'$, but there are two distinct sets of parameters, $\alpha^{(1)}, \beta^{(1)}$ and $\alpha^{(2)}, \beta^{(2)}$ that solve the moment equation in equation (19). Let $A^{(1)} = \text{supp}(\alpha^{(1)})$ and $A^{(2)} = \text{supp}(\alpha^{(2)})$ be the sets of invalid instruments for the two distinct parameter sets, not equal to each other; if the supports are equal to each other, we have the degenerate case whereby from equation (19), for any $j \in A^{(1)} = A^{(2)}$ $\gamma_j^* \beta^{(1)} = \Gamma_j^*$ and $\gamma_j^* \beta^{(2)} = \Gamma_j^*$, which implies that $\beta^{(1)} = \beta^{(2)}$ and $\alpha^{(1)} = \alpha^{(2)}$, a contradiction. Because the number of invalid instruments, s , is less than U , $s < U$, the number of valid instruments, $L - s$, must be greater than $L - U$, $L - s > L - U$. Thus, $|(A^{(1)})^C|, |(A^{(2)})^C| > L - U$.

Now, pick any subsets, $(A^{(1')})^C$ and $(A^{(2')})^C$, of $(A^{(1)})^C$ and $(A^{(2)})^C$, respectively, where $|(A^{(1')})^C| = |(A^{(2')})^C| = L - U + 1$. These subsets $(A^{(1')})^C$ and $(A^{(2')})^C$ inherit the following property from their larger sets $(A^{(1)})^C$ and $(A^{(2)})^C$, respectively.

$$\alpha_j^{(1)} + \gamma_j^* \beta^{(1)} = \gamma_j^* \beta^{(1)} = \Gamma_j^*, \quad j \in (A^{(1')})^C \subseteq (A^{(1)})^C$$

$$\alpha_k^{(2)} + \gamma_k^* \beta^{(2)} = \gamma_k^* \beta^{(2)} = \Gamma_k^*, \quad k \in (A^{(2')})^C \subseteq (A^{(2)})^C$$

The subspace condition on γ^* and Γ^* in Theorem 1 state that for any sets C_m with size $|C_m| = L - U + 1$ and with the property that $\gamma_j q_m = \Gamma_j, j \in C_m$, we have $q_m = q_{m'}$ for any m, m' . The subsets we constructed, $(A^{(1')})^C$ and $(A^{(2')})^C$, satisfy these subspace condition with constants $q_{1'} = \beta^{(1)}$ and $q_{2'} = \beta^{(2)}$. Hence, $\beta^{(1)} = q_{1'} = q_{2'} = \beta^{(2)}$, which is a contradiction. Hence, the two sets of parameters $\alpha^{(1)}, \beta^{(1)}$ and $\alpha^{(2)}, \beta^{(2)}$ are identical to each other and the solution is unique.

Now, suppose the solution is unique. Then, we show that the subspace conditions on γ^* and Γ^* must hold. Pick any two sets $A^{(1)}, A^{(2)} \subseteq \{1, \dots, L\}$ with their complements having the size $|(A^{(1)})^C| = |(A^{(2)})^C| = L - U + 1$ and corresponding constants q_1 and q_2 , respectively, defined in the Theorem. We have to show that $q_1 = q_2$ for any pair of two sets.

Note that at least one set of these sets and its corresponding constant q must exist because at the true parameter values, α^* and β^* , equation (7) is satisfied. Specifically, if $A^* = \text{supp}(\alpha^*)$ where, by $s < U$, $|(A^*)^C| = |\text{supp}(\alpha^*)^C| > L - U$, we can take any subset $(A^{(*)})^C \subseteq (A^*)^C$ of size $|(A^{(*)})^C| = L - U + 1$. For any $j \in (A^{(*)})^C$, by equation (7), $\gamma_j^* \beta^* = \Gamma_j^*$ and thus, its corresponding constant q_{*} is $q_{*} = \beta^*$. If there is exactly one set $A^{(1)}$, the subspace condition holds automatically.

Suppose there are two or more sets and let $A^{(1)}$ and $A^{(2)}$ be any pair of the sets. Based on the sets $A^{(1)}$ and $A^{(2)}$ and their corresponding constants q_1 and q_2 , we construct the following sets of parameters $\alpha^{(1)}, \beta^{(1)}$ and $\alpha^{(2)}, \beta^{(2)}$

$$\beta^{(1)} = q_1, \quad \alpha_j^{(1)} = \begin{cases} 0 & j \in (A^{(1)})^C \\ \Gamma_j^* - q_1 \gamma_j^* & j \in A^{(1)} \end{cases}$$

$$\beta^{(2)} = q_2, \quad \alpha_j^{(2)} = \begin{cases} 0 & j \in (A^{(2)})^C \\ \Gamma_j^* - q_2 \gamma_j^* & j \in A^{(2)} \end{cases}$$

The cardinality of $\alpha^{(1)}$ and $\alpha^{(2)}$ are less than U . In addition, they satisfy the moment

equation in equation (7).

$$\begin{aligned}\alpha_j^{(1)} + \gamma_j^* \beta^{(1)} &= \begin{cases} \gamma_j^* q_1 = \Gamma_j^* & j \in (A^{(1)})^C \\ \Gamma_j^* - q_1 \gamma_j^* + \gamma_j^* q_1 = \Gamma_j^* & j \in A^{(1)} \end{cases} \\ \alpha_j^{(2)} + \gamma_j^* \beta^{(2)} &= \begin{cases} \gamma_j^* q_2 = \Gamma_j^* & j \in (A^{(2)})^C \\ \Gamma_j^* - q_2 \gamma_j^* + \gamma_j^* q_2 = \Gamma_j^* & j \in A^{(2)} \end{cases}\end{aligned}$$

Since the equation has only one unique solution, this implies that $\beta^{(1)} = \beta^{(2)}$, or $q_1 = q_2$. Since this holds for any two sets $(A^{(1)})^C, (A^{(2)})^C$ with constants q_1 and q_2 and cardinality $L - U + 1$, we arrive at the subspace condition $q_m = q_{m'}$ for any m, m' . \square

5.2 Proof of Corollary 1

Consider any two sets C_m and $C_{m'}$ with the constants q_m and $q_{m'}$ in Theorem 1. Take an element j from the intersection $C_m \cap C_{m'}$; this intersection is non-empty because $|C_m| = |C_{m'}| = L - U + 1 \geq L/2 + 1$. At element $j \in C_m \cap C_{m'}$, we have $\gamma_j^* q_m = \Gamma_j^*$ and $\gamma_j^* q_{m'} = \Gamma_j^*$, which implies $q_m = q_{m'}$. Since this holds for any two sets C_m and $C_{m'}$, $q_m = q_{m'}$ for m, m' , the subspace restriction condition in Theorem 1 always holds whenever $U \geq L/2$ and we have identification. \square

5.3 Proof of Theorem 2

We begin by introducing some notations and terminologies. For $\alpha \in \mathbb{R}^p$ and $s \in \{1, \dots, p\}$, $\alpha_{\max(s)}$ is defined as the vector where all but the largest s elements set to zero and $\alpha_{-\max(s)}$ is defined as $\alpha - \alpha_{\max(s)}$.

Definition 1. The restricted orthogonal constant (ROC) of single matrix of order k_1 and k_2 , denoted as $\theta_{k_1, k_2}(\mathbf{M})$, is the smallest $\theta_{k_1, k_2}(\mathbf{M})$ where for any k_1 -sparse vector α_1 and

k_2 -sparse vector $\boldsymbol{\alpha}_2$ with non-overlapping support, we have

$$|\langle \mathbf{M}\boldsymbol{\alpha}_1, \mathbf{M}\boldsymbol{\alpha}_2 \rangle| \leq \theta_{k_1, k_2}(\mathbf{M}) \|\boldsymbol{\alpha}_1\|_2 \|\boldsymbol{\alpha}_2\|_2.$$

Next, we introduce two lemmas. The first Lemma relates the RIP and ROC constants.

Lemma 1. *For any matrix \mathbf{M} and positive integers s_1 and s_2 ,*

$$\theta_{s_1, s_2}(\mathbf{M}) \leq \frac{1}{2} (\delta_{s_1+s_2}^+(\mathbf{M}) - \delta_{s_1+s_2}^-(\mathbf{M})).$$

Proof. For any vectors x and y with disjoint supports and $\|x\|_2 = \|y\|_2 = 1$, we must have $x + y$, $x - y$ are both $(s_1 + s_2)$ -sparse and $\|x + y\|_2^2 = \|x - y\|_2^2 = 2$. Hence,

$$\begin{aligned} |\langle \mathbf{M}x, \mathbf{M}y \rangle| &= \frac{1}{4} \left| \|\mathbf{M}(x + y)\|_2^2 - \|\mathbf{M}(x - y)\|_2^2 \right| \\ &= \frac{1}{4} \max \left\{ \|\mathbf{M}(x + y)\|_2^2 - \|\mathbf{M}(x - y)\|_2^2, \|\mathbf{M}(x - y)\|_2^2 - \|\mathbf{M}(x + y)\|_2^2 \right\} \\ &\leq \frac{1}{4} \max \left\{ \delta_{s_1+s_2}^+(\mathbf{M}) \|x + y\|_2^2 - \delta_{s_1+s_2}^-(\mathbf{M}) \|x - y\|_2^2, \right. \\ &\quad \left. \delta_{s_1+s_2}^+(\mathbf{M}) \|x - y\|_2^2 - \delta_{s_1+s_2}^-(\mathbf{M}) \|x + y\|_2^2 \right\} \\ &\leq \frac{1}{2} (\delta_{s_1+s_2}^+(\mathbf{M}) - \delta_{s_1+s_2}^-(\mathbf{M})), \end{aligned}$$

which implies $\theta_{s_1, s_2}(\mathbf{M}) \leq \frac{1}{2} (\delta_{s_1+s_2}^+(\mathbf{M}) - \delta_{s_1+s_2}^-(\mathbf{M}))$. □

The second Lemma proves a standard property of the Lasso.

Lemma 2. *Suppose we have the model $Y_i = \mathbf{Z}_{i\cdot}^T \boldsymbol{\alpha}^* + \epsilon_i$ where $\boldsymbol{\alpha}^*$ is s -sparse. Further suppose that matrix \mathbf{Z} has upper and lower RIP constants $\delta_s^+(\mathbf{Z})$ and $\delta_s^-(\mathbf{Z})$, respectively. Define $\hat{\boldsymbol{\alpha}}$ as the Lasso estimator*

$$\hat{\boldsymbol{\alpha}}_\lambda = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \frac{1}{2} \|Y - \mathbf{Z}\boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1 \quad (20)$$

and let $h = \hat{\boldsymbol{\alpha}}_\lambda - \boldsymbol{\alpha}^*$ measure the errors of the estimator.

If $r\|\mathbf{Z}^T\boldsymbol{\epsilon}\|_\infty \leq \lambda$ for some $r > 1$, we have

$$\|h_{-\max(s)}\|_1 \leq \frac{r+1}{r-1}\|h_{\max(s)}\|_1. \quad (21)$$

Furthermore, if $(r+1)\delta_{2s}^+(\mathbf{Z}) < (3r-1)\delta_{2s}^-(\mathbf{Z})$,

$$\|h_{\max(s)}\|_2 \leq \frac{2\lambda\sqrt{s}(r-1)(r+1)/r}{(3r-1)\delta_{2s}^-(\mathbf{Z}) - (r+1)\delta_{2s}^+(\mathbf{Z})}. \quad (22)$$

Proof. Since $\hat{\boldsymbol{\alpha}}_\lambda$ is the minimizer of (20), we have

$$\frac{1}{2}\|Y - \mathbf{Z}\hat{\boldsymbol{\alpha}}_\lambda\|_2^2 + \lambda\|\hat{\boldsymbol{\alpha}}_\lambda\|_1 \leq \frac{1}{2}\|y - \mathbf{Z}\boldsymbol{\alpha}^*\|_2^2 + \lambda\|\boldsymbol{\alpha}^*\|_1.$$

By the assumed model $Y_i = \mathbf{Z}_i^T \boldsymbol{\alpha}^* + \epsilon_i$, we have

$$\frac{1}{2}(\|\boldsymbol{\epsilon} - \mathbf{Z}h\|_2^2 - \|\boldsymbol{\epsilon}\|_2^2) \leq \lambda(\|\boldsymbol{\alpha}^*\|_1 - \|\hat{\boldsymbol{\alpha}}_\lambda\|_1). \quad (23)$$

For the upper bound of (23), the fact that $\boldsymbol{\alpha}^*$ is s -sparse gives a useful bound. Specifically,

$$\begin{aligned} \|\boldsymbol{\alpha}^*\|_1 - \|\hat{\boldsymbol{\alpha}}_\lambda\|_1 &= \|\boldsymbol{\alpha}_{\text{supp}(\boldsymbol{\alpha}^*)}^*\|_1 - \|\hat{\boldsymbol{\alpha}}_{\text{supp}(\boldsymbol{\alpha}^*)}\|_1 - \|\hat{\boldsymbol{\alpha}}_{\text{supp}(\boldsymbol{\alpha}^*)^c}\|_1 \\ &\leq \|\boldsymbol{\alpha}_{\text{supp}(\boldsymbol{\alpha}^*)}^* - \hat{\boldsymbol{\alpha}}_{\text{supp}(\boldsymbol{\alpha}^*)}\|_1 - \|h_{\text{supp}(\boldsymbol{\alpha}^*)^c}\|_1 \\ &\leq \|h_{\text{supp}(\boldsymbol{\alpha}^*)}\|_1 - \|h_{\text{supp}(\boldsymbol{\alpha}^*)^c}\|_1 \\ &\leq \|h_{\max(s)}\|_1 - \|h_{-\max(s)}\|_1. \end{aligned}$$

For the lower bound of (23), $\|\boldsymbol{\epsilon} - \mathbf{Z}h\|_2^2 - \|\boldsymbol{\epsilon}\|_2^2$, we can simplify as

$$\begin{aligned} \frac{1}{2}(\|\boldsymbol{\epsilon} - \mathbf{Z}h\|_2^2 - \|\boldsymbol{\epsilon}\|_2^2) &= -\frac{1}{2}(\mathbf{Z}h)^T(2\boldsymbol{\epsilon} - \mathbf{Z}h) \geq -h^T\mathbf{Z}^T\boldsymbol{\epsilon} \geq -\|\mathbf{Z}^T\boldsymbol{\epsilon}\|_\infty\|h\|_1 \\ &= -\|\mathbf{Z}^T\boldsymbol{\epsilon}\|_\infty(\|h_{\max(s)}\|_1 + \|h_{-\max(s)}\|_1). \end{aligned}$$

Hence, by (23) and the condition $r\|\mathbf{Z}^T\boldsymbol{\epsilon}\|_\infty \leq \lambda$ where $r > 1$, we have

$$r(\|h_{\max(s)}\|_1 - \|h_{-\max(s)}\|_1) \geq -(\|h_{\max(s)}\|_1 + \|h_{-\max(s)}\|_1).$$

which yields (21), the first part of the theorem.

For (22), the second part of the theorem, suppose $(r+1)\delta_{2s}^+(\mathbf{Z}) < (3r-1)\delta_{2s}^-(\mathbf{Z})$ holds. By the Karush-Kuhn-Tucker (KKT) condition of the minimization problem in (20), we have $\|\mathbf{Z}^T(y - \mathbf{Z}\hat{\boldsymbol{\alpha}})\|_\infty \leq \lambda$ and

$$\|\mathbf{Z}^T\mathbf{Z}h\|_\infty \leq \|\mathbf{Z}^T(y - \mathbf{Z}\hat{\boldsymbol{\alpha}})\|_\infty + \|\mathbf{Z}^T(y - \mathbf{Z}\boldsymbol{\alpha}^*)\|_\infty \leq \lambda + \|\mathbf{Z}^T\boldsymbol{\epsilon}\|_\infty.$$

Lemma 5.1 in Cai and Zhang (2013) with $\lambda = \max(\|h_{-\max(s)}\|_\infty, \|h_{-\max(s)}\|_1/s)$ implies

$$\begin{aligned} |\langle \mathbf{Z}h_{\max(s)}, \mathbf{Z}h_{-\max(s)} \rangle| &\leq \theta_{s,s}(\mathbf{Z})\|h_{\max(s)}\|_2 \cdot \sqrt{s} \cdot \max(\|h_{-\max(s)}\|_\infty, \|h_{-\max(s)}\|_1/s) \\ &\leq \sqrt{s}\theta_{s,s}(\mathbf{Z})\|h_{\max(s)}\|_2 \cdot \frac{r+1}{r-1}\|h_{\max(s)}\|_1/s \\ &\leq \theta_{s,s}(\mathbf{Z})\frac{r+1}{r-1}\|h_{\max(s)}\|_2^2, \end{aligned}$$

where the last inequality uses (21). We then have

$$\begin{aligned} \sqrt{s}(\lambda + \|\mathbf{Z}^T\boldsymbol{\epsilon}\|_\infty)\|h_{\max(s)}\|_2 &\geq (\lambda + \|\mathbf{Z}^T\boldsymbol{\epsilon}\|_\infty)\|h_{\max(s)}\|_1 \geq \langle \mathbf{Z}^T\mathbf{Z}h, h_{\max(s)} \rangle \\ &= \langle \mathbf{Z}h_{\max(s)}, \mathbf{Z}h_{\max(s)} \rangle + \langle \mathbf{Z}h_{\max(s)}, \mathbf{Z}h_{-\max(s)} \rangle \\ &\geq \|\mathbf{Z}h_{\max(s)}\|_2^2 - \theta_{s,s}\frac{r+1}{r-1}\|h_{\max(s)}\|_2^2 \\ &= \left(\delta_{2s}^-(\mathbf{Z}) - \theta_{s,s}(\mathbf{Z})\frac{r+1}{r-1} \right) \|h_{\max(s)}\|_2^2 \\ &\geq \left(\frac{3r-1}{2(r-1)}\delta_{2s}^-(\mathbf{Z}) - \frac{r+1}{2(r-1)}\delta_{2s}^+(\mathbf{Z}) \right) \|h_{\max(s)}\|_2^2, \end{aligned}$$

where the last inequality uses Lemma 1. Moving $\|h_{\max(s)}\|$ to the right hand side and using the condition $r\|\mathbf{Z}^T\boldsymbol{\epsilon}\|_\infty \leq \lambda$ where $r > 1$ yields (22). \square

Now we move on to the proof of Theorem 2. Section 3.5 in the main paper states that the original estimation method can be reinterpreted as a two-step method where the first step is the Lasso step and the second step is a dot product. The proof will first analyze step 1 using the lemmas about Lasso performance and use it to analyze step 2.

