Supplementary Material for “Statistical test for consistent estimation of causal effects in linear non-Gaussian models”

Doris Entner[*], **Patrik O. Hoyer**[*], **Peter Spirtes**[**]
[*]HIIT and Department of Computer Science, University of Helsinki
[**]Department of Philosophy, Carnegie Mellon University

Abstract

This document contains supplementary material to the article ‘Statistical test for consistent estimation of causal effects in linear non-Gaussian models’, AISTATS 2012. A table of contents is given below.

Contents

1 Definitions and background

In this section we give some standard graph-specific terms, and the definitions of blocked and active paths, and back-door paths, and we restate the back-door criterion (Pearl, 2009), which is a graphical criterion for when adjustment is guaranteed to yield a consistent estimate of a causal effect from observational data.

A directed graph $\mathcal{G}$ is a pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with $\mathcal{V}$ a set of variables and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ a set of edges. If an ordered pair $(v_i, v_j) \in \mathcal{E}$, we write $v_i \to v_j$ and say that v_i is a parent of v_j, and v_j is a child of v_i. A path is a sequences of nodes $(v_1, \ldots, v_i, v_{i+1}, \ldots, v_m)$ such that for each pair of consecutive nodes either $(v_i, v_{i+1}) \in \mathcal{E}$ or $(v_{i+1}, v_i) \in \mathcal{E},\ i = 1, \ldots, m-1$. A directed path from v_1 to v_m is a path where all arrows point towards v_m ($v_1 \to \ldots v_i \to v_{i+1} \ldots \to v_m$), i.e. $(v_i, v_{i+1}) \in \mathcal{E}$ for all $i = 1, \ldots, m-1$. If there is a directed path from v_i to v_j, then v_i is called an ancestor of v_j, and v_j a descendant of v_i. A directed acyclic graph (DAG) is a directed graph that does not contain any directed paths from a node to itself.

A path is *blocked* (*d-separated*) by a variable set $\mathcal{Z}$ if the path contains (i) a triple of the form $v_i \to v_k \to v_j$ or $v_i \leftarrow v_k \to v_j$ with $v_k \in \mathcal{Z}$ or (ii) a triple $v_i \to v_l \leftarrow v_j$ (called a *collider*) with neither v_l nor any descendant of v_l in $\mathcal{Z}$. A path which is not blocked is called an *active* path. A set $\mathcal{Z} \subseteq \mathcal{V} \setminus \{v_i, v_j\}$ *d-separates* two variables v_i and v_j if $\mathcal{Z}$ blocks all paths between v_i and v_j (Pearl, 2009, Ch. 1).

A *back-door path* from v_i to v_j is a path leaving v_i via one of its parents, i.e. the first edge in the path is $v_i \leftarrow v_k$ for some k.

Given a DAG over a set of random variables (including x, y and $\mathcal{Z}$), the *back-door criterion* (Pearl, 2009, Ch. 3) states that the total causal effect of x on y is identifiable from observational data if there exists a set $\mathcal{Z}$ of observed variables such that no variable in $\mathcal{Z}$ is a descendant of x, and $\mathcal{Z}$ blocks (d-separates) all back-door paths from x to y. A set $\mathcal{Z}$ satisfying the back-door criterion is called *admissible* (Pearl, 2009, Ch. 3) (or 'sufficient', Greenland et al., 1999). In the linear case, the causal effect of x on y can be consistently estimated by including an admissible set $\mathcal{Z}$ in the regression of y on x, and the causal effect is obtained as the regression coefficient of x (Pearl, 2009, Ch. 5).

2 Detailed proof of Theorem 1

To recap, the model over the variable set $\mathcal{V} = \{x, y\} \cup \mathcal{W} \cup \mathcal{U}$ defined in Equation (1) of Section 2.2 of the paper is given by

$$\boldsymbol{v} := \mathbf{B}\boldsymbol{v} + \boldsymbol{e} \tag{1}$$

where $\mathbf{B}$ can be permuted to lower triangularity (acyclicity assumption) and the disturbance terms in $\boldsymbol{e}$ are mutually independent. We assume that the covariates in $\mathcal{W}$ precede the 'treatment' variable x which in turn precedes the 'outcome' variable y.

We here provide a detailed proof of the main result (Theorem 1), which is used to judge the outcome of the procedure of Algorithm 1. Both are restated for convenience.

Theorem 1. *Given the model of Equation* (1) *and using the procedure described in Algorithm 1 with a fixed conditioning set $\mathcal{Z}$, the following statements hold:*

(a) Under the assumption that the disturbance term *of x, denoted by e_x, has a non-Gaussian distribution, if the residuals r_x and r_y are asymptotically mutually independent, then a is a consistent estimator of the true causal effect α.*

(b) Under the assumption that the distribution over the variables in $\mathcal{V}$ is linearly faithful to the generating DAG (Spirtes et al., 2000, p. 47), if the residual *r_x is non-Gaussian, and r_x and r_y are asymptotically mutually independent, then a is a consistent estimator of the true causal effect α.*

(c) If the residuals r_x and r_y are asymptotically statistically dependent, then the set $\mathcal{Z}$ is not admissible.

Algorithm 1 (Statistical test for consistency)

Given a dataset over the observed variables $\{x, y\} \cup \mathcal{W}$ (from a model as in Equation (1)) and a set $\mathcal{Z} \subseteq \mathcal{W}$ of covariates to adjust for.

Estimate the following two regressions using ordinary least squares (OLS):

$$x = \boldsymbol{b}^T \boldsymbol{z} + r_x \tag{2}$$

$$y = ax + \boldsymbol{c}^T \boldsymbol{z} + r_y. \tag{3}$$

Perform a statistical test for Gaussianity of the residual r_x, proceed only if Gaussianity is rejected.

Perform a statistical test for independence of r_x and r_y, interpret the result as follows (Theorem 1):

```
    If independence is *not rejected* at threshold $p_t$:
        The estimated effect $a$ is inferred to be consistent
    If independence is *rejected* at threshold $p_t$:
        The estimated effect $a$ is inferred to be inconsistent
```

In the proofs we will use the 'reduced form' representation of the model in Equation (1), given by

$$\boldsymbol{v} = \mathbf{A}\boldsymbol{e} \tag{4}$$

with $\mathbf{A} = (\mathbf{I} - \mathbf{B})^{-1}$ containing the total effects of the disturbances $\boldsymbol{e}$ on the variables $\boldsymbol{v}$. That is, each variable $v \in \mathcal{V}$ can be written as a linear combination of the disturbance terms:

$$v = \mathbf{A}_v \boldsymbol{e} \tag{5}$$

with $\mathbf{A}_v$ being the corresponding row of matrix $\mathbf{A}$.

Furthermore, we use a property about dependence and independence of two sums of independent random variables, known as the Darmois-Skitovitch Theorem:

Darmois-Skitovitch Theorem (Darmois, 1953; Skitovitch, 1953). Let $e_1, \ldots, e_n$ be independent random variables ($n \geq 2$), $s_1 = \beta_1 e_1 + \ldots + \beta_n e_n$ and $s_2 = \gamma_1 e_1 + \ldots + \gamma_n e_n$ with constants $\beta_i, \gamma_i,\ i = 1, \ldots, n$. If s_1 and s_2 are independent, then those e_j which influence both sums s_1 and s_2 (i.e. $\beta_j \gamma_j \neq 0$) are Gaussian.

Proof of Theorem 1(a). We show that if a is an inconsistent estimator of the true causal effect α then the residuals r_x and r_y are dependent.

Let the variables in $\mathcal{V}$ be arranged in a causal order (i.e. no "later" variable has an effect on any "earlier" variable in this order) such that $\boldsymbol{v} = (v_1, \ldots, v_{n-2}, x, y)^T$ and the disturbances $\boldsymbol{e} = (e_{v_1}, \ldots, e_{v_{n-2}}, e_x, e_y)^T$ which implies that the matrix $\mathbf{A}$ in Equation (4) is lower triangular with a unit diagonal (possible by the acyclicity and partial temporal ordering assumption).

Using Equation (5) we can express the variable x as a linear combination of the underlying disturbances, $x = \mathbf{A}_x \boldsymbol{e}$, where $\mathbf{A}_x = (A_{x,1}, \ldots, A_{x,n-2}, 1, 0)$, meaning that the coefficient of e_x is equal to 1, the coefficient of e_y is equal to 0, and the coefficients $A_{x,i},\ i = 1, \ldots, n-2$, from the other disturbances can be either zero or non-zero, depending on the graph and the parameters of the model. Similarly, y is represented by $y = \mathbf{A}_y \boldsymbol{e}$ with $\mathbf{A}_y = (A_{y,1}, \ldots, A_{y,n-2}, \alpha, 1)$ with a coefficient of strength α for e_x, unit coefficient for e_y and zero or non-zero coefficients $A_{y,i},\ i = 1, \ldots, n-2$, for the other disturbances. The effect α of e_x on y is the same as the true causal effect of x on y because of the acyclicity assumption. Furthermore, for all other variables $v_i,\ i = 1, \ldots, n-2$, the representation $v_i = \mathbf{A}_{v_i} \boldsymbol{e}$ is such that the coefficients of e_x and e_y are zero (because of the lower triangularity of $\mathbf{A}$).

