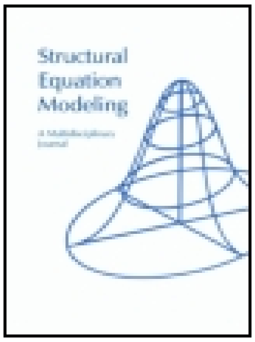

Structural Equation Modeling: A Multidisciplinary Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hsem20

Testing piecewise structural equations models in the presence of latent variables and including correlated errors

Bill Shipley & Jacob C. Douma

To cite this article: Bill Shipley & Jacob C. Douma (2021): Testing piecewise structural equations models in the presence of latent variables and including correlated errors, Structural Equation Modeling: A Multidisciplinary Journal, DOI: 10.1080/10705511.2020.1871355

To link to this article: https://doi.org/10.1080/10705511.2020.1871355

Published online: 04 Feb 2021.

STRUCTURAL EQUATION MODELING: A MULTIDISCIPLINARY JOURNAL
https://doi.org/10.1080/10705511.2020.1871355

Testing piecewise structural equations models in the presence of latent variables and including correlated errors

Bill Shipley[a] and Jacob C. Douma[b]

[a]Université de Sherbrooke; [b]Wageningen University

ABSTRACT

Path models, expressed as Directed Acyclic Graphs (DAGs), and the testing of such DAGs via a d-sep test, have become popular because they can incorporate complicated data structures that are difficult or impossible to accommodate in classical structural equation modeling. However, d-sep tests cannot accommodate DAGs that include unmeasured (latent) variables. We describe (i) how to convert a DAG with latent variables into an observationally equivalent graph without latents (a Mixed Acyclic Graph, MAG), (ii) how this MAG identifies which latents can/cannot be ignored without changing the causal meaning of the original DAG, and (iii) how to perform the MAG equivalent of a d-sep test.

KEYWORDS

Causal hypothesis; correlated errors; d-separation; d-sep tests; latent variables; m-separation; m-sep test; mixed graphs; structural causal modeling

Introduction

A common goal in many disciplines is to pose hypotheses involving a network of direct and indirect causal relationships between several variables and then to conduct falsifiable empirical tests of these hypotheses. Ideally, this is done using controlled manipulative experiments but such controlled manipulative experiments are often not possible. Structural equation modeling (SEM) is an alternative method, based on statistical rather than physical control of variables but using the same inferential logic. The classical version of SEM (Bollen, 1989) involves maximizing the likelihood of a series of linear equations involving normally distributed random variables measured over mutually independent observations. Although there are many variants of this classical method, they are "simultaneous" and "global" because both parameter estimation and model testing are done simultaneously over the entire (global) set of structural equations. Although much effort has gone into relaxing the statistical assumptions of such global estimation methods, it is difficult (sometimes impossible) to accommodate combinations of small sample sizes, non-normal distributions, nesting or cross-classification, and nonlinear relationships between variables that are often encountered and this is much easier with local estimation methods (Shipley, 2009; Shipley & Douma, 2020).

However, an important advantage of classical SEM over the local method described next is that, under certain conditions, it is possible to include "latent" variables in the causal model; i.e. variables that have not been directly observed or measured. When facing complex hierarchical designs or other complex data structures, one has to choose between local estimation methods that can take these complex data structures into account but cannot include latent variables or classical SEM that can include latent variables but perhaps not the complex data structures. In some cases, we can safely ignore latents. For instance, since a path diagram such as X→Y→Z is implicitly ignoring the causes (say, L_1) of X and additional intervening variables (say L_2) between X and Y, we can further augment the initial DAG: L_1→X→L_2→Y→Z. Whenever we conduct a test on X→Y→Z we are implicitly assuming that any latent variables that might be lurking in nature, but that are not included in it (here, L_1, L_2), will not change the independence relationships among the variables of interest (X, Y, Z). When is this assumption reasonable? When is it not reasonable? When can we safely ignore latents while maintaining the causal assumptions among the observed variables and when not? If we ignore latents then how does this affect the relationships among the measured variables? We answer this question by converting DAGs containing latent variables into Mixed Acyclic Graphs that omit these latents while maintaining all of the (conditional) independence relationships among the observed variables.

Shipley (2000, 2009, 2016) developed an alternative method of SEM that is based on Directed Acyclic Graphs (DAGs) and the graph theoretic notion of "d-separation" (Pearl, 2009; Pearl & Mackenzie, 2018). Although this method has been applied mostly in the fields of ecology and evolution in situations not requiring latent variables, it has also been applied in medicine (Buffart et al., 2018; Gordon et al., 2014), psychology (Thoemmes et al., 2018; Van Kampen, 2014) and sociology (Schweiger & Cress, 2019; Warach et al., 2018) when latent variables were not invoked. A DAG is a graphical depiction of the direct cause-effect links between variables (X→Y) and d-separation is defined below. It is local (or "piecewise" (Lefcheck, 2016)) rather than global because a local Markov decomposition of the overall model-implied multivariate probability distribution results in a series of "local" parent–child causal links that are estimated separately from one another. It is sequential rather than simultaneous because the causal hypotheses implied by the model (the DAG) are first tested via a "d-sep" test while the statistical fitting of the causal links

CONTACT Bill Shipley Bill.Shipley@USherbrooke.ca Département de Biologie, Université de Sherbrooke, Sherbrooke QC J1K 2R9, Canada.