First, in lieu of step 1, the model in equation (3) from the original paper can be modified to

$$\mathbf{P}_{\hat{\mathbf{D}}^\perp} \mathbf{P}_{\mathbf{Z}} Y = \mathbf{P}_{\hat{\mathbf{D}}^\perp} \mathbf{Z} \boldsymbol{\alpha}^* + \mathbf{P}_{\hat{\mathbf{D}}^\perp} \mathbf{P}_{\mathbf{Z}} \boldsymbol{\epsilon}. \quad (24)$$

Here, $\mathbf{P}_{\hat{\mathbf{D}}^\perp} \mathbf{Z}$ becomes the design matrix, $\mathbf{P}_{\hat{\mathbf{D}}^\perp} \mathbf{P}_{\mathbf{Z}} Y$ becomes the outcome, and $\mathbf{P}_{\hat{\mathbf{D}}^\perp} \mathbf{P}_{\mathbf{Z}} \boldsymbol{\epsilon}$ is the new error term. In addition, from the condition $3\|\mathbf{Z}^T \mathbf{P}_{\hat{\mathbf{D}}^\perp} \boldsymbol{\epsilon}\| \leq \lambda$, we have

$$\lambda \geq 3\|\mathbf{Z}^T (I - \mathbf{P}_{\hat{\mathbf{D}}}) \boldsymbol{\epsilon}\|_\infty = 3\|\mathbf{Z}^T (\mathbf{P}_{\mathbf{Z}} - \mathbf{P}_{\hat{\mathbf{D}}}) \boldsymbol{\epsilon}\|_\infty = 3\|\mathbf{Z}^T (I - \mathbf{P}_{\hat{\mathbf{D}}}) \mathbf{P}_{\mathbf{Z}} \boldsymbol{\epsilon}\|_\infty = 3\|(\mathbf{P}_{\hat{\mathbf{D}}^\perp} \mathbf{Z})^T \mathbf{P}_{\mathbf{Z}} \boldsymbol{\epsilon}\|_\infty.$$

Second, note that (27) is in terms of the RIP constants of $\mathbf{P}_{\hat{\mathbf{D}}^\perp} \mathbf{Z}$. To relate the RIP constants of $\mathbf{P}_{\hat{\mathbf{D}}^\perp} \mathbf{Z}$ with that of \mathbf{Z} , we see that for any $2s$ -sparse vector $x \in \mathbb{R}^L$, $\|\mathbf{P}_{\hat{\mathbf{D}}^\perp} \mathbf{Z} x\|_2^2 = \|\mathbf{Z} x\|_2^2 - \|\mathbf{P}_{\hat{\mathbf{D}}} \mathbf{Z} x\|_2^2 \leq \|\mathbf{Z} x\|_2^2 \leq \delta_{2s}^+(\mathbf{Z}) \|x\|_2^2$. By the definition of $\delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}^\perp} \mathbf{Z})$, this implies

$$\delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}^\perp} \mathbf{Z}) \leq \delta_{2s}^+(\mathbf{Z}). \quad (25)$$

In addition, we have $\|\mathbf{P}_{\hat{\mathbf{D}}^\perp} \mathbf{Z} x\|_2^2 = \|\mathbf{Z} x\|_2^2 - \|\mathbf{P}_{\hat{\mathbf{D}}} \mathbf{Z} x\|_2^2 \geq \delta_{2s}^-(\mathbf{Z}) \|x\|_2^2 - \delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}} \mathbf{Z}) \|x\|_2^2$. By the definition of $\delta_{2s}^-(\mathbf{P}_{\hat{\mathbf{D}}^\perp} \mathbf{Z})$, this also implies

$$\delta_{2s}^-(\mathbf{P}_{\hat{\mathbf{D}}^\perp} \mathbf{Z}) \geq \delta_{2s}^-(\mathbf{Z}) - \delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}} \mathbf{Z}). \quad (26)$$

Combining (25), (26) with assumption that $2\delta_{2s}^-(\mathbf{Z}) > \delta_{2s}^+(\mathbf{Z}) + 2\delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}} \mathbf{Z})$, we know $2\delta_{2s}^-(\mathbf{P}_{\hat{\mathbf{D}}^\perp} \mathbf{Z}) > \delta_{2s}^-(\mathbf{P}_{\hat{\mathbf{D}}^\perp} \mathbf{Z})$. By Lemma 2, where we set $r = 3$ in assumption $r\|\mathbf{Z}^T \boldsymbol{\epsilon}\|_\infty \leq \lambda$

and the model is rewritten as (24),

$$\|h_{\max(s)}\|_2 \leq \frac{4/3\lambda\sqrt{s}}{2\delta_{2s}^-(\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{Z}) - \delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{Z})} \quad (27)$$

and

$$\|h_{-\max(s)}\|_1 \leq 2\|h_{\max(s)}\|_1. \quad (28)$$

Combining the RIP relations established by (25) and (26), we can rewrite (27) as

$$\|h_{\max(s)}\|_2 \leq \frac{4/3\lambda\sqrt{s}}{2\delta_{2s}^-(\mathbf{Z}) - \delta_{2s}^+(\mathbf{Z}) - 2\delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z})}. \quad (29)$$

Third, we establish a bound for $\|\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z}h\|_2$. This bound is needed to bound step 2 in Section 3.5 of the original paper because

$$\hat{\beta}_\lambda = \frac{\hat{\mathbf{D}}^T \mathbf{P}_{\hat{\mathbf{D}}} (Y - \mathbf{Z}\hat{\alpha}_\lambda)}{\|\hat{\mathbf{D}}\|_2^2} = \frac{\hat{\mathbf{D}}^T \mathbf{P}_{\hat{\mathbf{D}}} (\mathbf{Z}\alpha^* + \mathbf{D}\beta^* + \epsilon - \mathbf{Z}\hat{\alpha}_\lambda)}{\|\hat{\mathbf{D}}\|_2^2} = \beta^* - \frac{\hat{\mathbf{D}}^T \mathbf{P}_{\hat{\mathbf{D}}} \mathbf{Z}h}{\|\hat{\mathbf{D}}\|_2^2} + \frac{\hat{\mathbf{D}}^T \mathbf{P}_{\hat{\mathbf{D}}} \epsilon}{\|\hat{\mathbf{D}}\|_2^2}.$$

Rearranging terms and taking norms on both sides give

$$\|\hat{\beta}_\lambda - \beta^*\|_2 \leq \frac{\|\hat{\mathbf{D}}^T \mathbf{P}_{\hat{\mathbf{D}}} \mathbf{Z}h\|_2}{\|\hat{\mathbf{D}}\|_2^2} + \frac{\|\hat{\mathbf{D}}^T \mathbf{P}_{\hat{\mathbf{D}}} \epsilon\|_2}{\|\hat{\mathbf{D}}\|_2^2} \leq \frac{\|\mathbf{P}_{\hat{\mathbf{D}}} \mathbf{Z}h\|_2}{\|\hat{\mathbf{D}}\|_2} + \frac{|\hat{\mathbf{D}}^T \epsilon|}{\|\hat{\mathbf{D}}\|_2^2}. \quad (30)$$

Hence, a bound on $\|\mathbf{P}_{\hat{\mathbf{D}}} \mathbf{Z}h\|_2$ is necessary to bound $\|\hat{\beta}_\lambda - \beta^*\|_2$. To start off, we apply Lemma 1.1 in Cai and Zhang (2014) to represent $h_{-\max(s)}$ as a weighted mean of s -sparse vectors. This lemma allows us to convert the bound for $h_{\max(s)}$ in (29) to the bound for $\|\mathbf{P}_{\hat{\mathbf{D}}} \mathbf{Z}h\|_2$. Specifically, the lemma states we can find $\lambda_i \geq 0$ and s -sparse $v_i \in \mathbb{R}^L$ where $i = 1, \dots, N$ such that $\sum_{i=1}^N \lambda_i = 1$ and $h_{-\max(s)} = \sum_{i=1}^N \lambda_i v_i$. Hence, $h = \sum_{i=1}^N \lambda_i (h_{\max(s)} + v_i)$. Furthermore, we have

$$\text{supp}(v_i) \subseteq \text{supp}(h_{-\max(s)}), \quad \|v_i\|_\infty \leq \max \left(\|h_{-\max(s)}\|_\infty, \frac{\|h_{-\max(s)}\|_1}{s} \right), \quad \|v_i\|_1 = \|h_{-\max(s)}\|_1,$$

which yields

$$\|v_i\|_\infty \leq \max\left(\frac{\|h_{\max(s)}\|_1}{s}, \frac{2\|h_{\max(s)}\|_1}{s}\right) = \frac{2\|h_{\max(s)}\|_1}{s}, \quad \|v_i\|_1 \leq 2\|h_{\max(s)}\|_1$$

and $\|h_{\max(s)} + v_i\|_2^2 = \|h_{\max(s)}\|_2^2 + \|v_i\|_2^2 \leq \|h_{\max(s)}\|_2^2 + \|v_i\|_1\|v_i\|_\infty \leq 5\|h_{\max(s)}\|_2^2$. Combining all these together with (29), we have

$$\begin{aligned} \|\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z}h\|_2 &\leq \sum_{i=1}^N \lambda_i \|\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z}(h_{\max(s)} + v_i)\|_2 \leq \sum_{i=1}^N \lambda_i \sqrt{5\delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z})} \|h_{\max(s)}\|_2 \\ &\leq \sqrt{5\delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z})} \frac{4/3\lambda\sqrt{s}}{2\delta_{2s}^-(\mathbf{Z}) - \delta_{2s}^+(\mathbf{Z}) - 2\delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z})} \\ &= \frac{4\sqrt{5}/3\lambda\sqrt{s\delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z})}}{2\delta_{2s}^-(\mathbf{Z}) - \delta_{2s}^+(\mathbf{Z}) - 2\delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z})}. \end{aligned}$$

Finally, using the relation (30) gives us the desired bound for Theorem 2. \square

Of independent interest is that the proof of Theorem 2 can be generalized to a matrix of \mathbf{D} instead of a vector of \mathbf{D} . That is, the proof can consider models where there are more than one endogenous variables in the data-generating model. However, for clarity of presentation, we don't explore this route.

5.4 Proof of Corollary 2

Now, we establish Corollary 2 as a Corollary to Theorem 2. Specifically, the task is to convert the RIP constants $\delta_{2s}^+(\mathbf{Z})$, $\delta_{2s}^-(\mathbf{Z})$, $\delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z})$ and the constraint of $2\delta_{2s}^-(\mathbf{Z}) - \delta_{2s}^+(\mathbf{Z}) - 2\delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z}) > 0$ into μ and a similar constraint on s . To do this, note that for any s -sparse vector $\boldsymbol{\alpha}$

$$\begin{aligned} \|\mathbf{Z}\boldsymbol{\alpha}\|_2^2 &= \sum_{j \in \text{supp}(\boldsymbol{\alpha})} \|\mathbf{Z}_{\cdot j}\|_2^2 \alpha_j^2 + \sum_{i < j, i, j \in \text{supp}(\boldsymbol{\alpha})} 2\alpha_i \alpha_j \langle \mathbf{Z}_{\cdot i}, \mathbf{Z}_{\cdot j} \rangle \leq \sum_{j \in \text{supp}(\boldsymbol{\alpha})} \alpha_j^2 + \sum_{i < j, i, j \in \text{supp}(\boldsymbol{\alpha})} (\alpha_i^2 + \alpha_j^2) \mu \\ &= (1 + (s-1)\mu) \sum_{j \in \text{supp}(\boldsymbol{\alpha})} \alpha_j^2 = (1 + (s-1)\mu) \|\boldsymbol{\alpha}\|_2^2 \end{aligned}$$

and

$$\begin{aligned}\|\mathbf{Z}\boldsymbol{\alpha}\|_2^2 &= \sum_{j \in \text{supp}(\boldsymbol{\alpha})} \|\mathbf{Z}_{\cdot j}\|_2^2 \alpha_j^2 + \sum_{i < j, i, j \in \text{supp}(\boldsymbol{\alpha})} 2\alpha_i \alpha_j \langle \mathbf{Z}_{\cdot i}, \mathbf{Z}_{\cdot j} \rangle \geq \sum_{j \in \text{supp}(\boldsymbol{\alpha})} \alpha_j^2 - \sum_{i < j, i, j \in \text{supp}(\boldsymbol{\alpha})} (\alpha_i^2 + \alpha_j^2) \mu \\ &= (1 - (s-1)\mu) \|\boldsymbol{\alpha}\|_2^2.\end{aligned}$$

The upper and lower bounds on $\|\mathbf{Z}\boldsymbol{\alpha}\|_2^2$ imply

$$\delta_s^+(\mathbf{Z}) \leq (1 + (s-1)\mu), \quad \text{and} \quad \delta_s^-(\mathbf{Z}) \geq (1 - (s-1)\mu);$$

For $\mathbf{P}_{\hat{\mathbf{D}}^\perp} \mathbf{Z}$ and all $2s$ -sparse vector x , we have

$$\begin{aligned}\|\mathbf{P}_{\hat{\mathbf{D}}^\perp} \mathbf{Z} x\|_2^2 &\leq \left(\sum_{j \in \text{supp}(x)} \|\mathbf{P}_{\hat{\mathbf{D}}^\perp} \mathbf{Z}_{\cdot j} x_j\|_2 \right)^2 \leq 2s \sum_{j \in \text{supp}(x)} \|\mathbf{P}_{\hat{\mathbf{D}}^\perp} \mathbf{Z}_{\cdot j} x_j\|_2^2 \\ &= 2s \sum_{j \in \text{supp}(x)} \|\mathbf{P}_{\hat{\mathbf{D}}^\perp} \mathbf{Z}_{\cdot j}\|_2^2 x_j^2 = 2s \sum_{j \in \text{supp}(x)} \frac{\|\mathbf{P}_{\hat{\mathbf{D}}^\perp} \mathbf{Z}_{\cdot j}\|_2^2}{\|\mathbf{Z}_{\cdot j}\|_2^2} \|\mathbf{Z}_{\cdot j} x_j\|_2^2 \\ &\leq 2s \rho^2 \delta_1^+(\mathbf{Z}) \sum_{j \in \text{supp}(x)} x_j^2 \leq 2s \rho^2 \delta_{2s}^+(\mathbf{Z}) \|x\|_2^2.\end{aligned}$$

Again, by the definition of $\delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}^\perp} \mathbf{Z})$, this implies that

$$\delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}^\perp} \mathbf{Z}) \leq 2s \rho^2 \delta_{2s}^+(\mathbf{Z}). \quad (31)$$

Under the condition $s < \min\left(\frac{1}{12\mu}, \frac{1}{10\rho^2}\right)$, the denominator of the bound in Theorem 2

becomes

$$\begin{aligned}
2\delta_{2s}^-(\mathbf{Z}) - \delta_{2s}^+(\mathbf{Z}) - 2\delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z}) &\geq 2\delta_{2s}^-(\mathbf{Z}) - (1 + 4s\rho^2)\delta_{2s}^+(\mathbf{Z}) \\
&\geq 2(1 - (2s - 1)\mu) - (1 + 4s\rho^2)(1 + (2s - 1)\mu) \\
&= 1 - 6s\mu + 3\mu - 4s\rho^2 - 8s^2\rho^2\mu + 4s\rho^2\mu \\
&\geq 1 - 6s\mu - 5s\rho^2 > 0.
\end{aligned}$$

For the numerator of the bound in Theorem 2, we have

$$\begin{aligned}
\frac{4\sqrt{5}}{3}\lambda\sqrt{s\delta_{2s}^+(\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{Z})} &\leq \frac{4\sqrt{5}}{3}\lambda\sqrt{2s^2\rho^2\delta_{2s}^+(\mathbf{Z})} \leq \frac{4\sqrt{10}}{3}\lambda s\rho\sqrt{1 + (2s - 1)\mu} \\
&\leq \frac{4\sqrt{10}}{3}\lambda s\rho\sqrt{1 + 2s\mu} \leq \frac{4\sqrt{10}}{3}\lambda s\rho\sqrt{1 + 1/6} = \frac{4\sqrt{105}}{9}\lambda s\rho.
\end{aligned}$$

Combining them together leads to the desired bound. Note that one can improve the constants in the constraint of s with a bit more care on the above inequalities. \square

5.5 Proof of Theorem 3

The original estimation method can be rewritten as follows

$$\begin{aligned}
\hat{\alpha}_\lambda, \hat{\beta}_\lambda &= \underset{\alpha, \beta}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{P}_{\mathbf{Z}}(\mathbf{Y} - \mathbf{Z}\alpha - \mathbf{D}\beta)\|_2^2 + \lambda \|\alpha\|_1 \\
&= \underset{\alpha, \beta}{\operatorname{argmin}} \frac{1}{2} \|(\mathbf{P}_{\hat{\mathbf{D}}} + \mathbf{P}_{\hat{\mathbf{D}}^\perp})\mathbf{P}_{\mathbf{Z}}(\mathbf{Y} - \mathbf{Z}\alpha - \mathbf{D}\beta)\|_2^2 + \lambda \|\alpha\|_1 \\
&= \underset{\alpha, \beta}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{P}_{\hat{\mathbf{D}}}\mathbf{P}_{\mathbf{Z}}(\mathbf{Y} - \mathbf{Z}\alpha - \mathbf{D}\beta)\|_2^2 + \frac{1}{2} \|\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{P}_{\mathbf{Z}}(\mathbf{Y} - \mathbf{Z}\alpha - \mathbf{D}\beta)\|_2^2 + \lambda \|\alpha\|_1 \\
&= \underset{\alpha, \beta}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{P}_{\hat{\mathbf{D}}}(\mathbf{Y} - \mathbf{Z}\alpha) - \hat{\mathbf{D}}\beta\|_2^2 + \frac{1}{2} \|\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{P}_{\mathbf{Z}}\mathbf{Y} - \mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{Z}\alpha\|_2^2 + \lambda \|\alpha\|_1.
\end{aligned}$$

The first term, $\frac{1}{2} \|\mathbf{P}_{\hat{\mathbf{D}}}(\mathbf{Y} - \mathbf{Z}\alpha) - \hat{\mathbf{D}}\beta\|_2^2$ is always zero for any given $\alpha \in \mathbb{R}^L$ because $\mathbf{P}_{\hat{\mathbf{D}}}(\mathbf{Y} - \mathbf{Z}\alpha)$ lies in the span of $\hat{\mathbf{D}}$ and thus, we can pick β such that the first term is zero.

The second term, $\frac{1}{2}||\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{P}_{\mathbf{Z}}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\alpha})||_2^2 + \lambda||\boldsymbol{\alpha}||_1$, is the traditional Lasso problem where the outcome is $\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{P}_{\mathbf{Z}}\mathbf{Y}$ and the design matrix is $\mathbf{P}_{\hat{\mathbf{D}}^\perp}\mathbf{Z}$. Hence, the minimizer for this Lasso problem is also the minimizer for the original method. \square

References

- Baraniuk, R., Davenport, M., DeVore, R., and Wakin, M. (2008), “A Simple Proof of the Restricted Isometry Property for Random Matrices,” *Constructive Approximation*, 28, 253–263.
- Bühlmann, P. and van de Geer, S. (2011), *Statistics for High-Dimensional Data*, New York: Springer-Verlag.
- Cai, T. T., Fan, J., and Jiang, T. (2013), “Distributions of Angels in Random Packing of Spheres,” *The Journal of Machine Learning Research*, 14, 1837–1864.
- Cai, T. T., and Zhang, A. (2013), “Compressed Sensing and Affine Rank Minimization Under Restricted Isometry,” *IEEE Transactions on Signal Processing*, 61, 3279–3290.
- (2014), “Sparse Representation of a Polytope and Recovery of Sparse Signals and Low-rank Matrices,” *IEEE Transactions on Information Theory*, 60, 122–132.
- Hastie, T., Tibshirani, R., and Friedman, H. (2009), *The Elements of Statistical Learning*, New York, NY: Springer, 2nd ed.
- Koopmans, T. C., Rubin, H. and Leipnik, R. B. (1950), “Measuring the Equation Systems of Dynamic Economics,” in *Statistical Inference in Dynamic Economic Models*, New York, NY: John Wiley and Sons, pp. 54-237.
- National Institute of Health (1998), “Clinical Guidelines on the Identification, Evaluation, and Treatment of Overweight and Obesity in Adults: The Evidence Report,” *Obesity Research*, 2, 51S–209S.
- Rothenberg, T. J., “Identification in Parametric Models,” *Econometrica*, 39, 577-591.