Using the above we can express the residuals from Equations (2) and (3) in terms of the disturbances:

$$r_x = x - \boldsymbol{b}^T \boldsymbol{z} = (\mathbf{A}_x - \sum_{z \in \mathcal{Z}} b_z \mathbf{A}_z)\, \boldsymbol{e} = (\ldots, 1, 0)\, \boldsymbol{e}$$

$$r_y = y - ax - \boldsymbol{c}^T \boldsymbol{z} = (\mathbf{A}_y - a\mathbf{A}_x - \sum_{z \in \mathcal{Z}} c_z \mathbf{A}_z)\, \boldsymbol{e} = (\ldots, \alpha - a, 1)\, \boldsymbol{e}$$

where the dots indicate the entries of the disturbances other than e_x and e_y. Given the premise that the estimator a is inconsistent (i.e. $(a - \alpha) \overset{P}{\nrightarrow} 0$) the disturbance e_x has a non-vanishing coefficient in the representation of *both* residuals r_x and r_y. By the assumed non-Gaussianity of e_x the Darmois-Skitovitch theorem ensures that r_x and r_y are statistically dependent. □

For the proofs of Theorem 1(b) and (c) we make use of the following lemma, which gives a criterion when a disturbance variable e_w has a non-vanishing coefficient in the representation of a residual r_v.

Lemma 1. *We are given a set of variables $\mathcal{V} = v \cup \mathcal{V}'$, where v is a single variable and $\mathcal{V}'$ a non-empty set of variables not including v, following the model in Equation* (1)*, and we assume that the distribution over these variables is linearly faithful to the generating DAG. Regressing v on a set $\mathcal{Z}' \subseteq \mathcal{V}'$ not containing any descendants of v yields*

$$\begin{aligned} r_v = v - \sum_{z \in \mathcal{Z}'} \hat{c}_z z &= \mathbf{A}_v \boldsymbol{e} - \sum_{z \in \mathcal{Z}'} \hat{c}_z \mathbf{A}_z \boldsymbol{e} = \boldsymbol{d}_v \boldsymbol{e} \\ &= (d_{v,1}, \ldots, d_{v,w}, \ldots, d_{v,n})(e_1, \ldots, e_w, \ldots, e_n)^T \end{aligned}$$

with $\mathbf{A}_v$ and $\mathbf{A}_z,\ z \in \mathcal{Z}'$, as in Equation (5) *and $\boldsymbol{d}_v = \mathbf{A}_v - \sum_{z \in \mathcal{Z}'} \hat{c}_z \mathbf{A}_z$. When estimating the regression using OLS, for $w \in \mathcal{V}'$ holds that $d_{v,w} \overset{P}{\nrightarrow} 0$ (with $d_{v,w}$ the coefficient of e_w in $\boldsymbol{d}_v$) if and only if*

1. *for $w \in \mathcal{Z}'$ there is an active back-door path (not blocked by $\mathcal{Z}' \setminus \{w\}$) from w to v pointing into v (i.e. the last edge on the path is "$\rightarrow v$"),*
2. *for $w \in \mathcal{V}' \setminus \mathcal{Z}'$ there is*
   (i) a directed active path from w to v (not blocked by $\mathcal{Z}'$) or
   (ii) a directed active path from w to some $z \in \mathcal{Z}'$ (not blocked by $\mathcal{Z}' \setminus \{z\}$) for which there is an active back-door path to v (not blocked by $\mathcal{Z}' \setminus \{z\}$) pointing into v.

Note that the linear faithfulness assumption is only necessary for the "if-direction", the "only-if-statement" is valid without this assumption as can be seen from the proof.

Proof of Lemma 1. We will use in both directions of the proof the fact that points 1 and 2, respectively, are equivalent to e_w being d-connected to v given $\mathcal{Z}'$, by a path pointing into v (for $w \in \mathcal{Z}'$ and $w \in \mathcal{V}' \setminus \mathcal{Z}'$, respectively), which follows straight from the definition of d-separation. Furthermore, in points 1 and 2(ii) any potentially active back-door path is pointing into v, since any other back-door path contains a collider at some descendant of v, which is not in the conditioning set (by assumption) and hence such paths are blocked.

"$\Leftarrow$" We show that if either point 1 or 2 holds, then $d_{v,w} \overset{P}{\nrightarrow} 0$.

First we note that for any $w \in \mathcal{Z}'$ which has no parents, i.e. $w = e_w$, point 1 never holds.

For any other disturbance variable e_w (independent of whether w is in the conditioning set $\mathcal{Z}'$), if point 1 or 2 is fulfilled, we know that e_w is not d-separated from v given $\mathcal{Z}'$ which implies using the faithfulness assumption that the partial correlation of e_w and v given $\mathcal{Z}'$ is non-zero. Hence, in the regression $v = \sum_{z \in \mathcal{Z}'} \tilde{c}_z z + b e_w + \tilde{r}_v$ (which has non-collinear regressors since $e_w \neq z, z \in \mathcal{Z}'$) the coefficient $b \overset{P}{\nrightarrow} 0$. We now show that in the regression $v = \sum_{z \in \mathcal{Z}'} \hat{c}_z z + r_v$ the contribution of the disturbance term e_w to the residual r_v is non-vanishing. We rewrite the regressions in matrix form, with $\hat{\mathbf{c}}$ and $\tilde{\mathbf{c}}$ collecting the coefficients $\hat{c}_z$ and $\tilde{c}_z,\ z \in \mathcal{Z}'$, respectively, $\mathbf{Z}$ the data matrix over the variables in $\mathcal{Z}'$ and E_w the data vector of the disturbance term e_w:

$$\begin{aligned} v &= \mathbf{Z}\hat{\mathbf{c}} + r_v \\ v &= \mathbf{Z}\tilde{\mathbf{c}} + E_w b + \tilde{r}_v \end{aligned}$$

The coefficients $\tilde{\mathbf{c}}$ of the second regression can be expressed in terms of the coefficients $\hat{\mathbf{c}}$ and b as $\tilde{\mathbf{c}} = \hat{\mathbf{c}} - (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T E_w b$ (Seber and Lee, 2003, p. 54), which yields for the second regression

$$\begin{aligned} v &= \mathbf{Z}(\hat{\mathbf{c}} - (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T E_w b) + E_w b + \tilde{r}_v \\ &= \mathbf{Z}\hat{\mathbf{c}} + (I - \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T)E_w b + \tilde{r}_v. \end{aligned}$$

It follows that the residual r_v of the first regression is equal to

$$r_v = (I - \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T)E_w b + \tilde{r}_v.$$

The (1x1) matrix $E'_w(I - \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T)E_w$ is positive definite (Seber and Lee, 2003, p. 54), which for the scalar case means it is a positive number, and hence the vector $(I - \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T)E_w$ is non-zero. Since $b \overset{P}{\nrightarrow} 0$, the residual r_v has a non-vanishing contribution from the disturbance term e_w, i.e. $d_{v,w} \overset{P}{\nrightarrow} 0$.

"⇒" We show that if neither point 1 nor 2 holds, then $d_{v,w} \xrightarrow{P} 0$.

First, for any $w \in \mathcal{Z}'$ with $w = e_w$ point 1 never holds. In this case, since in the regression of v on the variables in $\mathcal{Z}'$ the regressors are uncorrelated with the residual r_v, we obtain $0 = cov(w, r_v) = cov(e_w, r_v) = E(e_w\, \boldsymbol{d}_v \boldsymbol{e}) = \sum_{i=1}^n d_{v,i}\, E(e_w\, e_i)$ and $E(e_w\, e_i) \xrightarrow{P} 0$ for $w \neq i$ (since all e_i are independent). Thus, $\sum_{i=1}^n d_{v,i}\, E(e_w\, e_i) \xrightarrow{P} d_{v,w}E(e_w\, e_w) = d_{v,w}V(e_w) \xrightarrow{P} 0$, implying that $d_{v,w} \xrightarrow{P} 0$.

For any other disturbance variable e_w (independent of whether w is in the conditioning set $\mathcal{Z}'$) we know that the negation of points 1 and 2 imply that e_w is d-separated from v given $\mathcal{Z}'$, which in the linear model family implies that the partial correlation of e_w and v given $\mathcal{Z}'$ is zero. Thus, in the regression $v = \sum_{z \in \mathcal{Z}'} \tilde{c}_z z + b e_w + \tilde{r}_v$ (which has non-collinear regressors since $e_w \neq z, z \in \mathcal{Z}'$) the regression coefficient $b \xrightarrow{P} 0$, and hence for the regression $v = \sum_{z \in \mathcal{Z}'} \hat{c}_z z + r_v$ we get that for all $z \in \mathcal{Z}'$ the coefficients $\tilde{c}_z$ and $\hat{c}_z$ converge in probability to the same value and hence also $\tilde{r}_v$ and r_v. Because $0 = cov(e_w, \tilde{r}_v) = cov(e_w, r_v)$ we can, as above, conclude that $d_{v,w} \xrightarrow{P} 0$. □

Proof of Theorem 1(b). As in the proof of Theorem 1(a), we show that if the estimator is inconsistent, then the residuals are dependent.

First, the inconsistent estimator for the effect of x on y implies that there is an active back-door path from x to y (not blocked by the set $\mathcal{Z}$) pointing into y.

