TEST (2015) 24:489–509
DOI 10.1007/s11749-014-0421-3

ORIGINAL PAPER

Identification of causal effects in linear models: beyond instrumental variables

Elena Stanghellini · Eduwin Pakpahan

Received: 9 April 2014 / Accepted: 19 November 2014 / Published online: 6 December 2014

Abstract The instrumental variable (IV) formula has become widely used to address the issue of identification of a causal effect in linear systems with an unobserved variable that acts as direct confounder. We here propose two alternative formulations to achieve identification when the assumptions underlying the use of IV are violated. Parallel to the IV, the proposed formulas exploit the conditional independence structure of a directed acyclic graph and can be obtained via a series of univariate regressions, a feature that renders the results particularly attractive and easy to implement. By exploiting the notion of Markov equivalence, the derivations can also be applied to regression graphs, thereby enlarging the class of models to which the results are of use.

Keywords Causal effect · Confounder · Directed acyclic graph · Identification · Latent variable · Regression graph · Structural equation model

Mathematics Subject Classification Primary 62H99; Secondary 62H20

1 Introduction

Latent variables are ubiquitous in applied research. They may arise in observational studies where self-selection possibly occurs or in randomized studies with post-assignment complications, such as non compliance. In most situations, latent variables

E. Stanghellini (✉)
Department of Economics, University of Perugia, Perugia, Italy
e-mail: elena.stanghellini@stat.unipg.it; elena.stanghellini@unipg.it

E. Pakpahan
Department of Political and Social Science, European University Institute, Florence, Italy

act as direct confounders, i.e., they influence both the explanatory (X) variable and the response variable (Y). If not taken into account, conclusions on the effect of the explanatory variable on the response may be severely distorted.

The first way to tackle the problem was suggested in a seminal paper by Theil (1953) that introduced the concept of instrumental variables (IV), see also Bowden and Turkington (1984). With reference to a system of linear equations, if an additional variable that satisfies some conditional independence assumptions is available, the so-called instrument, the regression coefficient of the explanatory variable X on the outcome Y is identified from the distribution of the observable variables, and therefore can be estimated.

More recently, several contributions have appeared that embed the notion of identification of causal effects within graphical models, that are natural tools to describe conditional independence assumptions. The first contribution comes from the seminal paper of Pearl, see Pearl (1995), that best clarifies that a causal effect of X on Y is the effect of an hypothetical intervention on X. In words, it is assumed that the data-generating process is described through a causal directed acyclic graph (DAG), that is a DAG which includes all relevant variables, irrespective of whether they are observable or unobservable. It is also assumed that the DAG is stable under an external intervention that sets to x the value of X, i.e., the intervention does not destroy the basic features of the DAG. A causal effect is then defined as the conditional distribution of Y after an external intervention is performed to set to x the value of the random variable X. This notion is known as conditioning by intervention, as opposed to conditioning by observation, which is the usual concept. In a non-parametric framework, Pearl and coauthors have derived graphical conditions under which the causal effect can be identified. These are known as *back-door* and *front-door* conditions, see Pearl (2009 Chap. 3). Later contributions allow to determine a complete algorithm to find which causal effects in a DAG are identifiable, see Huang and Valtorta (2006), Shpitser and Pearl (2008), Tian and Pearl (2002).

If some parametric assumptions are made, the abovementioned criteria can be enlarged: under the additional assumption that the DAG is causal, the IV itself is an instance where the assumption of linearity allows for identification where the non-parametric criteria fail. More instances for the identification of causal effect in the linear case are in Kuroki and Pearl (2014), where results on identification in the binary case have also been presented, see also Stanghellini and Vantaggi (2013).

We here confine our research to linear regression models. In Fig. 1, the building block of our investigation: we are interested in the linear regression coefficient of X

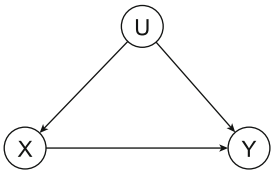


Fig. 1 A DAG showing the simplest example of a confounding problem: when U is associated with an unmeasured random variable the linear regression coefficient of X on Y, after conditioning on U, $\beta_{yx.u}$, is not identified from the observable distribution of X and Y

on Y after conditioning on U, that we denote by $\beta_{yx \cdot u}$. However, U is not observable and the effect of interest cannot be identified from the marginal distribution of the observable variables only. This work complements the work of Wermuth and Cox, see Wermuth and Cox (2008), where conditions of identification of linear regression parameters under marginalization and conditioning on indirect confounders are given. It also builds on results by Stanghellini and Wermuth and Kuroki and Pearl, see Stanghellini and Wermuth (2005) and Kuroki and Pearl (2014). Other examples are in Kuroki (2007) and Chan and Kuroki (2010). A similar work is also in Brito and Pearl (2002). However, our approach here is to assume that there is one latent variable responsible of the correlation induced between the observable variables. As such, some models that are not identified according to Theorem 1 of Brito and Pearl (2002) become identifiable. An instance is presented in Fig. 6a. Marginalization over U induces a model such that the corresponding graph violates conditions of Theorem 1 of Brito and Pearl (2002), but it is however identified.

The structure of the paper is as follows. In Sect. 2, we introduce the notion of DAG models, while in Sect. 3, we give the definition of causal graphs and introduce the concept of identification in the non-parametric context. We then turn into the parametric notion of identification in Sect. 4, and give details on the instrumental variable result. Regression graphs extend DAGs' models. They are introduced in Sect. 5, together with their pairwise Markov property. The notion of parameter equivalent regression graphs is introduced in Sect. 6, while the essential problem of the paper is presented in Sect. 7, and addressed in Sect. 8, which contains the main results. In Sect. 9, we report the problem of identification of a causal effect in the well-known Coleman's study and in Sects. 10 and 11 the derivations presented in the previous Sections are used to readdress the problem and propose new solutions. In Sect. 12, we draw some conclusions.

2 Directed acyclic graph models

A directed acyclic graph $G^V_{\text{dag}} = (V, E)$ is a mathematical object composed by V, the set of vertices or nodes, and $E \subseteq (V \times V)$ the set of edges. An edge is *directed* if and only if $(a, b) \subseteq E \Rightarrow (b, a) \notin E$. A directed edge between two vertices a and b such that $(a, b) \subseteq E$ is denoted by an arrow, pointing from a to b, i.e., $a \to b$. A graph $G^V_{\text{dag}} = (V, E)$ is said a *directed* graph if all edges in E are directed.

A *path* between a and b is a sequence of nodes $a_0 = a, a_1, \ldots, a_n = b$ such as $(a_{i-1}, a_i) \in E$ or $(a_i, a_{i-1}) \in E$. A *direction preserving* path is that every directed edge in the path points to the same direction. A *cycle* is a direction preserving path from a node directed to itself, or back to itself. An *acyclic* directed graph $G^V_{\text{dag}} = (V, E)$ is a directed graph without directed cycles.

We can use here the terminology of kinship: let a and b are two nodes in DAG G. If $a \to b$ then a is called a *parent* of b, and b is called the *child* of a. The set of parent of b is denoted *pa(b)*. A node a is said *transition* node if there are two nodes b and c such that $b \to a \to c$; it is said *source* node if $b \leftarrow a \to c$.

We have a $|V|$-vector of random variables $Y = (Y_1, \ldots, Y_{|V|})$ in a one-to-one correspondence with the set V with joint distribution function $P(y)$. We assume $P(y)$

admits a density $f(y)$ (this assumption can be removed). Then we say that the density $f(y)$ factorizes according to $G^V_{\text{dag}} = (V, E)$ if:

$$f(y) = \prod_{v \in V} f(y_v | y_{pa(v)}).$$

Assuming that Y is a vector of mean-centered random variables with Gaussian joint distribution with covariance matrix Σ, the recursive system can be written as:

$$AY = \epsilon, \quad \text{and} \quad cov(\epsilon) = \Delta, \tag{2.1}$$

where A is an upper triangular matrix with ones along the main diagonal and the negative of the off-diagonal element $-\alpha_{ij} = \beta_{ij.pa(i)\setminus j}$ corresponds to the partial regression coefficient of Y_j in the regression of Y_i against its parents, and is associated with a directed edge, $Y_j \to Y_i$. The Δ is a nonsingular diagonal covariance matrix of the residuals, with the partial variances $\delta_{ii} = \sigma_{ii.pa(i)}$ along the diagonal. Markov Properties of DAGs are presented in Lauritzen (1996, Chap. 3).