- Sach, T. H., Barton, G. R., Doherty, M., Muir, K. R., Jenkinson, C., and Avery, A. J. (2007), “The Relationship Between Body Mass Index and Health-related Quality of Life: Comparing the EQ-5D, EuroQol VAS and SF-6D,” *International Journal of Obesity*, 31, 189–196.
- Sargan, J. D. (1958), “The Estimation of Economic Relationships Using Instrumental Variables”, *Econometrica*, 26, 393–415.
- Trakas, K., Oh, P. I., Singh, S., Risebrough, N., and Shear, N. H. (2001), “The Health Status of Obese Individuals in Canada,” *International Journal of Obesity and Related Metabolic Disorders: Journal of the International Association for the Study of Obesity*, 25, 662–668.
- Wooldridge, J. M. (2010), *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MA: MIT Press, 2nd ed.

6 Figures

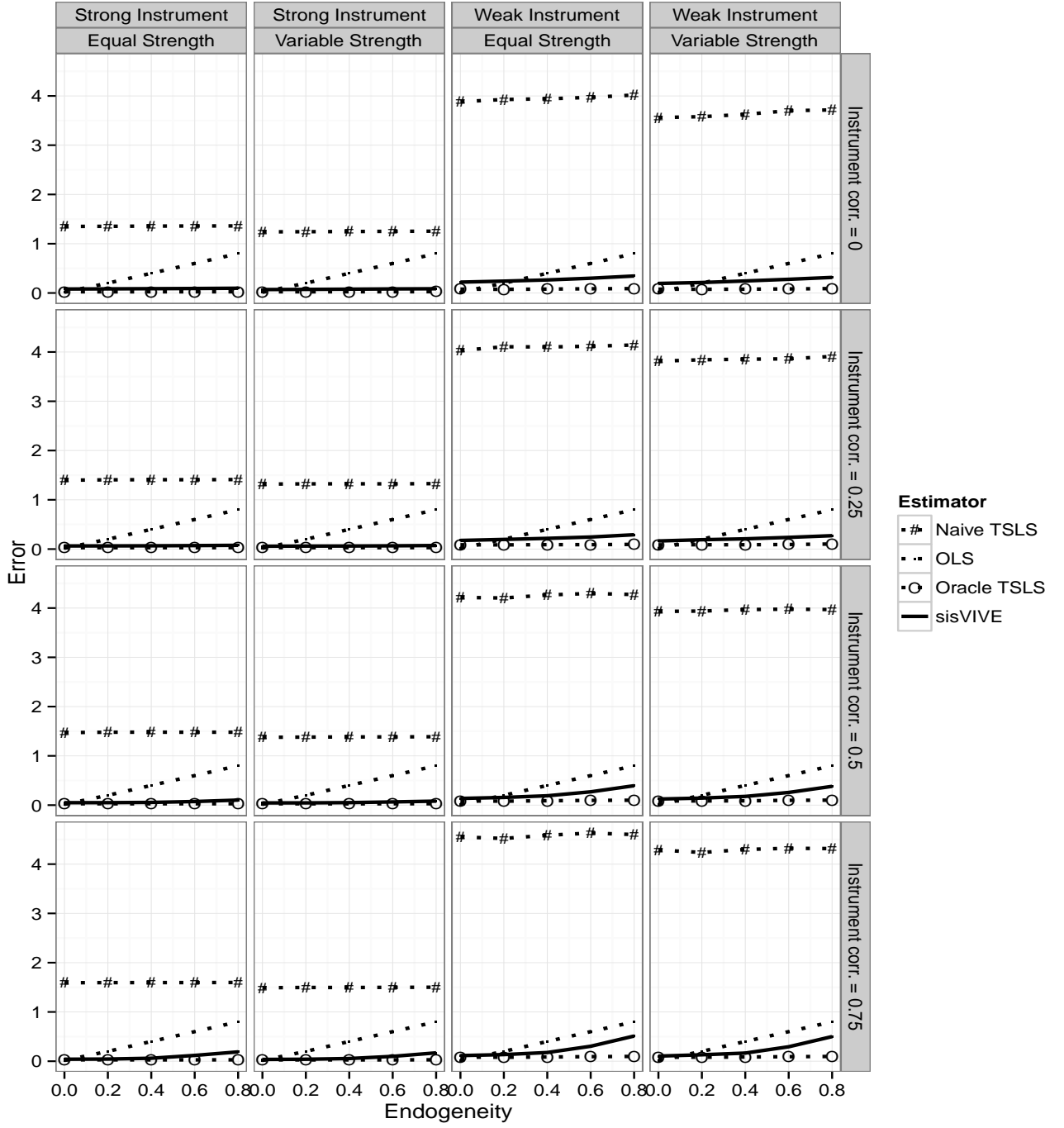


Figure 1: Simulation Study of Estimation Performance Varying Endogeneity and Correlation Only Exists Within Valid and Invalid Instruments. There are ten ($L = 10$) instruments. Each line represents the median absolute estimation error ($|\beta^* - \hat{\beta}|$) after 500 simulations. We fix the number of invalid instruments to $s = 3$. Each column in the plot corresponds to different variation of instruments' absolute and relative strength. There are two types of absolute strengths, “Strong” and “Weak”, measured by the concentration parameter. There are two types of relative strengths, “Equal” and “Variable”, measured by varying γ^* while holding the absolute strength (i.e. concentration parameter) fixed. Each row corresponds to the maximum correlation between instruments, but correlation only exists within valid and invalid instruments.

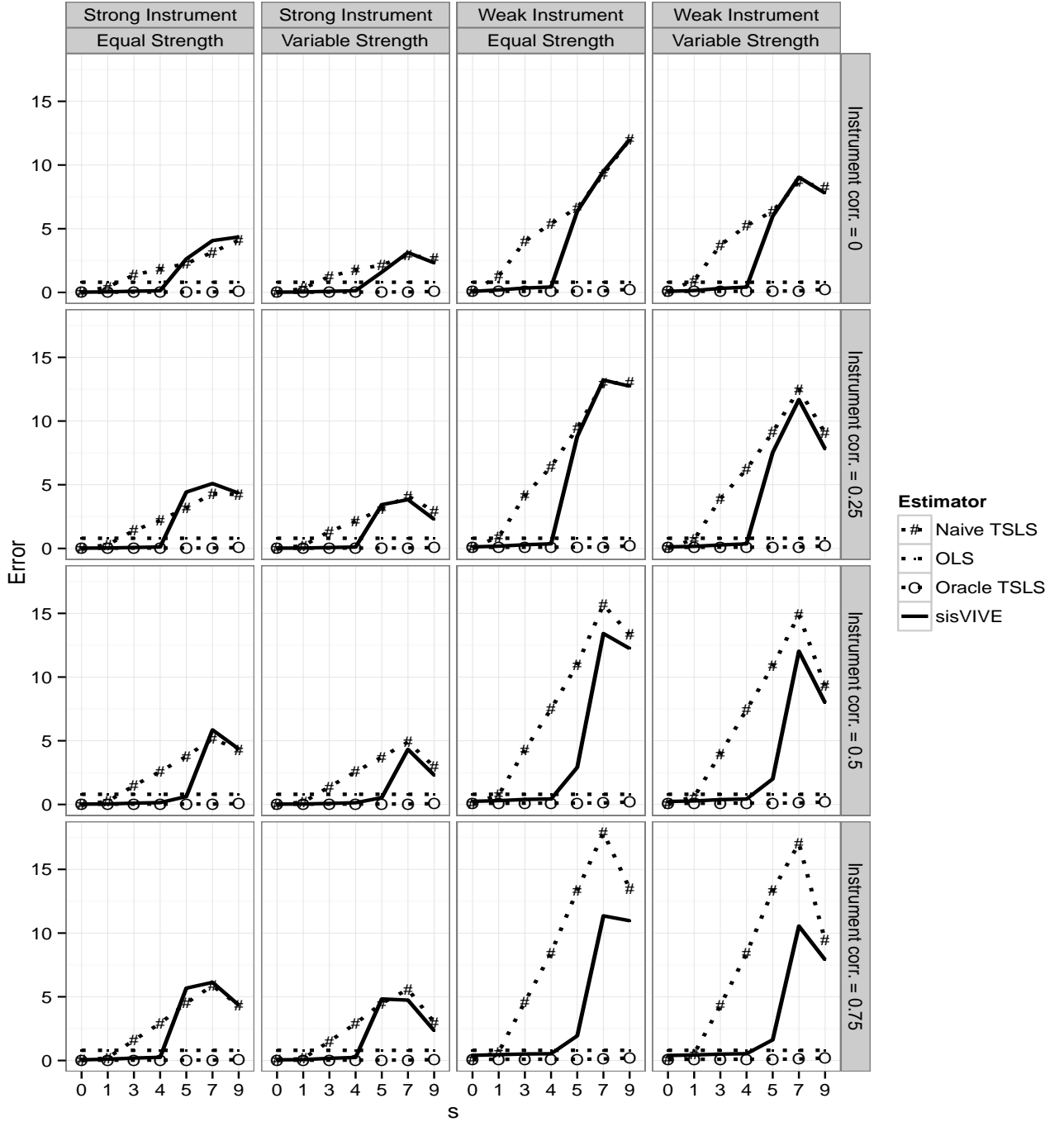


Figure 2: Simulation Study of Estimation Performance Varying the Number of Invalid Instruments (s) and Correlation Only Exists Within Valid and Invalid Instruments. There are ten ($L = 10$) instruments. Each line represents the median absolute estimation error ($|\beta^* - \hat{\beta}|$) after 500 simulations. We fix the endogeneity $\sigma_{\epsilon\xi}^*$ to $\sigma_{\epsilon\xi}^* = 0.8$. Each column in the plot corresponds to a different variation of instruments' absolute and relative strength. There are two types of absolute strengths, "Strong" and "Weak", measured by the concentration parameter. There are two types of relative strengths, "Equal" and "Variable", measured by varying γ^* while holding the absolute strength fixed. Each row corresponds to the maximum correlation between instruments, but correlation only exists within valid and invalid instruments.

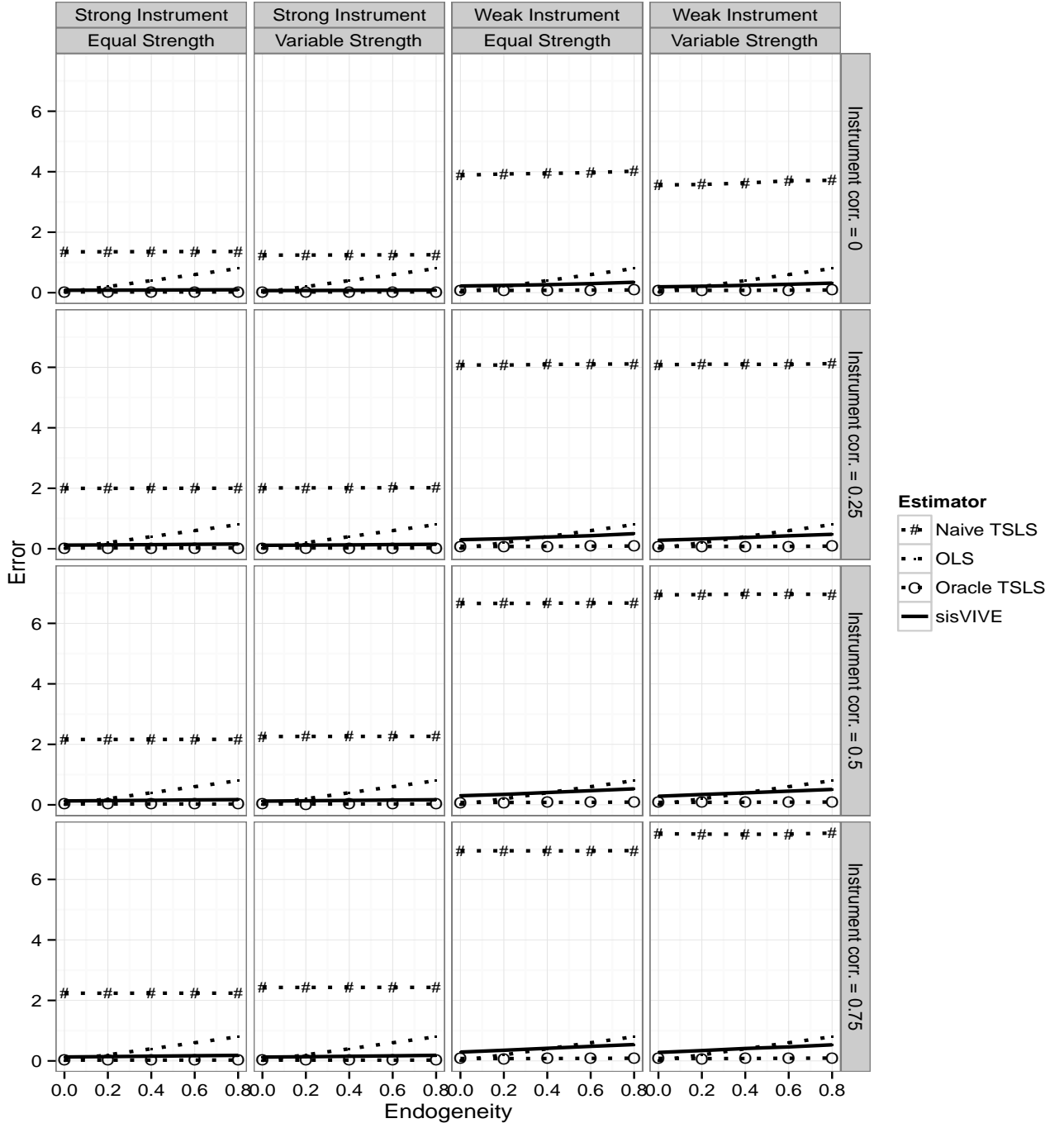


Figure 3: Simulation Study of Estimation Performance Varying Endogeneity and Correlation Only Exists Between Valid and Invalid Instruments. There are ten ($L = 10$) instruments. Each line represents the median absolute estimation error ($|\beta^* - \hat{\beta}|$) after 500 simulations. We fix the number of invalid instruments to $s = 3$. Each column in the plot corresponds to a different variation of instruments' absolute and relative strength. There are two types of absolute strengths, "Strong" and "Weak", measured by the concentration parameter. There are two types of relative strengths, "Equal" and "Variable", measured by varying γ^* while holding the absolute strength (i.e. concentration parameter) fixed. Each row corresponds to the maximum correlation between instruments, but correlation only exists between valid and invalid instruments.

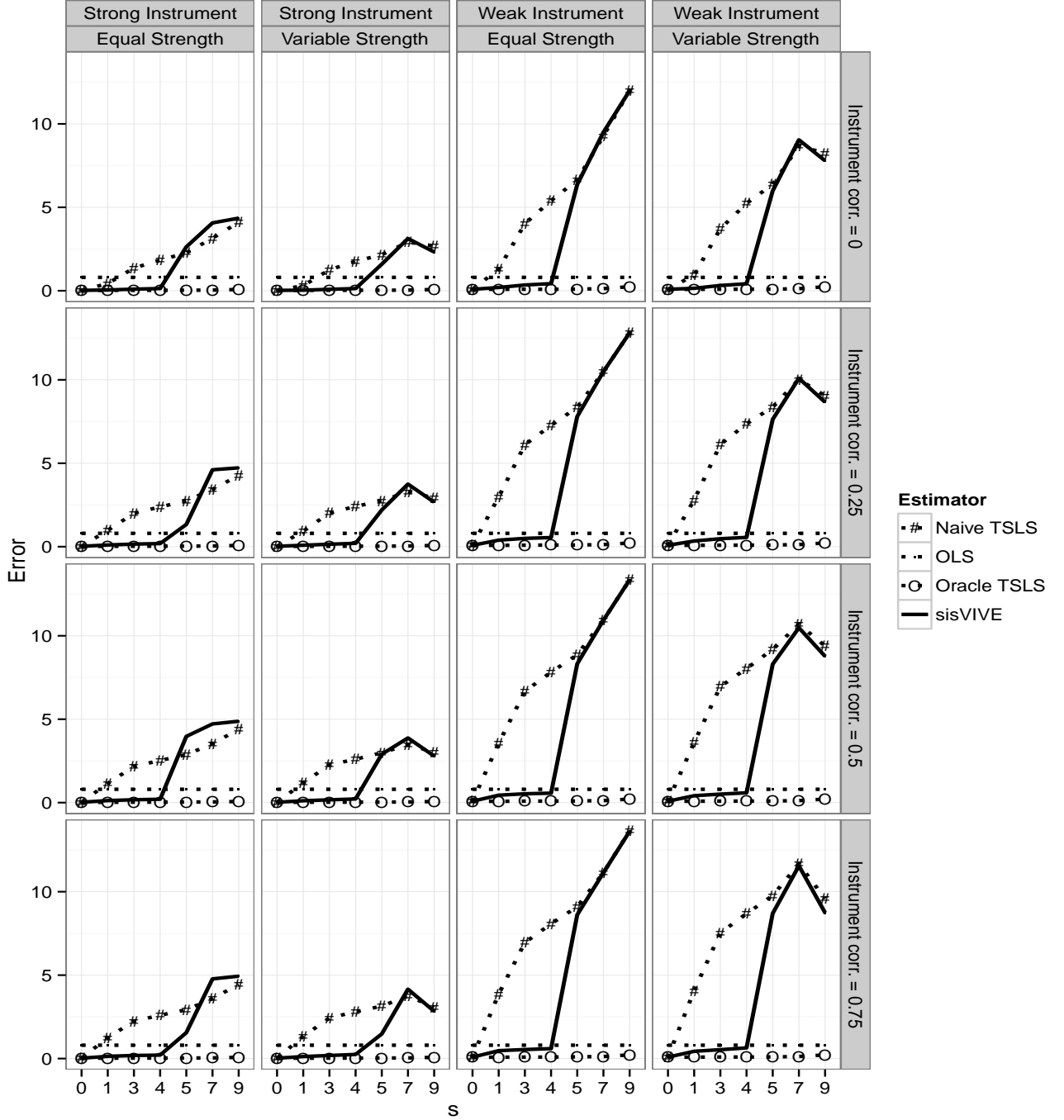


Figure 4: Simulation Study of Estimation Performance Varying the Number of Invalid Instruments (s) and Correlation Only Exists Between Valid and Invalid Instruments. There are ten ($L = 10$) instruments. Each line represents the median absolute estimation error ($|\beta^* - \hat{\beta}|$) after 500 simulations. We fix the endogeneity $\sigma_{\epsilon\xi}^*$ to $\sigma_{\epsilon\xi}^* = 0.8$. Each column in the plot corresponds to a different variation of instruments' absolute and relative strength. There are two types of absolute strengths, "Strong" and "Weak", measured by the concentration parameter. There are two types of relative strengths, "Equal" and "Variable", measured by varying γ^* while holding the absolute strength fixed. Each row corresponds to the maximum correlation between instruments, but correlation only exists between valid and invalid instruments.

