A Robustness Test for Estimating Total Effects with Covariate Adjustment

Abstract

Suppose we want to estimate a total effect with covariate adjustment in a linear structural equation model. We have a causal graph to decide what covariates to adjust for, but are uncertain about the graph. Here, we propose a testing procedure, that exploits the fact that there are multiple valid adjustment sets for the target total effect in the causal graph, to perform a robustness check on the graph. If the test rejects, it is a strong indication that we should not rely on the graph. We discuss what mistakes in the graph our testing procedure can detect and which ones it cannot and develop two strategies on how to select a list of valid adjustment sets for the procedure. We also connect our result to the related econometrics literature on coefficient stability tests.

1 INTRODUCTION

Suppose we are interested in estimating the total (causal) effect of a treatment X on an outcome Y from observational data. One popular approach to estimate such an effect is covariate adjustment, also known as adjusting for confounding. Deciding which covariates to adjust for is a difficult problem, but it can be answered precisely if we have knowledge of the underlying causal structure in the form of a graph [Pearl, 2009]. In particular, the class of covariate sets we may adjust for has been fully graphically characterised [Shpitser et al., 2010, Perković et al., 2018]. We refer to sets in this class as valid adjustment sets.

In some cases there is more than one valid adjustment set, which raises the question how we can exploit this. One approach is to try and select from the available valid adjustment sets the one that provides the most statistically efficient estimator [Kuroki and Miyakawa, 2003, Henckel et al., 2019, Rotnitzky and Smucler, 2020].

Another natural approach is to use multiple valid adjustment sets and test whether they all in fact lead to estimators of the same quantity. Such a test would be a simple and targeted robustness check on the causal graph we are relying on. Here we say targeted in the sense that the test would only detect mistakes in the graph that are relevant to our goal of estimating the total effect of interest, which is easier than checking whether the entire graph is correct.

In the econometrics literature, it is already common practice to estimate the total effect with multiple estimators and then check whether the estimates differ by a large margin [e.g. Dikova et al., 2019, Yigezu and El-Shater, 2021, Schlegel et al., 2021]. This approach is often called testing for coefficient stability [Walter and Tiemeier, 2009] or simply called a robustness check [Lu and White, 2014]. Practitioners often verify coefficient stability in a heuristic manner, but there is also a theoretical literature on the topic, especially for instrumental variable estimators [Frank et al., 2013, Oster, 2019].

For covariate adjustment estimators, Lu and White [2014] have proposed a formal test for coefficient stability. Their framework is not based on graphical models and therefore it is harder to decide which adjustment sets to use for their test. They propose to fix what they call a core of covariates and then create additional adjustment sets by adding what they call non-core covariates to the core. Here, the status of being core or non-core depends on certain conditional independences. In the graphical framework it becomes clear that this approach is too restrictive as, for example, two valid adjustment sets may be disjoint. As a result their approach may consider too few sets which leads to a loss of power.

There also exists a more general literature on validation tests for structural equation models. This literature, however, has focused on tests that either validate the entire model [Bollen, 1989, Bollen and Ting, 1993, Spirtes et al., 2000] or rely on instrumental variables [Kirby and Bollen, 2009].

In this paper, we adopt the framework of a linear structural equation model compatible with an unknown directed

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

acyclic graph (DAG). We propose a targeted robustness test that given a pair (X, Y) and a candidate DAG $\mathcal{G}$ tests whether the valid adjustment sets with respect to (X, Y) in $\mathcal{G}$ lead to estimates of the same quantity.

We first discuss, which mistakes in the candidate graph our robustness test has power for and how this depends on the valid adjustment sets we use for the test. We then propose a simple χ^2-test, similar to the one proposed by Lu and White [2014], although it differs in that we do not require a fixed core of covariates. We show that in general the joint asymptotic covariance matrix of the estimators we wish to compare is degenerate and that we need to know its rank for the test. This problem was acknowledged but not addressed by Lu and White [2014].

In response to the problem of the degenerate asymptotic covariance matrix, we propose two strategies. The first is to estimate the rank. This is a difficult statistical problem and may be unstable, especially in small samples. The upside of this approach, however, is that it allows us to use all valid adjustment sets for our test, which maximises its power.

The second strategy carefully selects a subset of the valid adjustment sets in a way that ensures the following two properties are likely to hold. First, the asymptotic covariance matrix is not degenerate. Second, we do not lose power completely against mistakes in the graph we had power for when using all valid adjustment sets. These two strategies represent different trade-offs between the stability of our testing procedure and its power to detect mistakes in the candidate graph.

Finally, we investigate with a simulation study how well our testing procedure controls the type-I error rate and how much power it has in finite samples. We do so for both of the strategies we propose, in order to compare and contrast their respective advantages and disadvantages. We also illustrate our testing procedure on a real data problem. All proofs are given in the supplementary materials. An implementation of our testing procedure and the code for our simulation study are made available at `https://anonymous.4open.science/r/RoCA-UAI`.

2 PRELIMINARIES

We consider a linear structural equation model compatible with a DAG, where nodes represent random variables and edges represent direct effects. We now provide the most important definitions. The remaining definitions are provided in Section A of the supplementary materials.

Linear structural equation models. Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a DAG. Then $\mathbf{V} = (V_1, \ldots, V_p)$ follows a *linear structural equation model* compatible with $\mathcal{G}$ if for all $i = 1, \ldots, p$

$$V_i \leftarrow \sum_{V_j \in \mathrm{pa}(V_i, \mathcal{G})} \alpha_{ij} V_j + \epsilon_i,$$

with edge coefficients α_{ij} and jointly independent errors ϵ_i with zero mean and finite variance. We do not assume that the errors are normally distributed.

Total effects. Consider a pair (X, Y) of random variables. The *total effect* of X on Y is the partial derivative of the expectation $\mathrm{E}(Y \mid do(X = x))$ with respect to x. This is the instantaneous change of the average of Y in the world where X is set to x [Pearl, 2009]. In a linear structural equation model, the partial derivative is a constant slope that does not depend on x. As a result, the total effect is simply a number τ_{yx}.

Causal and forbidden nodes. Consider two nodes X and Y in a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$. The *causal nodes* relative to (X, Y) in $\mathcal{G}$, denoted $\mathrm{cn}(X, Y, \mathcal{G})$, are all nodes on directed paths from X to Y, excluding X. The descendants of X in $\mathcal{G}$, denoted $\mathrm{de}(X, \mathcal{G})$, are all nodes V such that there exists a directed path from X to V in $\mathcal{G}$. The *forbidden nodes* relative to (X, Y) in $\mathcal{G}$, denoted $\mathrm{forb}(X, Y, \mathcal{G})$, are all nodes that are descendants of causal nodes, including X. The *non-forbidden nodes* relative to (X, Y) in $\mathcal{G}$, denoted $\mathrm{nonforb}(X, Y, \mathcal{G})$, are the nodes in $\mathbf{V} \setminus \mathrm{forb}(X, Y, \mathcal{G})$.

Notation for regression coefficients. Consider random variables X and Y, random vectors $\mathbf{Z}_1, \ldots, \mathbf{Z}_k$ and the collection of adjustment sets $\mathcal{Z} = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_k\}$. Let $\beta_{yx.\mathbf{z}_i}$ indicate the population level regression coefficient of X in the ordinary least squares regression of Y on X and $\mathbf{Z}_i$. Let $\hat{\beta}_{yx.\mathbf{z}_i}$ denote the corresponding estimator. Let $\boldsymbol{\beta}_{yx.\mathcal{Z}}$ denote the stacked population regression coefficients $(\beta_{yx.\mathbf{z}_1}, \ldots, \beta_{yx.\mathbf{z}_k})^\top$ and $\hat{\boldsymbol{\beta}}_{yx.\mathcal{Z}}$ the corresponding estimator.

Valid adjustment sets. Consider nodes X and Y in a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$. A node set $\mathbf{Z}$ is a *valid adjustment set* relative to (X, Y) in $\mathcal{G}$ if for all linear structural equation models compatible with $\mathcal{G}$, $\beta_{yx.\mathbf{z}} = \tau_{yx}$. We say a valid adjustment set $\mathbf{Z} = \{Z_1, \ldots, Z_k\}$ is *minimal* if for all $i \in \{1, \ldots, k\}$, $\mathbf{Z} \setminus Z_i$ is not a valid adjustment set. The class of valid adjustment sets has been fully characterised as follows.

Adjustment criterion. [Shpitser et al., 2010, Perković et al., 2018] A (possibly empty) set $\mathbf{Z}$ is a valid adjustment set relative to (X, Y) in $\mathcal{G}$ if and only if

1. $\mathbf{Z}$ contains no node in $\mathrm{forb}(X, Y, \mathcal{G})$, and
2. $\mathbf{Z}$ blocks all paths between X and Y in $\mathcal{G}$ that are not directed from X to Y.

d-separation. Consider three disjoint node sets $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$ in a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, such that $\mathbf{V}$ follows a linear structural equation model compatible with $\mathcal{G}$. We can read off from $\mathcal{G}$ whether $\mathbf{X}$ is independent of $\mathbf{Y}$ given $\mathbf{Z}$ with a graphical criterion called *d-separation* [Pearl, 2009] which we define formally in the supplementary materials. We use the notation $\mathbf{X} \perp_{\mathcal{G}} \mathbf{Y} \mid \mathbf{Z}$ to denote that $\mathbf{X}$ is d-separated from $\mathbf{Y}$ given $\mathbf{Z}$ in $\mathcal{G}$.

3 A TARGETED ROBUSTNESS TEST FOR COVARIATE ADJUSTMENT

3.1 THE NULL HYPOTHESIS AND ITS PROPERTIES

Suppose we wish to estimate the total effect of a treatment X on a response variable Y. Let $\mathcal{G}_0$ denote the unknown true underlying casual graph and suppose we have a candidate causal graph $\mathcal{G}$ that describes our understanding of the underlying causal structure but that we are not certain about. We would like to check whether the candidate graph is plausible, so we can rely on it to estimate τ_{yx} with some confidence. In order to do so, we use $\mathcal{G}$ to identify a collection of valid adjustment sets $\mathcal{Z} = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_k\}$ with respect to (X, Y) in $\mathcal{G}$. If $\mathcal{G}$ is correct each of these sets corresponds to a consistent estimator of τ_{yx}, i.e.,

$$\tau_{yx} = \beta_{yx.\mathbf{z}_1} = \beta_{yx.\mathbf{z}_2} = \cdots = \beta_{yx.\mathbf{z}_k}. \tag{1}$$

If $\mathbf{Z}$ consists of more than one set, then equation (1) imposes an over-identifying constraint on the total effect τ_{yx}. We use this constraint to test the plausibility of the candidate graph $\mathcal{G}$. The more valid adjustment sets $\mathbf{Z}_i$ we use, the more mistakes in $\mathcal{G}$ the test can detect.

It is generally not possible to directly test the constraint from equation (1) with observational data because we do not know the true total effect τ_{yx}. It is, however, possible to test the relaxed null hypothesis

$$H_0 : \beta_{yx.\mathbf{z}_1} = \beta_{yx.\mathbf{z}_2} = \cdots = \beta_{yx.\mathbf{z}_k}.$$

Let H_0^* denote the null hypothesis associated with equation (1). Any valid test for H_0 necessarily controls the type-I error for the null hypothesis H_0^*, so we can test H_0 as a proxy for testing H_0^*. In particular, any rejection of H_0 implies a rejection of H_0^* and therefore of the candidate graph $\mathcal{G}$.

There is an even more restrictive null hypothesis H_0^{**} : $\mathcal{G} = \mathcal{G}_0$. However, as we are interested in estimating one specific total effect, it is not necessary to validate the entire candidate graph $\mathcal{G}$, and the distinction between H_0^* and H_0^{**} is irrelevant for the purposes of this paper. It is in this sense that the testing procedure is targeted.

There are cases where H_0 holds but H_0^* does not, and as a result any test for H_0 will have no power to reject H_0^*. This occurs whenever H_0 holds but for all $\mathbf{Z}_i \in \mathcal{Z}$, $\beta_{yx.\mathbf{z}_i} \neq \tau_{yx}$. Whether this is the case depends on the choice of candidate sets $\mathcal{Z}$ and is more likely if $\mathcal{Z}$ contains few sets. In response, it is natural to use all available valid adjustment sets relative to (X, Y) in $\mathcal{G}$ to maximise the number of sets in $\mathcal{Z}$. However, even for a moderately sized $\mathcal{G}$ the number of valid adjustment sets relative to the pair (X, Y) can be very large, e.g., there are 72 in the graph $\mathcal{G}_0$ and 96 in the graph $\mathcal{G}_1$ from Figure 1. This raises the question whether it is possible

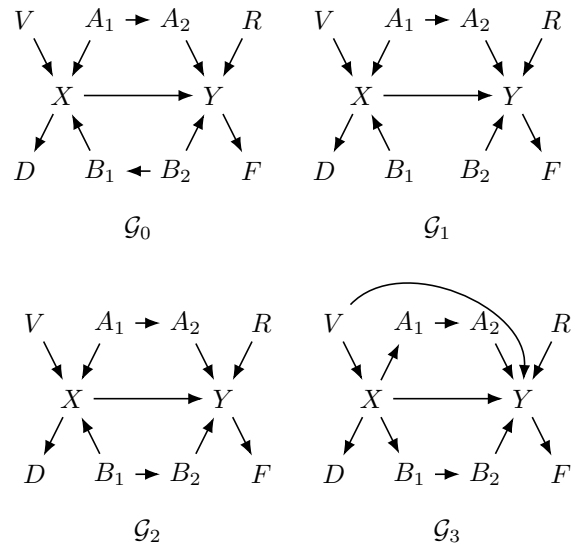

Figure 1: Graphs used in Examples 2, 4 and 9.

to select the collection $\mathcal{Z}$ in a way that minimises the risk of having no power for H_0^*, while simultaneously limiting its size.

We now provide a necessary condition on $\mathcal{Z}$ under which the problematic case that H_0 holds but H_0^* does not is rare, and as a result testing H_0 is a good proxy for testing H_0^*.

Theorem 1. *Consider nodes X and Y in a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ such that $Y \in \mathrm{de}(X, \mathcal{G})$. Let $\mathcal{Z} = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_k\}$ be a collection of node sets in $\mathcal{G}$. Suppose there exists a $\mathbf{Z}_i$, such that $\mathrm{forb}(X, Y, \mathcal{G}) \cap \mathbf{Z}_i = \emptyset$ and $\mathbf{Z}_i$ is not a valid adjustment set relative to (X, Y) in $\mathcal{G}$. Further, suppose that $(\mathbf{V} \setminus \mathrm{forb}(X, Y, \mathcal{G})) \subseteq \bigcup_{j=1}^{k} \mathbf{Z}_j$. If we sample the edge coefficients and error variances for a linear structural equation model compatible with $\mathcal{G}$ from a distribution P such that P is absolutely continuous with respect to the Lebesgue measure, then P-a.s. there exists a $\mathbf{Z}_j$ such that $\beta_{yx.\mathbf{z}_i} \neq \beta_{yx.\mathbf{z}_j}$.*

Theorem 1 gives two important but also intuitive insights on how to select $\mathcal{Z}$. First, the sets in $\mathcal{Z}$ should cover as many nodes in the candidate graph $\mathcal{G}$ as possible, i.e., ideally all non-forbidden nodes. Second, we should minimise the number of nodes that appear in all sets $\mathbf{Z}_i \in \mathcal{Z}$, and some of the candidate sets $\mathbf{Z}_i$ should be as small as possible. This minimises the risk that all adjustment sets $\mathbf{Z}_i$ contain a node that is a forbidden node in the true graph. Note that this is very different from the strategy proposed by Lu and White [2014].

Example 2. Consider the graphs from Figure 1 and the linear structural equation model compatible with $\mathcal{G}_0$, where all edge coefficients and error variances equal 1. We now illustrate for three candidate graphs that differ from $\mathcal{G}_0$, whether we can use tests for the null hypothesis H_0 to detect

the mistakes in the candidate graphs and how this depends on the choice of sets $\mathcal{Z}$.

Consider the candidate graph $\mathcal{G}_1$ and the collection

$$\mathcal{Z} = \{\{A_1\}, \{A_1, A_2\}, \{A_1, A_2, R\}\}$$

of three valid adjustment sets relative to (X, Y) in $\mathcal{G}_1$. A direct calculation shows that $\boldsymbol{\beta}_{yx.\mathcal{Z}} = (1.25, 1.25, 1.25)^\top$, none of which are equal to the total effect $\tau_{yx} = 1$. In this case, the null hypothesis H_0 is true even though the null hypothesis H_0^* is false, i.e., testing H_0 will not detect that there is a mistake in the candidate graph. However, if we add the set of non-forbidden nodes $\mathbf{Z}_4 = \{A_1, A_2, B_1, B_2, V, D, R\}$ to $\mathcal{Z}$, then H_0 no longer holds as $\mathbf{Z}_4$ is a valid adjustment set in $\mathcal{G}_0$ and therefore $\beta_{yx.\mathbf{z}_4} = 1$. In this case testing H_0 will detect that there is a mistake in the candidate graph.