3 Non-parametric identification of causal effects

The notion of identification varies with the different meaning and use of DAGs appeared in the literature. Broadly speaking, we can say that there is stream of literature that looks at a DAG as tool for causal inference and a stream of the literature that looks at it as a statistical model. The first stream originated from the work of Pearl and coauthors (see Pearl 2009), while the second originated from the work of Cox, Wermuth and coauthors (see Cox and Wermuth 1996). Pearl and coauthors give the definition of a *causal* DAG as a tool that graphically describes the data-generating process and is suitable to infer the effect of interventions as well as spontaneous changes on the variables. A causal DAG also allows to identify causes of reported events (see Pearl 2001). In this context, causal DAGs are also denoted as *Causal Bayesian Networks*.

More technically, let $X \subset Y$. We first introduce the notion of conditioning by intervention. This is an operation that forces a particular set of variables to take on specific values. Intervention will be denoted as $do(X = x)$ and the interventional distribution will be written as $P(Y|do(X = x))$. Let $P(y)$ denote the joint distribution that factorizes according to a DAG and $P_x(y)$ be the interventional distribution, i.e., the joint distribution of Y after intervention $do(X = x)$. Then we have the following definition (see Pearl 2009, p. 24).

Definition 1 (*Causal DAG*) Let $P(y)$ be a probability distribution on a set V of variables, and let $P_x(y)$ denote the distribution resulting from the intervention $do(X = x)$ that sets a subset X of variables to a constant x. Denote by P_* the set of all interventional distributions $P_x(y)$, $X \subseteq V$, including $P(y)$, which represents no intervention (i.e., $X = \varnothing$). A DAG G^V_{dag} is said to be a causal DAG compatible with P_* if and only if the following three conditions hold for every $P_x \subseteq P_*$:

1. $P_x(y)$ factorizes according to G^V_{dag};
2. $P_x(y_v) = 1$ for all $Y_v \subseteq X$ whenever y_v is consistent with $X = x$;
3. $P_x(y_v|y_{pa(v)}) = P(y_v|y_{pa(v)})$ for all $Y_v \notin X$ whenever $y_{pa(v)}$ is consistent with $X = x$, i.e., each $P(y_v|y_{pa(v)})$ remains invariant to interventions not involving Y_v.

Conditioning by intervention is opposed to conditioning by observation, denoted by $P(Y|X = x)$, that is the usual understanding of the concept (see Lauritzen 2001 for a discussion). As a matter of fact, the interventional and conditional distributions need not coincide. Interventional distributions are also known as truncated factorization, as:

$$P_x(y) = \prod_{Y_v \notin X} P\left(y_v|y_{pa(v)}\right) \quad \text{for all } y_v \text{ consistent with } x.$$

An instance of when they coincide is the following. Let $X = pa(Y)$. Then in a causal graph:

$$P(Y|do(X = x)) = P(Y|X = x).$$

The causal effect of X on Y is identifiable from a DAG if the quantity $P(Y|do(X = x))$ can be computed uniquely from any positive probability of the observed variables. The major results on identification are known as back-door, that we here recall, and front-door criteria, see Pearl (2009), Chaps. 3, 4.

Definition 2 (*Blocked path*) In a DAG G^V_{dag}, a path p between node a and b is blocked by a (possibly empty) set Z if one of the following conditions holds:

1. p contains at least one non-collider that is in Z;
2. p contains at least one collider that is not in Z and has no descendant in Z.

If all paths between a and b are blocked by Z, we say that Z d-separates a and b. Let $G^V_{\text{dag},\underline{X}}$ be the DAG obtained by deleting all arrows emerging from X. Then the following definition applies.

Definition 3 (*Back-door criterion*) Let $\{X, Y\}$ and S be three disjoint subsets of V in a DAG G^V_{dag}. S is said to meet the back-door criterion if it satisfies the following two conditions:

1. no vertex in S is a descendant of X;
2. S d-separates X and Y in $G^V_{\text{dag},\underline{X}}$.

A causal effect is identifiable if there exists a set S that satisfies the back-door criterion. Typically, in a causal graph, no reference to the parametric form of the joint distribution is made. Therefore, all results derived in this context are non-parametric. Since we will work on parametric models, we refer the reader to the above reference for more details.

4 Parametric identification

The second stream is more granted in the statistical literature. In this context, DAGs are probabilistic models suitable to describe associations. Also in this case, the ordering among the variables is given a priori and the joint distribution of the random variables is parametrically specified. The objective is to make inference on parameters from random samples drawn from the joint distribution of the observable variables. This requires the notion of parametric identification.

We assume to have a family $\mathcal{M}(\Theta) = \{P_\theta : \theta \in \Theta\}$ of probability distributions, with parameter space Θ, that is Markov with respect to a given DAG. The family is said to be *globally identifiable at* $\theta_0 \in \Theta$ if for any $\theta \neq \theta^0$, P_θ and P_{θ^0} are different (see Rothenberg 1971; Bowden 1973). If this condition holds for all $\theta_0 \in \Theta$ then it is said to be *globally identifiable*. Global identifiability is also referred to as *strict identifiability*.

The family $\mathcal{M}(\Theta)$ is said to be *locally identified at* $\theta^0 \in \Theta$ if there exists a neighborhood of θ^0, $N(\theta^0)$, such that for any $\theta \in N(\theta^0)$, the corresponding probability distributions P_θ and P_{θ^0} are different. If this condition holds for any $\theta^0 \in \Theta$, then the model is said to be *locally identified*.

We denote with $\psi : \theta \to P_\theta$ the parametrization map. In this paper, we restrict to the models with polynomial ψ. Global identification coincides with injectivity of ψ, while local identification corresponds to k-to-one map for finite value of k. For this reason, we say that a parametric model is identifiable if there exists a *finite-to-one* parametrization map.

As argued in Allman et al. (2009), also the above definition may be too restrictive from the statistical point of view. There may be models such that the parametrization mapping is finite-to-one almost everywhere (i.e., everywhere except in a subspace of null measure). In this case, we speak of *generically* (globally or locally) identifiable models. Generically identifiable models may be the object of inference, provided that we are aware of the existence of the subspace of null measure where identifiability breaks down (see also Stanghellini and Vantaggi 2013).

As already noticed, when $X = pa(Y)$, the non-parametric identification of the causal effects of X on Y coincides with identification of the conditional distribution $P(Y|X = x)$. Therefore, when dealing with a causal Gaussian DAG, the α_{ij}'s are the *causal* effects and the parameters of interest are (subsets of) A and Δ. It is trivial to show that a Gaussian DAG without latent variables is always strictly identified, provided that the covariance matrix Σ is positive definite.

When the DAG contains unobserved variables, things are more complex. An example is the well-known instrumental variable problem, as depicted in Fig. 4a: we are interested in the value of $\alpha_{yx} = \beta_{yx\cdot u}$. Notice that U is not observed, so we cannot regress Y on X and U. However, we have an instrumental variable W such that $W \perp\!\!\!\perp U$ and $Y \perp\!\!\!\perp W|\{X, U\}$. Therefore,

$$\sigma_{wy} = \sigma_{wx}\beta_{yx\cdot u}, \quad \text{and hence} \tag{4.1}$$

$$\beta_{yx\cdot u} = \frac{\sigma_{wy}}{\sigma_{wx}} = \frac{\beta_{yw}}{\beta_{xw}}. \tag{4.2}$$

Since the last equation is the ratio between identified entities, the coefficient $\beta_{yx \cdot u}$ is identified, provided that $\beta_{xw} \neq 0$. Moreover, since the mapping is one-to-one, this is an instance of a generically (globally) identified model. A single-factor model with three observable indicators is identified provided that $\lambda \neq 0$, with λ the factor loadings. Notice that the model is identified up to the sign of λ, therefore, the mapping is two-to-one. This is an instance of generically (locally) identified model.

5 Regression graphs

It often happens that the set of variables under study naturally partitions into groups, that we here denote by $Y_a, Y_b, Y_c, \ldots$. Variables Y_a are called primary responses, as they can potentially be responses of all other variables; variables in Y_b are intermediate variables, as they can be potentially explanatory variables of Y_a, but they can be also potential responses for all other variables, but Y_a. Variables in the last block of the ordering are not only context variables, but also background variables. Within blocks, variables are said to be of equal standing.