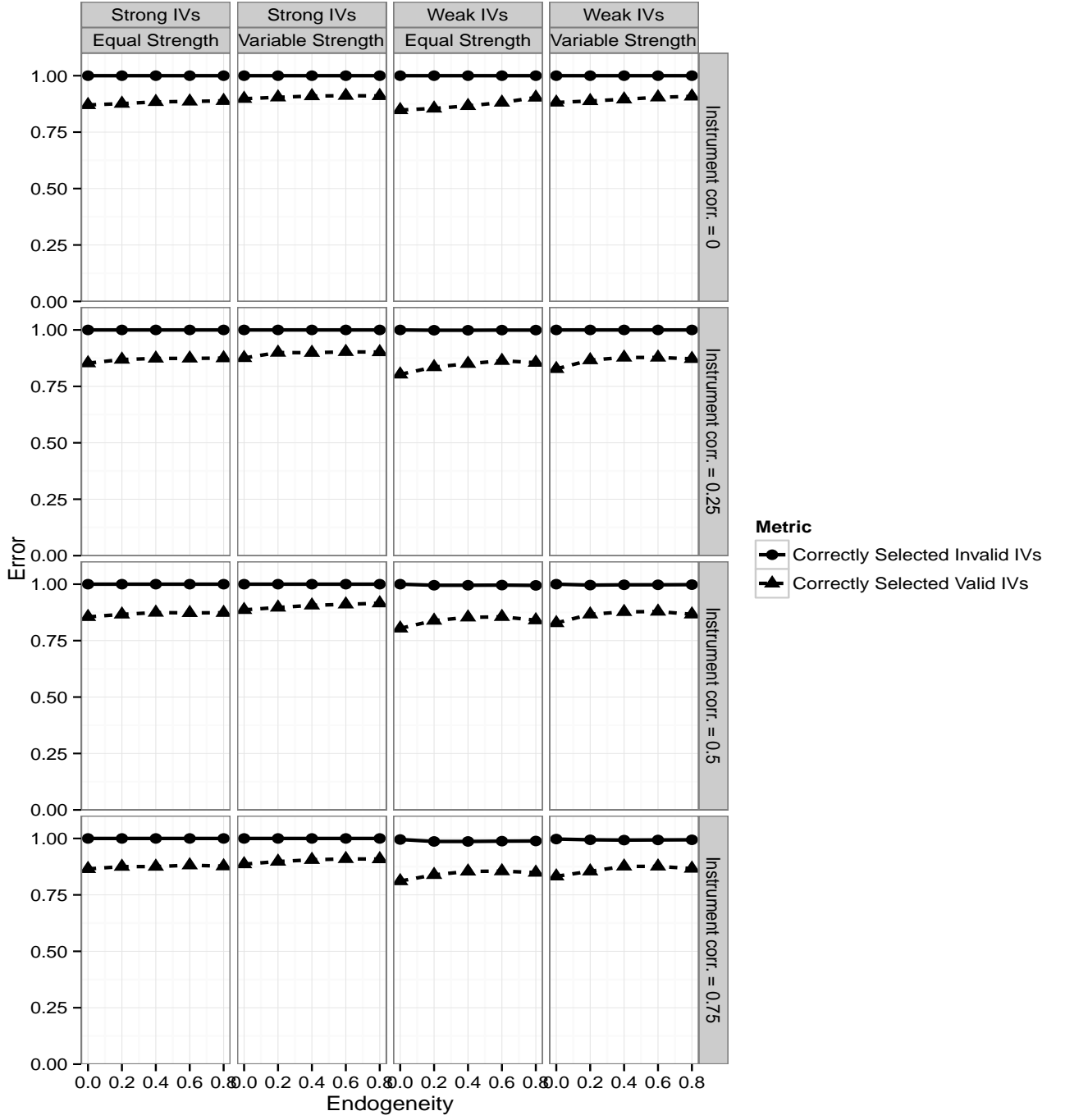


Figure 5: Simulation Study Varying Endogeneity and Correlation Exists Between All Instruments. There are ten ($L = 10$) instruments. Each line represents the average proportions of correctly selected valid instruments and correctly selected invalid instruments after 500 simulations. We fix the number of invalid instruments to $s = 3$. Each column in the plot corresponds to a different variation of instruments' absolute and relative strength. There are two types of absolute strengths, "Strong" and "Weak", measured by the concentration parameter. There are two types of relative strengths, "Equal" and "Variable", measured by varying γ^* while holding the absolute strength (i.e. concentration parameter) fixed. Each row corresponds to the maximum correlation between all instruments.

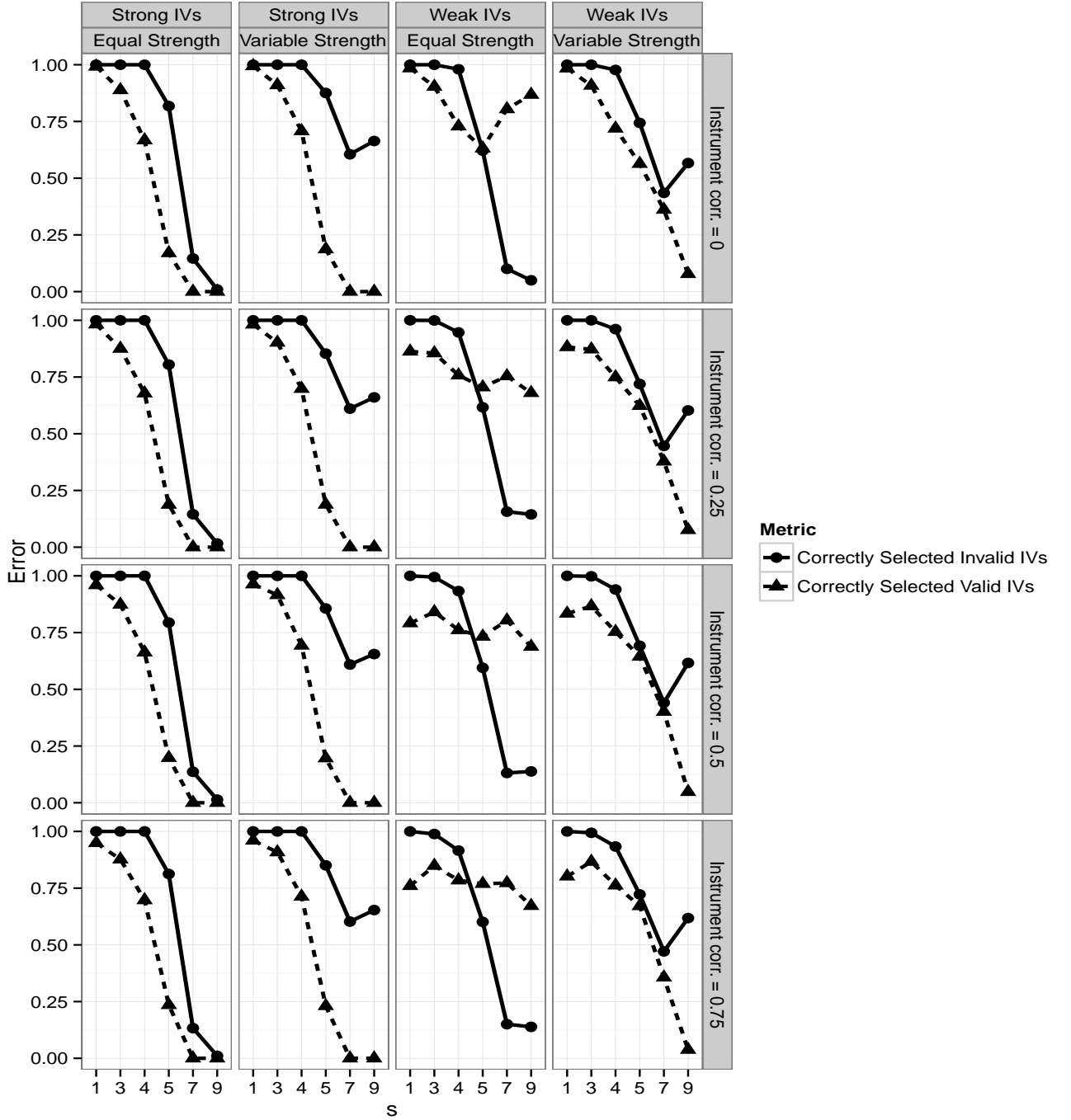


Figure 6: Simulation Study Varying the Number of Invalid Instruments (s) and Correlation Exists Between All Instruments. There are ten ($L = 10$) instruments. Each line represents the average proportions of correctly selected valid instruments and correctly selected invalid instruments after 500 simulations. We fix the endogeneity $\sigma_{\epsilon\xi}^*$ to $\sigma_{\epsilon\xi}^* = 0.8$. Each column in the plot corresponds to a different variation of instruments' absolute and relative strength. There are two types of absolute strengths, "Strong" and "Weak", measured by the concentration parameter. There are two types of relative strengths, "Equal" and "Variable", measured by varying γ^* while holding the absolute strength fixed. Each row corresponds to the maximum correlation between all instruments.

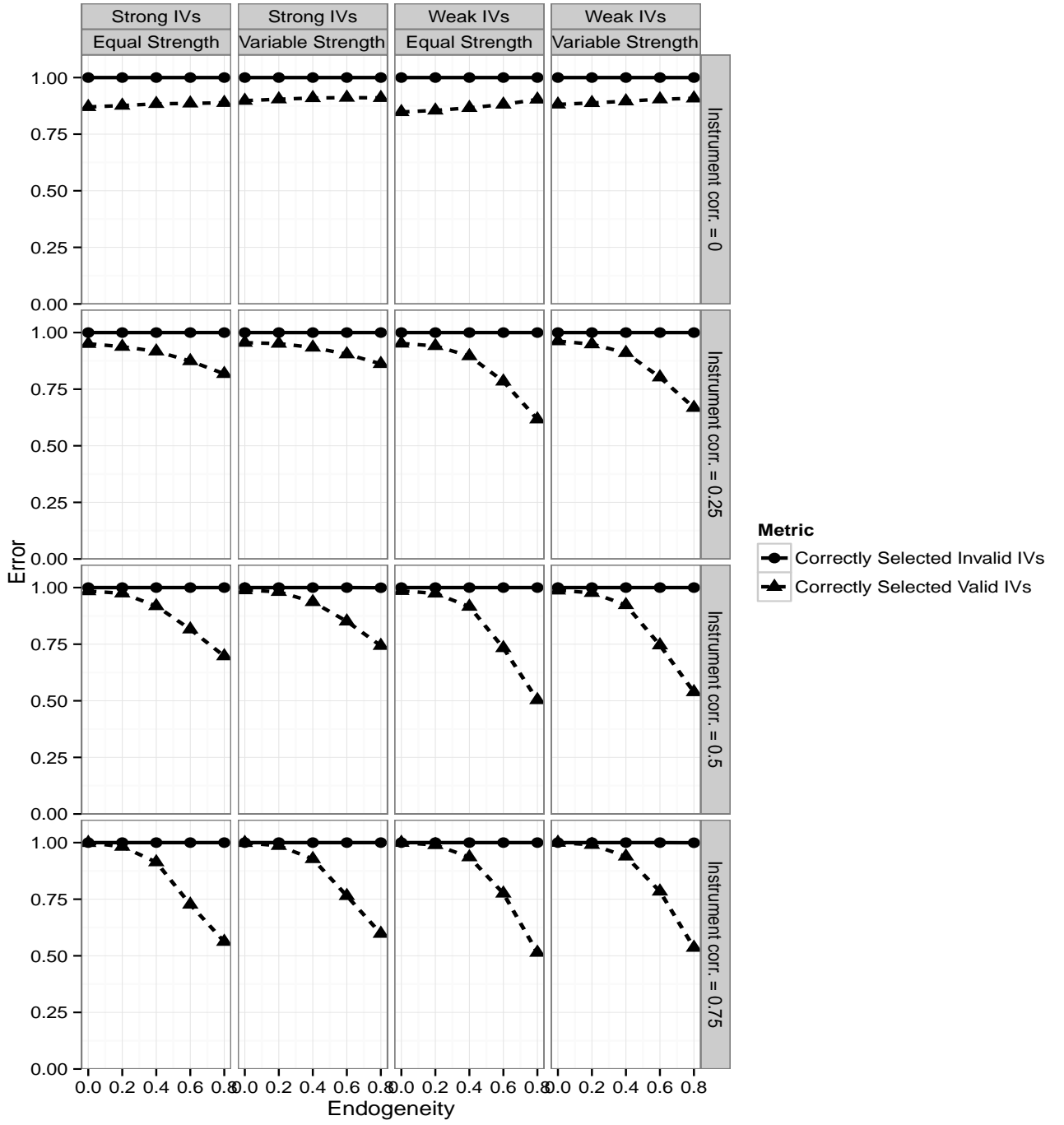


Figure 7: Simulation Study Varying Endogeneity and Correlation Only Exists Within Valid and Invalid Instruments. There are ten ($L = 10$) instruments. Each line represents the average proportions of correctly selected valid instruments and correctly selected invalid instruments after 500 simulations. We fix the number of invalid instruments to $s = 3$. Each column in the plot corresponds to a different variation of instruments' absolute and relative strength. There are two types of absolute strengths, "Strong" and "Weak", measured by the concentration parameter. There are two types of relative strengths, "Equal" and "Variable", measured by varying γ^* while holding the absolute strength (i.e. concentration parameter) fixed. Each row corresponds to the maximum correlation between instruments, but correlation only exists within valid and invalid instruments.

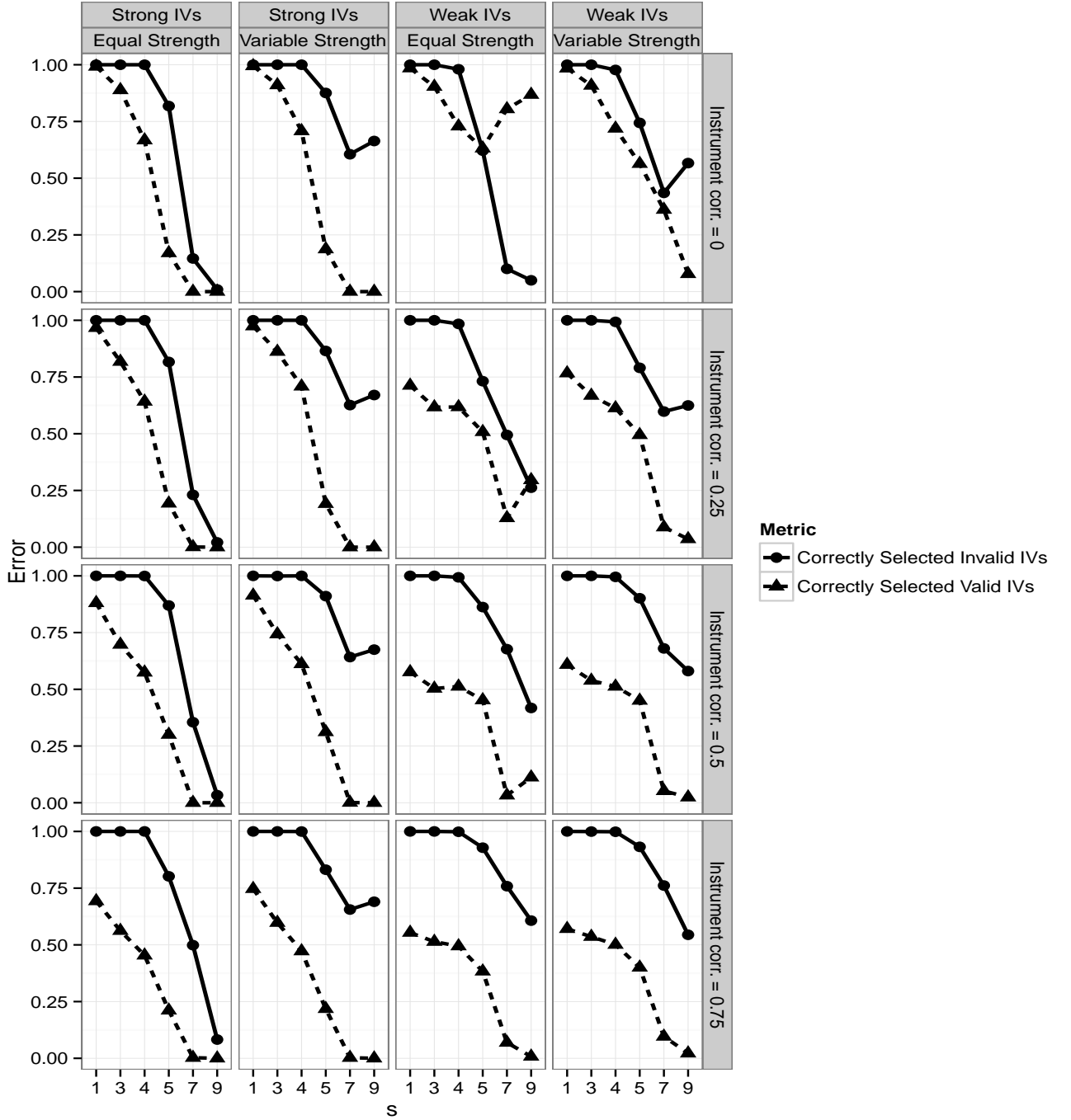


Figure 8: Simulation Study Varying the Number of Invalid Instruments (s) and Correlation Only Exists Within Valid and Invalid Instruments. There are ten ($L = 10$) instruments. Each line represents the average proportions of correctly selected valid instruments and correctly selected invalid instruments after 500 simulations. We fix the endogeneity $\sigma_{\epsilon\xi}^*$ to $\sigma_{\epsilon\xi}^* = 0.8$. Each column in the plot corresponds to a different variation of instruments' absolute and relative strength. There are two types of absolute strengths, "Strong" and "Weak", measured by the concentration parameter. There are two types of relative strengths, "Equal" and "Variable", measured by varying γ^* while holding the absolute strength fixed. Each row corresponds to the maximum correlation between instruments, but correlation only exists within valid and invalid instruments. 46

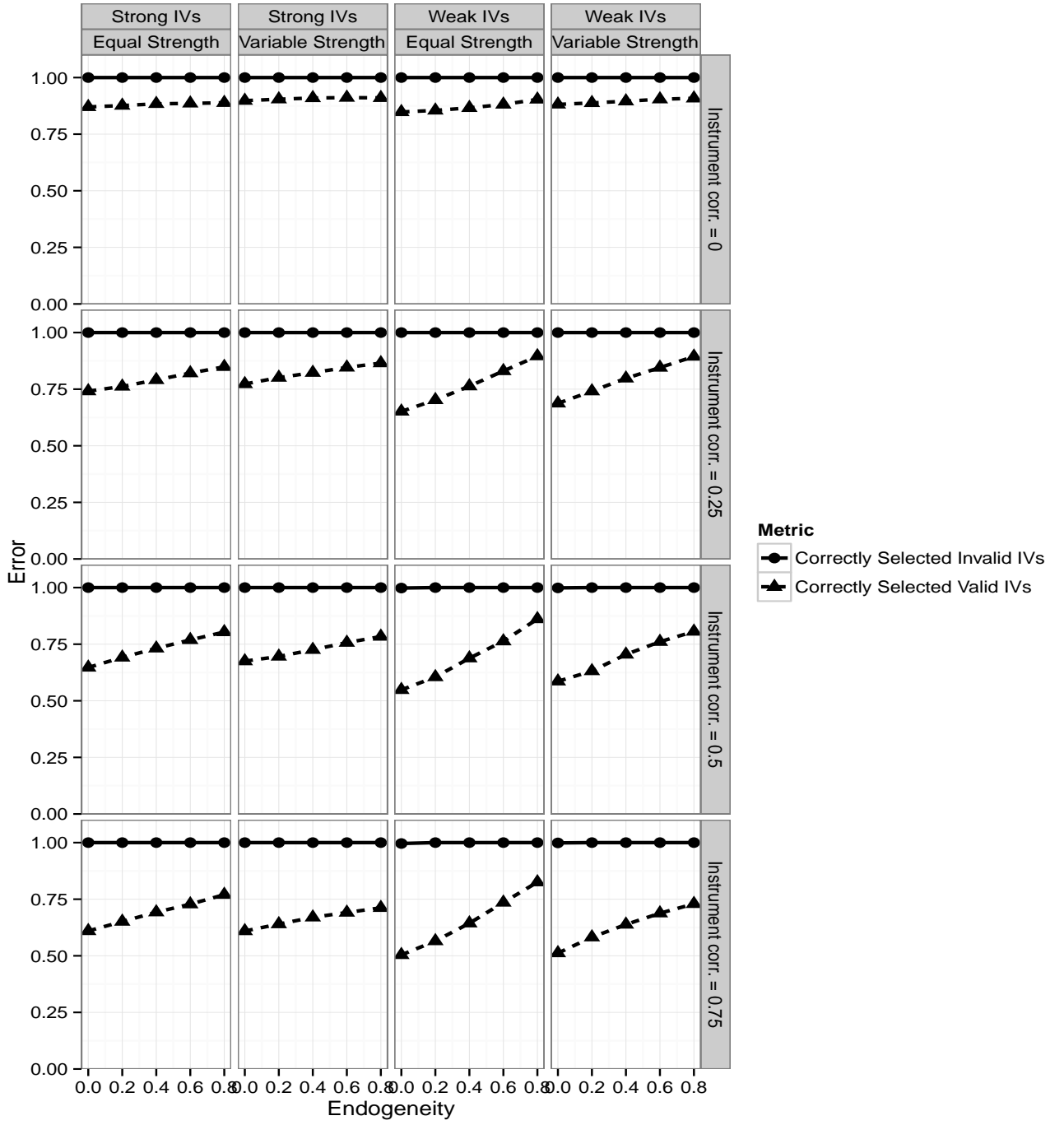


Figure 9: Simulation Study Varying Endogeneity and Correlation Only Exists Between Valid and Invalid Instruments. There are ten ($L = 10$) instruments. Each line represents the average proportions of correctly selected valid instruments and correctly selected invalid instruments after 500 simulations. We fix the number of invalid instruments to $s = 3$. Each column in the plot corresponds to a different variation of instruments' absolute and relative strength. There are two types of absolute strengths, "Strong" and "Weak", measured by the concentration parameter. There are two types of relative strengths, "Equal" and "Variable", measured by varying γ^* while holding the absolute strength (i.e. concentration parameter) fixed. Each row corresponds to the maximum correlation between instruments, but correlation only exists between valid and invalid instruments.

