New D-Separation Identification Results for Learning Continuous Latent Variable Models

Ricardo Silva and Richard Scheines
Center for Automated Learning and Discovery
rbas@cs.cmu.edu, scheines@andrew.cmu.edu

July 15, 2005
CMU-CALD-05-105

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

Learning the structure of graphical models is an important task, but one of considerable difficulty when latent variables are involved. Because conditional independences using hidden variables cannot be directly observed, one has to rely on alternative methods to identify the d-separations that define the graphical structure. This paper describes new distribution-free techniques for identifying d-separations in continuous latent variable models when non-linear dependencies are allowed among hidden variables.

Keywords: Causality discovery, graphical models, latent variable models

1 Introduction

Latent variable models are often represented as graphical models such as Bayesian networks. In a broad class of such models, sometimes called the measurement/structural model class (Bollen, 1989), the only constraint is that an observed variable cannot be a parent of a latent variable. This is especially useful in models where observed variables are *indicators* of latent concepts, such as in many models of economics, socials sciences and psychology. Factor analysis and its variations are standard models of such a class.

Learning the graphical structure of such models is of great interest. For causal analysis (Spirtes et al., 2000; Pearl, 2000), which is in fact the main motivation behind several latent variable models, knowing the model structure is essential. For probabilistic modeling (Bishop, 1998), a parsimonious structure that is as simple as possible but not simpler than the truth allows for more statistically efficient estimation of the joint.

A directed acyclic graphs (DAG) G can be defined in terms of conditional independencies among the random variables represented as nodes in G. Those independencies arise from the assumption that the Markov condition holds in such graphs: each node is independent of its non-descendants (and non-parents) conditioned on its parents. Many other conditional independencies are entailed from this local assumption. In special, d-separation is a sound and complete criterion for deriving conditional independencies entailed in a DAG by the Markov condition (Pearl, 2000). Therefore, one can also say that a DAG represents a set of d-separations among its nodes.

The contribution of this paper is theoretical: a set of *testable statistical conditions* that allows us to identify the presence of latent variables and several unobservable conditional independencies in the class of measurement/structural models. Such identification conditions can be used to create tests or search operators for learning the structure of Bayesian networks with latent variables, where non-independence constraints have to be used (Tian and Pearl, 2002).

While we will assume that observed variables are linear functions of their parents with additive noise, we will not assume any particular functional relationship among latents: any arbitrary non-linear function can link a latent to its parents. Indicators that are linear functions of their parents are acceptable in many situations (Bollen, 1989), but models where latents are linearly related are not as widely applicable.

In the next section we present a brief overview of previous work. Section 3 formalizes the problem and Section 4 presents an example on how to use our results. Section 5 provides the main theoretical results and Section 6 provides more details concerning the application of our results on learning the structure of latent graphical models. Section 7 describes some experimental results.

2 Related work

Many latent variable models assume latents are marginally independent as in, e.g., the mixture of factor analyzers of (Ghahramani and Hinton, 1996). For causal modeling this often makes no sense: see all examples given by Bollen (1989), for instance. For probabilistic modeling, this is also an inefficient representation: allowing latents to be dependent will eliminate many edges connecting observed variables and latents. This can be observed by applying "rotation methods" on factor analysis models with Gaussian variables (Bartholomew and Knott, 1999).

Nachman et al. (2004) describe computationally efficient heuristics to create continuous networks with hidden variables for a variety of practical uses, but with no theoretical guarantess about how close the resulting structures might be compared to the unknown true structure that generated the data. Our contributions are on the theoretical aspects and extend the work of Silva et al. (2003), one of the first principled approaches to introduce hidden variables in continuous networks with linear and non-linear relations. However, some extra structural assumptions were adopted in that work. Silva and Scheines (2004) introduced new results while removing such assumptions. However, several results in (Silva and Scheines, 2004) were established only for linear models. This report complements (Silva and Scheines, 2004) by presenting the corresponding results in the non-linear case and simplifies the description of previous results to match the presentation of (Silva et al., 2005). More related work is discussed in the given references.

3 Approach

We assume that the latent variable model to be discovered has a graphical structure and parameterization that obey the following constraints besides the Markov condition (Pearl, 2000; Spirtes et al., 2000):

A1. no observed variable is a parent of a latent variable;

A2. any observed variable is a linear function of its parents with additive noise of finite positive variance;

A3. all latent variables have finite positive variance, and the correlation of any two latents lies strictly in the open interval (-1, 1);

A4. there are no cycles that include an observed variable;

This means that observed variables can have observed parents, and that latents can be (noisy) non-linear functions of their parents, and that cycles are allowed among latents. These are more relaxed assumptions than those adopted in, e.g., factor analysis (Bartholomew and Knott, 1999), a standard tool in latent variable modeling.

In classic results concerning algorithms for learning the structure of directed acyclic graphs without hidden variables (Chickering, 2002; Pearl, 2000; Spirtes et al., 2000), an essential assumption is the *faithfulness* assumption: a conditional independence holds in the joint distribution if and only if it is entailed in the respective graphical model by d-separation. The movitation is that observed conditional independences should be the result of the graphical structure, not of an accidental choice of parameters defining the probability of a node given its parents.

Instead of assuming faithfulness, our results will have a measure-theoretical motivation. All results presented here have the following characteristics:

C1. they hold with probability 1 with respect to the Lebesgue measure over the set of linear coefficients and error variances that partially parameterize the density function of an observed variable given its parents;

C2. they hold for any distribution of the latent variables (that obey the given assumptions);

One can show that the Lebesgue argument is no different from the faithfulness assumption for typical families of graphical models, such as multinomial and Gaussian (Spirtes et al., 2000)[1].

Our goal is not to fully identify a graphical structure. The assumptions are too weak to reallistically accomplish this goal. Instead we will focus on a more restricted task:

- GOAL: *to identify d-separations between a pair of observed variables, or a pair of one observed and one latent variable, conditioned on sets of latent variables. These d-separations should be useful for existing algorithms that learn latent models.*

We do not aim at identifying d-separations between latents: this is a topic for future research, where specific assumptions concerning latent structure have to be adopted according to the problem at hand. This was accomplished for the linear case (Silva et al., 2005).

The strategy to accomplish our goal is to use *constraints in the observed covariance matrix* that will allow us to identify the following features of the unknown latent variable model:

F1. which hidden variables exist;

F2. that observed variable X cannot be an ancestor of observed variable Y;

F3. that observed variable X cannot have a common parent with observed variable Y;

In the next section we describe a way of putting together these pieces of information to learn a partial latent variable model structure, assuming features F1, F2 and F3 can be identified. Section 5 will describe testable methods that can in many cases identify the above features.

4 Application: learning latent model structure

Features F1, F2 and F3 compose all the information used in an algorithm described by Silva et al. (2003) that discovers latent variable structures. However, that algorithm was designed under a particular strong assumption: there is a subgraph G' of the true graph G where each latent has at least three unique indicators (that is, observed children that are not children of any other latent), and any two observed nodes in G' are d-separated given the latents.

[1] That is, in general no result concerning learning graphical models can be theoretically sound for all possible models. For some choice of parameter values (that generate constraints that are not a result of the graphical structure of the true model), several crucial results (Pearl, 2000; Spirtes et al., 2000) fail, and so do our results. Those parameter values, however, form a set of Lebesgue measure zero, which can be interpreted as having zero probability according to an uniform prior. The faithfulness condition is a way of excluding such parameter values by assumption.

We call this assumption the "3-clustering" assumption, because G' defines a clustering over its observed variables: each cluster is a set of observed nodes that share an unique common parent, and each cluster has at least three members.

The work of Silva et al. (2003) is one of the few theoretically sound approaches for learning latent graphs without imposing unrealistic restrictions on how latents are connected to other latents. However, it relies on this strong and generally untestable assumption. Our paper build on this previous result by proving which other guarantees the approach of Silva et al. (2003) can give when the "3-clustering" assumption is dropped:

1. we will show that in general there is no fully automated way of identifying latents individually (feature F1) using covariance information only, but some data-driven methods and generally weak prior knowledge can be combined to solve this issue;
2. we will show extra ways of identifying d-separations that were not discussed by Silva et al. (2003);
3. we will show the existence of empirically testable ways of discovering F3 features that are sound under fully linear models but not sound when non-linear relations among latents are allowed;
4. we will show how to approximate marginal distributions by using sparse latent variable models if this marginal can be approximated well by a mixture of Gaussians;

Our focus on using only the covariance matrix is motivated by a practical issue: since learning latent variable graphs is a difficult statistical problem, using only covariance information is desirable, since estimating second moments is easier than estimating higher order moments of the observed joint. Knowing the limits of what can be done using only covariance information is both of theoretical and practical interest.

5 Main results

Assume for now we know the true population covariance matrix. Without loss of generality, assume also that all variables have zero mean. Let $G(\mathbf{O})$ be the graph of the latent variable model with observed variables $\mathbf{O}$. The following lemma by Silva et al. (2003) illustrates a simple result that is intuitive but does not follow immediately from correlation analysis, since observed nodes can have non-linear dependencies:

Lemma 1 *If for $\{A, B, C\} \subseteq \mathbf{O}$ we have $\rho_{AB} = 0$ or $\rho_{AB.C} = 0$, then A and B cannot share a common latent parent in G.*

where $\rho_{XY.Z}$ is the partial correlation of X and Y given Z. In general, Z can be a set.