Situations like this are dealt with regression graphs (see Wermuth and Sadeghi 2011). They are constructed in a way that nodes representing variables on equal standing are put in a box. Starting with the response of primary interest, boxes are ordered, usually from right to left. Edges can be of three types: arrows if they connect nodes in two different boxes, originating from any node in the box to the right; undirected dashed edges if they connect nodes within any given box and full-line edges if they couple context nodes. Graphs containing full-line edges only are called undirected.

More formally: a regression graph is a graph $G^V_{\text{reg}} = (V, E)$ on given a set of nodes V, partitioned as $V = (u, v)$, where v are the context variables. The edge set $E \subset (V \times V)$ contains three types of edges: full lines, arrows or dashed lines. The connected components g_j, $j = \{1, 2, \ldots, J\}$, are the disconnected undirected graphs that remain after removing all arrows. Since different orderings are possible, we speak of *compatible ordering* if each arrow starting from a node in g_j points to the future, i.e., node in a box g_r, with $r < j$, and never to the past. Let $g_{>j}$ denote the past, i.e., the set $g_{>j} = g_{j+1} \cup \cdots \cup g_J$. We have a collection of random variables $Y = (Y_1, \ldots, Y_{|V|})$ with joint distribution function $P(y)$. We again assume $P(y)$ admits a density $f(y)$. Then we say that the density $f(y)$ factorizes according to a regression graph if:

$$f(y) = \prod_{j=1}^{J} f\left(y_{g_j} | y_{g>j}\right).$$

Furthermore, we let the marginal density of the variables corresponding to the nodes in c to factorize according to the undirected graph in the corresponding box. The factorization represents a sequence of regressions for the joint responses given the past. Regression graphs are studied in detail in Wermuth and Sadeghi (2011), where the interpretation of each edge in the graph is given. We here recall the *pairwise Markov* property.

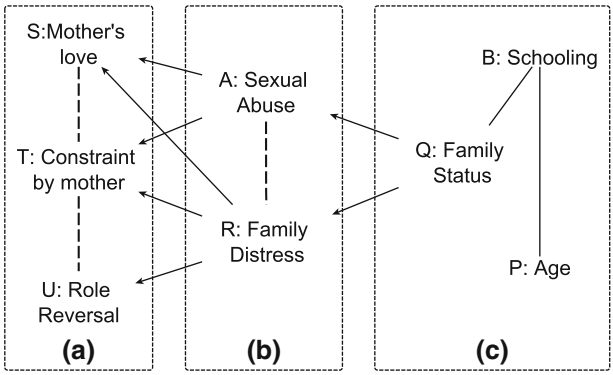


Fig. 2 An example of a regression graph with three blocks: in Y_a the primary responses, in Y_b the intermediate variables, and in Y_c the context variables

Definition 4 (*Pairwise Markov Property for regression graphs*) Pairwise Markov properties for regression graphs: Let $G^V_{\text{reg}} = (V, E)$ be a regression graph. A probability distribution P_V is said to satisfy the pairwise Markov property if, for any non-adjacent nodes $i, k \in V$, P_V satisfies:

(1) $i \perp\!\!\!\perp k | g_{>j}$ for i, k both in a response component g_j of u;
(2) $i \perp\!\!\!\perp k | g_{>j} \backslash \{k\}$ for i in g_j of u and k in $g_{>j}$;
(3) $i \perp\!\!\!\perp k | v \backslash \{i, k\}$ for i, k both in a context component g_j of v.

It then follows that, for the model the joint distribution of which is represented in Fig. 2, each variable in Y_a is conditionally independent from each other variable in Y_c given Y_b, as no arrow originating from a node in Y_c points directly to a node in Y_a. Furthermore, Q is independent of P given B, S is independent of U given A, R, Q, B and P.

Regression graphs extend directed acyclic graphs by allowing boxes with two types of undirected graph, one type for components of joint responses and the other for components of the context vector variable. DAG models are regression graphs with one variable per box. Regression graphs have an interest in their own, as possible models of the complex sets of data. They also arise after marginalization over source nodes in DAGs, provided that the children of the source node are not adjacent. More precisely, a dashed edge $a - - - b$ can always be seen as arising after marginalization over a source node v in the following structure: $a \leftarrow v \rightarrow b$. With this in mind, the collision nodes in a regression graphs are the inner nodes of the following three configurations: $\circ - - - \circ - - - \circ$, $\circ - - - \circ \longleftarrow \circ$ or $\circ \longrightarrow \circ \longleftarrow$. Each of these configurations is also called V-*structure*.

6 Markov equivalent and parameter equivalent models

Two models are Markov equivalent whenever their associated graphs capture the same independence structure, that is, the graphs imply the same set of independence statements. This implies that we never test a model but a whole class of observationally

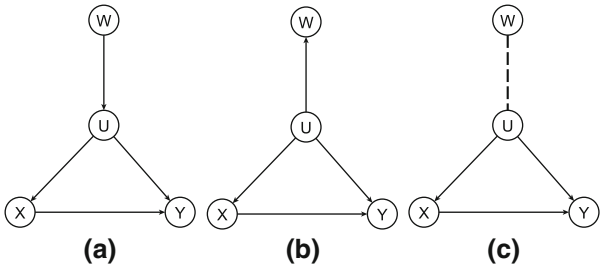


Fig. 3 Three Markov equivalent regression graphs

equivalent models that cannot be distinguished by any statistical means. It asserts as well that this equivalence class can be constructed from the graph, which thus provides the researcher with a clear representation of the compatible alternatives.

The following theorem is due to Wermuth and Sadeghi, see Wermuth and Sadeghi (2011):

Theorem 1 (Markov equivalent regression graphs) *Two regression graphs are Markov equivalent if and only if they have the same skeleton and the same* $\vee$*-structures, irrespective of the type of the edge.*

Three Markov equivalent regression graphs are in Fig. 3.

Since regression graphs contain DAGs, the theorem extends the result due to Frydenberg (1990) and Verma and Pearl (1991) on Markov equivalent DAGs. Parameter equivalence between two models concerns the existence of a one-to-one mapping from the parameters of the first model to the parameters of the second model. Another result, which is useful for our derivations, is the following, see Wermuth and Sadeghi (2011):

Theorem 2 (Parameter equivalent Gaussian regression graphs) *If two regression graphs are Markov equivalent for regular Gaussian distribution, then the distributions are also parameter equivalent.*

It follows from the fact that in the Gaussian case we have only one parameter attached to each edge, and the Gaussian distribution is closed under marginalization and conditioning, so that two models are parameter equivalent if they are Markov equivalent.

The results on Markov and parameter equivalent models allow to extend our derivations to regression graphs. As a matter of fact, we initially work with DAGs and establish rules for identification of a given causal effect, depicted in the DAG by a particular arrow. Then, we translate the identification rules into regression graphs that are Markov equivalent to the DAG we are dealing with. In doing so, we will make sure that the causal effect in the regression graph still remains depicted by an arrow. Since in the Gaussian case, Markov equivalence implies also parameter equivalence, then the rules for the identification of the causal effect of interest can be extended to regression graphs.

7 The confounding problem

With reference to the DAG represented in Fig. 1, the essence of the confounding problem lies on the fact that the linear least-squares regression coefficient $\beta_{yx \cdot u}$, that we shortly denote by α_{yx}, is not identified from the marginal distribution, i.e., the distribution of X and Y only. When the assumptions defining a causal DAG hold, this parameter is called the *causal effect* of X on Y.

We extend the graph of Fig. 1 by assuming that we have one more observable variable W. Figure 4 presents the six DAGs obtained when W is either the endpoint or the source node on one arrow only. Without additional information, only Fig. 4a leads to identification of α_{yx}; see Kuroki and Pearl (2014) for an example of additional information that permits identification of models as in Fig. 4e. In Fig. 5, we present the regression graph that is parameter equivalent to Fig. 4a, obtained by making use of the results on parameter equivalent between DAGs and regression graph models presented in the previous section. Notice that Fig. 4a is not parameter equivalent to Fig. 4d, as it contains a different set of V-structures.