Additionally, since the residual r_x from Equation (2) is non-Gaussian, when expressing $r_x = x - \boldsymbol{b}^T\boldsymbol{z} = \boldsymbol{d}_x\boldsymbol{e}$ as a linear combination of the disturbances $\boldsymbol{e}$, the coefficient of at least one non-Gaussian residual e_w has to be non-vanishing. By Lemma 1 with $v = x$ and $\mathcal{Z}' = \mathcal{Z}$ follows that there exists an active path from w to x of the type in point 1 or 2 of the lemma (which is pointing into x).

We now show that connecting this active path from w to x with the active back-door path from x to y yields an active path from w to y when conditioning on $\{x\} \cup \mathcal{Z}$, which has the form as one of the paths in point 1 or 2 of Lemma 1 (with $v = y$ and $\mathcal{Z}' = \{x\} \cup \mathcal{Z}$). We consider the following three cases, according to the lemma.

1. $w \in \mathcal{Z}$, active path as in point 1 of Lemma 1: Since this active back-door path from w to x is pointing into x, the concatenated path has a collider at x, which is in $\mathcal{Z}'$ and hence the path is an active back-door path from w to y (not blocked by $\mathcal{Z}'$) as in point 1 of the lemma. (Note that if the two paths have more than the node x in common, this is still valid using (Spirtes et al., 2000, Lemma 3.3.1), which states conditions under which a series of active paths from nodes v_1 to v_2, v_2 to $v_3, \ldots, v_{n-1}$ to v_n yield an active path from v_1 to v_n.)

2. $w \notin \mathcal{Z}$, active path as in point 2(i) of Lemma 1: We immediately obtain a path as in point 2(ii): a directed active path from w to x (with $x \in \mathcal{Z}'$) for which there is an active back-door path to y (not blocked by $\mathcal{Z} = \mathcal{Z}' \setminus \{x\}$).

3. $w \notin \mathcal{Z}$, active path as in point 2(ii) of Lemma 1: The active back-door path from some $z \in \mathcal{Z}$ to x is again pointing into x, and by the same argument as in point 1 by concatenating this active back-door path with the active back-door path from x to y yields an active back-door path (not blocked by $\mathcal{Z}'$) from the given z to y, pointing into y. Hence, by using the same directed path from w to z and the concatenated back-door path from z to y we obtain a path as in point 2(ii) of the lemma.

Thus, there exists an active path from w to y when conditioning on $\{x\} \cup \mathcal{Z}$ as in point 1 or 2 of Lemma 1 and applying the lemma with $v = y$ and $\mathcal{Z}' = \{x\} \cup \mathcal{Z}$ now implies that the effect of e_w on y is also non-vanishing (as is the effect of e_w on x). (Note that in this step we need the faithfulness assumption to apply Lemma 1.) From the non-Gaussianity of e_w and the Darmois-Skitovitch theorem then follows that the residuals r_x and r_y are dependent. □

Proof of Theorem 1(c). We show that if the residuals are dependent then there exists an active back-door path from x to y.

We rewrite the regressions from Equations (2) and (3) to obtain $r_x = x - \boldsymbol{b}^T\boldsymbol{z} = \boldsymbol{d}_x\boldsymbol{e}$ and $r_y = y - ax - \boldsymbol{c}^T\boldsymbol{z} = \boldsymbol{d}_y\boldsymbol{e}$. Since r_x and r_y are by assumption dependent there exists a $w \in \{x\} \cup \mathcal{W} \cup \mathcal{U}$ whose error term e_w has a non-vanishing coefficient in the representation of both residuals r_x and r_y. By Lemma 1 we thus know that there exist some kind of active paths (as in point 1 or 2 of the lemma) from w to x and from w to y, and we will show that by concatenating these paths we can always construct an active back-door path from x to y. (Note that here we only made use of the "only-if-statement" of Lemma 1 which does not need the linear faithfulness assumption.) Consider the following cases:

1. $w = x$: By point 1 of Lemma 1 (with $v = y$, $e_w = e_x$ and $\mathcal{Z}' = \{x\} \cup \mathcal{Z}$) there is an active back-door path from x to y.

2. $w \in Z$: This means by point 1 of Lemma 1 that there exist active back-door paths $p = (w \leftarrow \ldots \rightarrow x)$ from w to x, and $q = (w \leftarrow \ldots \rightarrow y)$ from w to y. Hence, the concatenated path $x \leftarrow \ldots \rightarrow w \leftarrow \ldots \rightarrow y$ is an active back-door path from x to y since $w \in \mathcal{Z}$ is an active collider. (Note that the two paths can have more than w in common, but this does not change the fact that there is a d-connecting back-door path, using (Spirtes et al., 2000, Lemma 3.3.1).)

3. $w \notin \{x\} \cup Z$: According to Lemma 1 there are four different possibilities:

   2(i) + 2(i): There are directed active paths from w to x and to y, which immediately implies an active back-door path from x to y.

   2(i) + 2(ii): There is a directed active path from w to x and a directed active path from w to some $z \in \mathcal{Z}$ for which there is an active back-door path to y: Since z cannot be along the directed active path from w to x we can concatenate the two paths and get an active back-door path from x to y: $x \leftarrow \ldots \leftarrow w \rightarrow \ldots \rightarrow z \leftarrow \ldots \rightarrow y$ (using (Spirtes et al., 2000, Lemma 3.3.1) if the paths have more nodes than w in common).

   2(ii) + 2(i): Completely analogous to 2(i) + 2(ii)

   2(ii) + 2(ii): There are directed active paths from w to some $z_1, z_2 \in Z$ and active back-door paths from z_1 to x and from z_2 to y, respectively. As before we can concatenate the paths at w to get an active back-door path from x to y: $x \leftarrow \ldots \rightarrow z_1 \leftarrow \ldots \leftarrow w \rightarrow \ldots \rightarrow z_2 \leftarrow \ldots \rightarrow y$, (using once more (Spirtes et al., 2000, Lemma 3.3.1) if the paths have more nodes than w in common). □

3 Pseudocode of the backward elimination search procedure

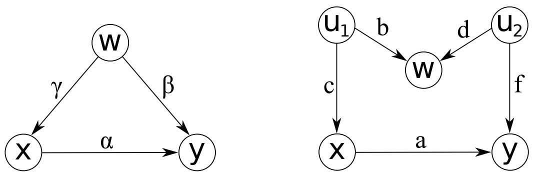
The pseudocode of the forward selection procedure was given in the paper in Algorithm 2. A similar piece of code is shown in Algorithm 3 for the backward elimination procedure, omitted from the main paper for reasons of space.

4 Details on the examples of Figure 1

In this section we present some details of the two example graphs in Figure 1 of the paper, redrawn here for convenience also in Figure 1. Note that for Example 2 we relabeled the edge coefficients with Latin letters, to be able to distinguish them from the coefficients of Example 1.

Algorithm 3 (Backward Elimination)

```
Let $\mathcal{Z}_{|W|} = \mathcal{W}$ and $m = |\mathcal{W}|$
If the residual $r_x$ from Algorithm 1 is Gaussian, set $p_{|\mathcal{W}|} =$ NaN,
Else otain a p-value $p_{|\mathcal{W}|}$ from the independence test of the residuals $r_x$ and $r_y$ from Algorithm 1
Repeat while $m > 0$
    $m = m - 1$
    For every set $\mathcal{Z} = \mathcal{Z}_{m+1} \setminus \{w\}$, $w \in \mathcal{Z}_{m+1}$, test whether $r_x$ is Gaussian
    For every such $\mathcal{Z}$ with non-Gaussian $r_x$ get a p-value $p_{\mathcal{Z}}$ from the independence test of $r_x$ and $r_y$
    Set $p_m = \max\{p_{\mathcal{Z}}\}$ and let $\mathcal{Z}_m$ be the corresponding set $\mathcal{Z}$
    If all $r_x$ were Gaussian, set $p_m =$ NaN and let $\mathcal{Z}_m$ be the set $\mathcal{Z}$ yielding the least Gaussian $r_x$
Return $\max_{m=0,\ldots,|\mathcal{W}|}\{p_m\}$ and the corresponding set $\mathcal{Z}_m$
```



(a) Example 1 (b) Example 2

Figure 1: Two example models with w, x, and y observed, and u_1 and u_2 hidden variables.

In Section 4.1 we first represent the models using Equations (1) and (4), and derive the covariance matrices over the observed variables w, x and y, which will be used throughout the whole section. In Section 4.2 we show what the effect of adjusting and not adjusting for variable w is in both examples, i.e. when the estimator of the causal effect of x on y is consistent and when it is inconsistent. We then prove in Section 4.3 that the two graphs in Figure 1 can model the same covariance matrices over the three observed variables w, x and y, which implies that they are indistinguishable from data over these variables when all disturbance variables are Gaussian. Finally, we show in Section 4.4 that if the disturbances deviate from Gaussianity, the two models can be told apart from observational data over w, x and y, using Theorem 1.

