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

Ricardo Silva RBAS@CS.CMU.EDU
Richard Scheines SCHEINES@ANDREW.CMU.EDU
CALD, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213 USA

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.

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 & 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.

Appearing in *Proceedings of the 22st International Conference on Machine Learning*, Bonn, Germany, 2005.

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 & 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. Our paper builds on this approach by removing such assumptions. 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 & 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

[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.

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.

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. For covariances, we will use the symbol σ_{XY}.

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 (Silva & Scheines, 2005).

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 sub-partition.

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.

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.

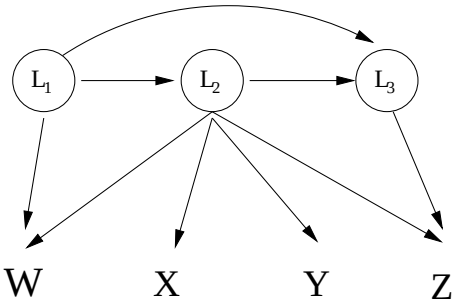

Figure 1. It is possible that $\rho_{L_1 L_3 . L_2} \neq 0$ even though L_2 does not d-separate L_1 and L_3. That happens, for instance, if $L_2 = \lambda_1 L_1 + \epsilon_2$, $L_3 = \lambda_2 L_1^2 + \lambda_3 L_2 + \epsilon_3$, where L_1, ϵ_2 and ϵ_3 are normally distributed with zero mean.

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 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.

[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 $\mathbf{C_i} = \mathbf{C_j}$.

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 a structure.

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 & 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).

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 estab-

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

lished 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 (Bach & 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

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

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

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

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

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

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

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

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

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

Silva, R., & Scheines, R. (2005). New d-separation identification results for learning continuous latent variable models. *Technical Report CMU-CALD-05-104, Carnegie Mellon University*.

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

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

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

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