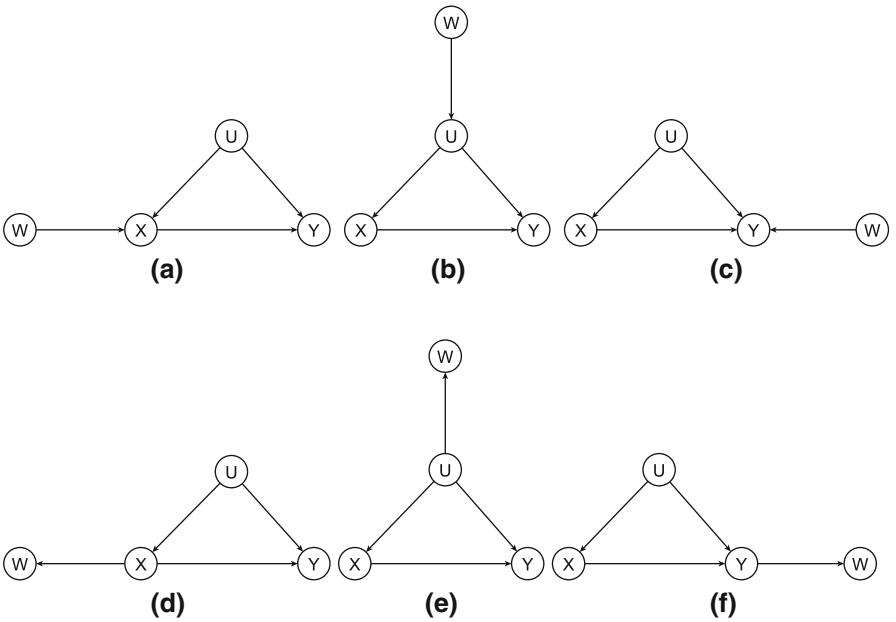
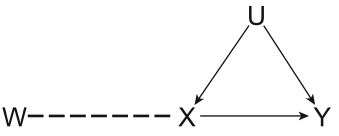


Fig. 4 Six DAGs obtained by adding node W to Fig. 1, with W having one neighbor node only. **a** corresponds to the instrumental variable model and is the only model that permits identification of $\beta_{yx \cdot u}$

Fig. 5 A regression graph that is Markov equivalent to the instrumental variable graph in Fig. 4a



8 A general rule for the identification of α_{yx}

We can now state our rule for DAGs models in which X is a parent of Y and both are children of latent variable U. We assume we have at least four observed variables.

Theorem 3 *In a Gaussian DAG $G^{V}_{dag} = (V, E)$ with $A \subseteq V$, $A = \{Y, U, X, Z, W\}$, such that the effect of X on Y is confounded by U, the parameter $\alpha_{yx} = \beta_{yx \cdot pa(y) \setminus x}$ is identified if:*

1. *in $V \setminus A$ there exists a set C of random variables such that $\{Z, U, W, C\}$ blocks every back-door path from X to Y;*
2. *$\{X, W, U, C\}$ d-separates Y from Z, i.e., $\beta_{yz \cdot xwuc} = 0$; and*
3. *at least one of the two following conditions hold:*
   a. *(U, C) d-separates W from $\{X, Z\}$;*
   b. *(U, C) d-separates Z from $\{W, X\}$.*

Furthermore, if Condition 3(a) holds:

$$\alpha_{yx} = \frac{\beta_{yx \cdot c}\beta_{wz \cdot c} - \beta_{wx \cdot c}\beta_{yz \cdot c}}{\beta_{wz \cdot c} - \beta_{xz \cdot c}\beta_{wx \cdot c}} \tag{8.1}$$

$$= \frac{\left(\beta_{yx \cdot zwc}\beta_{wz \cdot yxc}\right) - \left(\beta_{wx \cdot yzc}\beta_{yz \cdot xwc}\right)}{\beta_{wz \cdot yxc} + \left(\beta_{yz \cdot xwc}\beta_{wy \cdot xzc}\right)}. \tag{8.2}$$

Otherwise, if Condition 3(b) holds:

$$\alpha_{yx} = \frac{\beta_{zw \cdot c}\beta_{yx \cdot c} - \beta_{zx \cdot c}\beta_{yw \cdot c}}{\beta_{zw \cdot c} - \beta_{zx \cdot c}\beta_{xw \cdot c}} \tag{8.3}$$

$$= \frac{\left(\beta_{yx \cdot zwc}\beta_{zw \cdot yxc}\right) - \left(\beta_{yw \cdot xzc}\beta_{zx \cdot ywc}\right)}{\beta_{zw \cdot yxc} + \left(\beta_{yw \cdot xzc}\beta_{zy \cdot xwc}\right)}. \tag{8.4}$$

Proof See "Appendix". The above identification rules require the denominator of each expression not to vanish, and therefore lead to generic identifiability of α_{yx}. Notice that Eq. (8.1) is a straightforward extension of Kuroki and Pearl (2014) formula (6). See also Cai and Kuroki (2012). Also Eq. (8.4) is implicit in Stanghellini (2004).

The two formulas also have implications in terms of estimating procedures, as they provide the explicit expression of α_{yx} as a function of univariate linear regression coefficients. Therefore, estimation can be performed without resorting to ML methods, but simply using the methods of moments. Note further that the assumption of joint Gaussianity may be relaxed in favor to the standard assumptions of linear regression models, as in structural equation models (SEM), see Bollen (1989). In Fig. 6 the most complex DAGs with five variables are shown, for which identification of α_{yx} is possible according to Theorem 3. □

9 Coleman's research

In 1980, James Coleman and his colleagues conducted a comprehensive study about the students' performance in secondary schools, see Coleman et al. (1982). The study

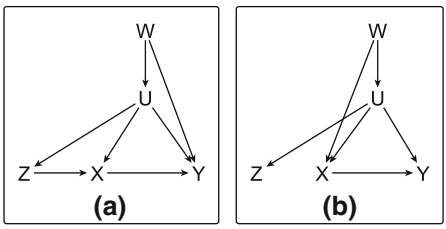


Fig. 6 The most complex DAGs with five nodes satisfying, in order, **a** conditions 3(a) and **b** conditions 3(b) of Theorem 3. In the corresponding models $\beta_{yx.u}$ is identified

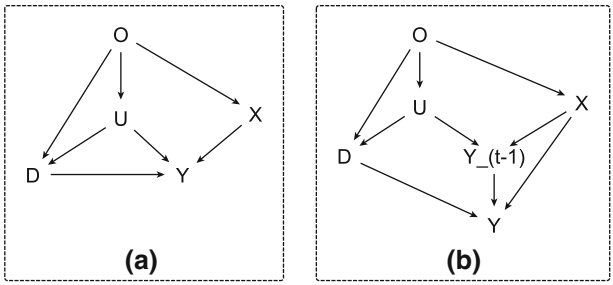


Fig. 7 Two DAGs representing criticism on Coleman's model on the effect of schools' choice (D) on students' test score at the 12th grade (Y): in **a** the effect is not identified while in **b** the use of the students' test score at the 10th grade (Y_{t-1}) as a proxy permits identification

has been extensively discussed in Morgan (2007). The objective of the study was to figure out what types of school characteristics are associated with students' success. Coleman and colleagues did focus on the single schools characteristic: whether schools are public or private. Given that Catholic schools constitute a large and relatively homogeneous group in the private school sector, they decided to direct most of their research to the differences between Catholic schools and public ones. They examined achievement test data and concluded that students in Catholic schools learn more than students in public ones. However, criticism arose around what is known as Coleman's conclusion, as the model did not take into account that best students may choose to attend the Catholic schools, and that the positive effect may be illusorily created by the self-selection mechanism. The model depicted in Fig. 7a represents the described criticism, with one variable U representing students' unmeasured ability, influencing both the choice of the school (D, an observed binary variable, taking value 1 for those attending Catholic schools and 0 for those who do not) and the standardized achievement test score at the 12th grade (Y). Then, U is an unobserved factor that confounds the effect of school on the test score in the direction that best students, i.e., students with high level of U, are more likely to choose Catholic schools. X is the determinant factor of achievement test scores that is not associated with the school selection, and O is the ultimate background that stimulate both score achievement test score and school selection (notice that X and O may be vectors of variables). The objective of Coleman's research is represented by the effect of D on Y. Given that there is one latent variable U indeed the effect is not identifiable.