4.1 Model equations and covariance matrices

Writing the equations for the two example graphs in Figure 1 as in Equation (1) we obtain for Example 1

$$\begin{pmatrix} w \\ x \\ y \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ \gamma & 0 & 0 \\ \beta & \alpha & 0 \end{pmatrix} \begin{pmatrix} w \\ x \\ y \end{pmatrix} + \begin{pmatrix} e_w \\ e_x \\ e_y \end{pmatrix} \tag{6}$$

and for Example 2

$$\begin{pmatrix} u_1 \\ u_2 \\ w \\ x \\ y \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ b & d & 0 & 0 & 0 \\ c & 0 & 0 & 0 & 0 \\ 0 & f & 0 & a & 0 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ w \\ x \\ y \end{pmatrix} + \begin{pmatrix} e_{u_1} \\ e_{u_2} \\ e_w \\ e_x \\ e_y \end{pmatrix}. \tag{7}$$

Rewriting both Equations (6) and (7) using the formula in Equation (4) yields for Example 1

$$\begin{pmatrix} w \\ x \\ y \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \gamma & 1 & 0 \\ \alpha\gamma + \beta & \alpha & 1 \end{pmatrix} \begin{pmatrix} e_w \\ e_x \\ e_y \end{pmatrix} \tag{8}$$

and for Example 2

$$\begin{pmatrix} u_1 \\ u_2 \\ w \\ x \\ y \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ b & d & 1 & 0 & 0 \\ c & 0 & 0 & 1 & 0 \\ ac & f & 0 & a & 1 \end{pmatrix} \begin{pmatrix} e_{u_1} \\ e_{u_2} \\ e_w \\ e_x \\ e_y \end{pmatrix}. \tag{9}$$

We can obtain the covariance matrices over all variables in Examples 1 and 2 by using the formula

$$C(\boldsymbol{v}) = E(\boldsymbol{v}\boldsymbol{v}^T) = E(\mathbf{A}\boldsymbol{e}\boldsymbol{e}^T\mathbf{A}^T) = \mathbf{A}C(\boldsymbol{e})\mathbf{A}^T \tag{10}$$

where $C(\boldsymbol{v})$ and $C(\boldsymbol{e})$ are the covariance matrices over the variables in $\boldsymbol{v}$ and $\boldsymbol{e}$, respectively. By the independence of the disturbance variables $\boldsymbol{e}$ the matrix $C(\boldsymbol{e})$ is diagonal, with the variances of the disturbances along the diagonal. Denoting these variances in Example 1 with σ_w^2, σ_x^2 and σ_y^2, respectively, and in Example 2 with $\tau_{u_1}^2, \tau_{u_2}^2, \tau_w^2, \tau_x^2$ and τ_y^2, respectively, for Example 1 the covariance matrix is given by (showing only the upper triangle because of symmetry)

$$C_1 = \begin{pmatrix} \sigma_w^2 & \gamma\sigma_w^2 & (\alpha\gamma+\beta)\sigma_w^2 \\ . & \gamma^2\sigma_w^2 + \sigma_x^2 & \gamma(\alpha\gamma+\beta)\sigma_w^2 + \alpha\sigma_x^2 \\ . & . & (\alpha\gamma+\beta)^2\sigma_w^2 + \alpha^2\sigma_x^2 + \sigma_y^2 \end{pmatrix} \tag{11}$$

and for Example 2 by

$$C_2 = \begin{pmatrix} \tau_{u_1}^2 & 0 & b\tau_{u_1}^2 & c\tau_{u_1}^2 & ac\tau_{u_1}^2 \\ . & \tau_{u_2}^2 & d\tau_{u_2}^2 & 0 & f\tau_{u_2}^2 \\ . & . & b^2\tau_{u_1}^2 + d^2\tau_{u_2}^2 + \tau_w^2 & bc\tau_{u_1}^2 & abc\tau_{u_1}^2 + df\tau_{u_2}^2 \\ . & . & . & c^2\tau_{u_1}^2 + \tau_x^2 & ac^2\tau_{u_1}^2 + a\tau_x^2 \\ . & . & . & . & a^2c^2\tau_{u_1}^2 + f^2\tau_{u_2}^2 + a^2\tau_x^2 + \tau_y^2 \end{pmatrix}.$$

Thus, in Example 2 the covariance matrix over the observed variables w, x and y results in

$$C_{2,obs} = \begin{pmatrix} b^2\tau_{u_1}^2 + d^2\tau_{u_2}^2 + \tau_w^2 & bc\tau_{u_1}^2 & abc\tau_{u_1}^2 + df\tau_{u_2}^2 \\ . & c^2\tau_{u_1}^2 + \tau_x^2 & ac^2\tau_{u_1}^2 + a\tau_x^2 \\ . & . & a^2c^2\tau_{u_1}^2 + f^2\tau_{u_2}^2 + a^2\tau_x^2 + \tau_y^2 \end{pmatrix}, \tag{12}$$

the submatrix over these three variables (i.e. u_1 and u_2 are marginalized out).

We can use the covariance matrices to calculate the exact values of the estimates. To see this, for the regression $v = \sum_{z \in \mathcal{Z}'} c_z z + r_v$ the coefficient vector $\boldsymbol{c}_z$ is obtained by the OLS estimate which converges for growing sample size to an expression depending only on the covariance matrix. Let $i_{\mathcal{Z}'}$ denote the indices of $\mathcal{Z}'$, and i_v the index of v (among all observed variables) and let $\mathcal{D}$ be the data matrix over the observed variables. Then we obtain for the coefficient vector

$$\boldsymbol{c}_z^T = (\mathcal{D}_{i_{\mathcal{Z}'}}^T \mathcal{D}_{i_{\mathcal{Z}'}})^{-1} \mathcal{D}_{i_{\mathcal{Z}'}}^T \mathcal{D}_{i_v} = (\frac{1}{k}\mathcal{D}_{i_{\mathcal{Z}'}}^T \mathcal{D}_{i_{\mathcal{Z}'}})^{-1} \frac{1}{k}\mathcal{D}_{i_{\mathcal{Z}'}}^T \mathcal{D}_{i_v} \tag{13}$$

$$\xrightarrow{P} (Cov[i_{\mathcal{Z}'}, i_{\mathcal{Z}'}])^{-1} Cov[i_{\mathcal{Z}'}, i_v] \text{ for } k \to \infty \tag{14}$$

with $\mathcal{D}_{inds}$ the data submatrix over the variables with indices $inds$, k the sample size, and $Cov([inds_1, inds_2])$ the submatrix of the covariance matrix containing the rows with indices $inds_1$ and columns with indices $inds_2$.

4.2 The effect of controlling for w in Figure 1 (a) and 1 (b)

We now show that in case of Example 1 of Figure 1, not controlling for w in the regression of y on x yields an inconsistent estimator, whereas controlling for w will render the estimator consistent. In Example 2, the situation is exactly the opposite.

Lets start with Example 1. When not including w in the regression, i.e. estimating

$$y = \hat{\alpha}x + r_y$$

the regression coefficient $\hat{\alpha}$ is obtained as (using Equations (13) and (14))

$$\hat{\alpha} \xrightarrow{P} \frac{cov(x,y)}{V(x)} = \frac{\gamma(\alpha\gamma+\beta)\sigma_w^2 + \alpha\sigma_x^2}{\gamma^2\sigma_w^2+\sigma_x^2} = \alpha + \frac{\beta\gamma\sigma_w^2}{\gamma^2\sigma_w^2+\sigma_x^2} \tag{15}$$

where the covariance of x and y and the variance of x can be read off the covariance matrix defined in Equation (11). We can see that the estimator $\hat{\alpha}$ converges to the true effect α plus a *non-zero* term, depending on β, γ and the variances of e_w and e_x (σ_w^2 and σ_x^2, respectively) i.e. the estimator is not consistent (for $\beta \neq 0,\ \gamma \neq 0$).

On the other hand, if w is included in the regression of y on x, that means we are estimating

$$y = \hat{\alpha}x + \hat{\beta}w + r_y,$$

the estimates can be obtained using the covariance matrix of Equation (11) as shown in Equations (13) and (14):

$$\begin{aligned}
\begin{pmatrix} \hat{\beta} \\ \hat{\alpha} \end{pmatrix} &\xrightarrow{P} C_1[1:2,1:2]^{-1}C_1[1:2,3] = \begin{pmatrix} \sigma_w^2 & \gamma\sigma_w^2 \\ \gamma\sigma_w^2 & \gamma^2\sigma_w^2+\sigma_x^2 \end{pmatrix}^{-1} \begin{pmatrix} (\alpha\gamma+\beta)\sigma_w^2 \\ \gamma(\alpha\gamma+\beta)\sigma_w^2+\alpha\sigma_x^2 \end{pmatrix} \\
&= \frac{1}{\sigma_x^2\sigma_w^2} \begin{pmatrix} \gamma^2\sigma_w^2+\sigma_x^2 & -\gamma\sigma_w^2 \\ -\gamma\sigma_w^2 & \sigma_w^2 \end{pmatrix} \begin{pmatrix} (\alpha\gamma+\beta)\sigma_w^2 \\ \gamma(\alpha\gamma+\beta)\sigma_w^2+\alpha\sigma_x^2 \end{pmatrix} \\
&= \begin{pmatrix} \frac{\gamma^2\sigma_w^2(\alpha\gamma+\beta)}{\sigma_x^2} + (\alpha\gamma+\beta) - \frac{\gamma^2(\alpha\gamma+\beta)\sigma_w^2}{\sigma_x^2} - \alpha\gamma \\ -\frac{\gamma(\alpha\gamma+\beta)\sigma_w^2}{\sigma_x^2} + \frac{\gamma(\alpha\gamma+\beta)\sigma_w^2}{\sigma_x^2} + \alpha \end{pmatrix} = \begin{pmatrix} \beta \\ \alpha \end{pmatrix}
\end{aligned} \tag{16}$$

where $C_1[1:2,1:2]$ is the submatrix of C_1 with rows 1 and 2, and columns 1 and 2, and similarly $C_1[1:2,3]$ with rows 1 and 2, and column 3. This calculation shows that the estimator $\hat{\alpha}$ of α is consistent (as is the estimator $\hat{\beta}$ of β).