Consider now the candidate graph $\mathcal{G}_2$. It has exactly the same valid adjustment sets relative to (X, Y) as the true graph $\mathcal{G}_0$. Testing H_0 will therefore not detect that there is a mistake in the candidate graph, irrespective of the collection of valid adjustment sets. Since the two graphs are equivalent with respect to estimating the total effect τ_{yx} with covariate adjustment this is not a concern.

Consider now the candidate graph $\mathcal{G}_3$. All valid adjustment sets relative to (X, Y) in $\mathcal{G}_3$ result in estimates of 1.5. Therefore testing H_0 will not detect that there is a mistake in the candidate graph, irrespective of the collection of valid adjustment sets. Interestingly, in the graph $\mathcal{G}_3'$ equal to $\mathcal{G}_3$ but with the edge $V \to Y$ removed, we can detect the mistakes by testing H_0 with, for example, the collection $\mathcal{Z}$ of all valid adjustment sets in $\mathcal{G}_3'$. This is an example of using an instrument to detect omitted variables bias [cf. Chen and Pearl, 2015], which our test implicitly exploits.

3.2 THE TEST STATISTIC

As a preparatory result and for completeness, we first derive the asymptotic distribution of the estimator $\hat{\boldsymbol{\beta}}_{yx.\mathcal{Z}}$.

Lemma 3. *Consider a p-dimensional random vector $\mathbf{V} = (V_1, V_2, \ldots, V_p)$ that follows a distribution where $\mathrm{E}(V_\ell^4) < \infty$ for all $1 \le \ell \le p$. Given two random variables $X, Y \in \mathbf{V}$, let $\mathcal{Z} = \{\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_k\}$, $k \ge 2$, be a collection of random subvectors of $\mathbf{V}$ that do not contain X or Y, and let $\mathbf{Z}_i' = (X, \mathbf{Z}_i^\top)^\top$ for $i = 1, 2, \ldots, k$. Then the random vector $\sqrt{n}(\hat{\boldsymbol{\beta}}_{yx.\mathcal{Z}} - \boldsymbol{\beta}_{yx.\mathcal{Z}})$ converges in distribution to a multivariate normal random variable with mean zero and covariance matrix*

$$(\boldsymbol{\Sigma}_{\mathcal{Z}})_{ij} = \frac{\mathrm{E}(\delta_{x\mathbf{z}_i}\delta_{y\mathbf{z}_i'}\delta_{x\mathbf{z}_j}\delta_{y\mathbf{z}_j'})}{\mathrm{E}(\delta_{x\mathbf{z}_i}^2)\,\mathrm{E}(\delta_{x\mathbf{z}_j}^2)} \quad 1 \le i, j \le k, \tag{2}$$

where $\delta_{y\mathbf{z}_i'} = Y - \boldsymbol{\beta}_{y\mathbf{z}_i'}\mathbf{Z}_i'$ and $\delta_{x\mathbf{z}_i} = X - \boldsymbol{\beta}_{x\mathbf{z}_i}\mathbf{Z}_i$.

In general the covariance matrix $\boldsymbol{\Sigma}_{\mathcal{Z}}$ of the limiting normal distribution will not be of full rank, that is, the distribution will be degenerate. To illustrate this, we now give an example.

Example 4. Consider again the DAG $\mathcal{G}_0$ and the linear structural equation model from Example 2. Let

$$\mathcal{Z} = \{\{A_1, B_1\}, \{A_1, A_2, B_1\}, \{A_1, B_1, B_2\}, \{A_1, A_2, B_1, B_2\}\}.$$

The asymptotic covariance matrix $\boldsymbol{\Sigma}_{\mathcal{Z}}$ of $\hat{\boldsymbol{\beta}}_{yx.\mathcal{Z}}$ is the rank-3 matrix

$$\begin{pmatrix} 1.75 & 1.25 & 1.5 & 1 \\ 1.25 & 1.25 & 1 & 1 \\ 1.5 & 1 & 1.5 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

We now reformulate and slightly generalise the null hypothesis H_0 from Section 3.1 as follows. Consider a pair of random variables (X, Y) and a collection of random vectors $\mathcal{Z} = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_k\}$. Define a contrast matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{(k-1)\times k}$ such that $\boldsymbol{\Gamma}\mathbf{1} = \mathbf{0}$ and $\mathrm{rank}(\boldsymbol{\Gamma}) = k - 1$ and consider the null hypothesis: $H_0 : \boldsymbol{\Gamma}\boldsymbol{\beta}_{yx.\mathcal{Z}} = \mathbf{0}$. Based on the joint asymptotic normality of $\hat{\boldsymbol{\beta}}_{yx.\mathcal{Z}}$ we now construct an asymptotically χ^2-distributed test statistic for this null hypothesis.

Definition 5 (Rank-r Moore-Penrose inverse)**.** Consider the spectral decomposition of an $l \times l$ positive semidefinite matrix $\boldsymbol{\Delta} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^\top$, where $\mathbf{P}$ is the orthonormal matrix of eigenvectors and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_l)$ with $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_l$ the ordered eigenvalues of $\boldsymbol{\Delta}$. The rank-r Moore-Penrose inverse of $\boldsymbol{\Delta}$ is the matrix $\boldsymbol{\Delta}_r^\dagger = \mathbf{P}\boldsymbol{\Lambda}_r^\dagger\mathbf{P}^\top$, where $r \le \mathrm{rank}(\boldsymbol{\Delta})$ and $\boldsymbol{\Lambda}_r^\dagger = \mathrm{diag}(1/\lambda_1, \ldots, 1/\lambda_r, 0, \ldots, 0)$.

Theorem 6. *Assume the same conditions as in Lemma 3. Let $\boldsymbol{\Sigma}_{\mathcal{Z}}$ be the covariance matrix of the limiting distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}}_{yx.\mathcal{Z}} - \boldsymbol{\beta}_{yx.\mathcal{Z}})$ and given a $(k-1)\times k$ contrast matrix $\boldsymbol{\Gamma}$, define $\boldsymbol{\Delta}_{\mathcal{Z}} = \boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\mathcal{Z}}\boldsymbol{\Gamma}^\top$. Suppose that $\hat{\boldsymbol{\Sigma}}_{\mathcal{Z}}$ is a consistent estimator of $\boldsymbol{\Sigma}_{\mathcal{Z}}$ and that $\hat{r}$ is a consistent estimator of $\mathrm{rank}(\boldsymbol{\Delta}_{\mathcal{Z}}) = r_0, 1 \le r_0 \le k - 1$. Let $\hat{\boldsymbol{\Delta}}_{\mathcal{Z},\hat{r}}^\dagger$ denote the rank-$\hat{r}$ Moore-Penrose inverse of $\hat{\boldsymbol{\Delta}}_{\mathcal{Z}} = \boldsymbol{\Gamma}\hat{\boldsymbol{\Sigma}}_{\mathcal{Z}}\boldsymbol{\Gamma}^\top$. Then under the null hypothesis $H_0 : \boldsymbol{\Gamma}\boldsymbol{\beta}_{yx.\mathcal{Z}} = 0$, the test statistic*

$$T_{\hat{r}}^2 = n(\boldsymbol{\Gamma}\hat{\boldsymbol{\beta}}_{yx.\mathcal{Z}})^\top \hat{\boldsymbol{\Delta}}_{\mathcal{Z},\hat{r}}^\dagger (\boldsymbol{\Gamma}\hat{\boldsymbol{\beta}}_{yx.\mathcal{Z}}) \tag{3}$$

converges in distribution to a $\chi_{r_0}^2$-distributed random variable.

We can can estimate $\boldsymbol{\Sigma}_{\mathcal{Z}}$ consistently by plugging in sample residuals for the population level residuals in equation (2). We refer to this estimator as the plug-in estimator and denote it $\hat{\boldsymbol{\Sigma}}_{\mathcal{Z}}$. For a detailed argument, see Lemma 12 in the supplementary materials. Note also that for simplicity, we only consider the contrast matrix $\boldsymbol{\Gamma}$ with 1 at the entries (j, j) and -1 at the entries $(j, j+1)$ for $j = 1, 2, \ldots, k-1$, and with zeroes at the remaining entries in this paper.

3.3 THE DEGREES OF FREEDOM

To compute the Moore-Penrose inverse and the degrees of freedom for the test statistic in equation (3), it is necessary to know the rank of $\mathbf{\Delta}_{\mathcal{Z}}$. There are two possible approaches to this problem. The first is to estimate the rank r_0 with some estimate $\hat{r}$. The second approach is to select the candidate sets $\mathcal{Z}$ in a way that ensures the matrix $\mathbf{\Delta}_{\mathcal{Z}}$ is invertible. We now develop tools for both approaches.

3.3.1 Estimating the Degrees of Freedom

A standard approach to estimating the rank of a matrix from a noisy observation is information criterion based model selection. This is equivalent to conducting sequential hypothesis tests [Camba-Méndez and Kapetanios, 2009] for the possible ranks. In order to apply such model selection to the rank estimation of $\mathbf{\Delta}_{\mathcal{Z}}$, we first derive that the half-vectorised plug-in estimator $\operatorname{vech}(\hat{\mathbf{\Delta}}_{\mathcal{Z}})$ based on the plug-in estimator $\hat{\mathbf{\Sigma}}_{\mathcal{Z}}$ is asymptotically normal.

Proposition 7. *Under the same conditions as in Lemma 3,* $\sqrt{n}\operatorname{vech}(\hat{\mathbf{\Delta}}_{\mathcal{Z}} - \mathbf{\Delta}_{\mathcal{Z}}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \mathbf{C})$*, where* $\mathbf{C} = \mathbf{\Pi}\mathbf{F}\mathbf{\Pi}^\top$ *for some positive semidefinite matrix* $\mathbf{F}$*, with scaling matrix* $\mathbf{\Pi} = \mathbf{E}_l(\mathbf{\Gamma} \otimes \mathbf{\Gamma})\mathbf{D}_k$*. Here,* $\mathbf{E}_l$ *is the* $l(l+1)/2 \times l^2$ *elimination matrix,* $l = k-1$ *and* $\mathbf{D}_k$ *is the* $k^2 \times k(k+1)/2$ *duplication matrix.*

Based on Proposition 7 and a consistent estimator $\hat{\mathbf{C}}$ of the matrix $\mathbf{C}$, we may construct a rank estimation procedure from the minimum discrepancy function (MDF) test statistic [Cragg and Donald, 1997, Donald et al., 2007], which has the form

$$n \min_{\operatorname{rank}(\tilde{\mathbf{\Delta}}_{\mathcal{Z}}) \leq r} \operatorname{vech}(\hat{\mathbf{\Delta}}_{\mathcal{Z}} - \tilde{\mathbf{\Delta}}_{\mathcal{Z}})^\top \hat{\mathbf{C}}^{-1} \operatorname{vech}(\hat{\mathbf{\Delta}}_{\mathcal{Z}} - \tilde{\mathbf{\Delta}}_{\mathcal{Z}}). \tag{4}$$

This procedure, however, has only been shown to be consistent if either $\mathbf{C}$ is of full rank [Cragg and Donald, 1997] or, in slightly adapted form, if the true rank of $\mathbf{C}$ is known [Ratsimalahelo, 2003].

As we cannot estimate the rank of $\mathbf{C}$ to estimate the rank of $\mathbf{\Delta}_{\mathcal{Z}}$, we instead propose using a simplified estimator based on the MDF statistic from equation (4) which we call the pseudo-MDF estimator:

$$\hat{r} = \operatorname*{argmin}_{r \in \{1,\ldots,k-1\}} \Big\{ n \| \operatorname{vech}(\hat{\mathbf{\Delta}}_{\mathcal{Z}} - \tilde{\mathbf{\Delta}}_{\mathcal{Z},r}) \|_2^2 + \log(n) r (k - 1 - (r-1)/2) \Big\}, \tag{5}$$

where $\tilde{\mathbf{\Delta}}_{\mathcal{Z},r}$ is the best rank-r reconstruction of $\hat{\mathbf{\Delta}}_{\mathcal{Z}}$ based on spectral decomposition such that $\tilde{\mathbf{\Delta}}_{\mathcal{Z},r}\tilde{\mathbf{\Delta}}_{\mathcal{Z},r}^\dagger = \mathbf{I}$. Note that we effectively assume that the matrix $\hat{\mathbf{C}}^{-1}$ is the identity matrix. In doing so, we ignore the covariance structure between the elements of $\hat{\mathbf{\Delta}}_{\mathcal{Z}}$. Although the elements of $\hat{\mathbf{\Delta}}_{\mathcal{Z}}$ are likely correlated, the pseudo-MDF rank estimate nonetheless works well empirically (see Section 4).

Algorithm 1 Testing procedure

Input: Candidate graph $\mathcal{G}$, vertices (X, Y), data $\mathcal{D}_n$, testing strategy $S \in \{\mathrm{All}, \mathrm{Min}+\}$
Output: p-value
if $S = \mathrm{All}$ **then**
    Set $\mathcal{Z}$ as the collection of all valid adjustment sets relative to (X, Y) in $\mathcal{G}$
if $S = \mathrm{Min}+$ **then**
    Set $\mathcal{Z}$ as a pruned collection of all minimal valid adjustment sets relative to (X, Y) in $\mathcal{G}$ plus the set of non-forbidden nodes
for each adjustment set $\mathbf{Z}$ in $\mathcal{Z}$ **do**
    Get sample regression residuals $\mathbf{r}_{x\mathbf{z}}$ and $\mathbf{r}_{y\mathbf{z}'}$ from data $\mathcal{D}_n$
Compute $\hat{\mathbf{\Sigma}}_{\mathcal{Z}}$ and $\hat{\mathbf{\Delta}}_{\mathcal{Z}}$ with regression residuals
if $S = \mathrm{All}$ **then**
    Estimate optimal rank $\hat{r}$ from $\hat{\mathbf{\Delta}}_{\mathcal{Z}}$ based on (5)
if $S = \mathrm{Min}+$ **then**
    Set $\hat{r}$ as the cardinality of $\mathcal{Z}$ minus one
Compute test statistic

$$T^2_{\mathrm{obs}} = n(\mathbf{\Gamma}\hat{\boldsymbol{\beta}}_{yx})^\top \hat{\mathbf{\Delta}}^\dagger_{\mathcal{Z},\hat{r}}(\mathbf{\Gamma}\hat{\boldsymbol{\beta}}_{yx})$$

Calculate p-value $= 1 - F(T^2_{\mathrm{obs}})$ where $F(\cdot)$ is the cumulative distribution function of $\chi^2_{\hat{r}}$

3.3.2 Selecting $\mathcal{Z}$ to Ensure Full Rank

Depending on the choice of candidate sets $\mathcal{Z}$, the asymptotic covariance matrix $\mathbf{\Sigma}_{\mathcal{Z}}$ may be of full rank. This is, for example, trivially true if there is only one set in $\mathcal{Z}$. Whenever $\mathbf{\Sigma}_{\mathcal{Z}}$ is invertible, the matrix $\mathbf{\Delta}_{\mathcal{Z}}$ is also invertible. We now propose a strategy to select $\mathcal{Z}$, such that $\mathbf{\Sigma}_{\mathcal{Z}}$ is likely to be of full rank and that also follows the guidelines derived from Theorem 1 in Section 3.1.

Lemma 8. *Consider nodes* X *and* Y *in a DAG* $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ *such that* $Y \in \mathrm{de}(X, \mathcal{G})$*. Consider a collection of*

$$\mathcal{Z} = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_k\} \cup \{\mathrm{nonforb}(X, Y, \mathcal{G})\}$$

where $\mathbf{Z}_i$*,* $i = 1, \ldots, k$*, are minimal valid adjustment sets relative to* (X, Y) *in* $\mathcal{G}$*. If* $\mathbf{Z}_i \setminus (\cup_{j \neq i}\mathbf{Z}_j) \neq \emptyset$ *for all* $i = 1, \ldots, k$*,* $\mathrm{nonforb}(X, Y, \mathcal{G}) \setminus (\cup_i \mathbf{Z}_i) \not\perp_{\mathcal{G}} X$ *and we sample the edge coefficients and error variances for a linear structural equation model compatible with* $\mathcal{G}$ *from a distribution* P*, such that* P *is absolutely continuous with respect to the Lebesgue measure, then the asymptotic covariance matrix* $\mathbf{\Sigma}_{\mathcal{Z}}$ *for the random vector* $\hat{\boldsymbol{\beta}}_{yx.\mathcal{Z}}$ *is* P*-a.s. of full rank.*

In general, the collection of all minimal valid adjustment sets will not fulfil the distinct node condition of Lemma 8. It is, however, easy to prune the set of all minimal valid adjustment sets to obtain a subset that fulfils the conditions of Lemma 8 and still covers the same set of nodes as the collection of all minimal valid adjustment sets.

3.4 THE TESTING PROCEDURE

We propose a testing procedure that, given a pair of nodes (X, Y) in a candidate graph $\mathcal{G}$ and a data set, tests whether adjusting for the valid adjustment sets relative to (X, Y) in $\mathcal{G}$ leads to estimators that converge to the same quantity. The procedure uses the test statistic from equation (3) and we propose two strategies to select the collection of valid adjustment sets $\mathcal{Z}$.