occurs after this step and only if the d-sep test does not detect any significant lack-of-fit. The causal hypotheses are captured by a "basis set" of d-separation claims, which is defined later. One particular basis set, the "union" basis set, has the property (Shipley, 2000) that the statistical tests of each of the k d-separation claims in the set are mutually independent, allowing the null probabilities of each test ($p_1, \ldots, p_k$) to be combined into an omnibus test of the full model using Fisher's C statistic, $C = -2\sum_{i=1}^{k} p_i \ln p_i$, which is distributed as a chi-squared variate with 2k degrees of freedom given the null hypotheses. The main advantage of this method is that it can easily incorporate non-normally distributed variables, complex nesting and cross-classification, as well as nonlinear functional relationships. However, the d-sep test cannot be applied to DAGs involving latent variables. Here we first present a method of modifying a DAG that includes latent variables such that the modified graph, a Mixed Acyclic Graph (MAG), correctly captures the dependence and independence relationships of the observed variables in the original DAG, and then modify the original d-sep test based on such a MAG. We do not discuss parameter estimation, which occurs only after the causal topology of the observed variables in the DAG has been tested and not rejected.

DAGs and d-separation

A DAG (G) is a mathematical object consisting of vertices (variables) and directed edges (arrows) between pairs of vertices. A *path between α and β* in G is a sequence of vertices having α and β as endpoints that are connected by arrows, irrespective of the direction of these arrows. For instance, A→X←L→Y→B is a path between vertices A and B in DAG (I) in Figure 1. A *directed path from α to β* is a sequence of vertices, starting at α and ending at β, which allows one to move from α to β while respecting the direction of the arrows. In such a directed path, α is an *ancestor* of β and β is a *descendent* of α. In DAG (I) of Figure 1 there is a directed path from L to B (L→Y→B) in which L is an ancestor of B and B is a descendent of L. A DAG is "acyclic" because there can be no directed paths from a vertex that loops back into it (i.e. no cycles). A *collider vertex along a path* is a vertex along a path having arrows pointing into it from both directions; a vertex along a path that is not a collider is a *non-collider*. X is a collider, while L and Y are non-colliders, along the path A→X←L→Y→B between vertices A and B in DAG (I) of Figure 1. A vertex can be a collider along one path and a non-collider along a different path. Given a DAG G, a path p between two vertices (α, β) in G is *d-separated* (Verma & Pearl, 1990), or blocked, by a set of other vertices **Z** (which can be the empty set) if and only if:

(i) p contains a chain i→m→j or a fork i←m→j such that the middle vertex m is in the set **Z**, or
(ii) p contains a collider i→m←j such that the middle vertex m is not in the set **Z** and such that no descendent of m is in **Z**.

The two vertices (α, β) are d-separated (or blocked) by the set **Z** if the very path between them is d-separated.

If the vertices of a DAG represent random variables then the full DAG represents the topological structure of the process generating the multivariate probability distribution (or density) over these random variables. More specifically, the DAG expresses a multivariate causal hypothesis concerning this generating process in nature (Cox & Wermuth, 1996). If two vertices (X,Y) are d-separated given a set **Z** of other vertices in a DAG then the associated random variables (X,Y) will be statistically independent given the set **Z** in the resulting multivariate probability distribution that is generated by the DAG (Pearl, 2009, theorem 1.2.5). This is true irrespective of the actual distributional form of the random variables or of the functional form (linear or nonlinear) of the links representing the arrows between the random variables. Thus, DAGs and d-separation allow us to use the logic of the controlled experiment while replacing experimental control of variables with statistical control. One first expresses the multivariate causal hypothesis as a DAG. The operation of d-separation deduces how the pattern of dependence and independence between every pair of variables in the causal system will change as one statistically (rather than physically) holds constant any combination of other variables. Finally, one compares the full pattern of observed and predicted independencies to the observational data using the d-sep statistical test.

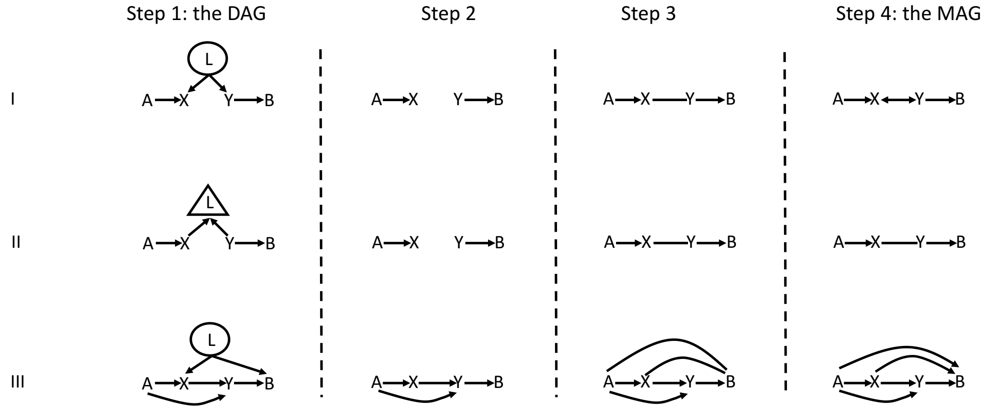

Figure 1. Three different DAGS (I, II, III) and the steps by which these DAGs are converted to a mixed acyclic graph (MAG). Observed variables are {A, X, Y, B} ∈ **O** and the latent variable is L ∈ **L**. Sampling independently of the value of L represents marginalizing (open circle, L ∈ $\mathbf{L}_M$) while sampling conditional on the value of L represents implicit conditioning (open triangle, L ∈ $\mathbf{L}_C$).