To account for the criticism, Coleman later proposed a solution to answer the unidentifiability of that effect. He added another observed variable that could be a

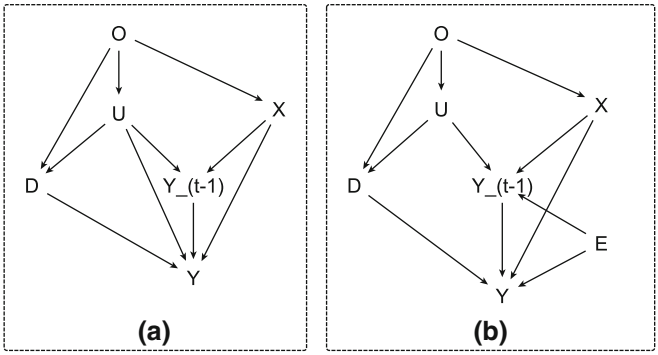


Fig. 8 Two DAGs representing the criticism raised on Coleman's model with the proxy variable: **a** U is influencing both test scores (Y and Y_{t-1}) and **b** there is an unobserved factor E such that conditioning on Y_{t-1} creates a back-door path. In both cases, the causal effect of D on Y is not identified

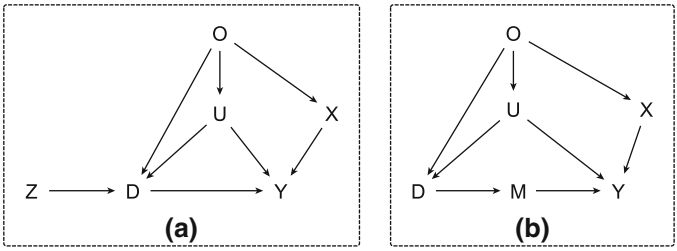


Fig. 9 Two alternative DAGs as proposed by Morgan and Winship (2007), with **a** Z is an instrumental variable or **b** M a mediator variable. In both cases, the effect of D on Y is identified

proxy for the test score of the 12th grade, that is the 10th grade score Y_{t-1}. He argued that those who obtain good score at the 10th grade would also obtain good score at the 12th grade, see Fig. 7b. Therefore, the test score in the 10th grade would be an effective pre-test variable for the 12th grade. Then, the effect is identified by repeated use of back-door criterion.

However, this solution is not without criticism too. It is argued in Morgan (2007, Chap. 6), that the latent variable U should also influence the 12th grade score. The unobserved factor U would affect both 10th and 12th grade levels, as in Fig. 8a, therefore confounding the influence of the choice on the school. Another criticism concerns the possible existence of another unobserved confounder, denoted by E, for 10th and 12th grade students, such that conditioning on Y_{t-1} creates a back-door path that is not blocked by any observed variable, see Fig. 8b.

In Morgan (2007, Chap. 6) two solutions are proposed. The first uses an instrumental variable, see Fig. 9a.

The instrumental variable may be the geographic area, denoted by Z, in a hypothetical study that randomly assigns tuition vouchers that could be redeemed at the Catholic schools. The second solution is by adding a mediator variable, that represents *a full mechanism* of the effect of D on Y, denoted by M, see Fig. 9b. Note that the front-door criterion holds here.

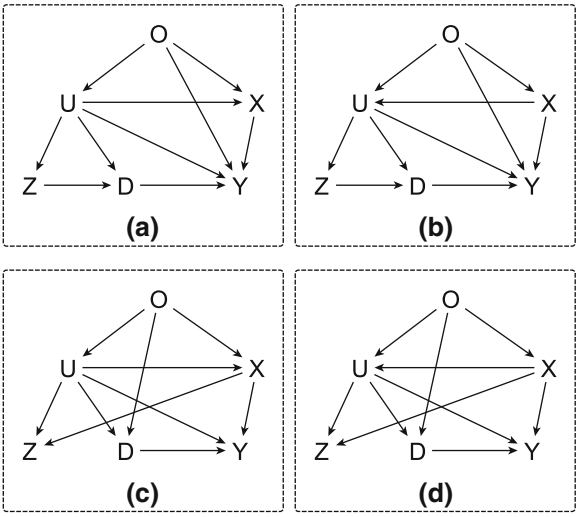


Fig. 10 Four alternative DAGs to solve Coleman's problem for which: **a**, **b** (8.2) and **c**, **d** (8.4) applies and therefore the effect of D on Y is identified

10 Coleman's research revisited

Previous solutions are in some sense questionable too. As an instance, the possibility to perform a random assignment of tuition vouchers that renders the geographic area Z an exogenous variable is very much theoretical. In practice, the geographic area Z can be influenced by factors, such as parents' wealth and education, that can be very much related to the unobserved factor U. However, taking into account all covariates plus the Z variable, we have 6 variables of which 5 are observed and 1 is latent. The range of models leading to identification of the effect of D on Y can be enlarged by making use of Theorem 3. In particular, by making use of Fig. 10, we lay down the assumptions that permit to use (8.2) or (8.4).

In Fig. 10a, b, the most complex model allowing to use (8.2) is presented, while Fig. 10(c, d) the most complex model allowing to use (8.4) is presented. Notice that Fig. 10b, d are obtained from Fig. 10(a, c) by flipping the arrow (U, X) and thereby Markov equivalence is preserved. In the first case, we have:

$$\alpha_{yd} = \frac{\left(\beta_{yd\cdot oxz}\beta_{oz\cdot dyx}\right) - \left(\beta_{od\cdot yxz}\beta_{yz\cdot dox}\right)}{\beta_{oz\cdot dyx} + \left(\beta_{yz\cdot dox}\beta_{oy\cdot dxz}\right)} \tag{10.1}$$

while in the second case, we have:

$$\alpha_{yd} = \frac{\left(\beta_{yd\cdot oxz}\beta_{zo\cdot dyx}\right) - \left(\beta_{yo\cdot dxz}\beta_{zd\cdot yxo}\right)}{\beta_{zo\cdot dyx} + \left(\beta_{yo\cdot dxz}\beta_{zy\cdot dox}\right)}. \tag{10.2}$$

Notice that both formulas imply conditioning on D, so it does not make much of practical difference here if D is continuous or binary. Notice further that all models obtained from Fig. 10a, b by deleting arrows are also identified, provided that the denominator of α_{yd} does not vanish.

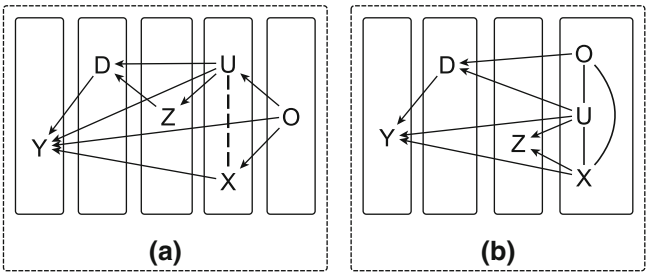


Fig. 11 Two regression graphs that are Markov equivalent to DAGs of Fig. 10, and such that the effect of D on Y is identified

The models entail different conditional independence assumptions. In particular models that satisfy condition 3(a) of Theorem 3 (for which therefore (8.2) can be applied) imply that O is an explanatory variable of Y but not of D, whereas models that satisfy condition 3(b) (for which (8.4) can be applied) allow O to be explanatory variable of D but not of Y. Furthermore, the second class of models does not allow Z to be directly explanatory variable of D and endogeneity of Z is accounted by the arrow emanating from U. Choice between the two strategies must be made on subject matter considerations.

11 Regression graphs for Coleman's problem

Having found various DAGs structures that provide solutions to the Coleman's problem, we can also derive the regression graphs that are Markov equivalent to those DAGs. Using the results in previous sections, this implies that they are also parameter equivalent. Therefore, if a DAG is such that one of the previous rule leads to an identification of the effect of D on Y, the formula will also apply to their Markov equivalent regression graphs. In Fig. 11, some examples are presented.

Figure 11a shows a regression graph which is Markov equivalent to DAG Fig. 10(a, b) while Fig. 11b represents a regression graph which is Markov equivalent to Fig. 10c, d). In the first graph depicted in Fig. 11a O is a context variable, potentially explanatory to all the others, while U and X are on equal standing. The rest follows the original DAG structure. In the second graph depicted in Fig. 11b we consider variables $\{O, U, X\}$ to be context variables, while the rest are ordered according to the original structure.

12 Conclusions

Graphical models are natural tools to address issues of causality and identification. After clarifying under which conditions a parameter of a Gaussian DAG can be enhanced with the interpretation of a *causal effect*, we have enlarged the set of DAGs that permit identification of a linear causal effect. Moreover, thanks to results on parameter equivalent models, the results can be applied to regression graphs, i.e., graphs that specify the ordering only among groups of variables. Since the identification is

achieved throughout a series of univariate regressions, moment estimation methods can be used. Bootstrapping procedures can be easily implemented to estimate the variance of the estimates of the parameters of interest and this has been done in Pakpahan (2012). Notice, that the assumption of joint Gaussianity of the random variables can be relaxed in favor of the ones underlying SEM.