The first strategy, which we call All, considers all available valid adjustment sets. This strategy is likely to lead to the best power for the test, but it requires estimating the degrees of freedom for the test statistic's asymptotic distribution. This is a difficult problem (see Section 3.3.1) and as a result the solution we propose does not have a formal consistency guarantee, although it performs well in practice (see Section 4).

The second strategy, which we call $\mathrm{Min}+$, is to prune the collection of all minimal valid adjustment sets as explained in Section 3.3.2 and then add the set of all non-forbidden nodes, which under the assumption $Y \in \mathrm{de}(X, \mathcal{G})$ is always a valid adjustment set (see Lemma 17 in the supplementary materials). This approach avoids estimating the degrees of freedom but may lead to a loss of power. Note that if $Y \notin \mathrm{de}(X, \mathcal{G})$ we would need to replace the non-forbidden nodes with another large set such as $\mathrm{Adjust}(X, Y, \mathcal{G})$ [Perković et al., 2018]. However, if $Y \notin \mathrm{de}(X, \mathcal{G})$ every set that d-separates X and Y is a valid adjustment set and therefore our problem reduces to checking d-separation statements, for which there exists a wide literature on conditional independence tests [e.g. Spirtes et al., 2000]. Therefore, we disregard this case.

We summarise the testing procedure in Algorithm 1. As discussed in Section 3.1, the test cannot detect all types of mistakes in $\mathcal{G}$, but it nonetheless serves as a simple and targeted robustness check.

Example 9. To illustrate our testing procedure, we revisit the linear structural equation model from Example 2 as well as the true graph and candidate graphs shown in Figure 1. In addition, we also consider the candidate graph $\mathcal{G}_3'$ which is the graph $\mathcal{G}_3$ with the edge $V \to Y$ deleted. To each candidate graph we apply the testing procedure with both testing strategies (see Algorithm 1). Recall that for the candidate graphs $\mathcal{G}_0$ and $\mathcal{G}_3$ the null hypothesis is true, while it is false for the candidate graphs $\mathcal{G}_1$ and $\mathcal{G}_3'$. We sample 100 data sets with $n = 25$ observations, 100 sets with $n = 100$ observations, and another 100 sets with $n = 400$ observations from the underlying linear structural equation model and apply our testing procedure to each of these data sets. The resulting p-values are shown as probability-probability plots against the standard uniform distribution in Figure 2. When the null hypothesis is true, we see that both strategies lead to close to uniform p-values, especially when $n > 25$. We do not consider $\mathcal{G}_2$ as it is equivalent to $\mathcal{G}_0$ in terms of valid adjustment sets. We also observe reasonable power for $n > 25$ and $\mathcal{G}_1$, especially with strategy All. For $\mathcal{G}_3'$ on the other hand the power is mediocre except for All and $n = 400$.

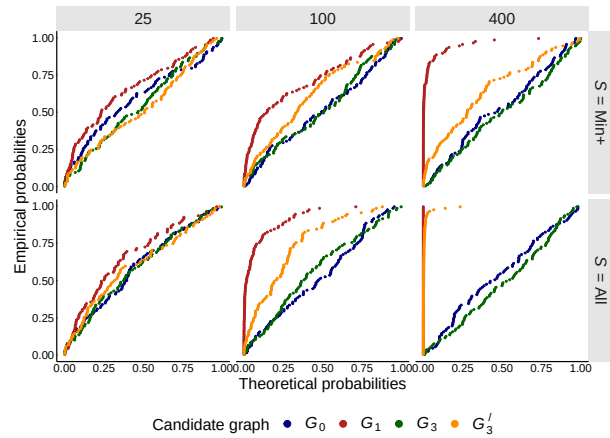

Figure 2: Probability-probability plots of the p-values in Example 9. Theoretical probabilities are from the cumulative distribution function of a standard uniform distribution. Rows: test strategies. Columns: sample sizes for the test.

4 SIMULATIONS

We investigate the finite sample performance of the testing procedure from Algorithm 1 in a simulation study. The study is structured as follows. We randomly generate 200 DAGs with the number of nodes and the expected neighbourhood size sampled uniformly from $\{10, 20\}$ and $\{2, 3, 4, 5\}$, respectively. For each DAG $\mathcal{G}_0$ we randomly generate a compatible linear structural equation model by (i) sampling the edge coefficients uniformly from the interval $[-2, -0.1] \cup [0, 1, 2]$, (ii) sampling the error distribution uniformly to either be normal, t, uniform or logistic for all errors and (iii) uniformly sampling scale parameters which depend on the error distribution such that the error variances are in the interval $[0.4, 1.6]$. We randomly choose a pair of nodes (X, Y) such that $Y \in \mathrm{de}(X, \mathcal{G}_0)$ and that there exist at least two valid adjustment sets relative to (X, Y) in $\mathcal{G}_0$.

For each true DAG $\mathcal{G}_0$ we then sample 20 data sets from the corresponding linear structural equation model. The sample size m for these data sets is sampled uniformly

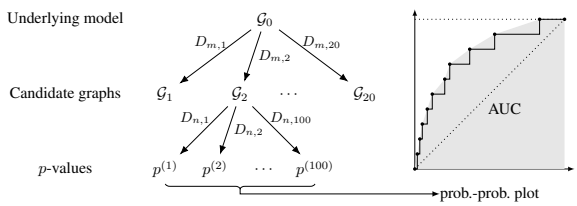

Figure 3: An illustration of the double simulation scheme for the simulation study and an illustration of the AUC metric.

from $\{100, 400\}$. With each of these data sets we estimate a causal graph $\mathcal{G}$ using either the Greedy Equivalence Search (GES) algorithm [Chickering, 2002], if the errors of the linear structural equation model are normal or the Linear Non-Gaussian Acyclic Models (LiNGAM) algorithm [Shimizu et al., 2006], otherwise. We do this in order to generate a large number of plausible candidate causal graphs for our testing procedure. We use the sample sizes 100 and 400 to ensure that some of the candidate graphs contain more errors and some fewer.

For each candidate graph $\mathcal{G}$ and each sample size $n = 50, 100, 200, 400$ we do the following procedure. We sample an additional 100 data sets with sample size n from the corresponding true linear structural equation model. Given these data sets, the pair (X, Y) and the candidate graph $\mathcal{G}$, we apply Algorithm 1 using both strategies. To measure the performance of our testing procedure we then compute the area under the curve (AUC) of the probability-probability plot of the 100 p-values against the uniform distribution. This means that in total we obtain 8 AUCs for each candidate graph, i.e., 4 for each testing strategy and 2 for each sample size. Figure 3 gives an illustration of the layered simulation scheme. Section C of the supplementary materials provides further details and analysis of the simulation study.

Figure 4 is an ensemble of violin plots of the AUCs from the simulation study. As we have access to the true graph, we can decide for each candidate graph and testing strategy, whether the null hypothesis for the test is true or false and plot these cases separately. Figure 4 shows that as the sample size of the data set used for our testing procedure increases, the AUCs in the cases where the null hypothesis is true become centred around 0.5. This indicates that both strategies control the type-I error rate asymptotically. There are, however, very large and small AUCs when we use the strategy $S = \text{All}$ with small sample sizes and for the more accurate candidate graphs. This is likely due to the rank estimation step required for this strategy and indicates that as expected the strategy Min+ is more stable than All.

In the cases where the null hypothesis is false, i.e., the candidate graph contains a mistake that the test can detect, the AUCs have a cluster close to 1 which is especially pronounced for the larger sample sizes. This indicates that our testing procedure has good power in many cases. Unsurprisingly, the AUCs are smaller for the candidate graphs with fewer errors, although they remain respectable and there continue to be AUCs close to 1. In general, the AUCs for strategy All are larger than those for Min+, although this gain is obtained at the price of a loss in stability.

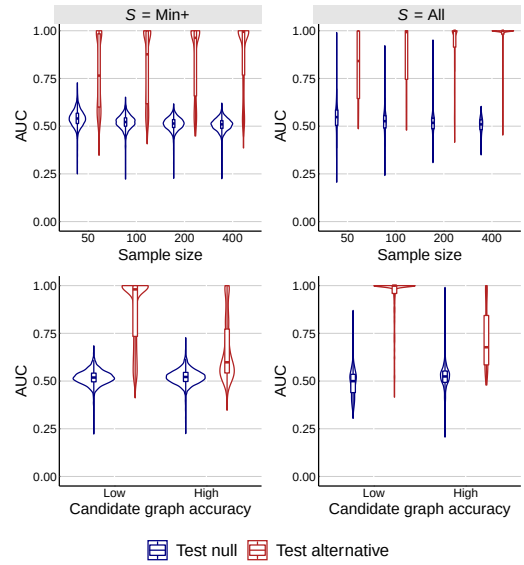

Figure 4: Violin plots (layered with boxplots) of the areas under the curve (AUC) from the simulation study. The AUCs are grouped by sample size for the testing procedure (first row) and the expected accuracy of the candidate graph (second row).

Cand. graph accuracy	n	Null $S = \text{Min+}$	Null $S = \text{All}$	Alternative $S = \text{Min+}$	Alternative $S = \text{All}$
Low	50	0.0854	0.0639	0.560	0.677
	100	0.0662	0.0511	0.656	0.801
	200	0.0574	0.0444	0.731	0.880
	400	0.0513	0.0401	0.801	0.929
High	50	0.0885	0.0844	0.182	0.200
	100	0.0658	0.0593	0.207	0.269
	200	0.0568	0.0519	0.273	0.366
	400	0.0533	0.0501	0.377	0.491

Table 1: Proportion of hypotheses rejected at level 0.05 in the simulation study.

The AUCs we consider do not capture the behaviour of our testing procedure fully, so we also calculate the proportion of tests rejected at level 0.05 among all tests performed in the simulation study as an additional metric. The results are given in Table 1. They indicate that for both strategies our testing procedure controls the type-I error rate asymptotically and at the same time has good power for the alternative.

5 REAL DATA EXAMPLE

We apply our testing procedure to the single cell data collected for the investigation of human primary naïve CD4$^+$

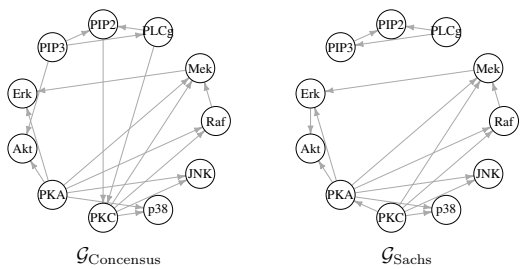

Figure 5: Causal DAGs representing intracellular signalling network among human primary naïve CD4$^+$ T cells.

T cell signalling networks by Sachs et al. [2005]. This data set consists of measurements from a total of 9 experimental conditions. We will only use the data from the observational regime, which corresponds to the experimental setting with reagent anti-CD3/CD28. This subset of the data consists of 853 measurements of 11 phosphorylated proteins and phospholipids. The observational data is thought to be consistent with the conventionally accepted molecular interaction network (also called the consensus graph, $\mathcal{G}_{\text{Consensus}}$, Figure 5 left). We use alternative graph proposed in Sachs et al. [2005] ($\mathcal{G}_{\text{Sachs}}$, Figure 5 right) to evaluate the results of the analysis.

We consider $\mathcal{G}_{\text{Consensus}}$ as the candidate graph $\mathcal{G}$. We extract all pairs of nodes (X, Y) that satisfy $Y \in \text{de}(X, \mathcal{G})$. There are 36 such node pairs in the graph. For every pair, we apply our testing procedure with strategy All to the log-transformed and centred observational data. After a Bonferroni adjustment only the p-values for the pairs (PKA, Erk) and (PKA, Akt) are significant at 0.05 level (1.19×10^{-14} and 4.91×10^{-14}).

We now take a closer look at these two node pairs. The collection $\mathcal{Z}$ of all valid adjustment sets relative to (PKA, Erk) in the consensus graph consists of 419 sets. The test rejects the null hypothesis that these adjustment sets lead to estimates of the same quantity. To illustrate a potential error in $\mathcal{G}_{\text{Concensus}}$, we consider the valid adjustment sets $\emptyset$ and {PLCg, PIP2, PIP3, Akt, PKC, p38, JNK}. If we consider the alternative graph $\mathcal{G}_{\text{Sachs}}$ as a more appropriate representation of the true data generating mechanism, the rejection is justifiable. The covariate Akt is a forbidden node in $\mathcal{G}_{\text{Sachs}}$ because it opens a collider path PKA $\rightarrow$ Akt $\leftarrow$ Erk. On the other hand, the empty set is a valid adjustment set also in $\mathcal{G}_{\text{Sachs}}$. A similar argument applies to the the pair (PKA, Akt). The collection $\mathcal{Z}$ relative to this node pair also has a size of 419, among which we can look at adjustment sets $\emptyset$ and {Raf, Mek, PLCg, PIP2, PIP3, Erk, PKC, p38, JNK}. Using Erk is problematic as it is a forbidden node in $\mathcal{G}_{\text{Sachs}}$ because it blocks the causal path PKA $\rightarrow$ Erk $\rightarrow$ Akt. This indicates that in both cases our testing procedure is detecting a mistake in the consensus graph.

Interestingly, our testing procedure produces rank estimates of $\mathbf{\Delta}_{\mathcal{Z}}$ mostly at 1 (30 out of 36 cases), even though the size of $\mathcal{Z}$ goes up to 419. This illustrates how a large number of adjustment sets does not necessarily mean a large number of effective over-identifying constraints on the total effect for the test. As a result, it is unsurprising that our testing procedure with the $\text{Min}+$ strategy detects the same two pairs of nodes as problematic (p-values 7.04×10^{-15} and 7.35×10^{-15}).

6 CONCLUSION AND DISCUSSION

In this paper, we propose a robustness test that checks whether it is reasonable to use a candidate causal graph to estimate a total effect of interest with covariate adjustment. This is a useful model validation tool for practitioners who wish to estimate a total effect with covariate adjustment and rely on causal graphs obtained from domain knowledge.

We develop our testing procedure assuming that the candidate graph is a DAG. In applications with unmeasured confounding between the covariates, it is more natural to assume that the candidate graph is an acyclic directed mixed graph (ADMG) with bi-directed edges that represent error correlations induced by the presence of unmeasured confounding. If the candidate ADMG contains at least two valid adjustment sets, it is also possible to apply our testing procedure in this setting with one limitation. The set $\text{nonforb}(X, Y, \mathcal{G})$ may not be a valid adjustment set and as a result the strategy $\text{Min}+$ fails. We believe it is possible to adapt $\text{Min}+$ to an ADMG by replacing $\text{nonforb}(X, Y, \mathcal{G})$ with a suitable alternative large valid adjustment set but we leave this for future research.

References

Donald WK Andrews. Asymptotic results for generalized wald tests. *Econometric Theory*, 3(3):348–358, 1987.

Kenneth A. Bollen. *Structural Equations with Latent Variables*. Wiley New York, 1989.

Kenneth A Bollen and Kwok-fai Ting. Confirmatory tetrad analysis. *Sociological methodology*, pages 147–175, 1993.

Andreas Buja, Lawrence Brown, Richard Berk, Edward George, Emil Pitkin, Mikhail Traskin, Kai Zhang, Linda Zhao, et al. Models as approximations I: Consequences illustrated with linear regression. *Statistical Science*, 34 (4):523–544, 2019.

Gonzalo Camba-Méndez and George Kapetanios. Statistical tests and estimators of the rank of a matrix and their applications in econometric modelling. *Econometric Reviews*, 28(6):581–611, 2009.

Bryant Chen and Judea Pearl. Exogeneity and robustness. Technical report, University of California, Los Angeles, 2015.

David Maxwell Chickering. Learning equivalence classes of Bayesian-network structures. *Journal of Machine Learning Research*, 2:445–498, 2002.

John G Cragg and Stephen G Donald. Inferring the rank of a matrix. *Journal of econometrics*, 76(1-2):223–250, 1997.

Desislava Dikova, Andrei Panibratov, and Anna Veselova. Investment motives, ownership advantages and institutional distance: An examination of russian cross-border acquisitions. *International Business Review*, 28(4):625–637, 2019.

Stephen G Donald, Natércia Fortuna, and Vladas Pipiras. On rank estimation in symmetric matrices: the case of indefinite matrix estimators. *Econometric Theory*, pages 1217–1232, 2007.

Claus Thorn Ekstrøm. *MESS: Miscellaneous Esoteric Statistical Scripts*, 2020. URL `https://CRAN.R-project.org/package=MESS`.

Kenneth A Frank, Spiro J Maroulis, Minh Q Duong, and Benjamin M Kelcey. What would it take to change an inference? using rubin's causal model to interpret the robustness of causal inferences. *Educational Evaluation and Policy Analysis*, 35(4):437–460, 2013.

Leonard Henckel, Emilija Perković, and Marloes H Maathuis. Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *arXiv preprint arXiv:1907.02435*, 2019.

Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H. Maathuis, and Peter Bühlmann. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26, 2012.

James B Kirby and Kenneth A Bollen. 10. using instrumental variable tests to evaluate model specification in latent variable structural equation models. *Sociological Methodology*, 39(1):327–355, 2009.