The d-sep test provides an empirical test of the independence relationships implied by a DAG. It does this by concentrating on a basis set of d-separation claims of a DAG, which is the smallest set of d-separation claims implied by the DAG and from which all other d-separation claims can be deduced. Each d-separation claim is associated with a bivariate conditional probability distribution. If X is d-separated from Y conditional on a set **Z** in the DAG, then this implies that X is probabilistically independent of Y conditional on the set **Z** in any data generated by the DAG, i.e. that the conditional bivariate distribution $P(X,Y|\mathbf{Z})$ generated by the DAG can be decomposed as $P(X|\mathbf{Z}) \times P(Y|\mathbf{Z})$. One must calculate the null probability of observing such conditional independence for each d-separation claim in the basis set and then combine these to produce the null probability of observing all of the d-separation claims in the DAG. An important advantage of d-sep tests over classical SEM is that one can use whatever test of conditional independence is appropriate for the nature of the variables involved in each d-separation claim. However, testing this claim requires that we have measured each of X, Y and **Z**. We cannot do this if any of these variables are unobserved (i.e. latent). We must modify the d-sep test so that it tests all and only those independence claims of the full DAG (including latents) that involve the observed variables of the full DAG. This requires transforming the DAG into a new type of graph that we call a Mixed Acyclic Graph (MAG) because, as will be seen, this new acyclic graph contains a mixture of directed (→), undirected (–) and bidirected (↔) edges involving only the observed variables.

Transforming a DAG into a MAG

Step 1. Start with a DAG representing the hypothesized data-generating mechanism. Consider a DAG G containing a set **V** of vertices that can be classified into two non-overlapping subsets: **O** and **L**: $\mathbf{V} = \mathbf{O} \cup \mathbf{L}$. The set **O** ("observed" variables) contains variables for which we have measurements. The set **L** ("latent" variables) contains variables for which we do not have measurements. The steps are illustrated by the three DAGs in Figure 1, each with vertices $\mathbf{O}$ = {A,X,Y,B} and $\mathbf{L}$ = {L}. The set **L** is further subdivided into latents that are to be implicitly marginalized ($\mathbf{L}_M$) and implicitly conditioned ($\mathbf{L}_C$) in the multivariate probability distribution or density.

Marginalizing over a variable (Z) in a multivariate probability distribution P(X, Z) means summing (or integrating) the probabilities of X over all values of Z: $P(X) = \sum_i P(X, Z_i)$. Imagine that you randomly sample observations but only measure variable X. You do not measure variable Z and may even not be aware of its existence. However, you sample in such a way that the inclusion (or not) of an observation in the sample is independent of the value of Z; in other words, the sampling protocol is not biased with respect to Z. If so, then then you have *implicitly* marginalized over Z when obtaining P(X). Conditioning on a variable (Z) in a multivariate probability distribution P(X, Z) means evaluating the probability of X after knowing the value of Z. Imagine again that you randomly sample observations and only measure variable X while ignoring Z but you sample in such a way that the inclusion (or not) of an observation in the sample is affected by the value of Z; in other words, the sampling protocol is biased with respect to Z. If so, then then you have *implicitly* conditioned over Z when obtaining P(X).

The following examples will make this distinction clearer based on a data-generating process in nature that involves two observed traits (X, Y) measured on a series of organisms. These two traits are causally independent of one another but each affects the status (L, i.e. dead/alive) of the adult organism so that the associated DAG is X→L←Y. We randomly sample these organisms without reference to their status (L), and the chance of either living or dead adults being included in our random sample is proportional to their occurrence in nature. This would represent an *implicit* marginalization of the joint probability distribution of the two observed variables (X, Y) over the latent variable L even though we did not actually record the status of our organisms. From d-separation, X and Y would remain independent. However, if we cannot find dead adult organisms (perhaps because they no longer exist) then our sample will only include living ones. If so then we have *implicitly* conditioned on L because, in our sample, the value of this latent is fixed (L = alive). This would represent a conditional joint probability distribution of X and Y given L. Even though X and Y are independent in the DAG, they would be dependent in this conditional distribution because conditioning on L, which is a collider variable along the path X→L←Y, opens up this path. This is called "selection bias" or Berkson's paradox (Berkson, 1946). This is true even if the sampling criterion is not exact; for instance, if our sampling design is only biased toward living organisms because we have more difficulty finding dead ones, then this still generates a conditional joint probability distribution of X and Y given L.

Step 2. Create a new graph G' containing only the vertices in **O**. The new graph G' will contain all and only the vertices **O** in G. In Figure 1 this is the set **O** = {A,X,Y,B}. Add arrows in the new graph G' corresponding to the arrows between pairs of adjacent observed variables in G. "Adjacent" vertices or variables are ones jointed by an arrow in G.

Step 3. If there are two vertices $(X, Y) \in \mathbf{O}$ in G' after step 2 that are not adjacent in G, but that are not d-separated in G given every possible subset of other vertices (i.e. are d-connected) excluding the vertices in $\mathbf{L}_M$ but including the vertices in $\mathbf{L}_C$, then add an undirected edge between X and Y (i.e. X – Y) in the new graph G'.