Acknowledgments The authors are grateful to Prof. Nanny Wermuth and to Prof. Giovanni Maria Marchetti for their constructive and detailed comments.

Appendix A: Proof of (8.1) and (8.2)

Without loss of generality assume $C = \emptyset$ (otherwise derivations hold conditionally on C). To see that Eq. (8.1) is a straightforward extension of Kuroki and Pearl (2014) formula, notice that their formula is:

$$\alpha_{yx} = \frac{\sigma_{xy}\sigma_{zw} - \sigma_{xw}\sigma_{zy}}{\sigma_{xx}\sigma_{zw} - \sigma_{xz}\sigma_{xw}} \tag{12.1}$$

and by multiplication with $\frac{\sigma_{xx}}{\sigma_{xx}}\frac{\sigma_{zz}}{\sigma_{zz}}$ the result follows.

The derivations are taken from Kuroki and Pearl (2014). See also Cai and Kuroki (2012). Since we are interested in parameter $\alpha_{yx} = \beta_{yx\cdot zuw}$, we will then make use the definition of least-squares regression:

$$\begin{bmatrix} \beta_{yx\cdot zuw} \\ \beta_{yz\cdot xuw} \\ \beta_{yu\cdot xzw} \\ \beta_{yw\cdot xuz} \end{bmatrix} = \begin{bmatrix} \sigma_{xx} & \sigma_{xz} & \sigma_{xu} & \sigma_{xw} \\ \sigma_{zx} & \sigma_{zz} & \sigma_{zu} & \sigma_{zw} \\ \sigma_{ux} & \sigma_{uz} & \sigma_{uu} & \sigma_{uw} \\ \sigma_{wx} & \sigma_{wz} & \sigma_{wu} & \sigma_{ww} \end{bmatrix}^{-1} \begin{bmatrix} \sigma_{yx} \\ \sigma_{yz} \\ \sigma_{yu} \\ \sigma_{yw} \end{bmatrix},$$

thus we take the first two expressions, that is σ_{yx} and σ_{yz}, as follows:

$$\begin{bmatrix} \sigma_{yx} \\ \sigma_{yz} \end{bmatrix} = \begin{bmatrix} \sigma_{xx} & \sigma_{xw} \\ \sigma_{zx} & \sigma_{zw} \end{bmatrix} \begin{bmatrix} \beta_{yx\cdot zuw} \\ \beta_{yw\cdot xzu} \end{bmatrix} + \begin{bmatrix} \sigma_{xu} \\ \sigma_{zu} \end{bmatrix} \beta_{yu\cdot xzw} + \begin{bmatrix} \sigma_{zx} \\ \sigma_{zw} \end{bmatrix} \beta_{yz\cdot xwu},$$

then we have

$$\begin{bmatrix} \sigma_{xx} & \sigma_{xw} \\ \sigma_{zx} & \sigma_{zw} \end{bmatrix} \begin{bmatrix} \beta_{yx\cdot zuw} \\ \beta_{yw\cdot xzu} \end{bmatrix} = \begin{bmatrix} \sigma_{yx} \\ \sigma_{yz} \end{bmatrix} - \begin{bmatrix} \sigma_{xu} \\ \sigma_{zu} \end{bmatrix} \beta_{yu\cdot xzw} - \begin{bmatrix} \sigma_{zx} \\ \sigma_{zw} \end{bmatrix} \beta_{yz\cdot xwu},$$

thus

$$\begin{bmatrix} \beta_{yx\cdot zuw} \\ \beta_{yw\cdot xzu} \end{bmatrix} = \begin{bmatrix} \sigma_{xx} & \sigma_{xw} \\ \sigma_{zx} & \sigma_{zw} \end{bmatrix}^{-1} \left\{ \begin{bmatrix} \sigma_{yx} \\ \sigma_{yz} \end{bmatrix} - \begin{bmatrix} \sigma_{xu} \\ \sigma_{zu} \end{bmatrix} \beta_{yu\cdot xzw} - \begin{bmatrix} \sigma_{zx} \\ \sigma_{zw} \end{bmatrix} \beta_{yz\cdot xwu} \right\},$$

$$= K \begin{bmatrix} \sigma_{zw} & -\sigma_{xw} \\ -\sigma_{zx} & \sigma_{xx} \end{bmatrix} \left\{ \begin{bmatrix} \sigma_{yx} \\ \sigma_{yz} \end{bmatrix} - \begin{bmatrix} \sigma_{xu} \\ \sigma_{zu} \end{bmatrix} \beta_{yu\cdot xzw} - \begin{bmatrix} \sigma_{zx} \\ \sigma_{zw} \end{bmatrix} \beta_{yz\cdot xwu} \right\},$$

where $K = (\sigma_{zw}\sigma_{xx} - \sigma_{xw}\sigma_{zx})^{-1}$. Given that $\alpha_{yx} = \beta_{yx \cdot zuw}$ and that from Condition 2 $\beta_{yz \cdot xwu} = 0$, thus we can calculate for the first element only, as follows:

$$\beta_{yx \cdot zuw} = K\left\{(\sigma_{zw}\sigma_{yx} - \sigma_{xw}\sigma_{yz}) - (\sigma_{zw}\sigma_{xu} - \sigma_{xw}\sigma_{zu})\beta_{yu \cdot xzw}\right\}.$$

By noting that from Condition 3(a):

$$\sigma_{wx} = \sigma_{wu}\sigma_{uu}^{-1}\sigma_{ux},$$
$$\sigma_{wz} = \sigma_{wu}\sigma_{uu}^{-1}\sigma_{uz},$$

thus $\sigma_{zw}\sigma_{xu} - \sigma_{xw}\sigma_{zu} = 0$. Then what remains lead to (12.1) and therefore to (8.1).

The equivalence between (8.1) and (8.2) can be proved by first principles, or using arguments similar to the ones in Stanghellini (2004) and Stanghellini and Wermuth (2005). We here follow the second strategy. We assume without loss of generality that $V \backslash A = \emptyset$. We partition $A = \{O, U\}$ with $O = \{Y, X, Z, W\}$. By denoting with λ the vector of covariances between O and U, the observable covariance matrix of the random variables in O is:

$$\Sigma_{OO} = \lambda\lambda^T\sigma_{uu} + \Sigma_{OO \cdot U},$$

with inverse:

$$\Sigma_{OO}^{-1} = -\delta\delta^T + \Sigma_{OO \cdot U}^{-1}. \tag{12.2}$$

where $\delta = \Sigma^{OU}/\sqrt{\sigma^{uu}}$, σ_{uu} and σ^{uu} are, in order, the marginal variance of U and the concentration of U after conditioning on O, Σ_{OO}^{-1} is the concentration matrix of the observable variables after marginalizing on U and $\Sigma_{OO \cdot U}^{-1}$ is the concentration matrix of the observable variables after conditioning on U. Elements of each matrix are denoted as follows:

$$\Sigma_{OO}^{-1} = \begin{bmatrix} \sigma^{yy(u)} & * & * & * \\ \sigma^{xy(u)} & \sigma^{xx(u)} & * & * \\ \sigma^{zy(u)} & \sigma^{zx(u)} & \sigma^{zz(u)} & * \\ \sigma^{wy(u)} & \sigma^{wx(u)} & \sigma^{wz(u)} & \sigma^{ww(u)} \end{bmatrix},$$

$$\Sigma_{OO \cdot U}^{-1} = \begin{bmatrix} \sigma^{yy} & * & * & * \\ \sigma^{xy} & \sigma^{xx} & * & * \\ \sigma^{zy} & \sigma^{zx} & \sigma^{zz} & * \\ \sigma^{wy} & \sigma^{wx} & \sigma^{wz} & \sigma^{ww} \end{bmatrix},$$

and

$$\delta\delta^T = \begin{bmatrix} \delta_y\delta_y & * & * & * \\ \delta_x\delta_y & \delta_x\delta_x & * & * \\ \delta_z\delta_y & \delta_z\delta_x & \delta_z\delta_z & * \\ \delta_w\delta_y & \delta_w\delta_x & \delta_w\delta_z & \delta_w\delta_w \end{bmatrix},$$