Although vanishing partial correlations (i.e., partial correlations constrained to be zero) can sometimes be useful, we are mostly motivated by problems where *all* observed variables have hidden common ancestors. Bartholomew and Knott (1999) describe several of such problems. In this case, vanishing partial correlations are useless. Instead, we will use rank constraints on the covariance matrix of the observed variables.

The following result, also by Silva et al. (2003), allows us to learn that observed variable X cannot be an ancestor of observed variable Y in many situations:

Lemma 2 *For any set $\mathbf{O}' = \{A, B, C, D\} \subseteq \mathbf{O}$, if $\sigma_{AB}\sigma_{CD} = \sigma_{AC}\sigma_{BD} = \sigma_{AD}\sigma_{BC}$ such that for all triplets $\{X, Y, Z\}$, $\{X, Y\} \subset \mathbf{O}', Z \in \mathbf{O}$, we have $\rho_{XY.Z} \neq 0$ and $\rho_{XY} \neq 0$, then no element in $X \in \mathbf{O}'$ is an ancestor of any element in $\mathbf{O}' \backslash X$ in G.*

Notice that this result allows us to identify the non-existence of several ancestral relations even when no conditional independences are observed and latents are non-linearly related. All of the next lemmas and theorems in this paper are new results not previously described by Silva et al. (2003). Detailed proofs are given in the Appendix.

A second way of learning how two observed variables can be d-separated conditioned on a latent is as follows: let $G(\mathbf{O})$ be a latent variable graph and $\{A, B\}$ be two elements of $\mathbf{O}$. Let the predicate $Factor_1(A, B, G)$ be true if and only there exists a set $\{C, D\} \subseteq \mathbf{O}$ such that the conditions of Lemma 2 are satisfied for $\mathbf{O}' = \{A, B, C, D\}$, i.e., $\sigma_{AB}\sigma_{CD} = \sigma_{AC}\sigma_{BD} = \sigma_{AD}\sigma_{BC}$ with the corresponding partial correlation constraints. The second approach for detecting lack of ancestral relations between two observed variables is given by the following lemma:

Lemma 3 *For any set $\mathbf{O}' = \{X_1, X_2, Y_1, Y_2\} \subseteq \mathbf{O}$, if $Factor_1(X_1, X_2, G) = true$, $Factor_1(Y_1, Y_2, G) = true$, $\sigma_{X_1Y_1}\sigma_{X_2Y_2} = \sigma_{X_1Y_2}\sigma_{X_2Y_1}$, and all elements of $\{X_1, X_2, Y_1, Y_2\}$ are correlated, then no element in $\{X_1, X_2\}$ is an ancestor of any element in $\{Y_1, Y_2\}$ in G and vice-versa.*

One can verify that Lemma 2 is a special case of our new lemma.

We define the predicate $Factor_2(A, B, G)$ to be true if and only it is possible to learn that A is not an ancestor of B in the unknown graph G that contains these nodes by using Lemma 3.

We now describe two ways of detecting if two observed variables have no (hidden) common parent in $G(\mathbf{O})$. Let first $\{X_1, X_2, X_3, Y_1, Y_2, Y_3\} \subseteq \mathbf{O}$. We define two identification conditions:

CS1. If $\sigma_{X_1Y_1}\sigma_{X_2X_3} = \sigma_{X_1X_2}\sigma_{X_3Y_1} = \sigma_{X_1X_3}\sigma_{X_2Y_1}, \sigma_{X_1Y_1}\sigma_{Y_2Y_3} = \sigma_{X_1Y_2}\sigma_{Y_1Y_3} = \sigma_{X_1Y_3}\sigma_{Y_1Y_2}$, $\sigma_{X_1X_2}\sigma_{Y_1Y_2} \neq \sigma_{X_1Y_2}\sigma_{X_2Y_1}$ and for all triplets $\{X, Y, Z\}, \{X, Y\} \subset \{X_1, X_2, X_3, Y_1, Y_2, Y_3\}$, $Z \in \mathbf{O}$, we have $\rho_{XY} \neq 0, \rho_{XY.Z} \neq 0$, then X_1 and Y_1 do not have a common parent in G.

CS2. If $Factor_1(X_1, X_2, G)$, $Factor_1(Y_1, Y_2, G)$, X_1 is not an ancestor of X_3, Y_1 is not an ancestor of Y_3, $\sigma_{X_1Y_1}\sigma_{X_2Y_2} = \sigma_{X_1Y_2}\sigma_{X_2Y_1}$, $\sigma_{X_2Y_1}\sigma_{Y_2Y_3} = \sigma_{X_2Y_3}\sigma_{Y_2Y_1}$, $\sigma_{X_1X_2}\sigma_{X_3Y_2} = \sigma_{X_1Y_2}\sigma_{X_3X_2}$, $\sigma_{X_1X_2}\sigma_{Y_1Y_2} \neq \sigma_{X_1Y_2}\sigma_{X_2Y_1}$ and for all triplets $\{X, Y, Z\}, \{X, Y\} \subset \{X_1, X_2, X_3, Y_1, Y_2, Y_3\}, Z \in \mathbf{O}$, we have $\rho_{XY} \neq 0, \rho_{XY.Z} \neq 0$, then X_1 and Y_1 do not have a common parent in G.

"CS" here stands for "constraint set," a set of constraints in the observable joint that are empirically verifiable. In the same way, call CS0 the separation rule of Lemma 1. The following lemmas state the correctness of CS1 and CS2:

Lemma 4 *CS1 is sound.*

Lemma 5 *CS2 is sound.*

It is clear that these identification conditions also hold in fully linear latent variable models, since they are just a special case of the non-linear models here described. One might conjecture that, as far as identifying ancestral relations among observed variables and hidden common parents goes, linear and non-linear latent variable models are identical (since any connection between a latent and an observed variable is always linear in our setup of non-linear models). However, this is not true.

Theorem 1 *Consider the problem of learning if two observed variables do not share a hidden common parent in a latent variable graph. There are identification rules for learning this information that are sound in linear models, but not sound for non-linear latent variable models.*

In other words, one gains more identification power if one is willing to assume full linearity of the latent variable model. We will see more of the implications of assuming linearity later.

Another important building block in our approach is the identification of which latents exist. Define an *immediate latent ancestor* of an observed node O in a latent variable graph G as a latent node L that is a parent of O or the source of a directed path $L \rightarrow V \rightarrow \cdots \rightarrow O$ where V is an observed variable. Notice that this implies that every element in this path, with the exception of L, is an observed node.

Lemma 6 *Let $\mathbf{S} \subseteq \mathbf{O}$ be any set such that, for all $\{A, B, C\} \subseteq \mathbf{S}$, there is a fourth variable $D \in \mathbf{O}$ where i. $\sigma_{AB}\sigma_{CD} = \sigma_{AC}\sigma_{BD} = \sigma_{AD}\sigma_{BD}$ and ii. for every set $\{X, Y\} \subset \{A, B, C, D\}, Z \in \mathbf{O}$ we have $\rho_{XY.Z} \neq 0$ and $\rho_{XY} \neq 0$. Then $\mathbf{S}$ can be partioned into two sets $\mathbf{S_1}, \mathbf{S_2}$ where*

1. *all elements in $\mathbf{S_1}$ share a common immediate latent ancestor, and no two elements in $\mathbf{S_1}$ have any other common immediate latent ancestor;*

2. *no element $S \in \mathbf{S_2}$ has any common immediate latent ancestor with any other element in $\mathbf{S}\backslash S$;*

3. *all elements in $\mathbf{S}$ are d-separated given the latents in G;*

We will see an application of our results in the next section, where they are used to identify interesting *clusters* of indicators, disjoint sets of observed variables that measure disjoint sets of latents.

6 Learning a semiparametric model

Our results can be used to learn graphical and probabilistical features of the true unknown model, as explained in the following subsections.

6.1 Structure learning

Given a set of observed variables $\mathbf{O}$, let $\mathbf{O}' \subseteq \mathbf{O}$, and let $\mathbf{C}$ be a partition of $\mathbf{O}'$ into k non-overlapping sets $\{\mathbf{C_1}, \ldots, \mathbf{C_k}\}$ such that

SC1. for any $\{X_1, X_2, X_3\} \subset \mathbf{C_i}$, there is some $X_4 \in \mathbf{O}'$ such that $\sigma_{X_1X_2}\sigma_{X_3X_4} = \sigma_{X_1X_3}\sigma_{X_2X_4} = \sigma_{X_1X_4}\sigma_{X_2X_3}$, $1 \leq i \leq k$ and X_4 is correlated with all elements in $\{X_1, X_2, X_3\}$;

SC2. for any $X_1 \in \mathbf{C_i}, X_2 \in \mathbf{C_j}$, $i \neq j$, we have that X_1 and X_2 are separated by CS0, CS1 or CS2;

SC3. for any $X_1, X_2 \in \mathbf{C_i}$, $Factor_1(X_1, X_2, G) = true$ or $Factor_2(X_1, X_2, G) = true$;

SC4. for any $\{X_1, X_2\} \subset \mathbf{C_i}$, $X_3 \in \mathbf{C_j}$, $\rho_{X_1X_3} \neq 0$ if and only if $\rho_{X_2X_3} \neq 0$;

Any partition with structural conditions SC1-SC4 has the following properties:

Theorem 2 *If a partition $\mathbf{C} = \{\mathbf{C_1}, \ldots, \mathbf{C_k}\}$ of $\mathbf{O}'$ respects structural conditions SC1-SC4, then the following should hold in the true latent variable graph G that generated the data:*

1. *for all $X \in \mathbf{C_i}, Y \in \mathbf{C_j}, i \neq j$, X and Y have no common parents, and X is d-separated from the latent parents of Y given the latent parents of X;*
2. *for all $X, Y \in \mathbf{O}'$, X is d-separated from Y given the latent parents of X;*
3. *every set $\mathbf{C_i}$ can be partitioned into two groups according to Lemma 6;*

An algorithm for learning such a partition is given by Silva et al. (2003) using statistical tests for deciding if the required constraints in the covariance matrix hold in the population. Notice that algorithm does not make use of CS2 (a less general form of CS1 is used), but it can be naturally added, as it was done in the algorithm for linear models introduced by Silva et al. (2005). Unlike the algorithm by Silva et al. (2003), we allow in principle partitions where some sets $\mathbf{C_i}$ are such that $|\mathbf{C_i}| = 1$ or $|\mathbf{C_i}| = 2$. In those cases, the properties established by Lemma 6 hold vacuously. A greedy Bayesian search algorithm can also be readily constructed by using the given identification rules. A particular algorithm will be a topic of future research.