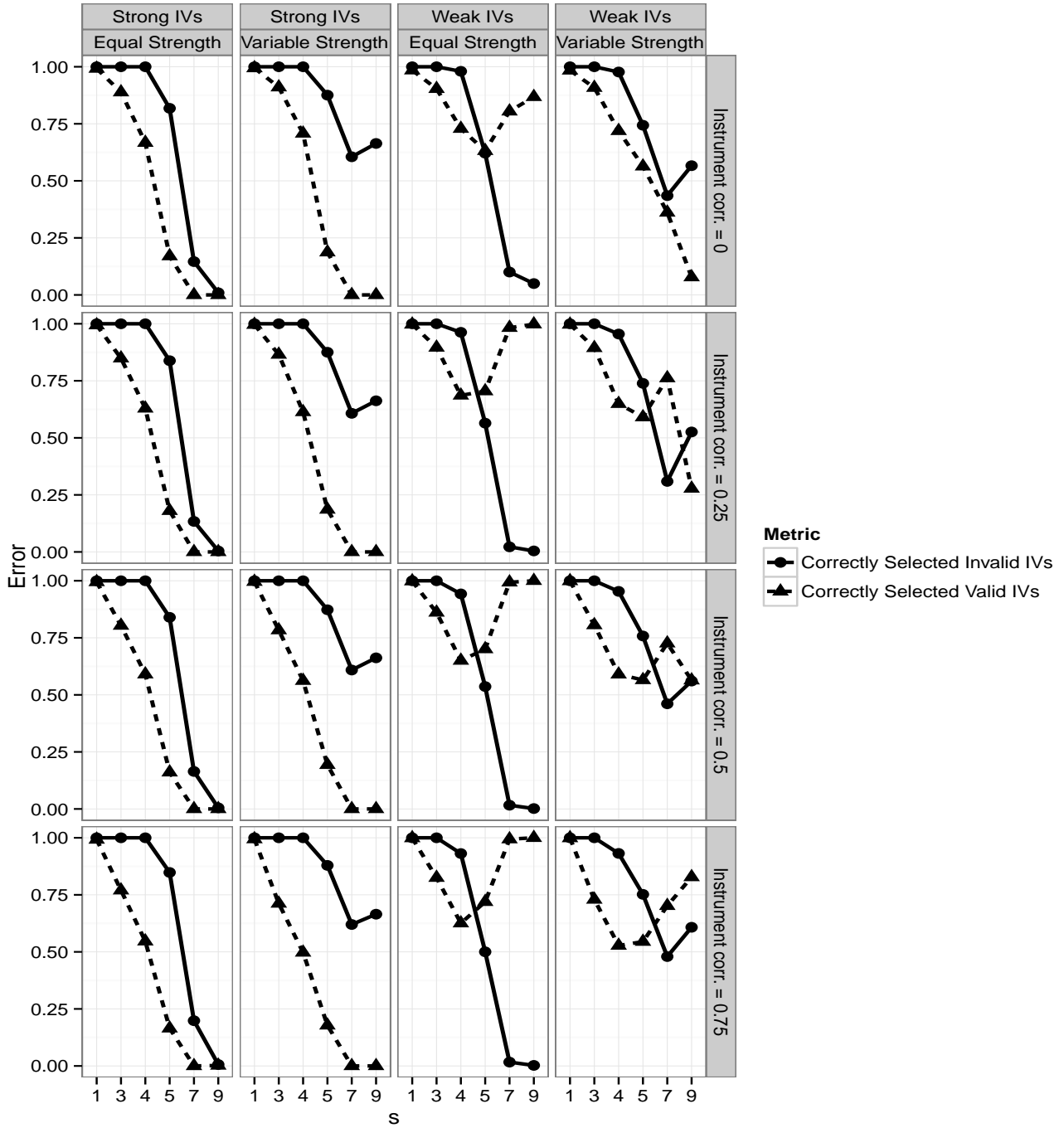


Figure 10: Simulation Study Varying the Number of Invalid Instruments (s) and Correlation Only Exists Between Valid and and Invalid Instruments. There are ten ($L = 10$) instruments. Each line represents the average proportions of correctly selected valid instruments and correctly selected invalid instruments after 500 simulations. We fix the endogeneity $\sigma_{\epsilon\xi}^*$ to $\sigma_{\epsilon\xi}^* = 0.8$. Each column in the plot corresponds to a different variation of instruments' absolute and relative strength. There are two types of absolute strengths, "Strong" and "Weak", measured by the concentration parameter. There are two types of relative strengths, "Equal" and "Variable", measured by varying γ^* while holding the absolute strength fixed. Each row corresponds to the maximum correlation between instruments, but correlation only exists between valid and invalid instruments.⁴⁸

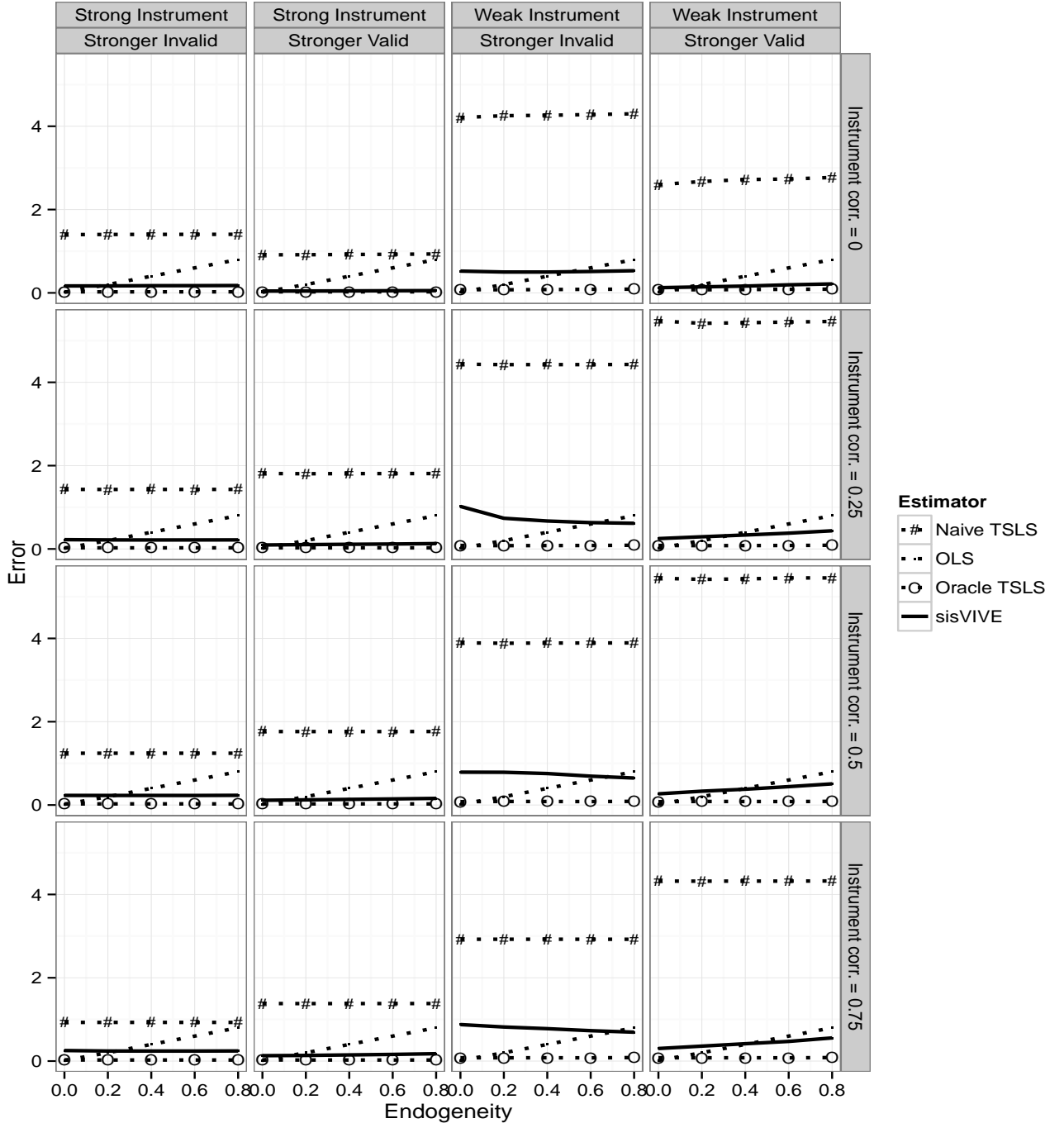


Figure 11: Simulation Study Varying Endogeneity and Correlation Exists Between All Instruments. We also vary the instrument strength of valid and invalid instruments. There are ten ($L = 10$) instruments. Each line represents the median absolute estimation error ($|\beta^* - \hat{\beta}|$) after 500 simulations. We fix the number of invalid instruments to $s = 3$. Each column in the plot corresponds to a different variation of instruments' absolute and relative strength. There are two types of absolute strengths, "Strong" and "Weak", measured by the concentration parameter. There are two types of strengths for valid and invalid instruments, "Stronger Invalid" and "Stronger Valid", determined by varying γ^* while holding the absolute strength fixed. Each row corresponds to the maximum correlation between instruments.

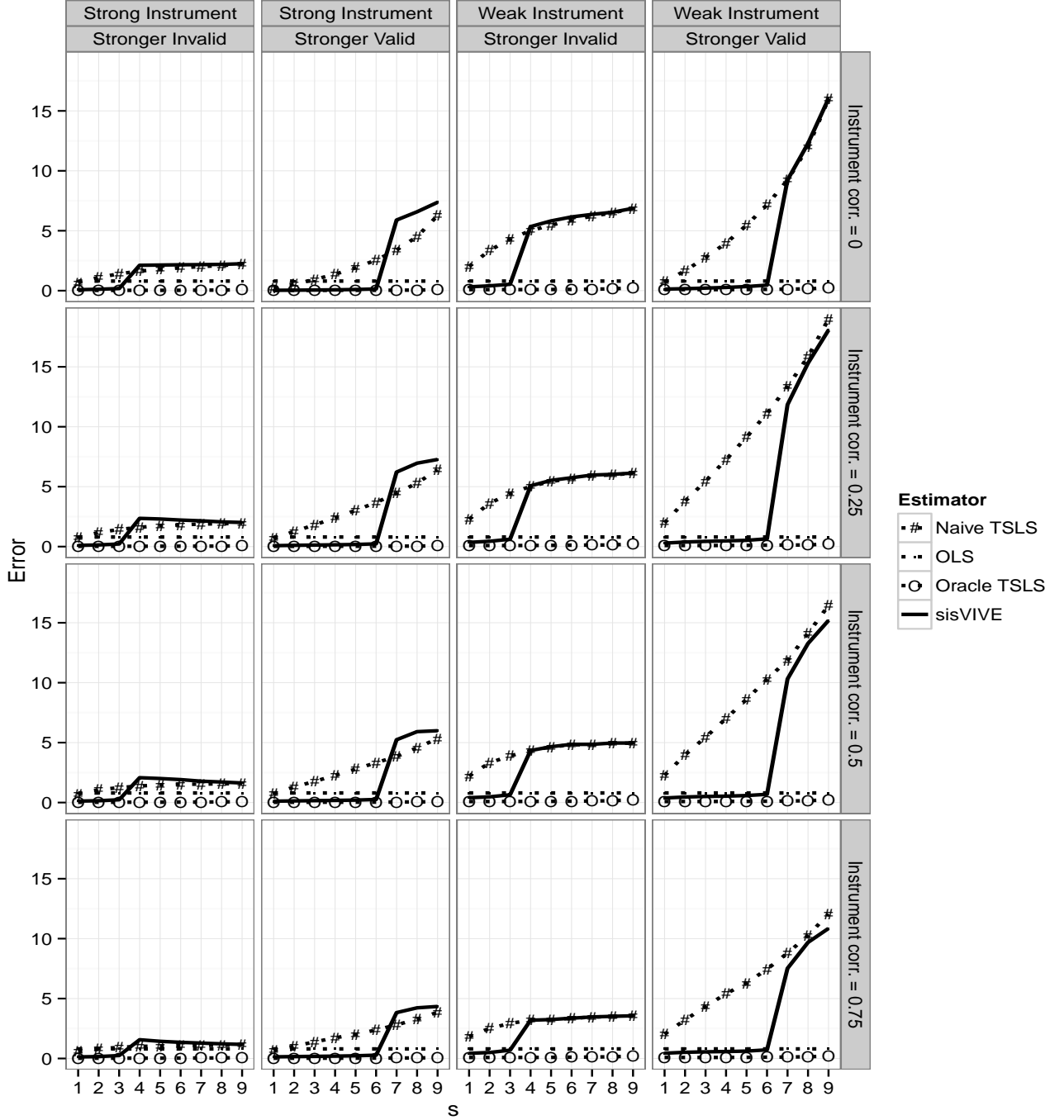


Figure 12: Simulation Study Varying the Number of Invalid Instruments (s) and Correlation Exists Between All Instruments. We also vary the instrument strength of valid and invalid instruments. There are ten ($L = 10$) instruments. Each line represents the median absolute estimation error ($|\beta^* - \hat{\beta}|$) after 500 simulations. We fix the endogeneity $\sigma_{\epsilon\xi}^*$ to $\sigma_{\epsilon\xi}^* = 0.8$. Each column in the plot corresponds to a different variation of instruments' absolute and relative strength. There are two types of absolute strengths, "Strong" and "Weak", measured by the concentration parameter. There are two types of strengths for valid and invalid instruments, "Stronger Invalid" and "Stronger Valid", determined by varying γ^* while holding the absolute strength fixed. Each row corresponds to the maximum correlation between instruments.

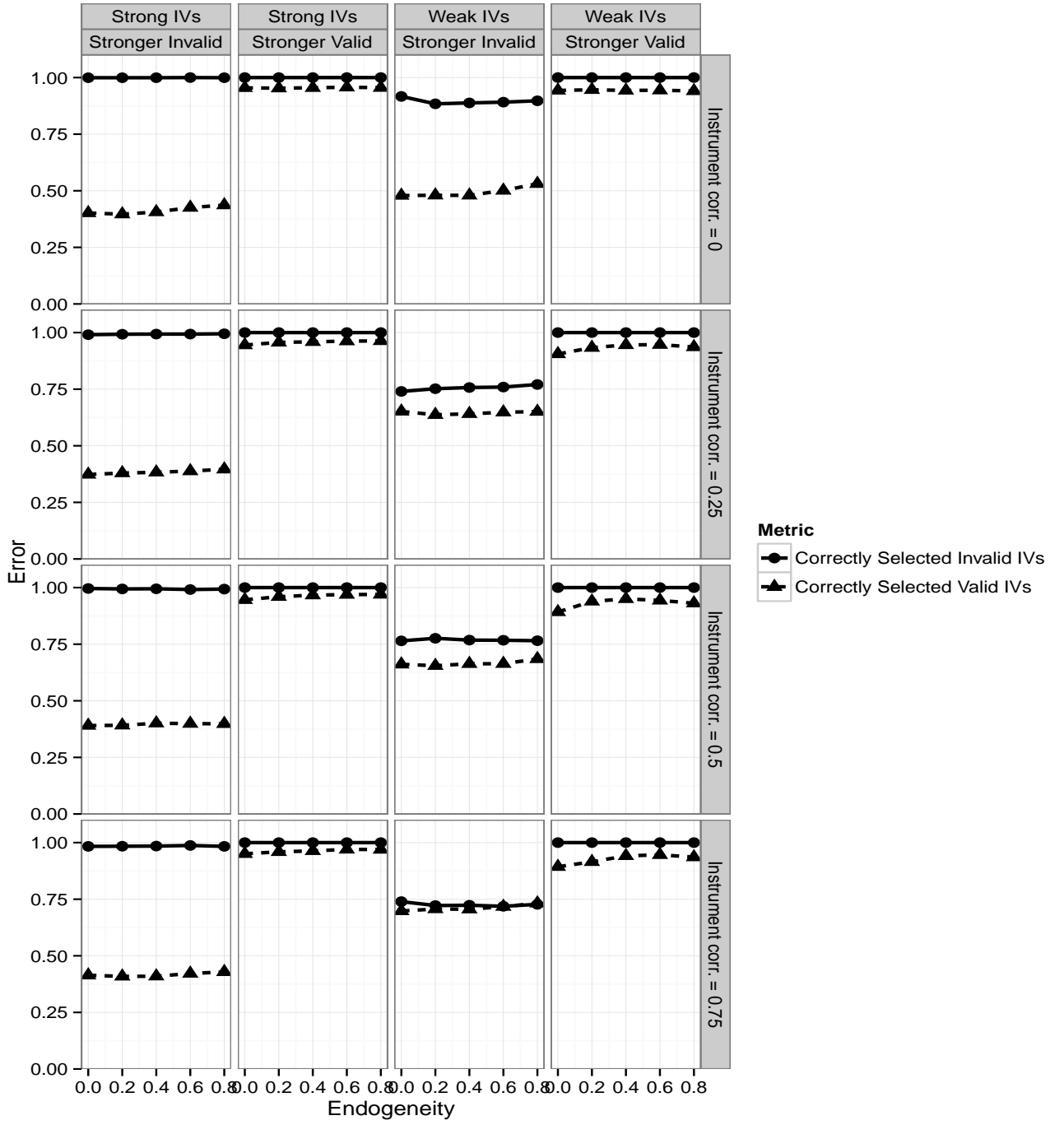


Figure 13: Simulation Study Varying Endogeneity and Correlation Exists Between All Instruments. We also vary the instrument strength of valid and invalid instruments. There are ten ($L = 10$) instruments. Each line represents the average proportions of correctly selected valid instruments and correctly selected invalid instruments after 500 simulations. We fix the number of invalid instruments to $s = 3$. Each column in the plot corresponds to a different variation of instruments' absolute and relative strength. There are two types of absolute strengths, "Strong" and "Weak", measured by the concentration parameter. There are two types of strengths for valid and invalid instruments, "Stronger Invalid" and "Stronger Valid", determined by varying γ^* while holding the absolute strength fixed. Each row corresponds to the maximum correlation between instruments.

