A Robustness Test for Estimating Total Effects with Covariate Adjustment (Supplementary Materials)

Zehao Su[1] **Leonard Henckel**[2]

[1]Section of Biostatistics, Department of Public Health, University of Copenhagen
[2]Department of Mathematical Sciences, University of Copenhagen

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 $(X, \dots, Y)$ 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 $(X, Z, \dots, Y)$ plus an edge between Y and X. A *directed cycle* is a directed path $(X, Z, \dots, Y)$ from X to Y plus an edge $Y \rightarrow X$. Given a path $p = (V_1, \dots, V_k)$, let $p(V_i, V_j)$, $i < j$ denote the path segment from V_i to V_j and let $-p = (V_k, \dots, V_1)$. Given two paths $p = (V_1, \dots, V_k)$ and $q = (V_k, \dots, V_q)$, let $p \oplus q = (V_1, \dots, V_k, \dots, V_q)$. We call any node V_i on a path $p = (V_1, \dots, V_k)$ 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 $\mathcal{G}$.

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

B PROOFS

B.1 PROOF OF THEOREM 1

Proof. Consider the set $\mathbf{Z}_i$. By assumption, $\text{forb}(X, Y, \mathcal{G}_0) \cap \mathbf{Z}_i = \emptyset$ 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 $\mathcal{G}_0$ that is open given $\mathbf{Z}_i$. Suppose that p is of the form $X \to C \leftarrow Y$. By assumption $Y \in \text{de}(X, \mathcal{G}_0)$ and therefore $Y \in \text{forb}(X, Y, \mathcal{G}_0)$ which in turn implies that $\text{de}(Y, \mathcal{G}_0) \subseteq \text{forb}(X, Y, \mathcal{G}_0)$. As a result, $C \in \text{forb}(X, Y, \mathcal{G}_0)$. But as $\mathbf{Z}_i \cap \text{forb}(X, Y, \mathcal{G}_0) = \emptyset$ it follows that $\text{de}(C, \mathcal{G}_0) \cap \mathbf{Z}_i = \emptyset$, which contradicts our assumption that p is open given $\mathbf{Z}_i$. We can therefore assume that p is not of the form $X \to C \leftarrow Y$ which implies that p must contain at least one non-collider. If every non-collider on p is in $\text{forb}(X, Y, \mathcal{G}_0)$, it follows that every node on p is in $\text{forb}(X, Y, \mathcal{G}_0)$. 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 $\text{forb}(X, Y, \mathcal{G}_0)$. But as $(\mathbf{V} \setminus \text{forb}(X, Y, \mathcal{G}_0)) \subseteq \cup_{j=1}^{k} \mathbf{Z}_j$ 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 therefore shown that there exists a linear structural equation model compatible with $\mathcal{G}$, such that $\beta_{yx.\mathbf{z}_i} - \beta_{yx.\mathbf{z}_j} \neq 0$.

Consider now the term $\beta_{yx.\mathbf{z}_i} - \beta_{yx.\mathbf{z}_j}$ as a function in the edge coefficients and error variances from the underlying linear structural equation model. By the same arguments as given in Section 13.3 of Spirtes et al. [2000] the function $\beta_{yx.\mathbf{z}_i} - \beta_{yx.\mathbf{z}_j}$ is equivalent to a polynomial in the edge coefficients and error variances of the linear structural equation model. As we have shown that there exists one linear structural equation model such that $\beta_{yx.\mathbf{z}_i} - \beta_{yx.\mathbf{z}_j} \neq 0$, this polynomial is non-trivial. Our claim then follows from the fact that the zero set of non-trivial polynomials has Lebesgue measure 0. □

B.2 PROOF OF LEMMA 3

Lemma 1 (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 2 (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 2 that for every set $\mathbf{Z}_i$, $i = 1, 2, \ldots, k$,

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

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 (1).

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

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 1 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 (1).

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 1.