Consider the three DAGS in Figure 1. The two versions of G' that are associated with DAGs (I) and (II) have an undirected edge between X and Y (X – Y) after step 3. This is because they will always be d-connected given every possible combination of observed conditioning sets that excludes L in DAG (I) (marginalizing over L) and given every possible combination of observed conditioning sets including L in DAG (II) (implicit conditioning on L). The undirected edge (X – B) in G' after step 3 between X and B in DAG (III) arises because both share a common latent cause (L) that is marginalized. The undirected edge in G' after step 3 between A and B (A – B) in DAG (III) is less obvious. This undirected edge exists because A and B will be d-connected through an open path given every possible conditioning set of other observed variables that do not include L: (i) A→X→Y→B and A→Y→B unconditionally, (ii) A→Y←X←L→B given {Y}; (iii) A→X←L→B given {X}; and

(iv) A→X←L→B given {X, Y}. A and B given Y are d-connected because Y is a child of the collider X (rule 2).

Step 4. Richardson and Spirtes (2002) define a transformation of an "ancestral graph" G, which includes DAGs (their proposition 3.4), into the graph G' that is produced at the end of step 3. Richardson and Spirtes (2002) further prove that undirected edges (X – Y) in G' represent the result of marginalizing or conditioning on latent variables in the DAG. They distinguish between "ancestral" vertices and "anterior" vertices in ancestral graphs (including both the DAG G and the transformed graph G'). X is an "ancestor" of Y in such graphs if there is at least one directed path from X to Y ($X \to \cdots \to Y$) or if X and Y are the same vertex. Note that this definition of an "ancestor" differs from that used by Pearl (2009) since, in that publication, a vertex is ancestral only with respect to a specific path. We will use the expressions "*ancestor in the graph*" and "*ancestor along the path*" to differentiate the two meanings of "ancestor". X is "anterior" to Y if there is at least one path from X to Y in which every edge is either of type X – W or X→W; here, W is an observed variable along such a path between X and Y. In other words, there can be no edge X↔W or X←W along such a path. In the context of DAGs, if X is anterior to Y then it is also ancestral to Y, but this is not necessarily the case for graphs like G'. Richardson and Spirtes (2002, lemma 3.9) prove the following orientation rule for orienting an edge (X – Y) in G' given the DAG G:

(i) orient X – Y as X→Y in G' if X is anterior (and ancestral) to Y in G and Y is not anterior (nor ancestral) to X in G.
(ii) orient X – Y as X←Y in G' if Y is anterior (and ancestral) to X in G and X is not anterior (nor ancestral) to Y in G.
(iii) orient X – Y as X↔Y in G' if neither X nor Y are anterior (nor ancestral) to the other in G.
(iv) keep X – Y oriented as X – Y in G' if neither X nor Y are anterior (nor ancestral) to the other in G but both X and Y are anterior (and ancestral) of the same latent causal descendent in the DAG G and some value of this latent variable defines a selection criterion for inclusion in the statistical population (i.e. conditioning or a common latent ancestor).

The application of these orientation rules leads to step 4 in constructing G'. The result is a mixed acyclic graph (MAG) containing any combination of the following edges (–, ←,→ or ↔). The interpretation of the two new edges in a MAG (i.e. – and ↔) that do not exist in a DAG is given later.

For example, in DAG (I) of Figure 1 there is an undirected edge (X – Y) in the graph G' associated with it after step 3. Since X is not anterior (ancestral) to Y, nor is Y anterior (ancestral) to X, this undirected edge is oriented as X↔Y in the MAG. The undirected edge (X – Y) in DAG (II) remains oriented as X – Y in the MAG since neither is anterior to the other in the DAG and the d-connection between them is generated by implicit conditioning on the latent variable (L) that is a common effect of both X and Y. The DAG (III) has two undirected edges in the graph G' associated with it after step 3: (A – B) and (X – B). However, A is anterior (ancestral) to B in the DAG via two different the directed paths (A→Y→B and A→X→Y→B) while B is not anterior (ancestral) to A in the DAG by any directed path. Therefore, the undirected path (A – B) is oriented as A→B in the resulting MAG. Similarly, X is anterior (ancestral) to B in the DAG via the directed path X→Y→B while B is not anterior (ancestral) to X in the DAG by any directed path. Therefore, the undirected path (X – B) is oriented as X→B in the resulting MAG. The MAG.to.DAG function of the CauseAndCorrelation R library (https://github.com/BillShipley/CauseAndCorrelation) performs the translation of a DAG with latents into its associated MAG.