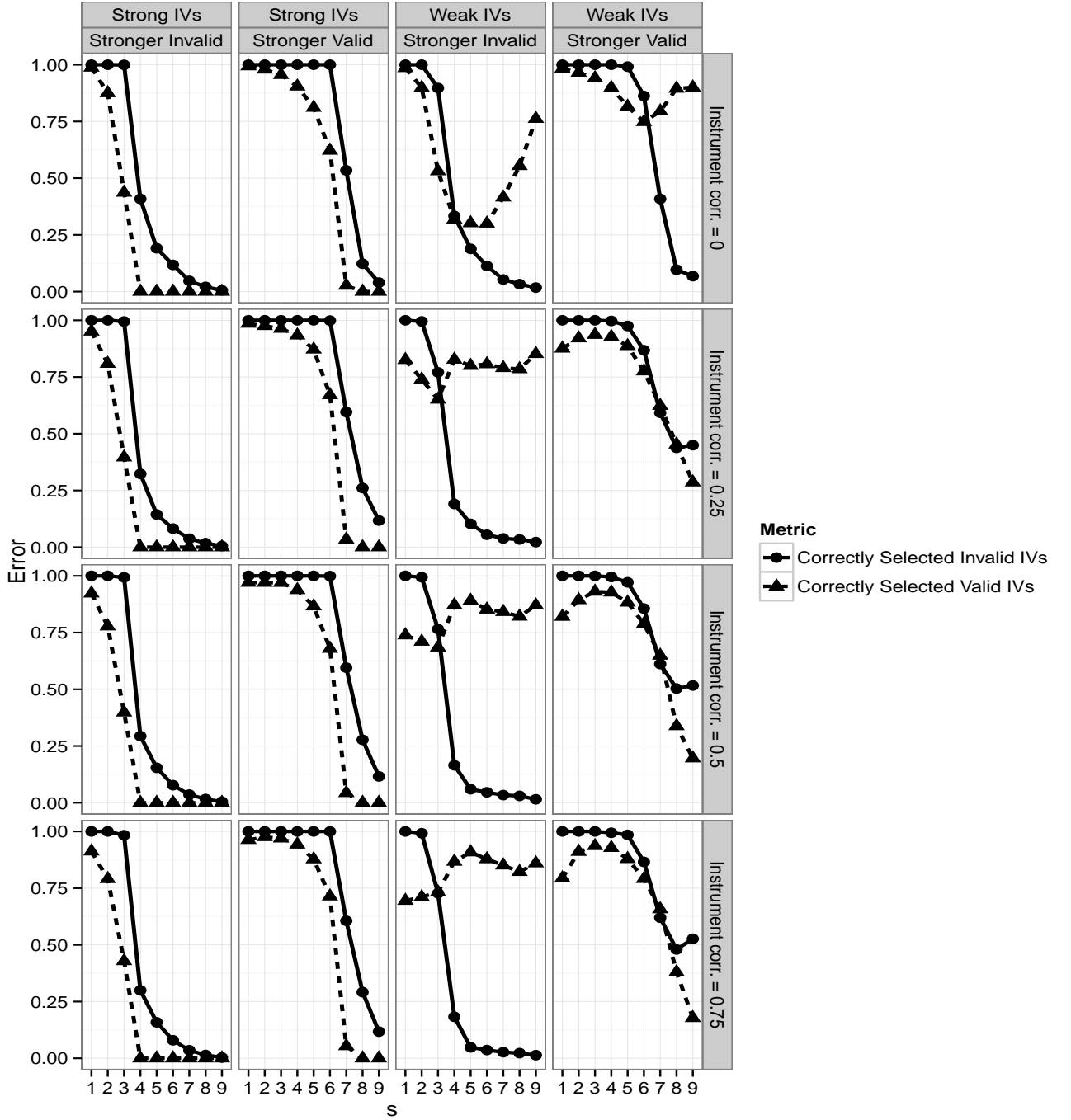


Figure 14: Simulation Study Varying the Number of Invalid Instruments (s) and Correlation Exists Between All Instruments. We also vary the instrument strength of valid and invalid instruments. There are ten ($L = 10$) instruments. Each line represents the average proportions of correctly selected valid instruments and correctly selected invalid instruments after 500 simulations. We fix the endogeneity $\sigma_{\epsilon\xi}^*$ to $\sigma_{\epsilon\xi}^* = 0.8$. Each column in the plot corresponds to a different variation of instruments' absolute and relative strength. There are two types of absolute strengths, "Strong" and "Weak", measured by the concentration parameter. There are two types of strengths for valid and invalid instruments, "Stronger Invalid" and "Stronger Valid", determined by varying γ^* while holding the absolute strength fixed. Each row corresponds to the maximum correlation between instruments.

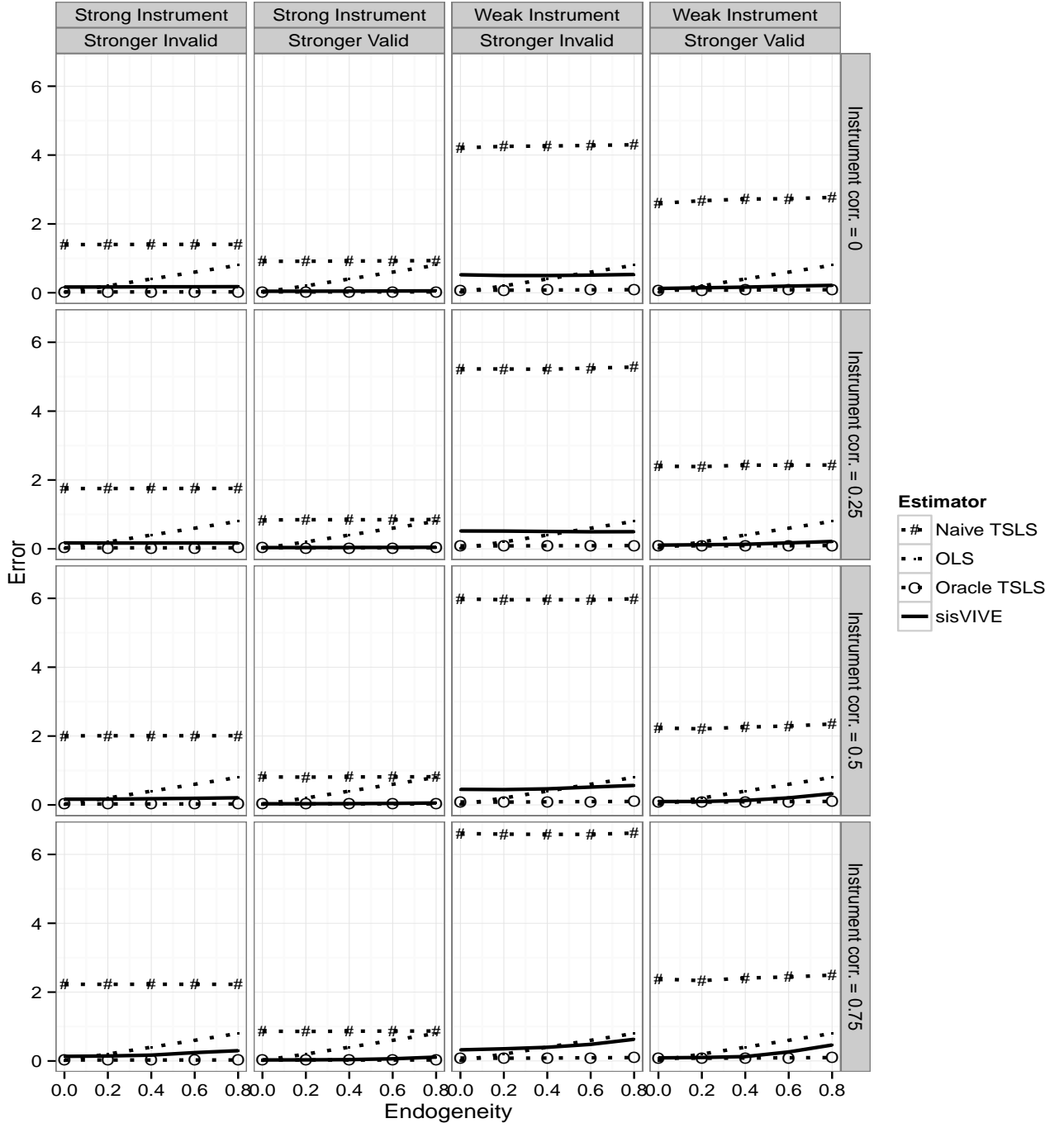


Figure 15: Simulation Study Varying Endogeneity and Correlation Only Exists Within Valid and Invalid Instruments. We also vary the instrument strength of valid and invalid instruments. There are ten ($L = 10$) instruments. Each line represents the median absolute estimation error ($|\beta^* - \hat{\beta}|$) after 500 simulations. We fix the number of invalid instruments to $s = 3$. Each column in the plot corresponds to a different variation of instruments' absolute and relative strength. There are two types of absolute strengths, "Strong" and "Weak", measured by the concentration parameter. There are two types of strengths for valid and invalid instruments, "Stronger Invalid" and "Stronger Valid", determined by varying γ^* while holding the absolute strength fixed. Each row corresponds to the maximum correlation between instruments, but correlation only exists within valid and invalid instruments.

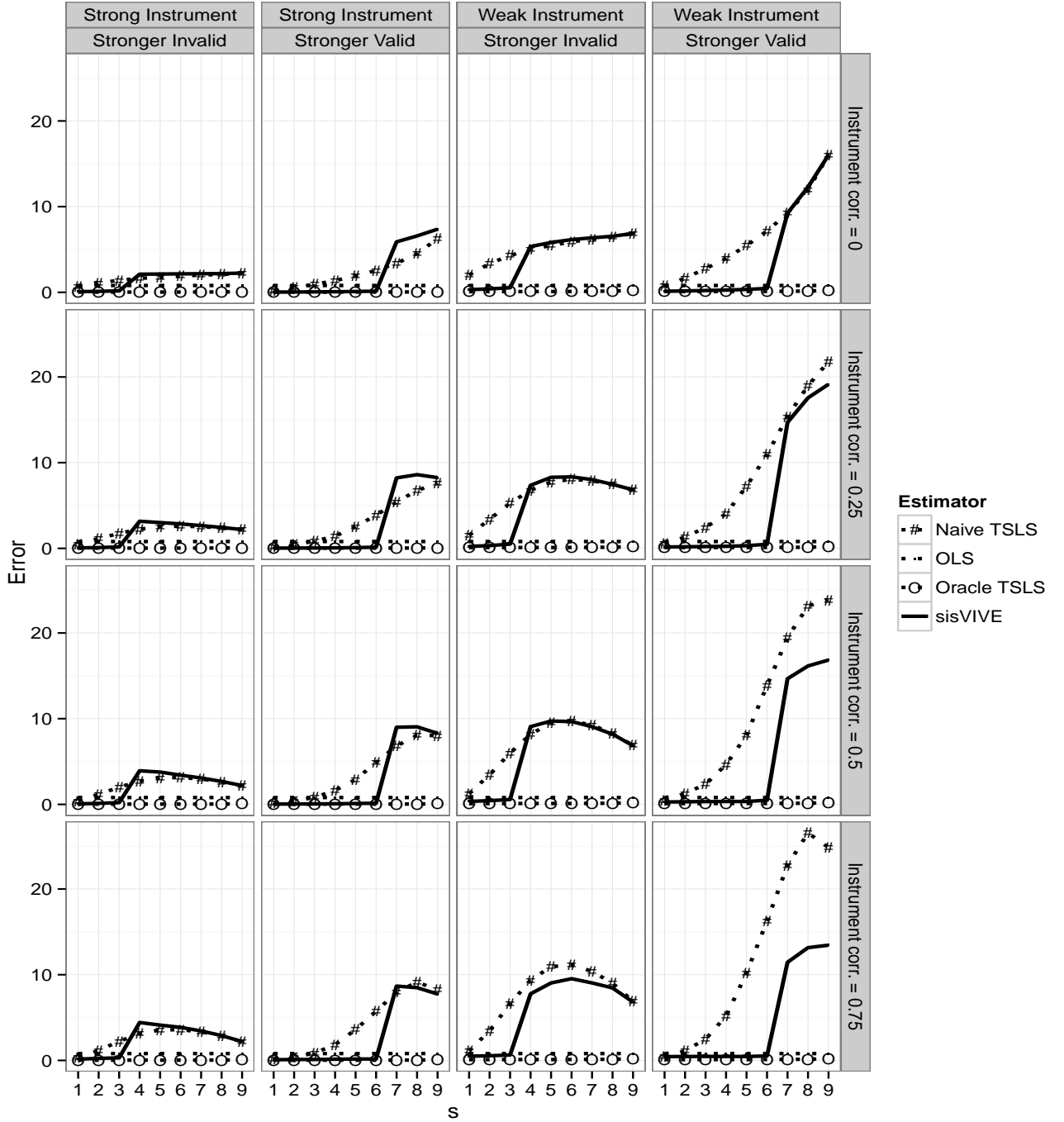


Figure 16: Simulation Study Varying the Number of Invalid Instruments (s) and Correlation Only Exists Within Valid and Invalid Instruments. We also vary the instrument strength of valid and invalid instruments. There are ten ($L = 10$) instruments. Each line represents the median absolute estimation error ($|\beta^* - \hat{\beta}|$) after 500 simulations. We fix the endogeneity $\sigma_{\epsilon\xi}^*$ to $\sigma_{\epsilon\xi}^* = 0.8$. Each column in the plot corresponds to a different variation of instruments' absolute and relative strength. There are two types of absolute strengths, “Strong” and “Weak”, measured by the concentration parameter. There are two types of strengths for valid and invalid instruments, “Stronger Invalid” and “Stronger Valid”, determined by varying γ^* while holding the absolute strength fixed. Each row corresponds to the maximum correlation between instruments, but correlation only exists

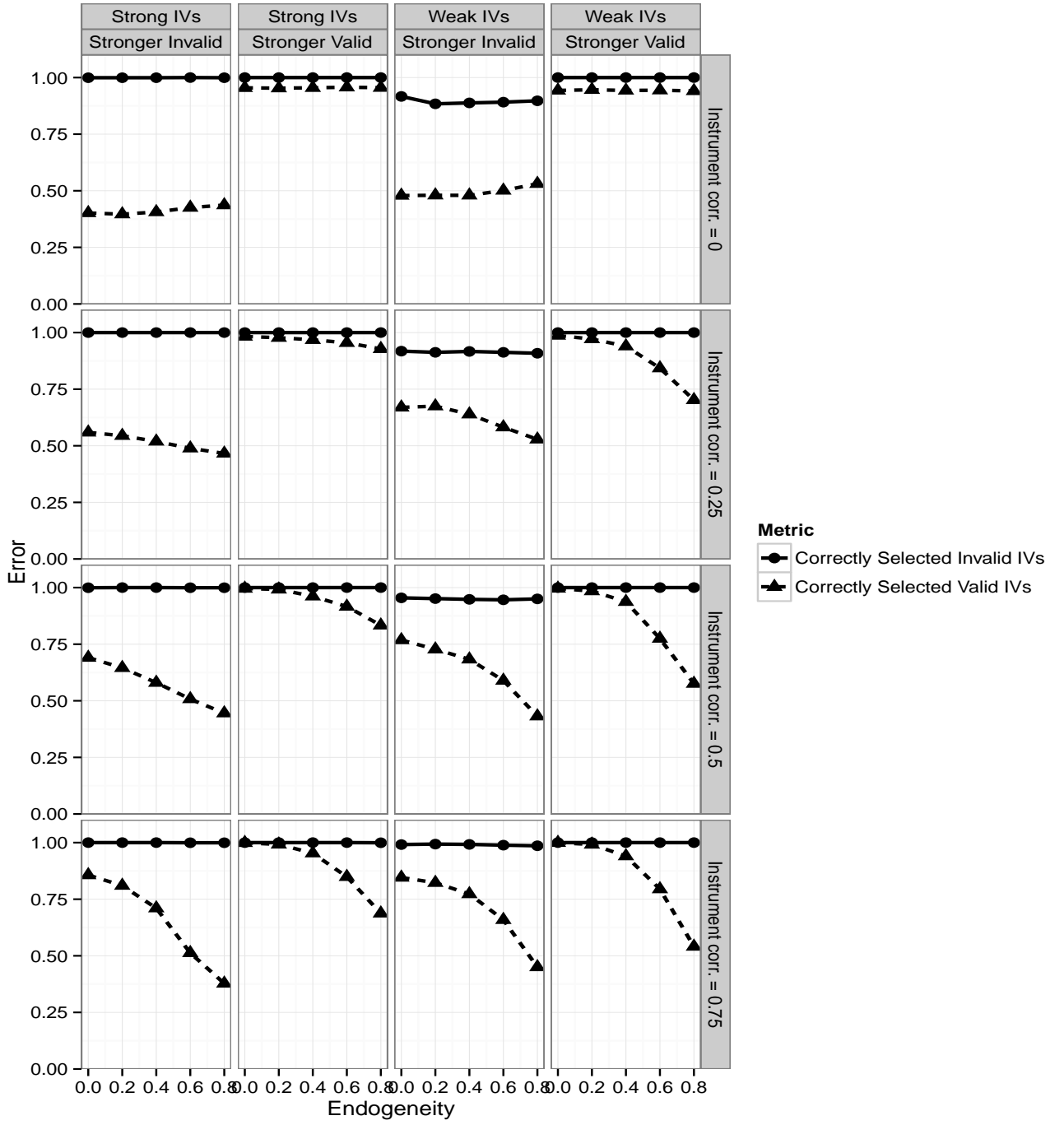


Figure 17: Simulation Study Varying Endogeneity and Correlation Only Exists Within Valid and Invalid Instruments. We also vary the instrument strength of valid and invalid instruments. There are ten ($L = 10$) instruments. Each line represents the average proportions of correctly selected valid instruments and correctly selected invalid instruments after 500 simulations. We fix the number of invalid instruments to $s = 3$. Each column in the plot corresponds to a different variation of instruments' absolute and relative strength. There are two types of absolute strengths, "Strong" and "Weak", measured by the concentration parameter. There are two types of strengths for valid and invalid instruments, "Stronger Invalid" and "Stronger Valid", determined by varying γ^* while holding the absolute strength fixed. Each row corresponds to the maximum correlation between instruments, but correlation only exists within valid and invalid instruments.