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

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

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

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

The target quantity can then be written as

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

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 $\mathbf{\Delta}_{\mathcal{Z},ii} = \mathrm{E}(\delta^2_{x\mathbf{z}_i}\delta^2_{y\mathbf{z}'_i}) = \mathrm{E}(\delta^2_{x\mathbf{z}_i})\,\mathrm{E}(\delta^2_{y\mathbf{z}'_i})$ 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. [2022]). Therefore, the corresponding terms are $\mathbf{\Sigma}_{\mathcal{Z},ii} = \mathrm{E}(\delta^2_{y\mathbf{z}'_i})/\,\mathrm{E}(\delta^2_{x\mathbf{z}_i})$. 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,

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

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 (2) 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 3 (Consistency of plug-in estimator of $\mathbf{\Sigma}_{\mathcal{Z}}$)**.** *Consider the setting in Lemma 3. The plug-in estimator $\hat{\mathbf{\Sigma}}_{\mathcal{Z}}$ of $\mathbf{\Sigma}_{\mathcal{Z}}$ with entries*

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

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

Proof of Lemma 3. 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

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

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.

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

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.

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

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 $\hat{\mathbf{\Sigma}}_{\mathcal{Z}}$, after subtracting their true values in $\mathbf{\Sigma}_{\mathcal{Z}}$, will converge to a zero-mean normal distribution.

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

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

The numerator N of the expression above is expanded as

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

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 3. 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 apply due to Slutsky's Theorem and the assumption that the fourth moments of the errors are finite.

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 $\mathrm{vech}(\boldsymbol{\Sigma}_{\mathcal{Z}})$ to the (i,j)-th entry of $\boldsymbol{\Sigma}_{\mathcal{Z}}$. The asymptotic covariance matrix $\mathbf{F}$ of $\mathrm{vech}(\hat{\boldsymbol{\Sigma}}_{\mathcal{Z}})$ 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 $\mathrm{vech}(\hat{\boldsymbol{\Delta}}_{\mathcal{Z}})$ follows from the linear relationship $\mathrm{vech}(\boldsymbol{\Delta}_{\mathcal{Z}}) = \boldsymbol{\Pi}\mathrm{vech}(\boldsymbol{\Sigma}_{\mathcal{Z}})$. □

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

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

The numerator

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

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 3. 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,

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

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 $\boldsymbol{\Delta}_{\mathcal{Z}} = \boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\mathcal{Z}}\boldsymbol{\Gamma}^\top$.

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

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

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

$$\begin{aligned} n(\boldsymbol{\Gamma}\hat{\boldsymbol{\beta}}_{yx.\mathcal{Z}})^\top \boldsymbol{\Delta}_{\mathcal{Z}}^\dagger (\boldsymbol{\Gamma}\hat{\boldsymbol{\beta}}_{yx.\mathcal{Z}}) &= \sum_{s=1}^{r_0} \phi_s^{-1} (\mathbf{q}_s^\top n^{1/2} \boldsymbol{\Gamma}\hat{\boldsymbol{\beta}}_{yx.\mathcal{Z}})^2 \\ &\xrightarrow{d} \sum_{s=1}^{r_0} \phi_s^{-1} (\mathbf{q}_s^\top \mathbf{G})^2 \sim \chi^2_{r_0}. \end{aligned} \tag{3}$$