from which we can see that, as an instance, that $\sigma^{xy(u)}$ is an element of the inverse of Σ_{OO}. From Condition 2 $\sigma^{zy} = 0$ and therefore $\delta_z\delta_y = \sigma^{zy(u)}$. Furthermore, from Condition 2 and 3(a) $\sigma^{wz} = 0$ and therefore $\delta_w\delta_z = \sigma^{wz(u)}$. Finally, from 3(a) $\sigma^{yw} = -\beta_{yw\cdot u}\sigma^{yy}$ and $\sigma^{xw} = \beta_{yw\cdot u}\sigma^{yy}\alpha_{yx}$, therefore:

$$\alpha_{yx} = -\frac{\sigma^{xw}}{\sigma^{yw}}. \tag{12.3}$$

We derive σ^{xw} first:

$$\begin{aligned}
\sigma^{xw} &= \sigma^{xw(u)} + \delta_x\delta_w, \quad \text{where} \\
\delta_x\delta_w &= \frac{\delta_x\delta_y\,\delta_z\delta_w}{\delta_z\delta_y} \\
&= \frac{\left(\sigma^{xy} - \sigma^{xy(u)}\right)\sigma^{zw(u)}}{\sigma^{zy(u)}} \\
&= \frac{\sigma^{xy}\sigma^{zw(u)} - \sigma^{xy(u)}\sigma^{zw(u)}}{\sigma^{zy(u)}}, \quad \text{and thus} \\
\sigma^{xw} &= \sigma^{xw(u)} + \frac{\sigma^{xy\cdot u}\sigma^{zw(u)} - \sigma^{xy(u)}\sigma^{zw(u)}}{\sigma^{zy(u)}} \\
&= \sigma^{xw(u)} - \frac{\sigma^{xy(u)}\sigma^{zw(u)}}{\sigma^{zy(u)}} + \frac{\sigma^{xy}\sigma^{zw(u)}}{\sigma^{zy(u)}} \\
&= \frac{\sigma^{xw(u)}\sigma^{zy(u)} - \sigma^{xy(u)}\sigma^{zw(u)}}{\sigma^{zy(u)}} + \frac{\sigma^{xy}\sigma^{zw(u)}}{\sigma^{zy(u)}}.
\end{aligned} \tag{12.4}$$

Now we do the same for σ^{yw}:

$$\begin{aligned}
\sigma^{yw\cdot u} &= \sigma^{yw(u)} + \delta_y\delta_w, \quad \text{where} \\
\delta_y\delta_w &= \frac{\delta_y\delta_y\,\delta_z\delta_w}{\delta_z\delta_y} \\
&= \frac{\left(\sigma^{yy\cdot u} - \sigma^{yy(u)}\right)\sigma^{zw(u)}}{\sigma^{zy(u)}} \\
&= \frac{\sigma^{yy\cdot u}\sigma^{zw(u)} - \sigma^{yy(u)}\sigma^{zw(u)}}{\sigma^{zy(u)}}, \quad \text{and thus} \\
\sigma^{yw\cdot u} &= \sigma^{yw(u)} + \frac{\sigma^{yy}\sigma^{zw(u)} - \sigma^{yy(u)}\sigma^{zw(u)}}{\sigma^{zy(u)}} \\
&= \sigma^{yw(u)} - \frac{\sigma^{yy(u)}\sigma^{zw(u)}}{\sigma^{zy(u)}} + \frac{\sigma^{yy}\sigma^{zw(u)}}{\sigma^{zy(u)}} \\
&= \frac{\sigma^{yw(u)}\sigma^{zy(u)} - \sigma^{yy(u)}\sigma^{zw(u)}}{\sigma^{zy(u)}} + \frac{\sigma^{yy}\sigma^{zw(u)}}{\sigma^{zy(u)}}.
\end{aligned} \tag{12.5}$$

Using (12.3)–(12.5) α_{yx} can be so derived:

$$\alpha_{yx} = -\frac{\sigma^{xw(u)} + \delta_x\delta_w}{\sigma^{yw(u)} + \delta_y\delta_w} \tag{12.6}$$

$$= -\frac{\sigma^{xw(u)}\sigma^{zy(u)} - \sigma^{xy(u)}\sigma^{zw(u)} + \sigma^{xy\cdot u}\sigma^{zw(u)}}{\sigma^{yw(u)}\sigma^{zy(u)} - \sigma^{yy(u)}\sigma^{zw(u)} + \sigma^{yy\cdot u}\sigma^{zw(u)}} \tag{12.7}$$

$$= -\frac{\sigma^{xw(u)}\sigma^{zy(u)} - \sigma^{xy(u)}\sigma^{zw(u)} - \alpha_{yx}\sigma^{yy}\sigma^{zw(u)}}{\sigma^{yw(u)}\sigma^{zy(u)} - \sigma^{yy(u)}\sigma^{zw(u)} + \sigma^{yy\cdot u}\sigma^{zw(u)}}, \tag{12.8}$$

and after some simple manipulations we have:

$$\alpha_{yx} = -\frac{\sigma^{xy(u)}\sigma^{zw(u)} - \sigma^{xw(u)}\sigma^{zy(u)}}{\sigma^{yw(u)}\sigma^{zy(u)} - \sigma^{yy(u)}\sigma^{zw(u)}}, \tag{12.9}$$

and the result follows after multiplication with: $\frac{\sigma^{ww(u)}}{\sigma^{yy(u)}}\frac{\sigma^{ww(u)}}{\sigma^{yy(u)}}$.

Appendix B: Proof of (8.3) and (8.4)

We first proof (8.4). With reference to (12.2) α_{yx} can be obtained by

$$\alpha_{yx} = -\frac{\sigma^{yx\cdot u}}{\sigma^{yy\cdot u}} = -\frac{\sigma^{yx(u)} + \delta_y\delta_x}{\sigma^{yy(u)} + \delta_y^2}.$$

From Conditions 2 and 3(b), it follows that $\sigma^{yz} = \sigma^{yw} = \sigma^{wz} = \sigma^{xz} = 0$, thus we can breakdown the *rhs* of formula above as follows:

$$\begin{aligned}
\sigma^{yx\cdot u} &= \sigma^{yx(u)} + \delta_y\delta_x \\
&= \sigma^{yx(u)} - \frac{\sigma^{yw(u)}\sigma^{xz(u)}}{\sigma^{zw(u)}} \\
&= \frac{\sigma^{yx(u)}\sigma^{zw(u)} - \sigma^{yw(u)}\sigma^{xz(u)}}{\sigma^{zw(u)}}, \\
\sigma^{yy\cdot u} &= \sigma^{yy(u)} + \delta_y^2 \\
&= \sigma^{yy(u)} - \frac{\sigma^{yw(u)}\sigma^{yz(u)}}{\sigma^{zw(u)}} \\
&= \frac{\sigma^{yy(u)}\sigma^{zw(u)} - \sigma^{yw(u)}\sigma^{yz(u)}}{\sigma^{zw(u)}},
\end{aligned}$$

and now we can obtain the parameter as follows:

$$\alpha_{yx} = -\frac{\sigma^{yx(u)}\sigma^{zw(u)} - \sigma^{yw(u)}\sigma^{xz(u)}}{\sigma^{yy(u)}\sigma^{zw(u)} - \sigma^{yw(u)}\sigma^{yz(u)}},$$

and if we multiply with $\frac{\sigma^{yy(u)}}{\sigma^{yy(u)}}\frac{\sigma^{zz(u)}}{\sigma^{zz(u)}}$ we arrive finally at (8.4).