This algorithm cannot identify how each set $\mathbf{C_i}$ can be further partitioned into two subsets, one where every node has an unique common immediate latent ancestor, and one where each node has no common immediate latent ancestor with any other node. It might be the case that no two nodes in $\mathbf{C_i}$ have a common immediate latent ancestor. It might be the case that all nodes in in $\mathbf{C_i}$ have an unique common immediate latent ancestor. The combination of Lemma 6 and domain knowledge can be useful to find the proper subpartition.

These are weaker results than the ones obtained for linear models, as described by Silva et al. (2005). There, each set $\mathbf{C_i}$ is associated with an unique latent variable L_i from G (as long as $|\mathbf{C_i}| > 2$). Furthermore, conditioned on L_i each node in $\mathbf{C_i}$ is d-separated from all other nodes in $\mathbf{O}'$, as well as from their respective latent parents. There might be no latent node in the non-linear case with these properties.

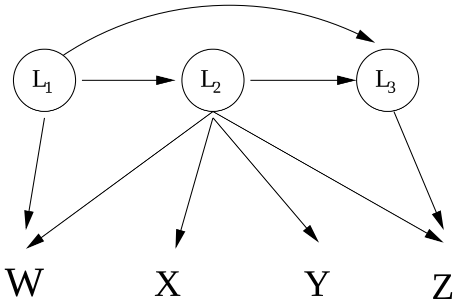

Figure 1: It is possible that $\rho_{L_1 L_3 . L_2} \neq 0$ even though L_2 does not d-separate L_1 and L_3. That happens, for instance, if $L_2 = \lambda_1 L_1 + \epsilon_2$, $L_3 = \lambda_2 L_1^2 + \lambda_3 L_2 + \epsilon_3$, where L_1, ϵ_2 and ϵ_3 are normally distributed with zero mean.

For instance, consider the graph in Figure 1, which depicts a latent variable graph with three latents L_1, L_2 and L_3, and four measured variables, W, X, Y, Z. L_2 does not d-separate L_1 and L_3, but there is no constraint in the assumptions that precludes the partial correlation of L_1 and L_3 given L_2 of being zero. If this is the case, the trivial partition $\mathbf{C} = \{\{W, X, Y, Z\}\}$, with a single element, will satisfy the structural conditions SC1-SC4, and therefore the properties of Theorem 2. However, there is no unique latent variable in this system that d-separates all elements of $\{W, X, Y, Z\}$. This would not be the case in a linear system.

There is an even more fundamental difference between the work presented here and the one developed by Silva et al. (2003). There, the 3-clustering assumption was used, i.e., each latent was assumed to have three observed children that were d-separated by it. In this way, it was possible to use a stronger version of CS1 and Lemma 2 to identify all latents and a bijective mapping between set $\{\mathbf{C_i}\}$ and the set of latents in the true graph[2].

Although one might adopt the 3-clustering assumption in studies where one already has a strong idea of which latents exist, this is in general an untestable assumption. This present work explores what is possible to achieve when minimal assumptions about the graphical structure are adopted, and expands it with extra identification rules. With the stronger assumptions of Silva et al. (2003), all latents could be identified, which highly simplified the problem. This is not the case here.

6.2 Parameter learning

As in the linear case, it is still possible to parameterize a latent variable model using the partition $\mathbf{C} = \{\mathbf{C_1}, \ldots, \mathbf{C_k}\}$ of a subset $\mathbf{O'}$ of the given observed variables such that the first two moments of the distribution of $\mathbf{O'}$ can still be represented. Given a graph G, a *linear parameterization* of G associates a parameter with each edge and two parameters with each node, such that each node V is functionally represented as a linear combination of its

[2] That is, every latent L_i in the true graph would be a hidden common cause d-separating elements in some set $\mathbf{C_i}$, and all observed nodes in some set $\mathbf{C_j}$ would be d-separated by a common hidden parent L_j in the true graph, where $L_i = L_j$ if and only if $\mathbf{C_i} = \mathbf{C_j}$.

parents plus an additive error: $V = \mu_V + \Sigma_i \lambda_i Pa_{V_i} + \epsilon_V$, where $\{Pa_{V_i}\}$ is the set of parents of V in G, and ϵ_V is a random variable with zero mean and variance ζ_V (μ_V and ζ_V are the two extra parameters by node). Notice that this parameterization might not be enough to represent all moments of a given family of probability distributions.

A *linear latent variable model* is a latent variable graph with a particular instance of a linear parameterization. In general, building a model that uses a particular set of constraints, such as the rank constraints of Section 5, might impose other constraints over the joint distribution that do not necessarily hold in the population. It is not obvious if a linear model obtained from the algorithm discussed in the previous section can be used to represent the population covariance matrix without any bias. We show this is true.

Theorem 3 *Given a partition* $\mathbf{C}$ *of a subset* $\mathbf{O'}$ *of the observed variables of a latent variable graph* G *such that* $\mathbf{C}$ *satisfies structural constraints SC1-SC4, there is a linear latent variable model for the first two moments of* $\mathbf{O'}$.

Consider the graph G_{linear} constructed by the following algorithm:

1. initialize G_{linear} with a node for each element in $\mathbf{O'}$;

2. for each $\mathbf{C_i} \in \mathbf{C}$, add a latent L_i to G, and for each $V \in \mathbf{C_i}$, add an edge $L_i \rightarrow V$

3. fully connect the latents in G_{linear} to form an arbitrary directed acyclic graph;

The constructive proof of Theorem 3 shows that G_{linear} can be used to parameterize a model of the first two moments of $\mathbf{O'}$. This has an important heuristic implication: if the joint distribution of the latents and observed variables can be reasonably approximated by a mixture of Gaussians, where each component has the same graphical structure, one can fit a mixture of G_{linear} graphical models. This can be motivated by assuming each mixture component represents a different subpopulation probabilitistic model where the same causal structures hold, and the distributions are close to normal (e.g., a drug might have different quantitative effects on different genders but with the same qualitative causal structure). Each model will provide unbiased estimates of the mean and covariance of the observed variables for a particular component of the mixture: since each component has the same graphical structure, the same required constraints in the component covariance matrix hold, and therefore the same parametric formulation can be used.

Notice this is less stringent than assuming that the causal model is fully linear. Assuming the distribution is fully linear can theoretically result in a wrong structure that might not be approximated well (e.g., if one applies unsound identification rules, as suggested by Theorem 1). Here, at least in principle the structure can be correctly induced. The joint distribution is approximated, and the quality of approximation will be dependendent on the domain.

6.3 Final remarks

Finally, it has to be stressed that there is no guarantee of how large the subset $\mathbf{O'}$ will be. It can be an empty set, for instance, if all observed variables are children of several latents.

An algorithm such as the one described by Silva et al. (2003) is still able to asymptotically find the largest submodel where each latent d-separates three or more of its children.

In principle, much of the limitations here described can be treated if one explores constraints that uses information besides the second moments of the observed variables. Still, it is of considerable interest to know what can be done with covariance information only, since using higher order moments highly increases the chance of commiting statistical mistakes. This is especially difficult concerning learning the structure of latent variable models.

7 Experiments

The main contribution of this paper is theoretical, but there are several aspects of our approach that can be evaluated empirically. For instance, if the correct qualitative causal relations are learned from data. This is usually accomplished through simulations, and an exhaustive study for linear models was done by Silva et al. (2005). For the non-linear case, some studies are shown in Silva et al. (2003).

In this paper, we will concentrate on evaluating our procedure as a way of finding good fitting submodels. We run the algorithm described by Silva et al. (2003) over some datasets from the UCI Machine Learning Repository to obtain a graphical structure analogous to G_{linear} described in the previous section. Following Silva et al. (2005), we call this algorithm a special version of BuildPureClusters (BPC). We then fit the data to such a structure by using a mixture of Gaussian latent DAGs with a standard EM algorithm. Each component has a full parameterization: different linear coefficients and error variances for each variable on each mixture component. The number of mixture components is chosen by fitting the model with 1 to up to 7 components and choosing the one that maximizes the BIC score (see, e.g., Chickering, 2002).