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 $\hat{\boldsymbol{\Delta}}_{\mathcal{Z}}$ is positive semidefinite, its spectral decomposition is $\hat{\mathbf{P}}\hat{\boldsymbol{\Lambda}}\hat{\mathbf{P}}^\top$, where $\hat{\boldsymbol{\Lambda}} = \operatorname{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_k)$ with $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_k \geq 0$. The rank-$\hat{r}$ spectral approximation of $\hat{\boldsymbol{\Delta}}_{\mathcal{Z}}$ is then $\hat{\mathbf{P}}\hat{\boldsymbol{\Lambda}}_{\hat{r}}\hat{\mathbf{P}}^\top$, where $\hat{\boldsymbol{\Lambda}}_{\hat{r}} = \operatorname{diag}(\hat{\lambda}_{\hat{r},1}, \ldots, \hat{\lambda}_{\hat{r},\hat{r}}, 0, \ldots, 0)$. Following Weyl's inequality [Stewart, 1998] and Proposition 7, we have $\hat{\boldsymbol{\Lambda}} \xrightarrow{p} \boldsymbol{\Lambda}$ since the asymptotic covariance matrix of $\operatorname{vech}(\hat{\boldsymbol{\Delta}}_{\mathcal{Z}})$ is finite. We now show that $\hat{\boldsymbol{\Lambda}}_{\hat{r}} \xrightarrow{p} \boldsymbol{\Lambda}$. 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, $\hat{\boldsymbol{\Delta}}_{\mathcal{Z},\hat{r}} - \hat{\boldsymbol{\Delta}}_{\mathcal{Z}} = \hat{\mathbf{P}}(\hat{\boldsymbol{\Lambda}}_{\hat{r}} - \boldsymbol{\Lambda})\hat{\mathbf{P}}^\top \xrightarrow{p} 0$. Then $\hat{\boldsymbol{\Delta}}_{\mathcal{Z},\hat{r}} \xrightarrow{p} \boldsymbol{\Delta}_{\mathcal{Z}}$ by consistency of $\hat{\boldsymbol{\Delta}}_{\mathcal{Z}}$.

The rank of $\hat{\boldsymbol{\Delta}}_{\mathcal{Z},\hat{r}}$ is equal to $\hat{r}$ by construction. With the condition that $\mathbb{P}(\operatorname{rank}(\hat{\boldsymbol{\Delta}}_{\mathcal{Z},\hat{r}}) = \operatorname{rank}(\boldsymbol{\Delta}_{\mathcal{Z}})) \to 1$, it follows from Theorem 2 in Andrews [1987] that $\hat{\boldsymbol{\Delta}}_{\mathcal{Z},\hat{r}}^\dagger \xrightarrow{p} \boldsymbol{\Delta}_{\mathcal{Z}}^\dagger$. By Slutsky's theorem, the convergence in distribution in (3) still holds if we use a consistent estimator $\hat{\boldsymbol{\Delta}}_{\mathcal{Z},\hat{r}}^\dagger$ of $\boldsymbol{\Delta}_{\mathcal{Z}}^\dagger$ instead. Therefore, $n(\boldsymbol{\Gamma}\hat{\boldsymbol{\beta}}_{yx.\mathcal{Z}})^\top \hat{\boldsymbol{\Delta}}_{\mathcal{Z},\hat{r}}^\dagger (\boldsymbol{\Gamma}\hat{\boldsymbol{\beta}}_{yx.\mathcal{Z}}) \xrightarrow{d} \chi^2_{r_0}$. □

B.5 PROOF OF LEMMA 8

Lemma 4 (Modified Lemma D.1 in Henckel et al. [2022])**.** *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 5 (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 6. *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 7. *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. [2022]. 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 $\delta_{x\mathbf{z}_1} = X - \boldsymbol{\beta}_{x\mathbf{z}_1}^\top \mathbf{Z}_1$, 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 6. It follows from Corollary 8 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

$$N = \frac{1}{\alpha_1 \beta_{xn.\mathbf{z}_{1,-n}}} \left[\alpha_1 \left(X - \sum_{V \in \mathbf{Z}_1 \setminus \{N\}} \beta_{xv.\mathbf{z}_{1,-v}} V \right) + \sum_{i \neq 1} \alpha_i (X - \boldsymbol{\beta}_{x\mathbf{z}_i}^\top \mathbf{Z}_i) \right] = \sum_{V \neq N} \gamma_v V, \tag{4}$$

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 (4) 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 9 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. [2022]). 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

$$\epsilon_Y \sum_i \xi_i \delta_{x\mathbf{z}_i} = 0, \tag{5}$$

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 $\mathbf{\Sigma}_{\mathcal{Z}}$ 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 $\mathbf{\Sigma}_{\mathcal{Z}}$ is invertible our claim follows. □