Manabu Kuroki and Masami Miyakawa. Covariate selection for estimating the causal effect of control plans by using causal diagrams. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 65(1):209–222, 2003.

Xun Lu and Halbert White. Robustness checks and robustness tests in applied economics. *Journal of econometrics*, 178:194–206, 2014.

Preetam Nandy, Marloes H Maathuis, and Thomas S Richardson. Estimating the effect of joint interventions from observational data in sparse high-dimensional settings. *Annals of Statistics*, 45(2):647–674, 2017.

Emily Oster. Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204, 2019.

Judea Pearl. *Causality*. Cambridge University Press, second edition, 2009.

Emilija Perković, Johannes Textor, Markus Kalisch, and Marloes H Maathuis. Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs. *Journal of Machine Learning Research*, 18(220):1–62, 2018.

Zaka Ratsimalahelo. Strongly consistent determination of the rank of matrix. Technical report, EERI Research Paper Series, 2003.

Andrea Rotnitzky and Ezequiel Smucler. Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *Journal of Machine Learning Research*, 21:188–1, 2020.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

Tobias Schlegel, Curdin Pfister, Dietmar Harhoff, and Uschi Backes-Gellner. Innovation effects of universities of applied sciences: an assessment of regional heterogeneity. *The Journal of Technology Transfer*, pages 1–56, 2021.

Shohei Shimizu. Lingam: Non-gaussian methods for estimating causal structures. *Behaviormetrika*, 41(1):65–98, 2014.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct): 2003–2030, 2006.

Ilya Shpitser, Tyler VanderWeele, and James Robins. On the validity of covariate adjustment for estimating causal effects. In *Proceedings of the Twenty-Sixth Annual Conference on Uncertainty in Artificial Intelligence (UAI-10)*, pages 527–536, Corvallis, Oregon, 2010. AUAI Press.

Peter Spirtes. Directed cyclic graphical representations of feedback models. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 491–498, San Francisco, USA, 1995. Morgan Kaufmann Publishers Inc.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, second edition, 2000.

Gilbert W Stewart. Perturbation theory for the singular value decomposition. Technical report, University of Maryland, 1998.

Stefan Walter and Henning Tiemeier. Variable selection: current practice in epidemiological studies. *European journal of epidemiology*, 24(12):733–736, 2009.

Yigezu A Yigezu and Tamer El-Shater. Socio-economic impacts of zero and reduced tillage in wheat fields of the moroccan drylands. *Agricultural Economics*, 2021.

Supplementary Materials

A GRAPHICAL PRELIMINARIES

Graphs A graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is a tuple of a node set $\mathbf{V}$ and an edge set $\mathbf{E}$. We consider simple directed graphs where there is at most one edge between any pair of vertices and the edges are of the form $\rightarrow$.

Walks, paths and cycles Two vertices are adjacent if there is an edge between them. A *walk* between X and Y is a sequence of vertices $(X, \ldots, Y)$ such that successive vertices are adjacent. A *path* between X and Y is a walk between X and Y where all vertices are distinct. A *directed path* from X to Y is a path between X and Y where all the edges point towards Y. A *cycle* is a path $(X, Z, \ldots, Y)$ plus an edge between Y and X. A *directed cycle* is a directed path $(X, Z, \ldots, Y)$ from X to Y plus an edge $Y \rightarrow X$. Given a path $p = (V_1, \ldots, V_k)$, let $p(V_i, V_j)$, $i < j$ denote the path segment from V_i to V_j and let $-p = (V_k, \ldots, V_1)$. Given two paths $p = (V_1, \ldots, V_k)$ and $q = (V_k, \ldots, V_q)$, let $p \oplus q = (V_1, \ldots, V_k, \ldots, V_q)$. We call any node V_i on a path $p = (V_1, \ldots, V_k)$ such that $V_{i-1} \rightarrow V_i \leftarrow V_{i+1}$ a collider on p and any node that is not a collider on p, a non-collider on p.

DAG A *directed acyclic graph* (DAG) is a directed graph without directed cycles.

Parents, children, ancestors and descendants If $X \rightarrow Y$, then X is a parent of Y and Y is a child of X. If there is a directed path from X to Y, then X is an ancestor of Y and Y is a descendant of X. Any node is an ancestor and a descendant of itself. For any node $X \in \mathbf{V}$, the sets of parents, children, ancestors and descendants of X in $\mathcal{G}$ are denoted by $\mathrm{pa}(X, \mathcal{G})$, $\mathrm{ch}(X, \mathrm{G})$, $\mathrm{an}(X, \mathrm{G})$ and $\mathrm{de}(X, \mathrm{G})$, respectively. This definition applies disjunctively to sets of nodes. For example, the parents of the set of vertices $\mathbf{X}$ are defined as $\mathrm{pa}(\mathbf{X}, \mathcal{G}) = \cup_{X \in \mathbf{X}} \mathrm{pa}(X, \mathcal{G})$. The non-descendants of $\mathbf{X}$ are $\mathrm{nonde}(\mathbf{X}, \mathcal{G}) = \mathbf{V} \setminus \mathrm{de}(\mathbf{X}, \mathcal{G})$.

***d*-separation** A path p between X and Y is blocked by a set $\mathbf{Z}$ if at least one of the following conditions holds:

(i) There is a non-collider on p that is in $\mathbf{Z}$;

(ii) There is a collider on p such that neither itself nor any other of its descendants are in $\mathbf{Z}$.

A path that is not blocked is said to be open. If all paths between $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$ are blocked by $\mathbf{Z}$, then $\mathbf{X}$ and $\mathbf{Y}$ are *d-separated* by $\mathbf{Z}$, denoted by $\mathbf{X} \perp_{\mathcal{G}} \mathbf{Y} \mid \mathbf{Z}$. Otherwise, they are said to be *d-connected* by $\mathbf{Z}$.

Faithfulness Consider a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ such that $\mathbf{V}$ follows a linear structural equation model compatible with $\mathcal{G}$. If for all disjoint subsets $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$ of $\mathbf{V}$ such that $\mathbf{X}$ is independent of $\mathbf{Y}$ given $\mathbf{Z}$, $\mathbf{X} \perp_{\mathcal{G}} \mathbf{Y} \mid \mathbf{Z}$ then we say that the distribution of $\mathbf{V}$ is faithful to $\mathcal{G}$.

B PROOFS

B.1 PROOF OF THEOREM 1

Proof. Consider the set $\mathbf{Z}_i$. By assumption, $\mathrm{forb}(X, Y, \mathcal{G}) \cap \mathbf{Z}_i = \emptyset$ and, nonetheless, $\mathbf{Z}_i$ is not a valid adjustment set. Thus, there must exist a non-causal path p from X to Y in $\mathcal{G}$ that is open given $\mathbf{Z}_i$. Clearly, p must contain at least one non-collider. If every non-collider is in $\mathrm{forb}(X, Y, \mathcal{G})$, it follows that every node on p is in $\mathrm{forb}(X, Y, \mathcal{G})$. But this contradicts our assumption that p is non-causal and open given $\mathbf{Z}_i$. We can therefore conclude that p contains at least one non-collider that is not in $\mathrm{forb}(X, Y, \mathcal{G})$. But as $(\mathbf{V} \setminus \mathrm{forb}(X, Y, \mathcal{G})) \subseteq \cup_{j=1}^{k} \mathbf{Z}_j$ by assumption, p must be blocked by some set $\mathbf{Z}_j$.

Consider the potential colliders $C_1, \ldots, C_m$ on p. As p is open given $\mathbf{Z}_i$ for each collider C_k, there must exist a causal path q_k to some node in $\mathbf{Z}_i$, where we choose q_k to be the shortest possible such path. If any of the q_k intersects, drop the longer of the two paths. If any q_k contains X, replace p with $-q_k(X, C_k) \oplus p(C_k, Y)$ and repeat our argument. Consider now the following linear structural equation: set all edge coefficients not on p or our list of paths $q_1, \ldots, q_{m'}$ to 0. The resulting model is clearly compatible with $\mathcal{G}$ but also to a pruned graph $\mathcal{G}'$ where we drop all edges with edge coefficient 0. Clearly, in $\mathcal{G}'$ the path p is still open given $\mathbf{Z}_i$ and closed given $\mathbf{Z}_j$. Furthermore, p is the only path from X to Y in $\mathcal{G}'$, and as a result, we can conclude that $\beta_{yx.\mathbf{z}_i} \neq 0$ and $\beta_{yx.\mathbf{z}_j} = 0$.

We have shown that there exists a linear structural equation model compatible with $\mathcal{G}$, such that $\beta_{yx.\mathbf{z}_i} \neq \beta_{yx.\mathbf{z}_j}$. By rewriting the ordinary least squares coefficients in this equation as functions of the covariances from the underlying linear causal model and multiplying out the denominators we can reformulate the equation $\beta_{yx.\mathbf{z}_i} - \beta_{yx.\mathbf{z}_j}$ as a polynomial in the covariances. This new equation has the same zero set as the original function and is therefore non-trivial. As the covariance matrix is an analytic function in the edge coefficients and error variances of the linear structural equation model the zero set of this reformulated equation has Lebesgue measure 0 and our claim follows. □

B.2 PROOF OF LEMMA 3

Lemma 10 (Orthogonality between covariates and regression residual, [Buja et al., 2019])**.** *In a least squares regression of X on $\mathbf{Z}$, the minimiser of the optimisation problem $\min_{\boldsymbol{\beta}} \mathrm{E}(X - \mathbf{Z}^\top \boldsymbol{\beta})^2$ is the population regression coefficient $\boldsymbol{\beta}_{x\mathbf{z}} = \boldsymbol{\Sigma}_{\mathbf{zz}}^{-1} \boldsymbol{\Sigma}_{\mathbf{z}x}$. The residual $\delta_{x\mathbf{z}} = X - \mathbf{Z}^\top \boldsymbol{\beta}_{x\mathbf{z}}$ is orthogonal to $\mathbf{Z}$, i.e., $\mathrm{E}(\mathbf{Z}\delta_{x\mathbf{z}}) = \mathbf{0}$.*

Unless specified otherwise, serif letters denote random samples for scalar random variables. For example, $\mathsf{X} = (X_1, X_2, \ldots, X_n)^\top$ is an n-dimensional vector containing n i.i.d. copies of X. Bold serif letters denote random samples for vector random variables. For example, $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)^\top$ is an $n \times p$ matrix where each row is i.i.d. as $\mathbf{Z} \in \mathbb{R}^p$.

Lemma 11 (Regression error representation of OLS coefficients, [Buja et al., 2019])**.** *The difference between sample and population regression coefficient of X from regressing Y on $\mathbf{Z}' = (X, \mathbf{Z}^\top)^\top$ is*

$$\hat{\beta}_{yx.\mathbf{z}} - \beta_{yx.\mathbf{z}} = \frac{\langle \mathbf{r}_{x\mathbf{z}}, \boldsymbol{\delta}_{y\mathbf{z}'} \rangle}{\|\mathbf{r}_{x\mathbf{z}}\|^2},$$

where $\mathbf{r}_{x\mathbf{z}} = \mathsf{X} - \mathbf{Z}\hat{\boldsymbol{\beta}}_{x\mathbf{z}}$ is the vector of sample residuals from regressing X on $\mathbf{Z}$.

Proof of Lemma 3. The proof is inspired by the results in Appendix E.5 of Buja et al. [2019]. We first observe from Lemma 11 that for every set $\mathbf{Z}_i$, $i = 1, 2, \ldots, k$,

$$n^{1/2}(\hat{\beta}_{yx.\mathbf{z}_i} - \beta_{yx.\mathbf{z}_i}) = \frac{n^{-1/2}\langle \mathbf{r}_{x\mathbf{z}_i}, \boldsymbol{\delta}_{y\mathbf{z}_i'} \rangle}{n^{-1}\|\mathbf{r}_{x\mathbf{z}_i}\|^2}, \tag{6}$$

where $\mathbf{r}_{x\mathbf{z}_i} = \mathsf{X} - \mathbf{Z}_i\hat{\boldsymbol{\beta}}_{x\mathbf{z}_i}$ is the sample residuals from regressing X on $\mathbf{Z}_i$.

Numerator of (6).

$$\begin{aligned} n^{-1/2}\langle \mathbf{r}_{x\mathbf{z}_i}, \boldsymbol{\delta}_{y\mathbf{z}_i'} \rangle &= n^{-1/2}\langle \mathsf{X} - \mathbf{Z}_i\hat{\boldsymbol{\beta}}_{x\mathbf{z}_i}, \boldsymbol{\delta}_{y\mathbf{z}_i'} \rangle \\ &= n^{-1/2}\langle \boldsymbol{\delta}_{x\mathbf{z}_i} - \mathbf{Z}_i(\hat{\boldsymbol{\beta}}_{x\mathbf{z}_i} - \boldsymbol{\beta}_{x\mathbf{z}_i}), \boldsymbol{\delta}_{y\mathbf{z}_i'} \rangle \\ &= n^{-1/2}\langle \boldsymbol{\delta}_{x\mathbf{z}_i}, \boldsymbol{\delta}_{y\mathbf{z}_i'} \rangle - n^{-1/2}\langle \mathbf{Z}_i(\hat{\boldsymbol{\beta}}_{x\mathbf{z}_i} - \boldsymbol{\beta}_{x\mathbf{z}_i}), \boldsymbol{\delta}_{y\mathbf{z}_i'} \rangle. \end{aligned}$$

For the second term on the last line it holds that

$$\begin{aligned} n^{-1/2}\langle \mathbf{Z}_i(\hat{\boldsymbol{\beta}}_{x\mathbf{z}_i} - \boldsymbol{\beta}_{x\mathbf{z}_i}), \boldsymbol{\delta}_{y\mathbf{z}_i'} \rangle &= \left(n^{-1}\boldsymbol{\delta}_{y\mathbf{z}_i'}^\top \mathbf{Z}_i\right) \cdot n^{1/2}(\hat{\boldsymbol{\beta}}_{x\mathbf{z}_i} - \boldsymbol{\beta}_{x\mathbf{z}_i}) \\ &= o_p(1) \cdot O_p(1) = o_p(1), \end{aligned}$$

since $\mathrm{E}(\delta_{y\mathbf{z}_i'}\mathbf{Z}_i) = \mathbf{0}$ by Lemma 10 and $n^{1/2}(\hat{\boldsymbol{\beta}}_{x\mathbf{z}_i} - \boldsymbol{\beta}_{x\mathbf{z}_i})$ converges in distribution to a multivariate normal random variable by the central limit theorem, which is appropriate since by assumption, the fourth moments of $\mathbf{V}$ are finite.

Denominator of (6).

Using the convention that the hat matrix $\mathbf{H}_n = \mathbf{Z}_i(\mathbf{Z}_i^\top \mathbf{Z}_i)^{-1}\mathbf{Z}_i^\top$, the average squared sample residuals

$$\begin{aligned} n^{-1}\|\mathbf{r}_{x\mathbf{z}_i}\|^2 &= n^{-1}\mathsf{X}^\top(\mathbf{I} - \mathbf{H}_n)\mathsf{X} \\ &= n^{-1}\|\mathsf{X}\|^2 - \left(n^{-1}\mathsf{X}^\top \mathbf{Z}_i\right)\left(n^{-1}\mathbf{Z}_i^\top \mathbf{Z}_i\right)^{-1}\left(n^{-1}\mathbf{Z}_i^\top \mathsf{X}\right) \\ &\xrightarrow{p} \mathrm{E}(X^2) - \mathrm{E}(X\mathbf{Z}_i^\top)[\mathrm{E}(\mathbf{Z}_i\mathbf{Z}_i^\top)]^{-1}\,\mathrm{E}(\mathbf{Z}_i X) \\ &= \mathrm{E}(X^2) - \mathrm{E}(X\mathbf{Z}_i^\top \boldsymbol{\beta}_{x\mathbf{z}_i}) \\ &= \mathrm{E}(X - \mathbf{Z}_i^\top \boldsymbol{\beta}_{x\mathbf{z}_i})^2 = \mathrm{E}(\delta_{x\mathbf{z}_i}^2). \end{aligned}$$

The second to last step follows because $\mathrm{E}[\mathbf{Z}_i(X - \mathbf{Z}_i^\top \boldsymbol{\beta}_{x\mathbf{z}_i})] = \mathrm{E}(\mathbf{Z}_i \delta_{x\mathbf{z}_i}) = \mathbf{0}$ by Lemma 10.