Transform (8.4) into (8.3) by noting that

$$\begin{bmatrix}\beta_{yx.zw}\\ \beta_{yz.xw}\\ \beta_{yw.xz}\end{bmatrix}=\begin{bmatrix}\sigma_{xx} & \sigma_{xz} & \sigma_{xw}\\ \sigma_{zx} & \sigma_{zz} & \sigma_{zw}\\ \sigma_{wx} & \sigma_{wz} & \sigma_{ww}\end{bmatrix}^{-1}\begin{bmatrix}\sigma_{xy}\\ \sigma_{zy}\\ \sigma_{wy}\end{bmatrix},$$

and

$$\begin{bmatrix}\beta_{zw\cdot yx}\\ \beta_{zx\cdot yw}\\ \beta_{zy\cdot xw}\end{bmatrix}=\begin{bmatrix}\sigma_{ww} & \sigma_{wx} & \sigma_{wy}\\ \sigma_{xw} & \sigma_{xx} & \sigma_{xy}\\ \sigma_{yw} & \sigma_{yx} & \sigma_{yy}\end{bmatrix}^{-1}\begin{bmatrix}\sigma_{zw}\\ \sigma_{zx}\\ \sigma_{zy}\end{bmatrix}.$$

Therefore, after some tedious derivations, we have

$$\beta_{yx\cdot zw}=\frac{\sigma_{xy}\sigma_{zz}\sigma_{ww}-\sigma_{xy}\sigma_{wz}^2-\sigma_{zy}\sigma_{xz}\sigma_{ww}+\sigma_{zy}\sigma_{xw}\sigma_{wz}+\sigma_{wy}\sigma_{xz}\sigma_{wz}-\sigma_{wy}\sigma_{xw}\sigma_{zz}}{\sigma_{xx}\sigma_{zz}\sigma_{ww}-\sigma_{xx}\sigma_{wz}^2-\sigma_{xz}^2\sigma_{ww}+2\,\sigma_{xz}\sigma_{xw}\sigma_{wz}-\sigma_{xw}^2\sigma_{zz}},$$

$$\beta_{yw\cdot xz}=\frac{\sigma_{xy}\sigma_{xz}\sigma_{wz}-\sigma_{xy}\sigma_{xw}\sigma_{zz}-\sigma_{zy}\sigma_{xx}\sigma_{wz}+\sigma_{zy}\sigma_{xz}\sigma_{xw}+\sigma_{wy}\sigma_{xx}\sigma_{zz}-\sigma_{wy}\sigma_{xz}^2}{\sigma_{xx}\sigma_{zz}\sigma_{ww}-\sigma_{xx}\sigma_{wz}^2-\sigma_{xz}^2\sigma_{ww}+2\,\sigma_{xz}\sigma_{xw}\sigma_{wz}-\sigma_{xw}^2\sigma_{zz}},$$

$$\beta_{zw\cdot yx}=-\frac{\sigma_{xz}\sigma_{xw}\sigma_{yy}-\sigma_{xz}\sigma_{xy}\sigma_{wy}-\sigma_{wz}\sigma_{xx}\sigma_{yy}+\sigma_{wz}\sigma_{xy}^2-\sigma_{yz}\sigma_{xy}\sigma_{xw}+\sigma_{yz}\sigma_{wy}\sigma_{xx}}{\sigma_{xx}\sigma_{ww}\sigma_{yy}-\sigma_{xx}\sigma_{wy}^2-\sigma_{xw}^2\sigma_{yy}+2\,\sigma_{xw}\sigma_{xy}\sigma_{wy}-\sigma_{xy}^2\sigma_{ww}},$$

$$\beta_{zx\cdot yw}=-\frac{-\sigma_{xz}\sigma_{ww}\sigma_{yy}+\sigma_{xz}\sigma_{wy}^2+\sigma_{wz}\sigma_{xw}\sigma_{yy}-\sigma_{wz}\sigma_{xy}\sigma_{wy}+\sigma_{yz}\sigma_{xy}\sigma_{ww}-\sigma_{yz}\sigma_{wy}\sigma_{xw}}{\sigma_{xx}\sigma_{ww}\sigma_{yy}-\sigma_{xx}\sigma_{wy}^2-\sigma_{xw}^2\sigma_{yy}+2\,\sigma_{xw}\sigma_{xy}\sigma_{wy}-\sigma_{xy}^2\sigma_{ww}},$$

$$\beta_{zy\cdot xw}=\frac{-\sigma_{xz}\sigma_{xy}\sigma_{ww}+\sigma_{xz}\sigma_{wy}\sigma_{xw}+\sigma_{wz}\sigma_{xy}\sigma_{xw}-\sigma_{wz}\sigma_{wy}\sigma_{xx}+\sigma_{yz}\sigma_{xx}\sigma_{ww}-\sigma_{yz}\sigma_{xw}^2}{\sigma_{xx}\sigma_{ww}\sigma_{yy}-\sigma_{xx}\sigma_{wy}^2-\sigma_{xw}{}^2\sigma_{yy}+2\,\sigma_{xw}\sigma_{xy}\sigma_{wy}-\sigma_{xy}^2\sigma_{ww}},$$

and applying these expressions to (8.4) we arrive at

$$\alpha_{yx}=\frac{\sigma_{wz}\sigma_{xy}-\sigma_{xz}\sigma_{wy}}{\sigma_{xx}\sigma_{wz}-\sigma_{xz}\sigma_{xw}},$$

which can be transformed into (8.3) by multiplication with $\frac{\sigma_{ww}}{\sigma_{ww}}\frac{\sigma_{xx}}{\sigma_{xx}}$.

References

Allman ES, Matias C, Rhodes JA (2009) Identifiability parameters in latent structure models with many observed variables. Ann Stat 37:3099–3132

Bollen K (1989) Structural equation models. Wiley, New York

Bowden R (1973) The theory of parametric identification. Econometrica 41:1069–1074

Bowden R, Darrel AT (1984) Instrumental variables. Cambridge University Press, Cambridge

Brito C, Pearl J (2002) Generalized instrumental variables, Technical report, R-303

Cai Z, Kuroki M (2012) On identifying total effects in the presence of latent variables and selection bias. In: Proceedings of the twenty-fourth conference on uncertainty in artificial intelligence (UAI-08), pp 62–69

Chan H, Kuroki M (2010) Using descendants as instrumental variables for the identification of directed causal effects in linear SEMs. J Mach Learn Res 9:73–80

Coleman JS, Hoffer T, Kilgore S (1982) High school achievement: public, catholic, and private schools compared. Basic Books, New York
Cox DR, Wermuth N (1996) Multivariate dependencies-models, analysis and interpretation. Chapman & Hall, London
Frydenberg M (1990) The chain graph Markov property. Scand J Stat 17:333–353
Huang Y, Valtorta M (2006) Pearls calculus of intervention is complete. In: Proceedings of the twenty-second conference on uncertainty in artificial intelligence (UAI-06), pp 217–224
Kuroki M (2007) Graphical identifiability criteria for causal effect in studies with an unobserved treatment/response variable. Biometrika 94:37–47
Kuroki M, Pearl J (2014) Measurement bias and effect restoration in causal inference. Biometrika 101:423–437
Lauritzen SL (1996) Graphical models. Oxford Science Publication, Oxford
Lauritzen SL (2001) Causal inference from graphical models. In: Barndoff-Nielsen OE, Cox DR, Klüpperberg C (eds) Complex stochastic system. CRC Press, London, pp 63–107
Morgan SL, Winship C (2007) Counterfactuals and causal inference: methods and principles for social research. Cambridge University Press, Cambridge
Pakpahan E (2012) Essays on causal inference in Gaussian graphical models. PhD thesis. University of Perugia. Available from the author by emailing to epakpahan@gmail.com
Pearl J (1995) Causal diagrams for empirical research. Biometrika 82:669–709
Pearl J (2001) Statistics and causal inference: a review. TEST 12:281–345
Pearl J (2009) Causality: models, reasoning, and inference, 2nd edn. Cambridge University Press, Cambridge
Rothenberg TJ (1971) Identification in parametric models. Econometrica 39:577–591
Shpitser I, Pearl J (2008) Complete identification methods for the causal hierarchy. J Mach Learn Res 9:1941–1979
Stanghellini E (2004) Instrumental variables in Gaussian directed acyclic graph models with an unobserved confounder. Environmetrics 15:463–469
Stanghellini E, Vantaggi B (2013) On the identification of discrete concentration graph models with one hidden binary variable. Bernoulli 19:1920–1937
Stanghellini E, Wermuth N (2005) On the identification of path analysis models with one hidden variable. Biometrika 92:337–350
Theil H (1953) Repeated least square applied to complete equation systems. Central Planning Bureau, The Hague, The Netherlands
Tian J, Pearl J (2002) A general identification condition for causal effects. In: Proceedings of the eighteenth national conference on artificial intelligence, AAAI Press/The MIT Press: Menlo Park, CA, pp 567–573
Verma T, Pearl J (1991) Equivalence and synthesis of causal models. In: Proceedings of the sixth annual conference on uncertainty in artificial intelligence (UAI 90), pp 255–270
Wermuth N, Cox DR (2008) Distortion of effects caused by indirect confounding. Biometrika 95:17–33
Wermuth N, Sadeghi K (2011) Sequences of regressions and their independences. TEST 21:215–252