B.6 LEMMA 10 AND ITS PROOF

Lemma 8. *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

C.1 SIMULATION IN EXAMPLE 9

The definition of the probability-probability plot that we employ in Example 6 is described as follows. Given a sample of p-values $p_1, p_2, \ldots, p_R$, we sort them in the increasing order: $p_{(1)}, \ldots, p_{(R)}$. Then we apply the empirical distribution function to get the empirical probabilities $\hat{P}_{(j)}$ for $j = 1, \ldots, R$, i.e., $\hat{P}_{(j)} = \sum_{i=1}^{R} I(p_{(i)} \leq p_{(j)})/R$. These are simply j/R assuming no ties. Since we wish to compare the sample to the standard uniform distribution, whose cumulative distribution function is $F(t) = t$ for $t \in [0, 1]$, we compute the theoretical probabilities $P_{(j)} = F(p_{(j)}) = p_{(j)}$. The plot is finally obtained by plotting $\hat{P}_{(j)}$ against $P_{(j)}$.

C.2 SIMULATION IN SECTION 4

True graph We generate causal DAGs as Erdős–Rényi random graphs. There are in total 50 DAGs with 10 nodes and 50 DAGs with 15 nodes. The expected neighbourhood size for each DAG is 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 $\mathcal{G}_0$.

Factor	Strategy	$S = \text{Min+}$			$S = \text{All}$		
	Hypothesis	H_0^*	$\neg H_0^* \wedge H_0$	$\neg H_0$	H_0^*	$\neg H_0^* \wedge H_0$	$\neg H_0$
Expected graph accuracy	Low	42.82	1.98	55.20	36.14	1.98	61.88
	High	85.64	5.48	8.88	84.71	5.27	10.02
Graph size	10	81.12	3.08	15.8	77.73	3.08	19.19
	15	55.24	7.46	37.3	54.31	6.99	38.69
Neighbourhood size	2	92.18	0.00	7.82	88.83	0.00	11.17
	3	93.31	1.49	5.20	88.61	1.24	10.15
	4	63.14	6.34	30.53	62.56	6.34	31.10
	5	54.52	7.47	38.01	52.49	7.24	40.27

Table 1: Percentage of true hypotheses in the simulation normalised within each combination of factor and strategy.

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 $Y \notin \text{de}(X, \mathcal{G})$ 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 and sample size for testing n, 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.3 EXTRA SIMULATION RESULTS

Figure 1 and Figure 2 show additional plots of the AUCs from the simulation study. In Figure 1 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. In Figure 1 they are additionally grouped by the sample size used for testing and the candidate graph accuracy. 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.

Table 1 summarises the proportions of candidate graphs (and strategies) where the null-hypothesis H_0^* is true, the null hypothesis H_0^* is false but the actual test null hypothesis H_0 is true and both are false, respectively. Unsurprisingly H_0^* is true more often for the high accuracy candidate graphs. We can also see that the strategy $S = \text{All}$ always result in a higher proportions of cases where H_0 is false when compared to $S = \text{Min+}$, which is due to the fact that $S = \text{Min+}$ consider a subset of the adjustment sets $S = \text{All}$ considers. The problematic cases where $\neg H_0^* \wedge H_0$ generally occur in around 10% of the cases, and interestingly are more common for the larger graphs than for the smaller graphs.

Minimal adjustment sets in large sparse graphs We ran a small simulation to demonstrate the scalability of the algorithm for minimal adjustment sets proposed by Van der Zander et al. [2014]. We simulated Erdős–Rényi graphs with graph size $100, 250, 500, 1000, 2500, 5000$ and expected neighbourhood size $2, 3, 4, 5$. For each combination above, we generated