Obtaining conditional independence claims from a MAG

Given a DAG involving a set **V** of vertices, Richardson and Spirtes (2002) define an "independence model" associated with it that consists of the full set of conditional independence relations implied by it. Each conditional independence relation is a triple, $\mathrm{I}(\alpha, \beta|\theta a)$, stating that vertex α is independent of vertex β, conditional on a set of vertices $\theta\mathbf{a}$; $\theta\mathbf{a}$ can be any subset of **V** (including the null subset) that does not include α or β. If some of the vertices in the DAG include latent vertices then marginalizing or implicitly conditioning over this set of latent vertices produces the associated MAG. The independence model of this MAG is the subset of triples of independence relations implied by the DAG after marginalizing and/or conditioning on latents. Each independence relation $\mathrm{I}(\mathrm{X}, \mathrm{Y}|\mathbf{Z} \cup \mathbf{L}_\mathrm{C})$ is a triple involving two observed vertices (X, Y) conditional on a set of observed vertices **Z** (excluding X and Y) plus $\mathbf{L}_\mathrm{C}$; both **Z** and $\mathbf{L}_\mathrm{C}$ can be empty sets. The union basis set of d-separation claims of a DAG that is defined in Shipley (2000) and used in the d-sep test of DAGs is an example of such an independence model. The union basis set consists of a set of independence relations $\mathrm{I}(\mathrm{X}, \mathrm{Y}|\mathbf{Z} \cup \mathbf{L}_\mathrm{C})$ such that each independence relation involves a pair of variables (X, Y) in the MAG that are not adjacent (i.e. do not have an edge between them), conditional on the set (**Z**) of observed causal parents of either X or Y plus all latent conditioning variables ($\mathbf{L}_\mathrm{C}$). An important result of Richardson and Spirtes (2002, theorem 4.18) is that the independence model corresponding to a DAG G, after marginalizing and/or conditioning on latents, is the independence model of the transformation of the DAG G into the MAG G'. In other words, we can test all of (and only) the independence relations involving *observed* variables implied by a DAG G that contains latent variables by obtaining the independence relations of the transformed MAG G'. Independence of vertices in a DAG is determined by the operation of d-separation. Independence of vertices in a MAG is determined by an extension of the d-separation operation, called "m-separation" by Richardson and Spirtes (2002).

M-separation in a MAG

Let a path, p_i, between any two vertices (α, β) in a MAG G' be a sequence of adjacent vertices linking α and β irrespective of the type of edge connecting adjacent pairs. A collider vertex along this path is a vertex that has an arrowhead pointing into it from both directions. Thus, a collider vertex along a path in a MAG is a vertex δ having orientations of types (i) $\rightarrow\delta\leftarrow$, (ii) $\rightarrow\delta\leftrightarrow$, (iii) $\leftrightarrow\delta\leftarrow$ or (iv) $\leftrightarrow\delta\leftrightarrow$. For instance, A→X↔Y is a path between A and Y in the MAG of DAG (I) in Figure 1 and X is a collider along this path. A vertex along a given path that is not a collider is a non-collider along this path. A path between two vertices α and β along path p_i in a MAG is "m-connecting" given a set **Z** (possibly empty, but not including α or β), if:

(i) Every non-collider along path p_i is not in **Z**, and
(ii) Every collider along path p_i is in **Z**, or is anterior to a member of **Z**.

Two variables are "m-separated" in a MAG if there are no m-connecting paths between them. Thus, m-separation is simply an extension of Pearl's d-separation operation, which is only defined for DAGs (i.e. without edges X – Y or X↔Y), to this wider class of mixed acyclic graphs. Richardson and Spirtes (2002, theorem 6.3) prove that, given a DAG G involving latent variables, observed variables X and Y are m-separated in the associated MAG G' given some subset of remaining observed variables **Z** if and only if X and Y are d-separated in G given **Z** plus the subset $\mathbf{L}_C$ of latent variables that are implicitly conditioned.

For example, consider the DAG (I) in Figure 1: A→X←L→Y→B with L being latent. The independence model of the DAG (i.e. the union basis set) is given in the first column of Table 1. Marginalizing over L produces the associated MAG: A→X↔Y→B. The independence model of the MAG, based on m-separation claims and using the union basis set, is given in the second column of Table 1. Of the six independence claims in the basis set of the DAG, there are three independence claims, involving only the observed variables **O** = {A,X,Y,B} of the DAG, in the basis set of the MAG. If any of these three independence claims from the MAG are rejected by the empirical data then we can reject the original causal hypothesis represented by the DAG.

It is important to emphasize that an empirical test of the m-separation claims of an MAG are tests of the independence constraints on the joint probability distribution over the observed variables. It is possible, depending on the additional statistical assumptions that one is willing to make, that the causal structure will additionally imply non-independence constraints on this joint probability distribution. For instance, it is well-known that the classic measurement model of SEM, in which a latent variable is a common cause of a series of observed indicator variables, implies non-independence constraints on the covariance matrix of these observed variables (Shipley, 2016, chapter 5). Such non-independence constraints are not reflected in the basis set of independence claims of the MAG, although they are reflected in the independence claims of the DAG via vanishing tetrad constraints (Spirtes et al., 1993, theorem 6.10). As well, an empirical test of a MAG is not a complete test of the independence claims in the basis set of the original DAG. Rather, it evaluates the *empirically testable* independence claims of the DAG involving the measured variables. The basis set of an MAG will always contain fewer elements that does the basis set of the DAG upon which it is based if the DAG contains latents. It is also possible that there are no empirically testable independence claims of the DAG without invoking the latent variables. The DAG (III) in Figure 1 is an example of this possibility since no observed variable is m-separated from any other observed variable in the MAG associated with it.