We are now ready to present the asymptotic joint normality of $\hat{\boldsymbol{\beta}}_{yx.\mathcal{Z}}$. Since $\mathrm{E}(\delta_{x\mathbf{z}_i}\delta_{y\mathbf{z}_i}) = \mathrm{E}[(X - \mathbf{Z}_i^\top \boldsymbol{\beta}_{x\mathbf{z}_i})\delta_{y\mathbf{z}'_i}] = 0$, together with the fact that the fourth moments of $\mathbf{V}$ are finite, we can apply the multivariate central limit theorem to conclude that

$$\left(n^{-1/2}\langle \boldsymbol{\delta}_{x\mathbf{z}_1}, \boldsymbol{\delta}_{y\mathbf{z}'_1}\rangle, \ldots, n^{-1/2}\langle \boldsymbol{\delta}_{x\mathbf{z}_k}, \boldsymbol{\delta}_{y\mathbf{z}'_k}\rangle\right) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{\Psi})$$

where the entries have the form $\boldsymbol{\Psi}_{ij} = \mathrm{E}(\delta_{x\mathbf{z}_i}\delta_{y\mathbf{z}'_i}\delta_{x\mathbf{z}_j}\delta_{y\mathbf{z}'_j})$ for all $1 \le i, j \le k$. Therefore, the random vector

$$\begin{pmatrix} n^{-1/2}\langle \mathbf{r}_{x\mathbf{z}_1}, \boldsymbol{\delta}_{y\mathbf{z}'_1}\rangle \\ n^{-1/2}\langle \mathbf{r}_{x\mathbf{z}_2}, \boldsymbol{\delta}_{y\mathbf{z}'_2}\rangle \\ \vdots \\ n^{-1/2}\langle \mathbf{r}_{x\mathbf{z}_k}, \boldsymbol{\delta}_{y\mathbf{z}'_k}\rangle \end{pmatrix} = \begin{pmatrix} n^{-1/2}\langle \boldsymbol{\delta}_{x\mathbf{z}_1}, \boldsymbol{\delta}_{y\mathbf{z}'_1}\rangle \\ n^{-1/2}\langle \boldsymbol{\delta}_{x\mathbf{z}_2}, \boldsymbol{\delta}_{y\mathbf{z}'_2}\rangle \\ \vdots \\ n^{-1/2}\langle \boldsymbol{\delta}_{x\mathbf{z}_k}, \boldsymbol{\delta}_{y\mathbf{z}'_k}\rangle \end{pmatrix} - \begin{pmatrix} n^{-1/2}\langle \mathbf{Z}_1(\hat{\boldsymbol{\beta}}_{x\mathbf{z}_1} - \boldsymbol{\beta}_{x\mathbf{z}_1}), \boldsymbol{\delta}_{y\mathbf{z}'_1}\rangle \\ n^{-1/2}\langle \mathbf{Z}_2(\hat{\boldsymbol{\beta}}_{x\mathbf{z}_2} - \boldsymbol{\beta}_{x\mathbf{z}_2}), \boldsymbol{\delta}_{y\mathbf{z}'_2}\rangle \\ \vdots \\ n^{-1/2}\langle \mathbf{Z}_k(\hat{\boldsymbol{\beta}}_{x\mathbf{z}_k} - \boldsymbol{\beta}_{x\mathbf{z}_k}), \boldsymbol{\delta}_{y\mathbf{z}'_k}\rangle \end{pmatrix}$$
$$\xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{\Psi}),$$

due to the fact that the second vector converge in distribution to a vector of zeroes. Based on the discussion of the denominator term, we can conclude that

$$n^{-1}\mathrm{diag}(\|\mathbf{r}_{x\mathbf{z}_1}\|^2, \|\mathbf{r}_{x\mathbf{z}_2}\|^2, \ldots, \|\mathbf{r}_{x\mathbf{z}_k}\|^2) \xrightarrow{p} \mathrm{diag}(\mathrm{E}(\delta^2_{x\mathbf{z}_1}), \mathrm{E}(\delta^2_{x\mathbf{z}_2}), \ldots, \mathrm{E}(\delta^2_{x\mathbf{z}_k})) = \boldsymbol{\Upsilon}.$$

The target quantity can then be written as

$$n^{1/2}(\hat{\boldsymbol{\beta}}_{yx.\mathcal{Z}} - \boldsymbol{\beta}_{yx.\mathcal{Z}}) = \begin{pmatrix} n^{-1}\|\mathbf{r}_{x\mathbf{z}_1}\|^2 & 0 & \cdots & 0 \\ 0 & n^{-1}\|\mathbf{r}_{x\mathbf{z}_2}\|^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & n^{-1}\|\mathbf{r}_{x\mathbf{z}_k}\|^2 \end{pmatrix}^{-1} \begin{pmatrix} n^{-1/2}\langle \mathbf{r}_{x\mathbf{z}_1}, \boldsymbol{\delta}_{y\mathbf{z}'_1}\rangle \\ n^{-1/2}\langle \mathbf{r}_{x\mathbf{z}_2}, \boldsymbol{\delta}_{y\mathbf{z}'_2}\rangle \\ \vdots \\ n^{-1/2}\langle \mathbf{r}_{x\mathbf{z}_k}, \boldsymbol{\delta}_{y\mathbf{z}'_k}\rangle \end{pmatrix}$$
$$\xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathcal{Z}}),$$

where the convergence follows from Slutsky's Theorem, and the asymptotic covariance matrix $\boldsymbol{\Sigma}_{\mathcal{Z}} = \boldsymbol{\Upsilon}^{-1}\boldsymbol{\Psi}\boldsymbol{\Upsilon}^{-1}$ is as specified in the theorem statement. □

Remark. If $\mathbf{Z}_1, \ldots, \mathbf{Z}_k$ are all valid adjustment sets relative to (X, Y) in $\mathcal{G}$ for a linear structural equation model compatible with a DAG $\mathcal{G}$, we can simplify the diagonal terms $\boldsymbol{\Delta}_{\mathcal{Z},ii} = \mathrm{E}(\delta^2_{x\mathbf{z}_i}\delta^2_{y\mathbf{z}'_i}) = \mathrm{E}(\delta^2_{x\mathbf{z}_i})\,\mathrm{E}(\delta^2_{y\mathbf{z}'_i})$ due to the independence between $\delta_{x\mathbf{z}_i}$ and $\delta_{y\mathbf{z}'_i}$ (see proof of Proposition 3.1 in Supplement from Henckel et al. [2019]). Therefore, the corresponding terms are $\boldsymbol{\Sigma}_{\mathcal{Z},ii} = \mathrm{E}(\delta^2_{y\mathbf{z}'_i})/\,\mathrm{E}(\delta^2_{x\mathbf{z}_i})$. It can also be shown that $\hat{\beta}_{yx.\mathbf{z}}$ is root-n consistent for the total effect τ_{yx} for any valid adjustment set $\mathbf{Z}$ [Nandy et al., 2017]. In this case, in order to apply the central limit theorem separately on every entry of $(n^{-1/2}\langle \boldsymbol{\delta}_{x\mathbf{z}_1}, \boldsymbol{\delta}_{y\mathbf{z}'_1}\rangle, \ldots, n^{-1/2}\langle \boldsymbol{\delta}_{x\mathbf{z}_k}, \boldsymbol{\delta}_{y\mathbf{z}'_k}\rangle)^\top$, we only need the finite variance assumption for the error terms $\boldsymbol{\epsilon}$ of the linear structural equation model. In such a model, both $\delta_{x\mathbf{z}_i}$ and $\delta_{y\mathbf{z}'_i}$ can be expressed as linear functions of the error terms, say $\boldsymbol{\theta}_i^\top \boldsymbol{\epsilon}$ and $\boldsymbol{\xi}_i^\top \boldsymbol{\epsilon}$. Furthermore,

$$\begin{aligned} \mathrm{Var}(\delta_{x\mathbf{z}_i}\delta_{y\mathbf{z}'_i}) &= \mathrm{E}(\delta^2_{x\mathbf{z}_i}\delta^2_{y\mathbf{z}'_i}) - [\underbrace{\mathrm{E}(\delta_{x\mathbf{z}_i}\delta_{y\mathbf{z}'_i})}_{=0}]^2 \\ &= \mathrm{E}(\delta^2_{x\mathbf{z}_i})\,\mathrm{E}(\delta^2_{y\mathbf{z}'_i}) = \mathrm{E}(\boldsymbol{\theta}_i^\top \boldsymbol{\epsilon})^2\,\mathrm{E}(\boldsymbol{\xi}_i^\top \boldsymbol{\epsilon})^2, \end{aligned} \tag{7}$$

due to the independence between $\delta_{x\mathbf{z}_i}$ and $\delta_{y\mathbf{z}'_i}$. The order of each ϵ_{v_i} term cannot be larger than 2 in (7) for all $V_i \in \mathbf{V}$. Therefore, $\mathrm{Var}(\delta_{x\mathbf{z}_i}\delta'_{y\mathbf{z}_i})$ is finite for all $\mathbf{Z}_i$ whenever $\mathrm{E}(\epsilon^2_{v_i}) < \infty$ for all $V_i \in \mathbf{V}$.

We also show that a consistent estimator of the covariance matrix can be obtained by plugging in the sample residuals.

Lemma 12 (Consistency of plug-in estimator of $\boldsymbol{\Sigma}_{\mathcal{Z}}$)**.** *Consider the setting in Lemma 3. The plug-in estimator $\hat{\boldsymbol{\Sigma}}_{\mathcal{Z}}$ of $\boldsymbol{\Sigma}_{\mathcal{Z}}$ with entries*

$$\hat{\boldsymbol{\Sigma}}_{\mathcal{Z},ij} = \frac{n\sum_{s=1}^n r_{x\mathbf{z}_i,s} \cdot r_{y\mathbf{z}'_i,s} \cdot r_{x\mathbf{z}_j,s} \cdot r_{y\mathbf{z}'_j,s}}{\|\mathbf{r}_{x\mathbf{z}_i}\|^2\|\mathbf{r}_{x\mathbf{z}_j}\|^2},$$

for all $1 \le i, j \le k$, is consistent.

Proof of Lemma 12. Consider

$$\hat{\mathbf{\Sigma}}_{\mathcal{Z},ij} = \frac{n^{-1}\sum_{s=1}^{n} r_{x\mathbf{z}_i,s} \cdot r_{y\mathbf{z}'_i,s} \cdot r_{x\mathbf{z}_j,s} \cdot r_{y\mathbf{z}'_j,s}}{n^{-1}\|\mathbf{r}_{x\mathbf{z}_i}\|^2 n^{-1}\|\mathbf{r}_{x\mathbf{z}_j}\|^2}.$$

The denominator converges in probability to $\mathrm{E}(\delta^2_{x\mathbf{z}_i})\,\mathrm{E}(\delta^2_{x\mathbf{z}_j})$ by the proof of Lemma 3. The numerator can be written as

$$\begin{aligned}
n^{-1}\sum_{s=1}^{n} r_{x\mathbf{z}_i,s} r_{y\mathbf{z}'_i,s} r_{x\mathbf{z}_j,s} r_{y\mathbf{z}'_j,s} &= n^{-1}\sum_{s=1}^{n}\Big[(\delta_{x\mathbf{z}_i,s} - \mathbf{Z}^\top_{i,s}(\hat{\boldsymbol{\beta}}_{x\mathbf{z}_i} - \boldsymbol{\beta}_{x\mathbf{z}_i}))(\delta_{y\mathbf{z}'_i,s} - \mathbf{Z}'^\top_{i,s}(\hat{\boldsymbol{\beta}}_{y\mathbf{z}'_i} - \boldsymbol{\beta}_{y\mathbf{z}'_i})) \\
&\qquad (\delta_{x\mathbf{z}_j,s} - \mathbf{Z}^\top_{j,s}(\hat{\boldsymbol{\beta}}_{x\mathbf{z}_j} - \boldsymbol{\beta}_{x\mathbf{z}_j}))(\delta_{y\mathbf{z}'_j,s} - \mathbf{Z}'^\top_{j,s}(\hat{\boldsymbol{\beta}}_{y\mathbf{z}'_j} - \boldsymbol{\beta}_{y\mathbf{z}'_j}))\Big] \\
&= n^{-1}\sum_{s=1}^{n} \delta_{x\mathbf{z}_i,s}\delta_{y\mathbf{z}'_i,s}\delta_{x\mathbf{z}_j,s}\delta_{y\mathbf{z}'_j,s} + R,
\end{aligned}$$

where the remainder term R contains the rest of the products from the expansion: 1 product with no δ-term, 4 products with 1 δ-term, 6 products with 2 δ-terms and 4 products with 3 δ-terms. Below we will show that the remainder term $R \xrightarrow{p} 0$, and it follows that the numerator converges in probability to $\mathrm{E}(\delta_{x\mathbf{z}_i}\delta_{y\mathbf{z}'_i}\delta_{x\mathbf{z}_j}\delta_{y\mathbf{z}'_j})$. By the continuous mapping theorem, $\hat{\mathbf{\Sigma}}_{\mathcal{Z},ij} \xrightarrow{p} \mathbf{\Sigma}_{\mathcal{Z},ij}$ follows.

We will discuss one case from each category, as the results can be shown similarly for other products in the same category. The use of parentheses in the subscript denotes a particular entry of a vector. For example, $Z_{i(t),s}$ is the t-th entry of the s-th observation $\mathbf{Z}_{i,s}$ and $\hat{\beta}_{x\mathbf{z}_i(t)}$ is the t-th entry of the vector $\hat{\boldsymbol{\beta}}_{x\mathbf{z}_i}$. With the finite fourth moment assumption on $\mathbf{V}$, $\hat{\beta}_{x\mathbf{z}_i(t)} \xrightarrow{p} \beta_{x\mathbf{z}_i(t)}$ and $\hat{\beta}_{y\mathbf{z}'_i(t)} \xrightarrow{p} \beta_{y\mathbf{z}'_i(t)}$ for any $\mathbf{Z}_i$ and $1 \le t \le |\mathbf{Z}_i|$.

No δ-term.

$$\begin{aligned}
& n^{-1}\sum_{s=1}^{n} \mathbf{Z}^\top_{i,s}(\hat{\boldsymbol{\beta}}_{x\mathbf{z}_i} - \boldsymbol{\beta}_{x\mathbf{z}_i})\mathbf{Z}'^\top_{i,s}(\hat{\boldsymbol{\beta}}_{y\mathbf{z}'_i} - \boldsymbol{\beta}_{y\mathbf{z}'_i})\mathbf{Z}^\top_{j,s}(\hat{\boldsymbol{\beta}}_{x\mathbf{z}_j} - \boldsymbol{\beta}_{x\mathbf{z}_j})\mathbf{Z}'^\top_{j,s}(\hat{\boldsymbol{\beta}}_{y\mathbf{z}'_j} - \boldsymbol{\beta}_{y\mathbf{z}'_j}) \\
= & \sum_{t,u,v,w}\left(n^{-1}\sum_{s=1}^{n} Z_{i(t),s}Z'_{i(u),s}Z_{j(v),s}Z'_{j(w),s}\right)(\hat{\beta}_{x\mathbf{z}_i(t)} - \beta_{x\mathbf{z}_i(t)})(\hat{\beta}_{y\mathbf{z}'_i(u)} - \beta_{y\mathbf{z}'_i(u)}) \\
& \qquad (\hat{\beta}_{x\mathbf{z}_j(v)} - \beta_{x\mathbf{z}_j(v)})(\hat{\beta}_{y\mathbf{z}'_j(w)} - \beta_{y\mathbf{z}'_j(w)}) \\
\xrightarrow{p} & \sum_{t,u,v,w} \mathrm{const} \cdot 0 \cdot 0 \cdot 0 \cdot 0 \\
= & \, 0,
\end{aligned}$$

where $1 \le t \le |\mathbf{Z}_i|, 1 \le u \le |\mathbf{Z}'_i|, 1 \le v \le |\mathbf{Z}_j|, 1 \le w \le |\mathbf{Z}'_j|$, the constant term $\mathrm{E}(Z_{i(t)}Z'_{i(u)}Z_{j(v)}Z'_{j(w)})$ exists due to the finite fourth moment assumption on $\mathbf{V}$.

One δ-term.

$$\begin{aligned}
& -n^{-1}\sum_{s=1}^{n} \delta_{x\mathbf{z}_i,s}\mathbf{Z}'^\top_{i,s}(\hat{\boldsymbol{\beta}}_{y\mathbf{z}'_i} - \boldsymbol{\beta}_{y\mathbf{z}'_i}))\mathbf{Z}^\top_{j,s}(\hat{\boldsymbol{\beta}}_{x\mathbf{z}_j} - \boldsymbol{\beta}_{x\mathbf{z}_j})\mathbf{Z}'^\top_{j,s}(\hat{\boldsymbol{\beta}}_{y\mathbf{z}'_j} - \boldsymbol{\beta}_{y\mathbf{z}'_j}) \\
= & \sum_{u,v,w}\left(n^{-1}\sum_{s=1}^{n} \delta_{x\mathbf{z}_i,s}Z'_{i(u),s}Z_{j(v),s}Z'_{j(w),s}\right)(\hat{\beta}_{y\mathbf{z}'_i(u)} - \beta_{y\mathbf{z}'_i(u)})(\hat{\beta}_{x\mathbf{z}_j(v)} - \beta_{x\mathbf{z}_j(v)})(\hat{\beta}_{y\mathbf{z}'_j(w)} - \beta_{y\mathbf{z}'_j(w)}) \\
\xrightarrow{p} & \sum_{u,v,w} \mathrm{const} \cdot 0 \cdot 0 \cdot 0 \\
= & \, 0,
\end{aligned}$$

where $1 \le u \le |\mathbf{Z}'_i|, 1 \le v \le |\mathbf{Z}_j|, 1 \le w \le |\mathbf{Z}'_j|$, the constant term $\mathrm{E}(\delta_{x\mathbf{z}_i}Z'_{i(u)}Z_{j(v)}Z'_{j(w)})$ exists due to the finite fourth moment assumption on $\mathbf{V}$.