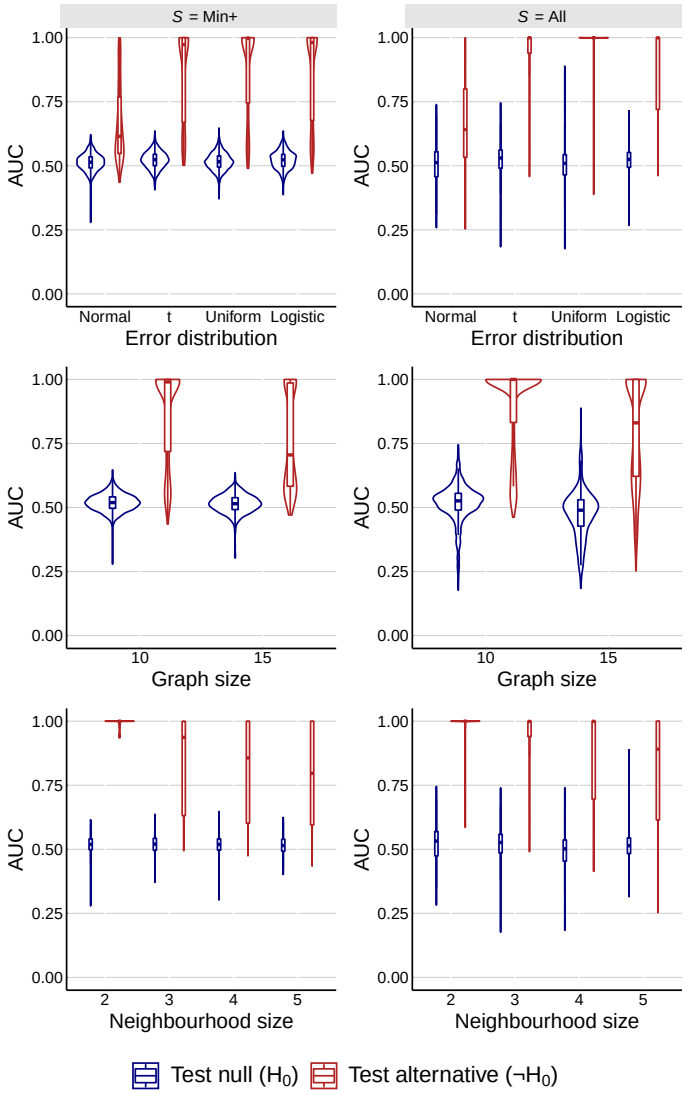

Figure 1: Extra violin plots (layered with boxplots) of AUCs from the simulation study.