Table 1. A comparison of the d-separation claims of the original DAG (A→X←L→Y→B) and the m-separation claims of the resulting MAG (A→X↔Y→B). I(α,β,**Z**) means "vertex α is independent of vertex β, conditional on the set **Z** of vertices".

d-separation claims of the DAG:	m-separation claims of the transformed MAG:
I(A,L|null)	
I(A,Y|L)	I(A,Y|null)
I(A,B|Y)	I(A,B|Y)
I(X,Y|A,L)	
I(X,B|A,Y,L)	I(X,B|A,Y)
I(L,B|Y)	

An m-sep test for DAGs involving latent variables

We can now extend the d-sep test of DAGs involving only observed variables to the equivalent test of DAGs involving latent variables. Since this involves m-separation of a MAG, rather than d-separation of a DAG, we call it an "m-sep" test, but the steps are identical to a d-sep test after replacing the d-separation operation by the m-separation operation. The steps are as follows:

(1) Express your causal hypothesis in the form of a DAG G.
(2) Identify (i) the vertices in G that are observed (**O**), (ii) the latents whose values do not implicitly affect the sampling of observational units and so will be marginalized ($\mathbf{L}_M$), and (iii) those latents whose values do implicitly affect the sampling of observational units and so will be conditioned ($\mathbf{L}_C$).
(3) Convert your DAG into a MAG by marginalizing and/or conditioning on the appropriate latents.
(4) Determine the m-separation claims of the MAG that define the union basis set; i.e. the independence model.
(5) Convert each of the m-separation claims in the basis set into claims of conditional independence in the empirical data.
(6) Calculate the null probability, p_i, associated with each of these claims of conditional independence.
(7) Combine these null probabilities via Fisher's C-statistic: $C = -2\sum_{i=1}^{k} p_i \ln(p_i)$
(8) Compare the resulting C-statistic to a chi-squared distribution with degrees of freedom equal to 2k, where k is the number of m-separation claims in the basis set of G'.

An R function to perform steps 1–4 (basiSet.mag) is available in the CauseAndCorrelation R library (https://github.com/BillShipley/CauseAndCorrelation). If the null probability of the C-statistic is below your significance level, reject your causal hypothesis; if not, then provisionally accept the causal hypothesis and proceed to the estimation of parameters in each of the local relationships, defined by each endogenous variable and those other variables to which it is connected via an edge. The method of parameter estimation can differ for different local relationships depending on the nature of the variables involved and on the type of edges linking them.

Discussion

In this paper, we have outlined a method for empirically testing multivariate causal hypotheses (DAGs) involving variables that are not observed. Two questions are pertinent. First, when can we ignore latent variables in our causal hypothesis? Second, when should we prefer this method to classical SEM?

A directed edge (X→Y) in a DAG, when expressed as a causal hypothesis, is simply a claim that a change in the value of X will provoke a change in the value of Y even when every combination of other variables in the DAG, including the null or empty combination, are held constant (either physically or statistically). It is never a claim that X will provoke a change in the value of Y even when every possible combination of other variables that might exist in nature are held constant since the number of such possible variables is uncountably large. Every mechanistic description of a natural process can be rendered more complex and so researchers must choose those variables to include in their causal hypothesis that are relevant to the level of complexity at which they are working. Stated equivalently, researchers always ignore latent variables when proposing DAGs as causal hypotheses. Imagine that a researcher proposes the following DAG as a description of a causal hypothesis involving three variables (X, Y, Z): G_1:X→Y→Z. This causal description ignores, amongst others, the latent causes of X (L_1), intervening latent variables between X and Y (L_2) and the latent effects of Z (L_3) in the more complete DAG G_2:L_1→X→L_2→Y→Z→L_3. We intuitively understand that ignoring these three latent variables does not change the causal hypothesis of interest involving X, Y and Z. Is this intuition correct? If our sampling protocol is not biased with respect to these latents then the MAG that results from marginalizing over them is X→Y→Z; the MAG of G_2 equals the DAG G_1. This confirms our intuition that ignoring these latents (i.e. marginalizing over them) does not change the conditional independence relationships involving the variables (X, Y, Z) of interest and encoded in the MAG of G_2. However, it is clear that in other situations the result of ignoring latents in a more complete DAG will change the (conditional) independence relationships involving the observed variables of interest.

Consider now a more complicated example from the ecological literature which is slightly modified from Juhasz et al. (2020, online supporting information). That paper proposed the following causal hypothesis (Figure 2a) linking annual climate variation in the Canadian Arctic (winter, spring and summer annual Arctic climate oscillation plus summer mean temperature and precipitation each year for 21 years) with the abundance of lemmings and the reproductive success of arctic foxes and snow geese. Since lemmings are the preferred prey of foxes then, when lemming abundance is high, the foxes will preferentially eat lemmings rather than goose eggs. If so, then the consumption rate of lemmings by foxes increases with increasing lemming abundance. By increasing their consumption of lemmings, this would reduce the consumption rate of goose eggs by foxes, which would improve the nesting success of geese. At the same time, a summer that is warmer and has more precipitation should increase plant production, which would improve food availability for geese and hence their incubation attentiveness, which would decrease both fox predation opportunity and success (*i.e., consumption rate of goose eggs*). A decrease in the rate of goose egg consumption would improve the nesting success of the geese. Finally, the annual arctic oscillation (AO), by influencing several interrelated climatic factors, could affect lemming reproduction and survival during the winter, which drives lemming abundance during the following summer. The spring AO could influence the nesting success of the geese by modifying the environment of goose breeders (food accessibility, space for nest).