Two δ-terms.

$$
\begin{aligned}
& n^{-1}\sum_{s=1}^{n}\delta_{x\mathbf{z}_i,s}\delta_{y\mathbf{z}'_i,s}\mathbf{Z}_{j,s}^{\top}(\hat{\boldsymbol{\beta}}_{x\mathbf{z}_j}-\boldsymbol{\beta}_{x\mathbf{z}_j})\mathbf{Z}_{j,s}'^{\top}(\hat{\boldsymbol{\beta}}_{y\mathbf{z}'_j}-\boldsymbol{\beta}_{y\mathbf{z}'_j}) \\
&= \sum_{v,w}\left(n^{-1}\sum_{s=1}^{n}\delta_{x\mathbf{z}_i,s}\delta_{y\mathbf{z}'_i,s}Z_{j(v),s}Z'_{j(w),s}\right)(\hat{\beta}_{x\mathbf{z}_j(v)}-\beta_{x\mathbf{z}_j(v)})(\hat{\beta}_{y\mathbf{z}'_j(w)}-\beta_{y\mathbf{z}'_j(w)}) \\
&\xrightarrow{p} \sum_{v,w}\text{const}\cdot 0\cdot 0 \\
&= 0,
\end{aligned}
$$

where $1 \le v \le |\mathbf{Z}_j|, 1 \le w \le |\mathbf{Z}'_j|$, the constant term $\mathrm{E}(\delta_{x\mathbf{z}_i}\delta_{y\mathbf{z}'_i}Z_{j(v)}Z'_{j(w)})$ exists due to the finite fourth moment assumption on $\mathbf{V}$.

Three δ-terms.

$$
\begin{aligned}
& -n^{-1}\sum_{s=1}^{n}\delta_{x\mathbf{z}_i,s}\delta_{y\mathbf{z}'_i,s}\delta_{x\mathbf{z}_j,s}\mathbf{Z}_{j,s}'^{\top}(\hat{\boldsymbol{\beta}}_{y\mathbf{z}'_j}-\boldsymbol{\beta}_{y\mathbf{z}'_j}) \\
&= \sum_{w}\left(n^{-1}\sum_{s=1}^{n}\delta_{x\mathbf{z}_i,s}\delta_{y\mathbf{z}'_i,s}\delta_{x\mathbf{z}_j,s}Z'_{j(w),s}\right)(\hat{\beta}_{y\mathbf{z}'_j(w)}-\beta_{y\mathbf{z}'_j(w)}) \\
&\xrightarrow{p} \sum_{w}\text{const}\cdot 0 \\
&= 0,
\end{aligned}
$$

where $1 \le w \le |\mathbf{Z}'_j|$, the constant term $\mathrm{E}(\delta_{x\mathbf{z}_i}\delta_{y\mathbf{z}'_i}\delta_{x\mathbf{z}_j}Z'_{j(w)})$ exists due to the finite fourth moment assumption on $\mathbf{V}$. □

Remark. If the $\mathbf{Z}_i$'s are valid adjustment sets, the diagonal terms simplify to $(\hat{\boldsymbol{\Sigma}}_{\mathcal{Z}})_{ii} = \|\mathbf{r}_{y\mathbf{z}'_i}\|_2^2/\|\mathbf{r}_{x\mathbf{z}_i}\|_2^2$, and their convergence follows by the proof of Lemma 3 on the denominator.

B.3 PROOF OF PROPOSITION 7

Proof. The proof aims to show that the half-vectorised asymptotic covariance matrix estimator $\hat{\boldsymbol{\Sigma}}_{\mathcal{Z}}$, after subtracting their true values in $\boldsymbol{\Sigma}_{\mathcal{Z}}$, will converge to a zero-mean normal distribution.

For the (i,j)-th entry, we write

$$
\begin{aligned}
n^{1/2}(\hat{\boldsymbol{\Sigma}}_{\mathcal{Z},ij}-\boldsymbol{\Sigma}_{\mathcal{Z},ij}) &= \frac{n^{-1/2}\sum_{s=1}^{n} r_{x\mathbf{z}_i,s}r_{y\mathbf{z}'_i,s}r_{x\mathbf{z}_j,s}r_{y\mathbf{z}'_i,s}}{n^{-1}\sum_{s=1}^{n} r_{x\mathbf{z}_i,s}^2 n^{-1}\sum_{s=1}^{n} r_{x\mathbf{z}_j,s}^2} - \frac{n^{1/2}\,\mathrm{E}(\delta_{x\mathbf{z}_i}\delta_{y\mathbf{z}'_i}\delta_{x\mathbf{z}_j}\delta_{y\mathbf{z}'_i})}{\mathrm{E}(\delta_{x\mathbf{z}_i}^2)\,\mathrm{E}(\delta_{x\mathbf{z}_j}^2)} \\
&= \frac{N}{\mathrm{E}(\delta_{x\mathbf{z}_i}^2)\,\mathrm{E}(\delta_{x\mathbf{z}_j}^2)n^{-1}\sum_{s=1}^{n} r_{x\mathbf{z}_i,s}^2 n^{-1}\sum_{s=1}^{n} r_{x\mathbf{z}_j,s}^2}.
\end{aligned}
$$

The numerator N of the expression above is expanded as

$$
\begin{aligned}
& \mathrm{E}(\delta_{x\mathbf{z}_i}^2)\,\mathrm{E}(\delta_{x\mathbf{z}_j}^2)n^{-1/2}\sum_{s=1}^{n} r_{x\mathbf{z}_i,s}r_{y\mathbf{z}'_i,s}r_{x\mathbf{z}_j,s}r_{y\mathbf{z}'_i,s} - \mathrm{E}(\delta_{x\mathbf{z}_i}\delta_{y\mathbf{z}'_i}\delta_{x\mathbf{z}_j}\delta_{y\mathbf{z}'_j})n^{1/2}n^{-1}\sum_{s=1}^{n} r_{x\mathbf{z}_i,s}^2 n^{-1}\sum_{s=1}^{n} r_{x\mathbf{z}_j,s}^2 \\
&= \mathrm{E}(\delta_{x\mathbf{z}_i}^2)\,\mathrm{E}(\delta_{x\mathbf{z}_j}^2)n^{-1/2}\sum_{s=1}^{n}\delta_{x\mathbf{z}_i,s}\delta_{y\mathbf{z}'_i,s}\delta_{x\mathbf{z}_j,s}\delta_{y\mathbf{z}'_i,s} - \mathrm{E}(\delta_{x\mathbf{z}_i}\delta_{y\mathbf{z}'_i}\delta_{x\mathbf{z}_j}\delta_{y\mathbf{z}'_j})n^{1/2}n^{-1}\sum_{s=1}^{n}\delta_{x\mathbf{z}_i,s}^2 n^{-1}\sum_{s=1}^{n}\delta_{x\mathbf{z}_j,s}^2 + R,
\end{aligned}
$$

where R collects the remainder term resulting from replacing the sample residuals with population residuals.

We now subtract and add back the expected squared population residuals from the average squared population residuals.

That is,

$$
\begin{aligned}
n^{1/2}(\hat{\mathbf{\Sigma}}_{\mathcal{Z},ij} - \mathbf{\Sigma}_{\mathcal{Z},ij}) &= \mathrm{E}(\delta^2_{x\mathbf{z}_i})\,\mathrm{E}(\delta^2_{x\mathbf{z}_j}) n^{-1/2} \sum_{s=1}^{n} \delta_{x\mathbf{z}_i,s}\delta_{y\mathbf{z}'_i,s}\delta_{x\mathbf{z}_j,s}\delta_{y\mathbf{z}'_i,s} \\
&\quad - \mathrm{E}(\delta_{x\mathbf{z}_i}\delta_{y\mathbf{z}'_i}\delta_{x\mathbf{z}_j}\delta_{y\mathbf{z}'_j}) n^{1/2} \left(n^{-1}\sum_{s=1}^{n} \delta^2_{x\mathbf{z}_i,s} - \mathrm{E}(\delta^2_{x\mathbf{z}_i}) \right) \left(n^{-1}\sum_{s=1}^{n} \delta^2_{x\mathbf{z}_j,s} - \mathrm{E}(\delta^2_{x\mathbf{z}_j}) \right) \\
&\quad - \mathrm{E}(\delta_{x\mathbf{z}_i}\delta_{y\mathbf{z}'_i}\delta_{x\mathbf{z}_j}\delta_{y\mathbf{z}'_j}) n^{-1/2} \sum_{s=1}^{n} \left(\mathrm{E}(\delta^2_{x\mathbf{z}_i})\delta^2_{x\mathbf{z}_j,s} + \mathrm{E}(\delta^2_{x\mathbf{z}_j})\delta^2_{x\mathbf{z}_i,s} \right) \\
&\quad + \mathrm{E}(\delta_{x\mathbf{z}_i}\delta_{y\mathbf{z}'_i}\delta_{x\mathbf{z}_j}\delta_{y\mathbf{z}'_j}) n^{1/2}\,\mathrm{E}(\delta^2_{x\mathbf{z}_i})\,\mathrm{E}(\delta^2_{x\mathbf{z}_j}) + R \\
&= n^{-1/2} \sum_{s=1}^{n} \Big[\mathrm{E}(\delta^2_{x\mathbf{z}_i})\,\mathrm{E}(\delta^2_{x\mathbf{z}_j}) \delta_{x\mathbf{z}_i,s}\delta_{y\mathbf{z}'_i,s}\delta_{x\mathbf{z}_j,s}\delta_{y\mathbf{z}'_i,s} \\
&\quad - \mathrm{E}(\delta_{x\mathbf{z}_i}\delta_{y\mathbf{z}'_i}\delta_{x\mathbf{z}_j}\delta_{y\mathbf{z}'_j}) \left(\mathrm{E}(\delta^2_{x\mathbf{z}_i})\delta^2_{x\mathbf{z}_j,s} + \mathrm{E}(\delta^2_{x\mathbf{z}_j})\delta^2_{x\mathbf{z}_i,s} \right) \\
&\quad + \mathrm{E}(\delta_{x\mathbf{z}_i}\delta_{y\mathbf{z}'_i}\delta_{x\mathbf{z}_j}\delta_{y\mathbf{z}'_j})\,\mathrm{E}(\delta^2_{x\mathbf{z}_i})\,\mathrm{E}(\delta^2_{x\mathbf{z}_j}) \Big] + R'.
\end{aligned}
$$

The first term converges to a zero-mean normal distribution by the central limit theorem and the finite fourth moment assumption on $\mathbf{V}$. The remainder term $R = o_p(1)$ by analogous arguments to the ones used in the proof of Lemma 12. The second term on the second to last line disappears asymptotically, which entails that $R' = o_p(1)$.

The asymptotic covariance between two entries in $\mathrm{vech}(\hat{\mathbf{\Sigma}}_{\mathcal{Z}})$

$$
a.\,\mathrm{Cov}(n^{1/2}(\hat{\mathbf{\Sigma}}_{\mathcal{Z},ij} - \mathbf{\Sigma}_{\mathcal{Z},ij}), n^{1/2}(\hat{\mathbf{\Sigma}}_{\mathcal{Z},kl} - \mathbf{\Sigma}_{\mathcal{Z},kl})) = \frac{\gamma_{ij,kl}}{\omega_{ij,kl}},
$$

where

$$
\begin{aligned}
\gamma_{ij,kl} := {} & \mathrm{E}(\delta^2_{x\mathbf{z}_i})\,\mathrm{E}(\delta^2_{x\mathbf{z}_j})\,\mathrm{E}(\delta^2_{x\mathbf{z}_k})\,\mathrm{E}(\delta^2_{x\mathbf{z}_l})\,\mathrm{Cov}(\delta_{x\mathbf{z}_i}\delta_{y\mathbf{z}'_i}\delta_{x\mathbf{z}_j}\delta_{y\mathbf{z}'_j}, \delta_{x\mathbf{z}_k}\delta_{y\mathbf{z}'_k}\delta_{x\mathbf{z}_l}\delta_{y\mathbf{z}'_l}) \\
& - \mathrm{E}(\delta^2_{x\mathbf{z}_i})\,\mathrm{E}(\delta^2_{x\mathbf{z}_j})\,\mathrm{E}(\delta^2_{x\mathbf{z}_k})\,\mathrm{E}(\delta_{x\mathbf{z}_k}\delta_{y\mathbf{z}'_k}\delta_{x\mathbf{z}_l}\delta_{y\mathbf{z}'_l})\,\mathrm{Cov}(\delta_{x\mathbf{z}_i}\delta_{y\mathbf{z}'_i}\delta_{x\mathbf{z}_j}\delta_{y\mathbf{z}'_j}, \delta^2_{x\mathbf{z}_l}) \\
& - \mathrm{E}(\delta^2_{x\mathbf{z}_i})\,\mathrm{E}(\delta^2_{x\mathbf{z}_j})\,\mathrm{E}(\delta^2_{x\mathbf{z}_l})\,\mathrm{E}(\delta_{x\mathbf{z}_k}\delta_{y\mathbf{z}'_k}\delta_{x\mathbf{z}_l}\delta_{y\mathbf{z}'_l})\,\mathrm{Cov}(\delta_{x\mathbf{z}_i}\delta_{y\mathbf{z}'_i}\delta_{x\mathbf{z}_j}\delta_{y\mathbf{z}'_j}, \delta^2_{x\mathbf{z}_k}) \\
& - \mathrm{E}(\delta^2_{x\mathbf{z}_i})\,\mathrm{E}(\delta^2_{x\mathbf{z}_k})\,\mathrm{E}(\delta^2_{x\mathbf{z}_l})\,\mathrm{E}(\delta_{x\mathbf{z}_i}\delta_{y\mathbf{z}'_i}\delta_{x\mathbf{z}_j}\delta_{y\mathbf{z}'_j})\,\mathrm{Cov}(\delta_{x\mathbf{z}_k}\delta_{y\mathbf{z}'_k}\delta_{x\mathbf{z}_l}\delta_{y\mathbf{z}'_l}, \delta^2_{x\mathbf{z}_j}) \\
& - \mathrm{E}(\delta^2_{x\mathbf{z}_j})\,\mathrm{E}(\delta^2_{x\mathbf{z}_k})\,\mathrm{E}(\delta^2_{x\mathbf{z}_l})\,\mathrm{E}(\delta_{x\mathbf{z}_i}\delta_{y\mathbf{z}'_i}\delta_{x\mathbf{z}_j}\delta_{y\mathbf{z}'_j})\,\mathrm{Cov}(\delta_{x\mathbf{z}_k}\delta_{y\mathbf{z}'_k}\delta_{x\mathbf{z}_l}\delta_{y\mathbf{z}'_l}, \delta^2_{x\mathbf{z}_i}) \\
& + \mathrm{E}(\delta_{x\mathbf{z}_i}\delta_{y\mathbf{z}'_i}\delta_{x\mathbf{z}_j}\delta_{y\mathbf{z}'_j})\,\mathrm{E}(\delta_{x\mathbf{z}_k}\delta_{y\mathbf{z}'_k}\delta_{x\mathbf{z}_l}\delta_{y\mathbf{z}'_l})\,\mathrm{E}(\delta^2_{x\mathbf{z}_i})\,\mathrm{E}(\delta^2_{x\mathbf{z}_k})\,\mathrm{Cov}(\delta^2_{x\mathbf{z}_j}, \delta^2_{x\mathbf{z}_l}) \\
& + \mathrm{E}(\delta_{x\mathbf{z}_i}\delta_{y\mathbf{z}'_i}\delta_{x\mathbf{z}_j}\delta_{y\mathbf{z}'_j})\,\mathrm{E}(\delta_{x\mathbf{z}_k}\delta_{y\mathbf{z}'_k}\delta_{x\mathbf{z}_l}\delta_{y\mathbf{z}'_l})\,\mathrm{E}(\delta^2_{x\mathbf{z}_i})\,\mathrm{E}(\delta^2_{x\mathbf{z}_l})\,\mathrm{Cov}(\delta^2_{x\mathbf{z}_j}, \delta^2_{x\mathbf{z}_k}) \\
& + \mathrm{E}(\delta_{x\mathbf{z}_i}\delta_{y\mathbf{z}'_i}\delta_{x\mathbf{z}_j}\delta_{y\mathbf{z}'_j})\,\mathrm{E}(\delta_{x\mathbf{z}_k}\delta_{y\mathbf{z}'_k}\delta_{x\mathbf{z}_l}\delta_{y\mathbf{z}'_l})\,\mathrm{E}(\delta^2_{x\mathbf{z}_j})\,\mathrm{E}(\delta^2_{x\mathbf{z}_k})\,\mathrm{Cov}(\delta^2_{x\mathbf{z}_i}, \delta^2_{x\mathbf{z}_l}) \\
& + \mathrm{E}(\delta_{x\mathbf{z}_i}\delta_{y\mathbf{z}'_i}\delta_{x\mathbf{z}_j}\delta_{y\mathbf{z}'_j})\,\mathrm{E}(\delta_{x\mathbf{z}_k}\delta_{y\mathbf{z}'_k}\delta_{x\mathbf{z}_l}\delta_{y\mathbf{z}'_l})\,\mathrm{E}(\delta^2_{x\mathbf{z}_j})\,\mathrm{E}(\delta^2_{x\mathbf{z}_l})\,\mathrm{Cov}(\delta^2_{x\mathbf{z}_i}, \delta^2_{x\mathbf{z}_k}) \text{ and} \\
\omega_{ij,kl} := {} & [\mathrm{E}(\delta^2_{x\mathbf{z}_i})]^2 [\mathrm{E}(\delta^2_{x\mathbf{z}_j})]^2 [\mathrm{E}(\delta^2_{x\mathbf{z}_k})]^2 [\mathrm{E}(\delta^2_{x\mathbf{z}_l})]^2.
\end{aligned}
$$

Analogous to the proof of Lemma 3, the joint normality follows by the multivariate Central Limit Theorem, which we can be apply as by assumption the fourth moments of the errors are finite, and Slutsky's Theorem.