We compare this model against the mixture of factor analyzers, MofFA (Ghahramani and Hinton, 1996). In this case, we want to compare what can be gained by fitting a model where latents are allowed to be dependent, even when we restrict the observed variables to be children of a single latent. Therefore, we fit mixtures of factor analyzers using the same number of latents we find with our algorithm. The number of mixture components is chosen independently, using the same BIC-based procedure. Since BPC can return only a model for a subset of the given observed variables, we run MofFA for the same subsets given by our algorithm.

In practice, our approach can be used in two ways. First, as a way of decomposing the full joint of a set $\mathbf{O}$ of observed variables by splitting it into two sets: one set where variables $\mathbf{X}$ can be modeled as a mixture of G_{linear} models, and another set of variables $\mathbf{Y} = \mathbf{O} \backslash \mathbf{X}$ whose conditional probability $f(\mathbf{Y}|\mathbf{X})$ can be modeled by some other representation of choice. Alternatively, if the observed variables are redundant (i.e., many variables are intended to measure the same latent concept), this procedure can be seen as a way of choosing a subset whose marginal is relatively easy to model with simple causal graphical structures. This is sometimes called "purification" and has several applications in sciences where designing proper indicators is of special concern, such as econometrics and psychology (Spirtes et al., 2000).

Table 1: The difference in average test log-likelihood of BPC and MofFA with respect to a multivariate mixture of Gaussians. Positive values indicate that a method gives a better fit that the mixture of Gaussians. The statistics are the average of the results over a 10-fold cross-validation. A standard deviation is provided. The average number of variables used by our algorithm is also reported.

Dataset	BPC	MofFA	**% variables**
iono	1.56 ± 1.10	-3.03 ± 2.55	0.37 ± 0.06
spectf	-0.33 ± 0.73	-0.75 ±0.88	0.34 ± 0.07
water	-0.01 ± 0.74	-0.90 ± 0.79	0.36 ± 0.04
wdbc	-0.88 ± 1.40	-1.96 ± 2.11	0.24 ± 0.13

As a baseline, we use a standard mixture of Gaussians (MofG), where an unconstrained multivariate Gaussian is used on each mixture component. Again, the number of mixture components is chosen independently by maximizing BIC. Since the number of variables used in our experiments are relatively small, we do not expect to perform significantly better than MofG in the task of density estimation, but a similar performance is an indication that our highly constrained models provide a good fit, and therefore our observed rank constraints can be reasonably expected to hold in the population.

We ran a 10-fold cross-validation experiment for each one of the following four UCI datasets: iono, specft, water and wdbc, all of which are measured over continuous or ordinal variables. We tried also the small dataset wine (13 variables), but we could not find any structure using our method. The chosen datasets have from 30 to 40 variables. The results given in Table 1 show the average log-likelihood per data point on the respective test sets, also averaged over the 10 splits. These results are subtracted from the baseline established by MofG. We also show the average percentage of variables that were selected by our algorithm. The outcome is that we can represent the joint of a significant portion of the observed variables as a simple latent variable model where observed variables have a single parent. Such models do not significantly lose information compared to the full mixture of Gaussians. In one case (iono) we were able to significantly improve over the mixture of factor analyzers when using the same number of latent variables.

We conjecture these results can be greatly improved by using Bayesian search algorithms (BPC is a very simple algorithm that tests hypothesis of rank constraints). We intend also to expand our method to allow the insertion of more observed variables, and not only those that have a single parent in a linearized graph.

8 Conclusion

We presented empirically testable conditions that allows one to learn structural features of latent variable models where latents are non-linearly related. These results can be used in an algorithm for learning the graphical structure of a subset of the observed variables without making any assumptions about the true graphical structure, besides the fairly general assumption by which observed variables cannot be parents of latent variables. We intend to extend this work in the future by exploring kernel methods to learn probabilistic models (Bach and Jordan, 2002) based on the discovered structures, to evaluate it as a technique to discover instrumental variables in non-linear regression problems with measurement error (Carroll et al., 1995) and, finally, as a fundamental step on discovering the causal structure among latent variables when non-linear relations are allowed.

References

F. Bach and M. Jordan. Learning graphical models with Mercer kernels. *Neural Information Processing Systems*, 2002.

D. Bartholomew and M. Knott. *Latent Variable Models and Factor Analysis.* Arnold Publishers, 1999.

C. Bishop. Latent variable models. *Learning in Graphical Models*, 1998.

K. Bollen. *Structural Equation Models with Latent Variables.* John Wiley & Sons, 1989.

R. Carroll, D. Ruppert, and L. Stefanski. *Measurement Error in Nonlinear Models.* Chapman & Hall, 1995.

D. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

Z. Ghahramani and G. Hinton. The EM algorithm for the mixture of factor analyzers. *Technical Report CRG-TR-96-1. Department of Computer Science, University of Toronto.*, 1996.

N. Nachman, G. Elidan, and N. Friedman. The "ideal parent" structure learning for continuous variable networks. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, 2004.

J. Pearl. *Causality: Models, Reasoning and Inference.* Cambridge University Press, 2000.

R. Silva and R. Scheines. Generalized measurement models. *Technical Report CMU-CALD-04-101, Carnegie Mellon University*, 2004.

R. Silva, R. Scheines, C. Glymour, and P. Spirtes. Learning measurement models for unobserved variables. *Proceedings of 19th Conference on Uncertainty in Artificial Intelligence*, pages 543–550, 2003.

R. Silva, R. Scheines, C. Glymour, and P. Spirtes. Learning the structure of linear latent variable models. *Submitted*, 2005.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search.* Cambridge University Press, 2000.

J. Tian and J. Pearl. On the testable implications of causal models with hidden variables. *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, 2002.

A Appendix

All of the following proofs hold with probability 1 with respect to the Lebesgue measure taken over the set of linear coefficients and error variances that partially parameterize the density function of an observed variable given its parents. The main idea used across most proofs is that some covariance constraints boil down to polynomial identities with probability 1. These identities will imply other identities that in many cases will be used to prove results by contradiction. A few of these proofs have appeared before in Silva and Scheines (2004).

The term "immediate latent ancestor," used in several points of this document, is defined in the paper. The symbol $\rho_{XY.Z}$ is the partial correlation of X and Y given Z.

In all of the following proofs, G is a latent variable graph with a set $\mathbf{O}$ of observable variables. In some of these proofs, we use the term "edge label" as a synonym of the coefficient associated with an edge that is into an observed node (e.g., as in linear Gaussian networks). Without loss of generality, we will also assume that all variables have zero mean, unless specified otherwise. The symbol $\{X_t\}$ will stand for a finitely indexed set of variables.

The following lemma will be useful to prove Lemma 3:

Lemma 7 *For any set* $\{A, B, C, D\} = \mathbf{O}' \subseteq \mathbf{O}$*, if* $\sigma_{AB}\sigma_{CD} = \sigma_{AC}\sigma_{BD} = \sigma_{AD}\sigma_{BD}$ *such that for every set* $\{X, Y\} \subset \mathbf{O}', Z \in \mathbf{O}$ *we have* $\rho_{XY.Z} \neq 0$ *and* $\rho_{XY} \neq 0$*, then no pair of elements in* $\mathbf{O}'$ *has an observed common ancestor.*

Proof: Assume for the sake of contradiction that some pair in $\mathbf{O}'$ has an observed common ancestor. Let K be a common ancestor of some pair of elements in $\mathbf{O}'$ such that no descendant of K is also a common ancestor of some pair in $\mathbf{O}'$.

Without loss of generality, assume K is a common ancestor of A and B. Let α be the concatenation of edge labels in some directed path from K to A, and β the concatenation of edge labels in some directed path from K to B. That is,

$$\begin{array}{rcl} A & = & \alpha K + R_A \\ B & = & \beta K + R_B \end{array}$$

where R_X is the remainder of the polynomial expression that describes node X as a function of its immediate latent ancestors and K.

By the given constraint $\sigma_{AB}\sigma_{CD} = \sigma_{AC}\sigma_{BD}$, it follows $\alpha\beta(\sigma_K^2\sigma_{CD} - \sigma_{CK}\sigma_{DK}) + f(G) = 0$, where

$$f(G) = (\alpha\sigma_{KR_B} + \beta\sigma_{KR_A} + \sigma_{R_AR_B})\sigma_{CD} - \sigma_{CR_A} - \sigma_{DR_B}$$

However, no term in $f(G)$ can contain the symbol $\alpha\beta$: by Lemma 2 no element X in $\mathbf{O}'$ can be an ancestor of any element in $\mathbf{O}'\backslash X$; also, by construction no descendant of K (with the possible exception of K) can be an ancestor of C or D and therefore no sequence α or β

can be generated from the polynomial f that is a function of $\sigma_{KR_B}, \sigma_{KR_A}, \sigma_{R_AR_B}, \sigma_{CD}, \sigma_{CR_A}$ or σ_{DR_B}.