Turning now to the graph of Example 2 in Figure 1, we prove that not conditioning on w yields a consistent estimator, and conditioning on w an inconsistent estimator of the causal effect of x on y.

When not including w in the regression, i.e. estimating

$$y = \hat{a}x + r_y$$

we obtain the estimate using the covariance matrix of Equation (12) as

$$\hat{a} \xrightarrow{P} \frac{cov(x,y)}{V(x)} = \frac{ac^2\tau_{u_1}^2 + a\tau_x^2}{c^2\tau_{u_1}^2+\tau_x^2} = a \tag{17}$$

which is consistent.

Including w in the regression, as in

$$y = \hat{a}x + \hat{h}w + r_y$$

we obtain the regression coefficient by using a similar calculation as in Example 1 (using Equations (13) and (14)):

$$\begin{aligned}
\begin{pmatrix} \hat{h} \\ \hat{a} \end{pmatrix} &\xrightarrow{P} C_{2,obs}[1:2,1:2]^{-1}C_{2,obs}[1:2,3] \\
&= \begin{pmatrix} b^2\tau_{u_1}^2 + d^2\tau_{u_2}^2 + \tau_w^2 & bc\tau_{u_1}^2 \\ bc\tau_{u_1}^2 & c^2\tau_{u_1}^2+\tau_x^2 \end{pmatrix}^{-1} \begin{pmatrix} abc\tau_{u_1}^2 + df\tau_{u_2}^2 \\ ac^2\tau_{u_1}^2 + a\tau_x^2 \end{pmatrix} \\
&= \frac{1}{(d^2\tau_{u_2}^2+\tau_w^2)(c^2\tau_{u_1}^2+\tau_x^2)+b^2\tau_{u_1}^2\tau_x^2} \begin{pmatrix} c^2\tau_{u_1}^2+\tau_x^2 & -bc\tau_{u_1}^2 \\ -bc\tau_{u_1}^2 & b^2\tau_{u_1}^2+d^2\tau_{u_2}^2+\tau_w^2 \end{pmatrix} \\
&\quad \begin{pmatrix} abc\tau_{u_1}^2 + df\tau_{u_2}^2 \\ ac^2\tau_{u_1}^2 + a\tau_x^2 \end{pmatrix} \\
&= \begin{pmatrix} 0 \\ a \end{pmatrix} + \frac{1}{(d^2\tau_{u_2}^2+\tau_w^2)(c^2\tau_{u_1}^2+\tau_x^2)+b^2\tau_{u_1}^2\tau_x^2} \begin{pmatrix} df\tau_{u_2}^2(c^2\tau_{u_1}^2+\tau_x^2) \\ -bcdf\tau_{u_1}^2\tau_{u_2}^2 \end{pmatrix}.
\end{aligned} \tag{18}$$

In particular, this calculation shows that the estimator $\hat{a}$ of a is inconsistent since the second expression on the right hand side is always non-zero (for non-zero edge coefficients b, c, d, f).

4.3 Models of Figure 1 (a) and 1 (b) are covariance-equivalent

We show that the two example graphs of Figure 1 can model the same covariance matrices over the three observed variables w, x and y. It follows that for Gaussian disturbance variables $\boldsymbol{e}$ we cannot distinguish the two models from data over the observed variables, and hence are not able to decide whether to control for the variable w to obtain a consistent estimate of the causal effect of x on y.

In particular, we claim that any covariance matrix that can be modeled with the graph of Example 1 can also be obtained from the model of Example 2 and vice versa. Therefore, we solve two sets of equations.

In the first set we solve the six equations from $C_1 = C_{2,obs}$ with regard to the parameters in the first model (i.e. $\alpha, \beta, \gamma, \sigma_w^2, \sigma_x^2, \sigma_y^2$) which shows that any covariance matrix created from the second model could have equally well been obtained from the first model.[1] We obtain

$$\alpha = a - \frac{bcdf\tau_{u_1}^2\tau_{u_2}^2}{c^2d^2\tau_{u_1}^2\tau_{u_2}^2 + c^2\tau_{u_1}^2\tau_w^2 + b^2\tau_{u_1}^2\tau_x^2 + d^2\tau_{u_2}^2\tau_x^2 + \tau_w^2\tau_x^2} \tag{19}$$

$$\beta = \frac{df\tau_{u_2}^2(c^2\tau_{u_1}^2 + \tau_x^2)}{c^2d^2\tau_{u_1}^2\tau_{u_2}^2 + c^2\tau_{u_1}^2\tau_w^2 + b^2\tau_{u_1}^2\tau_x^2 + d^2\tau_{u_2}^2\tau_x^2 + \tau_w^2\tau_x^2} \tag{20}$$

$$\gamma = \frac{bc\tau_{u_1}^2}{b^2\tau_{u_1}^2 + d^2\tau_{u_2}^2 + \tau_w^2} \tag{21}$$

$$\sigma_w^2 = \tau_w^2 + b^2\tau_{u_1}^2 + d^2\tau_{u_2}^2 \tag{22}$$

$$\sigma_x^2 = \tau_x^2 + \frac{c^2d^2\tau_{u_1}^2\tau_{u_2}^2 + c^2\tau_{u_1}^2\tau_w^2}{b^2\tau_{u_1}^2 + d^2\tau_{u_2}^2 + \tau_w^2} \tag{23}$$

$$\sigma_y^2 = \tau_y^2 + \frac{b^2f^2\tau_{u_1}^2\tau_{u_2}^2\tau_x^2 + c^2f^2\tau_{u_1}^2\tau_{u_2}^2\tau_w^2 + f^2\tau_{u_2}^2\tau_w^2\tau_x^2}{c^2d^2\tau_{u_1}^2\tau_{u_2}^2 + c^2\tau_{u_1}^2\tau_w^2 + b^2\tau_{u_1}^2\tau_x^2 + d^2\tau_{u_2}^2\tau_x^2 + \tau_w^2\tau_x^2} \tag{24}$$

where we can see from Formulas (22), (23) and (24) that the variances are always positive.

Conversely, to show that any covariance matrix created by the model in Example 1 could have been produced by the graph of Example 2 as well, we solve the six equations from $C_1 = C_{2,obs}$ with regard to the parameters of the second model (i.e. $a, b, c, d, f, \tau_{u_1}^2, \tau_{u_2}^2, \tau_w^2, \tau_x^2, \tau_y^2$) and obtain:

$$a = \alpha + \frac{\beta\gamma\sigma_w^2}{\gamma^2\sigma_w^2 + \sigma_x^2} \tag{25}$$

$$\tau_{u_1}^2 = \frac{\gamma\sigma_w^2}{bc} \tag{26}$$

$$\tau_{u_2}^2 = \frac{\beta\sigma_w^2\sigma_x^2}{df(\gamma^2\sigma_w^2 + \sigma_x^2)} \tag{27}$$

$$\tau_w^2 = \frac{\sigma_w^2}{\gamma^2\sigma_w^2 + \sigma_x^2}(\gamma^2\sigma_w^2 + \sigma_x^2 - \frac{b}{c}\gamma^3\sigma_w^2 - \frac{b}{c}\gamma\sigma_x^2 - \frac{d}{f}\beta\sigma_x^2) \tag{28}$$

$$\tau_x^2 = \gamma^2\sigma_w^2 + \sigma_x^2 - \frac{c}{b}\gamma\sigma_w^2 \tag{29}$$

$$\tau_y^2 = \frac{1}{\gamma^2\sigma_w^2 + \sigma_x^2}(\gamma^2\sigma_w^2\sigma_y^2 - \frac{f}{d}\beta\sigma_w^2\sigma_x^2 + \beta^2\sigma_w^2\sigma_x^2 + \sigma_x^2\sigma_y^2) \tag{30}$$

and b, c, d and f are set such that the variances in Equations (26) - (30) are positive. Since the coefficient α does not appear in the formulas for the variances it is enough to analyze the following four cases.

[1] It is well known that the graph of Example 1 can model *any* covariance matrix over three observed variables so it must also be able to produce the one from Example 2. We will show this here anyway.