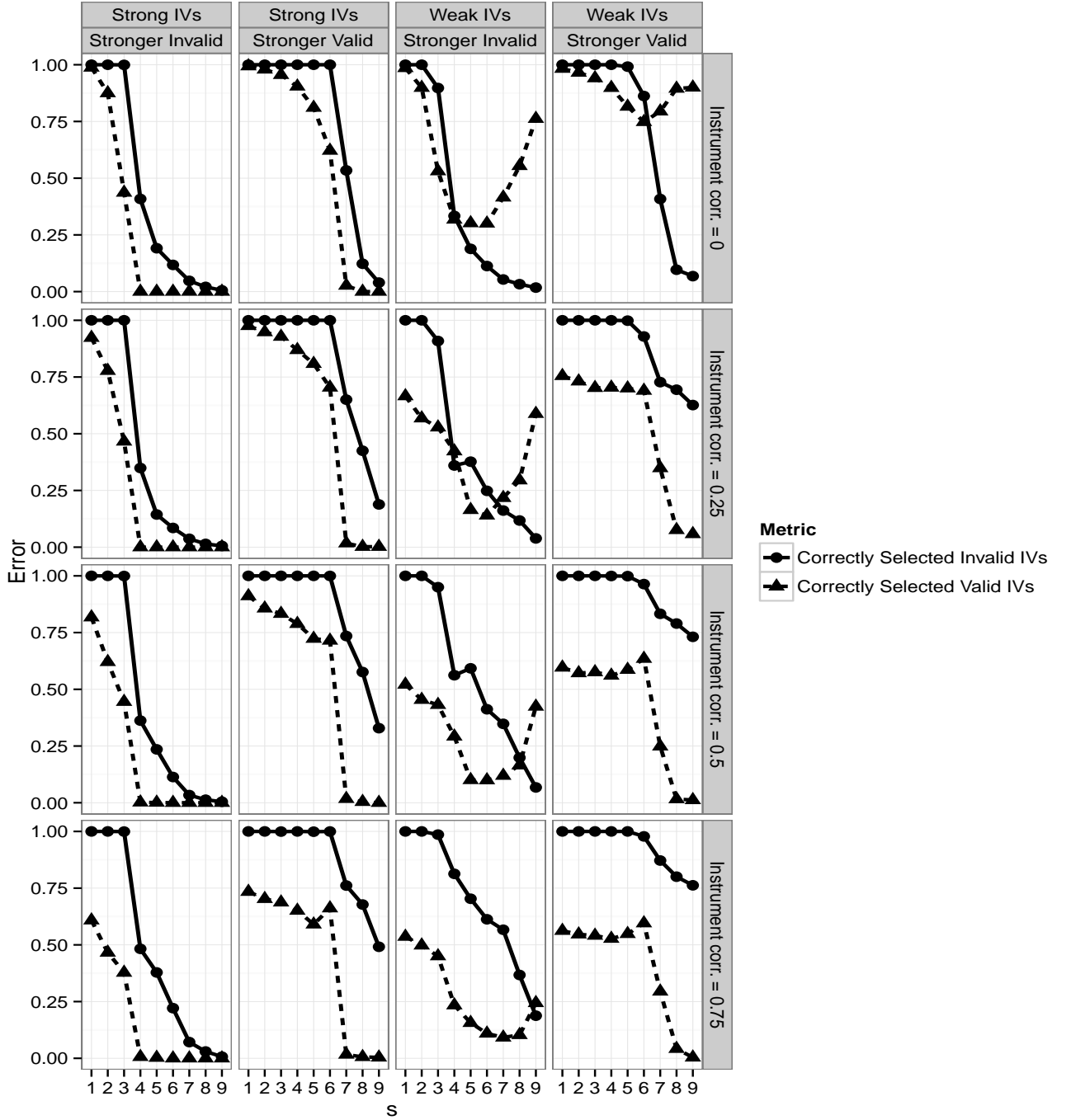


Figure 18: Simulation Study Varying the Number of Invalid Instruments (s) and Correlation Only Exists Within Valid and Invalid Instruments. We also vary the instrument strength of valid and invalid instruments. There are ten ($L = 10$) instruments. Each line represents the average proportions of correctly selected valid instruments and correctly selected invalid instruments after 500 simulations. We fix the endogeneity $\sigma_{\epsilon\xi}^*$ to $\sigma_{\epsilon\xi}^* = 0.8$. Each column in the plot corresponds to a different variation of instruments' absolute and relative strength. There are two types of absolute strengths, "Strong" and "Weak", measured by the concentration parameter. There are two types of strengths for valid and invalid instruments, "Stronger Invalid" and "Stronger Valid", determined by varying γ^* while holding the absolute strength fixed. Each row corresponds to the maximum correlation between instruments, but correlation only exists within valid and invalid instruments.

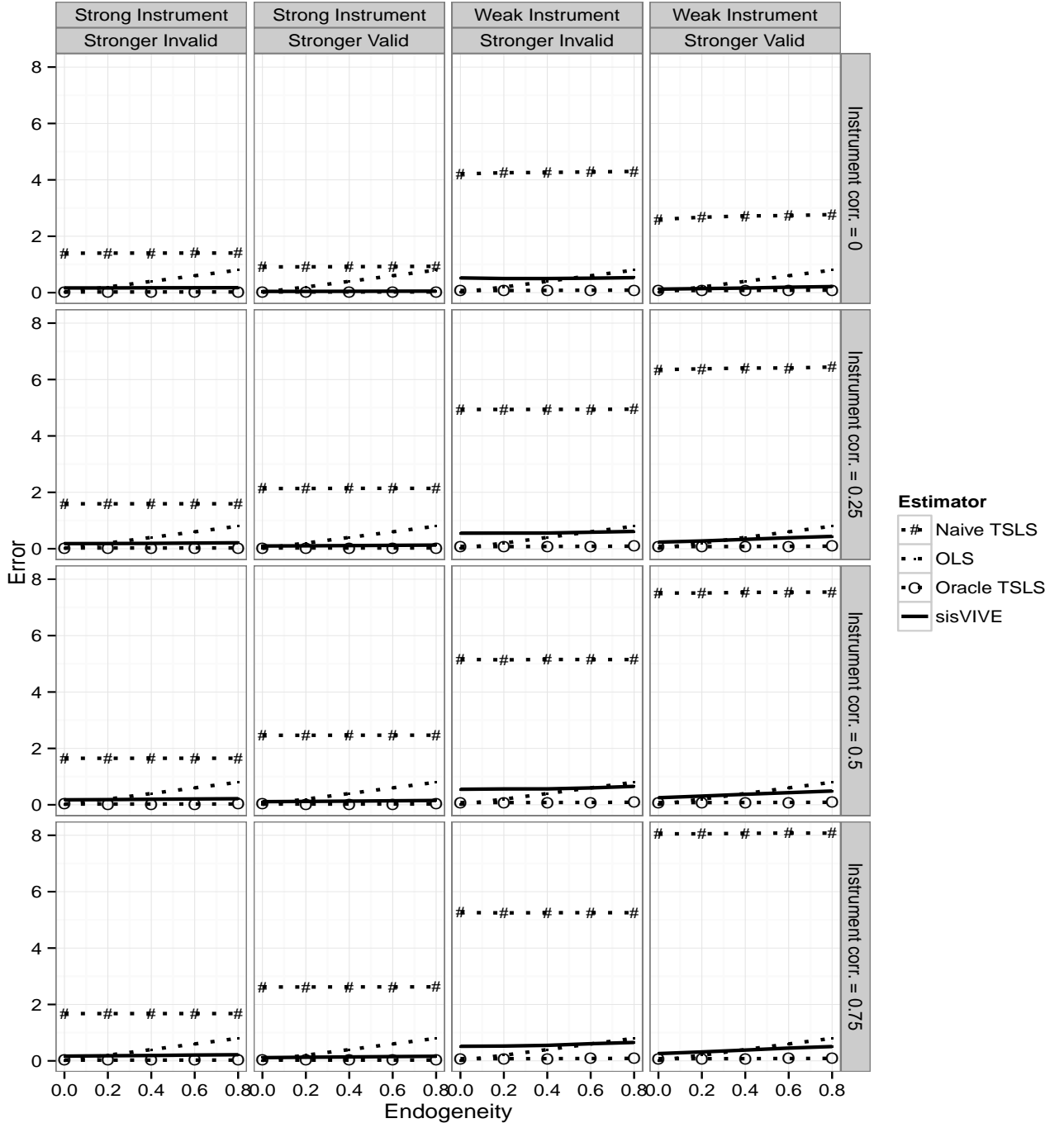


Figure 19: Simulation Study Varying Endogeneity and Correlation Only Exists Between Valid and Invalid Instruments. We also vary the instrument strength of valid and invalid instruments. There are ten ($L = 10$) instruments. Each line represents the median absolute estimation error ($|\beta^* - \hat{\beta}|$) after 500 simulations. We fix the number of invalid instruments to $s = 3$. Each column in the plot corresponds to a different variation of instruments' absolute and relative strength. There are two types of absolute strengths, "Strong" and "Weak", measured by the concentration parameter. There are two types of strengths for valid and invalid instruments, "Stronger Invalid" and "Stronger Valid", determined by varying γ^* while holding the absolute strength fixed. Each row corresponds to the maximum correlation between instruments, but correlation only exists between valid and invalid instruments.

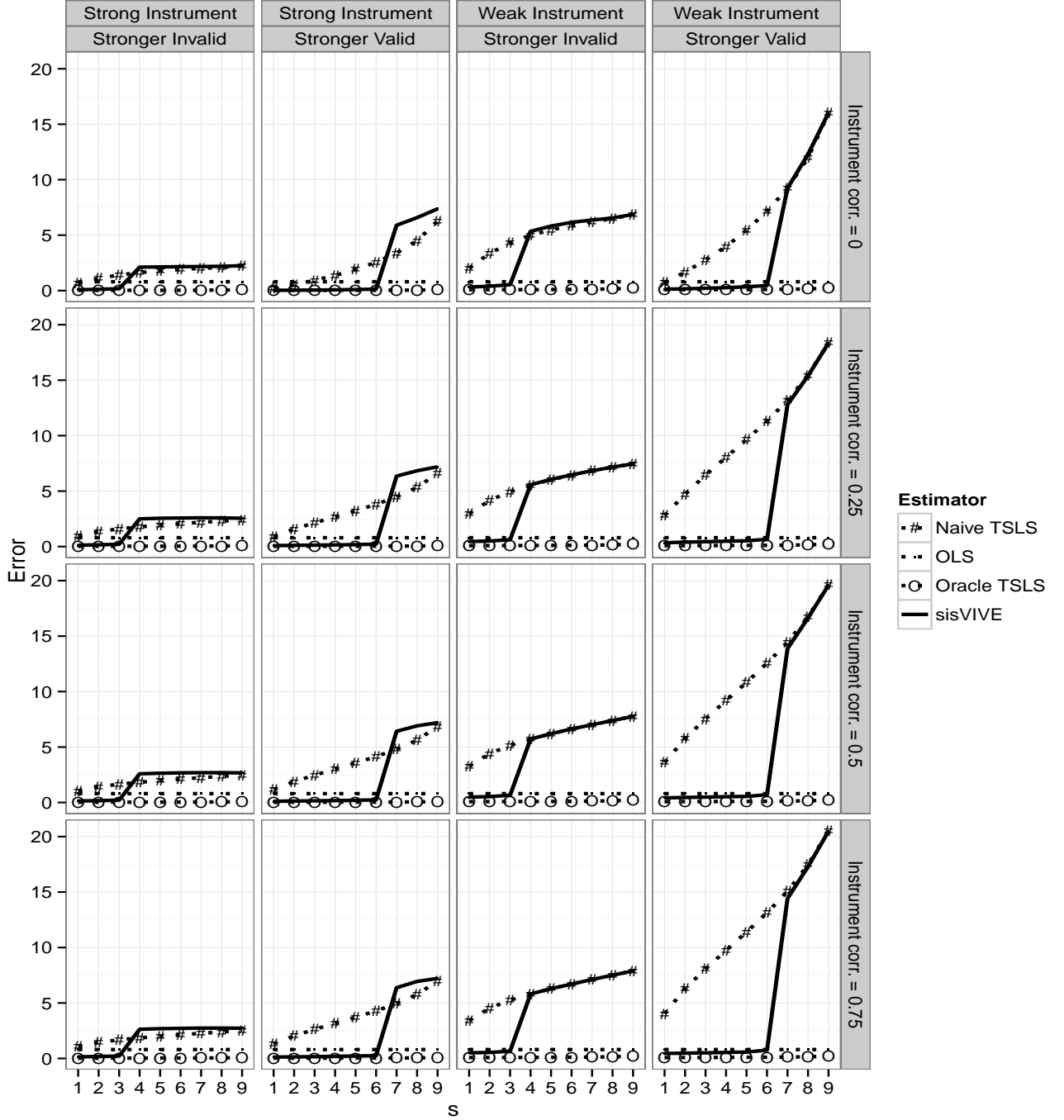


Figure 20: Simulation Study Varying the Number of Invalid Instruments (s) and Correlation Only Exists Between Valid and Invalid Instruments. We also vary the instrument strength of valid and invalid instruments. There are ten ($L = 10$) instruments. Each line represents median absolute estimation error ($|\beta^* - \hat{\beta}|$) after 500 simulations. We fix the endogeneity $\sigma_{\epsilon\xi}^*$ to $\sigma_{\epsilon\xi}^* = 0.8$. Each column in the plot corresponds to a different variation of instruments' absolute and relative strength. There are two types of absolute strengths, "Strong" and "Weak", measured by the concentration parameter. There are two types of strengths for valid and invalid instruments, "Stronger Invalid" and "Stronger Valid", determined by varying γ^* while holding the absolute strength fixed. Each row corresponds to the maximum correlation between instruments, but correlation only exists between valid and invalid instruments.

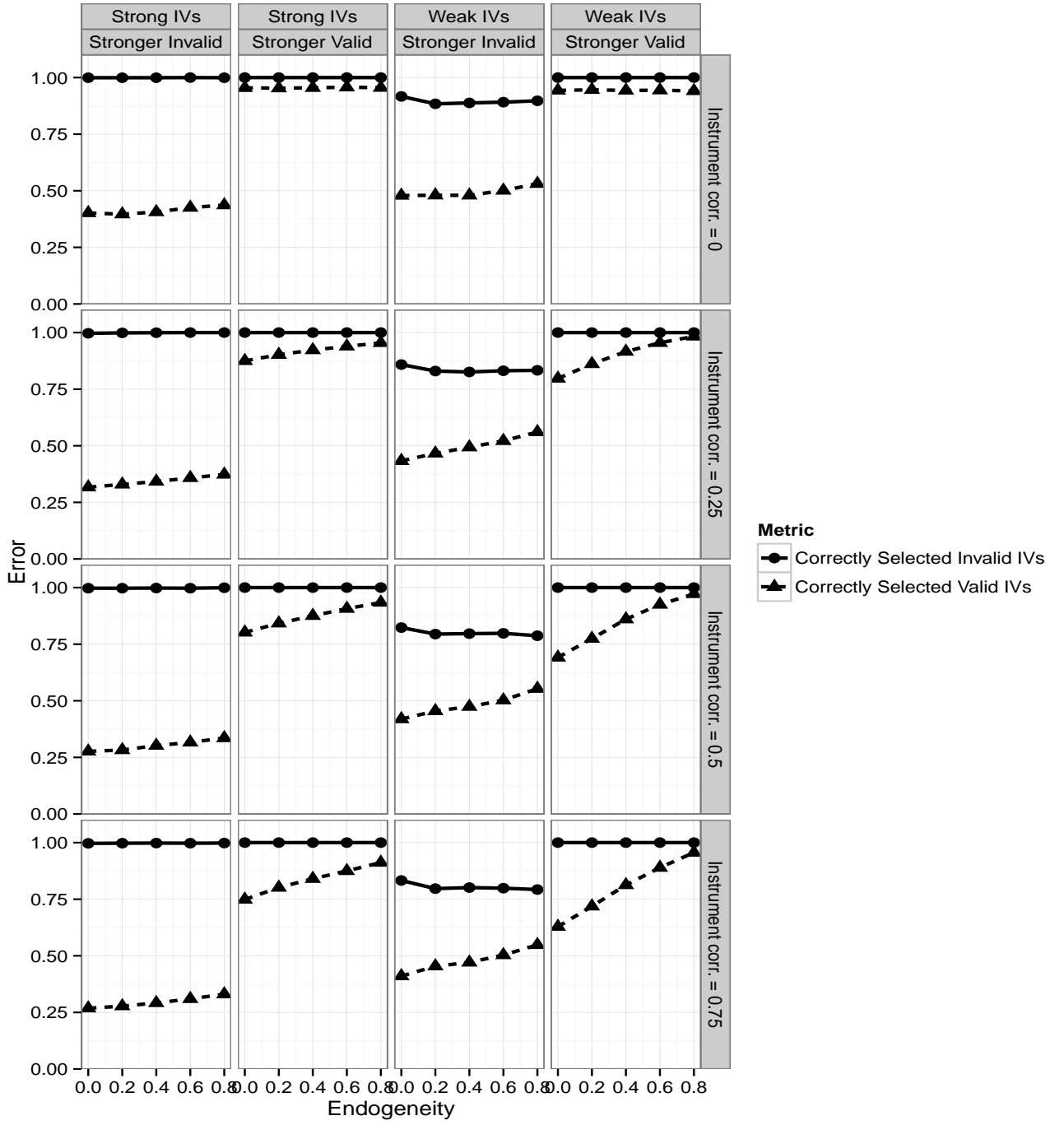


Figure 21: Simulation Study Varying Endogeneity and Correlation Only Exists Between Valid and Invalid Instruments. We also vary the instrument strength of valid and invalid instruments. There are ten ($L = 10$) instruments. Each line represents the average proportions of correctly selected valid instruments and correctly selected invalid instruments after 500 simulations. We fix the number of invalid instruments to $s = 3$. Each column in the plot corresponds to a different variation of instruments' absolute and relative strength. There are two types of absolute strengths, "Strong" and "Weak", measured by the concentration parameter. There are two types of strengths for valid and invalid instruments, "Stronger Invalid" and "Stronger Valid", determined by varying γ^* while holding the absolute strength fixed. Each row corresponds to maximum correlation between instruments, but correlation only exists between valid and invalid instruments.

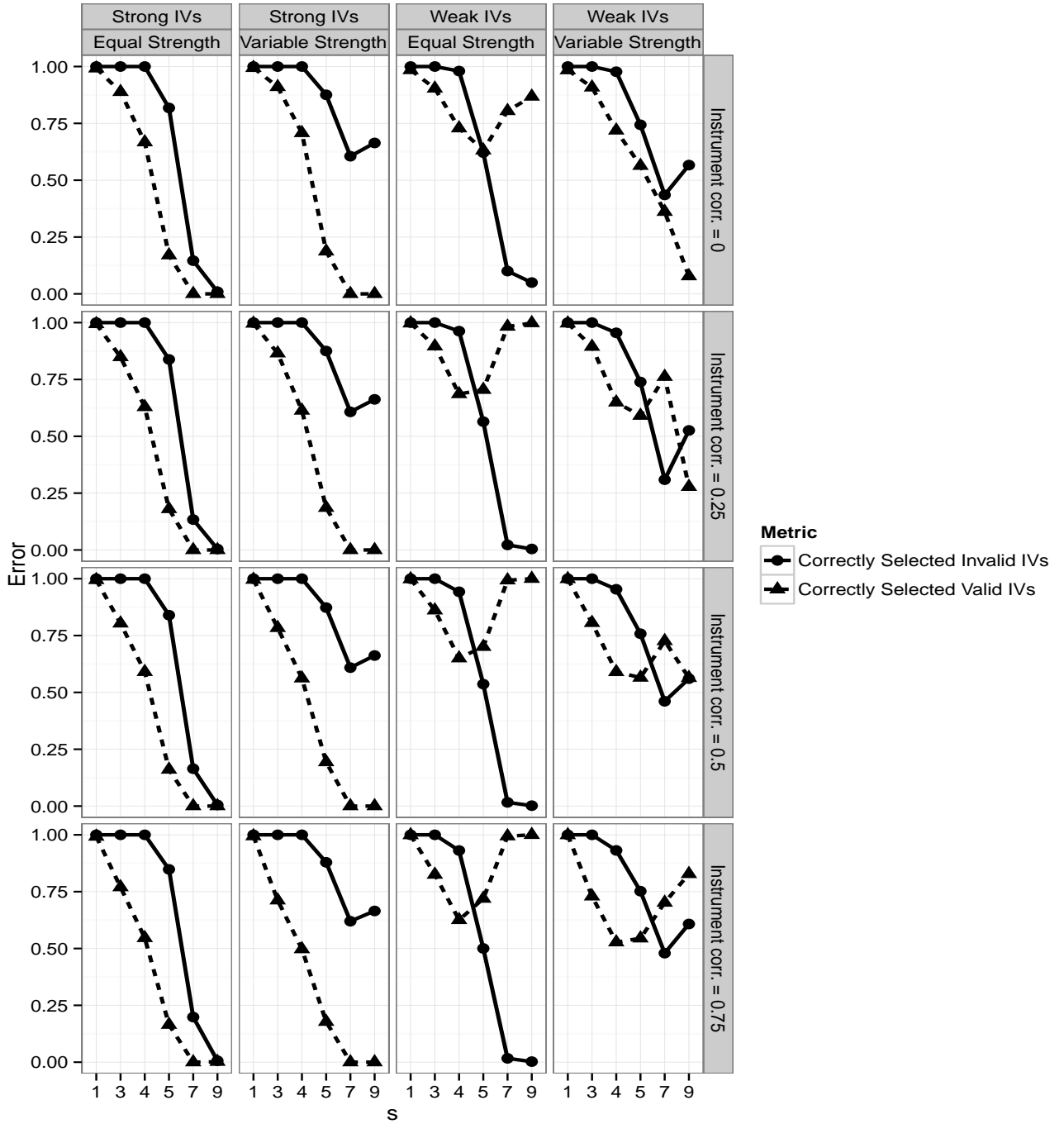


Figure 22: Simulation Study Varying the Number of Invalid Instruments (s) and Correlation Only Exists Between Valid and Invalid Instruments. We also vary the instrument strength of valid and invalid instruments. There are ten ($L = 10$) instruments. Each line represents the average proportions of correctly selected valid instruments and correctly selected invalid instruments after 500 simulations. We fix the endogeneity $\sigma_{\epsilon\xi}^*$ to $\sigma_{\epsilon\xi}^* = 0.8$. Each column in the plot corresponds to a different variation of instruments' absolute and relative strength. There are two types of absolute strengths, "Strong" and "Weak", measured by the concentration parameter. There are two types of strengths for valid and invalid instruments, "Stronger Invalid" and "Stronger Valid", determined by varying γ^* while holding the absolute strength fixed. Each row corresponds to a maximum correlation between instruments, but correlation only exists between valid and invalid instruments.