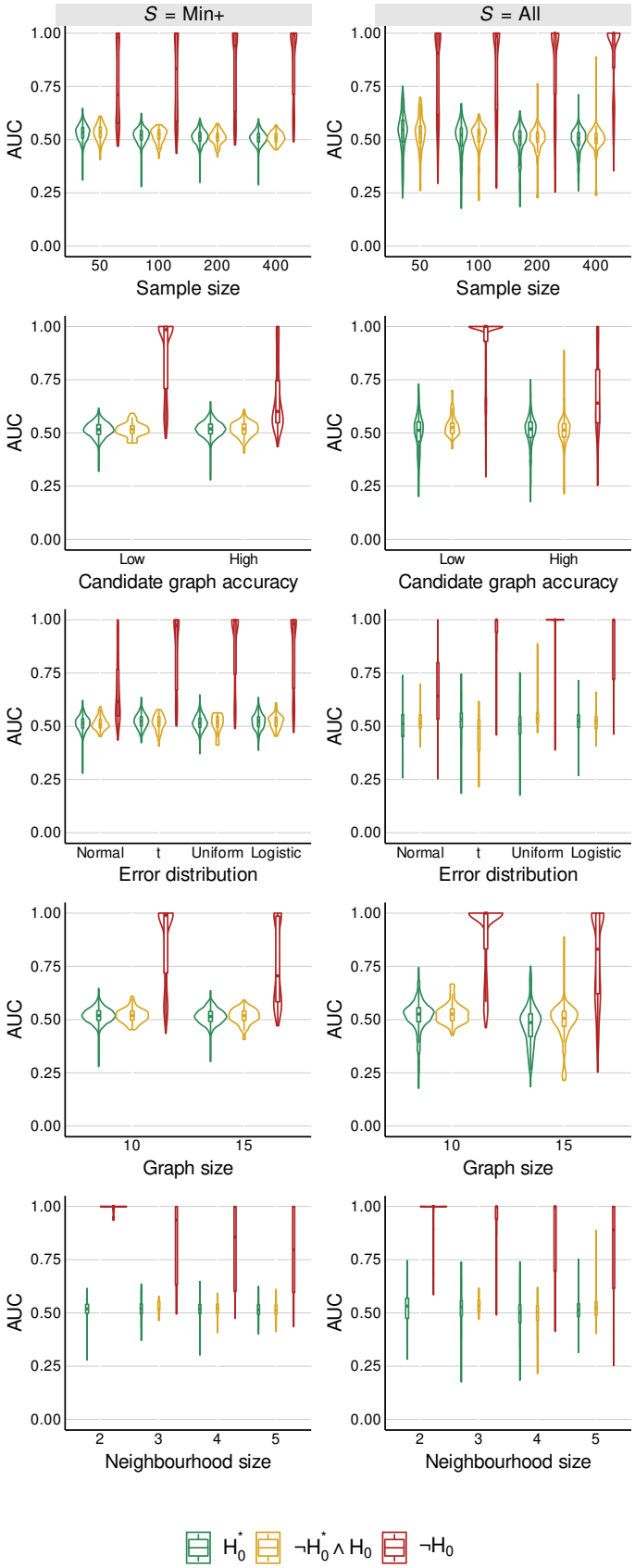

Figure 2: Extra violin plots (layered with boxplots) of AUCs from the simulation study, partitioning H_0 into H_0^* and $\neg H_0^* \wedge H_0$.