- For α arbitrary, $\beta > 0, \gamma > 0$ it is required that

  Eq. (26): $sign(b) = sign(c)$
  Eq. (27): $sign(d) = sign(f)$
  Eq. (29): $\frac{c}{b} < \frac{\gamma^2\sigma_w^2 + \sigma_x^2}{\gamma\sigma_w^2}$
  Eq. (30): $\frac{f}{d} < \frac{\gamma^2\sigma_w^2\sigma_y^2 + \beta^2\sigma_w^2\sigma_x^2 + \sigma_x^2\sigma_y^2}{\beta\sigma_w^2\sigma_x^2}$
  Eq. (28): b, c, d, f also such that $\gamma^2\sigma_w^2 + \sigma_x^2 > \frac{b}{c}(\gamma^3\sigma_w^2 + \gamma\sigma_x^2) + \frac{d}{f}\beta\sigma^2 x.$ (31)

  Note that such b, c, d and f always exist since setting $\frac{b}{c} = \frac{\gamma\sigma_w^2+\varepsilon}{\gamma^2\sigma_w^2+\sigma_x^2}$ and $\frac{d}{f} = \frac{\beta\sigma_w^2\sigma_x^2+\delta}{\gamma^2\sigma_w^2\sigma_y^2+\beta^2\sigma_w^2\sigma_x^2+\sigma_x^2\sigma_y^2}$ for some $\varepsilon > 0, \delta > 0$ yields for Equation (31)

  $$\begin{aligned}\gamma^2\sigma_w^2 + \sigma_x^2 &> \frac{\gamma\sigma_w^2 + \varepsilon}{\gamma^2\sigma_w^2 + \sigma_x^2}\frac{\gamma}{\gamma}(\gamma^3\sigma_w^2 + \gamma\sigma_x^2) + \frac{\beta\sigma_w^2\sigma_x^2 + \delta}{\gamma^2\sigma_w^2\sigma_y^2 + \beta^2\sigma_w^2\sigma_x^2 + \sigma_x^2\sigma_y^2}\beta\sigma_x^2 \\ &= \gamma^2\sigma_w^2 + \gamma\varepsilon + \frac{\beta^2\sigma_w^2\sigma_x^2 + \beta\delta}{\gamma^2\sigma_w^2\sigma_y^2 + \beta^2\sigma_w^2\sigma_x^2 + \sigma_x^2\sigma_y^2}\sigma_x^2\end{aligned}$$

  from which follows that

  $$1 > \varepsilon\frac{\gamma}{\sigma_x^2} + \delta\frac{\beta}{\gamma^2\sigma_w^2\sigma_y^2 + \beta^2\sigma_w^2\sigma_x^2 + \sigma_x^2\sigma_y^2} + \frac{\boldsymbol{\beta^2\sigma_w^2\sigma_x^2}}{\gamma^2\sigma_w^2\sigma_y^2 + \boldsymbol{\beta^2\sigma_w^2\sigma_x^2} + \sigma_x^2\sigma_y^2}$$

  and since the last term of the right hand side is always smaller than 1 (because of the bold parts), there always exist $\varepsilon > 0$ and $\delta > 0$ such that Equation (31) is fulfilled.

- For α arbitrary, $\beta < 0, \gamma > 0$ it is required that

  Eq. (26): $sign(b) = sign(c)$
  Eq. (27): $sign(d) = -sign(f)$
  Eq. (29): $\frac{c}{b} < \frac{\gamma^2\sigma_w^2 + \sigma_x^2}{\gamma\sigma_w^2}$
  Eq. (30): $\frac{f}{d} > \frac{\gamma^2\sigma_w^2\sigma_y^2 + \beta^2\sigma_w^2\sigma_x^2 + \sigma_x^2\sigma_y^2}{\beta\sigma_w^2\sigma_x^2}$
  Eq. (28): b, c, d, f also such that $\gamma^2\sigma_w^2 + \sigma_x^2 > \frac{b}{c}(\gamma^3\sigma_w^2 + \gamma\sigma_x^2) + \frac{d}{f}\beta\sigma^2 x$

  Similarly as for Equation (31) one can show that appropriate b, c, d and f always exist.

- For α arbitrary, $\beta > 0, \gamma < 0$ it is required that

  Eq. (26): $sign(b) = -sign(c)$
  Eq. (27): $sign(d) = sign(f)$
  Eq. (29): $\frac{c}{b} > \frac{\gamma^2\sigma_w^2 + \sigma_x^2}{\gamma\sigma_w^2}$
  Eq. (30): $\frac{f}{d} < \frac{\gamma^2\sigma_w^2\sigma_y^2 + \beta^2\sigma_w^2\sigma_x^2 + \sigma_x^2\sigma_y^2}{\beta\sigma_w^2\sigma_x^2}$
  Eq. (28): b, c, d, f also such that $\gamma^2\sigma_w^2 + \sigma_x^2 > \frac{b}{c}(\gamma^3\sigma_w^2 + \gamma\sigma_x^2) + \frac{d}{f}\beta\sigma^2 x$

  Again, such b, c, d and f always exist using a similar argument as above.

- For α arbitrary, $\beta < 0, \gamma < 0$ it is required that

  Eq. (26): $sign(b) = -sign(c)$

  Eq. (27): $sign(d) = -sign(f)$

  Eq. (29): $\frac{c}{b} > \frac{\gamma^2\sigma_w^2 + \sigma_x^2}{\gamma\sigma_w^2}$

  Eq. (30): $\frac{f}{d} > \frac{\gamma^2\sigma_w^2\sigma_y^2 + \beta^2\sigma_w^2\sigma_x^2 + \sigma_x^2\sigma_y^2}{\beta\sigma_w^2\sigma_x^2}$

  Eq. (28): b, c, d, f also such that $\gamma^2\sigma_w^2 + \sigma_x^2 > \frac{b}{c}(\gamma^3\sigma_w^2 + \gamma\sigma_x^2) + \frac{d}{f}\beta\sigma^2 x$

  An analog argument as in Equation (31) shows that appropriate b, c, d and f always exist.

This concludes the proof that the two graphs of Figure 1 can model the same covariance matrices.

4.4 Models of Figure 1 (a) and 1 (b) are distinguished for non-Gaussian disturbances

As shown in the previous section, the two graphs of Examples 1 and 2 in Figure 1 can generate the same covariance matrices. Hence, in the case of Gaussian disturbances these two models are indistinguishable and it is impossible to know whether to include w in the regression to obtain a consistent estimate of the causal effect of x on y. We show now, that if the disturbance terms $\boldsymbol{e}$ contain any kind of non-Gaussianity we can distinguish the two models from data over the observed variables w, x and y (asymptotically), and thus are able to obtain a consistent estimator of the causal effect of x on y by adjusting accordingly (using Theorem 1).

For the graph of Example 1, if we do not include w in the conditioning set $\mathcal{Z}$ which means that we apply Algorithm 1 with $\mathcal{Z} = \emptyset$, we estimate the two regressions

$$\begin{aligned} x &= r_x \\ y &= \hat{\alpha}x + r_y. \end{aligned}$$

It is well known that we obtain an inconsistent estimator, i.e. $(\hat{\alpha} - \alpha) \overset{P}{\nrightarrow} 0$ (as shown in Equation (15) in Section 4.2). The important point is that now the estimated error terms r_x and r_y are by necessity *statistically dependent* (and we can thus use Theorem 1(c) to detect the inconsistent estimator of α). This can be seen by expressing the two estimated residuals in terms of the original disturbances $\boldsymbol{e} = (e_w, e_x, e_y)$ using Equation (8):

$$\begin{aligned} r_x &= x = (\gamma\ ,\ 1\ ,\ 0)\,\boldsymbol{e} \\ r_y &= y - \hat{\alpha}x = (\alpha\gamma + \beta\ ,\ \alpha\ ,\ 1)\,\boldsymbol{e} - (\hat{\alpha}\gamma\ ,\ \hat{\alpha}\ ,\ 0)\,\boldsymbol{e} \\ &= ((\alpha - \hat{\alpha})\gamma + \beta\ ,\ \alpha - \hat{\alpha}\ ,\ 1)\,\boldsymbol{e}. \end{aligned}$$

Because $\hat{\alpha}$ is an inconsistent estimator of α (i.e. $(\hat{\alpha} - \alpha) \overset{P}{\nrightarrow} 0$) the contribution of e_x (and in general also of e_w) to both r_x and r_y is non-vanishing. By the Darmois-Skitovitch Theorem (Darmois, 1953; Skitovitch, 1953) we have that the non-Gaussianity of the elements in $\boldsymbol{e}$ is sufficient to ensure that r_x and r_y are statistically dependent. This is in accordance with Theorem 1, since, on the one hand an inconsistent estimate yields dependent residuals (parts (a) and (b) of the theorem), and on the other hand, part (c) of the theorem states that the dependent residuals imply the existence of an active back-door path from x to y ($x \leftarrow w \rightarrow y$), and hence we should not trust the estimate. Obviously, in practice only the latter part of the theorem is applicable.