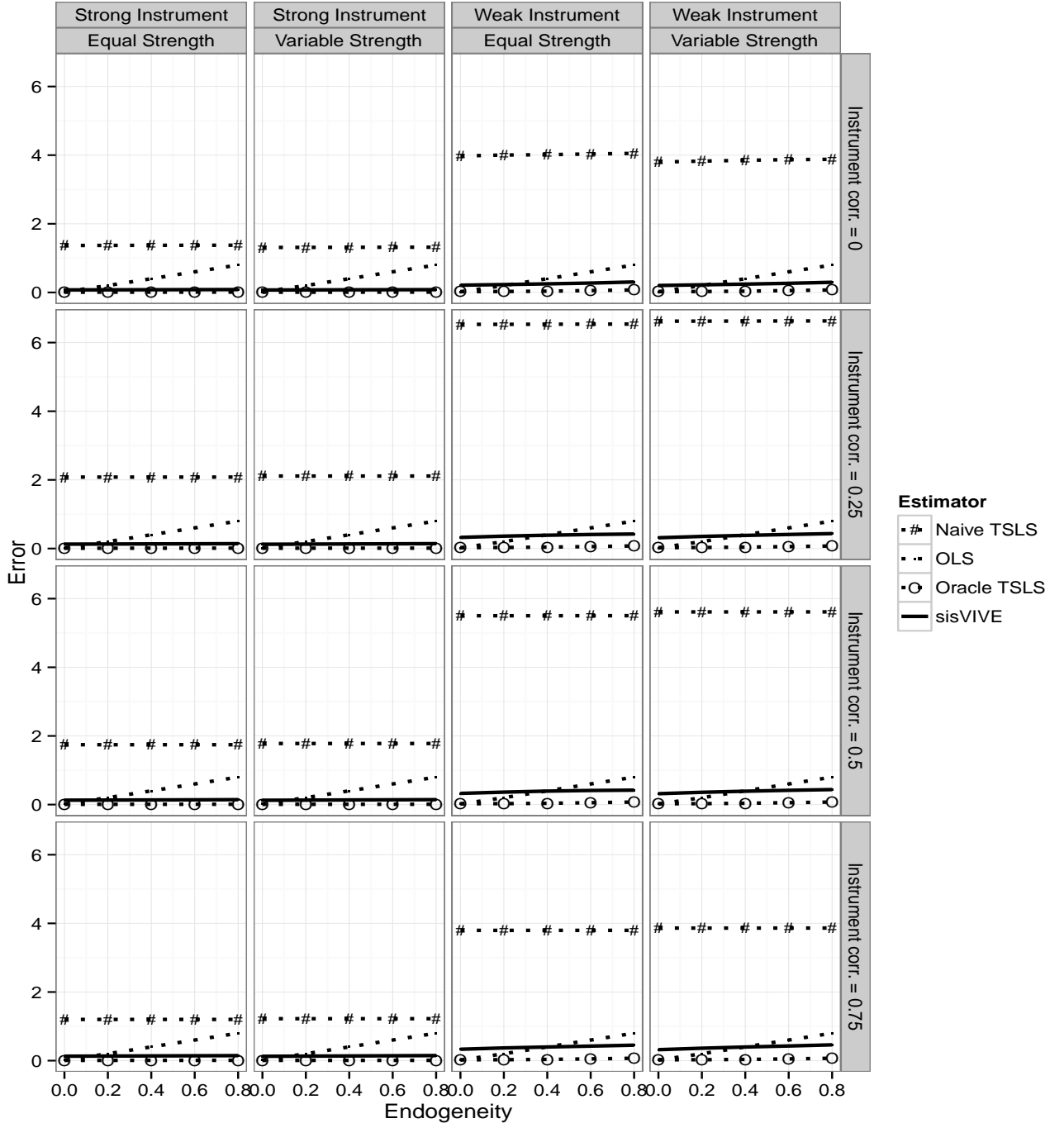


Figure 23: Simulation Study of Estimation Performance Varying Endogeneity and Correlation Exists Between All Instruments. There are 100 ($L = 100$) instruments. Each line represents the median absolute estimation error ($|\beta^* - \hat{\beta}|$) after 500 simulations. We fix the number of invalid instruments to $s = 30$. Each column in the plot corresponds to a different variation of instruments' absolute and relative strength. There are two types of absolute strengths, "Strong" and "Weak", measured by the concentration parameter. There are two types of relative strengths, "Equal" and "Variable", measured by varying γ^* while holding the absolute strength (i.e. concentration parameter) fixed. Each row corresponds to the maximum correlation between instruments.

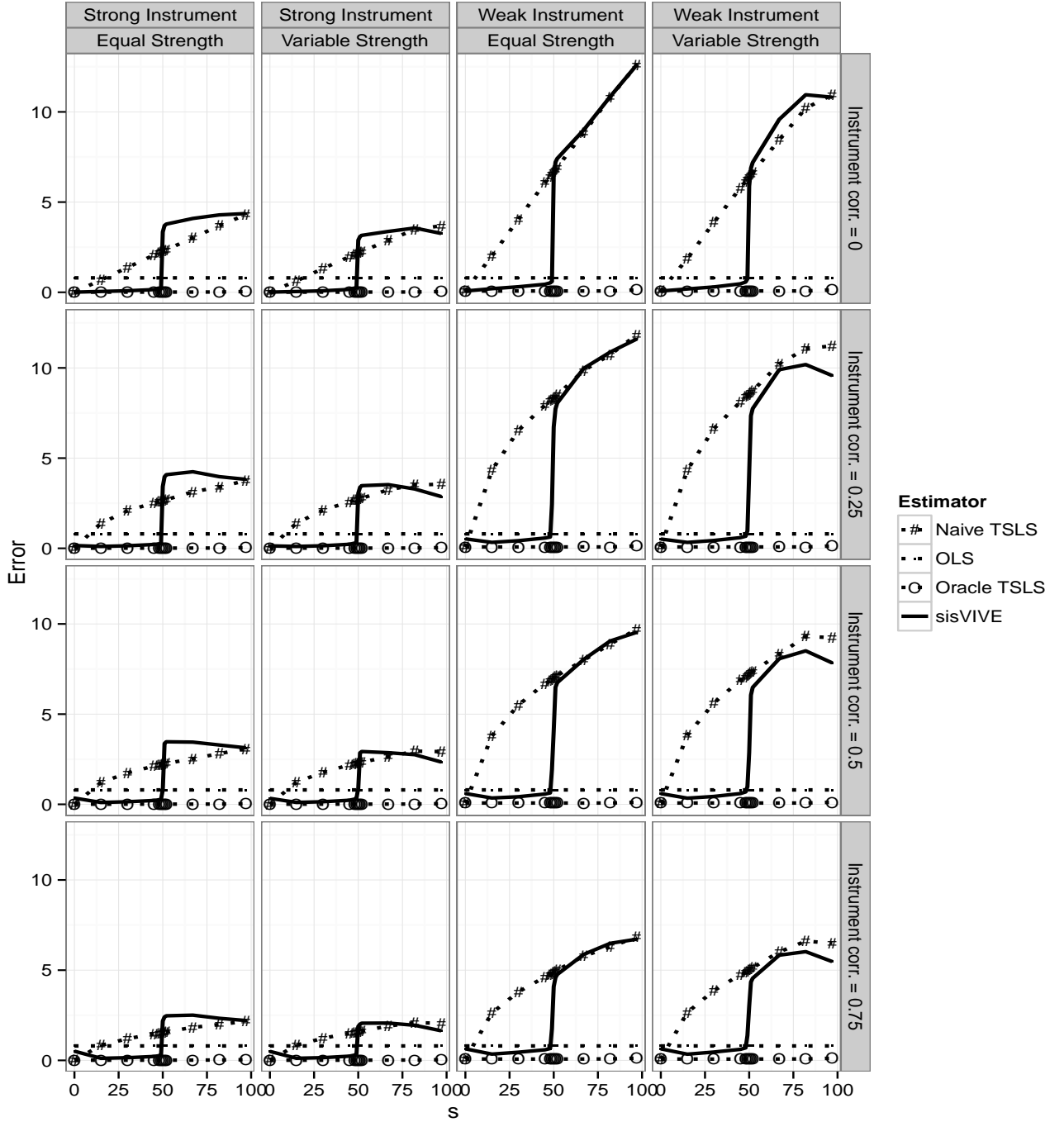


Figure 24: Simulation Study of Estimation Performance Varying the Number of Invalid Instruments (s) and Correlation Exists Between All Instruments. There are 100 ($L = 100$) instruments. Each line represents the median absolute estimation error ($|\beta^* - \hat{\beta}|$) after 500 simulations. We fix the endogeneity $\sigma_{\epsilon\xi}^*$ to $\sigma_{\epsilon\xi}^* = 0.8$. Each column in the plot corresponds to a different variation of instruments' absolute and relative strength. There are two types of absolute strengths, "Strong" and "Weak", measured by the concentration parameter. There are two types of relative strengths, "Equal" and "Variable", measured by varying γ^* while holding the absolute strength fixed. Each row corresponds to maximum correlation between instruments.

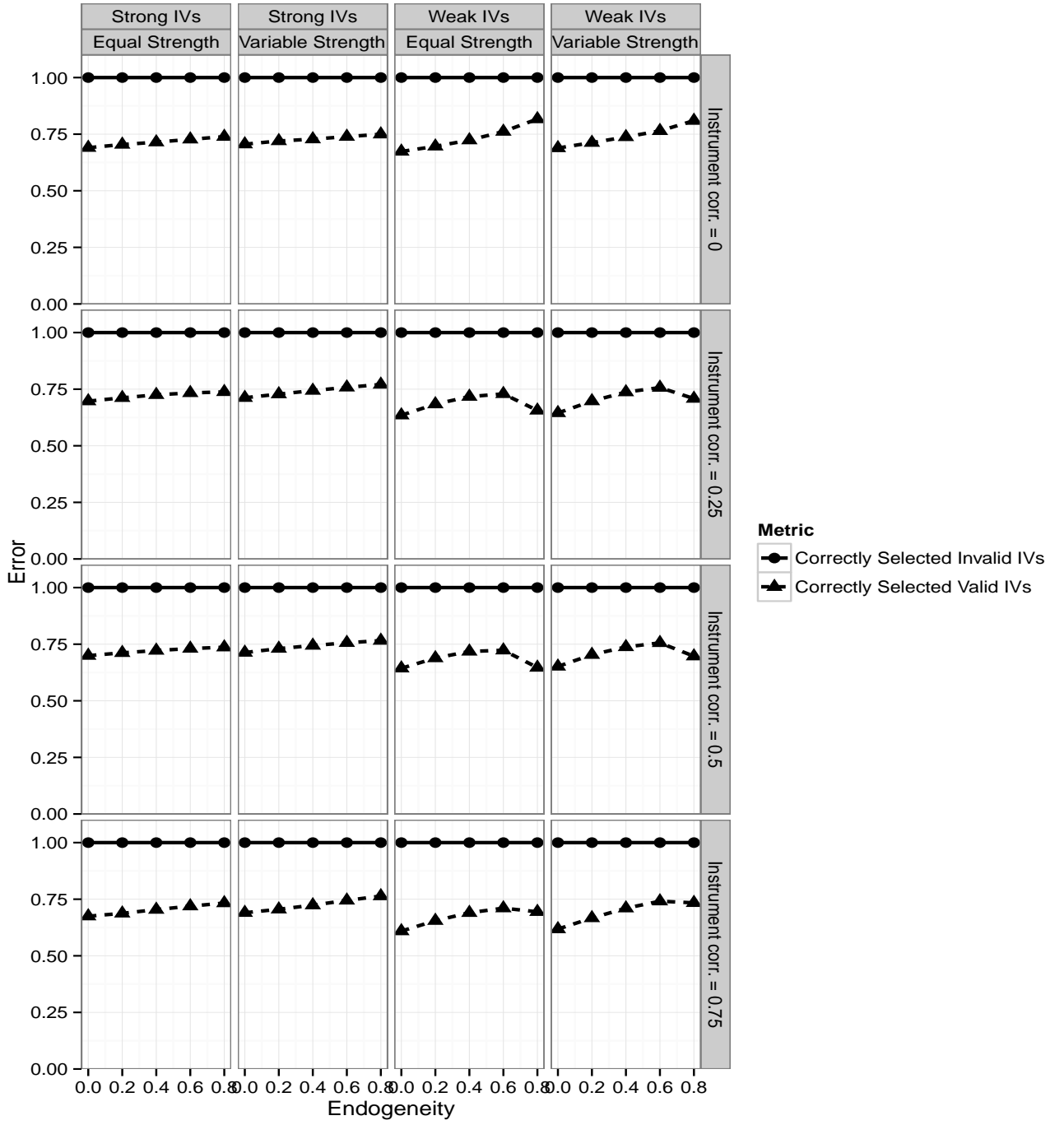


Figure 25: Simulation Study Varying Endogeneity and Correlation Exists Between All Instruments. There are ten ($L = 100$) instruments. Each line represents the average proportions of correctly selected valid instruments and correctly selected invalid instruments after 500 simulations. We fix the number of invalid instruments to $s = 30$. Each column in the plot corresponds to a different variation of instruments' absolute and relative strength. There are two types of absolute strengths, "Strong" and "Weak", measured by the concentration parameter. There are two types of relative strengths, "Equal" and "Variable", measured by varying γ^* while holding the absolute strength (i.e. concentration parameter) fixed. Each row corresponds to the maximum correlation between all instruments.

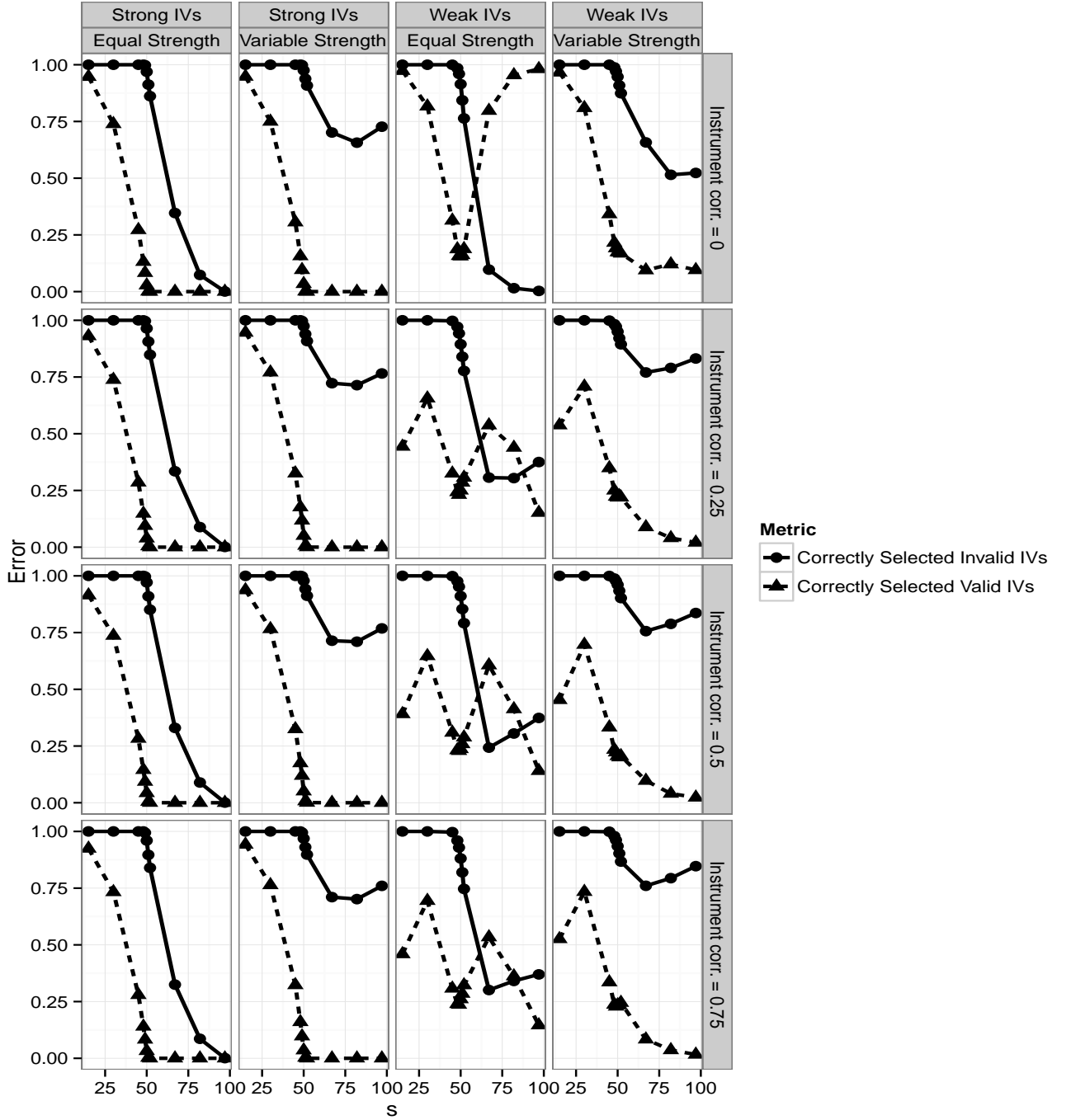


Figure 26: Simulation Study Varying the Number of Invalid Instruments (s) and Correlation Exists Between All Instruments. There are 100 ($L = 100$) instruments. Each line represents the average proportions of correctly selected valid instruments and correctly selected invalid instruments after 500 simulations. We fix the endogeneity $\sigma_{\epsilon\xi}^*$ to $\sigma_{\epsilon\xi}^* = 0.8$. Each column in the plot corresponds to a different variation of instruments' absolute and relative strength. There are two types of absolute strengths, "Strong" and "Weak", measured by the concentration parameter. There are two types of relative strengths, "Equal" and "Variable", measured by varying γ^* while holding the absolute strength fixed. Each row corresponds to maximum correlation between all instruments.