Define a deterministic mapping for subscript $\mathbf{g}(a) = (ij)$, $a = 1, 2, \ldots, k(k+1)/2$ such that it maps the a-th element of $\mathrm{vech}(\mathbf{\Sigma}_{\mathcal{Z}})$ to the (i,j)-th entry of $\mathbf{\Sigma}_{\mathcal{Z}}$. The asymptotic covariance matrix $\mathbf{F}$ of $\mathrm{vech}(\hat{\mathbf{\Sigma}}_{\mathcal{Z}})$ is a $k(k+1)/2 \times k(k+1)/2$ matrix whose entries are related to the expression of $\omega_{\cdot,\cdot}$ and $\gamma_{\cdot,\cdot}$ by the mapping $\mathbf{g}(\cdot)$ such that

$$
\mathbf{F}_{ab} = \frac{\gamma_{\mathbf{g}(a),\mathbf{g}(b)}}{\omega_{\mathbf{g}(a),\mathbf{g}(b)}},
$$

for $1 \leq a, b \leq k(k+1)/2$. The asymptotic covariance matrix $\mathbf{C}$ of $\mathrm{vech}(\hat{\mathbf{\Delta}}_{\mathcal{Z}})$ follows from the linear relationship $\mathrm{vech}(\mathbf{\Delta}_{\mathcal{Z}}) = \mathbf{\Pi}\mathrm{vech}(\mathbf{\Sigma}_{\mathcal{Z}})$. □

Remark. Again we discuss the special situation where the $\mathbf{Z}_i$'s are valid adjustments sets. In this case, the diagonal terms

$$n^{1/2}(\hat{\mathbf{\Sigma}}_{\mathcal{Z},ii} - \mathbf{\Sigma}_{\mathcal{Z},ii}) = \frac{n^{1/2}n^{-1}\sum_{s=1}^n r^2_{y\mathbf{z}'_i,s}}{n^{-1}\sum_{s=1}^n r^2_{x\mathbf{z}_i,s}} - \frac{\mathrm{E}(\delta^2_{y\mathbf{z}'_i})}{\mathrm{E}(\delta^2_{x\mathbf{z}_i})}$$
$$= \frac{n^{-1/2}\sum_{s=1}^n [\mathrm{E}(\delta^2_{x\mathbf{z}_i})r^2_{y\mathbf{z}'_i,s} - \mathrm{E}(\delta^2_{y\mathbf{z}'_i})r^2_{x\mathbf{z}_i,s}]}{\mathrm{E}(\delta^2_{x\mathbf{z}_i})n^{-1}\sum_{s=1}^n r^2_{x\mathbf{z}_i,s}}.$$

The numerator

$$n^{-1/2}\sum_{s=1}^n [\mathrm{E}(\delta^2_{x\mathbf{z}_i})r^2_{y\mathbf{z}'_i,s} - \mathrm{E}(\delta^2_{y\mathbf{z}'_i})r^2_{x\mathbf{z}_i,s}] = n^{-1/2}\sum_{s=1}^n \Big[\mathrm{E}(\delta^2_{x\mathbf{z}_i})(\delta_{x\mathbf{z}_i,s} - \mathbf{Z}'^{\top}_{i,s}(\hat{\boldsymbol{\beta}}_{y\mathbf{z}'_i} - \boldsymbol{\beta}_{y\mathbf{z}'_i}))^2$$
$$- \mathrm{E}(\delta^2_{y\mathbf{z}'_i})(\delta_{x\mathbf{z}_i,s} - \mathbf{Z}^{\top}_{i,s}(\hat{\boldsymbol{\beta}}_{x\mathbf{z}_i} - \boldsymbol{\beta}_{x\mathbf{z}_i}))^2\Big]$$
$$= n^{-1/2}\sum_{s=1}^n [\mathrm{E}(\delta^2_{x\mathbf{z}_i})\delta^2_{x\mathbf{z}_i,s} - \mathrm{E}(\delta^2_{y\mathbf{z}'_i})\delta^2_{x\mathbf{z}_i,s}] + R$$
$$\xrightarrow{d} \mathrm{N}(0, [\mathrm{E}(\delta^2_{x\mathbf{z}_i})]^2\,\mathrm{Var}(\delta^2_{y\mathbf{z}'_i}) + [\mathrm{E}(\delta^2_{y\mathbf{z}'_i})]^2\,\mathrm{Var}(\delta^2_{x\mathbf{z}_i})),$$

where we can apply the central limit theorem because to the first term because $\mathrm{E}(\mathrm{E}(\delta^2_{x\mathbf{z}_i})\delta^2_{y\mathbf{z}'_i,s} - \mathrm{E}(\delta^2_{y\mathbf{z}'_i})\delta^2_{x\mathbf{z}_i,s}) = 0$ and $\mathrm{Var}(\mathrm{E}(\delta^2_{x\mathbf{z}_i})\delta^2_{y\mathbf{z}'_i,s} - \mathrm{E}(\delta^2_{y\mathbf{z}'_i})\delta^2_{x\mathbf{z}_i,s}) = [\mathrm{E}(\delta^2_{x\mathbf{z}_i})]^2\,\mathrm{Var}(\delta^2_{y\mathbf{z}'_i}) + [\mathrm{E}(\delta^2_{y\mathbf{z}'_i})]^2\,\mathrm{Var}(\delta^2_{x\mathbf{z}_i})$. The remainder term $R = o_p(1)$ by analogous arguments used in the proof of Lemma 12. Similarly, the denominator $\mathrm{E}(\delta^2_{x\mathbf{z}_i})n^{-1}\sum_{s=1}^n r^2_{x\mathbf{z}_i,s}$ converges in probability to $[\mathrm{E}(\delta^2_{x\mathbf{z}_i})]^2$. Then by Slutsky's Theorem,

$$n^{1/2}(\hat{\mathbf{\Sigma}}_{\mathcal{Z},ii} - \mathbf{\Sigma}_{\mathcal{Z},ii}) \xrightarrow{d} \mathrm{N}\left(0, \frac{[\mathrm{E}(\delta^2_{x\mathbf{z}_i})]^2\,\mathrm{Var}(\delta^2_{y\mathbf{z}'_i}) + [\mathrm{E}(\delta^2_{y\mathbf{z}'_i})]^2\,\mathrm{Var}(\delta^2_{x\mathbf{z}_i})}{[\mathrm{E}(\delta^2_{x\mathbf{z}_i})]^4}\right).$$

B.4 PROOF OF THEOREM 6

Proof of Theorem 6. Lemma 3 states that $n^{1/2}(\hat{\boldsymbol{\beta}}_{yx.\mathcal{Z}} - \boldsymbol{\beta}_{yx.\mathcal{Z}})$ is asymptotically normal. We first show that to quantify the degrees of freedom of a Wald-type statistic, one only needs to look at the rank of covariance matrix $\mathbf{\Delta}_{\mathcal{Z}} = \mathbf{\Gamma}\mathbf{\Sigma}_{\mathcal{Z}}\mathbf{\Gamma}^{\top}$.

Suppose $\mathrm{rank}(\mathbf{\Delta}_{\mathcal{Z}}) = r_0 \le l$ where $l = k-1$. Consider the eigendecomposition of $\mathbf{\Delta}_{\mathcal{Z}} = \mathbf{Q}\mathbf{\Phi}\mathbf{Q}^{\top}$, where $\mathbf{Q} = (\mathbf{q}_1 \ \cdots \ \mathbf{q}_l)$ is the orthonormal matrix containing the eigenvectors of $\mathbf{\Delta}_{\mathcal{Z}}$, and $\mathbf{\Phi} = \mathrm{diag}(\phi_1, \ldots, \phi_l)$ with eigenvalues $\phi_1 \ge \cdots \ge \phi_{r_0} > \phi_{r_0+1} = \cdots = \phi_l = 0$. It can be verified that the (unique) Moore-Penrose inverse of $\mathbf{\Delta}_{\mathcal{Z}}$ is defined as

$$\mathbf{\Delta}^{\dagger}_{\mathcal{Z}} = \sum_{s=1}^{r_0} \phi_j^{-1}\mathbf{q}_s\mathbf{q}_s^{\top},$$

because of the semi-positive definiteness. Under $H_0 : \mathbf{\Gamma}\boldsymbol{\beta}_{yx.\mathcal{Z}} = \mathbf{0}$, denote $n^{1/2}\mathbf{\Gamma}\hat{\boldsymbol{\beta}}_{yx} \xrightarrow{d} \mathbf{G} \sim \mathrm{N}(\mathbf{0}, \mathbf{\Delta}_{\mathcal{Z}})$. For all $1 \le s \ne t \le r_0$, $\mathrm{Cov}(\mathbf{q}_s^{\top}\mathbf{G}, \mathbf{q}_t^{\top}\mathbf{G}) = \mathbf{q}_s^{\top}\mathbf{\Delta}_{\mathcal{Z}}\mathbf{q}_t = 0$. By joint normality of $\mathbf{G}$, $\mathbf{q}_s^{\top}\mathbf{G}$ and $\mathbf{q}_t^{\top}\mathbf{G}$ are independent. Moreover, since $\mathbf{q}_s^{\top}\mathbf{G} \sim \mathrm{N}(0, \phi_s)$,

$$n(\mathbf{\Gamma}\hat{\boldsymbol{\beta}}_{yx.\mathcal{Z}})^{\top}\mathbf{\Delta}^{\dagger}_{\mathcal{Z}}(\mathbf{\Gamma}\hat{\boldsymbol{\beta}}_{yx.\mathcal{Z}}) = \sum_{s=1}^{r_0}\phi_s^{-1}(\mathbf{q}_s^{\top}n^{1/2}\mathbf{\Gamma}\hat{\boldsymbol{\beta}}_{yx.\mathcal{Z}})^2$$
$$\xrightarrow{d} \sum_{s=1}^{r_0}\phi_s^{-1}(\mathbf{q}_s^{\top}\mathbf{G})^2 \sim \chi^2_{r_0}. \tag{8}$$

The consistency of $\hat{r}$, i.e., $\lim_{n\to\infty}\mathbb{P}(|\hat{r} - r_0| < \epsilon) = 1, \forall\epsilon > 0$, implies that $\lim_{n\to\infty}\mathbb{P}(\hat{r} = r_0) = 1$ when taking $\epsilon < 1$, since both $\hat{r}$ and r_0 are integer-valued.

Since $\hat{\mathbf{\Delta}}_{\mathcal{Z}}$ is positive semidefinite, its spectral decomposition is $\hat{\mathbf{P}}\hat{\mathbf{\Lambda}}\hat{\mathbf{P}}^{\top}$, where $\hat{\mathbf{\Lambda}} = \mathrm{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_k)$ with $\hat{\lambda}_1 \ge \cdots \ge \hat{\lambda}_k \ge 0$. The rank-$\hat{r}$ spectral approximation of $\hat{\mathbf{\Delta}}_{\mathcal{Z}}$ is then $\hat{\mathbf{P}}\hat{\mathbf{\Lambda}}_{\hat{r}}\hat{\mathbf{P}}^{\top}$, where $\hat{\mathbf{\Lambda}}_{\hat{r}} = \mathrm{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_{\hat{r}}, 0, \ldots, 0)$. Following

Weyl's inequality [Stewart, 1998] and Proposition 7, we have $\hat{\mathbf{\Lambda}} \xrightarrow{p} \mathbf{\Lambda}$ since the asymptotic covariance matrix of $\mathrm{vech}(\hat{\mathbf{\Delta}}_{\mathcal{Z}})$ is finite. We now show that $\hat{\mathbf{\Lambda}}_{\hat{r}} \xrightarrow{p} \mathbf{\Lambda}$. For any $\ell \in \{1, \ldots, k\}$,

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}(|\hat{\lambda}_{\hat{r},\ell} - \hat{\lambda}_\ell| < \epsilon) \\
&= \lim_{n\to\infty} \mathbb{P}(|\hat{\lambda}_{\hat{r},\ell} - \hat{\lambda}_\ell| < \epsilon \mid \hat{r} = r_0)\mathbb{P}(\hat{r} = r_0) \\
&\quad + \lim_{n\to\infty} \mathbb{P}(|\hat{\lambda}_{\hat{r},\ell} - \hat{\lambda}_\ell| < \epsilon \mid \hat{r} \neq r_0)\mathbb{P}(\hat{r} \neq r_0) \\
&= \lim_{n\to\infty} \mathbb{P}(|\hat{\lambda}_{r_0,\ell} - \hat{\lambda}_\ell| < \epsilon \mid \hat{r} = r_0).
\end{aligned}$$

If $\ell \leq r_0$, $\hat{\lambda}_{r_0,\ell} = \hat{\lambda}_\ell$ and $\mathbb{P}(|\hat{\lambda}_{r_0,\ell} - \hat{\lambda}_\ell| < \epsilon \mid \hat{r} = r_0) = 1$. Otherwise if $\ell > r_0$, $\lim_{n\to\infty} \mathbb{P}(|\hat{\lambda}_{r_0,\ell} - \hat{\lambda}_\ell| < \epsilon \mid \hat{r} = r_0) = \lim_{n\to\infty} \mathbb{P}(|\hat{\lambda}_\ell| < \epsilon \mid \hat{r} = r_0) = 1$ because $\hat{\lambda}_\ell \xrightarrow{p} \lambda_\ell = 0$. Hence, $\hat{\lambda}_{\hat{r},\ell} \xrightarrow{p} \hat{\lambda}_\ell$ for all ℓ. Since all entries of $\hat{\mathbf{P}}$ are bounded by 1, $\hat{\mathbf{\Delta}}_{\mathcal{Z},\hat{r}} - \hat{\mathbf{\Delta}}_{\mathcal{Z}} = \hat{\mathbf{P}}(\hat{\mathbf{\Lambda}}_{\hat{r}} - \mathbf{\Lambda})\hat{\mathbf{P}}^\top \xrightarrow{p} 0$. Then $\hat{\mathbf{\Delta}}_{\mathcal{Z},\hat{r}} \xrightarrow{p} \mathbf{\Delta}_{\mathcal{Z}}$ by consistency of $\hat{\mathbf{\Delta}}_{\mathcal{Z}}$.