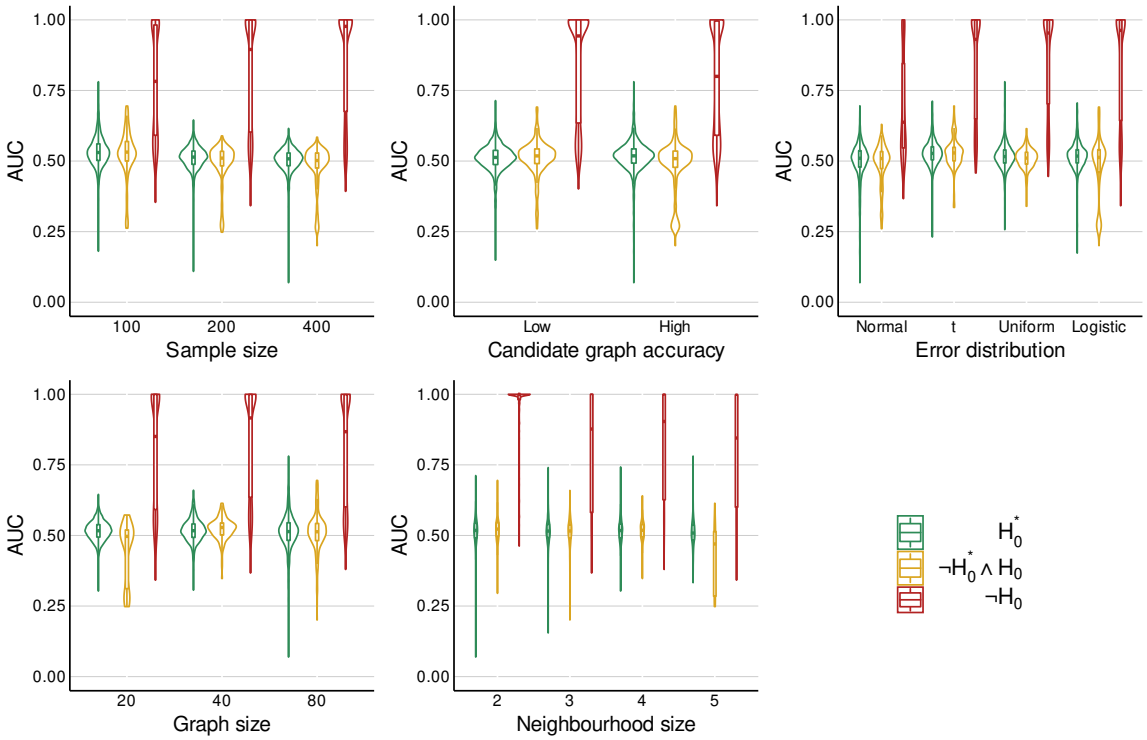

Figure 3: Violin plots (layered with boxplots) of AUCs from the simulation study using only the Min+ strategy, partitioning H_0 into H_0^* and $\neg H_0^* \wedge H_0$.

Cand. graph accuracy	n	H_0^*		$\neg H_0^* \wedge H_0$		$\neg H_0$	
		$S = \text{Min+}$	$S = \text{All}$	$S = \text{Min+}$	$S = \text{All}$	$S = \text{Min+}$	$S = \text{All}$
Low	50	0.0751	0.0909	0.0788	0.0759	0.5570	0.7396
	100	0.0636	0.0636	0.0587	0.0385	0.6352	0.7880
	200	0.0543	0.0510	0.0488	0.0516	0.7132	0.8341
	400	0.0493	0.0466	0.0525	0.0503	0.7887	0.8812
High	50	0.0786	0.0897	0.0711	0.0712	0.1543	0.1697
	100	0.0634	0.0587	0.0585	0.0557	0.2026	0.2094
	200	0.0559	0.0500	0.0558	0.0476	0.2838	0.3010
	400	0.0543	0.0471	0.0492	0.0475	0.3838	0.4152

Table 2: Proportion of hypotheses rejected at level 0.05 in the simulation study.

10 DAGs. For each DAG, we selected the pair of (X, Y) nodes in the same way as in the main simulation described in Section 4. We then ran the algorithm to extract minimal adjustment sets relative to (X, Y) and performed the rest of the testing procedure according to Algorithm 1. We allowed up to one hour on each DAG to finish the computation of minimal adjustment sets, and for the graph sizes $100, 250, 500, 1000, 2500, 5000$, the percentages of completed algorithm runs were $100\%, 57.5\%, 92.5\%, 100\%, 95\%, 35\%$, respectively. The results suggest that the extraction of minimal adjustment sets is possible even for graphs with sizes in the order of 1000s. We also noted, however, that the space required to store the adjustment sets can also exceed the 4 GB RAM allocated.

Min+ **strategy-only simulation on larger graphs** We conducted another simulation on graphs of size $20, 40$ and 80 with precisely the same setup as the simulation in Section 4 using only the Min+ strategy. As the Min+ strategy is computationally much fast than the All strategy we, we were able to increase the graph sizes while keeping the other configurations unchanged. It is worth pointing out that attempting to run the simulation on graphs of 20 nodes with the All strategy in the same setup almost always exceeded the one-hour timeout. Figure 3 contains violin plots of AUCs framed by different parameters used in the simulation and coloured by their respective true hypotheses. The results are very similar to

what we saw in the simulation in Section 4. The small bulks around AUC 0.25 to 0.3 for $\neg H_0^* \wedge H_0$ in Figure 3 are due to a specific DAG and structural equation model where our procedure was very conservative. One particular simulated graph of size 80 was not included in the plots due to memory overflow during the computation of the minimal adjustment sets, which indicated that for graphs larger than 80, memory might have to be taken into account.

References

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

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.

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

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

Leonard Henckel, Emilija Perković, and Marloes H. Maathuis. Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(2): 579–599, 2022.

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.

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.

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

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.

Benito Van der Zander, Maciej Liskiewicz, and Johannes Textor. Constructing separators and adjustment sets in ancestral graphs. In *Proceedings of the UAI 2014 Conference on Causal Inference: Learning and Prediction*, CI'14, pages 11–24, 2014.