It is clearly impossible to conduct a controlled manipulative experiment to test this hypothesis over the scale of the Canadian Arctic. Furthermore, four of the variables (circled)

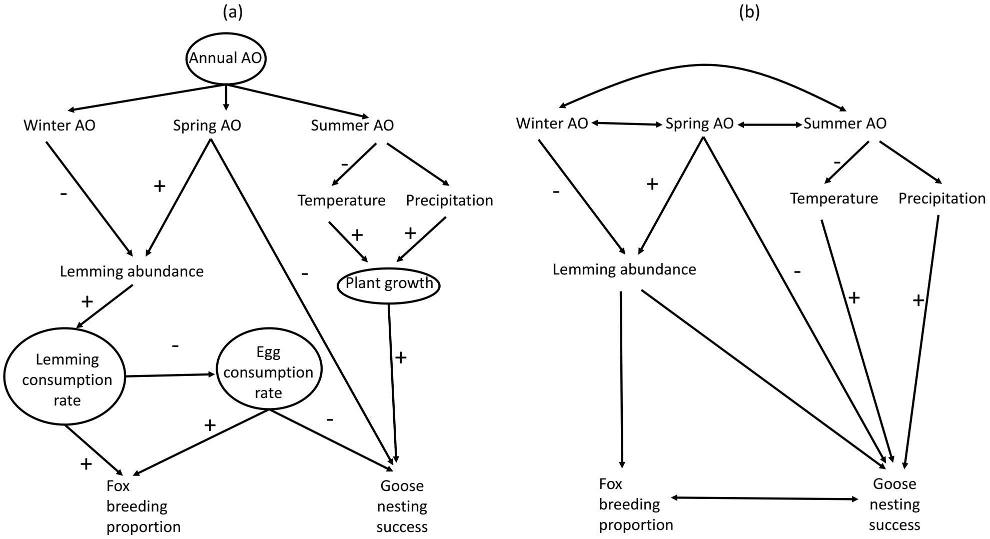

Figure 2. The full causal hypothesis of Juhasz et al. (2020, online supporting information), expressed as a DAG is show in panel (a). This causal hypothesis involves four latent (unobserved) variables enclosed in open circles: *Annual AO, Plant growth, Lemming consumption rate* and *Egg consumption rate.* All remaining variables are observed and measured. The result of marginalizing over the latent variables results in the MAG shown in panel (b).

in the hypothesized causal explanation were latent in that study. The nature of the data (repeated measurements over 21 years and non-normally distributed variables) and the fact that the latent variables did not have observed indicator variables make it impossible to test this hypothesis via classical SEM. The associated MAG of this DAG (Figure 2b) that results from marginalizing over the four latents is empirically testable with 16 m-separation claims in the union basis set. Furthermore, this MAG shows which latent variables can be safely ignored with respect to the causal hypothesis and which cannot. Removing *Annual AO* and *Plant growth* did not change the relationships between any of the other observed variables. If the interest is primarily in the animals, then these two latents could be safely ignored without affecting the hypothesized causal tests involving the variables of principal interest. Removing *Lemming consumption rate* and *Egg consumption rate* would not change the causal hypotheses linking *Lemming abundance* and either *Fox breeding proportion* or *Goose nesting success* but it would change the causal hypotheses linking these last two variables. This is because there are no directed paths from one of these variables to the other; rather, their statistical association is due to two common latent causes.

When substituting a MAG for the original DAG, the causal interpretation of the MAG must be based on the original DAG. An undirected edge in a MAG (X – Y) does not mean an unresolved causal relationship. Rather, it means that there is a spurious association between X and Y due to implicit conditioning from biased sampling on a common causal latent effect of both. The double-headed arrow in Figure 2b (*Fox breeding proportion↔Goose nesting success*) does not mean a feedback relationship between these two variables. Rather, it means that neither is a cause of the other but both variables share common latent causes (lemming and egg consumption rates) whose values have been implicitly marginalized. DAG (III) of Figure 1 provides an even more counter-intuitive example. There is a directed arrow (A→B) but A is not a direct cause of B in the associated DAG even with respect to the other observed variables. Here, A→B in the MAG simply means that A is a causal ancestor of B and that the statistical association between A and B cannot be removed by statistically conditioning on any combination of other variables in the MAG. Referring to the DAG from which it is derived, we see that while A is indeed a cause of B, it is not a direct cause, yet this causal effect cannot be blocked by conditioning any combination of other observed variables. This is called an "inducing path" by Spirtes et al. (1993) and is a good example of why the distinction between a "direct" and an "indirect" cause is always conditional on the other variables included in any causal explanation (Shipley, 2016, pp. 21–22). The same explanation holds for the orientation of the undirected path (X – B) at step 3 in the same MAG: X is a cause (an ancestor) of B and this causal effect cannot be blocked by conditioning any combination of other observed variables. Thus, in a MAG a directed arrow can represent both a direct cause, as well as a causal effect that cannot be blocked by conditioning on any combination of observed variables.

Many researchers, when testing path models using classical SEM without explicit latent variables, sometimes add "free covariances" between variables. In fact, SEM programs like MPLUS or lavaan add these free covariances by default between each pair of exogenous variables although other programs, like EQS, do not. However, a free covariance between two variables is equivalent to adding a common latent cause between them (Pearl, 2009, theorem 5.2.3). In other words, a path model with correlated errors between two observed variables (X, Y) is a MAG, obtained after marginalizing over a latent L in the associated DAG containing the edges (X←L→Y). The union basis set of such an MAG is simply the union basis set of the full DAG after removing the independence relation I(X,Y|L). In fact, the piecewiseSEM package in R (Lefcheck, 2016) performs the d-sep test while allowing the user to incorporate correlated errors by doing this even though no theoretical justification for this was provided. This paper provides this justification for this d-sep test implementation in the piecewiseSEM package. However, since classical SEM simultaneously estimates the path coefficients and the correlated errors, while the current implementation of piecewiseSEM does this separately, the latter approach to estimating path coefficients may produce different parameter estimates.