If we include w in the regressions, i.e. apply Algorithm 1 with $\mathcal{Z} = \{w\}$, we obtain a consistent estimator $\hat{\alpha}$ of α (compare Equation (16), Section 4.2), the only back-door path from x to y ($x \leftarrow w \rightarrow y$) is blocked, and for the residuals we obtain

$$\begin{aligned} r_x &= x - \hat{\gamma}w = (\gamma\ ,\ 1\ ,\ 0)\,\boldsymbol{e} - (\hat{\gamma}\ ,\ 0\ ,\ 0)\,\boldsymbol{e} = (\gamma - \hat{\gamma}\ ,\ 1\ ,\ 0)\,\boldsymbol{e} \xrightarrow{P} (0\ ,\ 1\ ,\ 0)\,\boldsymbol{e} \\ r_y &= y - \hat{\alpha}x - \hat{\beta}w = (\alpha\gamma + \beta\ ,\ \alpha\ ,\ 1)\,\boldsymbol{e} - (\hat{\alpha}\gamma\ ,\ \hat{\alpha}\ ,\ 0)\,\boldsymbol{e} - (\hat{\beta}\ ,\ 0\ ,\ 0)\,\boldsymbol{e} \\ &= ((\alpha - \hat{\alpha})\gamma + (\beta - \hat{\beta})\ ,\ \alpha - \hat{\alpha}\ ,\ 1)\,\boldsymbol{e} \xrightarrow{P} (0\ ,\ 0\ ,\ 1)\,\boldsymbol{e}, \end{aligned}$$

since $\hat{\gamma}$ is a consistent estimator of γ (using Equations (13) and (14), and the covariance matrix of Equation (11) we obtain $\hat{\gamma} \xrightarrow{P} \frac{cov(x,w)}{V(w)} = \frac{\gamma\sigma_w^2}{\sigma_w^2} = \gamma$) and, as shown in Equation (16), $\hat{\alpha}$ and $\hat{\beta}$ are consistent estimators of α and β, respectively. Thus, the residuals r_x and r_y are asymptotically the same as the disturbances e_x and e_y, respectively, which are by assumption independent. This is also in line with the claims of Theorem 1, since by parts (a) and (b) the independent residuals imply a consistent estimate, and from part (c) follows that if all back-door paths are blocked then the residuals are independent. In practice, we can of course only use the former statement to identify consistent estimators.

Considering now the graph of Example 2, if w is included in the conditioning set $\mathcal{Z}$ (i.e. $\mathcal{Z} = \{w\}$ in Algorithm 1) we obtain the regressions

$$\begin{aligned} x &= \hat{g}w + r_x \\ y &= \hat{a}x + \hat{h}w + r_y \end{aligned}$$

where the estimated regression coefficients are in general inconsistent, i.e. $\hat{g} \overset{P}{\nrightarrow} 0$ (using the covariance matrix in Equation (12) yields $\hat{g} \xrightarrow{P} \frac{cov(x,w)}{V(w)} = \frac{bc\tau_{u_1}^2}{b^2\tau_{u_1}^2 + d^2\tau_{u_2}^2 + \tau_w^2} \neq 0$), and $\hat{h} \overset{P}{\nrightarrow} 0$, and $\hat{a} \overset{P}{\nrightarrow} a$ (see Equation (18) in Section 4.2). The residuals can again be expressed in terms of the original disturbance variables $\boldsymbol{e} = (e_{u_1}, e_{u_2}, e_w, e_x, e_y)$ using Equation (9):

$$\begin{aligned} r_x &= x - \hat{g}w = (c\ ,\ 0\ ,\ 0\ ,\ 1\ ,\ 0)\,\boldsymbol{e} - (\hat{g}b\ ,\ \hat{g}d\ ,\ \hat{g}\ ,\ 0\ ,\ 0)\,\boldsymbol{e} \\ &= (c - \hat{g}b\ ,\ -\hat{g}d\ ,\ -\hat{g}\ ,\ 1\ ,\ 0)\,\boldsymbol{e} \\ r_y &= y - \hat{a}x - \hat{h}w = (ac\ ,\ f\ ,\ 0\ ,\ a\ ,\ 1)\,\boldsymbol{e} - (\hat{a}c\ ,\ 0\ ,\ 0\ ,\ \hat{a}\ ,\ 0)\,\boldsymbol{e} - (\hat{h}b\ ,\ \hat{h}d\ ,\ \hat{h}\ ,\ 0\ ,\ 0)\,\boldsymbol{e} \\ &= ((a - \hat{a})c - \hat{h}b\ ,\ f - \hat{h}d\ ,\ -\hat{h}\ ,\ a - \hat{a}\ ,\ 1)\,\boldsymbol{e}. \end{aligned}$$

As above, since $\hat{a}$ is an inconsistent estimator of a, by the Darmois-Skitovitch theorem and the non-Gaussianity of the variables in $\boldsymbol{e}$ follows that the two estimated residuals are dependent, because they both have non-vanishing contributions, for example, from e_x. Furthermore, there exists an active back-door path from x to y ($x \leftarrow u_1 \rightarrow w \leftarrow u_2 \rightarrow y$).

However, when excluding w from the analysis, we obtain a consistent estimator of the causal effect of x on y (i.e. $\hat{a} \xrightarrow{P} a$, see Equation (17), Section 4.2), the only back-door path from x to y ($x \leftarrow u_1 \rightarrow w \leftarrow u_2 \rightarrow y$) is blocked, and the estimated residuals are given by

$$\begin{aligned} r_x &= x = (c\ ,\ 0\ ,\ 0\ ,\ 1\ ,\ 0)\,\boldsymbol{e} \\ r_y &= y - \hat{a}x = ((\hat{a} - a)c\ ,\ f\ ,\ 0\ ,\ \hat{a} - a\ ,\ 1)\,\boldsymbol{e} \xrightarrow{P} (0\ ,\ f\ ,\ 0\ ,\ 0\ ,\ 1)\,\boldsymbol{e} \end{aligned}$$

which are independent by the assumption of mutually independent components of $\boldsymbol{e}$ since they do not share any components. All these facts are of course again in line with Theorem 1.

5 An example where r_x and r_y are independent, but x and r_y dependent

In our procedure (Algorithm 1) we test for independence between the residuals r_x and r_y to infer whether an estimator is consistent. We now show that it is necessary to use the two residuals in the independence test, and we cannot instead use the variable x and the residual r_y, since these may be dependent even though the estimator is consistent.

Consider the graph in Figure 2. Expressing the model in the form of Equation (4) yields

$$\begin{pmatrix} u \\ w \\ x \\ y \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \gamma & 1 & 0 & 0 \\ \beta\gamma & \beta & 1 & 0 \\ \alpha\beta\gamma + \delta & \alpha\beta & \alpha & 1 \end{pmatrix} \begin{pmatrix} e_u \\ e_w \\ e_x \\ e_y \end{pmatrix}. \tag{32}$$

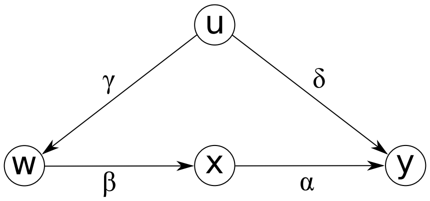


Figure 2: Example model with w, x and y observed variables, and u a hidden variable.

Using Equation (10), the covariance matrix over u, w, x and y is obtained as

$$C = \begin{pmatrix} \sigma_u^2 & \gamma\sigma_u^2 & \beta\gamma\sigma_u^2 & (\alpha\beta\gamma+\delta)\sigma_u^2 \\ \cdot & \gamma^2\sigma_u^2+\sigma_w^2 & \beta\gamma^2\sigma_u^2+\beta\sigma_w^2 & (\alpha\beta\gamma+\delta)\gamma\sigma_u^2+\alpha\beta\sigma_w^2 \\ \cdot & \cdot & \beta^2\gamma^2\sigma_u^2+\beta^2\sigma_w^2+\sigma_x^2 & (\alpha\beta\gamma+\delta)\beta\gamma\sigma_u^2+\alpha\beta^2\sigma_w^2+\alpha\sigma_x^2 \\ \cdot & \cdot & \cdot & (\alpha\beta\gamma+\delta)^2\sigma_u^2+\alpha^2\beta^2\sigma_w^2+\alpha^2\sigma_x^2+\sigma_y^2 \end{pmatrix} \quad (33)$$

with $\sigma_u^2, \sigma_w^2, \sigma_x^2$ and σ_y^2 the variances of e_u, e_w, e_x and e_y, respectively.

For $\mathcal{Z} = \{w\}$ we estimate the two regressions of Algorithm 1 as:

$$x = bw + r_x \quad (34)$$

$$y = ax + cw + r_y. \quad (35)$$

Since the set $\mathcal{Z}$ blocks all back-door paths from x to y it is admissible and thus, in the regression for y, the coefficient a of x is a *consistent estimator* of α (i.e. $a \xrightarrow{P} \alpha$, Back-door criterion, Pearl, 2009). The coefficient c of w is non vanishing because of the latent variable u. (This can also be shown formally using Equations (13) and (14) as in the previous section.) Furthermore, in the regression of x on w the coefficient b is a consistent estimator of β since $b \xrightarrow{P} \frac{cov(x,w)}{V(w)} = \frac{\beta\gamma^2\sigma_u^2+\beta\sigma_w^2}{\beta\gamma^2\sigma_u^2} = \beta$. Using Equation (32) to express r_x, r_y and x in terms of the disturbance variables e_u, e_w, e_x and e_y we obtain

$$\begin{aligned} r_x &= x - bw = (\beta\gamma,\ \ \beta,\ \ 1,\ \ 0)\,\boldsymbol{e} - (b\gamma,\ \ b,\ \ 0,\ \ 0)\,\boldsymbol{e} \xrightarrow{P} (0,\ \ 0,\ \ 1,\ \ 0)\,\boldsymbol{e} \\ r_y &= y - ax - cw \\ &= (\alpha\beta\gamma+\delta,\ \ \alpha\beta,\ \ \alpha,\ \ 1)\,\boldsymbol{e} - (a\beta\gamma,\ \ a\beta,\ \ a,\ \ 0)\,\boldsymbol{e} - (c\gamma,\ \ c,\ \ 0,\ \ 0)\,\boldsymbol{e} \\ &\xrightarrow{P} (\delta - c\gamma,\ \ -c,\ \ 0,\ \ 1)\,\boldsymbol{e} \\ x &= (\beta\gamma,\ \ \beta,\ \ 1,\ \ 0)\,\boldsymbol{e}. \end{aligned}$$

We can see that r_x and r_y are independent since they do not share any disturbance variables (asymptotically), and that x and r_y are dependent by the non-Gaussianity of the disturbances and the Darmois-Skitovitch Theorem. Hence, we cannot use a dependence between x and r_y to detect inconsistent estimators a of α (since in this example we obtained a consistent estimator a of α although x and r_y are dependent).