The rank of $\hat{\mathbf{\Delta}}_{\mathcal{Z},\hat{r}}$ is equal to $\hat{r}$ by construction. With the condition that $\mathbb{P}(\mathrm{rank}(\hat{\mathbf{\Delta}}_{\mathcal{Z},\hat{r}}) = \mathrm{rank}(\mathbf{\Delta}_{\mathcal{Z}})) \to 1$, it follows from Theorem 2 in Andrews [1987] that $\hat{\mathbf{\Delta}}^\dagger_{\mathcal{Z},\hat{r}} \xrightarrow{p} \mathbf{\Delta}^\dagger_{\mathcal{Z}}$. By Slutsky's theorem, the convergence in distribution in (8) still holds if we use a consistent estimator $\hat{\mathbf{\Delta}}^\dagger_{\mathcal{Z},\hat{r}}$ of $\mathbf{\Delta}^\dagger_{\mathcal{Z}}$ instead. Therefore, $n(\mathbf{\Gamma}\hat{\boldsymbol{\beta}}_{yx.\mathcal{Z}})^\top \hat{\mathbf{\Delta}}^\dagger_{\mathcal{Z},\hat{r}}(\mathbf{\Gamma}\hat{\boldsymbol{\beta}}_{yx.\mathcal{Z}}) \xrightarrow{d} \chi^2_{r_0}$. □

B.5 PROOF OF LEMMA 8

Lemma 13 (Modified Lemma D.1 in Henckel et al. [2019])**.** *Consider a causal DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ such that $X, Y \in \mathbf{V}$ and that $\mathbf{Z} \subset \mathbf{V} \setminus \{X, Y\}$ is a valid adjustment set relative to (X, Y) in $\mathcal{G}$. Given a partition $\mathbf{Z} = \mathbf{Z}_1 \cup \mathbf{Z}_2$, if $X \perp_{\mathcal{G}} \mathbf{Z}_1 \mid \mathbf{Z}_2$, then $\mathbf{Z}_2$ is a valid adjustment set relative to (X, Y) in $\mathcal{G}$.*

Theorem 14 (Spirtes, 1995)**.** *Consider DAG $\mathcal{G}$ containing X, Y and $\mathbf{Z}$, where $X \neq Y$ and $\mathbf{Z}$ does not contain X or Y, X is d-separated from Y given $\mathbf{Z}$ if and only if the partial correlation coefficient $\rho_{xy.\mathbf{z}} = 0$ for all linear structural equation models compatible with $\mathcal{G}$.*

Corollary 15. *Consider nodes X and Y, and a set $\mathbf{Z}$ in a DAG $\mathcal{G}$. Then X is d-separated from Y given $\mathbf{Z}$ if and only if $\beta_{yx.\mathbf{z}} = 0$ for some linear structural equation model compatible with and faithful to $\mathcal{G}$.*

Lemma 16. *Consider a causal DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ and let $\mathbf{V}$ follow a linear structural equation model compatible with $\mathcal{G}$. Let $\boldsymbol{\epsilon} = \{\epsilon_{v_1}, \epsilon_{v_2}, \ldots, \epsilon_{v_p}\}$ be the set of independent errors from the linear structural equation model, where p is the number of nodes in $\mathcal{G}$. Given two nodes $X, Y \in \mathbf{V}$ such that $Y \in \mathrm{de}(X, \mathcal{G})$ and any valid adjustment set $\mathbf{Z}$ relative to (X, Y) in $\mathcal{G}$, the population regression residual $\delta_{y\mathbf{z}'}$ is a linear combination of the error terms $\boldsymbol{\epsilon}$, in which the coefficient of ϵ_y is 1.*

Proof. We refer to the proof of Lemma B.4 in Henckel et al. [2019]. The residual $\delta_{y\mathbf{z}'}$ can be written as a linear combination of errors. In particular, the coefficient of ϵ_Y is

$$\tau_{yy} - \sum_{N \in \mathrm{de}(Y,\mathcal{G}) \cap \mathbf{Z}'} \beta_{yn.\mathbf{z}'_{-n}} \tau_{ny}.$$

Since $\mathbf{Z}$ is a valid adjustment set relative to (X, Y) in $\mathcal{G}$, it cannot contain descendants of Y, which are forbidden nodes. Then the set $\mathrm{de}(Y, \mathcal{G}) \cap \mathbf{Z}'$ is empty, because $X \notin \mathrm{de}(Y, \mathcal{G})$. The result is immediate using the convention that $\tau_{yy} = 1$. □

We are now ready to present the proof of Lemma 8.

Proof of Lemma 8. Consider a linear structural equation model that is faithful to $\mathcal{G}$. We will first only consider the minimal valid adjustment sets $\mathbf{Z}_1, \ldots, \mathbf{Z}_k$ in the collection $\mathcal{Z}$. The first step of the proof is to show that the regression residuals $(\delta_{x\mathbf{z}_1}, \ldots, \delta_{x\mathbf{z}_k})$ cannot be linearly dependent. Suppose on the contrary that there is a linear combination $\ell = \sum_i \alpha_i \delta_{x\mathbf{z}_i}$ such that $\ell = 0$ for some $\alpha_1, \ldots, \alpha_k$ not all equal to 0. Without loss of generality, suppose that $\alpha_1 \neq 0$. Consider the first minimal valid adjustment set $\mathbf{Z}_1$. It contains at least one unique node $N \notin \cup_{2 \leq j \leq k} \mathbf{Z}_j$. We can thus write $\delta_{x\mathbf{z}_1} = X - \boldsymbol{\beta}^\top_{x\mathbf{z}_1} \mathbf{Z}_1$, where $\boldsymbol{\beta}_{x\mathbf{z}_1}$ is the population OLS regression coefficient of X on $\mathbf{Z}_1$. Since $\mathbf{Z}_1$ is a minimal adjustment set, node N is d-connected with X in $\mathcal{G}$ given $\mathbf{Z}_1 \setminus \{N\}$ by Lemma 13. It follows from Corollary 15 that the regression coefficient $\beta_{xn.\mathbf{z}_{1,-n}}$ of N in

$\boldsymbol{\beta}_{x\mathbf{z}_1}$ cannot be zero. In this case, expanding $\delta_{x\mathbf{z}_i}$ into $X - \boldsymbol{\beta}_{x\mathbf{z}_i}^\top \mathbf{Z}_i$ and rearranging the terms, the equation $\ell = 0$ can be expressed equivalently as

$$N = \frac{1}{\alpha_1 \beta_{xn.\mathbf{z}_{1,-n}}} \left[\alpha_1 \left(X - \sum_{V \in \mathbf{Z}_1 \setminus \{N\}} \beta_{xv.\mathbf{z}_{1,-v}} V \right) + \sum_{i \neq 1} \alpha_i (X - \boldsymbol{\beta}_{x\mathbf{z}_i}^\top \mathbf{Z}_i) \right] = \sum_{V \neq N} \gamma_v V, \tag{9}$$

where $\gamma_v = -(\alpha_1 \beta_{xn.\mathbf{z}_{1,-n}})^{-1} \left(\sum_i I(V \in \mathbf{Z}_i) \beta_{xv.\mathbf{z}_{i,-v}} \right)$ for $V \neq X$ and $\gamma_x = (\alpha_1 \beta_{xn.\mathbf{z}_{1,-n}})^{-1} \sum_i \alpha_i$. Equation (9) cannot hold due to the fact that the covariance matrix of $\mathbf{V}$ is non-singular. Therefore, we conclude that $\ell \neq 0$ when $\alpha_1 \neq 0$. On the contrary, when $\alpha_1 = 0$, the argument above can be repeated for minimal adjustment sets $\mathbf{Z}_2$ with $\alpha_2 \neq 0$, so on and so forth until $\alpha_k \neq 0$. Since the linear combination ℓ cannot evaluate to zero whenever $\alpha_i \neq 0$ for any $i \in \{1, \ldots, k\}$, the inequality $\ell \neq 0$ holds generally for all α_i's not all equal to zero.

The second step is to show that the regression residual products $(\delta_{x\mathbf{z}_1} \delta_{y\mathbf{z}'_1}, \ldots, \delta_{x\mathbf{z}_k} \delta_{y\mathbf{z}'_k})$ cannot be linearly dependent either. Lemma 16 states that each $\delta_{y\mathbf{z}'_i}$ contains the error term ϵ_y. For any valid adjustment set $\mathbf{Z}_i$, $\delta_{y\mathbf{z}'_i} \perp\!\!\!\perp \delta_{x\mathbf{z}_i}$ (see proof of Proposition 3.1 in Supplement from Henckel et al. [2019]). Therefore, $\delta_{x\mathbf{z}_i}$, when written in the form of error terms only, cannot contain ϵ_y. Consider now another linear combination $\ell^* = \sum_i \xi_i \delta_{y\mathbf{z}'_i} \delta_{x\mathbf{z}_i}$. Suppose that $\ell^* = 0$ for some ξ_i's not all equal to 0. We can expand $\delta_{y\mathbf{z}'_i}$ into ϵ_y plus some linear combination of the other errors. Singling out the terms involving ϵ_y in ℓ^*, we have that

$$\epsilon_Y \sum_i \xi_i \delta_{x\mathbf{z}_i} = 0, \tag{10}$$

since $\ell^* = 0$ and ϵ_y is independent from the other errors. Due to the non-degeneracy of ϵ_y, the linear combination $\sum_i \epsilon_i \delta_{x\mathbf{z}_i}$ must evaluate to 0 for some ξ_i's not all equal to 0. However, this is impossible by independence between $\delta_{x\mathbf{z}_i}$'s shown in the first step, and we have reached a contradiction.

Following the proof of Lemma 3, the asymptotic covariance matrix $\boldsymbol{\Psi}$ is precisely the covariance matrix of $(\delta_{x\mathbf{z}_1} \delta_{y\mathbf{z}'_1}, \ldots, \delta_{x\mathbf{z}_k} \delta_{y\mathbf{z}'_k})^\top$, which is non-singular due to linear independence among $\delta_{x\mathbf{z}_i} \delta_{y\mathbf{z}'_i}$'s. Hence, the corresponding asymptotic covariance matrix $\boldsymbol{\Sigma}_{\mathcal{Z} \setminus \mathrm{nonforb}(X,Y,\mathcal{G})}$ also has full rank.

Now we consider the set of non-forbidden nodes. Let $\mathbf{N} = \mathrm{nonforb}(X, Y, \mathcal{G})$. The d-connection condition of a unique node $N \in \mathbf{N}$ and faithfulness ensures a non-zero coefficient in front of N in $\delta_{x\mathbf{n}}$. Since $\mathrm{nonforb}(X, Y, \mathcal{G})$ is a valid adjustment set relative to (X, Y) in $\mathcal{G}$, we can repeat the argument above and conclude that the enlarged asymptotic covariance matrix $\boldsymbol{\Sigma}_{\mathcal{Z}}$ is also non-singular.

When the edge coefficients and the error variances in the linear structural equation model are sampled from an absolutely continuous distribution P with respect to the Lebesgue measure, the model is faithful with probability 1 [Spirtes et al., 2000]. Therefore, since we showed that for all faithful models $\boldsymbol{\Sigma}_{\mathcal{Z}}$ is invertible our claim follows. □

B.6 LEMMA 17

Lemma 17. *Consider nodes X and Y in a DAG $\mathcal{G}$ such that $Y \in \mathrm{de}(X, \mathcal{G})$. Then $\mathrm{nonforb}(X, Y, \mathcal{G})$ is a valid adjustment set relative to (X, Y) in $\mathcal{G}$.*

Proof. Obviously, $\mathrm{nonforb}(X, Y, \mathcal{G})$ does not contain any forbidden nodes so it only remains to show that it blocks all paths from X to Y that are not directed. Note first the only possible path from X to Y that does not contain a non-collider is $X \to C \leftarrow Y$. By assumption $\mathrm{de}(Y, \mathcal{G}) \subseteq \mathrm{forb}(X, Y, \mathcal{G})$ and therefore this path is blocked by $\mathrm{nonforb}(X, Y, \mathcal{G})$. Let p be any other path from X to Y that is not directed. It must therefore contain at least one non-collider. If any non-collider on p is in $\mathrm{nonforb}(X, Y, \mathcal{G})$, p is blocked so suppose this is not the case, i.e., all non-collider on p are in $\mathrm{forb}(X, Y, \mathcal{G})$. Any collider on p must be a descendant of a non-collider on p and is therefore also in $\mathrm{forb}(X, Y, \mathcal{G})$. In this case p is again blocked given $\mathrm{nonforb}(X, Y, \mathcal{G})$ and therefore we can assume that p does not contain any colliders and is therefore of the form $X \leftarrow \cdots \leftarrow F \to \cdots \to Y$. But any node in $\mathrm{forb}(X, Y, \mathcal{G})$ that is not X is a descendant of X and therefore $F = X$ or we would have a violation of the acyclicity assumption. But then p is a directed path which contradicts out starting assumption for p. □

C SIMULATION SETUP

True graph We generate the causal DAGs as Erdős–Rényi random graphs with the graph size drawn uniformly from $\{10, 20\}$ nodes and the expected neighbourhood size drawn uniformly from $\{2, 3, 4, 5\}$, with the function randDAG in R package pcalg [Kalisch et al., 2012].

Linear structural equation model For our compatible linear structural equation we sample edge coefficients uniformly from $[-2, -0.1] \cup [0.1, 2]$. We then draw an error distribution uniformly from one of four distributions: normal, uniform, t, or logistic. Note that we use the same error distribution for all errors in the model. We than sample variances for each error in our model as follows. The variance parameter of the normal errors is sampled uniformly from 0.5 to 1.5. The location parameter of the uniform errors symmatric around zero is sampled uniformly from 1.2 to 2.1. The t-errors are sampled from a t-distribution with 5 degrees of freedom and then scaled by $\sqrt{3/5}$ times the square root of a uniformly sampled number from 0.5 to 1.5. The scale parameter of the logistic errors centred around zero is sampled uniformly from 0.4 to 0.7. By sampling our parameters this way we ensure that the variances are approximately in the interval from 0.4 to 1.6.

The pair (X, Y). The node X is randomly drawn from the true DAG $\mathcal{G}_0$, where we weight each node in $\mathcal{G}_0$ by the number of its descendants minus 1. Once X is fixed, we sample Y uniformly from the set $\mathrm{de}(X, \mathcal{G}_0) \setminus \{X\}$. The sampling procedure is repeated until there are at least two valid adjustment sets relative to the selected pair (X, Y) in the completed partially directed acyclic graph (CPDAG) of $\mathcal{G}_0$.

Causal structure learning We use causal structure learning algorithms to generate large numbers of reasonable candidate graphs for our test procedure. If the error distribution is normal, we apply Greedy Equivalence Search (GES, Chickering [2002]) to estimate a completed partially directed acyclic graph (CPDAG). Note that the adjustment criterion also applies to CPDAGs. Otherwise, we apply LiNGAM [Shimizu, 2014] and estimate a DAG. We use the functions ges and lingam from R package pcalg with default options [Kalisch et al., 2012].

Untestable cases If there is only one or no adjustment set in the candidate graph $\mathcal{G}$, the proposed test cannot be performed so we discard these cases. If $Y \notin \mathrm{de}(X, \mathcal{G})$ the valid adjustment sets are simply those sets that d-separate X from Y. As there is a large literature on conditional independence tests which are more suitable here than our test procedure, we discard this case. If the rank of $\mathbf{\Sigma}_{\mathcal{Z}}$ is estimated to be 1, there is no effective over identifying constraint for our test procedure, so we discard these cases as well.

AUC calculation Recall that for each candidate graph we perform our test 100 times. We plot the probability-probability plot between the corresponding 100 p-values and the standard uniform distribution. We compute the area under the curve (AUC) of this curve with the function auc from R package MESS [Ekstrøm, 2020].

Determining whether null hypothesis is true For every estimated graph and test strategy, we check using the true linear structural equation model whether the null hypothesis H_0 is true or false by computing the population level regression coefficients and checking whether they are all equal.

Version control The simulation studies were conducted using R version 4.1.1.

C.1 EXTRA SIMULATION RESULTS

Figure 6 shows additional plots of the AUCs from the simulation study. The AUCs are grouped by error distribution of the linear structural equation model, graph size of the true graph and expected neighbourhood size of the true graph, respectively. The plots show that of the three parameters only the error distribution seems to have an impact on the performance of our testing procedure. This is likely due to the fact that in cases with normally distributed errors we can only learn a CPDAG, which contain fewer valid adjustment sets than DAGs.

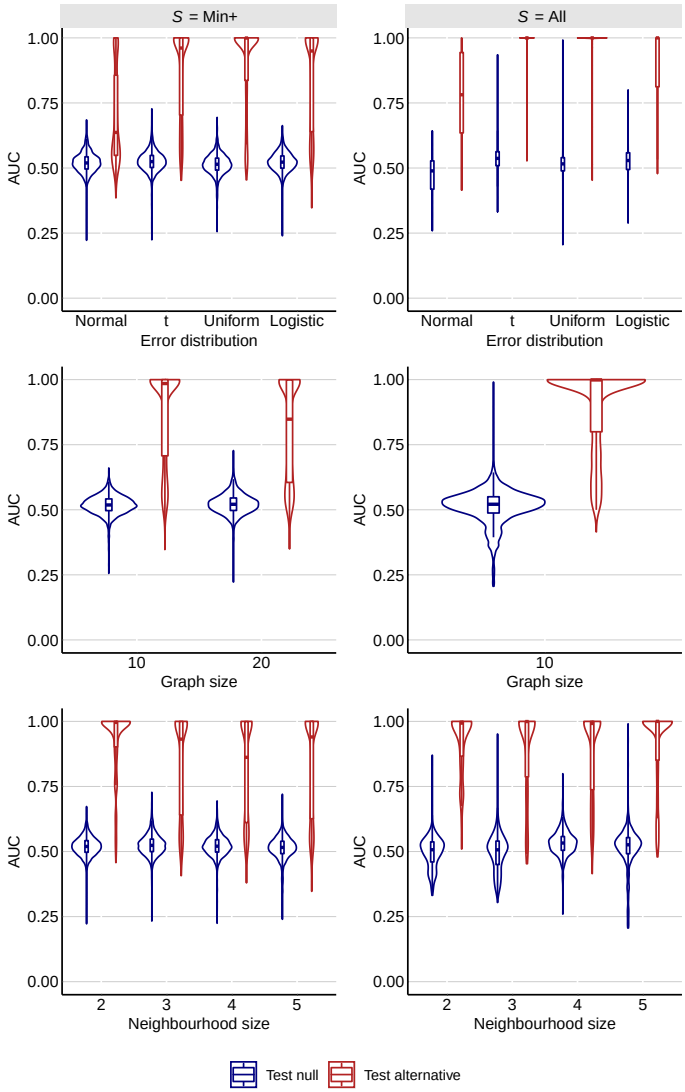

Figure 6: Extra violin plots (layered with boxplots) of AUCs from the simulation study.