It follows that with probability 1 we have $\alpha\beta(\sigma_K^2\sigma_{CD} - \sigma_{CK}\sigma_{DK}) = 0$, and since $\alpha\beta \neq 0$ by assumption, this implies $\sigma_K^2\sigma_{CD} - \sigma_{CK}\sigma_{DK} = 0 \Rightarrow \rho_{CD.K} = 0$. Contradiction. □

Lemma 3 *For any set* $\mathbf{O}' = \{X_1, X_2, Y_1, Y_2\} \subseteq \mathbf{O}$, *if* $Factor_1(X_1, X_2, G) = true$, $Factor_1(Y_1, Y_2, G) = true$, $\sigma_{X_1Y_1}\sigma_{X_2Y_2} = \sigma_{X_1Y_2}\sigma_{X_2Y_1}$, *and all elements of* $\{X_1, X_2, Y_1, Y_2\}$ *are correlated, then no element in* $\{X_1, X_2\}$ *is an ancestor of any element in* $\{Y_1, Y_2\}$ *in* G *and vice-versa.*

Proof: Assume for the sake of contradiction that X_1 is an ancestor of Y_1. Let P be an arbitrary directed path from X_1 to Y_1 of K edges such that the edge coefficients on this path are $\alpha_1 \ldots \alpha_K$. One can write the covariance of X_1 and Y_1 as $\sigma_{X_1Y_1} = c\alpha_1\sigma_{X_1}^2 + F(G)$, where $F(G)$ is a polynomial (in terms of edge coefficients and error variances) that does not contain any term that includes the symbol α_1, and $c = \alpha_2 \ldots \alpha_K$. Also, the polynomial corresponding to $\sigma_{X_1}^2$ cannot contain any term that includes the symbol α_1.

Also analogously, $\sigma_{X_2Y_1}$ can be written as $c\alpha_1\sigma_{X_1X_2} + F'(G)$, where $F'(G)$ does not contain α_1, since X_1 cannot be an ancestor of X_2 by the given hypothesis and Lemma 2.

By Lemma 2 and the given conditions, Y_2 cannot be an ancestor of Y_1 and therefore, not an ancestor of X_1. X_1 cannot be an ancestor of Y_2, by Lemma 7 applied to pair $\{Y_1, Y_2\}$. This implies that $\sigma_{X_1Y_2}$ cannot contain any term that includes α_1. By the same reason, the polynomial corresponding to $\sigma_{X_2Y_2}$ cannot contain any term that includes α_1.

This means that the constraint $\sigma_{X_1Y_1}\sigma_{X_2Y_2} = \sigma_{X_1Y_2}\sigma_{X_2Y_1}$ corresponds to the polynomial identity $\alpha_1(\sigma_{X_1}^2\sigma_{X_2Y_2} - \sigma_{X_1Y_2}\sigma_{X_1X_2}) + F''(G) = 0$, where the polynomial $F''(G)$ does not contain any term that includes α_1, and neither does any term in the factor $(\sigma_{X_1}^2\sigma_{X_2Y_2} - \sigma_{X_1Y_2}\sigma_{X_1X_2})$. This will imply with probability 1 that $\sigma_{X_1}^2\sigma_{X_2Y_2} - \sigma_{X_1Y_2}\sigma_{X_1X_2} = 0$ (which is the same of saying that the partial correlation of X_2 and Y_2 given X_1 is zero).

The expression $\sigma_{X_1}^2\sigma_{X_2Y_2}$ contains a term that include ζ_{X_1}, the error variance for X_1, while $\sigma_{X_1Y_2}\sigma_{X_1X_2}$ cannot contain such a term, since X_1 is not an ancestor of either X_2 or Y_2. That will then imply the term $\zeta_{X_1}\sigma_{X_2Y_2}$ should vanish, which is a contradiction since $\zeta_{X_1} \neq 0$ by assumption and $\sigma_{X_2Y_2} \neq 0$ by hypothesis. □

Lemma 4 *CS1 is sound.*

Proof: Analogous to a result given by Silva et al. (2003). □

Lemma 5 *CS2 is sound.*

Proof: Suppose X_1 and Y_1 have a common parent L in G. Let $X_1 = aL + \sum_p a_pA_p$ and $Y_1 = bL + \sum_p b_iB_i$. To simplify the presentation, we will represent $\sum_p a_pA_p$ by random variable P_x and $\sum_p b_iB_i$ by P_y, such that $X_1 = aL + P_x$ and $Y_1 = bL + P_y$. We will assume that $E[P_xP]$ and $E[P_yP]$ are not zero, for $P \in \{X_1, X_2, Y_1, Y_2\}$ to shorten the proof. The case where these expectations are zero can be derived in an analogous (and simpler) proof.

With probability 1 with respect to a Lebesgue measure over the linear coefficients parameterizing the graph, the constraint $\sigma_{X_1Y_1}\sigma_{X_2Y_2} - \sigma_{X_1Y_2}\sigma_{X_2Y_1} = 0$ corresponds to a polynomial identity where some terms contain the product ab, some contain only a, some contain only b, and some contain none of such symbols. Since this is a polynomial identity, all terms containing ab should sum to zero. The same holds for terms containing only a, only b and not containing a or b. This constraint can be rewritten as

$$\begin{array}{lc} ab(E[L^2]\sigma_{X_2Y_2} - E[LY_2]E[LX_2]) & + \\ a(E[LP_y]\sigma_{X_2Y_2} - E[LY_2]E[X_2P_y]) & + \\ b(E[LP_x]\sigma_{X_2Y_2} - E[Y_2P_x]E[LX_2]) & + \\ (E[P_xP_y]\sigma_{X_2Y_2} - E[P_xY_2]E[P_yX_2]) & \end{array}$$

From Lemmas 2 and 3 and the given hypothesis, X_1 cannot be an ancestor of any element of $\{X_2, Y_1, Y_2\}$ and Y_1 cannot be an ancestor of any element in $\{X_1, X_2, Y_2\}$. Therefore, the symbols a and b cannot appear inside any of the polynomial expressions obtained when terms such as $\sigma_{X_2Y_2}$ or $E[Y_2P_x]$ are expressed as functions of the latent covariance matrix and the linear coefficients and error variances of the measurement model. All symbols a and b of $\sigma_{X_1Y_1}\sigma_{X_2Y_2} - \sigma_{X_1Y_2}\sigma_{X_2Y_1}$ were therefore factorized as above. Therefore, with probability 1 we have:

$$E[L^2]\sigma_{X_2Y_2} = E[LX_2]E[LY_2] \tag{1}$$

$$E[LP_y]\sigma_{X_2Y_2} = E[LY_2]E[X_2P_y] \tag{2}$$

$$E[LP_x]\sigma_{X_2Y_2} = E[Y_2P_x]E[LX_2] \tag{3}$$

$$E[P_xP_y]\sigma_{X_2Y_2} = E[Y_2P_x]E[X_2P_Y] \tag{4}$$

Analogously, the constraint $\sigma_{X_2Y_1}\sigma_{Y_2Y_3} - \sigma_{X_2Y_3}\sigma_{Y_2Y_1} = 0$ will force other identities. Since Y_1 is also not an ancestor of Y_3, we can split the polynomial expression derived from $\sigma_{X_2Y_1}\sigma_{Y_2Y_3} - \sigma_{X_2Y_3}\sigma_{Y_2Y_1} = 0$ into two parts

$$\begin{array}{c} b\{E[LX_2]\sigma_{Y_2Y_3} - E[LY_2]\sigma_{X_2Y_3}\}+ \\ \{E[X_2P_Y]\sigma_{Y_2Y_3} - E[Y_2P_Y]\sigma_{X_2Y_3}\} = 0 \end{array}$$

where the second component, $E[X_2P_Y]\sigma_{Y_2Y_3} - E[Y_2P_Y]\sigma_{X_2Y_3}$, cannot contain any term that includes the symbol b, and neither can the second factor of the first component, $E[LX_2]\sigma_{Y_2Y_3} - E[LY_2]\sigma_{X_2Y_3}$. With probability 1, it follows that:

$$\begin{array}{lcl} E[LX_2]\sigma_{Y_2Y_3} & = & E[LY_2]\sigma_{X_2Y_3} \\ E[X_2P_Y]\sigma_{Y_2Y_3} & = & E[Y_2P_Y]\sigma_{X_2Y_3} \end{array}$$

Since we have that $\sigma_{Y_2Y_3} \neq 0$ and $\sigma_{X_2Y_3} \neq 0$, from the two equations above, we get:

$$E[LX_2]E[Y_2P_Y] = E[LY_2]E[X_2P_Y] \tag{5}$$

From the constraint $\sigma_{X_1X_2}\sigma_{X_3Y_2} = \sigma_{X_1Y_2}\sigma_{X_3X_2}$ and a similar reasoning, we get

$$E[LX_2]E[Y_2P_X] = E[LY_2]E[X_2P_X] \tag{6}$$

from which follows

$$E[X_2P_X]E[Y_2P_Y] = E[X_2P_Y]E[Y_2P_X] \tag{7}$$

Combining (2) and (5), we have

$$aE[LP_y]\sigma_{X_2Y_2} = aE[LX_2]E[Y_2P_Y] \tag{8}$$

Combining (3) and (6), we have

$$bE[LP_x]\sigma_{X_2Y_2} = bE[X_2P_X]E[LY_2] \tag{9}$$

Combining (4) and (7), we have

$$E[P_xP_y]\sigma_{X_2Y_2} = E[X_2P_X]E[Y_2P_Y] \tag{10}$$

From (1), (8), (9) and (10) and the given constraints:

$$\begin{aligned}\sigma_{X_1X_2}\sigma_{Y_1Y_2} &= abE[LX_2]E[LY_2] + aE[LX_2]E[Y_2P_x] + bE[X_2P_x]E[LY_2] + E[X_2P_X]E[Y_2P_Y] \\ &= abE[L^2]\sigma_{X_2Y_2} + E[LP_y]\sigma_{X_2Y_2} + E[LP_y]\sigma_{X_2Y_2} + E[P_xP_y]\sigma_{X_2Y_2} = \sigma_{X_1Y_1}\sigma_{X_2Y_2} = \sigma_{X_1Y_2}\sigma_{X_2Y_1}\end{aligned}$$