6 An example of a linearly unfaithful model for which the procedure fails

In this section we introduce an example that illustrates why the linear faithfulness assumption is needed in Theorem 1(b). Consider the graph in Figure 3 with two latent variables u_1 and u_2. We show that by choosing an unfaithful parametrization we can obtain an inconsistent estimator a of α even though the residual r_x is non-Gaussian and the residuals r_x and r_y are independent.

Writing the model in the form of Equation (4) yields

$$\begin{pmatrix} u_1 \\ u_2 \\ x \\ y \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \gamma & \delta & 1 & 0 \\ \beta+\alpha\gamma & \zeta+\alpha\delta & \alpha & 1 \end{pmatrix} \begin{pmatrix} e_{u_1} \\ e_{u_2} \\ e_x \\ e_y \end{pmatrix}. \quad (36)$$

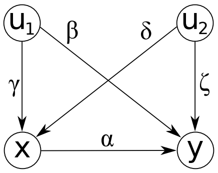


Figure 3: Example model with x and y observed, and u_1 and u_2 hidden variables.

Denoting the variances of the disturbance variables $\boldsymbol{e}$ as $\sigma^2_{u_1}, \sigma^2_{u_2}, \sigma^2_x$ and σ^2_y, respectively, we obtain the covariance matrix over all four variables using Equation (10) as

$$C = \begin{pmatrix} \sigma^2_{u_1} & 0 & \gamma\sigma^2_{u_1} & (\beta+\alpha\gamma)\sigma^2_{u_1} \\ \cdot & \sigma^2_{u_2} & \delta\sigma^2_{u_2} & (\zeta+\alpha\delta)\sigma^2_{u_2} \\ \cdot & \cdot & \gamma^2\sigma^2_{u_1}+\delta^2\sigma^2_{u_2}+\sigma^2_x & \gamma(\beta+\alpha\gamma)\sigma^2_{u_1}+\delta(\zeta+\alpha\delta)\sigma^2_{u_2}+\alpha\sigma^2_x \\ \cdot & \cdot & \cdot & (\beta+\alpha\gamma)^2\sigma^2_{u_1}+(\zeta+\alpha\delta)^2\sigma^2_{u_2}+\alpha^2\sigma^2_x+\sigma^2_y \end{pmatrix}. \tag{37}$$

When regressing y on x (i.e. $y = ax + r_y$, $\mathcal{Z} = \emptyset$), the estimator a of α can be read off from the covariance matrix C and is given by

$$a \xrightarrow{P} \frac{cov(x,y)}{V(x)} = \frac{\gamma(\beta+\alpha\gamma)\sigma^2_{u_1}+\delta(\zeta+\alpha\delta)\sigma^2_{u_2}+\alpha\sigma^2_x}{\gamma^2\sigma^2_{u_1}+\delta^2\sigma^2_{u_2}+\sigma^2_x} = \alpha + \frac{\gamma\beta\sigma^2_{u_1}+\delta\zeta\sigma^2_{u_2}}{\gamma^2\sigma^2_{u_1}+\delta^2\sigma^2_{u_2}+\sigma^2_x}. \tag{38}$$

To create an unfaithful parametrization we set the effect of one disturbance variable on r_y to zero. Therefore, we express the residual r_y in terms of the disturbances $\boldsymbol{e}$ using Equation (36):

$$\begin{aligned} r_y &= y - ax \\ &= (\beta+\alpha\gamma,\ \ \zeta+\alpha\delta,\ \ \alpha,\ \ 1)\,\boldsymbol{e} - (a\gamma,\ \ a\delta,\ \ a,\ \ 0)\,\boldsymbol{e} \\ &= (\beta+\alpha\gamma-a\gamma,\ \ \zeta+\alpha\delta-a\delta,\ \ \alpha-a,\ \ 1)\,\boldsymbol{e} \end{aligned}$$

and set for example the first entry $\beta+\alpha\gamma-a\gamma$ (the effect of e_{u_1} on r_y) to zero. Solving this equation with respect to the parameters yields

$$\sigma^2_x = \frac{\delta\sigma^2_{u_2}(\zeta\gamma-\beta\delta)}{\beta}. \tag{39}$$

For σ^2_x to be a valid variance it must be positive, which holds for example when $\zeta\gamma > \beta\delta$ and $\frac{\delta}{\beta} > 0$. The other parameters $(\alpha, \sigma^2_{u_1}, \sigma^2_{u_2}, \sigma^2_y)$ are free ones. Keeping this in mind, we obtain $\beta+\alpha\gamma-a\gamma = 0$ and hence we have

$$\begin{aligned} r_y &= (0,\ \ \zeta+\alpha\delta-a\delta,\ \ \alpha-a,\ \ 1)\,\boldsymbol{e} \\ r_x &= x = (\gamma,\ \ \delta,\ \ 1,\ \ 0)\,\boldsymbol{e}. \end{aligned}$$

Furthermore, the estimator a of α (given in Equation (38)) remains inconsistent when using σ^2_x from Equation (39) and is given by

$$a \xrightarrow{P} \alpha + \frac{\beta}{\gamma}. \tag{40}$$

If now e_{u_2} and e_x were Gaussian residuals, and e_{u_1} was a non-Gaussian residual (the distribution of e_y does not matter), then r_x is non-Gaussian, but the residuals r_x and r_y are independent: To see this, assume for a moment that r_x and r_y are both only influenced by e_{u_2} and e_x (i.e. the only non-vanishing coefficients are in the spots for these two residuals). Thus, r_x and r_y were sums of Gaussian residuals, and since r_x and r_y are uncorrelated they are in this case independent. Adding a (non-Gaussian) variable to only one of the two (for example e_{u_1} to r_x, or e_y to r_y) does not destroy the independence of r_x and r_y.

This is a case of a linearly unfaithful parametrization since from the graph we see that $u_1 = e_{u_1}$ is not d-separated from y given x, but the effect of u_1 on y given x is zero, and thus also the partial correlation of u_1 and y given x is zero. This can be calculated using the covariance matrix C in Equation (37) and the formula for partial correlation

$$\rho_{v_1,v_2.v_3} = \frac{\rho_{v_1,v_2} - \rho_{v_1,v_3}\rho_{v_2,v_3}}{\sqrt{1-\rho_{v_1,v_3}}\sqrt{1-\rho_{v_2,v_3}}}$$

where ρ_{v_i,v_j} denotes the correlation of v_i and v_j. Ignoring the denominator we get for the partial correlation of u_1 and y given x

$$\begin{aligned}\rho_{u_1,y.x} &\propto \frac{\sigma_{u_1}(\delta^2\sigma_{u_2}^2\beta + \sigma_x^2\beta - \gamma\delta\sigma_{u_2}^2\zeta)}{(\gamma^2\sigma_{u_1}^2 + \delta^2\sigma_{u_2}^2 + \sigma_x^2)\sqrt{(\beta+\alpha\gamma)^2\sigma_{u_1}^2 + (\zeta+\alpha\delta)^2\sigma_{u_2}^2 + \alpha^2\sigma_x^2 + \sigma_y^2}} \\ &\propto \delta^2\sigma_{u_2}^2\beta + \sigma_x^2\beta - \gamma\delta\sigma_{u_2}^2\zeta\end{aligned}$$

which is zero when using σ_x^2 from Equation (39).

To sum up, this example shows that even if the estimator a of α is inconsistent and the estimated residual of x, r_x, is non-Gaussian, there may still exist a parametrization which yields independent residuals r_x and r_y. Once we rule out these unfaithful cases, we can always conclude that if r_x is non-Gaussian and the residuals r_x and r_y are independent then the estimator is consistent, as stated in Theorem 1(b).

References

Darmois, G. (1953). Analyse générale des liaisons stochastiques. *Revue de l'Institut International de Statistique*, 21:2–8.

Greenland, S., Pearl, J., and Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology*, 10(1):37–48.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition.

Seber, G. A. and Lee, A. J. (2003). *Linear Regression Analysis*. Wiley Series in Probability and Statistics, 2nd edition.

Skitovitch, W. (1953). On a property of the normal distribution. *Doklady Akademii Nauk SSSR*, 89:217–219.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. Cambridge MA: MIT Press, 2nd edition.