There are many cases in which classical SEM with latent variables provides a more complete test of the underlying causal hypothesis provided that the additional statistical assumptions are reasonable. Latent variable modeling in classical SEM requires that the system of structural equations be identified. The most common way of insuring identification of a latent is via a "measurement model" in which the latent is a common cause of a minimal number of observed indicator variables; the minimal number depends on several other properties of the system of structural equations (Bollen, 1989; Grace, 2006; Shipley, 2016). When these conditions are met, then the presence of latents implies constraints on the covariance matrix of the observed variables that do not involve conditional independence relations between them. Given identification, as well as the other statistical assumptions of classical SEM mentioned above, then the chi-squared test of classical SEM is a more complete test of the full causal hypothesis because it includes any non-independence constraints that are imposed by the latent variables. An m-sep test would not include such non-independence constraints. In fact, the DAG of a measurement model having a single latent cause of a series of observed indicator variables would result in a MAG, upon marginalizing over the latent, in which each pair of indicator variables is joined by a double-headed arrow (←→) and with no conditional independence relationships between any of them. Classical SEM is preferable to m-sep tests when a DAG involves latent variables that are properly identified by observed indicator variables, and when the other substantive statistical assumptions of classical SEM are reasonable.

Acknowledgments

Clark Glymour, Françoise Cardou, Li Zhang and two anonymous reviewers provided helpful comments on this paper. This research was financially supported by an NSERC Discovery Grant to BS.

References

Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, *2*, 47–53. https://doi.org/10.2307/3002000

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley and Sons.

Buffart, L. M., de Bree, R., Altena, M., van der Werff, S., Drossaert, C. H. C., Speksnijder, C. M., … Stuiver, M. M. (2018). Demographic, clinical, lifestyle-related, and social-cognitive correlates of physical activity in head and neck cancer survivors. *Supportive Care in Cancer*, *26*, 1447–1456. https://doi.org/10.1007/s00520-017-3966-3

Cox, D. R., & Wermuth, N. (1996). *Multivariate dependencies: Models, analysis and interpretation*. Chapman and Hall.

Gordon, A. M., Rissman, J., Kiani, R., & Wagner, A. D. (2014). Cortical reinstatement mediates the relationship between content-specific encoding activity and subsequent recollection decisions. *Cerebral Cortex*, *24*, 3350–3364. https://doi.org/10.1093/cercor/bht194

Grace, J. B. (2006). *Structural equation modeling and natural systems*. Cambridge University Press.

Juhasz, C. C., Shipley, B., Gauthier, G., Berteaux, D., & Lecomte, N. (2020). Direct and indirect effects of regional and local climatic factors on trophic interactions in the Arctic tundra. *The Journal of Animal Ecology*, *89*, 704–715. https://doi.org/10.1111/1365-2656.13104

Lefcheck, J. S. (2016). piecewiseSEM: Piecewise structural equation modelling in r for ecology, evolution, and systematics. *Methods in Ecology and Evolution*, *7*, 573–579. https://doi.org/10.1111/2041-210X.12512

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.

Pearl, J., & Mackenzie, D. (2018). *The book of why. The new science of cause and effect*. Basic Books.

Richardson, T., & Spirtes, P. (2002). Ancestral graph Markov models. *The Annals of Statistics*, *30*, 962–1030. https://doi.org/10.1214/aos/1031689015

Schweiger, S., & Cress, U. (2019). Attitude confidence and source credibility in information foraging with social tags. *PLoS ONE*, *14*. https://doi.org/10.1371/journal.pone.0210423

Shipley, B. (2000). A new inferential test for path models based on directed acyclic graphs. *Structural Equation Modeling*, *7*, 206–218. https://doi.org/10.1207/S15328007SEM0702_4

Shipley, B. (2009). Confimatory path analysis in a gereralized multilevel context. *Ecology*, *90*, 363–368. https://doi.org/10.1890/08-1034.1

Shipley, B. (2016). *Cause and correlation in biology: A user's guide to path analysis, structural equations and causal inference with R* (2nd ed.). Cambridge University Press.

Shipley, B., & Douma, J. C. (2020). Generalized AIC and chi-squared statistics for path models consistent with directed acyclic graphs. *Ecology*, *101*. https://doi.org/10.1002/ecy.2960

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, Prediction, and Search*. Springer-Verlag.

Thoemmes, F., Rosseel, Y., & Textor, J. (2018). Local fit evaluation of structural equation models using graphical criteria. *Psychological Methods*, *23*, 27–41. https://doi.org/10.1037/met0000147

van Kampen, D. (2014). The SSQ model of schizophrenic prodromal unfolding revised: An analysis of its causal chains based on the language of directed graphs. *European Psychiatry: The Journal of the Association of European Psychiatrists*, *29*, 437–448. https://doi.org/10.1016/j.eurpsy.2013.11.001

Verma, T., & Pearl, J. (1990). Causal networks: Semantics and expressiveness. In R. Shachter, T. S. Levitt, & L. N. Kanal (Eds.), *Uncertainty in AI 4* (pp. 69–76). Elsevier Science Publishers.

Warach, B., Josephs, L., & Gorman, B. S. (2018). Pathways to infidelity: The roles of self-serving bias and betrayal trauma. *Journal of Sex and Marital Therapy*, *44*, 497–512. https://doi.org/10.1080/0092623X.2017.1416434