Contradiction. □

Theorem 1 *Consider the problem of learning if two observed variables do not share a hidden common parent in a latent variable graph. There are identification rules for learning this information that are sound in linear models, but not sound for non-linear latent variable models.*

Proof: Consider first the following test: let $G(\mathbf{O})$ be a linear latent variable model. Assume $\{X_1, X_2, X_3, Y_1, Y_2, Y_3\} \subseteq \mathbf{O}$ and $\sigma_{X_1Y_1}\sigma_{Y_2Y_3} = \sigma_{X_1Y_2}\sigma_{Y_1Y_3} = \sigma_{X_1Y_3}\sigma_{Y_1Y_2}$, $\sigma_{X_1Y_2}\sigma_{X_2X_3} = \sigma_{X_1X_2}\sigma_{Y_2X_3} = \sigma_{X_1X_3}\sigma_{X_2Y_2}$, $\sigma_{X_1Y_3}\sigma_{X_2X_3} = \sigma_{X_1X_2}\sigma_{Y_3X_3} = \sigma_{X_1X_3}\sigma_{X_2Y_3}$, $\sigma_{X_1X_2}\sigma_{Y_2Y_3} \neq \sigma_{X_1Y_2}\sigma_{X_2Y_3}$ and that for all triplets $\{A, B, C\}, \{A, B\} \subset \{X_1, X_2, X_3, Y_1, Y_2, Y_3\}, C \in \mathbf{O}$, we have $\rho_{AB} \neq 0, \rho_{AB.C} \neq 0$. Then X_1 and Y_1 do not have a common parent in G.

Call this test CS3. Test CS3 is sound for linear models: if its conditions are true, then X_1 and Y_1 do not have a common parent in G. The proof of this result is given by Silva et al. (2005). However, this is not a sound rule for the non-linear case. To show this, it is enough to come up with a latent variable model where X_1 and Y_1 have a common parent, and a latent covariance matrix such that, for any choice of linear coefficients and error variances, this test applies. Notice that the definition of a sound identification rule in non-linear models allows us to choose specific latent covariance matrices but the constraints should hold for any choice of linear coefficients and error variances (or, more precisely, with probability 1 with respect to the Lebesgue measure).

Consider the graph G with five latent variables $L_i, 1 \leq i \leq 5$, where L_1 has X_1 and Y_1 as its only children, X_2 is the only child of L_2, X_3 is the only child of L_3, Y_2 is the only child of L_4 and Y_3 is the only child of L_5. Also, $\{X_1, X_2, X_3, Y_1, Y_2, Y_3\}$, as defined in CS3, are the only observed variables, and each observed variable has only one parent besides its error term. Error variables are independent.

The following simple randomized algorithm will choose a covariance matrix Σ_L for $\{L_1, L_2, L_3, L_4, L_5\}$ that entails CS3. The symbol σ_{ij} will denote the covariance of L_i and L_j.

1. Choose positive random values for all $\sigma_{ii}, 1 \leq i \leq 5$
2. Choose random values for σ_{12} and σ_{13}
3. $\sigma_{23} \leftarrow \sigma_{12}\sigma_{13}/\sigma_{11}$
4. Choose random values for σ_{45}, σ_{25} and σ_{24}
5. $\sigma_{14} \leftarrow \sigma_{12}\sigma_{45}/\sigma_{25}$
6. $\sigma_{15} \leftarrow \sigma_{12}\sigma_{45}/\sigma_{24}$
7. $\sigma_{35} \leftarrow \sigma_{13}\sigma_{45}/\sigma_{14}$
8. $\sigma_{34} \leftarrow \sigma_{12}\sigma_{45}/\sigma_{15}$
9. Repeat from the beginning if Σ_L is not positive definite or if $\sigma_{14}\sigma_{23} = \sigma_{12}\sigma_{34}$

Notice that the intuition behind this example is to set the covariance matrix of the latent variables to have some vanishing partial correlations, even though one does not necessarily have any conditional independence. For linear models, both conditions are identical, and therefore this identification rule holds in such a case. □.

Lemma 8 *For any set $\{A, B, C, D\} = \mathbf{O}' \subseteq \mathbf{O}$, if $\sigma_{AB}\sigma_{CD} = \sigma_{AC}\sigma_{BD} = \sigma_{AD}\sigma_{BD}$ such that for every set $\{X, Y\} \subset \mathbf{O}', Z \in \mathbf{O}$ we have $\rho_{XY.Z} \neq 0$ and $\rho_{XY} \neq 0$, then A and B do not have more than one common immediate latent ancestor in G.*

Proof: Assume for the sake of contradiction that L_1 and L_2 are two common immediate latent ancestors of A and B in G. Let the structural equations for A, B, C and D be:

$$\begin{array}{rcl} A & = & \alpha_1 L_1 + \alpha_2 L_2 + R_A \\ B & = & \beta_1 L_1 + \beta_2 L_2 + R_B \\ C & = & \sum_j c_j C_j \\ D & = & \sum_k d_k D_k \end{array}$$

where α_1 is a sequence of labels of edges corresponding to some directed path connecting L_1 and A. Symbols α_2, β_1, β_2 are defined analogously. R_X is the remainder of the polynomial expression that describes node X as a function of its parents and the immediate latent ancestors L_1 and L_2.

Since the constraint $\sigma_{AB}\sigma_{CD} = \sigma_{AC}\sigma_{BD}$ is observed, we have $\sigma_{AB}\sigma_{CD} - \sigma_{AC}\sigma_{BD} = 0 \Rightarrow$ $(\alpha_1\beta_1\sigma^2_{L_1} + \alpha_1\beta_2\sigma_{L_1L_2} + \alpha_2\beta_1\sigma_{L_1L_2} + \alpha_2\beta_2\sigma^2_{L_2} + \alpha_1\sigma_{L_1R_B} + \alpha_2\sigma_{L_2R_B} + \beta_1\sigma_{L_1R_A} + \beta_2\sigma_{L_2R_A} + \sigma_{R_AR_B})\sigma_{CD} - (\alpha_1\sum_j c_j\sigma_{C_jL_1} + \alpha_2\sum_j c_j\sigma_{C_jL_2} + \sum_j c_j\sigma_{C_jR_A})$
$(\beta_1\sum_k d_k\sigma_{D_kL_1} + \beta_2\sum_k d_k\sigma_{D_kL_2} + \sum_k d_k\sigma_{D_kR_B})) = 0 \Rightarrow \alpha_1\beta_1(\sigma^2_{L_1}\sigma_{CD} - (\sum_j c_j\sigma_{C_jL_1})(\sum_k d_k\sigma_{D_kL_1})) +$

$f(G) = 0$, where

$$\begin{array}{l} f(G) = \\ (\alpha_1\beta_2\sigma_{L_1L_2} + \alpha_2\beta_1\sigma_{L_1L_2} + \alpha_2\beta_2\sigma^2_{L_2} + \alpha_1\sigma_{L_1R_B} + \\ \alpha_2\sigma_{L_2R_B} + \beta_1\sigma_{L_1R_A} + \beta_2\sigma_{L_2R_A} + \sigma_{R_AR_B})\sigma_{CD} - \\ \alpha_1\sum_j c_j\sigma_{C_jL_1}(\beta_2\sum_k d_k\sigma_{D_kL_2} + \sum_k d_k\sigma_{D_kR_B})) - \\ \alpha_2\sum_j c_j\sigma_{C_jL_2}(\beta_1\sum_k d_k\sigma_{D_kL_1} + \beta_2\sum_k d_k\sigma_{D_kL_2} + \\ \sum_k d_k\sigma_{D_kR_B})) - \sum_j c_j\sigma_{C_jR_A}(\beta_1\sum_k d_k\sigma_{D_kL_1} + \\ \beta_2\sum_k d_k\sigma_{D_kL_2} + \sum_k d_k\sigma_{D_kR_B})) \end{array}$$

No element in $\mathbf{O}'$ is an ancestor of any other element in this set (Lemma 2) and no observed node in any directed path from $L_i \in \{L_1, L_2\}$ to $X \in \{A, B\}$ can be an ancestor of any node in $\mathbf{O}'\backslash X$ (Lemma 7). That is, when fully expanding $f(G)$ as a function of the linear parameters of G, the product $\alpha_1\beta_1$ cannot possibly appear.

Therefore, since with probability 1 the polynomial constraint is identically zero and nothing in $f(G)$ can cancel the term $\alpha_1\beta_1$, we have:

$$\sigma^2_{L_1}\sigma_{CD} = \sum_j c_j\sigma_{C_jL_1}\sum_k d_k\sigma_{D_kL_1} \tag{11}$$

Using a similar argument for the coefficients of $\alpha_1\beta_2$, $\alpha_2\beta_1$ and $\alpha_2\beta_2$, we get:

$$\sigma_{L_1L_2}\sigma_{CD} = \sum_j c_j\sigma_{C_jL_1}\sum_k d_k\sigma_{D_kL_2} \tag{12}$$

$$\sigma_{L_1L_2}\sigma_{CD} = \sum_j c_j\sigma_{C_jL_2}\sum_k d_k\sigma_{D_kL_1} \tag{13}$$

$$\sigma^2_{L_2}\sigma_{CD} = \sum_j c_j\sigma_{C_jL_2}\sum_k d_k\sigma_{D_kL_2} \tag{14}$$

From (11),(12), (13), (14), it follows: $\sigma_{AC}\sigma_{AD} =$

$$\begin{array}{rl} = & [\alpha_1\sum_j c_j\sigma_{C_jL_1} + \alpha_2\sum_j c_j\sigma_{C_jL_2}]\times \\ & [\alpha_1\sum_k d_k\sigma_{D_kL_1} + \alpha_2\sum_k d_k\sigma_{D_kL_2}] \\ = & \alpha_1^2\sum_j c_g\sigma_{C_jL_1}\sum_k d_k\sigma_{D_kL_1} + \\ & \alpha_1\alpha_2\sum_j c_j\sigma_{C_jL_1}\sum_k d_k\sigma_{D_kL_2} + \\ & \alpha_1\alpha_2\sum_j c_j\sigma_{C_jL_2}\sum_k d_k\sigma_{D_kL_1} + \\ & \alpha_2^2\sum_j c_j\sigma_{C_jL_2}\sum_k d_k\sigma_{D_kL_2} \\ = & [\alpha_1^2\sigma^2_{L_1} + 2\alpha_1\alpha_2\sigma_{L_1L_2} + \alpha_2^2\sigma^2_{L_2}]\sigma_{CD} \\ = & \sigma^2_A\sigma_{CD} \end{array}$$

which implies $\sigma_{CD} - \sigma_{AC}\sigma_{AD}(\sigma^2_A)^{-1} = 0 \Rightarrow \rho_{CD.A} = 0$. Contradiction. □

Lemma 9 *For any set* $\{A, B, C, D\} = \mathbf{O}' \subseteq \mathbf{O}$*, if* $\sigma_{AB}\sigma_{CD} = \sigma_{AC}\sigma_{BD} = \sigma_{AD}\sigma_{BD}$ *such that for every set* $\{X, Y\} \subset \mathbf{O}', Z \in \mathbf{O}$ *we have* $\rho_{XY.Z} \neq 0$ *and* $\rho_{XY} \neq 0$*, then if* A *and* B *have a common immediate latent ancestor* L_1 *in* G*,* B *and* C *have a common immediate latent ancestor* L_2 *in* G*, we have* $L_1 = L_2$*.*

Proof: Assume A, B and C are parameterized as follows:

$$\begin{array}{rcl} A & = & aL_1 + \sum_p a_p A_p \\ B & = & b_1 L_1 + b_2 L_2 + \sum_i b_i B_i \\ C & = & cL_2 + \sum_j c_j C_j \end{array}$$

where as before $\{A_p\} \cup \{B_i\} \cup \{C_j\}$ represents the possible other parents of A, B and C, respectively. Assume $L_1 \neq L_2$. We will show that $\rho_{L_1L_2} = 1$, which contradicts our assumptions. From the given constraint $\sigma_{AB}\sigma_{CD} = \sigma_{AD}\sigma_{BC}$, and the fact that from Lemma 2 we have that, for no pair $\{X, Y\} \subset \mathbf{O}'$, X is an ancestor of Y, if we factorize the constraint according to which terms include ab_1c as a factor, we obtain with probability 1:

$$ab_1c[\sigma^2_{L_1}\sigma_{L_2D} - \sigma_{L_1D}\sigma_{L_1L_2}] \tag{15}$$

If we factorize such constraint according to ab_2c, it follows:

$$ab_2c[\sigma_{L_1L_2}\sigma_{L_2D} - \sigma_{L_1D}\sigma^2_{L_2}] \tag{16}$$

From (15) and (16), it follows that $\sigma^2_{L_1}\sigma^2_{L_2} = (\sigma_{L_1L_2})^2 \Rightarrow \rho_{L_1L_2} = 1$. Contradiction. □

Lemma 10 *For any set* $\{A, B, C, D\} = \mathbf{O}' \subseteq \mathbf{O}$*, if* $\sigma_{AB}\sigma_{CD} = \sigma_{AC}\sigma_{BD} = \sigma_{AD}\sigma_{BD}$ *such that for every set* $\{X, Y\} \subset \mathbf{O}', Z \in \mathbf{O}$ *we have* $\rho_{XY.Z} \neq 0$ *and* $\rho_{XY} \neq 0$*, then if* A *and* B *have a common immediate latent ancestor* L_1 *in* G*,* C *and* D *have a common immediate latent ancestor* L_2 *in* G*, we have* $L_1 = L_2$*.*

Proof: Assume for the sake of contradiction that $L_1 \neq L_2$. Let P_A be a directed path from L_1 to A, and α_1 the sequence of edge labels in this path. Analogously, define α_2 as the sequence of edge labels from L_1 to B by some arbitrary path P_B, β_1 a sequence from L_2 to C according to some path P_C and β_2 a sequence from L_2 to D according to some path P_D.

P_A and P_B cannot intersect, since it would imply the existance of an observed common cause for A and B, which is ruled out by the given assumptions and Lemma 7. Similarly, no pair of paths in $\{P_A, P_B, P_C, P_D\}$ can intersect. By Lemma 9, L_1 cannot be an ancestor of either C or D, or otherwise $L_1 = L_2$. Analogously, L_2 cannot be an ancestor of either A or B.

By Lemma 2 and the given constraints, no element X in $\mathbf{O}'$ can be ancestor of an element in $\mathbf{O}'\backslash X$.

It means that when expanding the given constraint $\sigma_{AB}\sigma_{CD} - \sigma_{AD}\sigma_{BC} = 0$, and keeping all and only the terms that include the sequence of symbols $\alpha_1\alpha_2\beta_1\beta_2$, we obtain $\alpha_1\alpha_2\beta_1\beta_2\sigma^2_{L_1}\sigma^2_{L_2} - \alpha_1\alpha_2\beta_1\beta_2\sigma^2_{L_1L_2} = 0$, which implies $\rho_{L_1L_2} = 1$ with probability 1. Contradiction. □

Lemma 6 *Let $\mathbf{S} \subseteq \mathbf{O}$ be any set such that, for all $\{A, B, C\} \subseteq \mathbf{S}$, there is a fourth variable $D \in \mathbf{O}$ where i. $\sigma_{AB}\sigma_{CD} = \sigma_{AC}\sigma_{BD} = \sigma_{AD}\sigma_{BD}$ and ii. for every set $\{X, Y\} \subset \{A, B, C, D\}, Z \in \mathbf{O}$ we have $\rho_{XY.Z} \neq 0$ and $\rho_{XY} \neq 0$. Then $\mathbf{S}$ can be partioned into two sets $\mathbf{S_1}, \mathbf{S_2}$ where*

1. *all elements in $\mathbf{S_1}$ share a common immediate latent ancestor, and no two elements in $\mathbf{S_1}$ have any other common immediate latent ancestor;*
2. *no element $S \in \mathbf{S_2}$ has any common immediate latent ancestor with any other element in $\mathbf{S}\backslash S$*
3. *all elements in $\mathbf{S}$ are d-separated given the latents in G;*

Proof: Follows immediately from the given constraints and Lemmas 2, 9 and 10. □

Theorem 2 *If a partition $\{\mathbf{C_1}, \ldots, \mathbf{C_k}\}$ of $\mathbf{O}'$ respects structural conditions SC1, SC2 and SC3, then the following should hold in the true latent variable graph G that generated the data:*

1. *for all $X \in \mathbf{C_i}, Y \in \mathbf{C_j}, i \neq j$, X and Y have no common parents, and X is d-separated from the latent parents of Y given the latent parents of X;*
2. *for all $X, Y \in \mathbf{O}'$, X is d-separated from Y given the latent parents of X;*
3. *every set $\mathbf{C_i}$ can be partitioned into two groups according to Lemma 6;*

Proof: Follows immediately from the given constraints and Lemmas 1, 4, 5 and 6. □

Before showing the proof of Theorem 3, the next two lemmas will be useful:

Lemma 11 *Let set $\{A, B, C, D\} = \mathbf{O}' \subseteq \mathbf{O}$ be such that $\sigma_{AB}\sigma_{CD} = \sigma_{AC}\sigma_{BD} = \sigma_{AD}\sigma_{BD}$ and for every set $\{X, Y\} \subset \mathbf{O}', Z \in \mathbf{O}$ we have $\rho_{XY.Z} \neq 0$ and $\rho_{XY} \neq 0$. If an immediate latent ancestor L_X of $X \in \mathbf{O}'$ is uncorrelated with some immediate latent ancestor L_Y of $Y \in \mathbf{O}'$, then L_X is uncorrelated with all immediate latent ancestors of all elements in $\mathbf{O}'\backslash X$ or L_Y is uncorrelated with all immediate latent ancestors of all elements in $\mathbf{O}'\backslash Y$.*

Proof: Since the immediate latent ancestors of $\mathbf{O}'$ are linked to $\mathbf{O}'$ in that set by directed paths that do not intersect (Lemma 7) other than at the sources, and the model is linear below the latents, we can treat them as parents of $\mathbf{O}'$ without loss of generality. We will prove the lemma in two steps.

Step 1: let $X, Y \in \mathbf{O}'$. If a parent L_X of X is uncorrelated with all parents of Y, then L_X is uncorrelated with all parents of all elements in $\mathbf{O}'\backslash X$. To see this, without loss of generality let $A = aL_A + \sum_p a_pA_p$, and let L_A be uncorrelated with all parents of B. Let $C = cL_C + \sum_j c_jC_j$. This means that when expanding the polynomial $\sigma_{AB}\sigma_{CD} - \sigma_{AC}\sigma_{BD} = 0$, the only terms containing the symbol ac will be $ac\sigma_{L_AL_C}\sigma_{BD}$. Since $ac \neq 0, \sigma_{BD} \neq 0$, this will force $\sigma_{L_AL_C} = 0$ with probability 1. By symmetry, L_A will be uncorrelated with all parents of C and D.

Step 2: now we show the result stated by the lemma. Without loss of generality let $A = aL_A + \sum_p a_p A_p$, $B = bL_B + \sum_i b_i B_i$ and let L_A be uncorrelated with L_B. Then no term in the polynomial corresponding to $\sigma_{AB}\sigma_{CD}$ can contain a term with the symbol ab, since $\sigma_{L_A L_B} = 0$. If L_B is uncorrelated with all parents of D, then L_B is uncorrelated will all parents of all elements in $\mathbf{O}'\backslash B$, and we are done. Otherwise, assume L_B is correlated with at least one parent of D. Then at least one term in $\sigma_{AC}\sigma_{BD}$ will contain the symbol ab if there is some parent of C that is correlated with L_A (because σ_{BD} will contain some term with b). It follows that L_A has to be uncorrelated with every parent of D, and by the result in Step 1, with all parents of all elements in $\mathbf{O}'\backslash A$. □

Lemma 12 *Let set $\{A, B, C, D\} = \mathbf{O}' \subseteq \mathbf{O}$ be such that $\sigma_{AB}\sigma_{CD} = \sigma_{AC}\sigma_{BD} = \sigma_{AD}\sigma_{BD}$ and for every set $\{X, Y\} \subset \mathbf{O}', Z \in \mathbf{O}$ we have $\rho_{XY.Z} \neq 0$ and $\rho_{XY} \neq 0$. Let $\{A_p\}$ be the set of immediate latent ancestors of A, $\{B_i\}$ be the set of immediate latent ancestors of B, $\{C_j\}$ be the set of immediate latent ancestors of C, $\{D_k\}$ be the set of immediate latent ancestors of D. Then $\sigma_{A_pB_i}\sigma_{C_jD_k} = \sigma_{A_pC_j}\sigma_{B_iD_k} = \sigma_{A_pD_k}\sigma_{B_iC_j}$ for all $\{A_p, B_i, C_j, D_k\} \in \{A_p\} \times \{B_i\} \times \{C_j\} \times \{D_k\}$.*

Proof: Since the immediate latent ancestors of $\mathbf{O}'$ are linked to $\mathbf{O}'$ in that set by directed paths that do not intersect (Lemma 7) other than at the sources, and the model is linear below the latents, we can treat them as parents of $\mathbf{O}'$ without loss of generality. Let a_p be the coefficient linking A and A_p. Define b_i, c_j, d_k analogously. The lemma follows immediately by the same measure theoretical arguments of previous lemmas applied to the terms that include $a_p b_i c_j d_k$. □

Theorem 3 *Given a partition $\mathbf{C}$ of a subset $\mathbf{O}'$ of the observed variables of a latent variable graph G such that $\mathbf{C}$ satisfies structural constraints SC1-SC4, there is a linear latent variable model for the first two moments of $\mathbf{O}'$.*

Proof: We will assume that all elements of all sets in $\mathbf{C}$ are correlated. Otherwise, $\mathbf{C}$ can be partitioned into subsets with this property (because of the SC4 condition), and the parameterization given below can be applied independently to each member of the partition without loss of generality.

Let $\mathbf{An_i}$ be the set of immediate latent ancestors of the elements in $\mathbf{C_i} \in \mathbf{C} = \{\mathbf{C_1}, \ldots, \mathbf{C_k}\}$. Split every $\mathbf{An_i}$ into two disjoint sets $\mathbf{An_i^0}$ and $\mathbf{An_i^1}$, such that $\mathbf{An_i^0}$ contains all and only the those elements of $\mathbf{An_i^0}$ that are uncorrelated with all elements in $\mathbf{An_1} \cup \cdots \cup \mathbf{An_k}$. This implies that all elements in $\mathbf{An_1^1} \cup \cdots \cup \mathbf{An_k^1}$ are pairwise correlated by Lemma 11.

Construct the graph G^L_{linear} as follows. For each set $\mathbf{An_i}$, add a latent L_{An_i} to G^L_{linear}, as well as all elements of $\mathbf{An_i^1}$. Add a directed edge from L_{An_i} to each element in $\mathbf{An_i^1}$. Let G^L_{linear} be also a linear latent variable model. We will define values for each parameter in this model.

Fully connected all elements in $\{L_{An_i}\}$ as an arbitrary directed acyclic graph (DAG). Instead of defining the parameters for the edges and error variances in the subgraph of

G^L_{linear} induced by $\{L_{An_i}\}$, we will directly define a covariance matrix Σ_L among these nodes. Standard results in linear models can be used to translate this covariance matrix to the parameters of an arbitrary fully connected DAG (Spirtes et al., 2000). Set the diagonal of Σ_L to be 1.

Define the intercept parameters μ_x of all elements in G^L_{linear} to be zero. For each V in $\mathbf{An^1_i}$ we have a set of parameters for the local equations $V = \lambda_V L_{An_i} + \epsilon_V$, where ϵ_V is a random variable with zero mean and variance ζ_V.

Choose any three arbitrary elements $\{X, Y, Z\} \subseteq \mathbf{An^1_i}$. Since the subgraph $L_{An_i} \rightarrow X, L_{An_i} \rightarrow Y, L_{An_i} \rightarrow Z$ has six parameters $(\lambda_X, \lambda_Y, \lambda_Z, \zeta_X, \zeta_Y, \zeta_Z)$ and the population covariance matrix of X, Y and Z has six entries, these parameters can be assigned an unique value (Bollen, 1989) such that $\sigma_{XY} = \lambda_X\lambda_Y$ and $\zeta_X = \lambda^2_X - \sigma^2_X$. Let W be any other element of $\mathbf{An^1_i}$: set $\lambda_W = \sigma_{WX}/\lambda_X$, $\zeta_W = \sigma^2_W - \lambda^2_W$. From Lemma 12, we have the constraint $\sigma_{WY}\sigma_{XZ} - \sigma_{WX}\sigma_{YZ} = 0$, from which one can verify that $\sigma_{WY} = \lambda_W\lambda_Y$ does hold in the population. By symmetry and induction, for every pair P, Q in $\mathbf{An^1_i}$, we have $\sigma_{PQ} = \lambda_P\lambda_Q$.

Let T be some element in $\mathbf{An^1_j}$, $i \neq j$: set the entry σ_{ij} of Σ_L to be $\sigma_{TX}/(\lambda_T\lambda_X)$. Let R and S be another elements in $\mathbf{An^1_j}$. From Lemma 12, we have the constraint $\sigma_{XT}\sigma_{RS} - \sigma_{XR}\sigma_{ST} = 0$, from which one can verify that $\sigma_{XR} = \lambda_X\lambda_R\sigma_{ij}$. Let Y and Z be another elements in $\mathbf{An^1_i}$. From Lemma 12, we have the constraint $\sigma_{XT}\sigma_{YZ} - \sigma_{XY}\sigma_{ZT} = 0$ from which one can verify that $\sigma_{ZT} = \lambda_Z\lambda_t\sigma_{ij}$. By symmetry and induction, for every pair P, Q in $\mathbf{An^1_i} \times \mathbf{An^1_j}$, we have $\sigma_{PQ} = \lambda_P\lambda_Q\sigma_{ij}$.

Finally, let G_{linear} be a graph constructed as follows:

1. start G_{linear} with a node for each element in $\mathbf{O'}$;

2. for each $\mathbf{C_i} \in \mathbf{C}$, add a latent L_i to G, and for each $V \in \mathbf{C_i}$, add an edge $L_i \rightarrow V$

3. fully connect the latents in G_{linear} to form an arbitrary directed acyclic graph

Parameterize a linear latent model based on G as follows: let $V \in \mathbf{C_i}$ such that V has immediate latent ancestors $\{L_{V_i}\}$. In the true model, let $V = \mu^G_V + \Sigma_i\lambda^G_{iV}L_{V_i} + \epsilon^G_V$, where every latent has zero mean. Construct the equation $V = \mu_V + \lambda_V L_i + \epsilon_V$ by instantiating $\mu_V = \mu^G_V$ and $\lambda_V = \Sigma_i\lambda^G_{iV}\lambda_{L_{V_i}}$, where $\lambda_{L_{V_i}}$ is the respective parameter for L_{V_i} in G^L_{linear} if $L_{V_i} \in \mathbf{An^1_i}$, and 0 otherwise. The variance for ϵ_V is defined as $\sigma^2_V - \lambda^2_V$. The L_i variables have covariance matrix Σ_L as defined above. One can then verify that the covariance matrix generated by this model equals the true covariance matrix of $\mathbf{O'}$. □